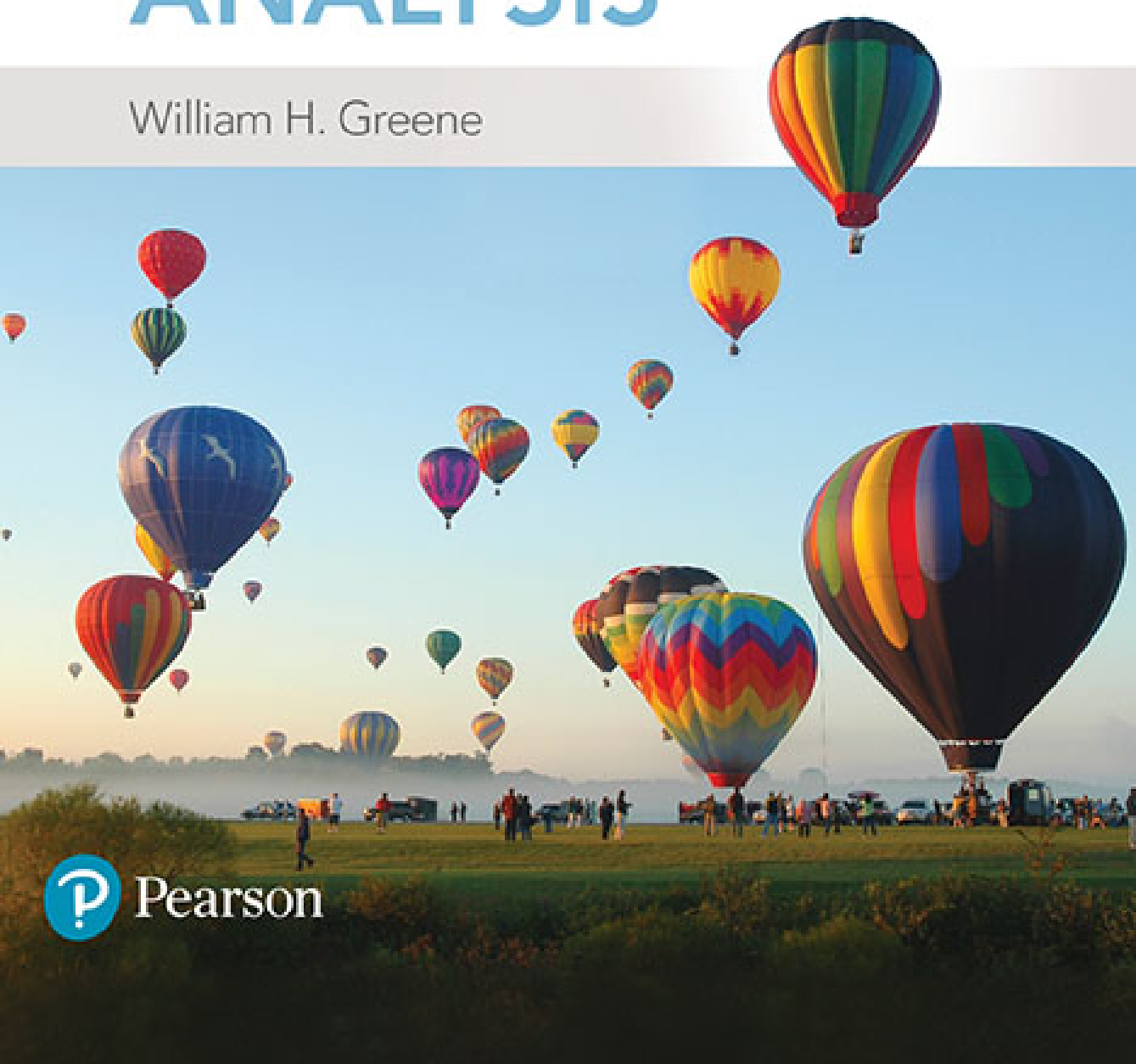
EIGHTH EDITION

ECONOMETRIC ANALYSIS

William H. Greene



Pearson

Percentiles of the Chi-Squared Distribution. Table Entry Is c Such That $\text{Prob}[\chi^2_n \le c] = P$

n	*.005*	*.010*	*.025*	*.050*	*.100*	*.250*	*.500*	*.750*	*.900*	*.950*	*.975*	*.990*	*.995*
1	.00004	.0002	.001	.004	.02	.10	.45	1.32	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	38.29	44.34	50.98	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49

EIGHTH EDITION

ECONOMETRIC ANALYSIS

William H. Greene

The Stern School of Business

New York University

New York, NY

For Margaret and Richard Greene

Vice President, Business Publishing: Donna Battista
Director of Portfolio Management: Adrienne D' Ambrosio
Director, Courseware Portfolio Management: Ashley Dodge
Senior Sponsoring Editor: Neeraj Bhalla
Editorial Assistant: Courtney Paganelli
Vice President, Product Marketing: Roxanne McCarley
Director of Strategic Marketing: Brad Parkins
Strategic Marketing Manager: Deborah Strickland
Product Marketer: Tricia Murphy
Field Marketing Manager: Ramona Elmer
Product Marketing Assistant: Jessica Quazza
Vice President, Production and Digital Studio, Arts and Business: Etain O'Dea
Director of Production, Business: Jeff Holcomb
Managing Producer, Business: Alison Kalil
Content Producer: Sugandh Juneja
Operations Specialist: Carol Melville
Creative Director: Blair Brown
Manager, Learning Tools: Brian Surette
Content Developer, Learning Tools: Lindsey Sloan
Managing Producer, Digital Studio, Arts and Business: Diane Lombardo
Digital Studio Producer: Melissa Honig
Digital Studio Producer: Alana Coles
Digital Content Team Lead: Noel Lotz
Digital Content Project Lead: Courtney Kamauf
Full-Service Project Management and Composition: SPi Global
Interior Design: SPi Global
Cover Design: SPi Global
Cover Art: Jim Lozouski/Shutterstock
Printer/Binder: RRD Crawfordsville
Cover Printer: Phoenix/Hagerstown

Library of Congress Cataloging-in-Publication Data on File

1 17

Pearson

ISBN 10: 0-13-446136-3
ISBN 13: 978-0-13-446136-6

BRIEF CONTENTS

Contents

Part II Generalized Regression Model and Equation Systems

EXAMPLES AND APPLICATIONS

PREFACE

ECONOMETRIC ANALYSIS

Econometric Analysis is a broad introduction to the field of econometrics. This field grows continually. A (not complete) list of journals devoted at least in part to econometrics now includes: *Econometric Reviews; Econometric Theory; Econometrica; Econometrics; Econometrics and Statistics; The Econometrics Journal; Empirical Economics; Foundations and Trends in Econometrics; The Journal of Applied Econometrics; The Journal of Business and Economic Statistics; The Journal of Choice Modelling; The Journal of Econometric Methods; The Journal of Econometrics; The Journal of Time Series Analysis; The Review of Economics and Statistics*. Constructing a textbook-style survey to introduce the topic at a graduate level has become increasingly ambitious. Nonetheless, that is what I seek to do here. This text attempts to present, at an entry graduate level, enough of the topics in econometrics that a student can comfortably move on from here to practice or to more advanced study. For example, the literature on "Treatment Effects" is already vast, rapidly growing, complex in the extreme, and occasionally even contradictory. But, there are a few bedrock principles presented in Chapter 8 that (I hope) can help the interested practitioner or student get started as they wade into this segment of the literature. The book is intended as a bridge between an introduction to econometrics and the professional literature.

The book has two objectives. The first is to introduce students to *applied econometrics*, including basic techniques in linear regression analysis and some of the rich variety of models that are used when the linear model proves inadequate or inappropriate. Modern software has made complicated modeling very easy to put into practice. The second objective is to present sufficient *theoretical background* so that the reader will (1) understand the advanced techniques that are made so simple in modern software and (2) recognize new variants of the models learned about here as merely natural extensions that fit within a common body of principles. This book contains a substantial amount of theoretical material, such as that on the GMM, maximum likelihood estimation, and asymptotic results for regression models.

One overriding purpose has motivated all eight editions of *Econometric Analysis*. The vast majority of readers of this book will be users, not developers, of econometrics. I believe that it is not sufficient to teach econometrics by reciting (and proving) the theories of estimation and inference. Although the often-subtle theory is extremely important, the application is equally crucial. To that end, I have provided hundreds of worked numerical examples and extracts from applications in the received empirical literature in many fields. My purpose in writing this work, and in my continuing efforts to update it, is to show readers how to *do* econometric analysis. But, I also believe that readers want (and need) to know what is going on behind the curtain when they use ever more sophisticated modern software for ever more complex econometric analyses.

I have taught econometrics at the level of *Econometric Analysis* at NYU for many years. I ask my students to learn how to use a (any) modern econometrics program as part of their study. I've lost track of the number of my students who recount to me their disappointment in a previous course in which they were taught how to use software, but not the theory and motivation of the techniques. In October, 2014, Google Scholar published its list of the 100 most cited works over all fields and all time. (www.nature.com/polopoly_fs/7.21245!/file/GoogleScholartop100.xlsx). *Econometric Analysis*, the only work in econometrics on the list, ranked number 34 with 48,100 citations. (As of this writing, November 2016, the number of citations to the first 7 editions in all languages approaches 60,000.) I take this extremely gratifying result as evidence that there are readers in many fields who agree that the practice of econometrics calls for an understanding of *why*, as well as *how* to use the tools in modern software. This book is for them.

THE EIGHTH EDITION OF *ECONOMETRIC ANALYSIS*

This text is intended for a one-year graduate course for social scientists. Prerequisites should include calculus, mathematical statistics, and an introduction to econometrics at the level of, say, Gujarati and Porter's (2011) *Basic Econometrics*, Stock and Watson's (2014) *Introduction to Econometrics*, Kennedy's (2008) *Guide to Econometrics*, or Wooldridge's (2015) *Introductory Econometrics: A Modern Approach*. I assume, for example, that the reader has already learned about the basics of econometric methodology including the fundamental role of economic and statistical assumptions; the distinctions between cross-section, time-series, and panel data sets; and the essential ingredients of estimation, inference, and prediction with the multiple linear regression model. Self-contained (for our purposes) summaries of the matrix algebra, mathematical statistics, and statistical theory used throughout the book are given in Appendices A through D. I rely heavily on matrix algebra throughout. This may be a bit daunting to some early on but matrix algebra is an indispensable tool and I hope the reader will come to agree that it is a means to an end, not an end in itself. With matrices, the unity of a variety of results will emerge without being obscured by a curtain of summation signs. Appendix E and Chapter 15 contain a description of numerical methods that will be useful to practicing econometricians (and to us in the later chapters of the book).

Estimation of advanced nonlinear models is now as routine as least squares. I have included five chapters on estimation methods used in current research and five chapters on applications in micro- and macroeconometrics. The nonlinear models used in these fields are now the staples of the applied econometrics literature. As a consequence, this book also contains a fair amount of material that will extend beyond many first courses in econometrics. Once again, I have included this in the hope of laying a foundation for study of the professional literature in these areas.

PLAN OF THE BOOK

The arrangement of the book is as follows:

Part I begins the formal development of econometrics with its fundamental pillar, the *linear multiple regression model*. Estimation and inference with the linear least squares estimator are analyzed in Chapters 2 through 6. The *nonlinear regression model* is introduced

in Chapter 7 along with quantile, semi- and nonparametric regression, all as extensions of the familiar linear model. *Instrumental variables estimation* is developed in Chapter 8.

Part II presents three major extensions of the regression model. Chapter 9 presents the consequences of relaxing one of the main assumptions of the linear model, homoscedastic nonautocorrelated disturbances, to introduce the *generalized regression model*. The focus here is on heteroscedasticity; autocorrelation is mentioned, but a detailed treatment is deferred to Chapter 20 in the context of time-series data. Chapter 10 introduces systems of regression equations, in principle, as the approach to modeling simultaneously a set of random variables and, in practical terms, as an extension of the generalized linear regression model. Finally, *panel data methods*, primarily fixed and random effects models of heterogeneity, are presented in Chapter 11.

The second half of the book is devoted to topics that extend the linear regression model in many directions. Beginning with Chapter 12, we proceed to the more involved methods of analysis that contemporary researchers use in analysis of "real-world" data. Chapters 12 to 16 in Part III present different estimation methodologies. Chapter 12 presents an overview by making the distinctions between *parametric*, *semiparametric* and *nonparametric methods*. The leading application of semiparametric estimation in the current literature is the *generalized method of moments (GMM) estimator* presented in Chapter 13. This technique provides the platform for much of modern econometrics. *Maximum likelihood estimation* is developed in Chapter 14. *Monte Carlo* and *simulation-based methods* such as *bootstrapping* that have become a major component of current research are developed in Chapter 15. Finally, *Bayesian methods* are introduced in Chapter 16.

Parts IV and V develop two major subfields of econometric methods, *microeconometrics*, which is typically based on cross-section and panel data, and *macroeconometrics*, which is usually associated with analysis of time-series data. In Part IV, Chapters 17 to 19 are concerned with models of *discrete choice, censoring, truncation, sample selection, duration* and the analysis of *counts of events*. In Part V, Chapters 20 and 21, we consider two topics in time-series analysis, models of *serial correlation* and regression models for *nonstationary data*—the usual substance of macroeconomic analysis.

REVISIONS

With only a couple exceptions noted below, I have retained the broad outline of the text. I have revised the presentation throughout the book (including this preface) to streamline the development of topics, in some cases (I hope), to improve the clarity of the derivations. Major revisions include:

- I have moved the material related to "causal inference" forward to the early chapters of the book – these topics are now taught earlier in the graduate sequence than heretofore and I've placed them in the context of the models and methods where they appear rather than as separate topics in the more advanced sections of the seventh edition. Difference in difference regression as a method, and regression discontinuity designs now appear in Chapter 6 with the discussion of functional forms and in the context of extensive applications extracted from the literature. The analysis of treatment effects has all been moved from Chapter 19 (on censoring and truncation) to Chapter 8 on endogeneity under the heading of "Endogenous

Dummy Variables." Chapter 8, as a whole, now includes a much more detailed treatment of instrumental variable methods.

- I have added many new examples, some as extracts from applications in the received literature, and others as worked numerical examples. I have drawn applications from many different fields including industrial organization, transportation, health economics, popular culture and sports, urban development and labor economics.
- Chapter 10 on systems of equations has been shifted (yet further) from its early emphasis on formal simultaneous linear equations models to systems of regression equations and the leading application, the single endogenous variable in a two equation recursive model – this is the implicit form of the regression model that contains one "endogenous" variable.
- The use of robust estimation and inference methods has been woven more extensively into the general methodology, in practice and throughout this text. The ideas of robust estimation and inference are introduced immediately with the linear regression model in Chapters 4 and 5, rather than as accommodations to nonspherical disturbances in Chapter 9. The role that a robust variance estimator will play in the Wald statistic is developed immediately when the result is first presented in Chapter 5.
- Chapters 4 (Least Squares), 6 (Functional Forms), 8 (Endogeneity), 10 (Equation Systems) and 11 (Panel Data) have been heavily revised to emphasize both contemporary econometric methods and the applications.
- I have moved Appendices A-F to the Companion Web site, at www.pearsonhighered.com/greene, that accompanies this text. Students can access them at no cost.

The first semester of study in a course based on Econometric Analysis would focus on Chapters 1-6 (the linear regression model), 8 (endogeneity and causal modeling), and possibly some of 11 (panel data). Most of the revisions in the eighth edition appear in these chapters.

SOFTWARE AND DATA

There are many computer programs that are widely used for the computations described in this book. All were written by econometricians or statisticians, and in general, all are regularly updated to incorporate new developments in applied econometrics. A sampling of the most widely used packages and Web sites where you can find information about them are

EViews	www.eviews.com	(QMS, Irvine, CA)
Gauss	www.aptech.com	(Aptech Systems, Kent, WA)
LIMDEP	www.limdep.com	(Econometric Software, Plainview, NY)
MATLAB	www.mathworks.com	(Mathworks, Natick, MA)
NLOGIT	www.nlogit.com	(Econometric Software, Plainview, NY)
R	www.r-project.org/	(The R Project for Statistical Computing)
RATS	www.estima.com	(Estima, Evanston, IL)
SAS	www.sas.com	(SAS, Cary, NC)
Shazam	econometrics.com	(Northwest Econometrics Ltd., Gibsons, Canada)
Stata	www.stata.com	(Stata, College Station, TX)

A more extensive list of computer software used for econometric analysis can be found at the resource Web site, http://www.oswego.edu/~economic/econsoftware.htm.

With only a few exceptions, the computations described in this book can be carried out with any of the packages listed. *NLOGIT* was used for the computations in most of the applications. This text contains no instruction on using any particular program or language. Many authors have produced *RATS*, *LIMDEP/NLOGIT*, *EViews*, *SAS*, or *Stata* code for some of the applications, including, in a few cases, in the documentation for their computer programs. There are also quite a few volumes now specifically devoted to econometrics associated with particular packages, such as Cameron and Trivedi's (2009) companion to their treatise on microeconometrics.

The data sets used in the examples are also available on the Web site for the text, http://people.stern.nyu.edu/wgreene/Text/econometricanalysis.htm. Throughout the text, these data sets are referred to "Table Fn.m," for example Table F4.1. The "F" refers to Appendix F available on the Companion web site which contains descriptions of the data sets. The actual data are posted in generic ASCII and portable formats on the Web site with the other supplementary materials for the text. There are now thousands of interesting Web sites containing software, data sets, papers, and commentary on econometrics. It would be hopeless to attempt any kind of a survey. One code/data site that is particularly agreeably structured and well targeted for readers of this book is the data archive for the *Journal of Applied Econometrics* (*JAE*). They have archived all the nonconfidential data sets used in their publications since 1988 (with some gaps before 1995). This useful site can be found at http://qed.econ.queensu.ca/jae/. Several of the examples in the text use the *JAE* data sets. Where we have done so, we direct the reader to the *JAE*'s Web site, rather than our own, for replication. Other journals have begun to ask their authors to provide code and data to encourage replication. Another easy-to-navigate site for aggregate data on the U.S. economy is https://datahub.io/dataset/economagic.

ACKNOWLEDGMENTS

It is a pleasure to express my appreciation to those who have influenced this work. I remain grateful to Arthur Goldberger (dec.), Arnold Zellner (dec.), Dennis Aigner, Bill Becker, and Laurits Christensen for their encouragement and guidance. After eight editions of this book, the number of individuals who have significantly improved it through their comments, criticisms, and encouragement has become far too large for me to thank each of them individually. I am grateful for their help and I hope that all of them see their contribution to this edition. Any number of people have submitted tips about the text. You can find many of them listed in the errata pages on the text Web site, http://people.stern.nyu.edu/wgreene/Text/econometricanalysis.htm, in particular: David Hoaglin, University of Massachusetts; Randall Campbell, Mississippi State University; Carter Hill, Louisiana State University; and Tom Doan, Estima Corp. I would also like to thank two colleagues who have worked on translations of *Econometric Analysis*, Marina Turuntseva (the Russian edition) and Umit Senesen (the Turkish translation). I must also acknowledge the mail I've received from hundreds of readers and practitioners from the world over who have given me a view into topics and questions that practitioners are interested in, and have provided a vast trove of helpful material for my econometrics courses.

I also acknowledge the many reviewers of my work whose careful reading has vastly improved the book through this edition: Scott Atkinson, University of Georgia; Badi Baltagi, Syracuse University; Neal Beck, New York University; William E. Becker (Ret.), Indiana University; Eric J. Belasko, Texas Tech University; Anil Bera, University of Illinois; John Burkett, University of Rhode Island; Leonard Carlson, Emory University; Frank Chaloupka, University of Illinois at Chicago; Chris Cornwell, University of Georgia; Craig Depken II, University of Texas at Arlington; Frank Diebold, University of Pennsylvania; Edward Dwyer, Clemson University; Michael Ellis, Wesleyan University; Martin Evans, Georgetown University; Vahagn Galstyan, Trinity College Dublin; Paul Glewwe, University of Minnesota; Ed Greenberg, Washington University at St. Louis; Miguel Herce, University of North Carolina; Joseph Hilbe, Arizona State University; Dr. Uwe Jensen, Christian-Albrecht University; K. Rao Kadiyala, Purdue University; William Lott, University of Connecticut; Thomas L. Marsh, Washington State University; Edward Mathis, Villanova University; Mary McGarvey, University of Nebraska–Lincoln; Ed Melnick, New York University; Thad Mirer, State University of New York at Albany; Cyril Pasche, University of Geneva; Paul Ruud, University of California at Berkeley; Sherrie Rhine, Federal Deposit Insurance Corp.; Terry G. Seaks (Ret.), University of North Carolina at Greensboro; Donald Snyder, California State University at Los Angeles; Steven Stern, University of Virginia; Houston Stokes, University of Illinois at Chicago; Dmitrios Thomakos, Columbia University; Paul Wachtel, New York University; Mary Beth Walker, Georgia State University; Mark Watson, Harvard University; and Kenneth West, University of Wisconsin. My numerous discussions with Bruce McCullough of Drexel University have improved Appendix E and at the same time increased my appreciation for numerical analysis. I am especially grateful to Jan Kiviet of the University of Amsterdam, who subjected my third edition to a microscopic examination and provided literally scores of suggestions, virtually all of which appear herein. Professor Pedro Bacao, University of Coimbra, Portugal, and Mark Strahan of Sand Hill Econometrics and Umit Senesen of Istanbul Technical University did likewise with the sixth and seventh editions.

I would also like to thank the many people at Pearson Education who have put this book together with me: Adrienne D' Ambrosio, Neeraj Bhalla, Sugandh Juneja, and Nicole Suddeth and the composition team at SPi Global.

For over 25 years since the first edition, I've enjoyed the generous support and encouragement of many people, some close to me, especially my family, and many not so close. I'm especially grateful for the help, support and priceless encouragement of my wife, Sherrie Rhine, whose unending enthusiasm for this project has made it much less daunting, and much more fun.

William H. Greene
February 2017

1
ECONOMETRICS

1.1 INTRODUCTION

This book will present an introductory survey of econometrics. We will discuss the fundamental ideas that define the methodology and examine a large number of specific models, tools, and methods that econometricians use in analyzing data. This chapter will introduce the central ideas that are the paradigm of econometrics. Section 1.2 defines the field and notes the role that theory plays in motivating econometric practice. Sections 1.3 and 1.4 discuss the types of applications that are the focus of econometric analyses. The process of econometric modeling is presented in Section 1.5 with a classic application, Keynes's consumption function. A broad outline of the text is presented in Section 1.6. Section 1.7 notes some specific aspects of the presentation, including the use of numerical examples and the mathematical notation that will be used throughout the text.

1.2 THE PARADIGM OF ECONOMETRICS

> In the first issue of *Econometrica*, the Econometric Society stated that its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking similar to that which has come to dominate the natural sciences. . . . But there are several aspects of the quantitative approach to economics, and no single one of these aspects taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonomous [sic] with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.

The Society responded to an unprecedented accumulation of statistical information. It saw a need to establish a body of principles that could organize what would otherwise become a bewildering mass of data. Neither the pillars nor the objectives of econometrics have changed in the years since this editorial appeared. Econometrics concerns itself with the

application of mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by an underlying theory.

It is interesting to observe the response to a contemporary, likewise unprecedented accumulation of massive amounts of quantitative information in the form of "Big Data." Consider the following assessment of what Kitchin (2014) sees as a paradigm shift in the analysis of data.

> This article examines how the availability of Big Data, coupled with new data analytics, challenges established epistemologies across the sciences, social sciences and humanities, and assesses the extent to which they are engendering paradigm shifts across multiple disciplines. In particular, it critically explores new forms of empiricism that declare 'the end of theory,' the creation of data-driven rather than knowledge-driven science, and the development of digital humanities and computational social sciences that propose radically different ways to make sense of culture, history, economy and society. It is argued that: (1) Big Data and new data analytics are disruptive innovations which are reconfiguring in many instances how research is conducted; and (2) there is an urgent need for wider critical reflection within the academy on the epistemological implications of the unfolding data revolution, a task that has barely begun to be tackled despite the rapid changes in research practices presently taking place.

We note the suggestion that data-driven analytics are proposed to replace theory (and econometrics as envisioned by Frisch) for providing the organizing principles to guide empirical research. (We will examine an example in Chapter 18 where we consider analyzing survey data with ordered choice models. Also, see Varian (2014) for a more balanced view.) The focus is driven partly by the startling computational power that would have been unavailable to Frisch. It seems likely that the success of this new paradigm will turn at least partly on the questions pursued. Whether the interesting features of an underlying data-generating process can be revealed by appealing to the data themselves without a theoretical platform seems to be a prospect raised by the author. The article does focus on the role of an underlying theory in empirical research—this is a central pillar of econometric methodology. As of this writing, the success story of Big Data analysis is still being written.

The crucial role that econometrics plays in economics has grown over time. The Nobel Prize in Economics has recognized this contribution with numerous awards to econometricians, including the first which was given to (the same) Ragnar Frisch in 1969. Lawrence Klein in 1980, Trygve Haavelmo in 1989, James Heckman and Daniel McFadden in 2000, Robert Engle and Clive Granger in 2003. Christopher Sims in 2011 and Lars Hansen in 2013 were recognized for their empirical research. The 2000 prize was noteworthy in that it celebrated the work of two scientists whose research was devoted to the marriage of behavioral theory and econometric modeling.

Example 1.1 Behavioral Models and the Nobel Laureates

The pioneering work by both James Heckman and Dan McFadden rests firmly on a theoretical foundation of utility maximization.

For Heckman's, we begin with the standard theory of household utility maximization over consumption and leisure. The textbook model of utility maximization produces a demand for leisure time that translates into a supply function of labor. When home production (i.e., work

in the home as opposed to the outside, formal labor market) is considered in the calculus, then desired *hours* of (formal) labor can be negative. An important conditioning variable is the *reservation wage*—the wage rate that will induce formal labor market participation. On the demand side of the labor market, we have firms that offer market wages that respond to such attributes as age, education, and experience. What can we learn about labor supply behavior based on observed market wages, these attributes, and observed hours in the formal market? Less than it might seem, intuitively because our observed data omit half the market—the data on formal labor market activity are not randomly drawn from the whole population.

Heckman's observations about this implicit truncation of the distribution of hours or wages revolutionized the analysis of labor markets. Parallel interpretations have since guided analyses in every area of the social sciences. The analysis of policy interventions such as education initiatives, job training and employment policies, health insurance programs, market creation, financial regulation, and a host of others is heavily influenced by Heckman's pioneering idea that when participation is part of the behavior being studied, the analyst must be cognizant of the impact of common influences in both the presence of the intervention and the outcome. We will visit the literature on sample selection and treatment/program evaluation in Chapters 5, 6, 8 and 19.

Textbook presentations of the theories of demand for goods that produce utility, because they deal in continuous variables, are conspicuously silent on the kinds of discrete choices that consumers make every day—what brand of product to choose, whether to buy a large commodity such as a car or a refrigerator, how to travel to work, whether to rent or buy a home, where to live, what candidate to vote for, and so on. Nonetheless, a model of *random utility* defined over the alternatives available to the consumer provides a theoretically sound platform for studying such choices. Important variables include, as always, income and relative prices. What can we learn about underlying preference structures from the discrete choices that consumers make? What must be assumed about these preferences to allow this kind of inference? What kinds of statistical models will allow us to draw inferences about preferences? McFadden's work on how commuters choose to travel to work, and on the underlying theory appropriate to this kind of modeling, has guided empirical research in discrete consumer choices for several decades. We will examine McFadden's models of discrete choice in Chapter 18.

1.3 THE PRACTICE OF ECONOMETRICS

We can make a useful distinction between *theoretical econometrics* and *applied econometrics*. Theorists develop new techniques for estimation and hypothesis testing and analyze the consequences of applying particular methods when the assumptions that justify those methods are not met. Applied econometricians are the users of these techniques and the analysts of data (real world and simulated). The distinction is far from sharp; practitioners routinely develop new analytical tools for the purposes of the study that they are involved in. This text contains a large amount of econometric theory, but it is directed toward applied econometrics. We have attempted to survey techniques, admittedly some quite elaborate and intricate, that have seen wide use in the field.

Applied econometric methods will be used for estimation of important quantities, analysis of economic outcomes such as policy changes, markets or individual behavior, testing theories, and for forecasting. The last of these is an art and science in itself that is the subject of a vast library of sources. Although we will briefly discuss some

aspects of forecasting, our interest in this text will be on estimation and analysis of models. The presentation, where there is a distinction to be made, will contain a blend of microeconometric and macroeconometric techniques and applications. It is also necessary to distinguish between *time-series analysis* (which is not our focus) and methods that primarily use *time-series data*. The former is, like forecasting, a growth industry served by its own literature in many fields. While we will employ some of the techniques of time-series analysis, we will spend relatively little time developing first principles.

1.4 MICROECONOMETRICS AND MACROECONOMETRICS

The connection between underlying behavioral models and the modern practice of econometrics is increasingly strong. Another distinction is made between *microeconometrics* and *macroeconometrics*. The former is characterized by its analysis of cross section and panel data and by its focus on individual consumers, firms, and micro-level decision makers. Practitioners rely heavily on the theoretical tools of microeconomics including utility maximization, profit maximization, and market equilibrium. The analyses are directed at subtle, difficult questions that often require intricate formulations. A few applications are as follows:

- What are the likely effects on labor supply behavior of proposed negative income taxes? [Ashenfelter and Heckman (1974)]
- Does attending an elite college bring an expected payoff in expected lifetime income sufficient to justify the higher tuition? [Kreuger and Dale (1999) and Kreuger (2000)]
- Does a voluntary training program produce tangible benefits? Can these benefits be accurately measured? [Angrist (2001)]
- Does an increase in the minimum wage lead to reduced employment? [Card and Krueger (1994)]
- Do smaller class sizes bring real benefits in student performance? [Hanuschek (1999), Hoxby (2000), and Angrist and Lavy (1999)]
- Does the presence of health insurance induce individuals to make heavier use of the health care system—is moral hazard a measurable problem? [Riphahn et al. (2003)]
- Did the intervention addressing anticompetitive behavior of a group of 50 boarding schools by the UK Office of Fair Trading produce a measurable impact on fees charged? [Pesaresi, Flanagan, Scott, and Tragear (2015)]

Macroeconometrics is involved in the analysis of time-series data, usually of broad aggregates such as price levels, the money supply, exchange rates, output, investment, economic growth, and so on. The boundaries are not sharp. For example, an application that we will examine in this text concerns spending patterns of municipalities, which rests somewhere between the two fields. The very large field of financial econometrics is concerned with long time-series data and occasionally vast panel data sets, but with a sharply focused orientation toward models of individual behavior. The analysis of market returns and exchange rate behavior is neither exclusively macro- nor microeconometric. [We will not be spending any time in this text on financial econometrics. For those with an interest in this field, We would recommend the celebrated work by Campbell, Lo, and Mackinlay (1997), or for a more time-series–oriented approach, Tsay (2005).]

Macroeconomic model builders rely on the interactions between economic agents and policy makers. For example:

- Does a monetary policy regime that is strongly oriented toward controlling inflation impose a real cost in terms of lost output on the U.S. economy? [Cecchetti and Rich (2001)]
- Did 2001's largest federal tax cut in U.S. history contribute to or dampen the concurrent recession? Or was it irrelevant?

Each of these analyses would depart from a formal model of the process underlying the observed data.

The techniques used in econometrics have been employed in a widening variety of fields, including political methodology, sociology,[1] health economics, medical research (e.g., how do we handle attrition from medical treatment studies?) environmental economics, economic geography, transportation engineering, and numerous others. Practitioners in these fields and many more are all heavy users of the techniques described in this text.

1.5 ECONOMETRIC MODELING

Econometric analysis usually begins with a statement of a theoretical proposition. Consider, for example, a classic application by one of Frisch's contemporaries:

Example 1.2 Keynes's Consumption Function

From Keynes's (1936) *General Theory of Employment, Interest and Money*:

We shall therefore define what we shall call the propensity to consume as the functional relationship f between X, a given level of income, and C, the expenditure on consumption out of the level of income, so that $C = f(X)$.

The amount that the community spends on consumption depends (i) partly on the amount of its income, (ii) partly on other objective attendant circumstances, and (iii) partly on the subjective needs and the psychological propensities and habits of the individuals composing it. The fundamental psychological law upon which we are entitled to depend with great confidence, both a priori from our knowledge of human nature and from the detailed facts of experience, is that men are disposed, as a rule and on the average, to increase their consumption as their income increases, but not by as much as the increase in their income. That is, . . . dC/dX is positive and less than unity.

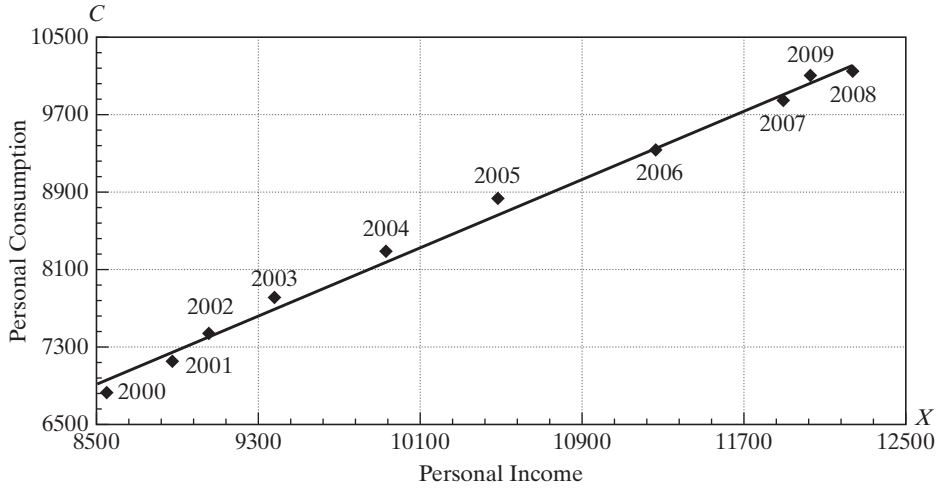
But, apart from short period changes in the level of income, it is also obvious that a higher absolute level of income will tend as a rule to widen the gap between income and consumption. . . . These reasons will lead, as a rule, to a greater proportion of income being saved as real income increases.

The theory asserts a relationship between consumption and income, $C = f(X)$, and claims in the second paragraph that the marginal propensity to consume (MPC), dC/dX, is between zero and one.[2] The final paragraph asserts that the average propensity to consume (APC), C/X, falls as income rises, or $d(C/X)/dX = (\text{MPC} - \text{APC})/X < 0$. It follows that MPC $<$ APC. The

[1] See, for example, Long (1997) and DeMaris (2004).

[2] Modern economists are rarely this confident about their theories. More contemporary applications generally begin from first principles and behavioral axioms, rather than simple observation.

FIGURE 1.1 Aggregate U.S. Consumption and Income Data, 2000–2009.



most common formulation of the consumption function is a linear relationship, $C = \alpha + X\beta$, that satisfies Keynes's "laws" if β lies between zero and one and if α is greater than zero.

These theoretical propositions provide the basis for an econometric study. Given an appropriate data set, we could investigate whether the theory appears to be consistent with the observed "facts." For example, we could see whether the linear specification appears to be a satisfactory description of the relationship between consumption and income, and, if so, whether α is positive and β is between zero and one. Some issues that might be studied are (1) whether this relationship is stable through time or whether the parameters of the relationship change from one generation to the next (a change in the average propensity to save, 1 − APC, might represent a fundamental change in the behavior of consumers in the economy); (2) whether there are systematic differences in the relationship across different countries, and, if so, what explains these differences; and (3) whether there are other factors that would improve the ability of the model to explain the relationship between consumption and income. For example, Figure 1.1 presents aggregate consumption and personal income in constant dollars for the United States for the 10 years of 2000–2009. (See Appendix Table F1.1.) Apparently, at least superficially, the data (the facts) are consistent with the theory. The relationship appears to be linear, albeit only approximately, the intercept of a line that lies close to most of the points is positive and the slope is less than one, although not by much. (However, if the line is fit by linear least squares regression, the intercept is negative, not positive.) Moreover, observers might disagree on what is meant by relationship in this description.

Economic theories such as Keynes's are typically sharp and unambiguous. Models of demand, production, labor supply, individual choice, educational attainment, income and wages, investment, market equilibrium, and aggregate consumption all specify precise, *deterministic relationships*. Dependent and independent variables are identified, a functional form is specified, and in most cases, at least a qualitative statement is made about the directions of effects that occur when independent variables in the model change. The model is only a simplification of reality. It will include the salient features of the relationship of interest but will leave unaccounted for influences that might well be present but are regarded as unimportant.

Correlations among economic variables are easily observable through descriptive statistics and techniques such as linear regression methods. The ultimate goal of the econometric model builder is often to uncover the deeper causal connections through elaborate structural, behavioral models. Note, for example, Keynes's use of the behavior of a *representative consumer* to motivate the behavior of macroeconomic variables, such as income and consumption. Heckman's model of labor supply noted in Example 1.1 is framed in a model of individual behavior. Berry, Levinsohn, and Pakes's (1995) detailed model of equilibrium pricing in the automobile market is another.

No model could hope to encompass the myriad essentially random aspects of economic life. It is thus also necessary to incorporate stochastic elements. As a consequence, observations on a variable will display variation attributable not only to differences in variables that are explicitly accounted for in the model, but also to the randomness of human behavior and the interaction of countless minor influences that are not. It is understood that the introduction of a random disturbance into a deterministic model is not intended merely to paper over its inadequacies. It is essential to examine the results of the study, in an ex post analysis, to ensure that the allegedly random, unexplained factor is truly unexplainable. If it is not, the model is, in fact, inadequate.[3] The stochastic element endows the model with its statistical properties. Observations on the variable(s) under study are thus taken to be the outcomes of a random process. With a sufficiently detailed stochastic structure and adequate data, the analysis will become a matter of deducing the properties of a probability distribution. The tools and methods of mathematical statistics will provide the operating principles.

A model (or theory) can never truly be confirmed unless it is made so broad as to include every possibility. But it may be subjected to ever more rigorous scrutiny and, in the face of contradictory evidence, refuted. A deterministic theory will be invalidated by a single contradictory observation. The introduction of stochastic elements into the model changes it from an exact statement to a probabilistic description about expected outcomes and carries with it an important implication. Only a preponderance of contradictory evidence can convincingly invalidate the probabilistic model, and what constitutes a preponderance of evidence is a matter of interpretation. Thus, the probabilistic model is less precise but at the same time, more robust.[4]

The process of econometric analysis departs from the specification of a theoretical relationship. We initially proceed on the optimistic assumption that we can obtain precise measurements on all the variables in a correctly specified model. If the ideal conditions are met at every step, the subsequent analysis will be routine. Unfortunately, they rarely are. Some of the difficulties one can expect to encounter are the following:

- The data may be badly measured or may correspond only vaguely to the variables in the model. "The interest rate" is one example.

[3] In the example given earlier, the estimated constant term in the linear least squares regression is negative. Is the theory wrong, or is the finding due to random fluctuation in the data? Another possibility is that the theory is broadly correct, but the world changed between 1936 when Keynes devised his theory and 2000–2009 when the data (outcomes) were generated. Or, perhaps linear least squares is not the appropriate technique to use for this model, and that is responsible for the inconvenient result (the negative intercept).

[4] See Keuzenkamp and Magnus (1995) for a lengthy symposium on testing in econometrics.

- Some of the variables may be inherently unmeasurable. "Expectations" is a case in point.
- The theory may make only a rough guess as to the correct form of the model, if it makes any at all, and we may be forced to choose from an embarrassingly long menu of possibilities.
- The assumed stochastic properties of the random terms in the model may be demonstrably violated, which may call into question the methods of estimation and inference procedures we have used.
- Some relevant variables may be missing from the model.
- The conditions under which data are collected lead to a sample of observations that is systematically unrepresentative of the population we wish to study.

The ensuing steps of the analysis consist of coping with these problems and attempting to extract whatever information is likely to be present in such obviously imperfect data. The methodology is that of mathematical statistics and economic theory. The product is an econometric model.

1.6 PLAN OF THE BOOK

Our objective in this survey is to develop in detail a set of tools, then use those tools in applications. The following set of applications will include many that readers will use in practice. But it is not exhaustive. We will attempt to present our results in sufficient generality that the tools we develop here can be extended to other kinds of situations and applications not described here.

One possible approach is to organize (and orient) the areas of study by the type of data being analyzed—cross section, panel, discrete data, then time series being the obvious organization.

Alternatively, we could distinguish at the outset between micro- and macroeconometrics.[5] Ultimately, all of these will require a common set of tools, including, for example, the multiple regression model, the use of moment conditions for estimation, instrumental variables (IV), and maximum likelihood estimation. With that in mind, the organization of this book is as follows: The first half of the text develops fundamental results that are common to all the applications. The concept of multiple regression and the linear regression model in particular constitutes the underlying platform of most modeling, even if the linear model itself is not ultimately used as the empirical specification. This part of the text concludes with developments of IV estimation and the general topic of panel data modeling. The latter pulls together many features of modern econometrics, such as, again, IV estimation, modeling heterogeneity, and a rich variety of extensions of the linear model. The second half of the text presents a variety

[5] An excellent reference on the former that is at a more advanced level than this text is Cameron and Trivedi (2005). There does not appear to be available a counterpart, large-scale pedagogical survey of macroeconometrics that includes both econometric theory and applications. The numerous more focused studies include books such as Bardsen et al. (2005).

of topics. Part III is an overview of estimation methods. Finally, Parts IV and V present results from microeconometrics and macroeconometrics, respectively. The broad outline is as follows:

I. Regression Modeling
Chapters 2 through 6 present the multiple linear regression model. We will discuss specification, estimation, and statistical inference. This part develops the ideas of estimation, robust analysis, functional form, and principles of model specification.

II. Generalized Regression, Instrumental Variables, and Panel Data
Chapter 7 extends the regression model to nonlinear functional forms. The method of instrumental variables is presented in Chapter 8. Chapters 9 and 10 introduce the generalized regression model and systems of regression models. This section ends with Chapter 11 on panel data methods.

III. Estimation Methods
Chapters 12 through 16 present general results on different methods of estimation including GMM, maximum likelihood, and simulation-based methods. Various estimation frameworks, including non- and semiparametric and Bayesian estimation, are presented in Chapters 12 and 16.

IV. Microeconometric Methods
Chapters 17 through 19 are about microeconometrics, discrete choice modeling, limited dependent variables, and the analysis of data on events—how many occur in a given setting and when they occur. Chapters 17 through 19 are devoted to methods more suited to cross sections and panel data sets.

V. Macroeconometric Methods
Chapters 20 and 21 focus on time-series modeling and macroeconometrics.

VI. Background Materials
Appendices A through E present background material on tools used in econometrics including matrix algebra, probability and distribution theory, estimation, and asymptotic distribution theory. Appendix E presents results on computation. The data sets used in the numerical examples are described in Appendix F. The actual data sets and other supplementary materials can be downloaded from the author's Web site for the text: http://people.stern.nyu.edu/wgreene/Text/.

1.7 PRELIMINARIES

Before beginning, we note some specific aspects of the presentation in the text.

1.7.1 NUMERICAL EXAMPLES

There are many numerical examples given throughout the discussion. Most of these are either self-contained exercises or extracts from published studies. In general, their purpose is to provide a limited application to illustrate a method or model. The reader can replicate them with the data sets provided. This will generally not entail attempting to replicate the full published study. Rather, we use the data sets to provide applications that relate to the published study in a limited fashion that also focuses on a particular

technique, model, or tool. Thus, Riphahn, Wambach, and Million (2003) provide a very useful, manageable (though relatively large) laboratory data set that the reader can use to explore some issues in health econometrics. The exercises also suggest more extensive analyses, again in some cases based on published studies.

1.7.2 SOFTWARE AND REPLICATION

There are now many powerful computer programs that can be used for the computations described in this text. In most cases, the examples presented can be replicated with any modern package, whether the user is employing a high level integrated program such as *Stata*, *SAS,* or *NLOGIT*, or writing his own programs in languages such as *R, MATLAB*, or *Gauss*. The notable exception will be exercises based on simulation. Because, essentially, every package uses a different random number generator, it will generally not be possible to replicate exactly the examples in this text that use simulation (unless you are using the same computer program with the same settings that we are). Nonetheless, the differences that do emerge in such cases should be largely attributable to minor random variation. You will be able to replicate the essential results and overall features in these applications with any of the software mentioned. We will return to this general issue of replicability at a few points in the text, including in Section 15.2 where we discuss methods of generating random samples for simulation-based estimators.

1.7.3 NOTATIONAL CONVENTIONS

We will use vector and matrix notation and manipulations throughout the text. The following conventions will be used: A scalar variable will be denoted with an italic lowercase letter, such as y or x_{nK}. A column vector of scalar values will be denoted by a boldface lowercase letter, such as $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$ and, likewise, for $\mathbf{x}$ and $\mathbf{b}$. The dimensions of a column vector are always denoted as those of a matrix with one column, such as $K \times 1$ or $n \times 1$ and so on. A matrix will always be denoted by a boldface uppercase letter, such as the $n \times K$ matrix, $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}$. Specific elements in a matrix are always subscripted so that the first subscript gives the row and the second gives the column. Transposition of a vector or a matrix is denoted with a prime. A row vector is obtained by transposing a column vector. Thus, $\boldsymbol{\beta}' = [\beta_1, \beta_2, \ldots, \beta_K]$. The product of a row and a column vector will always be denoted in a form such as $\boldsymbol{\beta}'\mathbf{x} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K$. The elements in a matrix, $\mathbf{X}$, form a set of vectors. In terms of its columns, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$—each column is an $n \times 1$ vector. The one possible, unfortunately unavoidable source of ambiguity is the notation necessary to denote a row of a matrix such as $\mathbf{X}$. The elements of the ith row of $\mathbf{X}$ are the row vector, $\mathbf{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{iK}]$. When the matrix, such as $\mathbf{X}$, refers to a data matrix, we

will prefer to use the "i" subscript to denote observations, or the rows of the matrix and "k" to denote the variables, or columns. As we note unfortunately, this would seem to imply that $\mathbf{x}_i$, the transpose of $\mathbf{x}_i'$, would be the ith column of $\mathbf{X}$, which will conflict with our notation. However, with no simple alternative notation available, we will maintain this convention, with the understanding that $\mathbf{x}_i'$, *always* refers to the row vector that is the ith row of an $\mathbf{X}$ matrix. A discussion of the matrix algebra results used in the text is given in Appendix A. A particularly important set of arithmetic results about summation and the elements of the matrix product, $\mathbf{X}'\mathbf{X}$, appears in Section A.2.8.

2
THE LINEAR REGRESSION MODEL

2.1 INTRODUCTION

Econometrics is concerned with *model building*. An intriguing point to begin the inquiry is to consider the question, "What is the model?" The statement of a "model" typically begins with an observation or a proposition that movement of one variable "is caused by" movement of another, or "a variable varies with another," or some qualitative statement about a relationship between a variable and one or more **covariates** that are expected to be related to the interesting variable in question. The model might make a broad statement about behavior, such as the suggestion that individuals' usage of the health care system depends on, for example, perceived health status, demographics (e.g., income, age, and education), and the amount and type of insurance they have. It might come in the form of a verbal proposition, or even a picture (e.g., a flowchart or **path diagram** that suggests directions of influence). The econometric model rarely springs forth in full bloom as a set of equations. Rather, it begins with an *idea* of some kind of relationship. The natural next step for the econometrician is to translate that idea into a set of equations, with a notion that some feature of that set of equations will answer interesting questions about the variable of interest. To continue our example, a more definite statement of the relationship between insurance and health care demanded might be able to answer *how* does health care system utilization depend on insurance coverage? Specifically, is the relationship "positive"—all else equal, is an insured consumer more likely to demand more health care than an uninsured one—or is it "negative"? And, ultimately, one might be interested in a more precise statement, "How much more (or less)?" This and the next several chapters will build the framework that model builders use to pursue questions such as these using data and econometric methods.

From a purely statistical point of view, the researcher might have in mind a variable, y, broadly "demand for health care, H," and a vector of covariates, $\mathbf{x}$ (income, I, insurance, T), and a joint probability distribution of the three, $p(H,I,T)$. Stated in this form, the "relationship" is not posed in a particularly interesting fashion—what is the statistical process that produces health care demand, income, and insurance coverage? However, it is true that $p(H,I,T) = p(H|I, T)p(I,T)$, which decomposes the probability model for the joint process into two outcomes, the joint distribution of income and insurance coverage in the population, $p(I,T)$, and the distribution of "demand for health care" for a specific income and insurance coverage, $p(H|I,T)$. From this perspective, the conditional distribution, $p(H|I,T)$, holds some particular interest, while $p(I,T)$, the distribution of income and insurance coverage in the population, is perhaps of secondary, or no interest. (On the other hand, from the same perspective, the conditional "demand" for insurance coverage, given income, $p(T|I)$, might also be interesting.) Continuing this line of thinking,

the model builder is often interested not in joint variation of all the variables in the model, but in **conditional variation** of one of the variables related to the others.

The idea of the conditional distribution provides a useful starting point for thinking about a relationship between a variable of interest, a "y," and a set of variables, "$\mathbf{x}$," that we think might bear some relationship to it. There is a question to be considered now that returns us to the issue of "What is the model?" What feature of the conditional distribution is of interest? The model builder, thinking in terms of features of the conditional distribution, often gravitates to the expected value, focusing attention on $E[y|\mathbf{x}]$, that is, the **regression function**, which brings us to the subject of this chapter. For the preceding example, this might be natural if y were "number of doctor visits" as in an application examined at several points in the chapters to follow. If we were studying incomes, I, however, which often have a highly skewed distribution, then the mean might not be particularly interesting. Rather, the **conditional median**, for given ages, $M[I|\mathbf{x}]$, might be a more interesting statistic. Still considering the distribution of incomes (and still conditioning on age), other quantiles, such as the 20th percentile, or a poverty line defined as, say, the 5th percentile, might be more interesting yet. Finally, consider a study in finance, in which the variable of interest is asset returns. In at least some contexts, means are not interesting at all—it is variances, and conditional variances in particular, that are most interesting.

The point is that we begin the discussion of the regression model with an understanding of what we mean by "the model." For the present, we will focus on the conditional mean, which is usually the feature of interest. Once we establish how to analyze the regression function, we will use it as a useful departure point for studying other features, such as quantiles and variances. The **linear regression model** is the single most useful tool in the econometrician's kit. Although to an increasing degree in contemporary research it is often only the starting point for the full investigation, it remains the device used to begin almost all empirical research. And it is the lens through which relationships among variables are usually viewed. This chapter will develop the linear regression model in detail. Here, we will detail the fundamental assumptions of the model. The next several chapters will discuss more elaborate specifications and complications that arise in the application of techniques that are based on the simple models presented here.

2.2 THE LINEAR REGRESSION MODEL

The **multiple linear regression model** is used to study the relationship between a **dependent variable** and one or more **independent variables**. The generic form of the linear regression model is

$$\begin{aligned} y &= f(x_1, x_2, \ldots, x_K) + \varepsilon \\ &= x_1\beta_1 + x_2\beta_2 + \ \ldots \ + x_K\beta_K + \varepsilon, \end{aligned} \tag{2-1}$$

where y is the dependent or **explained variable** and $x_1, \ldots, x_K$ are the independent or **explanatory variables**. (We will return to the meaning of "independent" shortly.) One's theory will specify $f(x_1, x_2, \ldots, x_K)$. This function is commonly called the **population regression equation** of y on $x_1, \ldots, x_K$. In this setting, y is the **regressand** and x_k, $k = 1, \ldots, K$ are the **regressors** or covariates. The underlying theory will specify the dependent and independent variables in the model. It is not always obvious which is

appropriately defined as each of these—for example, a demand equation, $quantity = \beta_1 + price \times \beta_2 + income \times \beta_3 + \varepsilon$, and an inverse demand equation, $price = \gamma_1 + quantity \times \gamma_2 + income \times \gamma_3 + u$ are equally valid representations of a market. For modeling purposes, it will often prove useful to think in terms of "autonomous variation." One can conceive of movement of the independent variables outside the relationships defined by the model while movement of the dependent variable is considered in response to some independent or exogenous stimulus.[1]

The term ε is a random **disturbance**, so named because it "disturbs" an otherwise stable relationship. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model, no matter how elaborate. The net effect, which can be positive or negative, of these omitted factors is captured in the disturbance. There are many other contributors to the disturbance in an empirical model. Probably the most significant is errors of measurement. It is easy to theorize about the relationships among precisely defined variables; it is quite another matter to obtain accurate measures of these variables. For example, the difficulty of obtaining reasonable measures of profits, interest rates, capital stocks, or, worse yet, flows of services from capital stocks, is a recurrent theme in the empirical literature. At the extreme, there may be no observable counterpart to the theoretical variable. The literature on the permanent income model of consumption [e.g., Friedman (1957)] provides an interesting example.

We assume that each observation in a sample $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK})$, $i = 1, \ldots, n$, is generated by an underlying process described by

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ \ldots \ + x_{iK}\beta_K + \varepsilon_i.$$

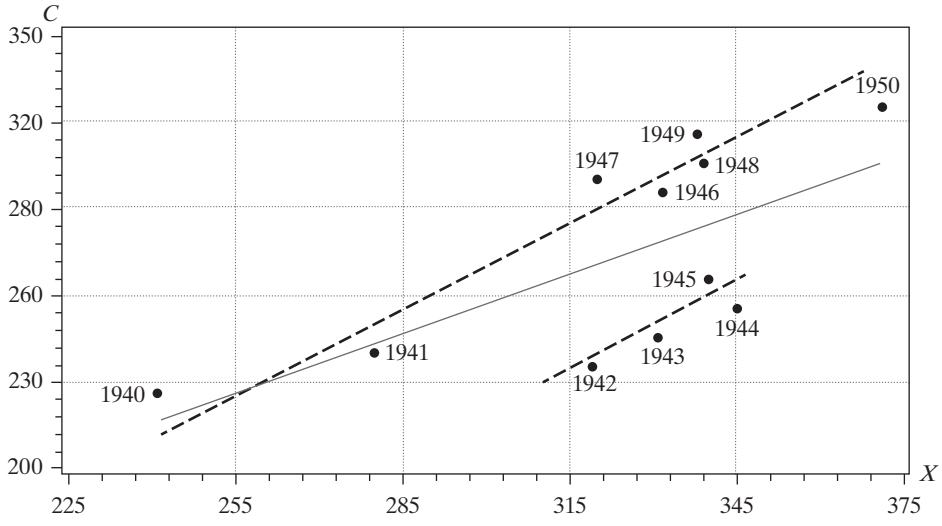
The observed value of y_i is the sum of two parts, the regression function and the disturbance, ε_i. Our objective is to estimate the unknown parameters of the model, use the data to study the validity of the theoretical propositions, and perhaps use the model to predict the variable y. How we proceed from here depends crucially on what we assume about the stochastic process that has led to our observations of the data in hand.

Example 2.1 Keynes's Consumption Function

Example 1.2 discussed a model of consumption proposed by Keynes in his *General Theory* (1936). The theory that consumption, C, and income, X, are related certainly seems consistent with the observed "facts" in Figures 1.1 and 2.1. (These data are in Data Table F2.1.) Of course, the linear function is only approximate. Even ignoring the anomalous wartime years, consumption and income cannot be connected by any simple **deterministic relationship**. The linear part of the model, $C = \alpha + \beta X$, is intended only to represent the salient features of this part of the economy. It is hopeless to attempt to capture every influence in the relationship. The next step is to incorporate the inherent randomness in its real-world counterpart. Thus, we write $C = f(X, \varepsilon)$, where ε is a stochastic element. It is important not to view ε as a catchall for the inadequacies of the model. The model including ε appears adequate for the data not including the war years, but for 1942–1945, something systematic clearly seems to be missing. Consumption in these years could not rise to rates historically consistent with these levels of income because of wartime rationing. A model meant to describe consumption in this period would have to accommodate this influence.

[1] By this definition, it would seem that in our demand relationship, only income would be an independent variable while both price and quantity would be dependent. That makes sense—in a market, equilibrium price and quantity *are* determined at the same time, and do change only when something outside the market equilibrium changes.

FIGURE 2.1 Consumption Data, 1940–1950.



It remains to establish how the stochastic element will be incorporated in the equation. The most frequent approach is to assume that it is *additive.* Thus, we recast the equation in stochastic terms: $C = \alpha + \beta X + \varepsilon$. This equation is an empirical counterpart to Keynes's theoretical model. But, what of those anomalous years of rationing? If we were to ignore our intuition and attempt to fit a line to all these data—the next chapter will discuss at length how we should do that—we might arrive at the solid line in the figure as our best guess. This line, however, is obviously being distorted by the rationing. A more appropriate specification for these data that accommodates both the stochastic nature of the data and the special circumstances of the years 1942–1945 might be one that shifts straight down in the war years, $C = \alpha + \beta X + d_{waryears}\delta_w + \varepsilon$, where the new variable, $d_{waryears}$, equals one in 1942–1945 and zero in other years, and $\delta_w < 0$. This more detailed model is shown by the parallel dashed lines.

One of the most useful aspects of the multiple regression model is its ability to identify the separate effects of a set of variables on a dependent variable. Example 2.2 describes a common application.

Example 2.2 Earnings and Education

Many studies have analyzed the relationship between earnings and education. We would expect, on average, higher levels of education to be associated with higher incomes. The simple regression model

$$earnings = \beta_1 + \beta_2\, education + \varepsilon,$$

however, neglects the fact that most people have higher incomes when they are older than when they are young, regardless of their education. Thus, β_2 will overstate the marginal impact of education. If age and education are positively correlated, then the regression model will associate all the observed increases in income with increases in education and none with, say, experience. A better specification would account for the effect of age, as in

$$earnings = \gamma_1 + \gamma_2\, education + \gamma_3\, age + \varepsilon.$$

It is often observed that income tends to rise less rapidly in the later earning years than in the early ones. To accommodate this possibility, we might further extend the model to

$$earnings = \delta_1 + \delta_2\, education + \delta_3\, age + \delta_4\, age^2 + \varepsilon.$$

We would expect δ_3 to be positive and δ_4 to be negative.

The crucial feature of this model is that it allows us to carry out a conceptual experiment that might not be observed in the actual data. In the example, we might like to (and could) compare the earnings of two individuals of the same age with different amounts of education even if the data set does not actually contain two such individuals. How education should be measured in this setting is a difficult problem. The study of the earnings of twins by Ashenfelter and Krueger (1994), which uses precisely this specification of the earnings equation, presents an interesting approach. [Studies of twins and siblings have provided an interesting thread of research on the education and income relationship. Two other studies are Ashenfelter and Zimmerman (1997) and Bonjour, Cherkas, Haskel, Hawkes, and Spector (2003).] The experiment embodied in the earnings model thus far suggested is a comparison of two otherwise identical individuals who have different years of education. Under this interpretation, the impact of education would be $\partial E[Earnings \,|\, Age, Education]/\partial Education = \beta_2$. But, one might suggest that the experiment the analyst really has in mind is the truly unobservable impact of the additional year of education on a particular individual. To carry out the experiment, it would be necessary to observe the individual twice, once under circumstances that actually occur, $Education_i$, and a second time under the hypothetical (**counterfactual**) *circumstance,* $Education_i + 1$. It is convenient to frame this in a **potential outcomes model** [Rubin (1974)] for individual i:

$$Potential\ Earning = \begin{cases} y_{i0} & \text{if Education} = \mathrm{E}_i, \\ y_{i1} & \text{if Education} = \mathrm{E}_i + 1. \end{cases}$$

By this construction, all other effects would indeed be held constant, and $(y_{i1} - y_{i0})$ could reasonably be labeled the **causal effect** of the additional year of education. If we consider Education in this example as a **treatment**, then the real objective of the experiment is to measure the ***effect of the treatment on the treated***. The ability to infer this result from nonexperimental data that essentially compares "otherwise similar individuals" will be examined in Chapters 8 and19.

A large literature has been devoted to another intriguing question on this subject. Education is not truly independent in this setting. Highly motivated individuals will choose to pursue more education (e.g., by going to college or graduate school) than others. By the same token, highly motivated individuals may do things that, on average, lead them to have higher incomes. If so, does a positive β_2 that suggests an association between income and education really measure the causal effect of education on income, or does it reflect the result of some underlying effect on both variables that we have not included in the regression model? We will revisit the issue in Chapter 19.[2]

2.3 ASSUMPTIONS OF THE LINEAR REGRESSION MODEL

The linear regression model consists of a set of assumptions about how a data set will be produced by an underlying "data-generating process." The theory will specify a relationship between a dependent variable and a set of independent variables. The

[2] This model lays yet another trap for the practitioner. In a cross section, the higher incomes of the older individuals in the sample might tell an entirely different, perhaps macroeconomic story (a cohort effect) from the lower incomes of younger individuals as time and their incomes evolve. It is not necessarily possible to deduce the characteristics of incomes of younger people in the sample *if they were older* by comparing the older individuals in the sample to the younger ones. A parallel problem arises in the analysis of treatment effects that we will examine in Chapter 8.

assumptions that describe the form of the model and relationships among its parts and imply appropriate estimation and inference procedures are listed in Table 2.1.

2.3.1 LINEARITY OF THE REGRESSION MODEL

Let the column vector $\mathbf{x}_k$ be the n observations on variable x_k, $k = 1, \ldots, K$, in a random sample of n observations, and assemble these data in an $n \times K$ data matrix, $\mathbf{X}$. In most contexts, the first column of $\mathbf{X}$ is assumed to be a column of 1s so that β_1 is

TABLE 2.1 Assumptions of the Linear Regression Model

A1. Linearity: We list the assumptions as a description of the joint distribution of y and a set of independent variables, $(x_1, x_2, \ldots, x_K) = \mathbf{x}$. The model specifies a linear relationship between y and $\mathbf{x}$; $y = x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \varepsilon = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$. We will be more specific and assume that this is the regression function, $E[y|x_1, x_2, \ldots, x_K] = E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. The difference between y and $E[y|\mathbf{x}]$ is the disturbance, ε.

A2. Full rank: There is no exact *linear* relationship among any of the independent variables in the model. One way to formulate this is to assume that $E[\mathbf{x}\mathbf{x}'] = \mathbf{Q}$, a $K \times K$ matrix that has full rank K. In practical terms, we wish to be sure that for a random sample of n observations drawn from this process, $(y_1,\mathbf{x}_1'), \ldots, (y_i,\mathbf{x}_i'), \ldots, (y_n,\mathbf{x}_n')$, that the $n \times K$ matrix $\mathbf{X}$ with n rows $\mathbf{x}_i'$ always has rank K if $n \geq K$. This assumption will be necessary for estimation of the parameters of the model.

A3. Exogeneity of the independent variables: $E[\varepsilon|x_1, x_2, \ldots, x_K] = E[\varepsilon|\mathbf{x}] = 0$. This states that the expected value of the disturbance in the regression is not a function of the independent variables observed. This means that the independent variables will not carry useful information for prediction of ε. The assumption is labeled **mean independence**. By the Law of Iterated Expectations (Theorem B.1), it follows that $E[\varepsilon] = 0$. An implication of the exogeneity assumption is that $E[y|x_1, x_2, \ldots, x_K] = \sum_{k=1}^{K} x_k\beta_k$. That is, the linear function in A1 is the **conditional mean function**, or **regression** of y on $x_1, \ldots, x_K$. In the setting of a random sample, we will also begin from an assumption that observations on ε in the sample are uncorrelated with information in other observations—that is, $E[\varepsilon_i|\mathbf{x}_1, \ldots, \mathbf{x}_n] = 0$. This is labeled **strict exogeneity**. An implication will be, for each observation in a sample of observations, $E[\varepsilon_i|\mathbf{X}] = 0$, and for the sample as a whole, $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$.

A4. Homoscedasticity: The disturbance in the regression has **conditional variance**, $\mathrm{Var}[\varepsilon|\mathbf{x}] = \mathrm{Var}[\varepsilon] = \sigma^2$. (The second equality follows from Theorem B.4.) This assumption limits the generality of the model, and we will want to examine how to relax it in the chapters to follow. Once again, considering a random sample, we will assume that the observations ε_i and ε_j are uncorrelated for $i \neq j$. With reference to a times-series setting, this will be labeled **nonautocorrelation**. The implication will be $\mathrm{E}[\varepsilon_i\varepsilon_j|\mathbf{x}_i,\mathbf{x}_j]$. We will strengthen this to $E[\varepsilon_i\varepsilon_j|\mathbf{X}] = 0$ for $i \neq j$ and $E[\varepsilon\varepsilon'|\mathbf{X}] = \sigma^2\mathbf{I}$.

A5. Data generation: The data in $(x_1, x_2, \ldots, x_K)$ (that is, the process by which $\mathbf{x}$ is generated) may be any mixture of constants and random variables. The crucial elements for present purposes are the exogeneity assumption, A3, and the variance and covariance assumption, A4. Analysis can be done conditionally on the observed $\mathbf{X}$, so whether the elements in $\mathbf{X}$ are fixed constants or random draws from a stochastic process will not influence the results. In later, more advanced treatments, we will want to be more specific about the possible relationship between ε_i and $\mathbf{x}_j$. Nothing is lost by assuming that the n observations in hand are a **random sample** of independent, identically distributed draws from a joint distribution of $(y,\mathbf{x})$. In some treatments to follow, such as panel data, some observations will be correlated by construction. It will be necessary to revisit the assumptions at that point, and revise them as necessary.

A6. Normal distribution: The disturbances are normally distributed. This is a convenience that we will dispense with after some analysis of its implications. The normality assumption is useful for defining the computations behind statistical inference about the regression, such as confidence intervals and hypothesis tests. For practical purposes, it will be useful then to extend those results and in the process develop a more flexible approach that does not rely on this specific assumption.

the constant term in the model. Let $\mathbf{y}$ be the n observations, $y_1, \ldots, y_n$, and let $\boldsymbol{\varepsilon}$ be the column vector containing the n disturbances. The model in (2-1) as it applies to each of and all n observations can now be written

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_K\beta_K + \boldsymbol{\varepsilon}, \tag{2-2}$$

or in the form of Assumption A1,

$$\boxed{\text{ASSUMPTION A1:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.} \tag{2-3}$$

A NOTATIONAL CONVENTION

Henceforth, to avoid a possibly confusing and cumbersome notation, we will use a boldface $\mathbf{x}$ to denote a column *or* a row of $\mathbf{X}$. Which of these applies will be clear from the context. In (2-2), $\mathbf{x}_k$ is the kth column of $\mathbf{X}$. Subscript k will usually be used to denote columns (variables). It will often be convenient to refer to a single observation in (2-3), which we would write

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{2-4}$$

Subscripts i, j, and t will generally be used to denote rows (observations) of $\mathbf{X}$. In (2-4), $\mathbf{x}_i'$ is a row vector that is the ith $1 \times K$ row of $\mathbf{X}$.

Our primary interest is in estimation and inference about the parameter vector $\boldsymbol{\beta}$. Note that the simple regression model in Example 2.1 is a special case in which $\mathbf{X}$ has only two columns, the first of which is a column of 1s. The assumption of linearity of the regression model includes the additive disturbance. For the regression to be linear in the sense described here, it must be of the form in (2-1) either in the original variables or after some suitable transformation. For example, the model

$$y = Ax^{\beta}e^{\varepsilon}$$

is linear (after taking logs on both sides of the equation), whereas

$$y = Ax^{\beta} + \varepsilon$$

is not. The observed dependent variable is thus the sum of two components, a deterministic element $\alpha + \beta x$ and a random variable ε. It is worth emphasizing that neither of the two parts is directly observed because α and β are unknown.

The linearity assumption is not so narrow as it might first appear. In the regression context, *linearity* refers to the manner in which the parameters and the disturbance enter the equation, not necessarily to the relationship among the variables. For example, the equations $y = \alpha + \beta x + \varepsilon$, $y = \alpha + \beta \cos(x) + \varepsilon$, $y = \alpha + \beta/x + \varepsilon$, and $y = \alpha + \beta \ln x + \varepsilon$ are all linear in some function of x by the definition we have used here. In the examples, only x has been transformed, but y could have been as well, as in $y = Ax^{\beta}e^{\varepsilon}$, which is a linear relationship in the logs of x and y; $\ln y = \alpha + \beta \ln x + \varepsilon$. The variety of functions is unlimited. This aspect of the model is used in a number of commonly used functional forms. For example, the **loglinear model** is

$$\ln y = \beta_1 + \beta_2 \ln x_2 + \beta_3 \ln x_3 + \cdots + \beta_K \ln x_K + \varepsilon.$$

This equation is also known as the **constant elasticity** form, as in this equation, the elasticity of y with respect to changes in x_k is $\partial \ln y / \partial \ln x_k = \beta_k$, which does not vary with x_k. The loglinear form is often used in models of demand and production. Different values of β_k produce widely varying functions.

Example 2.3 The U.S. Gasoline Market

Data on the U.S. gasoline market for the years 1953–2004 are given in Table F2.2 in Appendix F. We will use these data to obtain, among other things, estimates of the income, own price, and cross-price elasticities of demand in this market. These data also present an interesting question on the issue of holding "all other things constant," that was suggested in Example 2.2. In particular, consider a somewhat abbreviated model of per capita gasoline consumption:

$$\ln(G/pop) = \beta_1 + \beta_2 \ln(Income/pop) + \beta_3 \ln price_G + \beta_4 \ln P_{newcars} + \beta_5 \ln P_{usedcars} + \varepsilon.$$

This model will provide estimates of the income and price elasticities of demand for gasoline and an estimate of the elasticity of demand with respect to the prices of new and used cars. What should we expect for the sign of β_4? Cars and gasoline are complementary goods, so if the prices of new cars rise, *ceteris paribus*, gasoline consumption should fall. Or should it? If the prices of new cars rise, then consumers will buy fewer of them; they will keep their used cars longer and buy fewer new cars. If older cars use more gasoline than newer ones, then the rise in the prices of new cars would lead to higher gasoline consumption than otherwise, not lower. We can use the multiple regression model and the gasoline data to attempt to answer the question.

A **semilog** model is often used to model growth rates:

$$\ln y_t = \mathbf{x}_t' \boldsymbol{\beta} + \delta t + \varepsilon t.$$

In this model, the autonomous (at least not explained by the model itself) proportional, per period growth rate is $\partial \ln y / \partial t = \delta$. Other variations of the general form

$$f(y_t) = g(\mathbf{x}_t' \boldsymbol{\beta} + \varepsilon_t)$$

will allow a tremendous variety of functional forms, all of which fit into our definition of a linear model.

The linear regression model is sometimes interpreted as an approximation to some unknown, underlying function. (See Section A.8.1 for discussion.) By this interpretation, however, the linear model, even with quadratic terms, is fairly limited in that such an approximation is likely to be useful only over a small range of variation of the independent variables. The translog model discussed in Example 2.4, in contrast, has proven more effective as an approximating function.

Example 2.4 The Translog Model

Modern studies of demand and production are usually done with a **flexible functional form**. Flexible functional forms are used in econometrics because they allow analysts to model complex features of the production function, such as elasticities of substitution, which are functions of the second derivatives of production, cost, or utility functions. The linear model restricts these to equal zero, whereas the loglinear model (e.g., the Cobb–Douglas model) restricts the interesting elasticities to the uninteresting values of -1 or $+1$. The most popular flexible functional form is the **translog model**, which is often interpreted as a second-order approximation to an unknown functional form. [See Berndt and Christensen (1973).] One way to derive it is as follows. We first write $y = g(x_1, \ldots, x_K)$. Then, $\ln y = \ln g(\ldots) = f(\ldots)$. Since by a trivial transformation $x_k = \exp(\ln x_k)$, we interpret the function as a function of the logarithms of the x's. Thus, $\ln y = f(\ln x_1, \ldots, \ln x_K)$.

Now, expand this function in a second-order Taylor series around the point $\mathbf{x} = [1, 1, \ldots, 1]'$ so that at the expansion point, the log of each variable is a convenient zero. Then

$$\ln y = f(\mathbf{0}) + \sum_{k=1}^{K} [\partial f(\cdot)/\partial \ln x_k]_{|\ \ln \mathbf{x}=0} \ln x_k + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} [\partial^2 f(\cdot)/\partial \ln x_k \partial \ln x_l]_{|\ \ln \mathbf{x}=0} \ln x_k \ln x_l + \varepsilon.$$

The disturbance in this model is assumed to embody the familiar factors and the error of approximation to the unknown function. Because the function and its derivatives evaluated at the fixed value $\mathbf{0}$ are constants, we interpret them as the coefficients and write

$$\ln y = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_k + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} \gamma_{kl} \ln x_k \ln x_l + \varepsilon.$$

This model is linear by our definition but can, in fact, mimic an impressive amount of curvature when it is used to approximate another function. An interesting feature of this formulation is that the loglinear model is a special case, when $\gamma_{kl} = 0$. Also, there is an interesting test of the underlying theory possible because if the underlying function were assumed to be continuous and twice continuously differentiable, then by Young's theorem it must be true that $\gamma_{kl} = \gamma_{lk}$. We will see in Chapter 10 how this feature is studied in practice.

Despite its great flexibility, the linear model will not accommodate all the situations we will encounter in practice. In Example 14.13 and Chapter 18, we will examine the regression model for doctor visits that was suggested in the introduction to this chapter. An appropriate model that describes the number of visits has conditional mean function $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. It is tempting to linearize this directly by taking logs, because $\ln E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. But $\ln E[y|\mathbf{x}]$ is not equal to $E[\ln y|\mathbf{x}]$. In that setting, y can equal zero (and does for most of the sample), so $\mathbf{x}'\boldsymbol{\beta}$ (which can be negative) is not an appropriate model for $\ln y$ (which does not exist) or for y which cannot be negative. The methods we consider in this chapter are not appropriate for estimating the parameters of such a model. Relatively straightforward techniques have been developed for nonlinear models such as this, however. We shall treat them in detail in Chapter 7.

2.3.2 FULL RANK

Assumption A2 is that there are no exact *linear* relationships among the variables.

ASSUMPTION A2: $\mathbf{X}$ is an $n \times K$ matrix with rank K. **(2-5)**

Hence, $\mathbf{X}$ has full column rank; the columns of $\mathbf{X}$ are linearly independent and there are at least K observations. [See (A-42) and the surrounding text.] This assumption is known as an **identification condition**. To see the need for this assumption, consider an example.

Example 2.5 Short Rank

Suppose that a cross-section model specifies that consumption, C, relates to income as follows:

$$C = \beta_1 + \beta_2\, \textit{nonlabor income} + \beta_3\, \textit{salary} + \beta_4\, \textit{total income} + \varepsilon,$$

where *total income* is exactly equal to *salary* plus *nonlabor income*. Clearly, there is an exact linear relationship among the variables in the model. Now, let

$$\beta_2' = \beta_2 + a,$$
$$\beta_3' = \beta_3 + a,$$

and

$$\beta_4' = \beta_4 - a,$$

where a is any number. Then the exact same value appears on the right-hand side of C if we substitute β'_2, β'_3, and β'_4 for β_2, β_3, and β_4. Obviously, there is no way to estimate the parameters of this model.

If there are fewer than K observations, then $\mathbf{X}$ cannot have **full rank**. Hence, we make the assumption that n is at least as large as K.

In the simple linear model with a constant term and a single x, the full rank assumption means that there must be variation in the regressor, x. If there is no variation in x, then all our observations will lie on a vertical line. This situation does not invalidate the other assumptions of the model; presumably, it is a flaw in the data set. The possibility that this suggests is that we *could* have drawn a sample in which there was variation in x, but in this instance, we did not. Thus, the model still applies, but we cannot learn about it from the data set in hand.

Example 2.6 An Inestimable Model

In Example 3.4, we will consider a model for the sale price of Monet paintings. Theorists and observers have different models for how prices of paintings at auction are determined. One (naïve) student of the subject suggests the model

$$\begin{aligned}\ln \textit{Price} &= \beta_1 + \beta_2 \ln \textit{Size} + \beta_3 \ln \textit{Aspect Ratio} + \beta_4 \ln \textit{Height} + \varepsilon \\ &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon,\end{aligned}$$

where $\textit{Size} = \textit{Width} \times \textit{Height}$ and $\textit{Aspect Ratio} = \textit{Width}/\textit{Height}$. By simple arithmetic, we can see that this model shares the problem found with the consumption model in Example 2.5—in this case, $x_2 - x_4 = x_3 + x_4$. So, this model is, like the previous one, not estimable—it is not identified. It is useful to think of the problem from a different perspective here (so to speak). In the linear model, it must be possible for the variables in the model to vary linearly independently. But, in this instance, while it is possible for any pair of the three covariates to vary independently, the three together cannot. The "model," that is, the theory, is an entirely reasonable model as it stands. Art buyers might very well consider all three of these features in their valuation of a Monet painting. However, it is not possible to learn about that from the observed data, at least not with this linear regression model.

The full rank assumption is occasionally interpreted to mean that the variables in $\mathbf{X}$ must be able to vary independently from each other. This is clearly not the case in Example 2.6, which is a flawed model. But it is also not the case in the linear model

$$E[y|x,z] = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 z + \varepsilon.$$

There is nothing problematic with this model—nor with the model in Example 2.2 or the translog model in Example 2.4. Nonetheless, x and x^2 cannot vary independently. The resolution of this seeming contradiction is to sharpen what we mean by the variables in the model varying independently. First, it remains true that $\mathbf{X}$ must have full column rank to carry out the *linear* regression. But, independent variation of the variables in the model is a different concept. The columns of $\mathbf{X}$ are not necessarily the set of variables in the model. In the equation above, the "variables" are only x and z. The identification problem we consider here would state that it must be possible for z to vary independently

from x. If z is a deterministic function of x, then it is not possible to identify an effect in the model for variable z separately from that for x.

2.3.3 REGRESSION

The disturbance is assumed to have conditional expected value zero at every observation, which we write as

$$E[\varepsilon_i|\mathbf{X}] = 0. \tag{2-6}$$

For the full set of observations, we write Assumption A3 as

$$\text{ASSUMPTION A3:} \quad E[\boldsymbol{\varepsilon}|\mathbf{X}] = \begin{bmatrix} E[\varepsilon_1|\mathbf{X}] \\ E[\varepsilon_2|\mathbf{X}] \\ \vdots \\ E[\varepsilon_n|\mathbf{X}] \end{bmatrix} = \mathbf{0}. \tag{2-7}$$

There is a subtle point in this discussion that the observant reader might have noted. In (2-7), the left-hand side states, in principle, that the mean of each ε_i *conditioned on all observations* $\mathbf{x}_j$ is zero. This strict exogeneity assumption states, in words, that no observations on $\mathbf{x}$ convey information about the expected value of the disturbance. It is conceivable—for example, in a time-series setting—that although $\mathbf{x}_i$ might provide no information about $E[\varepsilon_i|\cdot]$, $\mathbf{x}_j$ *at some other observation*, such as in the previous time period, might. Our assumption at this point is that there is no information about $E[\varepsilon_i|\cdot]$ contained in *any* observation $\mathbf{x}_j$. Later, when we extend the model, we will study the implications of dropping this assumption. [See Wooldridge (1995).] We will also assume that the disturbances convey no information about each other. That is, $E[\varepsilon_i|\varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n] = 0$. In sum, at this point, we have assumed that the disturbances are purely random draws from some population.

The zero conditional mean implies that the unconditional mean is also zero, because by the **Law of Iterated Expectations** [Theorem B.1, (B-66)],

$$E[\varepsilon_i] = E_{\mathbf{x}}[E[\varepsilon_i|\mathbf{X}]] = E_{\mathbf{x}}[0] = 0.$$

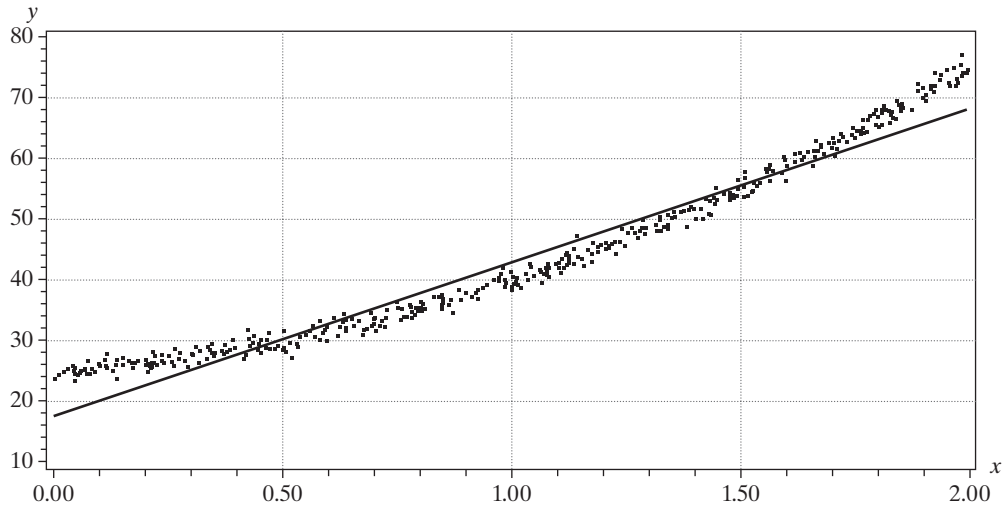
For each ε_i, by Theorem B.2, $\text{Cov}[E[\varepsilon_i|\mathbf{X}], \mathbf{X}] = \text{Cov}[\varepsilon_i, \mathbf{X}]$, Assumption A3 implies that $\text{Cov}[\varepsilon_i, \mathbf{x}] = 0$ for all i. The converse is not true; $E[\varepsilon_i] = 0$ does not imply that $E[\varepsilon_i|\mathbf{x}_i] = 0$. Example 2.7 illustrates the difference.

Example 2.7 Nonzero Conditional Mean of the Disturbances

Figure 2.2 illustrates the important difference between $E[\varepsilon_i] = 0$ and $E[\varepsilon_i|x_i] = 0$. The overall mean of the disturbances in the sample is zero, but the mean for specific ranges of x is distinctly nonzero. A pattern such as this in observed data would serve as a useful indicator that the specification of the linear regression should be questioned. In this particular case, the true conditional mean function (which the researcher would not know in advance) is actually $E[y|x] = 25 + 5x(1 + 2x)$. The sample data are suggesting that a linear specification is not appropriate for these data. A quadratic specification would seem to be a good candidate. This modeling strategy is pursued in an application in Example 6.6.

In most cases, the zero overall mean assumption is not restrictive. Consider a two-variable model and suppose that the mean of ε is $\mu \neq 0$. Then $\alpha + \beta x + \varepsilon$ is the same

FIGURE 2.2 Disturbances with Nonzero Conditional Mean and Zero Unconditional Mean.



as $(\alpha + \mu) + \beta x + (\varepsilon - \mu)$. Letting $\alpha' = \alpha + \mu$ and $\varepsilon' = \varepsilon - \mu$ produces the original model. For an application, see the discussion of frontier production functions in Section 19.2.4. But if the original model does not contain a constant term, then assuming $E[\varepsilon_i] = 0$ could be substantive. This suggests that there is a potential problem in models without constant terms. As a general rule, regression models should not be specified without constant terms unless this is specifically dictated by the underlying theory.[3] Arguably, if we have reason to specify that the mean of the disturbance is something other than zero, we should build it into the systematic part of the regression, leaving in the disturbance only the unknown part of ε. Assumption A3 also implies that

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}. \tag{2-8}$$

Assumptions A1 and A3 comprise the *linear regression model*. The regression of $\mathbf{y}$ on $\mathbf{X}$ is the conditional mean, $E[\mathbf{y}|\mathbf{X}]$, so that without Assumption A3, $\mathbf{X}\boldsymbol{\beta}$ is *not* the conditional mean function.

The remaining assumptions will more completely specify the characteristics of the disturbances in the model and state the conditions under which the sample observations on $\mathbf{x}$ are obtained.

2.3.4 HOMOSCEDASTIC AND NONAUTOCORRELATED DISTURBANCES

The fourth assumption concerns the variances and covariances of the disturbances:

$$\text{Var}[\varepsilon_i|\mathbf{X}] = \sigma^2, \qquad \text{for all } i = 1, \ldots, n,$$

[3] Models that describe first differences of variables might well be specified without constants. Consider $y_t - y_{t-1}$. If there is a constant term α on the right-hand side of the equation, then y_t is a function of αt, which is an explosive regressor. Models with linear time trends merit special treatment in the time-series literature. We will return to this issue in Chapter 21.

and

$$\text{Cov}[\varepsilon_i, \varepsilon_j | \mathbf{X}] = 0, \qquad \text{for all } i \neq j.$$

Constant variance is labeled **homoscedasticity**. Consider a model that describes the profits of firms in an industry as a function of, say, size. Even accounting for size, measured in dollar terms, the profits of large firms will exhibit greater variation than those of smaller firms. The homoscedasticity assumption would be inappropriate here. Survey data on household expenditure patterns often display marked **heteroscedasticity**, even after accounting for income and household size.

Uncorrelatedness across observations is labeled generically nonautocorrelation. In Figure 2.1, there is some suggestion that the disturbances might not be truly independent across observations. Although the number of observations is small, it does appear that, on average, each disturbance tends to be followed by one with the same sign. This "inertia" is precisely what is meant by **autocorrelation**, and it is assumed away at this point. Methods of handling autocorrelation in economic data occupy a large proportion of the literature and will be treated at length in Chapter 20. Note that nonautocorrelation does not imply that observations y_i and y_j are uncorrelated. The assumption is that *deviations* of observations from their expected values are uncorrelated.

The two assumptions imply that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1|\mathbf{X}] & E[\varepsilon_1\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n|\mathbf{X}] \\ E[\varepsilon_2\varepsilon_1|\mathbf{X}] & E[\varepsilon_2\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n|\mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\varepsilon_n\varepsilon_1|\mathbf{X}] & E\varepsilon_n\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n|\mathbf{X}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix},$$

which we summarize in Assumption A4:

$$\boxed{\text{Assumption A4:} \qquad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \sigma^2\mathbf{I}.} \tag{2-9}$$

By using the variance decomposition formula in (B-69), we find

$$\text{Var}[\boldsymbol{\varepsilon}] = E[\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}]] + \text{Var}[E[\boldsymbol{\varepsilon}|\mathbf{X}]] = \sigma^2\mathbf{I}.$$

Once again, we should emphasize that this assumption describes the information about the variances and covariances among the disturbances that is provided by the independent variables. For the present, we assume that there is none. We will also drop this assumption later when we enrich the regression model. We are also assuming that the disturbances themselves provide no information about the variances and covariances. Although a minor issue at this point, it will become crucial in our treatment of time-series applications. Models such as $\text{Var}[\varepsilon_t | \varepsilon_{t-1}] = \sigma^2 + \alpha\varepsilon_{t-1}^2$, a "GARCH" model (see Chapter 20), do not violate our conditional variance assumption, but do assume that $\text{Var}[\varepsilon_t | \varepsilon_{t-1}] \neq \text{Var}[\varepsilon_t]$.

2.3.5 DATA GENERATING PROCESS FOR THE REGRESSORS

It is common to assume that $\mathbf{x}_i$ is nonstochastic, as it would be in an experimental situation. Here the analyst chooses the values of the regressors and then observes y_i. This process might apply, for example, in an agricultural experiment in which y_i is yield and $\mathbf{x}_i$ is fertilizer concentration and water applied. The assumption of **nonstochastic regressors** at this point would be a mathematical convenience. With it, we could use the results of elementary statistics to obtain our results by treating the vector $\mathbf{x}_i$ simply as a known constant in the probability distribution of y_i. With this simplification, Assumptions A3 and A4 would be made unconditional and the counterparts would now simply state that the probability distribution of ε_i involves none of the constants in $\mathbf{X}$.

Social scientists are almost never able to analyze experimental data, and relatively few of their models are built around nonrandom regressors. Clearly, for example, in any model of the macroeconomy, it would be difficult to defend such an asymmetric treatment of aggregate data. Realistically, we have to allow the data on $\mathbf{x}_i$ to be random the same as y_i. So an alternative formulation is to assume that $\mathbf{x}_i$ is a random vector and our formal assumption concerns the nature of the random process that produces $\mathbf{x}_i$. If $\mathbf{x}_i$ is taken to be a random vector, then Assumptions A1 through A4 become a statement about the joint distribution of y_i and $\mathbf{x}_i$. The precise nature of the regressor and how we view the sampling process will be a major determinant of our derivation of the statistical properties of our estimators and test statistics. In the end, the crucial assumption is A3, the uncorrelatedness of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$. Now, we do note that this alternative is not completely satisfactory either, because $\mathbf{X}$ may well contain nonstochastic elements, including a constant, a time trend, and dummy variables that mark specific episodes in time. This makes for an ambiguous conclusion, but there is a straightforward and economically useful way out of it. We will allow $\mathbf{X}$ to be any mixture of constants and random variables, and the mean and variance of ε_i are both independent of all elements of $\mathbf{X}$.

ASSUMPTION A5: $\mathbf{X}$ may be fixed or random.	**(2-10)**

2.3.6 NORMALITY

It is convenient to assume that the disturbances are **normally distributed**, with zero mean and constant variance. That is, we add normality of the distribution to Assumptions A3 and A4.

ASSUMPTION A6: $\boldsymbol{\varepsilon} \mid \mathbf{X} \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$.	**(2-11)**

In view of our description of the source of ε, the conditions of the central limit theorem will generally apply, at least approximately, and the normality assumption will be reasonable in most settings. A useful implication of Assumption A6 is that it implies that observations on ε_i are statistically independent as well as uncorrelated. [See the third point in Section B.9, (B-97) and (B-99).]

Normality is usually viewed as an unnecessary and possibly inappropriate addition to the regression model. Except in those cases in which some alternative distribution is explicitly assumed, as in the stochastic frontier model discussed in Chapter 19, the normality assumption may be quite reasonable. But the assumption is not necessary

to obtain most of the results we use in multiple regression analysis. It will prove useful as a starting point in constructing confidence intervals and test statistics, as shown in Section 4.7 and Chapter 5. But it will be possible to discard this assumption and retain for practical purposes the important statistical results we need for the investigation.

2.3.7 INDEPENDENCE AND EXOGENEITY

The term *independent* has been used several ways in this chapter.

In Section 2.2, the right-hand-side variables in the model are denoted the independent variables. Here, the notion of independence refers to the sources of variation. In the context of the model, the variation in the independent variables arises from sources that are outside of the process being described. Thus, in our health services versus income example in the introduction, we have suggested a theory for how variation in demand for services is associated with variation in income and, possibly, variation in insurance coverage. But, we have not suggested an explanation of the sample variation in income; income is assumed to vary for reasons that are outside the scope of the model. Nor have we suggested a behavioral model for insurance take up. This will be a convenient definition to use for **exogeneity** of a variable x.

The assumption in (2-6), $E[\varepsilon_i|\mathbf{X}] = 0$, is mean independence. Its implication is that variation in the disturbances in our data is not explained by variation in the independent variables. Situations in which $E[\varepsilon_i|\mathbf{X}] \neq 0$ arise frequently, as we will explore in Chapter 8 and others. When $E[\varepsilon|x] \neq 0$, x is **endogenous** in the model. The most straightforward instance is a left-out variable. Consider the model in Example 2.2. In a simple model that contains only *Education* but which has inappropriately omitted *Age*, it would follow that *Age* implicitly appears in the disturbance:

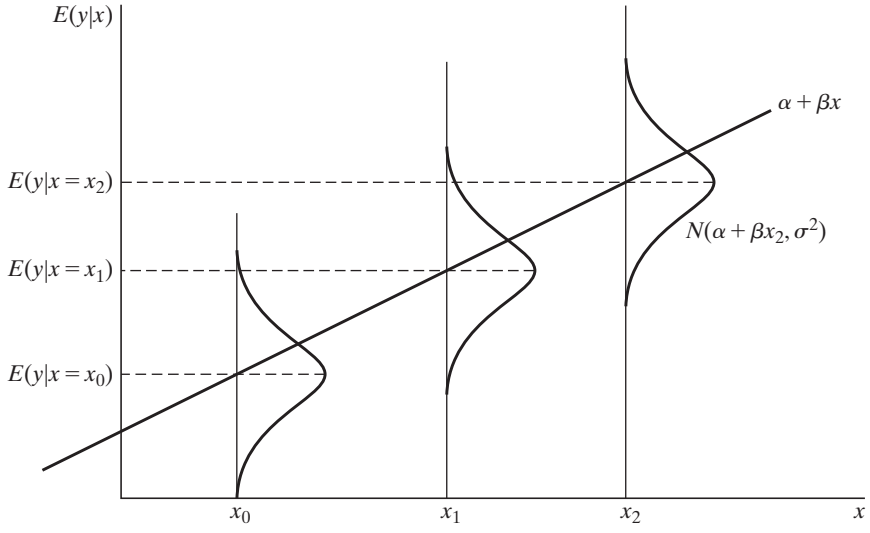
$$Income = \gamma_1 + \gamma_2 Education + (\gamma_3 Age + u) = \gamma_1 + \gamma_2 Education + \varepsilon.$$

If *Education* and (the hidden variable) *Age* are correlated, then *Education* is endogenous in this equation, which is no longer a regression because $E[\varepsilon|Education] = \gamma_3 E[Age|Education] + E[u|Education] \neq 0$.

We have also assumed in Section 2.3.4 that the disturbances are uncorrelated with each other (Assumption A4 in Table 2.1). This implies that $E[\varepsilon_i|\varepsilon_j] = 0$ when $i \neq j$—the disturbances are also mean independent of each other. Conditional normality of the disturbances assumed in Section 2.3.6 (Assumption A6) implies that they are statistically independent of each other, which is a stronger result than mean independence and stronger than we will need in most applications.

Finally, Section 2.3.2 discusses the **linear independence** of the columns of the data matrix, $\mathbf{X}$. The notion of independence here is an algebraic one relating to the column rank of $\mathbf{X}$. In this instance, the underlying interpretation is that it must be possible for the variables in the model to vary linearly independently of each other. Thus, in Example 2.6, we find that it is not possible for the logs of surface area, aspect ratio, and height of a painting all to vary independently of one another. The modeling implication is that, if the variables cannot vary independently of each other, then it is not possible to analyze them in a linear regression model that assumes the variables can each vary while holding the others constant. There is an ambiguity in this discussion of independence of the variables. We have both *age* and *age squared* in a model in Example 2.2. These cannot vary independently, but there is no obstacle to formulating a linear regression

FIGURE 2.3 The Normal Linear Regression Model.



model containing both *age* and *age squared*. The resolution is that *age* and *age squared*, though not *functionally* independent, are *linearly* independent in **X**. That is the crucial assumption in the linear regression model.

2.4 SUMMARY AND CONCLUSIONS

This chapter has framed the linear regression model, the basic platform for model building in econometrics. The assumptions of the classical regression model are summarized in Figure 2.3, which shows the two-variable case.

Key Terms and Concepts

- Autocorrelation
- Central limit theorem
- Conditional mean
- Conditional median
- Conditional variance
- Conditional variation
- Constant elasticity
- Counterfactual
- Covariate
- Dependent variable
- Deterministic relationship
- Disturbance
- Endogeneity
- Exogeneity
- Explained variable
- Explanatory variable
- Flexible functional form
- Full rank
- Heteroscedasticity
- Homoscedasticity
- Identification condition
- Impact of treatment on the treated
- Independent variable
- Law of Iterated Expectations
- Linear independence
- Linear regression model
- Loglinear model
- Mean independence
- Multiple linear regression model
- Nonautocorrelation
- Nonstochastic regressors
- Normality
- Normally distributed
- Path diagram
- Population regression equation
- Random sample
- Regressand
- Regression function
- Regressor
- Semilog
- Translog model

3
LEAST SQUARES REGRESSION

3.1 INTRODUCTION

This chapter examines the computation of the least squares regression model. A useful understanding of what is being computed when one uses least squares to compute the coefficients of the model can be developed before we turn to the statistical aspects. Section 3.2 will detail the computations of least squares regression. We then examine two particular aspects of the fitted equation:

- The crucial feature of the multiple regression model is its ability to provide the analyst a device for "holding other things constant." In an earlier example, we considered the "partial effect" of an additional year of education, holding age constant in

  $$Earnings = \gamma_1 + \gamma_2 Education + \gamma_3 Age + \varepsilon.$$

  The theoretical exercise is simple enough. How do we do this in practical terms? How does the actual computation of the linear model produce the interpretation of partial effects? An essential insight is provided by the notion of partial regression coefficients. Sections 3.3 and 3.4 use the **Frisch–Waugh theorem** to show how the regression model controls for (i.e., holds constant) the effects of intervening variables.
- The model is proposed to describe the movement of an explained variable. In broad terms, $y = \mu(\mathbf{x}) + \varepsilon$. How well does the model do this? How can we measure the success? Sections 3.5 and 3.6 examine fit measures for the linear regression.

3.2 LEAST SQUARES REGRESSION

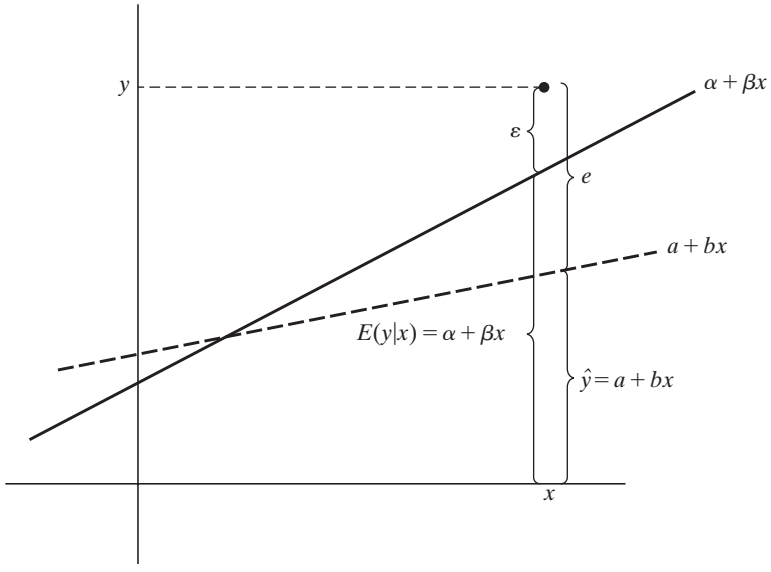
Consider a simple (the simplest) version of the model in the introduction,

$$Earnings = \alpha + \beta Education + \varepsilon.$$

The unknown parameters of the stochastic relationship, $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, are the objects of estimation. It is necessary to distinguish between unobserved population quantities, such as $\boldsymbol{\beta}$ and ε_i, and sample estimates of them, denoted $\mathbf{b}$ and e_i. The **population regression** is $E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$, whereas our estimate of $E[y_i|\mathbf{x}_i]$ is denoted $\hat{y}_i = \mathbf{x}_i'\mathbf{b}$. The **disturbance** associated with the ith data point is $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$. For any value of $\mathbf{b}$, we shall estimate ε_i with the **residual**

$$e_i = y_i - \mathbf{x}_i'\mathbf{b}.$$

FIGURE 3.1 Population and Sample Regression.



From the two definitions,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i = \mathbf{x}_i'\mathbf{b} + e_i.$$

These results are summarized for a two-variable regression in Figure 3.1.

The **population quantity**, $\boldsymbol{\beta}$, is a vector of unknown parameters of the joint probability distribution of $(y, \mathbf{x})$ whose values we hope to estimate with our sample data, $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$. This is a problem of statistical inference that is discussed in Chapter 4 and much of the rest of the book. It is useful, however, to begin by considering the algebraic problem of choosing a vector $\mathbf{b}$ so that the fitted line $\mathbf{x}_i'\mathbf{b}$ is close to the data points. The measure of closeness constitutes a **fitting criterion**. The one used most frequently is **least squares**.[1]

3.2.1 THE LEAST SQUARES COEFFICIENT VECTOR

The least squares coefficient vector minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} e_{i0}^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b}_0)^2, \tag{3-1}$$

where $\mathbf{b}_0$ denotes a choice for the coefficient vector. In matrix terms, minimizing the sum of squares in (3-1) requires us to choose $\mathbf{b}_0$ to

$$\text{Minimize}_{\mathbf{b}_0}\, S(\mathbf{b}_0) = \mathbf{e}_0'\mathbf{e}_0 = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'\,(\mathbf{y} - \mathbf{X}\mathbf{b}_0). \tag{3-2}$$

[1] We have yet to establish that the practical approach of fitting the line as closely as possible to the data by least squares leads to estimators with good statistical properties. This makes intuitive sense and is, indeed, the case. We shall return to the statistical issues in Chapter 4.

Expanding this gives

$$\mathbf{e}_0'\mathbf{e}_0 = \mathbf{y}'\mathbf{y} - \mathbf{b}_0'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0'\mathbf{X}'\mathbf{X}\mathbf{b}_0 \tag{3-3}$$

or

$$S(\mathbf{b}_0) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0'\mathbf{X}'\mathbf{X}\mathbf{b}_0.$$

The necessary condition for a minimum is

$$\frac{\partial S(\mathbf{b}_0)}{\partial \mathbf{b}_0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_0 = \mathbf{0}.^{2} \tag{3-4}$$

Let $\mathbf{b}$ be the solution (assuming it exists). Then, after manipulating (3-4), we find that $\mathbf{b}$ satisfies the **least squares normal equations**,

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{3-5}$$

If the inverse of $\mathbf{X}'\mathbf{X}$ exists, which follows from the full column rank assumption (Assumption A2 in Section 2.3), then the solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3-6}$$

For this solution to minimize the sum of squares, the second derivatives matrix,

$$\frac{\partial^2 S(\mathbf{b}_0)}{\partial \mathbf{b}_0\, \partial \mathbf{b}_0'} = 2\mathbf{X}'\mathbf{X},$$

must be a positive definite matrix. Let $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$ for some arbitrary nonzero vector $\mathbf{c}$. (The multiplication by 2 is irrelevant.) Then

$$q = \mathbf{v}'\mathbf{v} = \sum_{i=1}^{n} v_i^2, \quad \text{where } \mathbf{v} = \mathbf{X}\mathbf{c}.$$

Unless every element of $\mathbf{v}$ is zero, q is positive. But if $\mathbf{v}$ could be zero, then $\mathbf{v}$ would be a linear combination of the columns of $\mathbf{X}$ that equals $\mathbf{0}$, which contradicts Assumption A2, that $\mathbf{X}$ has full column rank. Because $\mathbf{c}$ is arbitrary, q is positive for every nonzero $\mathbf{c}$, which establishes that $2\mathbf{X}'\mathbf{X}$ is positive definite. Therefore, if $\mathbf{X}$ has full column rank, then the least squares solution $\mathbf{b}$ is unique and minimizes the sum of squared residuals.

3.2.2 APPLICATION: AN INVESTMENT EQUATION

To illustrate the computations in a multiple regression, we consider an example based on the macroeconomic data in Appendix Table F3.1. To estimate an investment equation, we first convert the investment series in Table F3.1 to real terms by dividing them by the GDP deflator and then scale the series so that they are measured in trillions of dollars. The real GDP series is the quantity index reported in the Economic Report of the President (2016). The other variables in the regression are a time trend $(1, 2, \ldots)$, an interest rate (the prime rate), and the yearly rate of inflation in the Consumer Price Index. These produce the data matrices listed in Table 3.1. Consider first a regression of real investment on a constant, the time trend, and real GDP, which correspond to x_1, x_2,

[2] See Appendix A.8 for discussion of calculus results involving matrices and vectors.

TABLE 3.1 Data Matrices

Real Investment (Y)	*Constant (1)*	*Trend (T)*	*Real GDP (G)*	*Interest Rate (R)*	*Inflation Rate (P)*
2.484	1	1	87.1	9.23	3.4
2.311	1	2	88.0	6.91	1.6
2.265	1	3	89.5	4.67	2.4
2.339	1	4	92.0	4.12	1.9
2.556	1	5	95.5	4.34	3.3
2.759	1	6	98.7	6.19	3.4
2.828	1	7	101.4	7.96	2.5
$\mathbf{y} =$ 2.717	$\mathbf{X} =$ 1	8	103.2	8.05	4.1
2.445	1	9	102.9	5.09	0.1
1.878	1	10	100.0	3.25	2.7
2.076	1	11	102.5	3.25	1.5
2.168	1	12	104.2	3.25	3.0
2.356	1	13	105.6	3.25	1.7
2.482	1	14	109.0	3.25	1.5
2.637	1	15	111.6	3.25	0.8

Notes:
1. Data from 2000–2014 obtained from Tables B-3, B-10, and B17 from Economic Report of the President: https://www.whitehouse.gov/sites/default/files/docs/2015_erp_appendix_b.pdf.
2. Results are based on the values shown. Slightly different results are obtained if the raw data on investment and the GNP deflator in Table F3.1 are input to the computer program and used to compute real investment = gross investment/(0.01*GNP deflator) internally.

and x_3. (For reasons to be discussed in Chapter 21, this is probably not a well-specified equation for these macroeconomic variables. It will suffice for a simple numerical example, however.) Inserting the specific variables of the example into (3-5), we have

$$\begin{aligned} b_1 n &+ b_2\Sigma_i T_i &+ b_3\Sigma_i G_i &= \Sigma_i Y_i, \\ b_1\Sigma_i T_i &+ b_2\Sigma_i T_i^2 &+ b_3\Sigma_i T_i G_i &= \Sigma_i T_i Y_i, \\ b_1\Sigma_i G_i &+ b_2\Sigma_i T_i G_i &+ b_3\Sigma_i G_i^2 &= \Sigma_i G_i Y_i. \end{aligned}$$

A solution for b_1 can be obtained by dividing the first equation by n and rearranging it to obtain

$$\begin{aligned} b_1 &= \overline{Y} - b_2\overline{T} - b_3\overline{G} \\ &= 2.41882 - b_2 \times 8 - b_3 \times 99.4133. \end{aligned} \quad \textbf{(3-7)}$$

Insert this solution in the second and third equations, and rearrange terms again to yield a set of two equations:

$$\begin{aligned} b_2\Sigma_i(T_i - \overline{T})^2 &+ b_3\Sigma_i(T_i - \overline{T})(G_i - \overline{G}) &= \Sigma_i(T_i - \overline{T})(Y_i - \overline{Y}), \\ b_2\Sigma_i(G_i - \overline{G})(T_i - \overline{T}) &+ b_3\Sigma_i(G_i - \overline{G})^2 &= \Sigma_i(G_i - \overline{G})(Y_i - \overline{Y}). \end{aligned}$$

This result shows the nature of the solution for the slopes, which can be computed from the sums of squares and cross products of the deviations of the variables from their

means. Letting lowercase letters indicate variables measured as deviations from the sample means, we find that the normal equations are

$$\begin{aligned} b_2\Sigma_i t_i^2 \quad &+ \quad b_3\Sigma_i t_i g_i \quad = \quad \Sigma_i t_i y_i, \\ b_2\Sigma_i g_i t_i \quad &+ \quad b_3\Sigma_i g_i^2 \quad = \quad \Sigma_i g_i y_i, \end{aligned}$$

and the least squares solutions for b_2 and b_3 are

$$b_2 = \frac{\Sigma_i t_i y_i \Sigma_i g_i^2 - \Sigma_i g_i y_i \Sigma_i t_i g_i}{\Sigma_i t_i^2 \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{-1.6351(792.857) - 4.22255(451.9)}{280(792.857) - (451.9)^2} = -0.180169,$$

$$b_3 = \frac{\Sigma_i g_i y_i \Sigma_i t_i^2 - \Sigma_i t_i y_i \Sigma_i t_i g_i}{\Sigma_i t_i^2 \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{4.22255(280) - (-1.6351)(451.9)}{280(792.857) - (451.9)^2} = 0.1080157. \quad \textbf{(3-8)}$$

With these solutions in hand, b_1 can now be computed using (3-7); $b_1 = -6.8780284$.

Suppose that we just regressed investment on the constant and GDP, omitting the time trend. At least some of the correlation between real investment and real GDP that we observe in the data will be explainable because both variables have an obvious time trend. (The trend in investment clearly has two parts, before and after the crash of 2007–2008.) Consider how this shows up in the regression computation. Denoting by "b_{yx}" the slope in the simple, **bivariate regression** of variable y on a constant and the variable x, we find that the slope in this reduced regression would be

$$b_{YG} = \frac{\Sigma_i g_i y_i}{\Sigma_i g_i^2} = 0.00533. \quad \textbf{(3-9)}$$

By manipulating the earlier expression for b_3 and using the definition of the sample correlation between G and T, $r_{GT}^2 = (\Sigma_i g_i t_i)^2/(\Sigma_i g_i^2 \Sigma_i t_i^2)$, we obtain

$$b_{YG|T} = \frac{b_{YG}}{1 - r_{GT}^2} - \frac{b_{YT} b_{TG}}{1 - r_{GT}^2} = b_{YG} - \left(\frac{b_{YT} b_{TG} - r_{GT}^2 b_{YG}}{1 - r_{GT}^2} \right) = 0.1080157. \quad \textbf{(3-10)}$$

(The notation "$b_{YG|T}$" used on the left-hand side is interpreted to mean the slope in the regression of Y on G and a constant "in the presence of T.") The slope in the **multiple regression** differs from that in the simple regression by a factor of 20, by including a correction that accounts for the influence of the additional variable T on both Y and G. For a striking example of this effect, in the simple regression of real investment on a time trend, $b_{YT} = -1.6351/280 = -0.00584$. But, in the multiple regression, after we account for the influence of GNP on real investment, the slope on the time trend is -0.180169. The general result for a three-variable regression in which x_1 is a constant term is

$$b_{Y2|3} = \frac{b_{Y2} - b_{Y3} b_{32}}{1 - r_{23}^2}. \quad \textbf{(3-11)}$$

It is clear from this expression that the magnitudes of $b_{y2|3}$ and b_{y2} can be quite different. They need not even have the same sign. The result just seen is worth emphasizing; the coefficient on a variable in the simple regression [e.g., Y on (1,G)] will generally not be the same as the one on that variable in the multiple regression [e.g., $>Y$ on (1,T,G)] if the new variable and the old one are correlated. But, note that b_{YG} in (3-9) *will* be the same as $b_3 = b_{YG|T}$ in (3-8) if $\Sigma_i t_i g_i = 0$, that is, if T and G are not correlated.

In practice, you will never actually compute a multiple regression by hand or with a calculator. For a regression with more than three variables, the tools of matrix algebra are indispensable (as is a computer). Consider, for example, an enlarged model of investment that includes—in addition to the constant, time trend, and GDP—an interest rate and the rate of inflation. Least squares requires the simultaneous solution of five normal equations. Letting **X** and **y** denote the full data matrices shown previously, the normal equations in (3-5) are

$$\begin{bmatrix} 15.000 & 120.00 & 1491.2 & 76.05 & 33.90 \\ 120.000 & 1240.0 & 12381.5 & 522.06 & 244.10 \\ 1491.2 & 12381.5 & 149038 & 7453.03 & 3332.83 \\ 76.06 & 522.06 & 7453.03 & 446.323 & 186.656 \\ 33.90 & 244.10 & 3332.83 & 186.656 & 93.33 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 36.28230 \\ 288.624 \\ 3611.17 \\ 188.176 \\ 82.7731 \end{bmatrix}.$$

The solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (-6.25441, -0.161342, 0.0994684, 0.0196656, -0.0107206)'.$$

3.2.3 ALGEBRAIC ASPECTS OF THE LEAST SQUARES SOLUTION

The normal equations are

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = -\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = -\mathbf{X}'\mathbf{e} = \mathbf{0}. \tag{3-12}$$

Hence, for every column $\mathbf{x}_k$ of $\mathbf{X}$, $\mathbf{x}_k'\mathbf{e} = 0$. If the first column of $\mathbf{X}$ is a column of 1s, which we denote $\mathbf{i}$, then there are three implications.

1. *The least squares residuals sum to zero*. This implication follows from $\mathbf{x}_1'\mathbf{e} = \mathbf{i}'\mathbf{e} = \Sigma_i e_i = 0$.
2. *The regression hyperplane passes through the point of means of the data*. The first normal equation implies that $\bar{y} = \bar{\mathbf{x}}'\mathbf{b}$. This follows from $\Sigma_i e_i = \Sigma_i (y_i - \mathbf{x}_i'\mathbf{b}) = 0$ by dividing by n.
3. *The mean of the fitted values from the regression equals the mean of the actual values*. This implication follows from point 2 because the fitted values are $\mathbf{x}_i'\mathbf{b}$.

It is important to note that none of these results need hold if the regression does not contain a constant term.

3.2.4 PROJECTION

The vector of least squares residuals is

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}. \tag{3-13}$$

Inserting the result in (3-6) for **b** gives

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{M}\mathbf{y}. \tag{3-14}$$

The $n \times n$ matrix **M** defined in (3-14) is fundamental in regression analysis. You can easily show that **M** is both symmetric ($\mathbf{M} = \mathbf{M}'$) and idempotent ($\mathbf{M} = \mathbf{M}^2$). In view of (3-13), we can interpret **M** as a matrix that produces the vector of least squares residuals

in the regression of $\mathbf{y}$ on $\mathbf{X}$ when it premultiplies any vector $\mathbf{y}$. It will be convenient later to refer to this matrix as a "**residual maker**." Matrices of this form will appear repeatedly in our development to follow.

DEFINITION 3.1: Residual Maker
Let the $n \times K$ full column rank matrix, $\mathbf{X}$ be composed of columns $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K)$, and let $\mathbf{y}$ be an $n \times 1$ column vector. The matrix, $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a "residual maker" in that when $\mathbf{M}$ premultiplies a vector, $\mathbf{y}$, the result, $\mathbf{My}$, is the column vector of residuals in the least squares regression of $\mathbf{y}$ on $\mathbf{X}$.

It follows from the definition that

$$\mathbf{MX} = \mathbf{0}, \tag{3-15}$$

because if a column of $\mathbf{X}$ is regressed on $\mathbf{X}$, a perfect fit will result and the residuals will be zero.

Result (3-13) implies that $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, which is the sample analog to Assumption A1, (2-3). (See Figure 3.1 as well.) The least squares results partition $\mathbf{y}$ into two parts, the fitted values $\hat{\mathbf{y}} = \mathbf{Xb}$ and the residuals, $\mathbf{e} = \mathbf{My}$. [See Section A.3.7, especially (A-54).] Because $\mathbf{MX} = \mathbf{0}$, these two parts are orthogonal. Now, given (3-13),

$$\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = \mathbf{Iy} - \mathbf{My} = (\mathbf{I} - \mathbf{M})\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Py}. \tag{3-16}$$

The matrix $\mathbf{P}$ is a **projection matrix**. It is the matrix formed from $\mathbf{X}$ such that when a vector $\mathbf{y}$ is premultiplied by $\mathbf{P}$, the result is the fitted values in the least squares regression of $\mathbf{y}$ on $\mathbf{X}$. This is also the **projection** of the vector $\mathbf{y}$ into the column space of $\mathbf{X}$. (See Sections A3.5 and A3.7.) By multiplying it out, you will find that, like $\mathbf{M}$, $\mathbf{P}$ is symmetric and idempotent. Given the earlier results, it also follows that $\mathbf{M}$ and $\mathbf{P}$ are orthogonal;

$$\mathbf{PM} = \mathbf{MP} = \mathbf{0}.$$

As might be expected from (3-15),

$$\mathbf{PX} = \mathbf{X}.$$

As a consequence of (3-14) and (3-16), we can see that least squares partitions the vector $\mathbf{y}$ into two orthogonal parts,

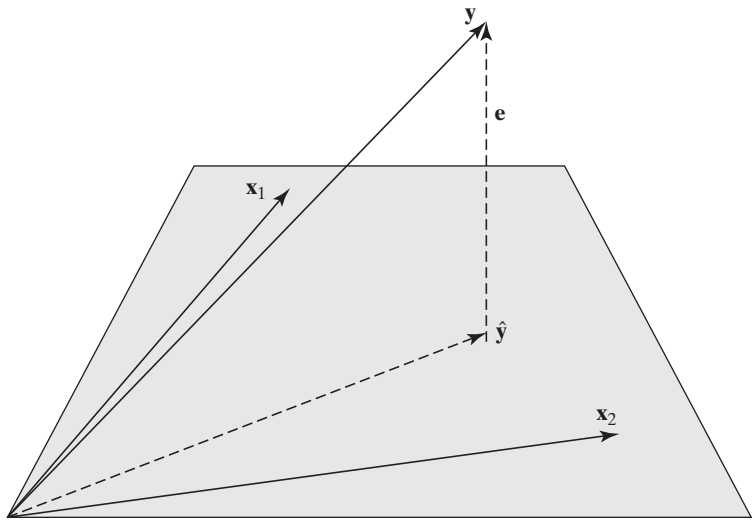
$$\mathbf{y} = \mathbf{Py} + \mathbf{My} = \textbf{projection} + \textbf{residual}.$$

The result is illustrated in Figure 3.2 for the two-variable case. The gray-shaded plane is the column space of $\mathbf{X}$. The projection and residual are the orthogonal dashed rays. We can also see the Pythagorean theorem at work in the sums of squares,

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= \mathbf{y}'\mathbf{P}'\mathbf{Py} + \mathbf{y}'\mathbf{M}'\mathbf{My} \\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}.\end{aligned}$$

The sample linear projection of y on $\mathbf{x}$, $Proj(y|\mathbf{x})$, is an extremely useful device in empirical research. Linear least squares regression is often the starting point for model development. We will find in developing the regression model that if the population conditional mean function in Assumption A1, $E[y|\mathbf{x}]$, is linear in $\mathbf{x}$, then $E[y|\mathbf{x}]$ is also

FIGURE 3.2 Projection of **y** into the Column Space of **X.**



the population counterpart to the projection of y on **x.** We will be able to show that $Proj(y|\mathbf{x})$ estimates $\mathbf{x}'\{E[\mathbf{x}\mathbf{x}']\}^{-1}E[\mathbf{x}y]$, which appears implicitly in (3-16), is also $E[y|\mathbf{x}]$. If the conditional mean function is not linear in **x**, then the projection of y on **x** will still estimate a useful descriptor of the joint distribution of y and **x**.

3.3 PARTITIONED REGRESSION AND PARTIAL REGRESSION

It is common to specify a multiple regression model when, in fact, interest centers on only one or a subset of the full set of variables—the remaining variables are often viewed as "controls." Consider the earnings equation discussed in the Introduction. Although we are primarily interested in the effect of education on earnings, age is, of necessity, included in the model. The question we consider here is what computations are involved in obtaining, in isolation, the coefficients of a subset of the variables in a multiple regression (e.g., the coefficient of education in the aforementioned regression).

Suppose that the regression involves two sets of variables, $\mathbf{X}_1$ and $\mathbf{X}_2$. Thus,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

What is the algebraic solution for $\mathbf{b}_2$? The **normal equations** are

$$\begin{matrix}(1)\\(2)\end{matrix}\quad\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1'\mathbf{y}\\ \mathbf{X}_2'\mathbf{y}\end{bmatrix}. \tag{3-17}$$

A solution can be obtained by using the partitioned inverse matrix of (A-74). Alternatively, (1) and (2) in (3-17) can be manipulated directly to solve for $\mathbf{b}_2$. We first solve (1) for $\mathbf{b}_1$:

$$\begin{aligned}&\mathbf{X}_1'\mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_1'\mathbf{y},\\ &\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2).\end{aligned} \tag{3-18}$$

This solution states that $\mathbf{b}_1$ is the set of coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}_1$, minus a correction vector. We digress briefly to examine an important result embedded in (3-18). Suppose that $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$. Then, $\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$, which is simply the coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}_1$. The general result is given in the following theorem.

THEOREM 3.1 Orthogonal Partitioned Regression
In the linear least squares multiple regression of $\mathbf{y}$ *on two sets of variables* $\mathbf{X}_1$ *and* $\mathbf{X}_2$, *if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of* $\mathbf{y}$ *on* $\mathbf{X}_1$ *alone and* $\mathbf{y}$ *on* $\mathbf{X}_2$ *alone.*
Proof: *The assumption of the theorem is that* $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ *in the normal equations in (3-17). Inserting this assumption into (3-18) produces the immediate solution for* $\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ *and likewise for* $\mathbf{b}_2$.

If the two sets of variables $\mathbf{X}_1$ and $\mathbf{X}_2$ are not orthogonal, then the solutions for $\mathbf{b}_1$ and $\mathbf{b}_2$ found by (3-17) and (3-18) are more involved than just the simple regressions in Theorem 3.1. The more general solution is suggested by the following theorem:

THEOREM 3.2 Frisch–Waugh (1933)–Lovell (1963) Theorem[3]
In the linear least squares regression of vector $\mathbf{y}$ *on two sets of variables,* $\mathbf{X}_1$ *and* $\mathbf{X}_2$, *the subvector* $\mathbf{b}_2$ *is the set of coefficients obtained when the residuals from a regression of* $\mathbf{y}$ *on* $\mathbf{X}_1$ *alone are regressed on the set of residuals obtained when each column of* $\mathbf{X}_2$ *is regressed on* $\mathbf{X}_1$.

To prove Theorem 3.2, begin from equation (2) in (3-17), which is

$$\mathbf{X}_2'\mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{y}.$$

Now, insert the result for $\mathbf{b}_1$ that appears in (3-18) into this result. This produces

$$\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{y}.$$

After collecting terms, the solution is

$$\begin{aligned}\mathbf{b}_2 &= [\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2]^{-1}[\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}] \\ &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2'\mathbf{M}_1\mathbf{y}). \qquad \textbf{(3-19)}\end{aligned}$$

[3] The theorem, such as it was, appeared in the first volume of *Econometrica*, in the introduction to the paper: "The partial trend regression method can never, indeed, achieve anything which the individual trend method cannot, because the two methods lead by definition to identically the same results." Thus, Frisch and Waugh were concerned with the (lack of) difference between a regression of a variable $\mathbf{y}$ on a time trend variable, $\mathbf{t}$, and another variable, $\mathbf{x}$, compared to the regression of a detrended $\mathbf{y}$ on a detrended $\mathbf{x}$, where detrending meant computing the residuals of the respective variable on a constant and the time trend, $\mathbf{t}$. A concise statement of the theorem and its matrix formulation were added later by Lovell (1963).

The $\mathbf{M}_1$ matrix appearing in the parentheses inside each set of parentheses is the "residual maker" defined in (3-14) and Definition 3.1, in this case defined for a regression on the columns of $\mathbf{X}_1$. Thus, $\mathbf{M}_1\mathbf{X}_2$ is a matrix of residuals; each column of $\mathbf{M}_1\mathbf{X}_2$ is a vector of residuals in the regression of the corresponding column of $\mathbf{X}_2$ on the variables in $\mathbf{X}_1$. By exploiting the fact that $\mathbf{M}_1$, like $\mathbf{M}$, is symmetric and idempotent, we can rewrite (3-19) as

$$\mathbf{b}_2 = (\mathbf{X}_2^{*\prime}\mathbf{X}_2^{*})^{-1}\mathbf{X}_2^{*\prime}\mathbf{y}_*, \tag{3-20}$$

where $\mathbf{X}_2^* = \mathbf{M}_1\mathbf{X}_2$ and $\mathbf{y}_* = \mathbf{M}_1\mathbf{y}$. This result is fundamental in regression analysis.

This process is commonly called **partialing out** or **netting out** the effect of $\mathbf{X}_1$. For this reason, the coefficients in a multiple regression are often called the **partial regression coefficients**. The application of Theorem 3.2 to the computation of a single coefficient as suggested at the beginning of this section is detailed in the following: Consider the regression of $\mathbf{y}$ on a set of variables $\mathbf{X}$ and an additional variable $\mathbf{z}$. Denote the coefficients $\mathbf{b}$ and c, respectively.

COROLLARY 3.2.1 Individual Regression Coefficients
The coefficient on $\mathbf{z}$ *in a multiple regression of* $\mathbf{y}$ *on* $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$ *is computed as* $c = (\mathbf{z}'\mathbf{M}_\mathbf{X}\mathbf{z})^{-1}(\mathbf{z}'\mathbf{M}_\mathbf{X}\mathbf{y}) = (\mathbf{z}_*'\mathbf{z}_*)^{-1}\mathbf{z}_*'\mathbf{y}_*$ *where* $\mathbf{z}_*$ *and* $\mathbf{y}_*$ *are the residual vectors from least squares regressions of* $\mathbf{z}$ *and* $\mathbf{y}$ *on* $\mathbf{X}$; $\mathbf{z}_* = \mathbf{M}_\mathbf{X}\mathbf{z}$ *and* $\mathbf{y}_* = \mathbf{M}_\mathbf{X}\mathbf{y}$ *where* $\mathbf{M}_\mathbf{X}$ *is defined in (3-14).*
Proof: *This is an application of Theorem 3.2 in which* $\mathbf{X}_1$ *is* $\mathbf{X}$ *and* $\mathbf{X}_2$ *is* $\mathbf{z}$.

In terms of Example 2.2, we could obtain the coefficient on education in the multiple regression by first regressing earnings and education on age (or age and age squared) and then using the residuals from these regressions in a simple regression. In the classic application of this latter observation, Frisch and Waugh (1933) noted that in a time-series setting, the same results were obtained whether a regression was fitted with a time-trend variable or the data were first "detrended" by netting out the effect of time, as noted earlier, and using just the detrended data in a simple regression.

Consider the case in which $\mathbf{X}_1$ is $\mathbf{i}$, a constant term that is a column of 1s in the first column of $\mathbf{X}$, and $\mathbf{X}_2$ is a set of variables. The solution for $\mathbf{b}_2$ in this case will then be the slopes in a regression that contains a constant term. Using Theorem 3.2 the vector of residuals for any variable, $\mathbf{x}$, in $\mathbf{X}_2$ will be

$$\begin{aligned}\mathbf{x}_* &= \mathbf{x} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'\mathbf{x} \\ &= \mathbf{x} - \mathbf{i}(1/n)\mathbf{i}'\mathbf{x} \\ &= \mathbf{x} - \mathbf{i}\bar{x} \\ &= \mathbf{M}^0\mathbf{x}. \end{aligned} \tag{3-21}$$

(See Section A.5.4 where we have developed this result purely algebraically.) For this case, then, the residuals are deviations from the sample mean. Therefore, each column of $\mathbf{M}_1\mathbf{X}_2$ is the original variable, now in the form of deviations from the mean. This general result is summarized in the following corollary.

COROLLARY 3.2.2 Regression with a Constant Term
The slopes in a multiple regression that contains a constant term can be obtained by transforming the data to deviations from their means and then regressing the variable y in deviation form on the explanatory variables, also in deviation form.

[We used this result in (3-8).] Having obtained the coefficients on $\mathbf{X}_2$, how can we recover the coefficients on $\mathbf{X}_1$ (the constant term)? One way is to repeat the exercise while reversing the roles of $\mathbf{X}_1$ and $\mathbf{X}_2$. But there is an easier way. We have already solved for $\mathbf{b}_2$. Therefore, we can use (3-18) in a solution for $\mathbf{b}_1$. If $\mathbf{X}_1$ is just a column of 1s, then the first of these produces the familiar result

$$b_1 = \bar{y} - \bar{x}_2 b_2 - \cdots - \bar{x}_K b_K$$

[which is used in (3-7)].

Theorem 3.2 and Corollaries 3.2.1 and 3.2.2 produce a useful interpretation of the **partitioned regression** when the model contains a constant term. According to Theorem 3.1, if the columns of $\mathbf{X}$ are orthogonal, that is, $\mathbf{X}_k'\mathbf{x}_m = 0$ for columns k and m, then the separate regression coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}$ when $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$ are simply $\mathbf{x}_k'\mathbf{y}/\mathbf{x}_k'\mathbf{x}_k$. When the regression contains a constant term, we can compute the multiple regression coefficients by regression of $\mathbf{y}$ in mean deviation form on the columns of $\mathbf{X}$, also in deviations from their means. In this instance, the *orthogonality* of the columns means that the sample covariances (and correlations) of the variables are zero. The result is another theorem:

THEOREM 3.3 Orthogonal Regression
If the multiple regression of $\mathbf{y}$ *on* $\mathbf{X}$ *contains a constant term and the variables in the regression are uncorrelated, then the multiple regression slopes are the same as the slopes in the individual simple regressions of* $\mathbf{y}$ *on a constant and each variable in turn.*
Proof: *The result follows from Theorems 3.1 and 3.2.*

3.4 PARTIAL REGRESSION AND PARTIAL CORRELATION COEFFICIENTS

The use of multiple regression involves a conceptual experiment that we might not be able to carry out in practice, the *ceteris paribus* analysis familiar in economics. To pursue the earlier example, a regression equation relating earnings to age and education enables us to do the experiment of comparing the earnings of two individuals of the same age with different education levels, *even if the sample contains no such pair of individuals*. It is this characteristic of the regression that is implied by the term partial regression coefficients. The way we obtain this result, as we have seen, is first to regress income and education on age and then to compute the residuals from this regression. By construction, age will not have any power in explaining variation in these residuals. Therefore, any

correlation between income and education after this "purging" is independent of (or after removing the effect of) age.

The same principle can be applied to the correlation between two variables. To continue our example, to what extent can we assert that this correlation reflects a direct relationship rather than that both income and education tend, on average, to rise as individuals become older? To find out, we would use a **partial correlation coefficient**, which is computed along the same lines as the partial regression coefficient. In the context of our example, the partial correlation coefficient between income and education, controlling for the effect of age, is obtained as follows:

1. $y_* =$ the residuals in a regression of income on a constant and age.
2. $z_* =$ the residuals in a regression of education on a constant and age.
3. The partial correlation r^*_{yz} is the simple correlation between y_* and z_*.

This calculation might seem to require a large amount of computation. Using Corollary 3.2.1, the two residual vectors in points 1 and 2 are $\mathbf{y}_* = \mathbf{My}$ and $\mathbf{z}_* = \mathbf{Mz}$ where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the residual maker defined in (3-14). We will assume that there is a constant term in $\mathbf{X}$ so that the vectors of residuals $\mathbf{y}_*$ and $\mathbf{z}_*$ have zero sample means. Then, the square of the partial correlation coefficient is

$$r^{*2}_{yz} = \frac{(\mathbf{z}'_*\mathbf{y}_*)^2}{(\mathbf{z}'_*\mathbf{z}_*)(\mathbf{y}'_*\mathbf{y}_*)}.$$

There is a convenient shortcut. Once the multiple regression is computed, the t ratio in (5-13) for testing the hypothesis that the coefficient equals zero (e.g., the last column of Table 4.6) can be used to compute

$$r^{*2}_{yz} = \frac{t_z^2}{t_z^2 + \text{degrees of freedom}}, \tag{3-22}$$

where the **degrees of freedom** is equal to $n-(K + 1)$; $K+1$ is the number of variables in the regression plus the constant term. The proof of this less than perfectly intuitive result will be useful to illustrate some results on partitioned regression. We will rely on two useful theorems from least squares algebra. The first isolates a particular diagonal element of the inverse of a **moment matrix** such as $(\mathbf{X}'\mathbf{X})^{-1}$.

THEOREM 3.4 Diagonal Elements of the Inverse of a Moment Matrix
Let $\mathbf{W}$ *denote the partitioned matrix* $[\mathbf{X}, \mathbf{z}]$*—that is, the* K *columns of* $\mathbf{X}$ *plus an additional column labeled* $\mathbf{z}$*. The last diagonal element of* $(\mathbf{W}'\mathbf{W})^{-1}$ *is* $(\mathbf{z}'\mathbf{M}_\mathbf{X}\mathbf{z})^{-1} = (\mathbf{z}'_*\mathbf{z}_*)^{-1}$ *where* $\mathbf{z}_* = \mathbf{M}_\mathbf{X}\mathbf{z}$ *and* $\mathbf{M}_\mathbf{X} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
Proof: *This is an application of the partitioned inverse formula in (A-74) where* $A_{11} = \mathbf{X}'\mathbf{X}$, $A_{12} = \mathbf{X}'\mathbf{z}$, $A_{21} = \mathbf{z}'\mathbf{X}$ *and* $A_{22} = \mathbf{z}'\mathbf{z}$*. Note that this theorem generalizes the development in Section A.2.8, where* $\mathbf{X}$ *contains only a constant term,* $\mathbf{i}$.

We can use Theorem 3.4 to establish the result in (3-22). Let c and $\mathbf{u}$ denote the coefficient on $\mathbf{z}$ and the vector of residuals in the multiple regression of $\mathbf{y}$ on $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$, respectively. Then, by definition, the squared t ratio that appears in (3-22) is

$$t_z^2 = \frac{c^2}{\left[\dfrac{\mathbf{u}'\mathbf{u}}{n-(K+1)}\right](\mathbf{W}'\mathbf{W})^{-1}_{K+1,K+1}},$$

where $(\mathbf{W}'\mathbf{W})^{-1}_{K+1,K+1}$ is the $(K+1)$ (last) diagonal element of $(\mathbf{W}'\mathbf{W})^{-1}$. [The bracketed term appears in (4-17).] The theorem states that this element of the matrix equals $(\mathbf{z}_*'\mathbf{z}_*)^{-1}$. From Corollary 3.2.1, we also have that $c^2 = [(\mathbf{z}_*'\mathbf{y}_*)/(\mathbf{z}_*'\mathbf{z}_*)]^2$. For convenience, let $DF = n - (K+1)$. Then, $t_z^2 = \dfrac{(\mathbf{z}_*'\mathbf{y}_*/\mathbf{z}_*'\mathbf{z}_*)^2}{(\mathbf{u}'\mathbf{u}/DF)(\mathbf{z}_*'\mathbf{z}_*)^{-1}} = \dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)^{-1}}$. It follows that the result in (3-22) is equivalent to

$$\frac{t_z^2}{t_z^2+DF} = \frac{\dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)}}{\dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)}+DF} = \frac{\dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)}}{\dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)}+1} = \frac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{z}_*'\mathbf{y}_*)^2+(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)}.$$

Divide numerator and denominator by $(\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*)$ to obtain

$$\frac{t_z^2}{t_z^2+DF} = \frac{(\mathbf{z}_*'\mathbf{y}_*)^2/((\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*))}{(\mathbf{z}_*'\mathbf{y}_*)^2/((\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*))+((\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*))/((\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*))} = \frac{r_{yz}^{*2}}{r_{yz}^{*2}+(\mathbf{u}'\mathbf{u})/(\mathbf{y}_*'\mathbf{y}_*)}. \tag{3-23}$$

We will now use a second theorem to manipulate $\mathbf{u}'\mathbf{u}$ and complete the derivation. The result we need is given in Theorem 3.5.

Returning to the derivation, then, $\mathbf{e}'\mathbf{e} = \mathbf{y}_*'\mathbf{y}_*$ and $c^2(\mathbf{z}_*'\mathbf{z}_*) = (\mathbf{z}_*'\mathbf{y}_*)^2/(\mathbf{z}_*'\mathbf{z}_*)$. Therefore,

$$\frac{\mathbf{u}'\mathbf{u}}{\mathbf{y}_*'\mathbf{y}_*} = \frac{\mathbf{y}_*'\mathbf{y}_* - (\mathbf{z}_*'\mathbf{y}_*)^2/\mathbf{z}_*'\mathbf{z}_*}{\mathbf{y}_*'\mathbf{y}_*} = 1 - r_{yz}^{*2}.$$

Inserting this in the denominator of (3-23) produces the result we sought.

THEOREM 3.5 Change in the Sum of Squares When a Variable Is Added to a Regression

If $\mathbf{e}'\mathbf{e}$ *is the sum of squared residuals when* $\mathbf{y}$ *is regressed on* $\mathbf{X}$ *and* $\mathbf{u}'\mathbf{u}$ *is the sum of squared residuals when* $\mathbf{y}$ *is regressed on* $\mathbf{X}$ *and* $\mathbf{z}$*, then*

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}_*'\mathbf{z}_*) \leq \mathbf{e}'\mathbf{e}, \tag{3-24}$$

where c *is the coefficient on* $\mathbf{z}$ *in the long regression of* $\mathbf{y}$ *on* $[\mathbf{X}, \mathbf{z}]$ *and* $\mathbf{z}_* = \mathbf{Mz}$ *is the vector of residuals when* $\mathbf{z}$ *is regressed on* $\mathbf{X}$*.*

Proof: *In the long regression of* $\mathbf{y}$ *on* $\mathbf{X}$ *and* $\mathbf{z}$*, the vector of residuals is* $\mathbf{u} = \mathbf{y} - \mathbf{Xd} - \mathbf{zc}$*. Note that unless* $\mathbf{X}'\mathbf{z} = \mathbf{0}$*,* $\mathbf{d}$ *will not equal* $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$*. (See Section 4.3.2.) Moreover, unless* $c = 0$*,* $\mathbf{u}$ *will not equal* $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$*. From Corollary 3.2.1,* $c = (\mathbf{z}_*'\mathbf{z}_*)^{-1}(\mathbf{z}_*'\mathbf{y}_*)$*. From (3-18), we also have that the coefficients on* $\boldsymbol{X}$ *in this long regression are*

$$\mathbf{d} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{z}c) = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c.$$

Inserting this expression for $\mathbf{d}$ *in that for* $\mathbf{u}$ *gives*

$$\mathbf{u} = \mathbf{y} - \mathbf{Xb} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c - \mathbf{z}c = \mathbf{e} - \mathbf{M}_{\mathbf{X}}\mathbf{z}c = \mathbf{e} - \mathbf{z}_*c.$$

Then,

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} + c^2(\mathbf{z}_*'\mathbf{z}_*) - 2c(\mathbf{z}_*'\mathbf{e}).$$

But, $\mathbf{e} = \mathbf{M}_{\mathbf{x}}\mathbf{y} = \mathbf{y}_*$ *and* $\mathbf{z}_*'\mathbf{e} = \mathbf{z}_*'\mathbf{y}_* = c(\mathbf{z}_*'\mathbf{z}_*)$. *Inserting this result in* $\mathbf{u}'\mathbf{u}$ *immediately above gives the result in the theorem.*

Example 3.1 Partial Correlations

For the data in the application in Section 3.2.2, the simple correlations between investment and the regressors, r_{Yk}, and the partial correlations, r^*_{Yk}, between investment and the four regressors (given the other variables) are listed in Table 3.2. As is clear from the table, there is no necessary relation between the simple and partial correlation coefficients. One thing worth noting is that the signs of the partial correlations are the same as those of the coefficients, but not necessarily the same as the signs of the raw correlations. Note the difference in the coefficient on *Inflation*.

3.5 GOODNESS OF FIT AND THE ANALYSIS OF VARIANCE

The original fitting criterion, the sum of squared residuals, suggests a measure of the fit of the regression line to the data. However, as can easily be verified, the sum of squared residuals can be scaled arbitrarily just by multiplying all the values of y by the desired scale factor. Because the fitted values of the regression are based on the values of $\mathbf{x}$, we might ask instead whether *variation* in $\mathbf{x}$ is a good predictor of *variation* in y. Figure 3.3 shows three possible cases for a simple linear regression model, $y = \beta_1 + \beta_2 x + \varepsilon$. The measure of fit described here embodies both the fitting criterion and the covariation of y and $\mathbf{x}$.

Variation of the dependent variable is defined in terms of deviations from its mean, $(y_i - \bar{y})$. The **total variation** in y is the sum of squared deviations:

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

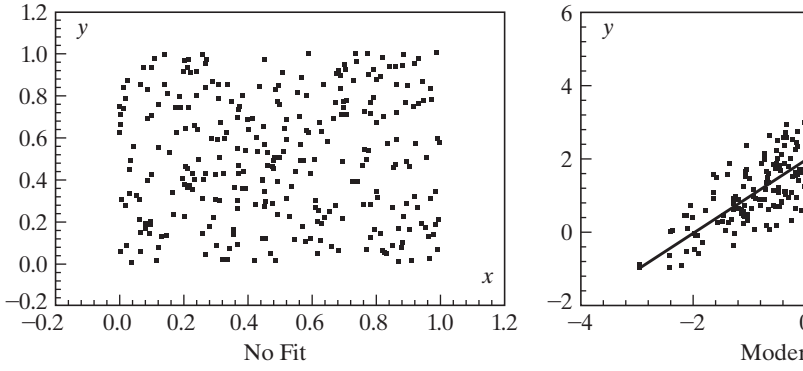
In terms of the regression equation, we may write the full set of observations as

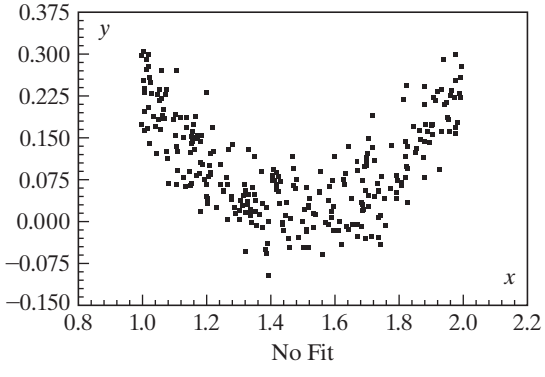
$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}.$$

TABLE 3.2 Correlations of Investment with Other Variables (DF = 10)

Variable	*Coefficient*	*t Ratio*	*Simple Correlation*	*Partial Correlation*
Trend	–0.16134	–3.42	–0.09965	–0.73423
RealGDP	0.09947	4.12	0.15293	0.79325
Interest	0.01967	0.58	0.55006	0.18040
Inflation	–0.01072	–0.27	0.19332	–0.08507

FIGURE 3.3 Sample Data.





For an individual observation, we have

$$y_i = \hat{y}_i + e_i = \mathbf{x}_i'\mathbf{b} + e_i.$$

If the regression contains a constant term, then the residuals will sum to zero and the mean of the predicted values of y_i will equal the mean of the actual values. Subtracting $\bar{y}$ from both sides and using this result and result 2 in Section 3.2.3 gives

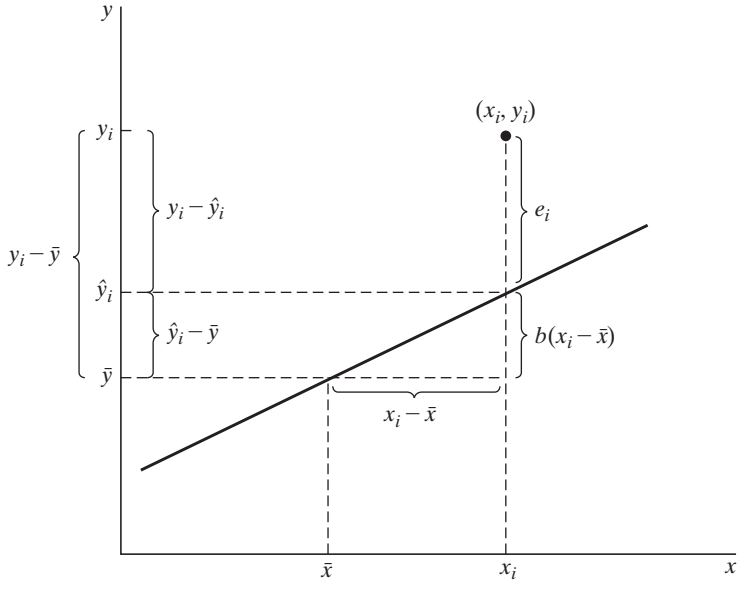
$$y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{b} + e_i.$$

Figure 3.4 illustrates the computation for the two-variable regression. Intuitively, the regression would appear to fit well if the deviations of y from its mean are more largely accounted for by deviations of x from its mean than by the residuals. Since both terms in this decomposition sum to zero, to quantify this fit, we use the sums of squares instead. For the full set of observations, we have

$$\mathbf{M}^0\mathbf{y} = \mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{M}^0\mathbf{e},$$

where $\mathbf{M}^0$ is the $n \times n$ idempotent matrix that transforms observations into deviations from sample means. [See (3-21)and Section A.2.8; $\mathbf{M}^0$ is a residual maker for $\mathbf{X} = \mathbf{i}$.] The column of $\mathbf{M}^0\mathbf{X}$ corresponding to the constant term is zero, and, since the residuals

FIGURE 3.4 Decomposition of y_i.



already have mean zero, $\mathbf{M}^0\mathbf{e} = \mathbf{e}$. Then, since $\mathbf{e}'\mathbf{M}^0\mathbf{X} = \mathbf{e}'\mathbf{X} = \mathbf{0}$, the total sum of squares is

$$\mathbf{y}'\mathbf{M}^0\mathbf{y} = \mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{e}'\mathbf{e}.$$

Write this as total sum of squares = regression sum of squares + error sum of squares, or

$$\text{SST} = \text{SSR} + \text{SSE}. \tag{3-25}$$

(Note that this is the same partitioning that appears at the end of Section 3.2.4.)

We can now obtain a measure of how well the regression line fits the data by using the

coefficient of determination:
$$\frac{\text{SSR}}{\text{SST}} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \frac{\Sigma_{i=1}^n e_i^2}{\Sigma_{i=1}^n (y_i - \bar{y})^2}. \tag{3-26}$$

The coefficient of determination is denoted R^2. As we have shown, it must be between 0 and 1, and it measures the proportion of the total variation in y that is accounted for by variation in the regressors. It equals zero if the regression is a horizontal line, that is, if all the elements of **b** except the constant term are zero. In this case, the predicted values of y are always $\bar{y}$, so deviations of **x** from its mean do not translate into different predictions for y. As such, **x** has no explanatory power. The other extreme, $R^2 = 1$, occurs if the values of **x** and y all lie in the same hyperplane (on a straight line for a two-variable regression) so that the residuals are all zero. If all the values of y_i lie on a vertical line, then R^2 has no meaning and cannot be computed.

Regression analysis is often used for forecasting. In this case, we are interested in how well the regression model predicts movements in the dependent variable. With this in mind, an equivalent way to compute R^2 is also useful. First, the sum of squares for the predicted values is

$$\Sigma_{i=1}^{n} (\hat{y}_i - \hat{\bar{y}})^2 = \hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}} = \mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b},$$

but $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$, $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$, $\mathbf{M}^0\mathbf{e} = \mathbf{e}$, and $\mathbf{X}'\mathbf{e} = \mathbf{0}$, so $\hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}} = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y} = \Sigma_{i=1}^{n} (\hat{y}_i - \hat{\bar{y}})(y_i - \bar{y})$. Multiply $R^2 = \hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}}/\mathbf{y}'\mathbf{M}^0\mathbf{y} = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ by $1 = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y}/\hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}}$ to obtain

$$R^2 = \frac{[\Sigma_i(y_i - \bar{y})(\hat{y}_i - \hat{\bar{y}})]^2}{[\Sigma_i(y_i - \bar{y})^2][\Sigma_i(\hat{y}_i - \hat{\bar{y}})^2]}, \tag{3-27}$$

which is the squared correlation between the observed values of y and the predictions produced by the estimated regression equation.

Example 3.2 Fit of a Consumption Function

The data plotted in Figure 2.1 are listed in Appendix Table F2.1. For these data, where y is C and x is X, we have $\bar{y} = 273.2727$, $\bar{x} = 323.2727$, $S_{yy} = 12{,}618.182$, $S_{xx} = 12{,}300.182$, and $S_{xy} = 8{,}423.182$, so SST $= 12{,}618.182$, $b = 8{,}423.182/12{,}300.182 = 0.6848014$, SSR $= b^2S_{xx} = 5{,}768.2068$, and SSE $=$ SST $-$ SSR $= 6{,}849.975$. Then $R^2 = b^2S_{xx} = 0.457135$. As can be seen in Figure 2.1, this is a moderate fit, although it is not particularly good for aggregate time-series data. On the other hand, it is clear that not accounting for the anomalous wartime data has degraded the fit of the model. This value is the R^2 for the model indicated by the solid line in the figure. By simply omitting the years 1942–1945 from the sample and doing these computations with the remaining seven observations—the dashed line—we obtain an R^2 of 0.93379. Alternatively, by creating a variable *WAR* which equals 1 in the years 1942–1945 and zero otherwise and including this in the model, which produces the model shown by the two dashed lines, the R^2 rises to 0.94450.

We can summarize the calculation of R^2 in an **analysis of variance** table, which might appear as shown in Table 3.3.

Example 3.3 Analysis of Variance for the Investment Equation

The analysis of variance table for the investment equation of Section 3.2.2 is given in Table 3.4.

3.5.1 THE ADJUSTED *R*-SQUARED AND A MEASURE OF FIT

There are some problems with the use of R^2 in analyzing **goodness of fit**. The first concerns the number of degrees of freedom used up in estimating the parameters.

TABLE 3.3 Analysis of Variance Table

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*
Regression	$\mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	$K - 1$ (assuming a constant term)	
Residual	$\mathbf{e}'\mathbf{e}$	$n - K$ (including the constant term)	s^2
Total	$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	$n - 1$	s_y^2
R^2	$1 - \mathbf{e}'\mathbf{e}/(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$		

TABLE 3.4 Analysis of Variance for the Investment Equation

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*
Regression	0.75621	4	
Residual	0.20368	10	0.02037
Total	0.95989	14	0.06856
R^2	0.78781		

[See (3-22) and Table 3.3.] R^2 *will never decrease when another variable is added to a regression equation*. Equation (3-24) provides a convenient means for us to establish this result. Once again, we are comparing a regression of $\mathbf{y}$ on $\mathbf{X}$ with sum of squared residuals $\mathbf{e}'\mathbf{e}$ to a regression of $\mathbf{y}$ on $\mathbf{X}$ and an additional variable $\mathbf{z}$, which produces sum of squared residuals $\mathbf{u}'\mathbf{u}$. Recall the vectors of residuals $\mathbf{z}_* = \mathbf{M}\mathbf{z}$ and $\mathbf{y}_* = \mathbf{M}\mathbf{y} = \mathbf{e}$, which implies that $\mathbf{e}'\mathbf{e} = (\mathbf{y}_*'\mathbf{y}_*)$. Let c be the coefficient on $\mathbf{z}$ in the longer regression. Then $c = (\mathbf{z}_*'\mathbf{z}_*)^{-1}(\mathbf{z}_*'\mathbf{y}_*)$, and inserting this in (3-24) produces

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - \frac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{z}_*'\mathbf{z}_*)} = \mathbf{e}'\mathbf{e}(1 - r_{yz}^{*2}), \tag{3-28}$$

where r_{yz}^* is the partial correlation between $\mathbf{y}$ and $\mathbf{z}$, controlling for $\mathbf{X}$. Now divide through both sides of the equality by $\mathbf{y}'\mathbf{M}^0\mathbf{y}$. From (3-26), $\mathbf{u}'\mathbf{u}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ is $(1 - R_{\mathbf{X}z}^2)$ for the regression on $\mathbf{X}$ and $\mathbf{z}$ and $\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ is $(1 - R_{\mathbf{X}}^2)$. Rearranging the result produces the following:

THEOREM 3.6 Change in R^2 When a Variable Is Added to a Regression
Let $R_{\mathbf{X}z}^2$ be the coefficient of determination in the regression of $\mathbf{y}$ on $\mathbf{X}$ and an additional variable $\mathbf{z}$, let $R_{\mathbf{X}}^2$ be the same for the regression of $\mathbf{y}$ on $\mathbf{X}$ alone, and let r_{yz}^* be the partial correlation between $\mathbf{y}$ and $\mathbf{z}$, controlling for $\mathbf{X}$. Then

$$R_{\mathbf{X}z}^2 = R_{\mathbf{X}}^2 + (1 - R_{\mathbf{X}}^2)\, r_{yz}^{*2}. \tag{3-29}$$

Thus, the R^2 in the longer regression cannot be smaller. It is tempting to exploit this result by just adding variables to the model; R^2 will continue to rise to its limit of 1.[4] The **adjusted R^2** (for degrees of freedom), which incorporates a penalty for these results, is computed as follows:

$$\overline{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n - K)}{\mathbf{y}'\mathbf{M}^0\mathbf{y}/(n - 1)}. \tag{3-30}$$

For computational purposes, the connection between R^2 and $\overline{R}^2$ is

$$\overline{R}^2 = 1 - \frac{n - 1}{n - K}(1 - R^2).$$

[4] This result comes at a cost, however. The parameter estimates become progressively less precise as we do so. We will pursue this result in Chapter 4.

The adjusted R^2 may decline when a variable is added to the set of independent variables. Indeed, $\overline{R}^2$ could even be negative. To consider an admittedly extreme case, suppose that $\mathbf{x}$ and $\mathbf{y}$ have a sample correlation of zero. Then the adjusted R^2 will equal $-1/(n-2)$. Whether $\overline{R}^2$ rises or falls when a variable is added to the model depends on whether the contribution of the new variable to the fit of the regression more than offsets the correction for the loss of an additional degree of freedom. The general result (the proof of which is left as an exercise) is as follows.

THEOREM 3.7 Change in $\overline{R}^2$ When a Variable Is Added to a Regression
In a multiple regression, $\overline{R}^2$ will fall (rise) when the variable x is deleted from the regression if the square of the t ratio associated with this variable is greater (less) than 1.

We have shown that R^2 will never fall when a variable is added to the regression. We now consider this result more generally. The change in the residual sum of squares when a set of variables $\mathbf{X}_2$ is added to the regression is

$$\mathbf{e}_1'\mathbf{e}_1 - \mathbf{e}_{1,2}'\mathbf{e}_{1,2} = \mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2,$$

where $\mathbf{e}_1$ is the residuals when $\mathbf{y}$ is regressed on $\mathbf{X}_1$ alone and $\mathbf{e}_{1,2}$ indicates regression on *both* $\mathbf{X}_1$ and $\mathbf{X}_2$. The coefficient vector $\mathbf{b}_2$ is the coefficients on $\mathbf{X}_2$ in the multiple regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$. [See (3-19) and (3-20) for definitions of $\mathbf{b}_2$ and $\mathbf{M}_1$.] Therefore,

$$R_{1,2}^2 = 1 - \frac{\mathbf{e}_1'\mathbf{e}_1 - \mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = R_1^2 + \frac{\mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

which is greater than R_1^2 unless $\mathbf{b}_2$ equals zero. ($\mathbf{M}_1\mathbf{X}_2$ could not be zero unless $\mathbf{X}_2$ is a linear function of $\mathbf{X}_1$, in which case the regression on $\mathbf{X}_1$ and $\mathbf{X}_2$ could not be computed.) This equation can be manipulated a bit further to obtain

$$R_{1,2}^2 = R_1^2 + \frac{\mathbf{y}'\mathbf{M}_1\mathbf{y}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}\frac{\mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}_1\mathbf{y}}.$$

But $\mathbf{y}'\mathbf{M}_1\mathbf{y} = \mathbf{e}_1'\mathbf{e}_1$, so the first term in the product is $1 - R_1^2$. The second is the **multiple correlation** in the regression of $\mathbf{M}_1\mathbf{y}$ on $\mathbf{M}_1\mathbf{X}_2$, or the partial correlation (after the effect of $\mathbf{X}_1$ is removed) in the regression of $\mathbf{y}$ on $\mathbf{X}_2$. Collecting terms, we have

$$R_{1,2}^2 = R_1^2 + (1 - R_1^2)r_{y2.1}^{*2}. \qquad \textbf{(3-31)}$$

[This is the multivariate counterpart to (3-29).]

It is possible to push R^2 as high as desired (up to one) just by adding regressors to the model. This possibility motivates the use of the adjusted R^2 in (3-30), instead of R^2 as a method of choosing among alternative models. Since $\overline{R}^2$ incorporates a penalty for reducing the degrees of freedom while still revealing an improvement in fit, one possibility is to choose the specification that maximizes $\overline{R}^2$. It has been suggested that

the adjusted R^2 does not penalize the loss of degrees of freedom heavily enough.[5] Some alternatives that have been proposed for comparing models (which we index by j) are a modification of the adjusted R squared, that minimizes Amemiya's (1985) **prediction criterion**,

$$PC_j = \frac{\mathbf{e}_j'\mathbf{e}_j}{n - K_j}\left(1 + \frac{K_j}{n}\right) = s_j^2\left(1 + \frac{K_j}{n}\right),$$

$$\overline{R}_j^2 = 1 - \frac{n + K_j}{n - K_j}(1 - R_j^2).$$

Two other fitting criteria are the Akaike and Bayesian information criteria discussed in Section 5.10.1,

$$AIC_j = \ln\left(\frac{\mathbf{e}_j'\mathbf{e}_j}{n}\right) + \frac{2K}{n},$$

$$BIC_j = \ln\left(\frac{\mathbf{e}_j'\mathbf{e}_j}{n}\right) + \frac{K \ln n}{n}.$$

3.5.2 *R*-SQUARED AND THE CONSTANT TERM IN THE MODEL

A second difficulty with R^2 concerns the constant term in the model. The proof that $0 \leq R^2 \leq 1$ requires $\mathbf{X}$ to contain a column of 1s. If not, then (1) $\mathbf{M}^0\mathbf{e} \neq \mathbf{e}$ and (2) $\mathbf{e}'\mathbf{M}^0\mathbf{X} \neq \mathbf{0}$, and the term $2\mathbf{e}'\mathbf{M}^0\mathbf{X}\mathbf{b}$ in $\mathbf{y}'\mathbf{M}^0\mathbf{y} = (\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{M}^0\mathbf{e})'(\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{M}^0\mathbf{e})$ in the expansion preceding (3-25) will not drop out. Consequently, when we compute

$$R^2 = 1 - \frac{\Sigma_{i=1}^n e_i^2}{\Sigma_{i=1}^n (y_i - \bar{y})^2},$$

the result is unpredictable. It will never be higher and can be far lower than the same figure computed for the regression with a constant term included. It can even be negative. Computer packages differ in their computation of R^2. An alternative computation,

$$R^2 = \frac{\Sigma_{i=1}^n (\hat{y}_i - \hat{\bar{y}})^2}{\Sigma_{i=1}^n (y_i - \bar{y})^2},$$

is equally problematic. Again, this calculation will differ from the one obtained with the constant term included; this time, R^2 may be larger than 1. Some computer packages bypass these difficulties by reporting a third "R^2," the squared sample correlation between the actual values of y and the fitted values from the regression. If the regression contains a constant term, then all three computations give the same answer. Even if not, this last one will always produce a value between zero and one. But it is not a proportion of variation explained. On the other hand, for the purpose of comparing models, this squared correlation might well be a useful descriptive device. It is important for users of computer packages to be aware of how the reported R^2 is computed.

[5] See, for example, Amemiya (1985, pp. 50–51).

3.5.3 COMPARING MODELS

The value of R^2 of 0.94450 that we obtained for the consumption function in Example 3.2 seems high in an absolute sense. Is it? Unfortunately, there is no absolute basis for comparison. In fact, in using aggregate time-series data, coefficients of determination this high are routine. In terms of the values one normally encounters in cross sections, an R^2 of 0.5 is relatively high. Coefficients of determination in cross sections of individual data as high as 0.2 are sometimes noteworthy. The point of this discussion is that whether a regression line provides a good fit to a body of data depends on the setting.

Little can be said about the relative quality of fits of regression lines in different contexts or in different data sets even if they are supposedly generated by the same data-generating mechanism. One must be careful, however, even in a single context, to be sure to use the same basis for comparison for competing models. Usually, this concern is about how the dependent variable is computed. For example, a perennial question concerns whether a linear or loglinear model fits the data better. Unfortunately, the question cannot be answered with a direct comparison. An R^2 for the linear regression model is different from an R^2 for the loglinear model. Variation in y is different from variation in $\ln y$. The latter R^2 will typically be larger, but this does not imply that the loglinear model is a better fit in some absolute sense.

It is worth emphasizing that R^2 is a measure of *linear* association between x and y. For example, the third panel of Figure 3.3 shows data that might arise from the model

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i.$$

The relationship between y and x in this model is nonlinear, and a linear regression of y on x would find no fit.

3.6 LINEARLY TRANSFORMED REGRESSION

As a final application of the tools developed in this chapter, we examine a purely algebraic result that is very useful for understanding the computation of linear regression models. In the regression of $\mathbf{y}$ on $\mathbf{X}$, suppose the columns of $\mathbf{X}$ are linearly transformed. Common applications would include changes in the units of measurement, say by changing units of currency, hours to minutes, or distances in miles to kilometers. Example 3.4 suggests a slightly more involved case. This is a useful practical, algebraic result. For example, it simplifies the analysis in the first application suggested, changing the units of measurement. If an independent variable is scaled by a constant, p, the regression coefficient will be scaled by $1/p$. There is no need to recompute the regression.

Example 3.4 Art Appreciation

Theory 1 of the determination of the auction prices of Monet paintings holds that the price is determined by the dimensions (width, *W,* and height, *H*) of the painting,

$$\begin{aligned}\ln \mathit{Price} &= \beta_1(1) + \beta_2 \ln W + \beta_3 \ln H + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.\end{aligned}$$

Theory 2 claims, instead, that art buyers are interested specifically in surface area and aspect ratio,

$$\begin{aligned}\ln \mathit{Price} &= \gamma_1(1) + \gamma_2 \ln (WH) + \gamma_3 \ln (W/H) + \varepsilon \\ &= \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + u.\end{aligned}$$

It is evident that $z_1 = x_1, z_2 = x_2 + x_3$, and $z_3 = x_2 - x_3$. In matrix terms, $\mathbf{Z} = \mathbf{XP}$ where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}, \mathbf{P}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & -1/2 \end{bmatrix}.$$

The effect of a transformation on the linear regression of $\mathbf{y}$ on $\mathbf{X}$ compared to that of $\mathbf{y}$ on $\mathbf{Z}$ is given by Theorem 3.8. Thus, $\beta_1 = \gamma_1, \beta_2 = 1/2(\gamma_2 + \gamma_3), \beta_3 = 1/2(\gamma_2 - \gamma_3)$.

THEOREM 3.8 Transformed Variables
In the linear regression of $\mathbf{y}$ *on* $\mathbf{Z} = \mathbf{XP}$ *where* $\mathbf{P}$ *is a nonsingular matrix that transforms the columns of* $\mathbf{X}$*, the coefficients will equal* $\mathbf{P}^{-1}\mathbf{b}$ *where* $\mathbf{b}$ *is the vector of coefficients in the linear regression of* $\mathbf{y}$ *on* $\mathbf{X}$*, and the* R^2 *will be identical.*
Proof: *The coefficients are*

$$\mathbf{d} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = [(\mathbf{XP})'(\mathbf{XP})]^{-1}(\mathbf{XP})'\mathbf{y} = (\mathbf{P}'\mathbf{X}'\mathbf{XP})^{-1}\mathbf{P}'\mathbf{X}'\mathbf{y}$$
$$= \mathbf{P}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{P}'^{-1}\mathbf{P}'\mathbf{X}'\mathbf{y} = \mathbf{P}^{-1}\mathbf{b}.$$

The vector of residuals is $\mathbf{u} = \mathbf{y} - \mathbf{Z}(\mathbf{P}^{-1}\mathbf{b}) = \mathbf{y} - \mathbf{XPP}^{-1}\mathbf{b} = \mathbf{y} - \mathbf{Xb} = \mathbf{e}$. *Since the residuals are identical, the numerator of* $1 - R^2$ *is the same, and the denominator is unchanged. This establishes the result.*

3.7 SUMMARY AND CONCLUSIONS

This chapter has described the exercise of fitting a line (hyperplane) to a set of points using the method of least squares. We considered the primary problem first, using a data set of n observations on K variables. We then examined several aspects of the solution, including the nature of the projection and residual maker matrices and several useful algebraic results relating to the computation of the residuals and their sum of squares. We also examined the difference between gross or simple regression and correlation and multiple regression by defining partial regression coefficients and partial correlation coefficients. The Frisch–Waugh–Lovell Theorem (3.2) is a fundamentally useful tool in regression analysis that enables us to obtain the expression for a subvector of a vector of regression coefficients. We examined several aspects of the partitioned regression, including how the fit of the regression model changes when variables are added to it or removed from it. Finally, we took a closer look at the conventional measure of how well the fitted regression line predicts or "fits" the data.

Key Terms and Concepts

- Adjusted R^2
- Analysis of variance
- Bivariate regression
- Coefficient of determination
- Degrees of freedom
- Disturbance
- Fitting criterion
- Frisch–Waugh theorem
- Goodness of fit
- Least squares
- Least squares normal equations
- Moment matrix
- Multiple correlation
- Multiple regression
- Netting out
- Normal equations
- Orthogonal regression
- Partial correlation coefficient

- Partial regression coefficient
- Partialing out
- Partitioned regression
- Prediction criterion
- Population quantity
- Population regression
- Projection
- Projection matrix
- Residual
- Residual maker
- Total variation

Exercises

1. *The two-variable regression.* For the regression model $y = \alpha + \beta x + \varepsilon$,
   a. Show that the least squares normal equations imply $\Sigma_i e_i = 0$ and $\Sigma_i x_i e_i = 0$.
   b. Show that the solution for the constant term is $a = \bar{y} - b\bar{x}$.
   c. Show that the solution for b is $b = \left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right] / \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$.
   d. Prove that these two values uniquely minimize the sum of squares by showing that the diagonal elements of the second derivatives matrix of the sum of squares with respect to the parameters are both positive and that the determinant is $4n\left[(\sum_{i=1}^{n}x_i^2) - n\bar{x}^2\right] = 4n\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$, which is positive unless all values of x are the same.
2. *Change in the sum of squares.* Suppose that $\mathbf{b}$ is the least squares coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}$ and that $\mathbf{c}$ is any other $K\times 1$ vector. Prove that the difference in the two sums of squared residuals is

$$(\mathbf{y} - \mathbf{Xc})'(\mathbf{y} - \mathbf{Xc}) - (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = (\mathbf{c} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{c} - \mathbf{b}).$$

   Prove that this difference is positive.
3. *Partial Frisch and Waugh.* In the least squares regression of $\mathbf{y}$ on a constant and $\mathbf{X}$, to compute the regression coefficients on $\mathbf{X}$, we can first transform $\mathbf{y}$ to deviations from the mean $\bar{y}$ and, likewise, transform each column of $\mathbf{X}$ to deviations from the respective column mean; second, regress the transformed $\mathbf{y}$ on the transformed $\mathbf{X}$ without a constant. Do we get the same result if we only transform $\mathbf{y}$? What if we only transform $\mathbf{X}$?
4. *Residual makers.* What is the result of the matrix product $\mathbf{M}_1\mathbf{M}$ where $\mathbf{M}_1$ is defined in (3-19) and $\mathbf{M}$ is defined in (3-14)?
5. *Adding an observation.* A data set consists of n observations contained in $\mathbf{X}_n$ and $\mathbf{y}_n$. The least squares estimator based on these n observations is $\mathbf{b}_n = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{y}_n$. Another observation, $\mathbf{x}_s$ and y_s, becomes available. Prove that the least squares estimator computed using this additional observation is

$$\mathbf{b}_{n,s} = \mathbf{b}_n + \frac{1}{1 + \mathbf{x}_s'(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{x}_s}(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{x}_s\,(y_s - \mathbf{x}_s'\mathbf{b}_n).$$

   Note that the last term is e_s, the residual from the prediction of y_s using the coefficients based on $\mathbf{X}_n$ and $\mathbf{y}_n$. Conclude that the new data change the results of least squares only if the new observation on y cannot be perfectly predicted using the information already in hand.
6. *Deleting an observation.* A common strategy for handling a case in which an observation is missing data for one or more variables is to fill those missing variables with 0s and add a variable to the model that takes the value 1 for that one observation and 0 for all other observations. Show that this strategy is equivalent to discarding the observation as regards the computation of $\mathbf{b}$ but it does have an

effect on R^2. Consider the special case in which $\mathbf{X}$ contains only a constant and one variable. Show that replacing missing values of x with the mean of the complete observations has the same effect as adding the new variable.

7. *Demand system estimation.* Let Y denote total expenditure on consumer durables, nondurables, and services and E_d, E_n, and E_s are the expenditures on the three categories. As defined, $Y = E_d + E_n + E_s$. Now, consider the expenditure system

$$E_d = \alpha_d + \beta_d Y + \gamma_{dd} P_d + \gamma_{dn} P_n + \gamma_{ds} P_s + \varepsilon_d,$$
$$E_n = \alpha_n + \beta_n Y + \gamma_{nd} P_d + \gamma_{nn} P_n + \gamma_{ns} P_s + \varepsilon_n,$$
$$E_s = \alpha_s + \beta_s Y + \gamma_{sd} P_d + \gamma_{sn} P_n + \gamma_{ss} P_s + \varepsilon_s.$$

Prove that if all equations are estimated by ordinary least squares, then the sum of the expenditure coefficients will be 1 and the four other column sums in the preceding model will be zero.

8. *Change in adjusted* R^2. Prove that the adjusted R^2 in (3-30) rises (falls) when variable $\mathbf{x}_k$ is deleted from the regression if the square of the t ratio on $\mathbf{x}_k$ in the multiple regression is less (greater) than 1.
9. *Regression without a constant.* Suppose that you estimate a multiple regression first with, then without, a constant. Whether the R^2 is higher in the second case than the first will depend in part on how it is computed. Using the (relatively) standard method $R^2 = 1 - (\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y})$, which regression will have a higher R^2?
10. Three variables, N, D, and Y, all have zero means and unit variances. A fourth variable is $C = N + D$. In the regression of C on Y, the slope is 0.8. In the regression of C on N, the slope is 0.5. In the regression of D on Y, the slope is 0.4. What is the sum of squared residuals in the regression of C on D? There are 21 observations and all moments are computed using $1/(n - 1)$ as the divisor.
11. Using the matrices of sums of squares and cross products immediately preceding Section 3.2.3, compute the coefficients in the multiple regression of real investment on a constant, GNP, and the interest rate. Compute R^2.
12. In the December 1969 *American Economic Review* (pp. 886–896), Nathaniel Leff reports the following least squares regression results for a cross section study of the effect of age composition on savings in 74 countries in 1964:

$$\ln S/Y = 7.3439 + 0.1596 \ln Y/N + 0.0254 \ln G - 1.3520 \ln D_1 - 0.3990 \ln D_2,$$
$$\ln S/N = 2.7851 + 1.1486 \ln Y/N + 0.0265 \ln G - 1.3438 \ln D_1 - 0.3966 \ln D_2,$$

where S/Y = domestic savings ratio, S/N = per capita savings, Y/N = per capita income, D_1 = percentage of the population under 15, D_2 = percentage of the population over 64, and G = growth rate of per capita income. Are these results correct? Explain.[6]

13. *Is it possible to partition* R^2? The idea of "hierarchical partitioning" is to decompose R^2 into the contributions made by each variable in the multiple regression. That is, if $x_1, \ldots, x_K$ are entered into a regression one at a time, then c_k is the incremental contribution of x_k such that given the order entered, $\Sigma_k c_k = R^2$ and the incremental

[6] See Goldberger (1973) and Leff (1973) for discussion.

contribution of x_k is then c_k/R^2. Of course, based on (3-31), we know that this is not a useful calculation.

a. Argue based on (3-31) why it is not useful.
b. Show using (3-31) that the computation is sensible if (and only if) all variables are orthogonal.
c. For the investment example in Section 3.2.2, compute the incremental contribution of T if it is entered first in the regression. Now compute the incremental contribution of T if it is entered last.

Application

The data listed in Table 3.5 are extracted from Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. (See Appendix Table F3.2.) Their data set is a panel of 2,178 individuals with a total of 17,919 observations. Shown in the table are the first year and the time-invariant variables for the first 15 individuals in the sample. The variables are defined in the article.

Let $\mathbf{X}_1$ equal a constant, education, experience, and ability (the individual's own characteristics). Let $\mathbf{X}_2$ contain the mother's education, the father's education, and the number of siblings (the household characteristics). Let $\mathbf{y}$ be the log of the hourly wage.

a. Compute the least squares regression coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}_1$. Report the coefficients.
b. Compute the least squares regression coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$. Report the coefficients.

TABLE 3.5 Subsample from Koop and Tobias Data

Person	*Education*	*ln Wage*	*Experience*	*Ability*	*Mother's Education*	*Father's Education*	*Siblings*
1	13	1.82	1	1.00	12	12	1
2	15	2.14	4	1.50	12	12	1
3	10	1.56	1	−0.36	12	12	1
4	12	1.85	1	0.26	12	10	4
5	15	2.41	2	0.30	12	12	1
6	15	1.83	2	0.44	12	16	2
7	15	1.78	3	0.91	12	12	1
8	13	2.12	4	0.51	12	15	2
9	13	1.95	2	0.86	12	12	2
10	11	2.19	5	0.26	12	12	2
11	12	2.44	1	1.82	16	17	2
12	13	2.41	4	−1.30	13	12	5
13	12	2.07	3	−0.63	12	12	4
14	12	2.20	6	−0.36	10	12	2
15	12	2.12	3	0.28	10	12	3

c. Regress each of the three variables in $\mathbf{X}_2$ on all the variables in $\mathbf{X}_1$ and compute the residuals from each regression. Arrange these new variables in the 15×3 matrix $\mathbf{X}_2^*$. What are the sample means of these three variables? Explain the finding.

d. Using (3-26), compute the R^2 for the regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$. Repeat the computation for the case in which the constant term is omitted from $\mathbf{X}_1$. What happens to R^2?

e. Compute the adjusted R^2 for the full regression including the constant term. Interpret your result.

f. Referring to the result in part c, regress $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2^*$. How do your results compare to the results of the regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$? The comparison you are making is between the least squares coefficients when $\mathbf{y}$ is regressed on $\mathbf{X}_1$ and $\mathbf{M}_1\mathbf{X}_2$ and when $\mathbf{y}$ is regressed on $\mathbf{X}_1$ and $\mathbf{X}_2$. Derive the result theoretically. (Your numerical results should match the theory, of course.)

4

ESTIMATING THE REGRESSION MODEL BY LEAST SQUARES

4.1 INTRODUCTION

In this chapter, we will examine least squares in detail as an **estimator** of the parameters of the linear regression model (defined in Table 4.1). There are other candidates for estimating $\boldsymbol{\beta}$. For example, we might use the coefficients that minimize the sum of absolute values of the residuals. We begin in Section 4.2 by considering the question "Why should we use least squares?" We will then analyze the estimator in detail. The question of which estimator to choose is based on the **statistical properties** of the candidates, such as unbiasedness, consistency, efficiency, and their **sampling distributions**. Section 4.3 considers **finite-sample properties** such as unbiasedness. The linear model is one of few settings in which the exact finite-sample properties of an estimator are known. In most cases, the only known properties are those that apply to large samples. We can approximate finite-sample behavior by using what we know about large-sample properties. In Section 4.4, we will examine the large-sample or **asymptotic properties** of the least squares estimator of the regression model.[1] Section 4.5 considers **robust inference**. The problem considered here is how to carry out inference when (real) data may not satisfy the assumptions of the basic linear model. Section 4.6 develops a method for inference based on functions of model parameters, rather than the estimates themselves.

Discussions of the properties of an estimator are largely concerned with **point estimation**—that is, in how to use the sample information as effectively as possible to produce the best single estimate of the model parameters. **Interval estimation**, considered in Section 4.7, is concerned with computing estimates that make explicit the uncertainty inherent in using randomly sampled data to estimate population quantities. We will consider some applications of interval estimation of parameters and some functions of parameters in Section 4.7. One of the most familiar applications of interval estimation is using the model to predict the dependent variable and to provide a plausible range of uncertainty for that prediction. Section 4.8 considers prediction and forecasting using the estimated regression model.

The analysis assumes that the data in hand correspond to the assumptions of the model. In Section 4.9, we consider several practical problems that arise in analyzing nonexperimental data. Assumption A2, full rank of $\mathbf{X}$, is taken as a given. As we noted in Section 2.3.2, when this assumption is not met, the model is not estimable, regardless of the sample size. **Multicollinearity**, the near failure of this assumption in real-world

[1]This discussion will use results on asymptotic distributions. It may be helpful to review Appendix D before proceeding to Section 4.4.

TABLE 4.1 Assumptions of the Classical Linear Regression Model

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$. For the sample, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

A2. Full rank: The $n \times K$ sample data matrix, $\mathbf{X}$, has full column rank for every $n \geq$ K.

A3. Exogeneity of the independent variables: $E[\varepsilon_i | x_{j1}, x_{j2}, \ldots, x_{jK}] = 0, i, j = 1, \ldots, n$. There is no correlation between the disturbances and the independent variables. $E[\boldsymbol{\varepsilon} | \mathbf{X}] = \mathbf{0}$.

A4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i, has the same finite variance; $\mathrm{E}[\varepsilon_i^2 | \mathbf{X}] = \sigma^2$. Every disturbance ε_i is uncorrelated with every other disturbance, ε_j, conditioned on $\mathbf{X}$; $E[\varepsilon_i \varepsilon_j | \mathbf{X}] = 0, i \neq j$. $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \sigma^2\mathbf{I}$.

A5. Stochastic or nonstochastic data: $(x_{i1}, x_{i2}, \ldots, x_{iK}), i = 1, \ldots, n$.

A6. Normal distribution: The disturbances, ε_i, are normally distributed. $\boldsymbol{\varepsilon} | \mathbf{X} \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$.

data, is examined in Sections 4.9.1 and 4.9.2. Missing data have the potential to derail the entire analysis. The benign case in which missing values are simply unexplainable random gaps in the data set is considered in Section 4.9.3. The more complicated case of nonrandomly missing data is discussed in Chapter 19. Finally, the problems of badly measured and outlying observations are examined in Section 4.9.4 and 4.9.5.

This chapter describes the properties of estimators. The assumptions in Table 4.1 will provide the framework for the analysis. (The assumptions are discussed in greater detail in Chapter 3.) For the present, it is useful to assume that the data are a cross section of independent, identically distributed random draws from the joint distribution of $(y_i, \mathbf{x}_i)$ with A1–A3 which defines $E[y_i | \mathbf{x}_i]$. Later in the text (and in Section 4.5), we will consider more general cases. The leading exceptions, which all bear some similarity, are stratified samples, cluster samples, **panel data**, and spatially correlated data. In these cases, groups of related individual observations constitute the observational units. The time-series case in Chapters 20 and 21 will deal with data sets in which potentially all observations are correlated. These cases will be treated later when they are developed in more detail. Under random (cross-section) sampling, with little loss of generality, we can easily obtain very general statistical results such as consistency and asymptotic normality. Later, such as in Chapter 11, we will be able to accommodate the more general cases fairly easily.

4.2 MOTIVATING LEAST SQUARES

Ease of computation is one reason that is occasionally offered to motivate least squares. But, with modern software, ease of computation is a minor (usually trivial) virtue. There are several theoretical justifications for this technique. First, least squares is a natural approach to estimation which makes explicit use of the structure of the model as laid out in the assumptions. Second, even if the true model is not a linear regression, the equation fit by least squares is an optimal linear predictor for the explained variable. Thus, it enjoys a sort of robustness that other estimators do not. Finally, under the specific assumptions of the classical model, by one reasonable criterion, least squares will be the most efficient use of the data.

4.2.1 POPULATION ORTHOGONALITY CONDITIONS

Let $\mathbf{x}$ denote the vector of independent variables in the population regression model. Assumption A3 states that $E[\varepsilon | \mathbf{x}] = 0$. Three useful results follow from this. First, by iterated expectations (Theorem B.1), $E_{\mathbf{x}}E[\varepsilon | \mathbf{x}]] = E_{\mathbf{x}}[0] = E[\varepsilon] = 0$; ε has

zero mean, conditionally and unconditionally. Second, by Theorem B.2, $\text{Cov}[\mathbf{x},\varepsilon] = \text{Cov}[\mathbf{x},E[\varepsilon|\mathbf{x}]] = \text{Cov}[\mathbf{x},0] = \mathbf{0}$ so $\mathbf{x}$ and ε are uncorrelated. Finally, combining the earlier results, $E[\mathbf{x}\varepsilon] = \text{Cov}[\mathbf{x},\varepsilon] + E[\varepsilon]E[\mathbf{x}] = \mathbf{0}$. We write the third of these as $E[\mathbf{x}\varepsilon] = E[\mathbf{x}(y - \mathbf{x}'\beta)] = \mathbf{0}$ or

$$E[\mathbf{x}y] = E[\mathbf{x}\mathbf{x}']\boldsymbol{\beta}. \tag{4-1}$$

Now, recall the least squares normal equations (3-5) based on the sample of n observations, $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$. Divide this by n and write it as a summation to obtain

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)\mathbf{b}. \tag{4-2}$$

Equation (4-1) is a population relationship. Equation (4-2) is a sample analog. Assuming the conditions underlying the laws of large numbers presented in Appendix D are met, the means in (4-2) are estimators of their counterparts in (4-1). Thus, by using least squares, we are mimicking in the sample the relationship that holds in the population.

4.2.2 MINIMUM MEAN SQUARED ERROR PREDICTOR

Consider the problem of finding an **optimal linear predictor** for y. Once again, ignore Assumption A6 and, in addition, drop Assumption A1. The conditional mean function, $E[y|\mathbf{x}]$, might be nonlinear. For the criterion, we will use the **mean squared error** rule, so we seek the minimum mean squared error *linear* predictor of y, which we'll denote $\mathbf{x}'\boldsymbol{\gamma}$. (The minimum mean squared error predictor would be the conditional mean function in all cases. Here, we consider only a linear predictor.) The expected squared error of the linear predictor is

$$\text{MSE} = E[y - \mathbf{x}'\boldsymbol{\gamma}]^2.$$

This can be written as

$$\text{MSE} = E\{y - E[y|\mathbf{x}]\}^2 + E\{E[y|\mathbf{x}] - \mathbf{x}'\boldsymbol{\gamma}\}^2.$$

We seek the γ that minimizes this expectation. The first term is not a function of γ, so only the second term needs to be minimized. The necessary condition is

$$\begin{aligned}\frac{\partial E\{E(y|\mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}\}^2}{\partial\boldsymbol{\gamma}} &= E\left\{\frac{\partial\{E(y|\mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}\}^2}{\partial\boldsymbol{\gamma}}\right\}\\ &= -2E\{\mathbf{x}[E(y|\mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]\} = \mathbf{0}.\end{aligned}$$

We arrive at the equivalent condition

$$E[\mathbf{x}E(y|\mathbf{x})] = E[\mathbf{x}\mathbf{x}']\boldsymbol{\gamma}.$$

The left-hand side of this result is $E[\mathbf{x}E(y|\mathbf{x})] = \text{Cov}[\mathbf{x}, E(y|\mathbf{x})] + E[\mathbf{x}]E[E(y|\mathbf{x})] = \text{Cov}[\mathbf{x},y] + E[\mathbf{x}]E[y] = E[\mathbf{x}y]$. (We have used Theorem B.2.) Therefore, the necessary condition for finding the minimum MSE predictor is

$$E[\mathbf{x}y] = E[\mathbf{x}\mathbf{x}']\boldsymbol{\gamma}. \tag{4-3}$$

This is the same as (4-1), which takes us back to the least squares condition. Assuming that these expectations exist, they would be estimated by the sums in (4-2), which means

THEOREM 4.1 Minimum Mean Squared Error Predictor
If the mechanism generating the data $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, *is such that the law of large numbers applies to the estimators in (4-2) of the matrices in (4-1), then the slopes of the minimum expected squared error linear predictor of y are estimated by the least squares coefficient vector.*

that regardless of the form of the conditional mean, least squares is an estimator of the coefficients of the minimum expected squared error linear predictor of $y|\mathbf{x}$.

4.2.3 MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION

Finally, consider the problem of finding a **linear unbiased estimator**. If we seek the one that has smallest variance, we will be led once again to least squares. This proposition will be proved in Section 4.3.5.

4.3 STATISTICAL PROPERTIES OF THE LEAST SQUARES ESTIMATOR

An *estimator* is a strategy, or formula, for using the sample data that are drawn from a population. The *properties* of that estimator are a description of how it can be expected to behave when it is applied to a sample of data. To consider an example, the concept of unbiasedness implies that on average an estimator (strategy) will correctly estimate the parameter in question; it will not be systematically too high or too low. It is not obvious how one could know this if they were only going to analyze a single sample of data from the population. The argument adopted in econometrics is provided by the **sampling properties** of the estimation strategy. A conceptual experiment lies behind the description. One imagines repeated sampling from the population and characterizes the behavior of the sample of samples. The underlying statistical theory of the estimator provides the basis of the description. Example 4.1 illustrates.

The development of the properties of least squares as an estimator can be viewed in three stages. The **finite sample properties** based on Assumptions A1–A6 are precise, and are independent of the sample size. They establish the essential characteristics of the estimator, such as unbiasedness and the broad approach to be used to estimate the sampling variance. Finite sample results have two limiting aspects. First, they can only be obtained for a small number of statistics—essentially only for the basic least squares estimator. Second, the sharpness of the finite sample results is obtained by making assumptions about the data-generating process that we would prefer not to impose, such as normality of the disturbances (Assumption A6 in Table 4.1). Asymptotic properties of the estimator are obtained by deriving reliable results that will provide good approximations in moderate sized or large samples. For example, the large sample property of consistency of the least squares estimator is looser than unbiasedness in one respect, but at the same time, is more informative about how the estimator improves as more sample data are used. Finally, robust inference methods are a refinement of the asymptotic results. The essential asymptotic theory for least squares modifies the finite sample results after relaxing certain assumptions, mainly A5 (data-generating process

for **X**) and A6 (normality). Assumption A4 (homoscedasticity and nonautocorrelation) remains a limitation on the generality of the model assumptions. Real-world data are likely to be heteroscedastic in ways that cannot be precisely quantified. They may also be autocorrelated as a consequence of the sample design, such as the within household correlation of panel data observations. These possibilities may taint the inferences that use standard errors that are based on A4. Robust methods are used to accommodate possible violations of Assumption A4 without redesigning the estimation strategy. That is, we continue to use least squares, but employ inference procedures that will be appropriate whether A4 is reasonable or not.

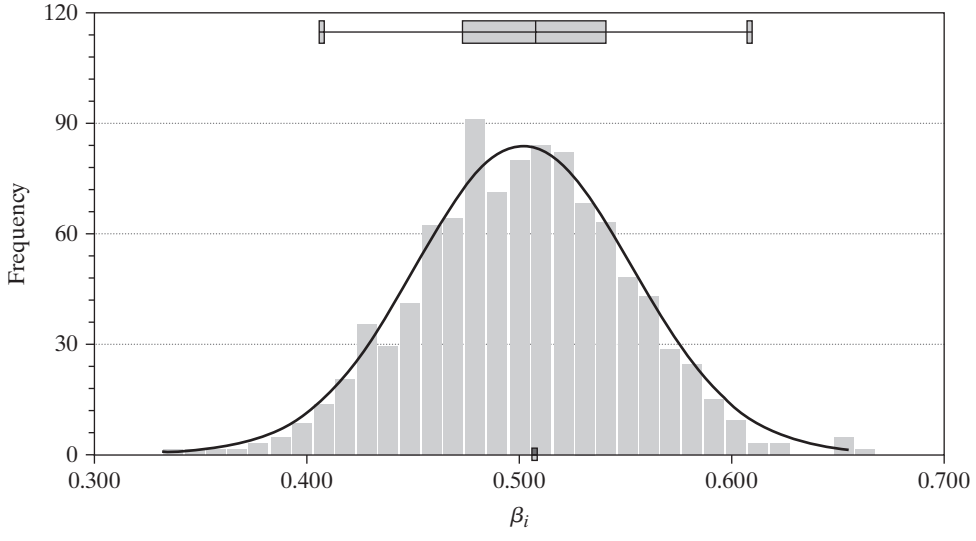
Example 4.1 The Sampling Distribution of a Least Squares Estimator

The following sampling experiment shows the nature of a sampling distribution and the implication of unbiasedness. We drew two samples of 10,000 random draws on variables w_i and x_i from the standard normal population (mean 0, variance 1). We generated a set of ε_i's equal to $0.5w_i$ and then $y_i = 0.5 + 0.5x_i + \varepsilon_i$. We take this to be our population. We then drew 1,000 random samples of 100 observations on (y_i,x_i) from this population (without replacement), and with each one, computed the least squares slope, using at replication r,

$$b_r = \left[\Sigma_{i=1}^{100}(x_{ir} - \bar{x}_r)y_{ir}\right] / \left[\Sigma_{i=1}^{100}(x_{ir} - \bar{x}_r)^2\right].$$

The histogram in Figure 4.1 shows the result of the experiment. Note that the distribution of slopes has mean and median roughly equal to the *true value* of 0.5, and it has a substantial variance, reflecting the fact that the regression slope, like any other statistic computed from the sample, is a random variable. The concept of unbiasedness relates to the central tendency of this distribution of values obtained in repeated sampling from the population. The shape of the histogram also suggests the normal distribution of the estimator that we will show theoretically in Section 4.3.6.

FIGURE 4.1 Histogram for Sampled Least Squares Regression Slopes.



4.3.1 UNBIASED ESTIMATION

The least squares estimator is unbiased in every sample. To show this, write

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-4}$$

Now, take expectations, iterating over $\mathbf{X}$:

$$E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}].$$

By Assumption A3, the expected value of the second term is $(\mathbf{X}'\mathbf{X}/n)^{-1}E[\Sigma_i \mathbf{x}_i\varepsilon_i/n|\mathbf{X}]$. Each term in the sum has expectation zero, which produces the result we need:

$$E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta}. \tag{4-5}$$

Therefore,

$$E[\mathbf{b}] = E_{\mathbf{x}}\{E[\mathbf{b}|\mathbf{x}]\} = E_{\mathbf{x}}[\boldsymbol{\beta}] = \boldsymbol{\beta}. \tag{4-6}$$

The interpretation of this result is that for any sample of observations, $\mathbf{X}$, the least squares estimator has expectation $\boldsymbol{\beta}$. When we average this over the possible values of $\mathbf{X}$, we find the unconditional mean is $\boldsymbol{\beta}$ as well.

4.3.2 OMITTED VARIABLE BIAS

Suppose that a correctly specified regression model would be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma + \boldsymbol{\varepsilon}, \tag{4-7}$$

where the two parts have K and 1 columns, respectively. If we regress $\mathbf{y}$ on $\mathbf{X}$ without including the relevant variable, $\mathbf{z}$, then the estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}\gamma + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-8}$$

(Note, "relevant" means $\gamma \neq 0$.) Taking the expectation, we see that unless $\mathbf{X}'\mathbf{z} = \mathbf{0}$, $\mathbf{b}$ is biased. The well-known result is the **omitted variable formula:**

$$E[\mathbf{b}|\mathbf{X},\mathbf{z}] = \boldsymbol{\beta} + \mathbf{p}_{\mathbf{X}.\mathbf{z}}\gamma, \tag{4-9}$$

where

$$\mathbf{p}_{\mathbf{X}.\mathbf{z}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}. \tag{4-10}$$

The vector $\mathbf{p}_{\mathbf{X}.\mathbf{z}}$ is the column of slopes in the least squares regression of $\mathbf{z}$ on $\mathbf{X}$. Theorem 3.2 (Frisch-Waugh) and Corollary 3.2.1 provide some insight for this result. For each coefficient in (4-9), we have

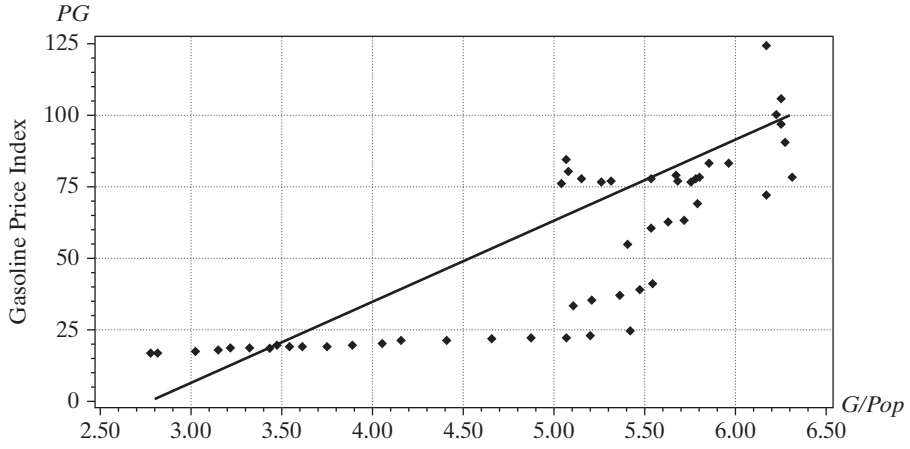
$$E[b_k|\mathbf{X},\mathbf{z}] = \beta_k + \gamma\left(\frac{\text{Cov}(z,x_k|\text{all other } x\text{'s})}{\text{Var}(x_k|\text{all other } x\text{'s})}\right) \tag{4-11}$$

Example 4.2 Omitted Variable in a Demand Equation

If a demand equation is estimated without the relevant income variable, then (4-11) shows how the estimated price elasticity will be biased. The gasoline market data we have examined in Example 2.3 provides a clear example. The base demand model is

$$\textit{Quantity} = \alpha + \beta\textit{Price} + \gamma\textit{Income} + \varepsilon.$$

FIGURE 4.2 Per Capita Gasoline Consumption Versus Price, 1953–2004.



Letting b be the slope coefficient in the regression of *Quantity* on *Price*, we obtain

$$E[b \mid Price, Income] = \beta + \gamma \frac{\text{Cov}[Price, Income]}{\text{Var}[Price]}.$$

In aggregate data, it is unclear whether the missing covariance would be positive or negative. The sign of the bias in b would be the same as this covariance, however, because Var[*Price*] and γ would both be positive for a normal good such as gasoline. Figure 4.2 shows a simple plot of per capita gasoline consumption, *G*/*Pop*, against the price index *PG* (in inverted Marshallian form). The plot disagrees with what one might expect. But a look at the data in Appendix Table F2.2 shows clearly what is at work. In these aggregate data, the simple correlations for (*G*/*Pop*, *Income*/*Pop*) and for (*PG*, *Income*/*Pop*) are 0.938 and 0.934, respectively. To see if the expected relationship between price and consumption shows up, we will have to purge our price and quantity data of the intervening effect of income. To do so, we rely on the Frisch–Waugh result in Theorem 3.2. In the simple regression of the log of per capita gasoline consumption on a constant and the log of the price index, the coefficient is 0.29904, which, as expected, has the *wrong* sign. In the multiple regression of the log of per capita gasoline consumption on a constant, the log of the price index and the log of per capita income, the estimated price elasticity, $\hat{\beta}$, is −0.16949 and the estimated income elasticity, $\hat{\gamma}$, is 0.96595. This agrees with expectations.

In this development, it is straightforward to deduce the directions of bias when there is a single included variable and one omitted variable, as in Example 4.2. It is important to note, however, that if more than one variable is included in $\mathbf{X}$, then the terms in the omitted variable formula, (4-9) and (4-10), involve *multiple* regression coefficients, which have the signs of partial, not simple correlations. For example, in the demand model of the previous example, if the price of a closely related product, say new cars, had been included as well, then the simple correlation between gasoline price and income would be insufficient to determine the direction of the bias in the price elasticity. What would be required is the sign of the correlation between price and income net of the effect of the other price:

$$E[b_{Gasoline\ Price}|\mathbf{X},\mathbf{z}] = \beta_{Gasoline\ Price} + \left(\frac{\text{Cov}(Income, Gasoline\ Price|New\ Cars\ Price)}{\text{Var}(Gasoline\ Price|New\ Cars\ Price)}\right)\gamma. \quad \textbf{(4-12)}$$

This sign might not be obvious, and it would become even less so as more regressors are added to the equation. However, (4-12) does suggest what would be needed for an argument that the least squares estimator remains unbiased, at least for coefficients that correspond to zero partial correlations.

4.3.3 INCLUSION OF IRRELEVANT VARIABLES

We can view the omission of a set of relevant variables as equivalent to imposing an incorrect restriction on (4-7). In particular, omitting $\mathbf{z}$ is equivalent to *incorrectly* estimating (4-7) subject to the restriction $\gamma = 0$. Incorrectly imposing a restriction produces a biased estimator. Suppose, however, that our error is a failure to use some information that is *correct*. If the regression model is correctly given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and we estimate it as if (4-7) were correct [i.e., we include an (or some) extra variable(s)], then the inclusion of the irrelevant variable $\mathbf{z}$ in the regression is equivalent to failing to impose $\gamma = 0$ on (4-7) in estimation. But (4-7) is not incorrect; it simply fails to incorporate $\gamma = 0$. The least squares estimator of $(\boldsymbol{\beta}, \gamma)$ in (4-7) is still unbiased *even given* the restriction:

$$E\left[\begin{pmatrix}\mathbf{b}\\ c\end{pmatrix}|\mathbf{X},\mathbf{z}\right] = \begin{pmatrix}\boldsymbol{\beta}\\ \gamma\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}\\ 0\end{pmatrix}. \quad \textbf{(4-13)}$$

The broad result is that *including irrelevant variables in the estimation equation does not lead to bias in the estimation of the nonzero coefficients*. Then where is the problem? It would seem that to be conservative, one might generally want to overfit the model. As we will show in Section 4.9.1, the covariance matrix in the regression that properly omits the irrelevant $\mathbf{z}$ is generally smaller than the covariance matrix for the estimator obtained in the presence of the superfluous variables. *The cost of overspecifying the model is larger variances (less precision) of the estimators.*

4.3.4 VARIANCE OF THE LEAST SQUARES ESTIMATOR

The least squares coefficient vector is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}, \quad \textbf{(4-14)}$$

where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. By Assumption A4, $\text{E}[\varepsilon\varepsilon'|\mathbf{X}] = \text{Var}[\varepsilon|\mathbf{X}] = \sigma^2\mathbf{I}$. The conditional covariance matrix of the least squares slope estimator is

$$\begin{aligned}\text{Var}[\mathbf{b}|\mathbf{X}] &= E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'|\mathbf{X}]\\ &= E[\mathbf{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{A}'|\mathbf{X}]\\ &= \mathbf{A}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\mathbf{A}'\\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned} \quad \textbf{(4-15)}$$

If we wish to use $\mathbf{b}$ to test hypotheses about $\boldsymbol{\beta}$ or to form confidence intervals, then we will require a sample estimate of this matrix. The population parameter σ^2 remains to be estimated. Because σ^2 is the expected value of ε_i^2 and e_i is an estimate of ε_i, $\hat{\sigma}^2 = (1/\text{n})\sum_{i=1}^{n} e_i^2$ would seem to be the natural estimator. But the least squares

residuals are imperfect estimates of their population counterparts; $e_i = y_i - \mathbf{x}_i'\mathbf{b} = \varepsilon_i - \mathbf{x}_i'(\mathbf{b} - \boldsymbol{\beta})$. The estimator $\hat{\sigma}^2$ is distorted because $\boldsymbol{\beta}$ must be estimated.

The least squares residuals are $\mathbf{e} = \mathbf{My} = \mathbf{M}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbf{M}\boldsymbol{\varepsilon}$, as $\mathbf{MX} = \mathbf{0}$. [See Definition 3.1 and (3-15).] An estimator of σ^2 will be based on the sum of squared residuals:

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}.$$

The expected value of this quadratic form is $E[\mathbf{e}'\mathbf{e}|\mathbf{X}] = E[\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}|\mathbf{X}]$. The scalar $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ is a 1×1 matrix, so it is equal to its trace. By using (A-94), $E[\text{tr}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})|\mathbf{X}] = E[\text{tr}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')|\mathbf{X}]$. Because $\mathbf{M}$ is a function of $\mathbf{X}$, the result is $\text{tr}(\mathbf{M}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]) = \text{tr}(\mathbf{M}\sigma^2\mathbf{I}) = \sigma^2\text{tr}(\mathbf{M})$. The trace of $\mathbf{M}$ is $\text{tr}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{tr}(\mathbf{I}_n) - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{I}_K) = n - K$. Therefore,

$$E[\mathbf{e}'\mathbf{e}|\mathbf{X}] = (n - K)\sigma^2. \tag{4-16}$$

The natural estimator is biased toward zero, but the bias becomes smaller as the sample size increases. An unbiased estimator of σ^2 is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K}. \tag{4-17}$$

Like $\mathbf{b}$, s^2 is unbiased unconditionally, because $E[s^2] = E_{\mathbf{x}}\{E[s^2|\mathbf{X}]\} = E_{\mathbf{x}}[\sigma^2] = \sigma^2$. The **standard error of the regression** is s, the square root of s^2. We can then compute

$$\text{Est. Var}[\mathbf{b}|\mathbf{X}] = s^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{4-18}$$

Henceforth, we shall use the notation Est.Var[Est.Var[·]] to indicate a sample estimate of the **sampling variance** of an estimator. The square root of the kth diagonal element of this matrix, $\{[s^2(\mathbf{X}'\mathbf{X})^{-1}]_{kk}\}^{1/2}$, is the **standard error** of the estimator b_k, which is often denoted simply the standard error of b_k.

4.3.5 THE GAUSS–MARKOV THEOREM

We will now obtain a general result for the class of linear unbiased estimators of $\boldsymbol{\beta}$. Because $\mathbf{b}|\mathbf{X} = \mathbf{Ay}$, where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, is a linear function of $\boldsymbol{\varepsilon}$, by the definition we will use here, it is a **linear estimator** of $\boldsymbol{\beta}$. Because $E[\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$, *regardless of the distribution of* $\boldsymbol{\varepsilon}$, *under our other assumptions,* $\mathbf{b}$ *is a linear, unbiased estimator of* $\boldsymbol{\beta}$.

THEOREM 4.2 Gauss–Markov Theorem
In the linear regression model with given regressor matrix $\mathbf{X}$, (1) *the least squares estimator,* $\mathbf{b}$, *is the minimum variance linear unbiased estimator of* $\boldsymbol{\beta}$ and (2) *for any vector of constants* $\mathbf{w}$, *the minimum variance linear unbiased estimator of* $\mathbf{w}'\boldsymbol{\beta}$ *is* $\mathbf{w}'\mathbf{b}$.

Note that the theorem makes no use of Assumption A6, normality of the distribution of the disturbances. Only A1 to A4 are necessary. Let $\mathbf{b}_0 = \mathbf{Cy}$ be a different linear unbiased estimator of $\boldsymbol{\beta}$, where $\mathbf{C}$ is a $K \times n$ matrix. If $\mathbf{b}_0$ is unbiased, then $E[\mathbf{Cy}|\mathbf{X}] = E[(\mathbf{CX}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon})|\mathbf{X}] = \boldsymbol{\beta}$, which implies that $\mathbf{CX} = \mathbf{I}$ and $\mathbf{b}_0 = \beta + \mathbf{C}\varepsilon$, so $\text{Var}[\mathbf{b}_0|\mathbf{X}] = \sigma^2\mathbf{CC}'$. Now, let $\mathbf{D} = \mathbf{C} - \mathbf{A}$ so $\mathbf{Dy} = \mathbf{b}_0 - \mathbf{b}$. Because $\mathbf{CX} = \mathbf{I}$ and

$\mathbf{AX} = \mathbf{I}, \mathbf{DX} = \mathbf{0}$ and $\mathbf{DA}' = \mathbf{0}$. Then, $\text{Var}[\mathbf{b}_0|\mathbf{X}] = \sigma^2[(\mathbf{D} + \mathbf{A})(\mathbf{D} + \mathbf{A})']$. By multiplying the terms, we find

$$\text{Var}[\mathbf{b}_0|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2\mathbf{DD}' = \text{Var}[\mathbf{b}|\mathbf{X}] + \sigma^2\mathbf{DD}'.$$

The quadratic form in $\mathbf{DD}'$ is $\mathbf{q}'\mathbf{DD}'\mathbf{q} = \mathbf{v}'\mathbf{v} \geq 0$. The conditional covariance matrix of $\mathbf{b}_0$ equals that of $\mathbf{b}$ plus a nonnegative definite matrix. Every quadratic form in $\text{Var}[\mathbf{b}_0|\mathbf{X}]$ is larger than the corresponding quadratic form in $\text{Var}[\mathbf{b}|\mathbf{X}]$, which establishes result (1).

The proof of result (2) of the theorem follows from the previous derivation, because the variance of $\mathbf{w}'\mathbf{b}$ is a quadratic form in $\text{Var}[\mathbf{b}|\mathbf{X}]$, and likewise for any $\mathbf{b}_0$, and implies that each individual slope estimator b_k is the best linear unbiased estimator of β_k. (Let $\mathbf{w}$ be all zeros except for a one in the kth position.) The result applies to every linear combination of the elements of $\boldsymbol{\beta}$. *The implication is that under Assumptions A1–A5,* $\mathbf{b}$ *is the most efficient (linear unbiased) estimator of* $\boldsymbol{\beta}$.

4.3.6 THE NORMALITY ASSUMPTION

To this point, the specification and analysis of the regression model are **semiparametric** (see Section 12.3). We have not used Assumption A6, normality of $\boldsymbol{\varepsilon}$, in any of the results. In (4-4), $\mathbf{b}$ is a linear function of the disturbance vector, $\boldsymbol{\varepsilon}$. If $\boldsymbol{\varepsilon}$ has a multivariate normal distribution, then we may use the results of Section B.10.2 and the mean vector and covariance matrix derived earlier to state that

$$\mathbf{b}|\mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

Each element of $\mathbf{b}|\mathbf{X}$ is normally distributed:

$$b_k|\mathbf{X} \sim N[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}].$$

We found evidence of this result in Figure 4.1 in Example 4.1.

The exact distribution of $\mathbf{b}$ is conditioned on $\mathbf{X}$. The normal distribution of $\mathbf{b}$ in a finite sample is a consequence of the specific assumption of normally distributed disturbances. The normality assumption is useful for constructing test statistics and for forming **confidence intervals**. But we will ultimately find that we will be able to establish the results we need for inference about $\boldsymbol{\beta}$ based only on the sampling behavior of the statistics without tying the analysis to a narrow assumption of normality of $\boldsymbol{\varepsilon}$.

4.4 ASYMPTOTIC PROPERTIES OF THE LEAST SQUARES ESTIMATOR

The finite sample properties of the least squares estimator are helpful in suggesting the range of results that can be obtained from a sample of data. But the list of settings in which exact finite sample results can be obtained is extremely small. The assumption of normality likewise narrows the range of the applications. Estimation and inference can be based on approximate results that will be reliable guides in even moderately sized data sets, and require fewer assumptions.

4.4.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR OF $\boldsymbol{\beta}$

Unbiasedness is a useful starting point for assessing the virtues of an estimator. It assures the analyst that their estimator will not persistently miss its target, either systematically too high or too low. However, as a guide to estimation strategy, unbiasedness has

two shortcomings. First, save for the least squares slope estimator we are discussing in this chapter, it is rare for an econometric estimator to be unbiased. In nearly all cases beyond the multiple linear regression model, the best one can hope for is that the estimator improves in the sense suggested by unbiasedness as more information (data) is brought to bear on the study. As such, we will need a broader set of tools to guide the econometric inquiry. Second, the property of unbiasedness does not, in fact, imply that more information is better than less in terms of estimation of parameters. The sample means of random samples of two, 20 and 20,000 are all unbiased estimators of a population mean—by this criterion all are equally desirable. Logically, one would hope that a larger sample is better than a smaller one in some sense that we are about to define. The property of **consistency** improves on unbiasedness in both of these directions.

To begin, we leave the data-generating mechanism for $\mathbf{X}$ unspecified—$\mathbf{X}$ may be any mixture of constants and random variables generated independently of the process that generates $\boldsymbol{\varepsilon}$. We do make two crucial assumptions. The first is a modification of Assumption A5; **A5a**. $(\mathbf{x}_i, \varepsilon_i)$, $i = 1, \ldots, n$ is a sequence of independent, identically distributed *observations*.

The second concerns the behavior of the data in large samples:

$$\operatorname*{plim}_{n \to \infty} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}, \text{ a positive definite matrix.} \tag{4-19}$$

Note how this extends A2. If every $\mathbf{X}$ has full column rank, then $\mathbf{X}'\mathbf{X}/n$ is a positive definite matrix in a specific sample of $n \geq K$ observations. Assumption (4-19) extends that to all samples with at least K observations. A straightforward way to reach (4-19) based on A5a is to assume

$$E[\mathbf{x}_i \mathbf{x}_i'] = \mathbf{Q},$$

so that by the law of large numbers, $(1/n)\Sigma_i \mathbf{x}_i \mathbf{x}_i'$ converges in probability to its expectation, $\mathbf{Q}$, and via Theorem D.14, $(\mathbf{X}'\mathbf{X}/n)^{-1}$ converges in probability to $\mathbf{Q}^{-1}$.

Time-series settings that involve trends, polynomial time series, and trending variables often pose cases in which the preceding assumptions are too restrictive. A somewhat weaker set of assumptions about $\mathbf{X}$ that is broad enough to include most of these is the **Grenander Conditions** listed in Table 4.2.[2] The conditions ensure that the data matrix is "well behaved" in large samples. The assumptions are very weak and likely to be satisfied by almost any data set encountered in practice.

At many points from here forward, we will make an assumption that the data are well behaved so that an estimator or statistic will converge to a result. Without repeating them in each instance, we will broadly rely on conditions such as those in Table 4.2.

The least squares estimator may be written

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right). \tag{4-20}$$

Then,

$$\operatorname{plim} \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\operatorname{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right).$$

[2]See Grenander (1956), Palma (2016, p. 373) and Judge et al. (1985, p. 162).

TABLE 4.2 Grenander Conditions for Well-Behaved Data

G1. For each column of $\mathbf{X}$, $\mathbf{x}_k$, if $d_{nk}^2 = \mathbf{x}_k'\mathbf{x}_k$, then $\lim_{n\to\infty} d_{nk}^2 = +\infty$. Hence, $\mathbf{x}_k$ does not degenerate to a sequence of zeros. *Sums of squares will continue to grow as the sample size increases.*

G2. $\text{Lim}_{n\to\infty}\, x_{ik}^2/d_{nk}^2 = 0$ for all $i = 1, \ldots, n$. No single observation will ever dominate $\mathbf{x}_k'\mathbf{x}_k$. *As* $n \to \infty$, *individual observations will become less important.*

G3. Let $\mathbf{C}_n$ be the sample correlation matrix of the columns of $\mathbf{X}$, excluding the constant term if there is one. Then $\lim_{n\to\infty}\mathbf{C}_n = \mathbf{C}$, a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that $\mathbf{X}$ has full rank in a finite sample. *This rank condition will not be violated as the sample size increases.*

We require the probability limit of the last term. In Section 4.2.1, we found that $E[\varepsilon|\mathbf{x}] = 0$ implies $E[\mathbf{x}\varepsilon] = \mathbf{0}$. Based on this result, again invoking D.4., we find $\mathbf{X}'\boldsymbol{\varepsilon}/n = (1/n)\Sigma_i \mathbf{x}_i\varepsilon_i$ converges in probability to its expectation of zero, so

$$\text{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = \mathbf{0}. \tag{4-21}$$

It follows that

$$\text{plim}\,\mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}. \tag{4-22}$$

This result establishes that under Assumptions A1–A4 and the additional assumption (4-19), $\mathbf{b}$ is a **consistent estimator** of $\boldsymbol{\beta}$ in the linear regression model. Note how consistency improves on unbiasedness. The asymptotic result does not insist that $\mathbf{b}$ be unbiased. But, by the definition of consistency (see Definition D.6), it will follow that $\lim_{n\to\infty} \text{Prob}[|b_k - \beta_k| > \delta] = 0$ for any positive δ. This means that with increasing sample size, the estimator will be ever closer to the target. This is sometimes (loosely) labeled "asymptotic unbiasedness."

4.4.2 THE ESTIMATOR OF Asy. Var[b]

To complete the derivation of the asymptotic properties of $\mathbf{b}$, we will require an estimator of Asy. $\text{Var}[\mathbf{b}] = (\sigma^2/n)\mathbf{Q}^{-1}$. With (4-19), it is sufficient to restrict attention to s^2, so the purpose here is to assess the consistency of s^2 as an estimator of σ^2. Expanding $s^2 = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}/(n-K)$ produces

$$s^2 = \frac{1}{n-K}[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] = \frac{n}{n-k}\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} - \left(\frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)\right].$$

The leading constant clearly converges to 1. We can apply (4-19), (4-21) (twice), and the product rule for **probability limits** (Theorem D.14) to assert that the second term in the brackets converges to 0. That leaves $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$. This is a narrow case in which the random variables ε_i^2 are independent with the same finite mean σ^2, so not much is required to get the mean to converge almost surely to $\sigma^2 = E[\varepsilon_i^2]$. By the Markov theorem (D.8), what is needed is for $E[(\varepsilon_i^2)^{1+\delta}]$ to be finite, so the minimal assumption thus far is that ε_i have finite moments up to slightly greater than 2. Indeed, if we further assume that every ε_i has the same distribution, then by the Khinchine theorem (D.5) or the corollary to D8, finite moments (of ε_i) up to 2 is sufficient. So, under

fairly weak conditions, the first term in brackets converges in probability to σ^2, which gives our result,

$$\text{plim } s^2 = \sigma^2,$$

and, by the product rule,

$$\text{plim } s^2(\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2\mathbf{Q}^{-1}. \tag{4-23}$$

The appropriate *estimator* of the **asymptotic covariance matrix** of **b** is the familiar one,

$$\text{Est.Asy.Var}[\mathbf{b}] = s^2 (\mathbf{X}'\mathbf{X})^{-1}. \tag{4-24}$$

4.4.3 ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR

By relaxing assumption A6, we will lose the exact normal distribution of the estimator that will enable us to form confidence intervals in Section 4.7. However, normality of the disturbances is not necessary for establishing the distributional results we need to allow statistical inference, including confidence intervals and testing hypotheses. Under generally reasonable assumptions about the process that generates the sample data, large sample distributions will provide a reliable foundation for statistical inference in the regression model (and more generally, as we develop more elaborate estimators later in the book).

To derive the **asymptotic distribution** of the least squares estimator, we shall use the results of Section D.3. We will make use of some basic central limit theorems, so in addition to Assumption A3 (uncorrelatedness), we will assume that observations are *independent*. It follows from (4-20) that

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-25}$$

If the limiting distribution of the random vector in (4-25) exists, then that limiting distribution is the same as that of

$$\left[\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\right]\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-26}$$

Thus, we must establish the limiting distribution of

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} = \sqrt{n}(\overline{\mathbf{w}} - E[\overline{\mathbf{w}}]), \tag{4-27}$$

where $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$ and $E[\mathbf{w}_i] = E[\overline{\mathbf{w}}] = \mathbf{0}$. The mean vector $\overline{\mathbf{w}}$ is the average of n independent identically distributed random vectors with means $\mathbf{0}$ and variances

$$\text{Var}[\mathbf{x}_i\varepsilon_i] = \sigma^2 E[\mathbf{x}_i\mathbf{x}_i'] = \sigma^2\mathbf{Q}. \tag{4-28}$$

The variance of $\sqrt{n}\,\overline{\mathbf{w}} = \dfrac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i$ is

$$\sigma^2\left(\frac{1}{n}\right)[\mathbf{Q} + \mathbf{Q} + \cdots + \mathbf{Q}] = \sigma^2\mathbf{Q}. \tag{4-29}$$

We may apply the Lindeberg–Levy central limit theorem (D.18) to the vector $\sqrt{n}\,\overline{\mathbf{w}}$, as we did in Section D.3 for the univariate case $\sqrt{n}\,\bar{x}$. If $[\mathbf{x}_i\varepsilon_i]$, $i = 1, \ldots, n$ are independent vectors, each distributed with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{Q} < \infty$, and if (4-19) holds, then

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}]. \tag{4-30}$$

It then follows that

$$\mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{Q}^{-1}\mathbf{0}, \mathbf{Q}^{-1}(\sigma^2\mathbf{Q})\mathbf{Q}^{-1}]. \tag{4-31}$$

Combining terms,

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}^{-1}]. \tag{4-32}$$

Using the technique of Section D.3, we then obtain the **asymptotic distribution** of $\mathbf{b}$:

THEOREM 4.3 Asymptotic Distribution of b with IID Observations
If $\{\varepsilon_i\}$ are independently distributed with mean zero and finite variance σ^2 and x_{ik} is such that the Grenander conditions are met, then

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right]. \tag{4-33}$$

The development here has relied on random sampling from $(\mathbf{x}_i, \varepsilon_i)$. If observations are not identically distributed, for example, if $E[\mathbf{x}_i\mathbf{x}_i'] = \boldsymbol{Q}_i$, then under suitable, more general assumptions, an argument could be built around the ***Lindeberg–Feller Central Limit Theorem*** *(D.19A). The essential results would be the same.*

In practice, it is necessary to estimate $(1/n)\mathbf{Q}^{-1}$ with $(\mathbf{X}'\mathbf{X})^{-1}$ and σ^2 with $\mathbf{e}'\mathbf{e}/(n - K)$.

If $\boldsymbol{\varepsilon}$ is normally distributed, then normality of $\mathbf{b}|\mathbf{X}$ holds in *every* sample, so it holds asymptotically as well. The important implication of this derivation is that *if the regressors are well behaved and observations are independent, then the* **asymptotic normality** *of the least squares estimator does not depend on normality of the disturbances; it is a consequence of the Central Limit Theorem.*

4.4.4 ASYMPTOTIC EFFICIENCY

It remains to establish whether the large-sample properties of the least squares estimator are optimal by any measure. The Gauss–Markov theorem establishes finite sample conditions under which least squares is optimal. The requirements that the estimator be linear and unbiased limit the theorem's generality, however. One of the main purposes of the analysis in this chapter is to broaden the class of estimators in the linear regression model to those which might be biased, but which are consistent. Ultimately, we will be interested in nonlinear estimators as well. These cases extend beyond the reach of the Gauss–Markov theorem. To make any progress in this direction, we will require an alternative estimation criterion.

DEFINITION 4.1 Asymptotic Efficiency
An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

We can compare estimators based on their asymptotic variances. The complication in comparing two consistent estimators is that both converge to the true parameter as the sample size increases. Moreover, it usually happens (as in our Example 4.3), that they converge at the same rate—that is, in both cases, the asymptotic variances of the two estimators are of the same order, such as $O(1/n)$. In such a situation, we can sometimes compare the asymptotic variances for the same n to resolve the ranking. The least absolute deviations estimator as an alternative to least squares provides a leading example.

Example 4.3 Least Squares Vs. Least Absolute Deviations—A Monte Carlo Study

Least absolute deviations (LAD) is an alternative to least squares. (The LAD estimator is considered in more detail in Section 7.3.1.) The LAD estimator is obtained as

$$\mathbf{b}_{\text{LAD}} = \text{the minimizer of} \sum_{i=1}^{n} |y_i - \mathbf{x}_i'\mathbf{b}_0|,$$

in contrast to the linear least squares estimator, which is

$$\mathbf{b}_{\text{LS}} = \text{the minimizer of} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b}_0)^2.$$

Suppose the regression model is defined by

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i,$$

where the distribution of ε_i has conditional mean zero, constant variance σ^2, and conditional *median* zero as well—the distribution is symmetric—and $\text{plim}(1/n)\mathbf{X}'\varepsilon = \mathbf{0}$. That is, all the usual regression assumptions, but with the normality assumption replaced by symmetry of the distribution. Then, under our assumptions, $\mathbf{b}_{\text{LS}}$ is a consistent and asymptotically normally distributed estimator with asymptotic covariance matrix given in Theorem 4.3, which we will call $\sigma^2\mathbf{A}$. As Koenker and Bassett (1978, 1982), Huber (1987), Rogers (1993), and Koenker (2005) have discussed, under these assumptions, $\mathbf{b}_{\text{LAD}}$ is also consistent. A good estimator of the asymptotic variance of $\mathbf{b}_{\text{LAD}}$ would be $(1/2)^2[1/f(0)]^2\,\mathbf{A}$ where $f(0)$ is the density of ε at its median, zero. This means that we can compare these two estimators based on their asymptotic variances. The ratio of the asymptotic variance of the kth element of $\mathbf{b}_{\text{LAD}}$ to the corresponding element of $\mathbf{b}_{\text{LS}}$ would be

$$q_k = \text{Var}(b_{k,\text{LAD}})/\text{Var}(b_{k,\text{LS}}) = (1/2)^2(1/\sigma^2)[1/f(0)]^2.$$

If ε did actually have a normal distribution with mean (and median) zero, then $f(\varepsilon) = (2\pi\sigma^2)^{-1/2}\exp(-\varepsilon^2/(2\sigma^2))$ so $f(0) = (2\pi\sigma^2)^{-1/2}$ and for this special case $q_k = \pi/2$. If the disturbances are normally distributed, then LAD will be asymptotically less efficient by a factor of $\pi/2 = 1.573$.

The usefulness of the LAD estimator arises precisely in cases in which we cannot assume normally distributed disturbances. Then it becomes unclear which is the better estimator. It has been found in a long body of research that the advantage of the LAD estimator is most likely to appear in small samples and when the distribution of ε has thicker tails than the

normal—that is, when outlying values of y_i are more likely. As the sample size grows larger, one can expect the LS estimator to regain its superiority. We will explore this aspect of the estimator in a small **Monte Carlo study**.

Examples 2.6 and 3.4 note an intriguing feature of the fine art market. At least in some settings, large paintings sell for more at auction than small ones. Appendix Table F4.1 contains the sale prices, widths, and heights of 430 Monet paintings. These paintings sold at auction for prices ranging from $10,000 to $33 million. A linear regression of the log of the price on a constant term, the log of the surface area, and the aspect ratio produces the results in the top line of Table 4.3. This is the focal point of our analysis. In order to study the different behaviors of the LS and LAD estimators, we will do the following Monte Carlo study: We will draw without replacement 100 samples of R observations from the 430. For each of the 100 samples, we will compute $\mathbf{b}_{LS,r}$ and $\mathbf{b}_{LAD,r}$. We then compute the average of the 100 vectors and the sample variance of the 100 observations.[3] The sampling variability of the 100 sets of results corresponds to the notion of "variation in repeated samples." For this experiment, we will do this for $R = 10$, 50, and 100. The overall sample size is fairly large, so it is reasonable to take the full sample results as at least approximately the "true parameters." The standard errors reported for the full sample LAD estimator are computed using **bootstrapping**. Briefly, the procedure is carried out by drawing B—we used $B = 100$—samples of n (430) observations *with replacement*, from the full sample of n observations. The estimated variance of the LAD estimator is then obtained by computing the mean squared deviation of these B estimates around the mean of the B estimates. This procedure is discussed in detail in Section 15.4.

TABLE 4.3 Estimated Equations for Art Prices

	Constant		*Log Area*		*Aspect Ratio*	
Full Sample	***Mean***	***Standard Error****	***Mean***	***Standard Error***	***Mean***	***Standard Error***
LS	−8.34327	0.67820	1.31638	0.09205	−0.09623	0.15784
LAD	−8.22726	0.82480	1.25904	0.13718	0.04195	0.22762
R = 10						
LS	−10.6218	8.39355	1.65525	1.21002	−0.07655	1.55330
LAD	−12.0635	11.1734	1.81531	1.53662	0.18269	2.11369
R = 50						
LS	−8.57755	1.94898	1.35026	0.27509	−0.08521	0.46600
LAD	−8.33638	2.18488	1.31408	0.36047	−0.06011	0.60910
R = 100						
LS	−8.38235	1.38332	1.32946	0.19682	−0.09378	0.33765
LAD	−8.37291	1.52613	1.31028	0.24277	−0.07908	0.47906

* For the full sample, standard errors for LS use (4-18). Standard errors for LAD are based on 100 bootstrap replications. For the $R = 10$, 50, and 100 experiments, standard errors are the sample standard deviations of the 100 sets of results from the runs of the experiments.

[3]The sample size R is not a negligible fraction of the population size, 430 for each replication. However, this does not call for a finite population correction of the variances in Table 4.3. We are not computing the variance of a sample of R observations drawn from a population of 430 paintings. We are computing the variance of a sample of R statistics, each computed from a different subsample of the full population. There about 10^{20} different samples of 10 observations we can draw. The number of different samples of 50 or 100 is essentially infinite.

If the assumptions underlying the regression model are correct, we should observe the following:

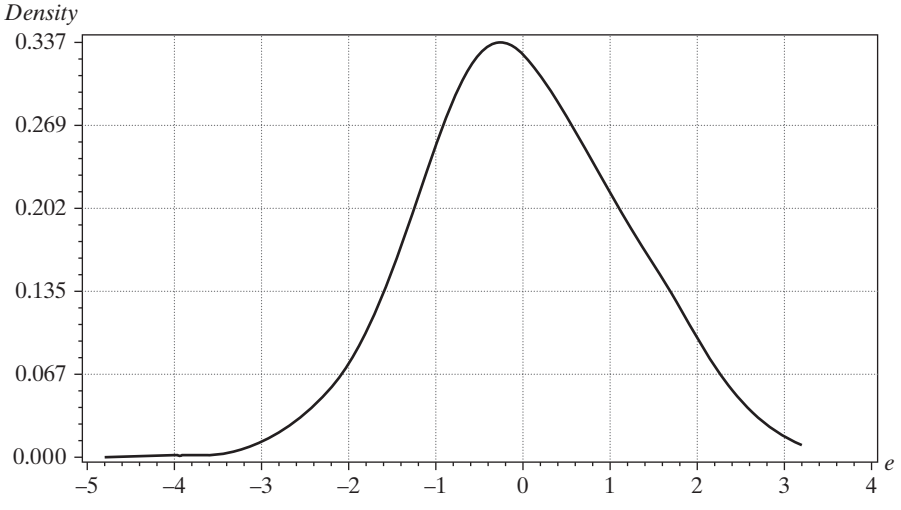
1. Because both estimators are consistent, the averages should resemble the full sample results, the more so as R increases.
2. As R increases, the sampling variance of the estimators should decline.
3. We should observe generally that the standard deviations of the LAD estimates are larger than the corresponding values for the LS estimator.
4. When R is small, the LAD estimator should compare more favorably to the LS estimator, but as R gets larger, the advantage of the LS estimator should become apparent.

A kernel density estimate for the distribution of the least squares residuals appears in Figure 4.3. There is a bit of skewness in the distribution, so a main assumption underlying our experiment may be violated to some degree. Results of the experiments are shown in Table 4.3. The force of the asymptotic results can be seen most clearly in the column for the coefficient on log Area. The decline of the standard deviation as R increases is evidence of the consistency of both estimators. In each pair of results (LS, LAD), we can also see that the estimated standard deviation of the LAD estimator is greater by a factor of about 1.2 to 1.4, which is also to be expected. Based on the normal distribution, we would have expected this ratio to be $\sqrt{\pi/2} = 1.253$.

4.4.5 LINEAR PROJECTIONS

Assumptions A1–A6 define the conditional mean function (CMF) in the joint distribution of $(y_i, \mathbf{x}_i)$, $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, and the conditional distribution of $y|\mathbf{x}$ (normal). Based on Assumptions A1–A6, we find that least squares is a consistent estimator of the slopes of the linear conditional mean under quite general conditions. A useful question for modeling is "What is estimated by linear least squares if the conditional mean function is not linear?" To consider this, we begin with a more

FIGURE 4.3 Kernel Density Estimator for Least Squares Residuals.



general statement of the structural model—this is sometimes labeled the "error form" of the model—in which

$$y = E[y|\mathbf{x}] + \varepsilon = \mu(\mathbf{x}) + \varepsilon.$$

We have shown earlier using the law of iterated expectations that $E[\varepsilon|\mathbf{x}] = E[\varepsilon] = 0$ regardless of whether $\mu(\mathbf{x})$ is linear or not. As a side result to modeling a conditional mean function without the linearity assumption, the modeler might use the results of linear least squares as an easily estimable, interesting feature of the population.

To examine the idea, we retain only the assumption of well-behaved data on $\mathbf{x}$, A2, and A5, and assume, as well, that $(y_i,\mathbf{x}_i)$, $i = 1, \ldots, n$ are a random sample from the joint population of $(y,\mathbf{x})$. We leave the marginal distribution of $\mathbf{x}$ and the conditional distribution of $y|\mathbf{x}$ both unspecified, but assume that all variables in $(y_i, \mathbf{x}_i)$ have finite means, variances, and covariances. The **linear projection** of y on $\mathbf{x}$, $Proj[y|\mathbf{x}]$, is defined by

$$\begin{aligned} & y = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma} + w = Proj[y|\mathbf{x}] + \mathrm{w}, \\ \text{where} \quad & \gamma_0 = E[y] - E[\mathbf{x}]'\boldsymbol{\gamma} \\ \text{and} \quad & \boldsymbol{\gamma} = (\mathrm{Var}[\mathbf{x}])^{-1}\mathrm{Cov}[\mathbf{x},y]. \end{aligned} \tag{4-34}$$

As noted earlier, if $E[w|\mathbf{x}] = 0$, then this would define the CMF, but we have not assumed that. It does follow by inserting the expression for γ_0 in $E[y] = \gamma_0 + E[\mathbf{x}]'\boldsymbol{\gamma} + E[w]$ that $E[w] = 0$, and by expanding $\mathrm{Cov}[\mathbf{x},\mathrm{y}]$ that $\mathrm{Cov}[\mathbf{x},w] = \mathbf{0}$. The linear projection is a characteristic of the joint distribution of $(y_i,\mathbf{x}_i)$. As we have seen, if the CMF in the joint distribution is linear, then the projection will be the conditional mean. But, in the more general case, the linear projection will simply be a feature of the joint distribution. Some aspects of the linear projection function follow from the specification of the model:

1. Because the linear projection is generally not a structural model—that would usually be the CMF—the coefficients in the linear projection will generally not have a *causal* interpretation; indeed, the elements of $\boldsymbol{\gamma}$ will usually not have any direct economic interpretation other than as approximations (of uncertain quality) to the slopes of the CMF.
2. As we saw in Section 4.2.1, linear least squares regression of $\mathbf{y}$ on $\mathbf{X}$ (under the assumed sampling conditions) always estimates the γ_0 and $\boldsymbol{\gamma}$ of the projection regardless of the form of the conditional mean.
3. The CMF is the **minimum mean squared error** predictor of y in the joint distribution of $(y,\mathbf{x})$. We showed in Section 4.2.2 that the linear projection would be the minimum mean squared error *linear* predictor of y. Because both functions are predicting the same thing, it is tempting to infer that the linear projection is a linear approximation to the conditional mean function—and the approximation is exact if the conditional mean is linear. This approximation aspect of the projection function is a common motivation for its use. How effective it is likely to be is obviously dependent on the CMF—a linear function is only going to be able to approximate a nonlinear function locally, and how accurate that is will depend generally on how much curvature there is in the CMF. No generality seems possible; this would be application specific.
4. The interesting features in a structural model are often the partial effects or derivatives of the CMF—in the context of a structural model these are generally the objects of a search for causal effects. A widely observed empirical regularity that

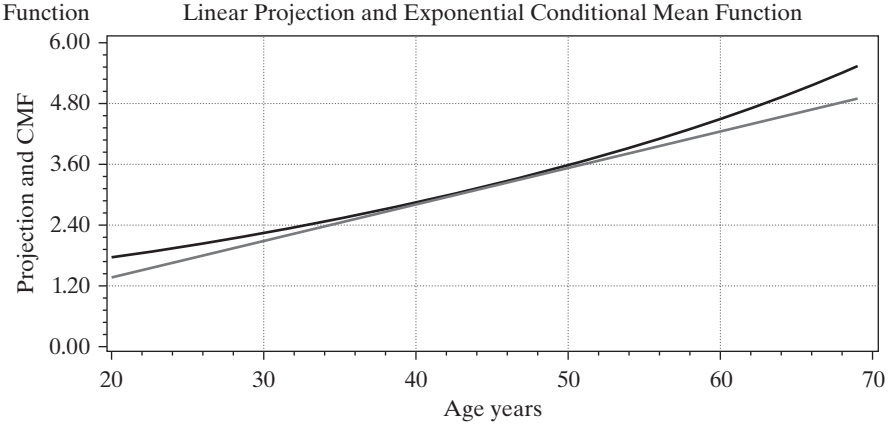
remains to be established with a firm theory is that $\boldsymbol{\gamma}$ in the linear projection often produces a good approximation to the average partial effects based on the CMF.

Example 4.4 Linear Projection: A Sampling Experiment

Table F7.1 describes panel data on 7,293 German households observed from 1 to 7 times for a total of 27,326 household-year observations. Looking ahead to Section 18.4, we examine a model with a nonlinear conditional mean function, a Poisson regression for the number of doctor visits by the household head, conditioned on the age of the survey respondent. We carried out the following experiment: Using all 27,326 observations, we fit a pooled Poisson regression by maximum likelihood in which the conditional mean function is $\lambda_i = \exp(\beta_0 + \beta_1 Age_i)$. The estimated values of (β_0,β_1) are [0.11384,0.02332]. We take this to be the population; $f(y_i|x_i) = \text{Poisson}(\lambda_i)$. We then used the observed data on age to (1) compute this true λ_i for each of the 27,326 observations. (2) We used a random number generator to draw 27,326 observations on y_i from the Poisson population with mean equal to this constructed λ_i. Note that the generated data conform exactly to the model with nonlinear conditional mean. The true value of the average partial effect is computed from $\partial E[y_i|\mathbf{x}_i]/\partial x_i = \beta_1\lambda_i$. We computed this for the full sample. The true APE is $(1/27{,}326)\Sigma_i\beta_1\lambda_i = 0.07384$. For the last step, we randomly sampled 1,000 observations from the population and fit the Poisson regression. The estimated coefficient was $b_1 = 0.02334$. The estimated average partial effect based on the MLEs is 0.07141. Finally, we linearly regressed the random draws y_i on Age_i using the 1,000 values. The estimated slope is 0.07163—nearly identical to the estimated average partial effect from the CMF. The estimated CMF and the linear projection are shown in Figure 4.4. The closest correspondence of the two functions occurs in the center of the data—the average age is 43 years. Several runs of the experiment (samples of 1,000 observations) produced the same result (not surprisingly).

As noted earlier, no firm theoretical result links the CMF to the linear projection save for the case when they are equal. As suggested by Figure 4.4, how good an approximation it provides will depend on the curvature of the CMF, and is an empirical question. For the present example, the fit is excellent in the middle of the data. Likewise, it is not possible to tie the slopes of the CMF at any particular point to the coefficients of the linear projection. The widely observed empirical regularity is that the linear projection can deliver good approximations to average partial effects in models with nonlinear CMFs. This is the underlying motivation

FIGURE 4.4 Nonlinear Conditional Mean Function and Linear Projection.



for recent applications of "linear probability models"—that is, for using linear least squares to fit a familiar nonlinear model. See Angrist and Pischke (2010) and Section 17.3 for further examination.

4.5 ROBUST ESTIMATION AND INFERENCE

Table 4.1 lists six assumptions that define the "Classical Linear Regression Model." A1–A3 define the linear regression framework. A5 suggests a degree of flexibility—the model is broad enough to encompass a wide variety of data generating processes. Assumptions A4 and A6, however, specifically narrow the situations in which the model applies. In particular, A4 seems to preclude the approach developed so far if the disturbances are heteroscedastic or autocorrelated, while A6 limits the stochastic specification to normally distributed disturbances. In fact, we have established all of the finite sample properties save for normality of $\mathbf{b}|\mathbf{X}$, and all of the asymptotic properties without actually using Assumption A6. As such, by these results, the least squares estimator is "robust" to violations of the normality assumption. In particular, it appears to be possible to establish the properties we need for least squares without any specific assumption about the distribution of ε (again, so long as the other assumptions are met).

An estimator of a model is said to be "robust" if it is insensitive to departures from the base assumptions of the model. In practical econometric terms, **robust estimators** retain their desirable properties in spite of violations of some of the assumptions of the model that motivate the estimator. We have seen, for example, that the unbiased least squares estimator is robust to a departure from the normality assumption, A6. In fact, the unbiasedness of least squares is also robust to violations of assumption A4. But, as regards unbiasedness, it is certainly not robust to violations of A3. Also, whether consistency for least squares can be established without A4 remains to be seen. Robustness is usually defined with respect to specific violations of the model assumptions. Estimators are not globally "robust." Robustness is not necessarily a precisely defined feature of an estimator, however. For example, the LAD estimator examined in Example 4.4 is often viewed as a more robust estimator than least squares, at least in small samples, because of its numerical insensitivity to the presence of outlying observations in the data.

For our practical purposes, we will take robustness to be a broad characterization of the asymptotic properties of certain estimators and procedures. We will specifically focus on and distinguish between **robust estimation** and **robust inference**. A robust estimator, in most settings, will be a consistent estimator that remains consistent in spite of violations of assumptions used to motivate it. To continue the example, with some fairly inocuous assumptions about the alternative specification, the least squares estimator will be robust to violations of the homoscedasticity assumption $\text{Var}[\varepsilon_i|\mathbf{x}_i] = \sigma^2$. In most applications, inference procedures are robust when they are based on estimators of asymptotic variances that are appropriate even when assumptions are violated.

Applications of econometrics rely heavily on robust estimation and inference. The development of robust methods has greatly simplified the development of models, as we shall see, by obviating assumptions that would otherwise limit their generality. We will develop a variety of robust estimators and procedures as we proceed.

4.5.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR

In the context of A1–A6, we established consistency of **b** by invoking two results. Assumption A2 is an assumption about existence. Without A2, discussion of consistency is moot, because if $\mathbf{X}'\mathbf{X}/n$ does not have full rank, **b** does not exist. We also relied on A4. The central result is plim $\mathbf{X}'\boldsymbol{\varepsilon}/n = \mathbf{0}$, which we could establish if $E[\mathbf{x}_i\varepsilon_i] = \mathbf{0}$. The remaining element would be a law of large numbers by which the sample mean would converge to its population counterpart. Collecting terms, it turns out that normality, homoscedasticity and nonautocorrelation are not needed for consistency of **b**, so, in turn, consistency of the least squares estimator is robust to violations of these three assumptions. Broadly, random sampling is sufficient.

4.5.2 A HETEROSCEDASTICITY ROBUST COVARIANCE MATRIX FOR LEAST SQUARES

The derivations in Sections 4.4.2 of Asy.Var[**b**] = $(\sigma^2/\mathrm{n})\mathbf{Q}^{-1}$ relied specifically on Assumption A4. In the analysis of a cross section, in which observations are uncorrelated, the issue will be the implications of violations of the homoscedasticity assumption. (We will consider the heteroscedasticity case here. Autocorrelation in time-series data is examined in Section 20.5.2.) For the most general case, suppose $\mathrm{Var}[\varepsilon_i|\mathbf{x}_i] = \sigma_i^2$, with variation assumed to be over $\mathbf{x}_i$. In this case,

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\sum_i \mathbf{x}_i\varepsilon_i.$$

Then,

$$\mathrm{Var}[\mathbf{b}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\left[\sum_i \sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right](\mathbf{X}'\mathbf{X})^{-1}. \tag{4-35}$$

Based on this finite sample result, the asymptotic variance will be

$$\mathrm{Asy.Var}[\mathbf{b}] = \frac{1}{n}\mathbf{Q}^{-1}\left[\mathrm{plim}\frac{1}{n}\sum_i \sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right]\mathbf{Q}^{-1} = \frac{1}{n}\mathbf{Q}^{-1}\mathbf{Q}^*\mathbf{Q}^{-1}. \tag{4-36}$$

Two points to consider are (1) is $s^2(\mathbf{X}'\mathbf{X})^{-1}$ likely to be a valid estimator of Asy.Var[**b**] in this case? and, if not, (2) is there a strategy available that is "robust" to unspecified heteroscedasticity? The first point is pursued in detail in Section 9.3. The answer to the second is yes. What is required is a feasible estimator of $\mathbf{Q}^*$. White's (1980) heteroscedasticity robust estimator of $\mathbf{Q}^*$ is

$$\mathbf{W}_{het} = \frac{1}{n}\sum_i e_i^2\mathbf{x}_i\mathbf{x}_i',$$

where e_i is the least squares residual, $y_i - \mathbf{x}_i'\mathbf{b}$. With $\mathbf{W}_{het}$ in hand, an estimator of Asy.Var[**b**] that is robust to unspecified heteroscedasticity is

$$\mathrm{Est.Asy.Var}[\mathbf{b}] = n(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{W}_{het}(\mathbf{X}'\mathbf{X})^{-1}. \tag{4-37}$$

The implication to this point will be that we can discard the homoscedasticity assumption in A4 and recover appropriate standard errors by using (4-37) to estimate the asymptotic standard errors for the coefficients.

4.5.3 ROBUSTNESS TO CLUSTERING

Settings in which the sample data consist of groups of related observations are increasingly common. Panel data applications such as that in Example 4.5 and in Chapter 11 are an obvious case. Samples of firms grouped by industries, students in schools, home prices in neighborhoods, and so on are other examples. In this application, we suppose that the sample consists of C groups, or "clusters" of observations, labeled $c = 1,...,C$. There are N_c observations in cluster c where N_c is one or more. The n observations in the entire sample therefore comprise $n = \sum_c N_c$ observations. The regression model is

$$y_{i,c} = \mathbf{x}_{i,c}'\boldsymbol{\beta} + \varepsilon_{i,c}.$$

The observations within a cluster are grouped by the correlation across observations within the group. Consider, for example, student test scores where students are grouped by their class. The common teacher will induce a cross-student correlation of $\varepsilon_{i,c}$. An intuitively appealing formulation of such teacher effects would be the "random effects" formulation,

$$y_{i,c} = \mathbf{x}_{i,c}'\boldsymbol{\beta} + w_c + u_{i,c}. \tag{4-38}$$

By this formulation, the common within cluster effect (e.g., the common teacher effect) would induce the same correlation across all members of the group. This random effects specification is considered in detail in Chapter 11. For present purposes, the assumption is stronger than necessary—note that in (4-38), assuming $u_{i,c}$ is independent across observations, $\text{Cov}(\varepsilon_{i,c},\varepsilon_{j,c}) = \sigma_w^2$. At this point, we prefer to allow the correlation to be unspecified, and possibly vary for different pairs of observations.

The least squares estimator is

$$\mathbf{b} = \boldsymbol{\beta} + \left(\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c\right)^{-1}\left[\sum_{c=1}^{C}\left(\sum_{i=1}^{N}\mathbf{x}_{i,c}\varepsilon_{i,c}\right)\right] = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[\sum_{c=1}^{C}\left(\mathbf{X}_c'\boldsymbol{\varepsilon}_c\right)\right],$$

where $\mathbf{X}_c$ is the $N_c \times K$ matrix of exogenous variables for cluster c and $\boldsymbol{\varepsilon}_c$ is the N_c disturbances for the group. Assuming that the clusters are independent,

$$\text{Var}[\mathbf{b}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[\sum_{c=1}^{C}\mathbf{X}_c\boldsymbol{\Omega}_c\mathbf{X}_c'\right]\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{4-39}$$

Like σ_i^2 before, $\boldsymbol{\Omega}_c$ is not meant to suggest a particular set of population parameters. Rather, $\boldsymbol{\Omega}_c$ represents the possibly unstructured correlations allowed among the N_c disturbances in cluster c. The construction is essentially the same as the White estimator, though $\boldsymbol{\Omega}_c$ is the matrix of variances and covariances for the full vector $\boldsymbol{\varepsilon}_c$. (It would be identical to the White estimator if each cluster contained one observation.) Taking the same approach as before, we obtain the asymptotic variance

$$\text{Asy.Var}[\mathbf{b}] = \frac{1}{C}\mathbf{Q}^{-1}\left[\text{plim}\frac{1}{C}\sum_{c=1}^{C}\mathbf{X}_c\boldsymbol{\Omega}_c\mathbf{X}_c'\right]\mathbf{Q}^{-1}.^{4} \tag{4-40}$$

[4]Since the observations in a cluster are not assumed to be independent, the number of observations in the sample is no longer n. Logically, the sample would now consist of C multivariate observations. In order to employ the asymptotic theory used to obtain Asy.Var[**b**], we are implicitly assuming that C is large while N_c is relatively small, and asymptotic results would relate to increasing C, not n. In practical applications, the number of clusters is often rather small, and the group sizes relatively large. We will revisit these complications in Section 11.3.3.

A feasible estimator of the bracketed matrix based on the least squares residuals is

$$\mathbf{W}_{cluster} = \frac{1}{C}\sum_{c=1}^{C}\left(\mathbf{X}_c'\mathbf{e}_c\right)\left(\mathbf{e}_c'\mathbf{X}_c\right) = \frac{1}{C}\sum_{c=1}^{C}\left(\sum_{i=1}^{N_c}\mathbf{x}_{ic}e_{ic}\right)\left(\sum_{i=1}^{N_c}\mathbf{x}_{ic}e_{ic}\right)'. \quad \textbf{(4-41)}$$

Then,

$$\text{Est.Asy.Var}[\mathbf{b}] = C(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{W}_{cluster}(\mathbf{X}'\mathbf{X})^{-1}. \quad \textbf{(4-42)}$$

[A refinement intended to accommodate a possible downward bias induced by a small number of clusters is to multiply $\mathbf{W}_{cluster}$ by $C/(C-1)$ (*SAS*) or by $[C/(C-1)] \times [(n-1)/(n-K)]$ (*Stata, NLOGIT*).]

Example 4.5 Robust Inference About the Art Market

The Monet paintings examined in Example 4.3 were sold at auction over 1989–2006. Our model thus far is

$$\ln Price_{it} = \beta_1 + \beta_2 \ln Area_{it} + \beta_3 AspectRatio_{it} + \varepsilon_{it}$$

The subscript "*it*" uniquely identifies the painting and when it was sold. Prices in open outcry auctions reflect (at least) three elements, the common (public), observable features of the item, the public unobserved (by the econometrician) elements of the asset, and the private unobservable preferences of the winning bidder. For example, it will turn out (in a later example) that whether the painting is signed or not has a large and significant influence on the price. For now, we assume (for sake of the example), that we do not observe whether the painting is signed or not, though, of course, the winning bidders do observe this. It does seem reasonable to suggest that the presence of a signature is uncorrelated with the two attributes we do observe, area and aspect ratio. We respecify the regression as

$$\ln Price_{it} = \beta_1 + \beta_2 \ln Area_{it} + \beta_3 AspectRatio_{it} + w_{it} + u_{it},$$

where w_{it} represents the intrinsic, unobserved features of the painting and u_{it} represents the unobserved preferences of the buyer. In fact, the sample of 430 sales involves 376 unique paintings. Several of the sales are repeat sales of the same painting. The numbers of sales per painting were one, 333; two, 34; three, 7; and four, 2. Figure 4.5 shows the configuration of the sample. For those paintings that sold more than once, the terms w_{it} do relate to the same *i,* and, moreover, would naturally be correlated. [They needn't be identical as in (4-38), however. The valuation of attributes of paintings or other assets sold at auction could vary over time.]

FIGURE 4.5 Repeat Sales of Monet Paintings.

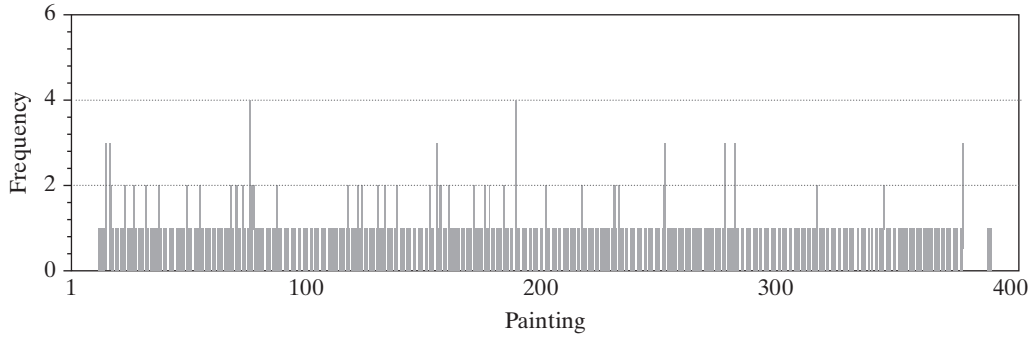


TABLE 4.4 Robust Standard Errors

Estimated		*LS Standard*	*Heteroscedasticity*	*Cluster Robust*
Variable	*Coefficient*	*Error*	*Robust Std.Error*	*Std.Error*
Constant	−8.34237	0.67820	0.73342	0.75873
ln *Area*	1.31638	0.09205	0.10598	0.10932
Aspect Ratio	−0.09623	0.15784	0.16706	0.17776

The least squares estimates and three sets of estimated standard errors are shown in Table 4.4. Even with only a small amount of clustering, the correction produces a tangible adjustment of the standard errors. Perhaps surprisingly, accommodating possible heteroscedasticity produces a more pronounced effect than the cluster correction. Note, finally, in contrast to common expectations, the robust covariance matrix does not always have larger standard errors. The standard errors do increase slightly in this example, however.

4.5.4 BOOTSTRAPPED STANDARD ERRORS WITH CLUSTERED DATA

The sampling framework that underlies the treatment of clustering in the preceding section assumes that the sample consists of a reasonably large number of clusters, drawn randomly from a very large population of clusters. Within each cluster reside a number of observations generated by the linear regression model. Thus,

$$y_{i,c} = \mathbf{x}_{i,c}'\boldsymbol{\beta} + \varepsilon_{i,c},$$

where within each cluster, $\mathrm{E}[\varepsilon_{\mathrm{i,c}},\varepsilon_{\mathrm{j,c}}]$ may be nonzero—observations may be freely correlated. Clusters are assumed to be independent. Each cluster consists of N_c observations, $(\mathbf{y}_c,\mathbf{X}_c,\boldsymbol{\varepsilon}_c)$ and the cluster is the unit of observation. For example, we might be examining student test scores in a state where students are grouped by classroom, and there are potentially thousands of classrooms in the state. The sample consists of a sample of classrooms. (Higher levels of grouping, such as classrooms in a school, and schools in districts, would require some extensions. We will consider this possibility later in Chapter 11.) The essential feature of the data is the likely correlation across observations in the group. Another natural candidate for this type of process would be a panel data set such as the labor market data examined in Example 4.6, where a sample of 595 individuals is each observed in 7 consecutive years. The common feature is the large number of relatively small or moderately sized clusters in the sample.

The method of estimating a robust asymptotic covariance matrix for the least squares estimator that was introduced in the preceding section involves a method of using the data and the least squares residuals to build a covariance matrix. **Bootstrapping** is another method that is likely to be effective under these assumed sampling conditions. (We emphasize, if the number of clusters is quite small and/or group sizes are very large relative to the number of clusters, then bootstrapping, like the previous method, is likely not to be effective.[5] Bootstrapping was introduced in Example 4.3 where we used the

[5]See, for example, Wooldridge (2010, Chapter 20).

method to estimate an asymptotic covariance matrix for the LAD estimator. The basic steps in the methodology are:

1. For R repetitions, draw a random sample of N_c observations from the full sample of N_c observations *with replacement*. Estimate the parameters of the regression model with each of the R constructed samples.
2. The estimator of the asymptotic covariance matrix is the sample variance of the R sets of estimated coefficients.

Keeping in mind that in the current case, the cluster is the unit of observation, we use a **block bootstrap**. In the example below, the block is the 7 observations for individual i, so each observation in the bootstrap replication is a block of 7 observations. Example 4.6 below illustrates the use of block bootstrap.

Example 4.6 Clustering and Block Bootstrapping

Cornwell and Rupert (1988) examined the returns to schooling in a panel data set of 595 heads of households observed in seven years, 1976–1982. The sample data (Appendix Table F8.1) are drawn from years 1976 to 1982 from the *Non-Survey of Economic Opportunity* from the Panel Study of Income Dynamics. A slightly modified version of their regression model is

$$\ln Wage_{it} = \beta_1 + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \beta_{11} Ed_i + \beta_{12} Fem_i + \beta_{13} Blk_i + \varepsilon_{it}.$$

The variables in the model are as follows:

Exp	= years of full time work experience,
Wks	= weeks worked,
Occ	= 1 if blue-collar occupation, 0 if not,
Ind	= 1 if the individual works in a manufacturing industry, 0 if not,
South	= 1 if the individual resides in the south, 0 if not,
SMSA	= 1 if the individual resides in an SMSA, 0 if not,
MS	= 1 if the individual is married, 0 if not,
Union	= 1 if the individual wage is set by a union contract, 0 if not,
Ed	= years of education as of 1976,
Fem	= 1 if the individual is female, 0 if not,
Blk	= 1 if the individual is black.

See Appendix Table F8.1 for the data source.

Table 4.5 presents the least squares and three sets of asymptotic standard errors. The first is the conventional results based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$. Compared to the other estimates, it appears that the uncorrected standard errors substantially understate the variability of the least squares estimator. The clustered standard errors are computed using (4-42). The values are 50%–100% larger. The bootstrapped standard errors are quite similar to the robust estimates, as would be expected.

4.6 ASYMPTOTIC DISTRIBUTION OF A FUNCTION OF b: THE DELTA METHOD

We can extend Theorem D.22 to functions of the least squares estimator. Let $\mathbf{f}(\mathbf{b})$ be a set of J continuous, linear, or nonlinear and continuously differentiable functions of the least squares estimator, and let

$$\mathbf{C}(\mathbf{b}) = \frac{\partial \mathbf{f}(\mathbf{b})}{\partial \mathbf{b}'},$$

TABLE 4.5 Clustered, Robust, and Bootstrapped Standard Errors

Variable	*Least Squares Estimate*	*Standard Error*	*Clustered Std.Error*	*Bootstrapped Std.Error*	*White Hetero. Robust Std.Error*
Constant	5.25112	0.07129	0.12355	0.11171	0.07435
Exp	0.00401	0.00216	0.00408	0.00434	0.00216
ExpSq	−0.00067	0.00005	0.00009	0.00010	0.00005
Wks	0.00422	0.00108	0.00154	0.00164	0.00114
Occ	−0.14001	0.01466	0.02724	0.02555	0.01494
Ind	0.04679	0.01179	0.02366	0.02153	0.01199
South	−0.05564	0.01253	0.02616	0.02414	0.01274
SMSA	0.15167	0.01207	0.02410	0.02323	0.01208
MS	0.04845	0.02057	0.04094	0.03749	0.02049
Union	0.09263	0.01280	0.02367	0.02553	0.01233
Ed	0.05670	0.00261	0.00556	0.00483	0.00273
Fem	−0.36779	0.02510	0.04557	0.04460	0.02310
Blk	−0.16694	0.02204	0.04433	0.05221	0.02075

where $\mathbf{C}$ is the $J \times K$ matrix whose jth row is the vector of derivatives of the jth function with respect to $\mathbf{b}'$. By the Slutsky theorem (D.12),

$$\text{plim } \mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta})$$

and

$$\text{plim } \mathbf{C}(\mathbf{b}) = \frac{\partial \mathbf{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \boldsymbol{\Gamma}.$$

Using a linear Taylor series approach, we expand this set of functions in the approximation

$$\mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta}) + \boldsymbol{\Gamma} \times (\mathbf{b} - \boldsymbol{\beta}) + \text{higher-order terms.}$$

The higher-order terms become negligible in large samples if plim $\mathbf{b} = \boldsymbol{\beta}$. Then, the asymptotic distribution of the function on the left-hand side is the same as that on the right. The mean of the asymptotic distribution is plim $\mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta})$, and the asymptotic covariance matrix is $\{\boldsymbol{\Gamma}[\text{Asy.Var}(\mathbf{b} - \boldsymbol{\beta})]\boldsymbol{\Gamma}'\}$, which gives us the following theorem:

THEOREM 4.4 Asymptotic Distribution of a Function of b
If $\mathbf{f}(\mathbf{b})$ *is a set of continuous and continuously differentiable functions of* $\mathbf{b}$ *such that* $\mathbf{f}(\text{plim } \mathbf{b})$ exists and $\boldsymbol{\Gamma} = \partial \mathbf{f}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$ *and if Theorem 4.4 holds, then*

$$\mathbf{f}(\mathbf{b}) \xrightarrow{a} N\left[\mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\Gamma}\left(\text{Asy.Var}[\mathbf{b}]\right)\boldsymbol{\Gamma}'\right]. \tag{4-43}$$

In practice, the estimator of the asymptotic covariance matrix would be

$$\text{Est.Asy.Var}[\mathbf{f}(\mathbf{b})] = \mathbf{C}\{\text{Est. Asy.Var}[\mathbf{b}]\}\mathbf{C}'.$$

If any of the functions are nonlinear, then the property of unbiasedness that holds for $\mathbf{b}$ may not carry over to $\mathbf{f}(\mathbf{b})$. Nonetheless, $\mathbf{f}(\mathbf{b})$ is a consistent estimator of $\mathbf{f}(\boldsymbol{\beta})$, and the asymptotic covariance matrix is readily available.

Example 4.7 Nonlinear Functions of Parameters: The Delta Method

A dynamic version of the demand for gasoline model in Example 2.3 would be used to separate the short- and long-term impacts of changes in income and prices. The model would be

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(Income/Pop)_t + \beta_4 \ln P_{nc,t} + \beta_5 \ln P_{uc,t} + \gamma \ln(G/Pop)_{t-1} + \varepsilon_t,$$

where P_{nc} and P_{uc} are price indexes for new and used cars. In this model, the short-run price and income elasticities are β_2 and β_3. The long-run elasticities are $\phi_2 = \beta_2/(1 - \gamma)$ and $\phi_3 = \beta_3/(1 - \gamma)$, respectively. To estimate the long-run elasticities, we will estimate the parameters by least squares and then compute these two nonlinear functions of the estimates. We can use the delta method to estimate the standard errors.

Least squares estimates of the model parameters with standard errors and t ratios are given in Table 4.6. (Because these are aggregate time-series data, we have not computed a robust covariance matrix.) The estimated short-run elasticities are the estimates given in the table. The two estimated long-run elasticities are $f_2 = b_2/(1 - c) = -0.069532/(1 - 0.830971) = -0.411358$ and $f_3 = 0.164047/(1 - 0.830971) = 0.970522$. To compute the estimates of the standard errors, we need the estimated partial derivatives of these functions with respect to the six parameters in the model:

$$\hat{\boldsymbol{\Gamma}}_2' = \partial\phi_2(\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}' = [0,1/(1 - \hat{\gamma}), 0, 0, 0, \hat{\boldsymbol{\beta}}_2/(1 - \hat{\gamma})^2] = [0, 5.91613, 0, 0, 0, -2.43365],$$

$$\hat{\boldsymbol{\Gamma}}_3' = \partial\phi_3(\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}' = [0, 0,1/(1 - \hat{\gamma}), 0, 0, \hat{\boldsymbol{\beta}}_3/(1 - \hat{\gamma})^2] = [0, 0, 5.91613, 0, 0, 5.74174].$$

Using (4-43), we can now compute the estimates of the asymptotic variances for the two estimated long-run elasticities by computing $\mathbf{g}_2'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{g}_2$ and $\mathbf{g}_3'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{g}_3$. The results are 0.023194 and 0.0263692, respectively. The two asymptotic standard errors are the square roots, 0.152296 and 0.162386.

TABLE 4.6 Regression Results for a Demand Equation

Sum of squared residuals:	0.0127352
Standard error of the regression:	0.0168227
R^2 based on 51 observations	0.9951081

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	−3.123195	0.99583	−3.136
ln P_G	−0.069532	0.01473	−4.720
ln *Income / Pop*	0.164047	0.05503	2.981
ln P_{nc}	−0.178395	0.05517	−3.233
ln P_{uc}	0.127009	0.03577	3.551
last period ln *G / Pop*	0.830971	0.04576	18.158

Estimated Covariance Matrix for **b** (*e*−*n* = *times* 10^{-n})

Constant	*ln* P_G	*ln (Income/Pop)*	*ln* P_{nc}	*ln* P_{uc}	*ln* $(G/Pop)_{t-1}$
0.99168					
−0.0012088	0.00021705				
−0.052602	1.62165e–5	0.0030279			
0.0051016	−0.00021705	−0.00024708	0.0030440		
0.0091672	−4.0551e−5	−0.00060624	−0.0016782	0.0012795	
0.043915	−0.0001109	−0.0021881	0.00068116	8.57001e–5	0.0020943

4.7 INTERVAL ESTIMATION

The objective of interval estimation is to present the best estimate of a parameter with an explicit expression of the uncertainty attached to that estimate. A general approach for estimation of a parameter θ would be

$$\hat{\theta} \pm \text{sampling variability.} \tag{4-44}$$

(We are assuming that the interval of interest would be symmetric around $\hat{\theta}$.) Following the logic that the range of the sampling variability should convey the degree of (un) certainty, we consider the logical extremes. We can be absolutely (100%) certain that the true value of the parameter we are estimating lies in the range $\hat{\theta} \pm \infty$. Of course, this is not particularly informative. At the other extreme, we should place no certainty (0.0%) on the range $\hat{\theta} \pm 0$. The probability that our estimate precisely hits the true parameter value should be considered zero. The point is to choose a value of α—0.05 or 0.01 is conventional—such that we can attach the desired confidence (probability), $100(1 - \alpha)\%$, to the interval in (4-44). We consider how to find that range and then apply the procedure to three familiar problems, calculating an interval for one of the regression parameters, estimating a function of the parameters, and predicting the value of the dependent variable in the regression using a specific setting of the independent variables. For this latter purpose, we will rely on the asymptotic normality of the estimator.

4.7.1 FORMING A CONFIDENCE INTERVAL FOR A COEFFICIENT

If the disturbances are normally distributed, then for any particular element of **b**,

$$b_k \sim N[\beta_k, \sigma^2 S^{kk}],$$

where S^{kk} denotes the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. By standardizing the variable, we find

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \tag{4-45}$$

has a standard normal distribution. Note that z_k, which is a function of b_k, β_k, σ^2, and S^{kk}, nonetheless has a distribution that involves none of the model parameters or the data. Using

the conventional 95% confidence level, we know that Prob$[-1.96 \leq z_k \leq 1.96] = 0.95$. By a simple manipulation, we find that

$$\text{Prob}[b_k - 1.96\sqrt{\sigma^2 S^{kk}} \leq \beta_k \leq b_k + 1.96\sqrt{\sigma^2 S^{kk}}] = 0.95. \quad \textbf{(4-46)}$$

This states the probability that the random interval, $[b_k \pm$ the sampling variability], contains β_k, not the probability that β_k lies in the specified interval. If we wish to use some other level of confidence, not 95%, then the 1.96 in (4-46) is replaced by the appropriate $z_{(1-\alpha/2)}$. (We are using the notation $z_{(1-\alpha/2)}$ to denote the value of z such that for the standard normal variable z, Prob$[z \leq z_{(1-\alpha/2)}] = 1 - \alpha/2$. Thus, $z_{0.975} = 1.96$, which corresponds to $\alpha = 0.05$.)

We would have the desired confidence interval in (4-46), save for the complication that σ^2 is not known, so the interval is not operational. Using s^2 from the regression instead, the ratio

$$t_k = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \quad \textbf{(4-47)}$$

has a t distribution with $(n - K)$ degrees of freedom.[6] We can use t_k to test hypotheses or form confidence intervals about the individual elements of β. A confidence interval for β_k would be formed using

$$\text{Prob}\left[b_k - t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}} \leq \beta_k \leq b_k + t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}}\right] = 1 - \alpha, \quad \textbf{(4-48)}$$

where $t_{(1-\alpha/2),[n-K]}$ is the appropriate critical value from the t distribution. The distribution of the pivotal statistic depends on the sample size through $(n - K)$, but, once again, not on the parameters or the data.

If the disturbances are not normally distributed, then the theory for the t distribution in (4-48) does not apply. But, the large sample results in Section 4.4 provide an alternative approach. Based on the development that we used to obtain Theorem 4.3 and (4-33), the limiting distribution of the statistic

$$z_k = \frac{\sqrt{n}(b_k - \beta_k)}{\sqrt{\sigma^2 Q^{kk}}}$$

is standard normal, where $\mathbf{Q} = [\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$ and Q^{kk} is the kth diagonal element of $\mathbf{Q}$. Based on the Slutsky theorem (D.16), we may replace σ^2 with a consistent estimator, s^2, and obtain a statistic with the same limiting distribution. We estimate $\mathbf{Q}$ with $(\mathbf{X}'\mathbf{X}/n)^{-1}$. This gives us precisely (4-47), which states that under the assumptions in Section 4.4, the "t" statistic in (4-47) converges to standard normal even if the disturbances are not normally distributed. The implication would be that to employ the asymptotic distribution of $\mathbf{b}$, we should use (4-48) to compute the confidence interval but use the critical values from the standard normal table (e.g., 1.96) rather than from the t distribution. In practical terms, if the degrees of freedom in (4-48) are moderately large, say greater than 100, then the t distribution will be indistinguishable from the standard normal, and this large sample result would apply in any event. For smaller sample sizes, however, in the interest of conservatism, one might be advised to use the critical

[6] See (B-36) in Section B.4.2. It is the ratio of a standard normal variable to the square root of a chi-squared variable divided by its degrees of freedom.

values from the *t* table rather the standard normal, even in the absence of the normality assumption. In the application in Example 4.8, based on a sample of 52 observations, we form a confidence interval for the income elasticity of demand using the critical value of 2.012 from the *t* table with 47 degrees of freedom. If we chose to base the interval on the asymptotic normal distribution, rather than the standard normal, we would use the 95% critical value of 1.96. One might think this is a bit optimistic, however, and retain the value 2.012, again, in the interest of conservatism.

The preceding analysis starts from Assumption A6, normally distributed disturbance, then shows how the procedure is adjusted to rely on the asymptotic properties of the estimator rather than the narrow possibly unwarranted assumption of normally distributed disturbances. It continues to rely on the homoscedasticity assumption in A4. (For the present, we are assuming away possible autocorrelation.) Section 4.5 showed how the estimator of the asymptotic covariance matrix can be refined to allow for unspecified heteroscedasticity or cluster effects. The final adjustment of the confidence intervals would be to replace (4-48) with

$$\text{Prob}[b_k - z_{(1-\alpha/2)}\sqrt{Est.Asy.Var[b_k]} \leq \beta_k \leq b_k + z_{(1-\alpha/2)}\sqrt{Est.Asy.Var[b_k]}] = 1 - \alpha, \qquad \textbf{(4-49)}$$

Example 4.8 Confidence Interval for the Income Elasticity of Demand for Gasoline

Using the gasoline market data discussed in Examples 4.2 and 4.4, we estimated the following demand equation using the 52 observations:

$$\ln(G/Pop) = \beta_1 + \beta_2 \ln P_G + \beta_3 \ln(Income/Pop) + \beta_4 \ln P_{nc} + \beta_5 \ln P_{uc} + \varepsilon.$$

Least squares estimates of the model parameters with standard errors and *t* ratios are given in Table 4.7. To form a confidence interval for the income elasticity, we need the critical value from the *t* distribution with $n - K = 52 - 5 = 47$ degrees of freedom. The 95% critical value is 2.012. Therefore a 95% confidence interval for β_3 is $1.095874 \pm 2.012\ (0.07771) = [0.9395, 1.2522]$.

4.7.2 CONFIDENCE INTERVAL FOR A LINEAR COMBINATION OF COEFFICIENTS: THE OAXACA DECOMPOSITION

In Example 4.8, we showed how to form a confidence interval for one of the elements of $\boldsymbol{\beta}$. By extending those results, we can show how to form a confidence interval for a

TABLE 4.7 Regression Results for a Demand Equation

Sum of squared residuals:	0.120871
Standard error of the regression:	0.050712
R^2 based on 52 observations	0.958443

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	−21.21109	0.75322	−28.160
ln P_G	−0.02121	0.04377	−0.485
ln *Income/Pop*	1.09587	0.07771	14.102
ln P_{nc}	−0.37361	0.15707	−2.379
ln P_{uc}	0.02003	0.10330	0.194

linear function of the parameters. **Oaxaca's (1973) and Blinder's (1973) decomposition** provides a frequently used application.[7]

Let $\mathbf{w}$ denote a $K \times 1$ vector of known constants. Then, the linear combination $c = \mathbf{w}'\mathbf{b}$ is asymptotically normally distributed with mean $\gamma = \mathbf{w}'\boldsymbol{\beta}$ and variance $\sigma_c^2 = \mathbf{w}'[Asy.Var[\mathbf{b}]]\mathbf{w}$, which we estimate with $s_c^2 = \mathbf{w}'[Est.Asy.Var[\mathbf{b}]]\mathbf{w}$. With these in hand, we can use the earlier results to form a confidence interval for γ :

$$\text{Prob}[c - z_{(1-\alpha/2)}s_c \leq \gamma \leq c + z_{(1-\alpha/2)}s_c] = 1 - \alpha. \tag{4-50}$$

This general result can be used, for example, for the sum of the coefficients or for a difference.

Consider, then, Oaxaca's (1973) application. In a study of labor supply, separate wage regressions are fit for samples of n_m men and n_f women. The underlying regression models are

$$\ln wage_{m,i} = \mathbf{x}'_{m,i}\boldsymbol{\beta}_m + \varepsilon_{m,i}, \quad i = 1, \ldots, n_m$$

and

$$\ln wage_{f,j} = \mathbf{x}'_{f,j}\boldsymbol{\beta}_f + \varepsilon_{f,j}, \quad j = 1, \ldots, n_f.$$

The regressor vectors include sociodemographic variables, such as age, and human capital variables, such as education and experience. We are interested in comparing these two regressions, particularly to see if they suggest wage discrimination. Oaxaca suggested a comparison of the regression functions. For any two vectors of characteristics,

$$\begin{aligned} E[\ln \text{wage}_{m,i}|\mathbf{x}_{m,i}] - E[\ln \text{wage}_{f,j}|\mathbf{x}_{f,i}] &= \mathbf{x}'_{m,i}\boldsymbol{\beta}_m - \mathbf{x}'_{f,j}\boldsymbol{\beta}_f \\ &= \mathbf{x}'_{m,i}\boldsymbol{\beta}_m - \mathbf{x}'_{m,i}\boldsymbol{\beta}_f + \mathbf{x}'_{m,i}\boldsymbol{\beta}_f - \mathbf{x}'_{f,j}\boldsymbol{\beta}_f \\ &= \mathbf{x}'_{m,i}(\boldsymbol{\beta}_m - \boldsymbol{\beta}_f) + (\mathbf{x}_{m,i} - \mathbf{x}_{f,i})'\boldsymbol{\beta}_f. \end{aligned}$$

The second term in this decomposition is identified with differences in human capital that would explain wage differences naturally, assuming that labor markets respond to these differences in ways that we would expect. The first term shows the differential in log wages that is attributable to differences unexplainable by human capital; holding these factors constant at $\mathbf{x}_m$ makes the first term attributable to other factors. Oaxaca suggested that this decomposition be computed at the means of the two regressor vectors, $\bar{\mathbf{x}}_m$ and $\bar{\mathbf{x}}_f$, and the least squares coefficient vectors, $\mathbf{b}_m$ and $\mathbf{b}_f$. If the regressions contain constant terms, then this process will be equivalent to analyzing $\overline{\ln y_m} - \overline{\ln y_f}$.

We are interested in forming a confidence interval for the first term, which will require two applications of our result. We will treat the two vectors of sample means as known vectors. Assuming that we have two independent sets of observations, our two estimators, $\mathbf{b}_m$ and $\mathbf{b}_f$, are independent with means $\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_f$ and estimated asymptotic covariance matrices $Est.Asy.Var[\mathbf{b}_m]$ and $Est.Asy.Var[\mathbf{b}_f]$. The covariance matrix of the difference is the sum of these two matrices. We are forming a confidence interval for $\bar{\mathbf{x}}'_m\,\mathbf{d}$ where $\mathbf{d} = \mathbf{b}_m - \mathbf{b}_f$. The estimated covariance matrix is

$$Est.Asy.Var[\mathbf{d}] = Est.Asy.Var[\mathbf{b}_m] + Est.Asy.Var[\mathbf{b}_f]. \tag{4-51}$$

Now we can apply the result above. We can also form a confidence interval for the second term; just define $\mathbf{w} = \bar{\mathbf{x}}_m - \bar{\mathbf{x}}_f$ and apply the earlier result to $\mathbf{w}'\mathbf{b}_f$.

[7]See Bourgignon et al. (2002) for an extensive application.

Example 4.9 Oaxaca Decomposition of Home Sale Prices

The town of Shaker Heights, Ohio, a suburb of Cleveland, developed in the twentieth century as a patchwork of neighborhoods associated with neighborhood-based school districts. Responding to changes in the demographic composition of the city, in 1987, Shaker Heights redistricted the neighborhoods. Some houses in some neighborhoods remained in the same school district while others in the same neighborhood were removed to other school districts. Bogart and Cromwell (2000) examined how this abrupt policy change affected home values in Shaker Heights by studying sale prices of houses before and after the change. Several econometric approaches were used.

- **Difference in Differences Regression:** Houses that did not change districts constituted a control group while those that did change constitute a treatment group. Sales take place both before and after the treatment date, 1987. A hedonic regression of home sale prices on attributes and the treatment and policy dummy variables reveals the causal effect of the policy change. (We will examine this method in Chapter 6.)
- **Repeat Sales:** Some homes were sold more than once. For those that sold both before and after the redistricting, a regression of the form

$$\ln\text{Price}_{i1} - \ln\text{Price}_{i0} = \text{time effects} + \text{school effects} + \Delta\text{redistricted}.$$

  The advantage of the first difference regression is that it effectively controls for and eliminates the characteristics of the house, and leaves only the persistent school effects and the effect of the policy change.
- **Oaxaca Decomposition:** Two hedonic regressions based on house characteristics are fit for different parts of neighborhoods where there are both houses that are in the neighborhood school areas and houses that are districted to other schools. The decomposition approach described above is applied to the two groups. The differences in the means of the sale prices are decomposed into a component that can be explained by differences in the house attributes and a residual effect that is suggested to be related to the benefit of having a neighborhood school. Figure 4.6 below shows the authors' main results for this part of the analysis.[8]

FIGURE 4.6 Results of Oaxaca Decomposition.

TABLE 6
Within Neighborhood Estimates of Neighborhood Schools Effect, Lomond Neighborhood (1987–1994)

Difference in mean house value	$6,545
Percent of difference due to district change	52.9%–59.1%
Effect of district change on mean house value	$3462–$3868
(decrease)	$3779
Dummy variable estimate of effect of district	476—same district
change	186—change district
Number of observations (662 total sales)	

Note: Percent of difference due to district change equals 100% minus the percent explained by differences in observable characteristics. Included characteristics are *heavy traffic, ln(frontage), ln(living area), ln(lot size), ln(age of house), average room size, plumbing fixtures, attached garage, finished attic, construction grade AA/A+, construction grade A, construction grade B or C or D, bad or fair condition, excellent condition,* and a set of year dummies. Regressions estimated using data from 1987 to 1994. Complete regression results available on request.

[8]Bogart and Cromwell (2000, p. 298).

4.8 PREDICTION AND FORECASTING

After the estimation of the model parameters, a common use of regression modeling is for prediction of the dependent variable. We make a distinction between *prediction* and *forecasting* most easily based on the difference between cross section and time-series modeling. **Prediction** (which would apply to either case) involves using the regression model to compute fitted (predicted) values of the dependent variable, either within the sample or for observations outside the sample. The same set of results will apply to cross sections, panels, and time series. We consider these methods first. **Forecasting**, while largely the same exercise, explicitly gives a role to time and often involves lagged dependent variables and disturbances that are correlated with their past values. This exercise usually involves predicting future outcomes. An important difference between predicting and forecasting (as defined here) is that for predicting, we are usually examining a scenario of our own design. Thus, in the example below in which we are predicting the prices of Monet paintings, we might be interested in predicting the price of a hypothetical painting of a certain size and aspect ratio, or one that actually exists in the sample. In the time-series context, we will often try to forecast an event such as real investment next year, not based on a hypothetical economy but based on our best estimate of what economic conditions will be next year. We will use the term **ex post prediction** (or **ex post forecast**) for the cases in which the data used in the regression equation to make the prediction are either observed or constructed experimentally by the analyst. This would be the first case considered here. An **ex ante forecast** (in the time-series context) will be one that requires the analyst to forecast the independent variables first before it is possible to forecast the dependent variable. In an exercise for this chapter, real investment is forecasted using a regression model that contains real GDP and the consumer price index. In order to forecast real investment, we must first forecast real GDP and the price index. Ex ante forecasting is considered briefly here and again in Chapter 20.

4.8.1 PREDICTION INTERVALS

Suppose that we wish to predict the value of y^0 associated with a regressor vector $\mathbf{x}^0$. The actual value would be

$$y^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0.$$

It follows from the Gauss–Markov theorem that

$$\hat{y}^0 = \mathbf{x}^{0\prime}\mathbf{b} \tag{4-52}$$

is the minimum variance linear unbiased estimator of $E[y^0|\mathbf{x}^0] = \mathbf{x}^{0\prime}\boldsymbol{\beta}$. The **prediction error** is

$$e^0 = \hat{y}^0 - y^0 = (\mathbf{b} - \boldsymbol{\beta})'\mathbf{x}^0 - \varepsilon^0.$$

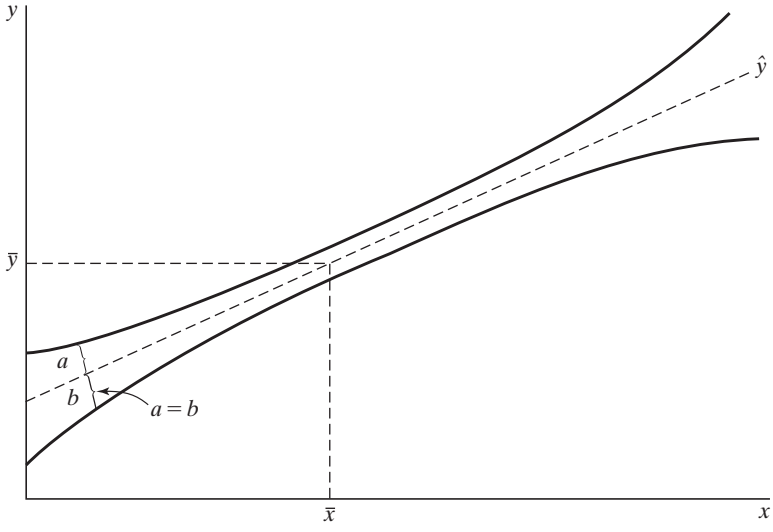
The **prediction variance** of this estimator based on (4-15) is

$$\text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \text{Var}[(\mathbf{b} - \boldsymbol{\beta})'\mathbf{x}^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \mathbf{x}^{0\prime}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x}^0. \tag{4-53}$$

If the regression contains a constant term, then an equivalent expression is

$$\text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2\left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}(x_j^0 - \bar{x}_j)(x_k^0 - \bar{x}_k)(\mathbf{Z}'\mathbf{M}^0\mathbf{Z})^{jk}\right], \tag{4-54}$$

FIGURE 4.7 Prediction Intervals.



where $\mathbf{Z}$ is the $K - 1$ columns of $\mathbf{X}$ not including the constant, $\mathbf{Z}'\mathbf{M}^0\mathbf{Z}$ is the matrix of sums of squares and products for the columns of $\mathbf{X}$ in deviations from their means [see (3-21)], and the "jk" superscript indicates the jk element of the inverse of the matrix. This result suggests that the width of a confidence interval (i.e., a **prediction interval**) depends on the distance of the elements of $\mathbf{x}^0$ from the center of the data. Intuitively, this idea makes sense; the farther the forecasted point is from the center of our experience, the greater is the degree of uncertainty. Figure 4.7 shows the effect for the bivariate case. Note that the prediction variance is composed of three parts. The second and third become progressively smaller as we accumulate more data (i.e., as n increases). But, the first term, σ^2 is constant, which implies that no matter how much data we have, we can never predict perfectly.

The prediction variance can be estimated by using s^2 in place of σ^2. A confidence (prediction) interval for y^0 would then be formed using

$$\text{prediction interval} = \hat{y}^0 \pm t_{(1-\alpha/2),[n-K]} se(e^0), \qquad \textbf{(4-55)}$$

where $t_{(1-\alpha/2),[n-K]}$ is the appropriate critical value for $100(1 - \alpha)$ % significance from the t table for $n - K$ degrees of freedom and $se(e^0)$ is the square root of the estimated prediction variance.

4.8.2 PREDICTING *Y* WHEN THE REGRESSION MODEL DESCRIBES LOG *y*

It is common to use the regression model to describe a function of the dependent variable, rather than the variable, itself. In Example 4.5 we model the sale prices of Monet paintings using

$$\ln Price = \beta_1 + \beta_2 \ln Area + \beta_3 Aspect\ Ratio + \varepsilon.$$

The log form is convenient in that the coefficient provides the elasticity of the dependent variable with respect to the independent variable, that is, in this model,

$\beta_2 = \partial E\,[\ln Price | \ln Area, AspectRatio]/\partial \ln Area$. However, the equation in this form is less interesting for prediction purposes than one that predicts the price itself. The natural approach for a predictor of the form

$$\ln y^0 = \mathbf{x}^{0\prime}\mathbf{b}$$

would be to use

$$\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b}).$$

The problem is that $E[y|\mathbf{x}^0]$ is not equal to $\exp(E[\ln y|\mathbf{X}^0])$. The appropriate conditional mean function would be

$$E[y|\mathbf{x}^0] = E[\exp(\mathbf{x}^{0\prime}\beta + \varepsilon^0)|\mathbf{x}^0] = \exp(\mathbf{x}^{0\prime}\beta)\,E[\exp(\varepsilon^0)|\mathbf{x}^0].$$

The second term is not $\exp(E[\varepsilon^0|\mathbf{x}^0]) = 1$ in general. The precise result if $\varepsilon^0|\mathbf{x}^0$ is normally distributed with mean zero and variance σ^2 is $E[\exp(\varepsilon^0)|\mathbf{x}^0] = \exp(\sigma^2/2)$. (See Section B.4.4.) The implication for normally distributed disturbances would be that an appropriate predictor for the conditional mean would be

$$\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b} + s^2/2) > \exp(\mathbf{x}^{0\prime}\mathbf{b}), \tag{4-56}$$

which would seem to imply that the naïve predictor would systematically underpredict y. However, this is not necessarily the appropriate interpretation of this result. The inequality implies that the naïve predictor will systematically underestimate the conditional mean function, not necessarily the realizations of the variable itself. The pertinent question is whether the conditional mean function is the desired predictor for the exponent of the dependent variable in the log regression. The conditional median might be more interesting, particularly for a financial variable such as income, expenditure, or the price of a painting. If the distribution of the variable in the log regression is symmetrically distributed (as they are when the disturbances are normally distributed), then the exponent will be asymmetrically distributed with a long tail in the positive direction, and the mean will exceed the median, possibly vastly so. In such cases, the median is often a preferred estimator of the center of a distribution. For estimating the median, rather then the mean, we would revert to the original naïve predictor, $\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b})$.

Given the preceding, we consider estimating $E[\exp(y)|\mathbf{x}^0]$. If we wish to avoid the normality assumption, then it remains to determine what one should use for $E[\exp(\varepsilon^0)|\mathbf{x}^0]$. Duan (1983) suggested the consistent estimator (assuming that the expectation is a constant, that is, that the regression is homoscedastic),

$$\hat{E}[\exp(\varepsilon^0)|\mathbf{x}^0] = h^0 = \frac{1}{n}\sum_{i=1}^{n}\exp(e_i), \tag{4-57}$$

where e_i is a least squares residual in the original log form regression. Then, Duan's **smearing estimator** for prediction of y^0 is

$$\hat{y}^0 = h^0 \exp(\mathbf{x}^{0\prime}\mathbf{b}).$$

4.8.3 PREDICTION INTERVAL FOR *Y* WHEN THE REGRESSION MODEL DESCRIBES LOG *y*

We obtained a prediction interval in (4-55) for $\ln y|\mathbf{x}^0$ in the loglinear model $\ln y = \mathbf{x}'\beta + \varepsilon$,

$$[\ln \hat{y}^0_{LOWER}, \ln \hat{y}^0_{UPPER}] = \left[\mathbf{x}^{0\prime}\mathbf{b} - t_{(1-\alpha/2),[n-K]}se(e^0), \mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha/2),[n-K]}se(e^0) \right].$$

For a given choice of α, say, 0.05, these values give the 0.025 and 0.975 quantiles of the distribution of $\ln y|\mathbf{x}^0$. If we wish specifically to estimate these quantiles of the distribution of $y|\mathbf{x}^0$, not $\ln y|\mathbf{x}^0$, then we would use:

$$\left[\hat{y}^0_{LOWER}, \hat{y}^0_{UPPER} \right] = \left\{ \exp\left[\mathbf{x}^{0\prime}\mathbf{b} - t_{(1-\alpha/2),[n-K]}se(e^0) \right], \exp\left[\mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha/2),[n-K]}se(e^0) \right] \right\}. \quad \textbf{(4-58)}$$

This follows from the result that if $\text{Prob}[\ln y \le \ln L] = 1 - \alpha/2$, then $\text{Prob}[y \le L] = 1 - \alpha/2$. The result is that the natural estimator is the right one for estimating the specific quantiles of the distribution of the original variable. However, if the objective is to find an interval estimator for $y|\mathbf{x}^0$ that is as narrow as possible, then this approach is not optimal. If the distribution of y is asymmetric, as it would be for a loglinear model with normally distributed disturbances, then the naïve interval estimator is longer than necessary. Figure 4.8 shows why. We suppose that (L, U) in the figure is the prediction interval formed by (4-58). Then the probabilities to the left of L and to the right of U each equal $\alpha/2$. Consider alternatives $L_0 = 0$ and U_0 instead. As we have constructed the figure, the area (probability) between L_0 and L equals the area between U_0 and U. But, because the density is so much higher at L, the distance $(0, U_0)$, the dashed interval, is visibly shorter than that between (L, U). The sum of the two tail probabilities is still equal to α, so this provides a shorter prediction interval. We could improve on (4-58) by using, instead, $(0, U_0)$, where U_0 is simply $\exp[\mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha),[n-K]}se(e^0)]$ (i.e., we put the entire tail area to the right of the upper value). However, while this is an improvement, it goes too far, as we now demonstrate.

Consider finding directly the shortest prediction interval. We treat this as an optimization problem,

$$\text{Minimize}(L, U): I = U - L \text{ subject to } F(L) + [1 - F(U)] = \alpha,$$

where F is the cdf of the random variable y (not $\ln y$). That is, we seek the shortest interval for which the two tail probabilities sum to our desired α (usually 0.05). Formulate this as a Lagrangean problem,

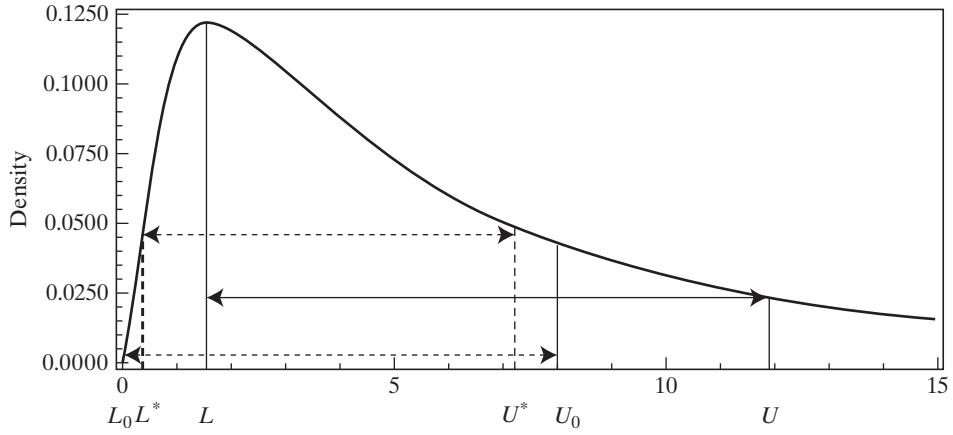
$$\text{Minimize}(L, U, \lambda): I^* = U - L + \lambda[F(L) + (1 - F(U)) - \alpha].$$

The solutions are found by equating the three partial derivatives to zero:

$$\partial I^*/\partial L = -1 + \lambda f(L) = 0,$$
$$\partial I^*/\partial U = 1 - \lambda f(U) = 0,$$
$$\partial I^*/\partial \lambda = F(L) + [1 - F(U)] - \alpha = 0,$$

where $f(L) = F'(L)$ and $f(U) = F'(U)$ are the derivatives of the cdf, which are the densities of the random variable at L and U, respectively. The third equation enforces the restriction that the two tail areas sum to α but does not force them to be equal. By adding the first two equations, we find that $\lambda[f(L) - f(U)] = 0$, which, if λ is not zero, means that the solution is obtained by locating (L^*, U^*) such that the tail areas sum to α

FIGURE 4.8 Lognormal Distribution for Prices of Monet Paintings.



and the densities are equal. Looking again at Figure 4.8, we can see that the solution we would seek is (L^*, U^*) where $0 < L^* < L$ and $U^* < U_0$. This is the shortest interval, and it is shorter than both $[0, U_0]$ and $[L, U]$.

This derivation would apply for any distribution, symmetric or otherwise. For a symmetric distribution, however, we would obviously return to the symmetric interval in (4-58). It provides the correct solution for when the distribution is asymmetric. In Bayesian analysis, the counterpart when we examine the distribution of a parameter conditioned on the data, is the **highest posterior density interval**. (See Section 16.4.2.) For practical application, this computation requires a specific assumption for the distribution of $y|\mathbf{x}^0$, such as lognormal. Typically, we would use the smearing estimator specifically to avoid the distributional assumption. There also is no simple formula to use to locate this interval, even for the lognormal distribution. A crude grid search would probably be best, though each computation is very simple. What this derivation does establish is that one can do substantially better than the naïve interval estimator, for example, using $[0, U_0]$.

Example 4.10 Pricing Art

In Examples 4.3 and 4.5, we examined an intriguing feature of the market for Monet paintings, that larger paintings sold at auction for more than smaller ones. Figure 4.9 shows a histogram for the sample of sale prices (in $million). Figure 4.10 shows a histogram for the logs of the prices. Results of the linear regression of lnPrice on lnArea (height times width) and Aspect Ratio (height divided by width) are given in Table 4.8.

We consider using the regression model to predict the price of one of the paintings, a 1903 painting of Charing Cross Bridge that sold for $3,522,500. The painting is 25.6″ high and 31.9″ wide. (This is observation 58 in the sample.) The log area equals $\ln(25.6 \times 31.9) = 6.705198$ and the aspect ratio equals $31.9/25.6 = 1.246094$. The prediction for the log of the price would be

$$\ln P|\mathbf{x}^0 = -8.34327 + 1.31638(6.705198) - 0.09623(1.246094) = 0.3643351$$

Note that the mean log price is 0.33274, so this painting is expected to sell for roughly 9.5% more than the average painting, based on its dimensions. The estimate of the prediction variance is computed using (4-53); $s_p = 1.105640$ The sample is large enough to use the

FIGURE 4.9 Histogram for Sale Prices of 430 Monet Paintings ($million).

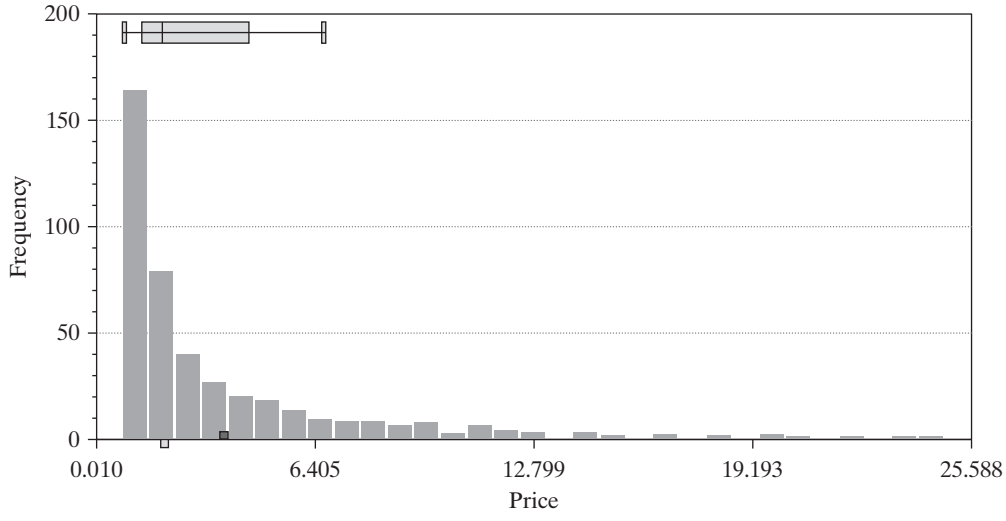
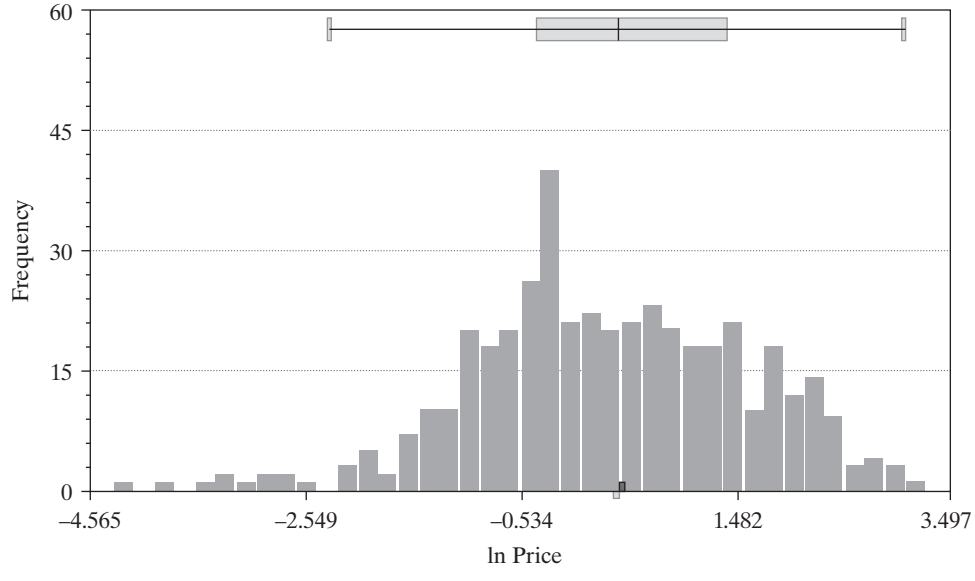


FIGURE 4.10 Histogram of Logs of Auction Prices for Monet Paintings.



critical value from the standard normal table, 1.96, for a 95% confidence interval. A prediction interval for the log of the price is therefore

$$0.364331 \pm 1.96(1.10564) = [-1.80272, 2.53140].$$

For predicting the price, the naïve predictor would be $\exp(0.3643351) = \$1.43956$M, which is far under the actual sale price of $3,522,500. To compute the smearing estimator, we require

TABLE 4.8 Estimated Equation for ln Price

Mean of ln Price	0.33274
Sum of squared residuals	520.765
Standard error of regression	1.10435
R-squared	0.33417
Adjusted *R*-squared	0.33105
Number of observations	430

Variable	*Coefficient*	*Standard Error*	*t Ratio*	*Mean of X*
Constant	−8.34327	0.67820	−12.30	1.00000
ln *Area*	1.31638	0.09205	14.30	6.68007
Aspect Ratio	−0.09623	0.15784	−0.61	1.23066

Estimated Asymptotic Covariance Matrix

	Constant	*ln Area*	*Aspect Ratio*
Constant	0.45996		
ln *Area*	−0.05969	0.00847	
Aspect Ratio	−0.04744	0.00251	0.02491

the mean of the exponents of the residuals, which is 1.81661. The revised point estimate for the price would thus be 1.81661 × 1.43956 = $2.61511M—this is better, but still fairly far off. This particular painting seems to have sold for relatively more than history (the data) would have predicted.

4.8.4 FORECASTING

The preceding discussion assumes that $\mathbf{x}^0$ is known with certainty, ex post, or has been forecast perfectly, ex ante. If $\mathbf{x}^0$ must, itself, be forecast (an ex ante forecast), then the formula for the forecast variance in (4-46) would have to be modified to incorporate the uncertainty in forecasting $\mathbf{x}^0$. This would be analogous to the term σ^2 in the prediction variance that accounts for the implicit prediction of ε^0. This will vastly complicate the computation. Many authors view it as simply intractable. Beginning with Feldstein (1971), derivation of firm analytical results for the correct forecast variance for this case remain to be derived except for simple special cases. The one qualitative result that seems certain is that (4-53) will understate the true variance. McCullough (1996) presents an alternative approach to computing appropriate forecast standard errors based on the method of bootstrapping. (See Chapter 15.)

Various measures have been proposed for assessing the predictive accuracy of forecasting models.[9] Most of these measures are designed to evaluate ex post forecasts; that is, forecasts for which the independent variables do not themselves have to be forecast. Two measures that are based on the residuals from the forecasts are the **root mean squared error**,

$$\text{RMSE} = \sqrt{\frac{1}{n^0}\sum_i (y_i - \hat{y}_i)^2},$$

[9]See Theil (1961) and Fair (1984).

and the **mean absolute error**,

$$\text{MAE} = \frac{1}{n^0}\sum_i |y_i - \hat{y}_i|,$$

where n^0 is the number of periods being forecasted. (Note that both of these, as well as the following measure below, are backward looking in that they are computed using the observed data on the independent variable.) These statistics have an obvious scaling problem—multiplying values of the dependent variable by any scalar multiplies the measure by that scalar as well. Several measures that are scale free are based on the **Theil *U* statistic**:[10]

$$U = \sqrt{\frac{(1/n^0)\sum_i (y_i - \hat{y}_i)^2}{(1/n^0)\sum_i y_i^2}}.$$

This measure is related to R^2 but is not bounded by zero and one. Large values indicate a poor forecasting performance.

4.9 DATA PROBLEMS

The analysis to this point has assumed that the data in hand, **X** and **y**, are well measured and correspond to the assumptions of the model and to the variables described by the underlying theory. At this point, we consider several ways that real-world observed nonexperimental data fail to meet the assumptions. Failure of the assumptions generally has implications for the performance of the estimators of the model parameters—unfortunately, none of them good. The cases we will examine are:

- **Multicollinearity:** Although the full rank assumption, A2, is met, it almost fails. (*Almost* is a matter of degree, and sometimes a matter of interpretation.) Multicollinearity leads to imprecision in the estimator, though not to any systematic biases in estimation.
- **Missing values:** Gaps in **X** and/or **y** can be harmless. In many cases, the analyst can (and should) simply ignore them, and just use the complete data in the sample. In other cases, when the data are missing for reasons that are related to the outcome being studied, ignoring the problem can lead to inconsistency of the estimators.
- **Measurement error:** Data often correspond only imperfectly to the theoretical construct that appears in the model—individual data on income and education are familiar examples. Measurement error is never benign. The least harmful case is measurement error in the dependent variable. In this case, at least under probably reasonable assumptions, the implication is to degrade the fit of the model to the data compared to the (unfortunately hypothetical) case in which the data are accurately measured. Measurement error in the regressors is malignant—it produces systematic biases in estimation that are difficult to remedy.

[10]Theil (1961).

4.9.1 MULTICOLLINEARITY

The Gauss–Markov theorem states that among all linear unbiased estimators, the least squares estimator has the smallest variance. Although this result is useful, it does not assure us that the least squares estimator has a small variance in any absolute sense. Consider, for example, a model that contains two explanatory variables and a constant. For either slope coefficient,

$$\text{Var}[b_k|\ \mathbf{X}] = \frac{\sigma^2}{(1 - r_{12}^2)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2} = \frac{\sigma^2}{(1 - r_{12}^2)S_{kk}}, k = 1, 2. \quad \textbf{(4-59)}$$

If the two variables are perfectly correlated, then the variance is infinite. The case of an exact linear relationship among the regressors is a serious failure of the assumptions of the model, not of the data. The more common case is one in which the variables are highly, but not perfectly, correlated. In this instance, the regression model retains all its assumed properties, although potentially severe statistical problems arise. The problem faced by applied researchers when regressors are highly, although not perfectly, correlated include the following symptoms:

- Small changes in the data produce wide swings in the parameter estimates.
- Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high.
- Coefficients may have the "wrong" sign or implausible magnitudes.

For convenience, define the data matrix, $\mathbf{X}$, to contain a constant and $K - 1$ other variables measured in deviations from their means. Let $\mathbf{x}_k$ denote the kth variable, and let $\mathbf{X}_{(k)}$ denote all the other variables (including the constant term). Then, in the inverse matrix, $(\mathbf{X}'\mathbf{X})^{-1}$, the kth diagonal element is

$$\begin{aligned}(\mathbf{x}_k'\mathbf{M}_{(k)}\mathbf{x}_k)^{-1} &= [\mathbf{x}_k'\mathbf{x}_k - \mathbf{x}_k'\mathbf{X}_{(k)}(\mathbf{X}_{(k)}'\mathbf{X}_{(k)})^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k]^{-1} \\ &= \left[\mathbf{x}_k'\mathbf{x}_k\left(1 - \frac{\mathbf{x}_k'\mathbf{X}_{(k)}(\mathbf{X}_{(k)}'\mathbf{X}_{(k)})^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k}{\mathbf{x}_k'\mathbf{x}_k}\right)\right]^{-1} \\ &= \frac{1}{(1 - R_{k.}^2)\, S_{kk}}, \end{aligned} \quad \textbf{(4-60)}$$

where $R_{k.}^2$ is the R^2 in the regression of x_k on all the other variables. In the multiple regression model, the variance of the kth least squares coefficient estimator is σ^2 times this ratio. It then follows that the more highly correlated a variable is with the other variables in the model (collectively), the greater its variance will be. In the most extreme case, in which $\mathbf{x}_k$ can be written as a linear combination of the other variables, so that $R_{k.}^2 = 1$, the variance becomes infinite. The result,

$$\text{Var}[b_k|\ \mathbf{X}] = \frac{\sigma^2}{(1 - R_{k.}^2)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}, \quad \textbf{(4-61)}$$

shows the three ingredients of the precision of the kth least squares coefficient estimator:

- Other things being equal, the greater the correlation of x_k with the other variables, the higher the variance will be, due to multicollinearity.
- Other things being equal, the greater the variation in x_k, the lower the variance will be.
- Other things being equal, the better the overall fit of the regression, the lower the variance will be. This result would follow from a lower value of σ^2.

Because nonexperimental data will never be orthogonal ($R_{k.}^2 = 0$), to some extent multicollinearity will always be present. When is multicollinearity a problem? That is, when are the variances of our estimates so adversely affected by this intercorrelation that we should be "concerned"? Some computer packages report a **variance inflation factor (VIF)**, $1/(1 - R_{k.}^2)$, for each coefficient in a regression as a diagnostic statistic. As can be seen, the VIF for a variable shows the increase in $\text{Var}[b_k]$ that can be attributable to the fact that this variable is not orthogonal to the other variables in the model. Another measure that is specifically directed at $\mathbf{X}$ is the **condition number** of $\mathbf{X}'\mathbf{X}$, which is the square root of the ratio of the largest characteristic root of $\mathbf{X}'\mathbf{X}$ to the smallest after scaling each column so that it has unit length. Values in excess of 20 are suggested as indicative of a problem [Belsley, Kuh, and Welsh (1980)]. (The condition number for the Longley data of Example 4.11 is over 15,000!)

Example 4.11 Multicollinearity in the Longley Data

The data in Appendix Table F4.2 were assembled by J. Longley (1967) for the purpose of assessing the accuracy of least squares computations by computer programs. (These data are still widely used for that purpose.[11]) The Longley data are notorious for severe multicollinearity. Note, for example, the last year of the data set. The last observation does not appear to be unusual. But the results in Table 4.9 show the dramatic effect of dropping this single observation from a regression of employment on a constant and the other variables. The last coefficient rises by 600%, and the third rises by 800%.

Several strategies have been proposed for finding and coping with multicollinearity.[12] Under the view that a multicollinearity problem arises because of a shortage of information, one suggestion is to obtain more data. One might argue that if analysts had such additional information available at the outset, they ought to have used it before reaching this juncture. More information need not mean more observations,

TABLE 4.9 Longley Results: Dependent Variable Is Employment

	1947–1961	*Variance Inflation*	*1947–1962*
Constant	1,459,415		1,169,087
Year	−721.756	143.4638	−576.464
GNP Deflator	−181.123	75.6716	−19.7681
GNP	0.0910678	132.467	0.0643940
Armed Forces	−0.0749370	1.55319	−0.0101453

[11]Computing the correct least squares coefficients with the Longley data is not a particularly difficult task by modern standards. The current standard benchmark is set by the NIST's "Filipelli Data." See www.itl.nist.gov/div898/strd/data/Filip.shtml. This application is considered in the Exercises.

[12]See Hill and Adkins (2001) for a description of the standard set of tools for diagnosing collinearity.

however. The obvious practical remedy (and surely the most frequently used) is to drop variables suspected of causing the problem from the regression—that is, to impose on the regression an assumption, possibly erroneous, that the *problem* variable does not appear in the model. If the variable that is dropped actually belongs in the model (in the sense that its coefficient, β_k, is not zero), then estimates of the remaining coefficients will be biased, possibly severely so. On the other hand, overfitting—that is, trying to estimate a model that is too large—is a common error, and dropping variables from an excessively specified model might have some virtue.

Using diagnostic tools to detect multicollinearity could be viewed as an attempt to distinguish a bad model from bad data. But, in fact, the problem only stems from a prior opinion with which the data seem to be in conflict. A finding that suggests multicollinearity is adversely affecting the estimates seems to suggest that, but for this effect, all the coefficients would be statistically significant and of the right sign. Of course, this situation need not be the case. If the data suggest that a variable is unimportant in a model, then, the theory notwithstanding, the researcher ultimately has to decide how strong the commitment is to that theory. Suggested remedies for multicollinearity might well amount to attempts to force the theory on the data.

As a response to what appears to be a multicollinearity problem, it is often difficult to resist the temptation to drop what appears to be an offending variable from the regression. This strategy creates a subtle dilemma for the analyst. Consider the partitioned multiple regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma + \varepsilon.$$

If we regress $\mathbf{y}$ only on $\mathbf{X}$, the estimator is biased:

$$E[\mathbf{b} \mid \mathbf{X}] = \boldsymbol{\beta} + \mathbf{p}_{\mathrm{X.z}}\gamma.$$

The covariance matrix of this estimator is

$$\mathrm{Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

(Keep in mind, this variance is around $E[\mathbf{b} \mid \mathbf{X}]$, not around $\boldsymbol{\beta}$.) If γ is not actually zero, then in the multiple regression of $\mathbf{y}$ on $(\mathbf{X}, \mathbf{z})$, the variance of $\mathbf{b}_{\mathrm{X.z}}$ around its mean, β would be

$$\begin{aligned}\mathrm{Var}[\mathbf{b}_{\mathrm{X.z}} \mid \mathbf{X},\mathbf{z}] &= \sigma^2(\mathbf{X}'\mathbf{M}_{\mathrm{z}}\mathbf{X})^{-1} \\ &= \sigma^2[\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{X}]^{-1}.\end{aligned}$$

To compare the two covariance matrices, it is simpler to compare their inverses. [See result (A-120).] Thus,

$$\{\mathrm{Var}[\mathbf{b} \mid \mathbf{X}]\}^{-1} - \{\mathrm{Var}[\mathbf{b}_{\mathrm{X.z}} \mid \mathbf{X},\mathbf{z}]\}^{-1} = (1/\sigma^2)\mathbf{X}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{X},$$

which is a nonnegative definite matrix. The implication is that the variance of $\mathbf{b}$ is not larger than the variance of $\mathbf{b}_{\mathbf{X.z}}$ (because its inverse is at least as large). It follows that although $\mathbf{b}$ is biased, its variance is never larger than the variance of the unbiased estimator. In any realistic case (i.e., if $\mathbf{X}'\mathbf{z}$ is not zero), in fact, it will be smaller. We get a useful comparison from a simple regression with two variables, x and z, measured as deviations from their means. Then, $\mathrm{Var}[b \mid \mathbf{x}] = \sigma^2/S_{xx}$ where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ while

$\text{Var}[b_{\mathbf{x},\mathbf{z}}|\mathbf{x},\mathbf{z}] = \sigma^2/[S_{xx}(1 - r_{xz}^2)]$ where r_{xz}^2 is the squared correlation between x and z. Clearly, $\text{Var}[b_{\text{x.z}}|\mathbf{x},\mathbf{z}]$ is larger.

The result in the preceding paragraph poses a bit of a dilemma for applied researchers. The situation arises frequently in the search for a model specification. Faced with a variable that a researcher suspects should be in the model, but that is causing a problem of multicollinearity, the analyst faces a choice of omitting the relevant variable or including it and estimating its (and all the other variables') coefficient imprecisely. This presents a choice between two estimators, the biased but precise b_1 and the unbiased but imprecise $b_{1.2}$. There is no accepted right answer to this dilemma, but as a general rule, the methodology leans away from estimation strategies that include ad hoc remedies for multicollinearity. For this particular case, there would be a general preference to retain z in the estimated model.

4.9.2 PRINCIPAL COMPONENTS

A device that has been suggested for reducing multicollinearity is to use a small number, say L, of **principal components** constructed as linear combinations of the K original variables.[13] (The mechanics are illustrated in Example 4.11.) The argument against using this approach is that if the original specification in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ were correct, then it is unclear what one is estimating when one regresses $\mathbf{y}$ on some small set of linear combinations of the columns of $\mathbf{X}$. For a set of $L < K$ principal components, if we regress $\mathbf{y}$ on $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$ to obtain $\mathbf{d}$, it follows that $E[\mathbf{d}] = \boldsymbol{\delta} = \mathbf{C}_L'\boldsymbol{\beta}$. (The proof is considered in the exercises.) In an economic context, if $\boldsymbol{\beta}$ has an interpretation, then it is unlikely that $\boldsymbol{\delta}$ will. For example, how do we interpret the price elasticity minus twice the income elasticity?

This orthodox interpretation cautions the analyst about mechanical devices for coping with multicollinearity that produce uninterpretable mixtures of the coefficients. But there are also situations in which the model is built on a platform that might well involve a mixture of some measured variables. For example, one might be interested in a regression model that contains *ability*, ambiguously defined. As a measured counterpart, the analyst might have in hand standardized scores on a set of tests, none of which individually has any particular meaning in the context of the model. In this case, a mixture of the measured test scores might serve as one's preferred proxy for the underlying variable. The study in Example 4.11 describes another natural example.

Example 4.12 Predicting Movie Success

Predicting the box office success of movies is a favorite exercise for econometricians.[14] The traditional predicting equation takes the form

$$\textit{Box Office Receipts} = f(\textit{Budget}, \textit{Genre}, \textit{MPAA Rating}, \textit{Star Power}, \textit{Sequel}, \textit{etc.}) + \varepsilon.$$

Coefficients of determination on the order of 0.4 are fairly common. Notwithstanding the relative power of such models, the common wisdom in Hollywood is "nobody knows." There is tremendous randomness in movie success, and few really believe they can forecast it with any reliability. Versaci (2009) added a new element to the model, "Internet buzz."

[13]See, for example, Gurmu, Rilstone, and Stern (1999).

[14]See, for example, Litman (1983), Ravid (1999), De Vany (2003), De Vany and Walls (1999, 2002, 2003), and Simonoff and Sparrow (2000).

Internet buzz is vaguely defined to be Internet traffic and interest on familiar Web sites such as RottenTomatoes.com, ImDB.com, Fandango.com, and traileraddict.com. None of these by itself defines Internet buzz. But, collectively, activity on these Web sites, say three weeks before a movie's opening, might be a useful predictor of upcoming success. Versaci's data set (Table F4.3) contains data for 62 movies released in 2009, including four Internet buzz variables, all measured three weeks prior to the release of the movie:

$buzz_1$ = number of Internet views of movie trailer at traileraddict.com

$buzz_2$ = number of message board comments about the movie at ComingSoon.net

$buzz_3$ = total number of "can't wait" (for release) plus "don't care" votes at Fandango.com

$buzz_4$ = percentage of Fandango votes that are "can't wait"

We have aggregated these into a single principal component as follows: We first computed the logs of $buzz_1 - buzz_3$ to remove the scale effects. We then standardized the four variables, so z_k contains the original variable minus its mean, $\bar{z}_k$, then divided by its standard deviation, s_k. Let $\mathbf{Z}$ denote the resulting 62 $\times$ 4 matrix (z_1, z_2, z_3, z_4). Then $\mathbf{V} = (1/61)\mathbf{Z}'\mathbf{Z}$ is the sample correlation matrix. Let c_1 be the characteristic vector of $\mathbf{V}$ associated with the largest characteristic root. The first principal component (the one that explains most of the variation of the four variables) is $\mathbf{Zc}_1$. (The roots are 2.4142, 0.7742, 0.4522, and 0.3585, so the first principal component explains 2.4142/4 or 60.3% of the variation. Table 4.10 shows the regression results for the sample of 62 2009 movies. It appears that Internet buzz adds substantially to the predictive power of the regression. The R^2 of the regression nearly doubles, from 0.34 to 0.59, when Internet buzz is added to the model. As we will discuss in Chapter 5, buzz is also a highly significant predictor of success.

4.9.3 MISSING VALUES AND DATA IMPUTATION

It is common for data sets to have gaps for a variety of reasons. Perhaps the most frequent occurrence of this problem is in survey data, in which respondents may simply

TABLE 4.10 Regression Results for Movie Success

	Internet Buzz Model			*Traditional Model*		
$e'e$	*22.30215*			*35.66514*		
R^2	*0.58883*			*0.34247*		
Variable	***Coefficient***	***Std.Error***	***t***	***Coefficient***	***Std.Error***	***t***
Constant	15.4002	0.64273	23.96	13.5768	0.68825	19.73
Action	−0.86932	0.29333	−2.96	−0.30682	0.34401	−0.89
Comedy	−0.01622	0.25608	−0.06	−0.03845	0.32061	−0.12
Animated	−0.83324	0.43022	−1.94	−0.82032	0.53869	−1.52
Horror	0.37460	0.37109	1.01	1.02644	0.44008	2.33
G	0.38440	0.55315	0.69	0.25242	0.69196	0.36
PG	0.53359	0.29976	1.78	0.32970	0.37243	0.89
PG13	0.21505	0.21885	0.98	0.07176	0.27206	0.26
ln *Budget*	0.26088	0.18529	1.41	0.70914	0.20812	3.41
Sequel	0.27505	0.27313	1.01	0.64368	0.33143	1.94
Star Power	0.00433	0.01285	0.34	0.00648	0.01608	0.40
Buzz	0.42906	0.07839	5.47	–	–	–

fail to respond to the questions. In a time series, the data may be missing because they do not exist at the frequency we wish to observe them; for example, the model may specify monthly relationships, but some variables are observed only quarterly. In panel data sets, the gaps in the data may arise because of **attrition** from the study. This is particularly common in health and medical research, when individuals choose to leave the study—possibly because of the success or failure of the treatment that is being studied.

There are several possible cases to consider, depending on why the data are missing. The data may be simply unavailable, for reasons unknown to the analyst and unrelated to the completeness or the values of the other observations in the sample. This is the most benign situation. If this is the case, then the complete observations in the sample constitute a usable data set, and the only issue is what possibly helpful information could be salvaged from the incomplete observations. Griliches (1986) calls this the **ignorable case** in that, for purposes of estimation, if we are not concerned with efficiency, then we may simply delete the incomplete observations and ignore the problem. Rubin (1976, 1987), Afifi and Elashoff (1966, 1967), and Little and Rubin (1987, 2002) label this case **missing completely at random (MCAR)**. A second case, which has attracted a great deal of attention in the econometrics literature, is that in which the gaps in the data set are not benign but are systematically related to the phenomenon being modeled. This case happens most often in surveys when the data are self-selected or self-reported. For example, if a survey were designed to study expenditure patterns and if high-income individuals tended to withhold information about their income, then the gaps in the data set would represent more than just missing information. The clinical trial case is another instance. In this (worst) case, the complete observations would be qualitatively different from a sample taken at random from the full population. The missing data in this situation are termed **not missing at random (NMAR)**. We treat this second case in Chapter 19 with the subject of **sample selection**, so we shall defer our discussion until later.

The intermediate case is that in which there is information about the missing data contained in the complete observations that can be used to improve inference about the model. The incomplete observations in this **missing at random (MAR)** case are also ignorable, in the sense that unlike the NMAR case, simply using the complete data does not induce any biases in the analysis, as long as the underlying process that produces the missingness in the data does not share parameters with the model that is being estimated, which seems likely.[15] This case is unlikely, of course, if "missingness" is based on the values of the dependent variable in a regression. Ignoring the incomplete observations when they are MAR but not MCAR does ignore information that is in the sample and therefore sacrifices some efficiency. Researchers have used a variety of **data imputation** methods to fill gaps in data sets. The (by far) simplest case occurs when the gaps occur in the data on the regressors. For the case of missing data on the regressors, it helps to consider the simple regression and multiple regression cases separately. In the first case, $\mathbf{X}$ has two columns: the column of 1s for the constant and a column with some blanks where the missing data would be if we had them. The **zero-order method** of replacing each missing x with $\bar{x}$ based on the observed data results in no changes and is equivalent to dropping the incomplete data. (See Exercise 7 in Chapter 3.) However, the R^2 will be lower. An alternative, **modified zero-order regression**, fills the second column of $\mathbf{X}$ with zeros and adds a variable that takes the value one for **missing observations**

[15]See Allison (2002).

and zero for complete ones. We leave it as an exercise to show that this is algebraically identical to simply filling the gaps with $\bar{x}$. These same methods can be used when there are multiple regressors. Once again, it is tempting to replace missing values of $\mathbf{x}_k$ with simple means of complete observations or with the predictions from linear regressions based on other variables in the model for which data are available when $\mathbf{x}_k$ is missing. In most cases in this setting, a general characterization can be based on the principle that for any missing observation, the *true* unobserved x_{ik} is being replaced by an erroneous proxy that we might view as $\hat{x}_{ik} = x_{ik} + u_{ik}$, that is, in the framework of **measurement error**. Generally, the least squares estimator is biased (and inconsistent) in the presence of measurement error such as this. (We will explore the issue in Chapter 8.) A question does remain: Is the bias likely to be reasonably small? As intuition should suggest, it depends on two features of the data: (1) how good the prediction of x_{ik} is in the sense of how large the variance of the measurement error, u_{ik}, is compared to that of the actual data, x_{ik}, and (2) how large a proportion of the sample the analyst is filling.

The regression method replaces each missing value on an $\mathbf{x}_k$ with a single prediction from a linear regression of $\mathbf{x}_k$ on other exogenous variables—in essence, replacing the missing x_{ik} with an estimate of it based on the regression model. In a Bayesian setting, some applications that involve unobservable variables (such as our example for a binary choice model in Chapter 17) use a technique called **data augmentation** to treat the unobserved data as unknown parameters to be estimated with the structural parameters, such as $\boldsymbol{\beta}$ in our regression model. Building on this logic researchers, for example, Rubin (1987) and Allison (2002), have suggested taking a similar approach in classical estimation settings. The technique involves a data imputation step that is similar to what was suggested earlier, but with an extension that recognizes the variability in the estimation of the regression model used to compute the predictions. To illustrate, we consider the case in which the independent variable, $\mathbf{x}_k$, is drawn in principle from a normal population, so it is a continuously distributed variable with a mean, a variance, and a joint distribution with other variables in the model. Formally, an imputation step would involve the following calculations:

1. Using as much information (complete data) as the sample will provide, linearly regress $\mathbf{x}_k$ on other variables in the model (and/or outside it, if other information is available), $\mathbf{Z}_k$, and obtain the coefficient vector $\mathbf{d}_k$ with associated asymptotic covariance matrix $\mathbf{A}_k$ and estimated disturbance variance s_k^2.
2. For purposes of the imputation, we draw an observation from the estimated asymptotic normal distribution of $\mathbf{d}_k$; that is, $\mathbf{d}_{k,m} = \mathbf{d}_k + \mathbf{v}_k$ where v_k is a vector of random draws from the normal distribution with mean zero and covariance matrix $\mathbf{A}_k$.
3. For each missing observation in $\mathbf{x}_k$ that we wish to impute, we compute $x_{i,k,m} = \mathbf{d}'_{k,m}\mathbf{z}_{i,k} + s_{k,m}u_{i,k}$, where $s_{k,m}$ is s_k divided by a random draw from the chi-squared distribution with degrees of freedom equal to the number of degrees of freedom in the imputation regression.

At this point, the iteration is the same as considered earlier, where the missing values are imputed using a regression, albeit a much more elaborate procedure. The regression is then computed, using the complete data and the imputed data for the missing observations, to produce coefficient vector, $\mathbf{b}_m$, and estimated covariance matrix, $\mathbf{V}_m$. This constitutes a single round. The technique of *multiple imputation* involves repeating

this set of steps M times. The estimators of the parameter vector and the appropriate asymptotic covariance matrix are

$$\hat{\boldsymbol{\beta}} = \bar{\mathbf{b}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{b}_{\mathbf{m}}, \tag{4-61}$$

$$\hat{\mathbf{V}} = \bar{\mathbf{V}} + \mathbf{B} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{V}_m + \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{m=1}^{M}(\mathbf{b}_m - \bar{\mathbf{b}})(\mathbf{b}_m - \bar{\mathbf{b}})'. \tag{4-62}$$

Researchers differ on the effectiveness or appropriateness of multiple imputation. When all is said and done, the measurement error in the imputed values remains. It takes very strong assumptions to establish that the multiplicity of iterations will suffice to average away the effect of this error. Very elaborate techniques have been developed for the special case of joint normally distributed cross sections of regressors such as those suggested above. However, the typical application to survey data involves gaps due to nonresponse to qualitative questions with binary answers. The efficacy of the theory is much less well developed for imputation of binary, ordered, count, or other qualitative variables.

Example 4.13 Imputation in the Survey of Consumer Finances[16]

The Survey of Consumer Finances (SCF) is a survey of U.S. households sponsored every three years by the Board of Governors of the Federal Reserve System with the cooperation of the U.S. Department of the Treasury. SCF interviews are conducted by NORC at the University of Chicago. Data from the SCF are used to inform monetary policy, tax policy, consumer protection, and a variety of other policy issues. The most recent release of the survey was in 2013. The 2016 survey is in process as of this writing. Missing data in the survey have been imputed five times using a multiple imputation technique. The information is stored in five separate imputation replicates (implicates). Thus, for the 6,026 families interviewed for the current survey, there are 30,130 records in the data set.[17] Rhine et al. (2016) used the Survey of Consumer Finances to examine savings behavior in the United States during the Great Recession of 2007–2009.

The more manageable case is missing values of the dependent variable, y_i. Once again, it must be the case that y_i is at least MAR and that the mechanism that is determining presence in the sample does not share parameters with the model itself. Assuming the data on $\mathbf{x}_i$ are complete for all observations, one might consider filling the gaps in the data on y_i by a two-step procedure: (1) estimate $\boldsymbol{\beta}$ with $\mathbf{b}_c$ using the complete observations, $\mathbf{X}_c$ and $\mathbf{y}_c$, then (2) fill the missing values, $\mathbf{y}_m$, with predictions, $\hat{\mathbf{y}}_m = \mathbf{X}_m\mathbf{b}_c$, and recompute the coefficients. We leave as an exercise (Exercise 17) to show that the second step estimator is exactly equal to the first. However, the variance estimator at the second step, s^2, must underestimate σ^2, intuitively because we are adding to the sample a set of observations that are fit perfectly.[18] So, this is not a beneficial way to proceed.

[16]See http://www.federalreserve.gov/econresdata/scf/scfindex.htm

[17]The Federal Reserve's download site for the SCF provides the following caution: *WARNING: Please review the following PDF for instructions on how to calculate correct standard errors. As a result of multiple imputation, the dataset you are downloading contains five times the number of actual observations. Failure to account for the imputations and the complex sample design will result in incorrect estimation of standard errors.* (Ibid.)

[18]See Cameron and Trivedi (2005, Chapter 27).

The flaw in the method comes back to the device used to impute the missing values for y_i. Recent suggestions that appear to provide some improvement involve using a randomized version, $\hat{\mathbf{y}}_m = \mathbf{X}_m\mathbf{b}_c + \hat{\boldsymbol{\varepsilon}}_m$, where $\hat{\boldsymbol{\varepsilon}}_m$ are random draws from the (normal) population with zero mean and estimated variance $s^2[\mathbf{I} + \mathbf{X}_m(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_m']$. (The estimated variance matrix corresponds to $\mathbf{X}_m\mathbf{b}_c + \boldsymbol{\varepsilon}_m$.) This defines an iteration. After reestimating $\boldsymbol{\beta}$ with the augmented data, one can return to re-impute the augmented data with the new $\hat{\boldsymbol{\beta}}$, then recompute $\mathbf{b}$, and so on. The process would continue until the estimated parameter vector stops changing. (A subtle point to be noted here: The same random draws should be used in each iteration. If not, there is no assurance that the iterations would ever converge.)

In general, not much is known about the properties of estimators based on using predicted values to fill missing values of y. Those results we do have are largely from simulation studies based on a particular data set or pattern of missing data. The results of these Monte Carlo studies are usually difficult to generalize. The overall conclusion seems to be that in a single-equation regression context, filling in missing values of y leads to biases in the estimator which are difficult to quantify. The only reasonably clear result is that imputations are more likely to be beneficial if the proportion of observations that are being filled is small—the smaller the better.

4.9.4 MEASUREMENT ERROR

There are any number of cases in which observed data are imperfect measures of their theoretical counterparts in the regression model. Examples include income, education, ability, health, the interest rate, output, capital, and so on. Mismeasurement of the variables in a model will generally produce adverse consequences for least squares estimation. Remedies are complicated and sometimes require heroic assumptions. In this section, we will provide a brief sketch of the issues. We defer to Section 8.8 for a more detailed discussion of the problem of measurement error, the most common solution (instrumental variables estimation), and some applications.

It is convenient to distinguish between measurement error in the dependent variable and measurement error in the regressor(s). For the second case, it is also useful to consider the simple regression case and then extend it to the multiple regression model. Consider a model to describe expected income in a population,

$$I^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \tag{4-63}$$

where I^* is the intended total income variable. Suppose the observed counterpart is I, earnings. How I relates to I^* is unclear; it is common to assume that the measurement error is additive, so $I = I^* + w$. Inserting this expression for I into (4-63) gives

$$\begin{aligned} I &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon + w \\ &= \mathbf{x}'\boldsymbol{\beta} + \upsilon, \end{aligned} \tag{4-64}$$

which appears to be a slightly more complicated regression, but otherwise similar to what we started with. As long as w and $\mathbf{x}$ are uncorrelated, that is the case. If w is a homoscedastic zero mean error that is uncorrelated with $\mathbf{x}$, then the only difference between the models in (4-63) and (4-64) is that the disturbance variance in (4-64) is $\sigma_w^2 + \sigma_\varepsilon^2 > \sigma_\varepsilon^2$. Otherwise both are regressions and evidently $\boldsymbol{\beta}$ can be estimated consistently by least squares in either case. The cost of the measurement error is in the

precision of the estimator because the asymptotic variance of the estimator in (4-64) is $(\sigma_v^2/n)[\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$, while it is $(\sigma_\varepsilon^2/n)[\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$ if $\boldsymbol{\beta}$ is estimated using (4-63). The measurement error also costs some fit. To see this, note that the R^2 in the sample regression in (4-63) is

$$R_*^2 = 1 - (\mathbf{e}'\mathbf{e}/n)/(\mathbf{I}^{*'}\mathbf{M}^0\mathbf{I}^*/n).$$

The numerator converges to σ_ε^2 while the denominator converges to the total variance of I^*, which would approach $\sigma_\varepsilon^2 + \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}$ where $\mathbf{Q} = \text{plim}(\mathbf{X}'\mathbf{X}/n)$. Therefore,

$$\text{plim}R_*^2 = \boldsymbol{\beta}'Q\boldsymbol{\beta}/[\sigma_\varepsilon^2 + \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}.$$

The counterpart for (4-64), R^2, differs only in that σ_ε^2 is replaced by $\sigma_v^2 > \sigma_\varepsilon^2$ in the denominator. It follows that

$$\text{plim } R_*^2 - \text{plim } R^2 > 0.$$

This implies that the fit of the regression in (4-64) will, at least broadly in expectation, be inferior to that in (4-63). (The preceding is an asymptotic approximation that might not hold in every finite sample.)

These results demonstrate the implications of measurement error in the dependent variable. We note, in passing, that if the measurement error is not additive, if it is correlated with $\mathbf{x}$, or if it has any other features such as heteroscedasticity, then the preceding results are lost, and nothing in general can be said about the consequence of the measurement error. Whether there is a *solution* is likewise an ambiguous question. The preceding explanation shows that it would be better to have the underlying variable if possible. In the absence, would it be preferable to use a proxy? Unfortunately, I is already a proxy, so unless there exists an available I' which has smaller measurement error variance, we have reached an impasse. On the other hand, it does seem that the outcome is fairly benign. The sample does not contain as much information as we might hope, but it does contain sufficient information consistently to estimate β and to do appropriate statistical inference based on the information we do have.

The more difficult case occurs when the measurement error appears in the independent variable(s). For simplicity, we retain the symbols I and I^* for our observed and theoretical variables. Consider a simple regression,

$$y = \beta_1 + \beta_2 I^* + \varepsilon,$$

where y is the perfectly measured dependent variable and the same measurement equation, $I = I* + w$, applies now to the independent variable. Inserting I into the equation and rearranging a bit, we obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 I + (\varepsilon - \beta_2 w) \\ &= \beta_1 + \beta_2 I + \upsilon. \end{aligned} \tag{4-65}$$

It appears that we have obtained (4-64) once again. Unfortunately, this is not the case, because $\text{Cov}[I, \upsilon] = \text{Cov}[I^* + w, \varepsilon - \beta_2 w] = -\beta_2\sigma_w^2$. Because the regressor in (4-65) is correlated with the disturbance, least squares regression in this case is inconsistent. There is a bit more that can be derived—this is pursued in Section 8.5, so we state it here without proof. In this case,

$$\text{plim } b_2 = \beta_2[\sigma_*^2/(\sigma_*^2 + \sigma_w^2)],$$

where σ_*^2 is the marginal variance of I^*. The scale factor is less than one, so the least squares estimator is biased toward zero. The larger the measurement error variance, the worse is the bias. (This is called **least squares attenuation**.) Now, suppose there are additional variables in the model:

$$y = \mathbf{x}'\boldsymbol{\beta}_1 + \beta_2 I^* + \varepsilon.$$

In this instance, almost no useful theoretical results are forthcoming. The following fairly general conclusions can be drawn—once again, proofs are deferred to Section 8.5:

1. The least squares estimator of β_2 is still biased toward zero.
2. All the elements of the estimator of $\boldsymbol{\beta}_1$ are biased, in unknown directions, even though the variables in $\mathbf{x}$ are not measured with error.

Solutions to the "measurement error problem" come in two forms. If there is outside information on certain model parameters, then it is possible to deduce the scale factors (using the **method of moments**) and undo the bias. For the obvious example, in (4-65), if σ_w^2 were known, then it would be possible to deduce σ_*^2 from $\text{Var}[I] = \sigma_*^2 + \sigma_w^2$ and thereby compute the necessary scale factor to undo the bias. This sort of information is generally not available. A second approach that has been used in many applications is the technique of instrumental variables. This is developed in detail for this application in Section 8.5.

4.9.5 OUTLIERS AND INFLUENTIAL OBSERVATIONS

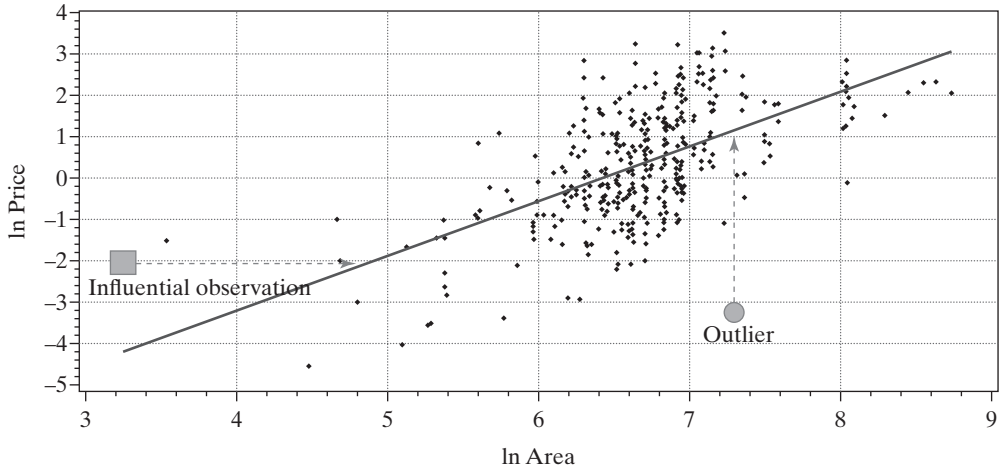
Figure 4.10 shows a scatter plot of the data on sale prices of Monet paintings that were used in Example 4.5. Two points have been highlighted. The one noted with the square overlay shows the smallest painting in the data set. The circle highlights a painting that fetched an unusually low price, at least in comparison to what the regression would have predicted. (It was not the least costly painting in the sample, but it was the one most poorly predicted by the regression.) Because least squares is based on squared deviations, the estimator is likely to be strongly influenced by extreme observations such as these, particularly if the sample is not very large.

An *influential observation* is one that is likely to have a substantial impact on the least squares regression coefficient(s). For a simple regression such as the one shown in Figure 4.11, Belsley, Kuh, and Welsh (1980) defined an influence measure, for observation x_i,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_{(i)})^2}{\sum_{j=1, j\neq i}^{n}(x_j - \bar{x}_{(i)})^2}, \tag{4-66}$$

where $\bar{x}_{(i)}$ and the summation in the denominator of the fraction are computed without this observation. (The measure derives from the difference between $\mathbf{b}$ and $\mathbf{b}_{(i)}$ where the latter is computed without the particular observation. We will return to this shortly.) It is suggested that an observation should be noted as influential if $h_i > 2/n$. The decision is whether to drop the observation or not. We should note observations with high leverage are arguably not outliers (which remains to be defined) because the analysis is conditional on x_i. To underscore the point, referring to Figure 4.11, this observation would be marked even if it fell precisely on the regression line—the source of the influence is the numerator of the second term in h_i, which is unrelated to the distance of the point from the line. In our example, the influential observation happens to be the

FIGURE 4.11 Log Price Versus Log Area for Monet Paintings.



result of Monet's decision to paint a small painting. The point is that in the absence of an underlying theory that explains (and justifies) the extreme values of x_i, eliminating such observations is an algebraic exercise that has the effect of forcing the regression line to be fitted with the values of x_i closest to the means.

The change in the linear regression coefficient vector in a multiple regression when an observation is added to the sample is

$$\mathbf{b} - \mathbf{b}_{(i)} = \Delta\mathbf{b} = \frac{1}{1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i}(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i(\mathbf{y}_i - \mathbf{x}_i'\mathbf{b}_{(i)}), \tag{4-67}$$

where $\mathbf{b}$ is computed with observation i in the sample, $\mathbf{b}_{(i)}$ is computed without observation i, and $\mathbf{X}_{(i)}$ does not include observation i. (See Exercise 5 in Chapter 3.) It is difficult to single out any particular feature of the observation that would drive this change. The influence measure,

$$\begin{aligned} h_{ii} &= \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i \\ &= \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}(x_{i,j} - \bar{x}_{n,j})(x_{i,k} - \bar{x}_k)(\mathbf{Z}_{(i)}'\mathbf{M}^0\mathbf{Z}_{(i)})^{jk}, \end{aligned} \tag{4-68}$$

has been used to flag influential observations.[19] In this instance, the selection criterion would be $h_{ii} > 2(K - 1)/n$. Squared deviations of the elements of $\mathbf{x}_i$ from the means of the variables appear in h_{ii}, so it is also operating on the difference of $\mathbf{x}_i$ from the center of the data. (See expression (4-54) for the forecast variance in Section 4.8.1 for an application.)

In principle, an outlier is an observation that appears to be outside the reach of the model, perhaps because it arises from a different data-generating process. The outlier

[19]See, once again, Belsley, Kuh, and Welsh (1980) and Cook (1977).

in Figure 4.11 appears to be a candidate. Outliers could arise for several reasons. The simplest explanation would be actual data errors. Assuming the data are not erroneous, it then remains to define what constitutes an outlier. Unusual residuals are an obvious choice. But, because the distribution of the disturbances would anticipate a certain small percentage of extreme observations in any event, simply singling out observations with large residuals is actually a dubious exercise. On the other hand, one might suspect that the outlying observations are actually generated by a different population. *Studentized residuals* are constructed with this in mind by computing the regression coefficients and the residual variance without observation i for each observation in the sample and then standardizing the modified residuals. The ith studentized residual is

$$e(i) = \frac{e_i}{\sqrt{1 - h_{ii}}} \Bigg/ \sqrt{\frac{\mathbf{e}'\mathbf{e} - e_i^2/(1 - h_{ii})}{n - 1 - K}}, \tag{4-69}$$

where $\mathbf{e}$ is the residual vector for the full sample, based on $\mathbf{b}$, including e_i the residual for observation i. In principle, this residual has a t distribution with $n - 1 - K$ degrees of freedom (or a standard normal distribution asymptotically). Observations with large studentized residuals, that is, greater than 2.0, would be singled out as outliers.

There are several complications that arise with isolating outlying observations in this fashion. First, there is no a priori assumption of which observations are from the alternative population, if this is the view. From a theoretical point of view, this would suggest a skepticism about the model specification. If the sample contains a substantial proportion of outliers, then the properties of the estimator based on the reduced sample are difficult to derive. In the next application, the suggested procedure deletes 4.2% of the sample (18 observations). Finally, it will usually occur that observations that were not outliers in the original sample will become outliers when the original set of outliers is removed. It is unclear how one should proceed at this point. (Using the Monet paintings data, the first round of studentizing the residuals removes 18 observations. After 11 iterations, the sample size stabilizes at 364 of the original 430 observations, a reduction of 15.3%.) Table 4.11 shows the original results (from Table 4.4) and the modified results with 18 outliers removed. Given that the 430 is a relatively large sample, the modest change in the results is to be expected.

TABLE 4.11 Estimated Equations for Log Price

Number of observations	430	412
Mean of log price	0.33274	0.36328
Sum of squared residuals	520.765	393.845
Standard error of regression	1.10435	0.98130
R-squared	0.33417	0.38371
Adjusted R-squared	0.33105	0.38070

	Coefficient		*Standard Error*		*t*	
Variable	$n = 430$	$n = 412$	$n = 430$	$n = 412$	$n = 430$	$n = 412$
Constant	−8.34237	−8.62152	0.67820	0.62524	−12.30	−13.79
ln *Area*	1.31638	1.35777	0.09205	0.08612	14.30	15.77
Aspect Ratio	−0.09623	−0.08346	0.15784	0.14569	−0.61	−0.57

It is difficult to draw firm general conclusions from this exercise. It remains likely that in very small samples, some caution and close scrutiny of the data are called for. If it is suspected at the outset that a process prone to large observations is at work, it may be useful to consider a different estimator altogether, such as least absolute deviations, or even a different model specification that accounts for this possibility. For example, the idea that the sample may contain some observations that are generated by a different process lies behind the latent class model that is discussed in Chapters 14 and 18.

4.10 SUMMARY AND CONCLUSIONS

This chapter has examined a set of properties of the least squares estimator that will apply in all samples, including unbiasedness and efficiency among unbiased estimators. The formal assumptions of the linear model are pivotal in the results of this chapter. All of them are likely to be violated in more general settings than the one considered here. For example, in most cases examined later in the book, the estimator has a possible bias, but that bias diminishes with increasing sample sizes. For purposes of forming confidence intervals and testing hypotheses, the assumption of normality is narrow, so it was necessary to extend the model to allow nonnormal disturbances. These and other "large-sample" extensions of the linear model were considered in Section 4.4. The crucial results developed here were the consistency of the estimator and a method of obtaining an appropriate covariance matrix and large-sample distribution that provides the basis for forming confidence intervals and testing hypotheses. Statistical inference in the form of interval estimation for the model parameters and for values of the dependent variable was considered in Sections 4.6 and 4.7. This development will continue in Chapter 5 where we will consider hypothesis testing and model selection.

Finally, we considered some practical problems that arise when data are less than perfect for the estimation and analysis of the regression model, including multicollinearity, missing observations, measurement error, and outliers.

Key Terms and Concepts

- Assumptions
- Asymptotic covariance matrix
- Asymptotic distribution
- Asymptotic efficiency
- Asymptotic normality
- Asymptotic properties
- Attrition
- Bootstrapping
- Condition number
- Confidence intervals
- Consistency
- Consistent estimator
- Data imputation
- Efficient scale
- Estimator
- Ex ante forecast
- Ex post forecast
- Ex post predication
- Finite sample properties
- Gauss–Markov theorem
- Grenander conditions
- Highest posterior density interval
- Ignorable case
- Interval estimation
- Least squares attenuation
- Lindeberg–Feller Central Limit Theorem
- Linear estimator
- Linear unbiased estimator
- Mean absolute error
- Mean squared error
- Measurement error
- Method of moments
- Minimum mean squared error
- Minimum variance linear unbiased estimator
- Missing at random (MAR)
- Missing completely at random (MCAR)
- Missing observations
- Modified zero-order regression

- Monte Carlo study
- Multicollinearity
- Not missing at random (NMAR)
- Oaxaca's and Blinder's decomposition
- Optimal linear predictor
- Panel data
- Point estimation
- Prediction error
- Prediction interval
- Prediction variance
- Principal components
- Probability limit
- Root mean squared error
- Sample selection
- Sampling distribution
- Sampling variance
- Semiparametric
- Smearing estimator
- Standard error
- Standard error of the regression
- Statistical properties
- Theil U statistic
- Variance inflation factor (VIF)
- Zero-order method

Exercises

1. Suppose that you have two independent unbiased estimators of the same parameter θ, say $\hat{\theta}_1$ and $\hat{\theta}_2$, with different variances v_1 and v_2. What linear combination $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ is the minimum variance unbiased estimator of θ?
2. Consider the simple regression $y_i = \beta x_i + \varepsilon_i$ where $E[\varepsilon|x] = 0$ and $E[\varepsilon^2|x] = \sigma^2$
   a. What is the minimum mean squared error linear estimator of β? [*Hint:* Let the estimator be $(\hat{\beta} = \mathbf{c}'\mathbf{y})$. Choose $\mathbf{c}$ to minimize $\text{Var}(\hat{\beta}) + (E(\hat{\beta} - \beta))^2$. The answer is a function of the unknown parameters.]
   b. For the estimator in part a, show that ratio of the mean squared error of $\hat{\beta}$ to that of the ordinary least squares estimator b is
   $$\frac{\text{MSE}[\hat{\beta}]}{\text{M}SE[b]} = \frac{\tau^2}{(1 + \tau^2)}, \text{ where } \tau^2 = \frac{\beta^2}{[\sigma^2/\mathbf{X}'\mathbf{X}]}.$$
   Note that τ is the population analog to the "t ratio" for testing the hypothesis that $\beta = 0$, which is given in (5-11). How do you interpret the behavior of this ratio as $\tau \to \infty$?
3. Suppose that the classical regression model applies but that the true value of the constant is zero. Compare the variance of the least squares slope estimator computed without a constant term with that of the estimator computed with an unnecessary constant term.
4. Suppose that the regression model is $y_i = \alpha + \beta x_i + \varepsilon_i$, where the disturbances ε_i have $f(\varepsilon_i) = (1/\lambda)\exp(-\varepsilon_i/\lambda)$, $\varepsilon_i \geq 0$. This model is rather peculiar in that all the disturbances are assumed to be nonnegative. Note that the disturbances have $E[\varepsilon_i|x_i] = \lambda$ and $\text{Var}[\varepsilon_i|x_i] = \lambda^2$. Show that the least squares slope estimator is unbiased but that the intercept estimator is biased.
5. Prove that the least squares intercept estimator in the classical regression model is the minimum variance linear unbiased estimator.
6. As a profit-maximizing monopolist, you face the demand curve $Q = \alpha + \beta P + \varepsilon$. In the past, you have set the following prices and sold the accompanying quantities:

Q	3	3	7	6	10	15	16	13	9	15	9	15	12	18	21
P	18	16	17	12	15	15	4	13	11	6	8	10	7	7	7

   Suppose that your marginal cost is 10. Based on the least squares regression, compute a 95% confidence interval for the expected value of the profit-maximizing output.

7. The following sample moments for $x = [1, x_1, x_2, x_3]$ were computed from 100 observations produced using a random number generator:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 100 & 123 & 96 & 109 \\ 123 & 252 & 125 & 189 \\ 96 & 125 & 167 & 146 \\ 109 & 189 & 146 & 168 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 460 \\ 810 \\ 615 \\ 712 \end{bmatrix}, \quad \mathbf{y}'\mathbf{y} = 3924.$$

The true model underlying these data is $y = x_1 + x_2 + x_3 + \varepsilon$.
   a. Compute the simple correlations among the regressors.
   b. Compute the ordinary least squares coefficients in the regression of y on a constant x_1, x_2, and x_3.
   c. Compute the ordinary least squares coefficients in the regression of y on a constant, x_1 and x_2, on a constant, x_1 and x_3, and on a constant, x_2 and x_3.
   d. Compute the variance inflation factor associated with each variable.
   e. The regressors are obviously badly collinear. Which is the problem variable? Explain.
8. Consider the multiple regression of $\mathbf{y}$ on K variables $\mathbf{X}$ and an additional variable $\mathbf{z}$. Prove that under the assumptions A1 through A6 of the classical regression model, the true variance of the least squares estimator of the slopes on $\mathbf{X}$ is larger when $\mathbf{z}$ is included in the regression than when it is not. Does the same hold for the sample estimate of this covariance matrix? Why or why not? Assume that $\mathbf{X}$ and $\mathbf{z}$ are nonstochastic and that the coefficient on $\mathbf{z}$ is nonzero.
9. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, assuming that the true value of β is zero, what is the exact expected value of $F[K, n - K] = (R^2/K)/[(1 - R^2)/(n - K)]$?
10. Prove that $E[\mathbf{b}'\mathbf{b}] = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{k=1}^{K} (1/\lambda_k)$, where $\mathbf{b}$ is the ordinary least squares estimator and λ_k is a characteristic root of $\mathbf{X}'\mathbf{X}$.
11. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, what is plim $F[K, n - K] = \text{plim} \frac{R^2/K}{(1 - R^2)/(n - K)}$, assuming that the true value of $\boldsymbol{\beta}$ is zero?
12. Let e_i be the ith residual in the ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ in the classical regression model, and let ε_i be the corresponding true disturbance. Prove that plim$(e_i - \varepsilon_i) = 0$.
13. For the simple regression model $y_i = \mu + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2]$, prove that the sample mean is consistent and asymptotically normally distributed. Now consider the alternative estimator $\hat{\mu} = \sum_i w_i y_i, w_i = \frac{i}{(n(n + 1)/2)} = \frac{i}{\sum_i i}$. Note that $\sum_i w_i = 1$. Prove that this is a consistent estimator of μ and obtain its asymptotic variance. [*Hint:* $\sum_i i^2 = n(n + 1)(2n + 1)/6$.]
14. Consider a data set consisting of n observations, n_c complete and n_m incomplete, for which the dependent variable, y_i, is missing. Data on the independent variables, $\mathbf{x}_i$, are complete for all n observations, $\mathbf{X}_c$ and $\mathbf{X}_m$. We wish to use the data to estimate the parameters of the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Consider the following the imputation strategy: Step 1: Linearly regress $\mathbf{y}_c$ on $\mathbf{X}_c$ and compute $\mathbf{b}_c$.

Step 2: Use $\mathbf{X}_m$ to predict the missing $\mathbf{y}_m$ with $\mathbf{X}_m\mathbf{b}_c$. Then regress the full sample of observations, $(\mathbf{y}_c, \mathbf{X}_m\mathbf{b}_c)$, on the full sample of regressors, $(\mathbf{X}_c, \mathbf{X}_m)$.

a. Show that the first and second step least squares coefficient vectors are identical.
b. Is the second step coefficient estimator unbiased?
c. Show that the sum of squared residuals is the same at both steps.
d. Show that the second step estimator of σ^2 is biased downward.

15. In (4-13), we find that when superfluous variables $\mathbf{X}_2$ are added to the regression of $\mathbf{y}$ on $\mathbf{X}_1$ the least squares coefficient estimator is an unbiased estimator of the true parameter vector, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \mathbf{0}')'$. Show that, in this long regression, $\mathbf{e}'\mathbf{e}/(n - K_1 - K_2)$ is also unbiased as estimator of σ^2.
16. In Section 4.9.2, we consider regressing $\mathbf{y}$ on a set of principal components, rather than the original data. For simplicity, assume that $\mathbf{X}$ does not contain a constant term, and that the K variables are measured in deviations from the means and are standardized by dividing by the respective standard deviations. We consider regression of $\mathbf{y}$ on L principal components, $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$, where $L < K$. Let $\mathbf{d}$ denote the coefficient vector. The regression model is $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. In the discussion, it is claimed that $E[\mathbf{d}] = \mathbf{C}_L'\beta$. Prove the claim.
17. Example 4.10 presents a regression model that is used to predict the auction prices of Monet paintings. The most expensive painting in the sample sold for \$33.0135M (ln = 17.3124). The height and width of this painting were 35″ and 39.4″, respectively. Use these data and the model to form prediction intervals for the log of the price and then the price for this painting.

Applications

1. Data on U.S. gasoline consumption for the years 1953 to 2004 are given in Table F2.2. Note the consumption data appear as total expenditure. To obtain the per capita quantity variable, divide GASEXP by GASP times Pop. The other variables do not need transformation.
   a. Compute the multiple regression of per capita consumption of gasoline on per capita income, the price of gasoline, the other prices, and a time trend. Report all results. Do the signs of the estimates agree with your expectations?
   b. Test the hypothesis that at least in regard to demand for gasoline, consumers do not differentiate between changes in the prices of new and used cars.
   c. Estimate the own price elasticity of demand, the income elasticity, and the cross-price elasticity with respect to changes in the price of public transportation. Do the computations at the 2004 point in the data.
   d. Reestimate the regression in logarithms so that the coefficients are direct estimates of the elasticities. (Do not use the log of the time trend.) How do your estimates compare with the results in the previous question? Which specification do you prefer?
   e. Compute the simple correlations of the price variables. Would you conclude that multicollinearity is a problem for the regression in part a or part d?
   f. Notice that the price index for gasoline is normalized to 100 in 2000, whereas the other price indices are anchored at 1983 (roughly). If you were to renormalize the indices so that they were all 100.00 in 2004, then how would the results of

the regression in part a change? How would the results of the regression in part d change?

g. This exercise is based on the model that you estimated in part d. We are interested in investigating the change in the gasoline market that occurred in 1973. First, compute the average values of log of per capita gasoline consumption in the years 1953–1973 and 1974–2004 and report the values and the difference. If we divide the sample into these two groups of observations, then we can decompose the change in the expected value of the log of consumption into a change attributable to change in the regressors and a change attributable to a change in the model coefficients, as shown in Section 4.7.2. Using the Oaxaca–Blinder approach described there, compute the decomposition by partitioning the sample and computing separate regressions. Using your results, compute a confidence interval for the part of the change that can be attributed to structural change in the market, that is, change in the regression coefficients.

2. Christensen and Greene (1976) estimated a "generalized Cobb–Douglas" cost function for electricity generation of the form

$$\ln C = \alpha + \beta \ln Q + \gamma[\tfrac{1}{2}(\ln Q)^2] + \delta_k \ln P_k + \delta_l \ln P_l + \delta_f \ln P_f + \varepsilon.$$

P_k, P_l, and P_f indicate unit prices of capital, labor, and fuel, respectively, Q is output, and C is total cost. To conform to the underlying theory of production, it is necessary to impose the restriction that the cost function be homogeneous of degree one in the three prices. This is done with the restriction $\delta_k + \delta_l + \delta_f = 1$, or $\delta_f = 1 - \delta_k - \delta_l$. Inserting this result in the cost function and rearranging terms produces the estimating equation,

$$\ln(C/P_f) = \alpha + \beta \ln Q + \gamma[\tfrac{1}{2}(\ln Q)^2] + \delta_k \ln(P_k/P_f) + \delta_l \ln(P_l/P_f) + \varepsilon.$$

The purpose of the generalization was to produce a U-shaped average total cost curve. We are interested in the **efficient scale**, which is the output at which the cost curve reaches its minimum. That is the point at which $(\partial \ln C/\partial \ln Q)_{|Q=Q^*} = 1$ or $Q^* = \exp[(1 - \beta)/\gamma]$.

a. Data on 158 firms extracted from Christensen and Greene's study are given in Table F4.4. Using all 158 observations, compute the estimates of the parameters in the cost function and the estimate of the asymptotic covariance matrix.

b. Note that the cost function does not provide a direct estimate of δ_f. Compute this estimate from your regression results, and estimate the asymptotic standard error.

c. Compute an estimate of Q^* using your regression results and then form a confidence interval for the estimated efficient scale.

d. Examine the raw data and determine where in the sample the efficient scale lies. That is, determine how many firms in the sample have reached this scale, and whether, in your opinion, this scale is large in relation to the sizes of firms in the sample. Christensen and Greene approached this question by computing the proportion of total output in the sample that was produced by firms that had not yet reached efficient scale. (*Note:* There is some double counting in the data set—more than 20 of the largest "firms" in the sample we are using for this exercise are holding companies and power pools that are aggregates of other

firms in the sample. We will ignore that complication for the purpose of our numerical exercise.)

3. The Filipelli data mentioned in Footnote 11 are used to test the accuracy of computer programs in computing least squares coefficients. The 82 observations on (x,y) are given in Appendix Table F4.5. The regression computation involves regression of y on a constant and the first 10 powers of x. (The condition number for this 11-column data matrix is 0.3×10^{10}.) The correct least squares solutions are given on the NIST Website. Using the software you are familiar with, compute the regression using these data.

5

HYPOTHESIS TESTS AND MODEL SELECTION

5.1 INTRODUCTION

The linear regression model is used for three major purposes: estimation and prediction, which were the subjects of the previous chapter, and hypothesis testing. In this chapter, we examine some applications of hypothesis tests using the linear regression model. We begin with the methodological and statistical theory. Some of this theory was developed in Chapter 4 (including the idea of a pivotal statistic in Section 4.7.1) and in Appendix C.7. In Section 5.2, we will extend the methodology to hypothesis testing based on the regression model. After the theory is developed, Sections 5.3 through 5.5 will examine some applications in regression modeling. This development will be concerned with the implications of restrictions on the parameters of the model, such as whether a variable is relevant (i.e., has a nonzero coefficient) or whether the regression model itself is supported by the data (i.e., whether the data seem consistent with the hypothesis that all of the coefficients are zero). We will primarily be concerned with **linear restrictions** in this discussion. We will turn to **nonlinear restrictions** in Section 5.5. Section 5.6 considers some broader types of hypotheses, such as choosing between two competing models, for example, whether a linear or a loglinear model is better suited to the data. In each of the cases so far, the testing procedure attempts to resolve a competition between two theories for the data; in Sections 5.2 through 5.5 between a narrow model and a broader one and in Section 5.6, between two arguably equal models. Section 5.7 illustrates a particular **specification test**, which is essentially a test of a proposition such as *the model is correct* versus *the model is inadequate*. This test pits the theory of the model against *some other unstated theory*. Finally, Section 5.8 presents some general principles and elements of a strategy of model testing and selection.

5.2 HYPOTHESIS TESTING METHODOLOGY

We begin the analysis with the regression model as a statement of a proposition,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{5-1}$$

To consider a specific application, Examples 4.3 and 4.5 depicted the auction prices of paintings,

$$\ln \mathit{Price} = \beta_1 + \beta_2 \ln \mathit{Size} + \beta_3 \mathit{Aspect\ Ratio} + \varepsilon. \tag{5-2}$$

Some questions might be raised about the model in (5-2), fundamentally, about the variables. It seems natural that fine art enthusiasts would be concerned about aspect ratio, which is an element of the aesthetic quality of a painting. But the idea

that size should be an element of the price is counterintuitive, particularly weighed against the surprisingly small sizes of some of the world's most iconic paintings such as the *Mona Lisa* (30″ high and 21″ wide) or Dali's *Persistence of Memory* (only 9.5″ high and 13″ wide). A skeptic might question the presence of ln*Size* in the equation or, equivalently, the nonzero coefficient, β_2. To settle the issue, the relevant empirical question is whether the equation specified appears to be consistent with the data—that is, the observed sale prices of paintings. In order to proceed, the obvious approach for the analyst would be to fit the regression first and then examine the estimate of β_2. The test, at this point, is whether b_2 in the least squares regression is zero or not. Recognizing that the least squares slope is a random variable that will never be exactly zero even if β_2 really is, we would soften the question to be whether the sample estimate seems to be close enough to zero for us to conclude that its population counterpart is actually zero, that is, that the nonzero value we observe is nothing more than noise that is due to sampling variability. Remaining to be answered are questions including: How close to zero is close enough to reach this conclusion? What metric is to be used? How certain can we be that we have reached the right conclusion? (Not absolutely, of course.) How likely is it that our decision rule, whatever we choose, will lead us to the wrong conclusion? This section will formalize these ideas. After developing the methodology in detail, we will construct a number of numerical examples.

5.2.1 RESTRICTIONS AND HYPOTHESES

The approach we will take is to formulate a hypothesis as a restriction on a model. Thus, in the classical methodology considered here, the model is a general statement and a hypothesis is a proposition that narrows that statement. In the art example in (5-2), the narrower statement is (5-2) with the additional statement that $\beta_2 = 0$—without comment on β_1 or β_3. We define the **null hypothesis** as the statement that narrows the model and the **alternative hypothesis** as the broader one. In the example, the broader model allows the equation to contain both ln *Size* and *Aspect Ratio*—it admits the possibility that either coefficient might be zero but does not insist upon it. The null hypothesis insists that $\beta_2 = 0$ while it also makes no comment about β_1 or β_3. The formal notation used to frame this hypothesis would be

$$\begin{aligned} &\ln Price = \beta_1 + \beta_2 \ln Size + \beta_3 AspectRatio + \varepsilon, \\ &H_0: \beta_2 = 0, \\ &H_1: \beta_2 \neq 0. \end{aligned} \qquad \textbf{(5-3)}$$

Note that the null and alternative hypotheses, together, are exclusive and exhaustive. There is no third possibility; either one or the other of them is true, not both.

The analysis from this point on will be to measure the null hypothesis against the data. The data might persuade the econometrician to reject the null hypothesis. It would seem appropriate at that point to accept the alternative. However, in the interest of maintaining flexibility in the methodology, that is, an openness to new information, the appropriate conclusion here will be either to reject the null hypothesis or not to reject it. Not rejecting the null hypothesis is not equivalent to accepting it—though the language might suggest so. By accepting the null hypothesis, we would implicitly be closing off further investigation. Thus, the traditional, classical methodology leaves

open the possibility that further evidence might still change the conclusion. Our testing methodology will be constructed so as either to

Reject H_0: The data appear to be inconsistent with the hypothesis with a reasonable degree of certainty.
Do not reject H_0: The data appear to be consistent with the null hypothesis.

5.2.2 NESTED MODELS

The general approach to testing a hypothesis is to formulate a statistical model that contains the hypothesis as a restriction on its parameters. A theory is said to have **testable implications** if it implies some testable restrictions on the model. Consider, for example, a model of investment, I_t,

$$\ln I_t = \beta_1 + \beta_2 i_t + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \quad \textbf{(5-4)}$$

which states that investors are sensitive to nominal interest rates, i_t, the rate of inflation, Δp_t, (the log of) real output, $\ln Y_t$, and other factors that trend upward through time, embodied in the time trend, t. An alternative theory states that "investors care about real interest rates." The alternative model is

$$\ln I_t = \beta_1 + \beta_2(i_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t. \quad \textbf{(5-5)}$$

Although this new model does embody the theory, the equation still contains both nominal interest and inflation. The theory has no testable implication for our model. But, consider the stronger hypothesis, "investors care *only* about real interest rates." The resulting equation,

$$\ln I_t = \beta_1 + \beta_2(i_t - \Delta p_t) + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \quad \textbf{(5-6)}$$

is now restricted; in the context of (5-4), the implication is that $\beta_2 + \beta_3 = 0$. The stronger statement implies something specific about the parameters in the equation that may or may not be supported by the empirical evidence.

The description of testable implications in the preceding paragraph suggests (correctly) that testable restrictions will imply that only some of the possible models contained in the original specification will be valid; that is, consistent with the theory. In the example given earlier, (5-4) specifies a model in which there are five unrestricted parameters $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$. But (5-6) shows that only some values are consistent with the theory, that is, those for which $\beta_3 = -\beta_2$. This subset of values is contained within the unrestricted set. In this way, the models are said to be **nested**. Consider a different hypothesis, "investors do not care about inflation." In this case, the smaller set of coefficients is $(\beta_1, \beta_2, 0, \beta_4, \beta_5)$. Once again, the restrictions imply a valid **parameter space** that is "smaller" (has fewer dimensions) than the unrestricted one. The general result is that the hypothesis specified by the restricted model is contained within the unrestricted model.

Now, consider an alternative pair of models: Model_0: "Investors care only about inflation"; Model_1: "Investors care only about the nominal interest rate." In this case, the two parameter vectors are $(\beta_1, 0, \beta_3, \beta_4, \beta_5)$ by Model_0 and $(\beta_1, \beta_2, 0, \beta_4, \beta_5)$ by Model_1. The two specifications are both subsets of the unrestricted model, but neither model is obtained as a restriction on the other. They have the same number of parameters; they just contain different variables. These two models are **nonnested**. For the present, we are concerned only with **nested models**. **Nonnested models** are considered in Section 5.6.

5.2.3 TESTING PROCEDURES

In the example in (5-2), intuition suggests a testing approach based on measuring the data against the hypothesis. The essential methodology provides a reliable guide to testing hypotheses in the setting we are considering in this chapter. Broadly, the analyst follows the logic, "What type of data will lead me to reject the hypothesis?" Given the way the hypothesis is posed in Section 5.2.1, the question is equivalent to asking what sorts of data will support the model. The data that one can observe are divided into a **rejection region** and an **acceptance region**. The testing procedure will then be reduced to a simple up or down examination of the statistical evidence. Once it is determined what the rejection region is, if the observed data appear in that region, the null hypothesis is rejected. To see how this operates in practice, consider, once again, the hypothesis about size in the art price equation. Our test is of the hypothesis that β_2 equals zero. We will compute the least squares slope. We will decide in advance how far the estimate of β_2 must be from zero to lead to rejection of the null hypothesis. Once the rule is laid out, the test, itself, is mechanical. In particular, for this case, b_2 is far from zero if $b_2 > \beta_2^{0+}$ or $b_2 < \beta_2^{0-}$. If either case occurs, the hypothesis is rejected. The crucial element is that the rule is decided upon in advance.

5.2.4 SIZE, POWER, AND CONSISTENCY OF A TEST

Because the testing procedure is determined in advance and the estimated coefficient(s) in the regression are random, there are two ways the Neyman–Pearson method can make an error. To put this in a numerical context, the sample regression corresponding to (5-2) appears in Table 4.7. The estimate of the coefficient on ln *Area* is 1.31638 with an estimated standard error of 0.09205. Suppose the rule to be used to test is decided arbitrarily (at this point—we will formalize it shortly) to be: If b_2 is greater than $+1.0$ or less than -1.0, then we will reject the hypothesis that the coefficient is zero (and conclude that art buyers really do care about the sizes of paintings). So, based on this rule, we will, in fact, reject the hypothesis. However, because b_2 is a random variable, there are the following possible errors:

Type I error: $\beta_2 = 0$, but we reject the hypothesis that $\beta_2 = 0$.
The null hypothesis is incorrectly rejected.

Type II error: $\beta_2 \neq 0$, but we do not reject the hypothesis that $\beta_2 = 0$.
The null hypothesis is incorrectly retained.

The probability of a Type I error is called the **size of the test**. The size of a test is the probability that the test will incorrectly reject the null hypothesis. As will emerge later, the analyst determines this in advance. One minus the probability of a Type II error is called the **power of a test**. The power of a test is the probability that it will correctly reject a false null hypothesis. The power of a test depends on the alternative. It is not under the control of the analyst. To consider the example once again, we are going to reject the hypothesis if $|b_2| > 1$. If β_2 is actually 1.5, then based on the results we've seen, we are quite likely to find a value of b_2 that is greater than 1.0. On the other hand, if β_2 is only 0.3, then it does not seem likely that we will observe a sample value greater than 1.0. Thus, again, the power of a test depends on the actual parameters that underlie the data. The idea of power of a test relates to its ability to find what it is looking for.

A test procedure is **consistent** if its power goes to 1.0 as the sample size grows to infinity. This quality is easy to see, again, in the context of a single parameter, such as the one being considered here. Because least squares is consistent, it follows that as the sample size grows, we will be able to learn the exact value of β_2, so we will know if it is zero or not. Thus, for this example, it is clear that as the sample size grows, we will know with certainty if we should reject the hypothesis. For most of our work in this text, we can use the following guide: A testing procedure about the parameters in a model is consistent if it is based on a consistent estimator of those parameters. Nearly all our work in this book is based on consistent estimators. Save for the latter sections of this chapter, where our tests will be about the parameters in nested models, our tests will be consistent as well.

5.2.5 A METHODOLOGICAL DILEMMA: BAYESIAN VERSUS CLASSICAL TESTING

As we noted earlier, the testing methodology we will employ here is an all-or-nothing proposition. We will determine the testing rule(s) in advance, gather the data, and either reject or not reject the null hypothesis. There is no middle ground. This presents the researcher with two uncomfortable dilemmas. First, the testing outcome, that is, the sample data might be uncomfortably close to the boundary of the rejection region. Consider our example. If we have decided in advance to reject the null hypothesis if $b_2 > 1.00$, and the sample value is 0.9999, it will be difficult to resist the urge to reject the null hypothesis anyway, particularly if we entered the analysis with a strongly held belief that the null hypothesis is false. That is, intuition notwithstanding, we are unconvinced that art buyers really do care about size. Second, the methodology we have laid out here has no way of incorporating other studies. To continue our example, if we were the tenth team of analysts to study the art market, and the previous nine had decisively rejected the hypothesis that $\beta_2 = 0$, we will find it very difficult not to reject that hypothesis even if our evidence suggests, based on our testing procedure, that we should not.

This dilemma is built into the classical testing methodology. There is a middle ground. The Bayesian methodology that we will discuss in Chapter 16 does not face this dilemma because Bayesian analysts never reach a firm conclusion. They merely update their priors. Thus, in the first case noted, in which the observed data are close to the boundary of the rejection region, the analyst will merely be updating the prior with slightly less persuasive evidence than might be hoped for. But the methodology is comfortable with this. For the second instance, we have a case in which there is a wealth of prior evidence in favor of rejecting H_0. It will take a powerful tenth body of evidence to overturn the previous nine conclusions. The results of the tenth study (the posterior results) will incorporate not only the current evidence, but the wealth of prior data as well.

5.3 THREE APPROACHES TO TESTING HYPOTHESES

We will consider three approaches to testing hypotheses, Wald tests, fit based tests, and **Lagrange multiplier tests**. The hypothesis characterizes the population. If the hypothesis is correct, then the sample statistics should mimic that description. To continue our earlier example, if the hypothesis that states that a certain coefficient in a regression model equals zero is correct, then the least squares estimate of that coefficient should

be close to zero, at least within sampling variability. The tests will follow that logic as follows:

- **Wald tests:** The hypothesis states that $\boldsymbol{\beta}$ obeys some restriction(s), which we might state generally as $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{0}$. The least squares estimator, $\mathbf{b}$, is a consistent estimator of $\boldsymbol{\beta}$. If the hypothesis is correct, then $\mathbf{c}(\mathbf{b})$ should be close to zero. For the example of a single coefficient, if the hypothesis that β_k equals zero is correct, then b_k should be close to zero. The Wald test measures how close $\mathbf{c}(\mathbf{b})$ is to zero. The Wald test is based on estimation of the unrestricted model—the test measures how close the estimated unrestricted model is to the hypothesized restrictions.
- **Fit based tests:** We obtain the best possible fit—highest R^2 (or smallest sum of squared residuals)—by using least squares without imposing the restrictions. Imposing the restrictions will degrade the fit of the model to the data. For example, when we impose $\beta_k = 0$ by leaving x_k out of the model, we should expect R^2 to fall. The empirical device to use for testing the hypothesis will be a measure of how much R^2 falls when we impose the restrictions. This test procedure compares the fit of the restricted model to that of the unrestricted model.
- **Lagrange multiplier (LM) tests:** The LM test is based on the restricted model. The logic of the test is based on the general result that with the restrictions imposed, if those restrictions are incorrect, then we will be able to detect that failure in a measurable statistic. For the example of a single coefficient, β_k, in a multiple regression, the LM approach for the test will be based on the residuals from the regression that omits x_k. If β_k actually is not zero, then those residuals, say $e_{i(k)}$, which contain $\beta_k x_{ik}$, will be correlated with x_k. The test statistic will be based on that correlation. The test procedure is based on the estimates of the restricted model.

IMPORTANT ASSUMPTIONS

To develop the testing procedures in this section, we will begin by assuming homoscedastic, normally distributed disturbances—Assumptions A4 and A6 in Table 4.1. As we saw in Chapter 4, with these assumptions, we are able to obtain the exact distributions of the test statistics. In Section 5.4, we will develop an alternative set of results that allows us to proceed without Assumptions A4 and A6. It is useful to keep the distinction between the underlying theory of the testing procedures and the practical mechanics of inferences based on asymptotic approximations and robust covariance matrices. *Robust inference* is an improvement on the received procedures based on large-sample approximations to conventional statistics that allow conclusions to be drawn in a broader set of circumstances. For example, the conventional "*F* statistic" examined in Section 5.3.1B derives specifically from Assumptions A4 and A6. Cameron and Miller (2015, Sec. VII.A) in their survey of cluster robust inference (see Section 4.5.3) examine reconstruction of the *F* statistic in the broader context of **nonnormality** and clustered sampling.

The *general linear hypothesis* is a set of J restrictions on the linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The restrictions are written

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1K}\beta_K &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2K}\beta_K &= q_2 \\ &\cdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{JK}\beta_K &= q_J. \end{aligned} \tag{5-7}$$

The general case can be written in the matrix form,

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}. \tag{5-8}$$

Each row of $\mathbf{R}$ is the coefficients in one of the restrictions. Typically, $\mathbf{R}$ will have only one or a few rows and numerous zeros in each row. The hypothesis implied by the restrictions is written

$$H_0: \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}, H_1: \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}.$$

Some examples would be as follows:

1. One of the coefficients is zero, $\beta_j = 0$,

$$\mathbf{R} = [0 \quad 0 \quad \cdots \quad 1 \quad 0 \quad \cdots \quad 0]; \mathbf{q} = 0.$$

2. Two of the coefficients are equal, $\beta_k = \beta_j$,

$$\mathbf{R} = [0 \quad 0 \quad 1 \quad \cdots \quad -1 \quad \cdots \quad 0]; \mathbf{q} = 0.$$

3. A set of the coefficients sum to one, $\beta_2 + \beta_3 + \beta_4 = 1$,

$$\mathbf{R} = [0 \quad 1 \quad 1 \quad 1 \quad 0 \quad \cdots]; \mathbf{q} = 1.$$

4. A subset of the coefficients are all zero, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$,

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} = [\mathbf{I} \mid \mathbf{0}]; \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

5. Several linear restrictions, $\beta_2 + \beta_3 = 1$, $\beta_4 + \beta_6 = 0$, and $\beta_5 + \beta_6 = 0$,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}; \quad \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

6. All the coefficients in the model except the constant term are zero,

$$\mathbf{R} = [\mathbf{0} \mid \mathbf{I}_{K-1}]; \quad \mathbf{q} = \mathbf{0}.$$

The matrix $\mathbf{R}$ has K columns to be conformable with $\boldsymbol{\beta}$, J rows for a total of J restrictions, and *full row rank*, so J must be less than or equal to K. The rows of $\mathbf{R}$ must be linearly independent. Although it does not violate the condition, the case of $J = K$ must also be ruled out. If the K coefficients satisfy $J = K$ restrictions, then $\mathbf{R}$ is square and nonsingular and $\boldsymbol{\beta} = \mathbf{R}^{-1}\mathbf{q}$. There is no estimation or inference problem. The restriction $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ imposes J restrictions on K otherwise free parameters. Hence, with the restrictions imposed, there are, in principle, only $K - J$ free parameters remaining.

We will want to extend the methods to nonlinear restrictions. In example 5.6 below, the hypothesis takes the form $H_0: \beta_j/\beta_k = \beta_l/\beta_m$. The **general nonlinear hypothesis** involves a set of J possibly nonlinear restrictions,

$$\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}, \tag{5-9}$$

where $\mathbf{c}(\boldsymbol{\beta})$ is a set of J nonlinear functions of $\boldsymbol{\beta}$. The linear hypothesis is a special case. The counterpart to our requirements for the linear case are that, once again, J be strictly less than K, and the matrix of derivatives,

$$\mathbf{G}(\boldsymbol{\beta}) = \partial \mathbf{c}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}', \tag{5-10}$$

have full row rank. This means that the restrictions are **functionally independent**. In the linear case, $\mathbf{G}(\boldsymbol{\beta})$ is the matrix of constants, $\mathbf{R}$, that we saw earlier and functional independence is equivalent to linear independence. We will consider nonlinear restrictions in detail in Section 5.5. For the present, we will restrict attention to the general linear hypothesis.

5.3.1 WALD TESTS BASED ON THE DISTANCE MEASURE

The **Wald test** is the most commonly used procedure. It is often called a *significance test*. The operating principle of the procedure is to fit the regression without the restrictions, and then assess whether the results appear, within sampling variability, to agree with the hypothesis.

5.3.1.a Testing a Hypothesis about a Coefficient

The simplest case is a test of the value of a single coefficient. Consider, once again, the art market example in Section 5.2. The null hypothesis is

$$H_0: \beta_2 = \beta_2^0,$$

where β_2^0 is the hypothesized value of the coefficient, in this case, zero. The **Wald distance** of a coefficient estimate from a hypothesized value is the distance measured in standard deviation units. For this case, the distance of b_k from β_k^0 would be

$$W_k = \frac{b_k - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}. \tag{5-11}$$

As we saw in (4-45), W_k (which we called z_k before) has a standard normal distribution assuming that $E[b_k] = \beta_k^0$. Note that if $E[b_k]$ is not equal to β_k^0, then W_k still has a normal distribution, but the mean is not zero. In particular, if $E[b_k]$ is β_k^1 which is different from β_k^0, then

$$E\{W_k \,|\, E[b_k] = \beta_k^1\} = \frac{\beta_k^1 - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}. \tag{5-12}$$

(For example, if the hypothesis is that $\beta_k = \beta_k^0 = 0$, and β_k does not equal zero, then the expected value of $W_k = b_k/\sqrt{\sigma^2 S^{kk}}$ will equal $\beta_k^1/\sqrt{\sigma^2 S^{kk}}$, which is not zero.) For purposes of using W_k to test the hypothesis, our interpretation is that if β_k does equal β_k^0, then b_k will be close to β_k^0, with the distance measured in standard error units. Therefore, the logic of the test, to this point, will be to conclude that H_0 is incorrect—should be rejected—if W_k is "large" in absolute value.

Before we determine a benchmark for large, we note that the Wald measure suggested here is not usable because σ^2 is not known. It is estimated by s^2. Once again, invoking our results from Chapter 4, if we compute W_k using the sample estimate of σ^2, we obtain

$$t_k = \frac{b_k - \beta_k^0}{\sqrt{s^2 S^{kk}}}. \tag{5-13}$$

Assuming that β_k does indeed equal β_k^0, that is, "under the assumption of the null hypothesis," t_k has a t distribution with $n - K$ degrees of freedom. [See (4-47).] We can now construct the testing procedure. The test is carried out by determining in advance the desired confidence with which we would like to draw the conclusion—the standard value is 95%. Based on (5-13), we can say that

$$\text{Prob}\{-t^*_{(1-\alpha/2),[n-K]} < t_k < +t^*_{(1-\alpha/2),[n-K]}\},$$

where $t^*_{(1-\alpha/2),[n-K]}$ is the appropriate critical value from the t table. By this construction, if the null hypothesis is true, then finding a sample value of t_k that falls outside this range is unlikely. The test procedure states that it is so unlikely that we would conclude that it could not happen if the hypothesis were correct, so the hypothesis must be incorrect.

A common test is the hypothesis that a parameter equals zero—equivalently, this is a test of the relevance of a variable in the regression. To construct the test statistic, we set β_k^0 to zero in (5-13) to obtain the standard t ratio, $t_k = b_k/s_{bk}$. This statistic is reported in the regression results in several of our earlier examples, such as Example 4.10 where the regression results for the model in (5-2) appear. This statistic is usually labeled the ***t* ratio** for the estimator b_k. If $|b_k|/s_{bk} > t_{(1-\alpha/2),[n-K]}$, where $t_{(1-\alpha/2),[n-K]}$ is the $100(1 - \alpha/2)$ % critical value from the t distribution with $(n - K)$ degrees of freedom, then the null hypothesis that the coefficient is zero is rejected and the coefficient (actually, the associated variable) is said to be statistically significant. The value of 1.96, which would apply for the 95% significance level in a large sample, is often used as a benchmark value when a table of critical values is not immediately available. The t ratio for the test of the hypothesis that a coefficient equals zero is a standard part of the regression output of most computer programs.

Another view of the testing procedure is useful. Also based on (4-48) and (5-13), we formed a confidence interval for β_k as $b_k \pm t \times s_k$. We may view this interval as the set of plausible values of β_k with a confidence level of $100(1-\alpha)\%$, where we choose α, typically 5%. The confidence interval provides a convenient tool for testing a hypothesis about β_k, because we may simply ask whether the hypothesized value, β_k^0, is contained in this range of plausible values. The complement of the confidence interval is the rejection region for this test.

Example 5.1 Art Appreciation

Regression results for the model in (5-3) based on a sample of 430 sales of Monet paintings appear in Table 4.7 in Example 4.9. The estimated coefficient on ln *Area* is 1.33372 with an estimated standard error of 0.09205. The distance of the estimated coefficient from zero is 1.31638/0.092 − 5 = 14.16. Because this is far larger than the 95% critical value of 1.96, we reject the hypothesis that β_2 equals zero; evidently buyers of Monet paintings do care about size. In contrast, the coefficient on *Aspect Ratio* is −0.09623 with an estimated standard error of 0.16706, so the associated t ratio for the test of H_0: $\beta_3 = 0$ is only −0.61. Given that this is well under 1.96, we conclude that art buyers (of Monet paintings) do not care about the

aspect ratio of the paintings. As a final consideration, we examine another (equally bemusing) hypothesis, whether auction prices are inelastic H_0: $\beta_2 \leq 1$ or elastic H_1: $\beta_2 > 1$ with respect to area. This is a **one-sided test**. Using our guideline for formulating the test, we will reject the null hypothesis if the estimated coefficient is sufficiently larger than 1.0. To maintain a test of size 0.05, we will then place all of the area for the rejection region to the right of 1.0; the critical value from the table is 1.645. The test statistic is (1.31638 − 1)/0.09205 = 3.437 > 1.645. Thus, we will reject this null hypothesis as well.

Example 5.2 Earnings Equation

Appendix Table F5.1 contains the 753 observations used in Mroz's (1987) study of the labor supply behavior of married women. Of the 753 individuals in the sample, 428 were participants in the formal labor market. For these individuals, we will fit a semilog earnings equation of the form suggested in Example 2.2:

$$\ln earnings = \beta_1 + \beta_2 age + \beta_3 age^2 + \beta_4 education + \beta_5 kids + \varepsilon,$$

where *earnings* is *hourly wage* times *hours worked*, *education* is measured in years of schooling, and *kids* is a binary variable which equals one if there are children under 18 in the household. (See the data description in Appendix F for details.) Regression results are shown in Table 5.1. There are 428 observations and 5 parameters, so the *t* statistics have (428−5) = 423 degrees of freedom. For 95% significance levels, the standard normal value of 1.96 is appropriate when the degrees of freedom are this large. By this measure, all variables are statistically significant and signs are consistent with expectations. It will be interesting to investigate whether the effect of *kids* is on the wage or hours, or both. We interpret the schooling variable to imply that an additional year of schooling is associated with a 6.7% increase in earnings. The quadratic age profile suggests that for a given education level and family size, earnings rise to a peak at $-b_2/(2b_3)$ which is about 43 years of age, at which point they begin to decline. Some points to note: (1) Our selection of only those individuals who had positive hours worked is not an innocent sample selection mechanism. Because individuals

TABLE 5.1 Regression Results for an Earnings Equation

Sum of squared residuals:	599.4582
R^2 based on 428 observations	0.040944
Standard error of the regression:	1.19044

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age^2	−0.0023147	0.00098688	−2.345
Education	0.067472	0.025248	2.672
Kids	−0.35119	0.14753	−2.380

Estimated Covariance Matrix for b(e−n = times 10^{-n})

	Constant	*Age*	Age^2	*Education*	*Kids*
Constant	3.12381				
Age	−0.13409	0.0070325			
Age^2	0.0016617	−8.23237e−5	9.73928e−7		
Education	−0.0092609	5.08549e−5	−4.96761e−7	0.00063729	
Kids	0.026749	−0.0026412	3.84102e−5	−5.46193e−5	0.021766

chose whether or not to be in the labor force, it is likely (almost certain) that earnings potential was a significant factor, along with some other aspects we will consider in Chapter 19. (2) The earnings equation is a mixture of a labor supply equation—hours worked by the individual—and a labor demand outcome—the wage is, presumably, an accepted offer. As such, it is unclear what the precise nature of this equation is. Presumably, it is a hash of the equations of an elaborate structural equation system. (See Example 10.1 for discussion.)

5.3.1.b The *F* Statistic

We now consider testing a set of J linear restrictions stated in the null hypothesis,

$$H_0: \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0},$$

against the alternative hypothesis,

$$H_1: \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}.$$

Given the least squares estimator $\mathbf{b}$, our interest centers on the **discrepancy vector** $\mathbf{Rb} - \mathbf{q} = \mathbf{m}$. It is unlikely that $\mathbf{m}$ will be exactly $\mathbf{0}$. The statistical question is whether the deviation of $\mathbf{m}$ from $\mathbf{0}$ can be attributed to sampling variability or whether it is significant. Because $\mathbf{b}$ is normally distributed [see Section 4.3.6] and $\mathbf{m}$ is a linear function of $\mathbf{b}$, $\mathbf{m}$ is also normally distributed. If the null hypothesis is true, then $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ and $\mathbf{m}$ has mean vector

$$E[\mathbf{m}|\mathbf{X}] = \mathbf{R}E[\mathbf{b}|\mathbf{X}] - \mathbf{q} = \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$$

and covariance matrix

$$\text{Var}[\mathbf{m}|\mathbf{X}] = \text{Var}[\mathbf{Rb} - \mathbf{q}|\mathbf{X}] = \mathbf{R}\{\text{Var}[\mathbf{b}|\mathbf{X}]\}\mathbf{R}' = \mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'.$$

We can base a test of H_0 on the **Wald criterion**. Conditioned on $\mathbf{X}$, we find:

$$\begin{aligned} W &= \mathbf{m}'\{\text{Var}[\mathbf{m}|\mathbf{X}]\}^{-1}\mathbf{m} \\ &= (\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\}^{-1}(\mathbf{Rb} - \mathbf{q}) \\ &\sim \chi^2[J]. \end{aligned} \tag{5-14}$$

The statistic W has a chi-squared distribution with J degrees of freedom if the hypothesis is correct.[1] Intuitively, the larger $\mathbf{m}$ is—that is, the worse the failure of least squares to satisfy the restrictions—the larger the chi-squared statistic. Therefore, a large chi-squared value will weigh against the hypothesis.

The chi-squared statistic in (5-14) is not usable because of the unknown σ^2. By using s^2 instead of σ^2 and dividing the result by J, we obtain a usable F statistic with J and $n - K$ degrees of freedom,

$$F = \frac{W}{J}\frac{\sigma^2}{s^2} = (\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]R'\}^{-1}(\mathbf{Rb} - \mathbf{q})/J. \tag{5-15}$$

The F statistic for testing the general linear hypothesis is simply the feasible Wald statistic, divided by J:

$$F[J, n - K|\mathbf{X}] = \frac{(\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\}^{-1}(\mathbf{Rb} - \mathbf{q})}{J}. \tag{5-16}$$

[1]This calculation is an application of the full rank quadratic form of Section B.11.6. Note that although the chi-squared distribution is conditioned on $\mathbf{X}$, it is also free of $\mathbf{X}$.

For testing one linear restriction of the form

$$H_0: r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = \mathbf{r}'\boldsymbol{\beta} = q,$$

(usually, some of the r's will be zero), the F statistic is

$$F[1, n - K] = \frac{(\Sigma_j r_j b_j - q)^2}{\Sigma_j \Sigma_k r_j r_k \text{ Est. Cov}[b_j, b_k]}.$$

If the hypothesis is that the jth coefficient is equal to a particular value, then $\mathbf{R}$ has a single row with a one in the jth position and zeros elsewhere, $\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'$ is the jth diagonal element of the estimated covariance matrix, and $\mathbf{Rb} - \mathbf{q}$ is $(b_j - q)$. The F statistic is then

$$F[1, n - K] = \frac{(b_j - q)^2}{\text{Est. Var}[b_j]}.$$

Consider an alternative approach. The sample estimate of $\mathbf{r}'\boldsymbol{\beta}$ is

$$r_1b_1 + r_2b_2 + \cdots + r_Kb_K = \mathbf{r}'\mathbf{b} = \hat{q}.$$

If $\hat{q}$ differs significantly from q, then we conclude that the sample data are not consistent with the hypothesis. It is natural to base the test on

$$t = \frac{\hat{q} - q}{\text{se}(\hat{q})}. \tag{5-17}$$

We require an estimate of the standard error of $\hat{q}$. Because $\hat{q}$ is a linear function of $\mathbf{b}$ and we have an estimate of the covariance matrix of $\mathbf{b}$, $s^2(\mathbf{X}'\mathbf{X})^{-1}$, we can estimate the variance of $\hat{q}$ with

$$\text{Est. Var}[\hat{q}\,|\,\mathbf{X}] = \mathbf{r}'[s^2\,(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{r}.$$

The denominator of t is the square root of this quantity. In words, t is the distance in standard error units between the hypothesized function of the true coefficients and the same function of the estimates of them. If the hypothesis is true, then the estimates should reflect that, at least within the range of sampling variability. Thus, if the absolute value of the preceding t ratio is larger than the appropriate critical value, then doubt is cast on the hypothesis.

There is a useful relationship between the statistics in (5-16) and (5-17). We can write the square of the t statistic as

$$t^2 = \frac{(\hat{q} - q)^2}{\text{Var}(\hat{q} - q\,|\,\mathbf{X})} = \frac{(\mathbf{r}'\mathbf{b} - q)\{\mathbf{r}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{r}\}^{-1}(\mathbf{r}'\mathbf{b} - q)}{1}. \tag{5-18}$$

It follows, therefore, that for testing a single restriction, the t statistic is the square root of the F statistic that would be used to test that hypothesis. (The sign of the t statistic is lost, of course.)

Example 5.3 Restricted Investment Equation

Section 5.2.2 suggested a theory about the behavior of investors: They care only about real interest rates. If investors were only interested in the real rate of interest, then equal increases in interest rates and the rate of inflation would have no independent effect on investment. The null hypothesis is

$$H_0: \beta_2 + \beta_3 = 0.$$

Estimates of the parameters of equations (5-4) and (5-6) using 1950I to 2000IV quarterly data on real investment, real GDP, an interest rate (the 90-day T-bill rate), and inflation measured by the change in the log of the CPI given in Appendix Table F5.2 are presented in Table 5.2. (One observation is lost in computing the change in the CPI.)

To form the appropriate test statistic, we require the standard error of $\hat{q} = b_2 + b_3$, which is

$$\text{se}(\hat{q}) = [0.00319^2 + 0.00234^2 + 2(-3.718 \times 10^{-6})]^{1/2} = 0.002866.$$

The t ratio for the test is therefore

$$t = \frac{-0.00860 + 0.00331}{0.002866} = -1.846.$$

Using the 95% critical value from $t[203–5] = 1.96$ (the standard normal value), we conclude that the sum of the two coefficients is not significantly different from zero, so the hypothesis should not be rejected.

There will usually be more than one way to formulate a restriction in a regression model. One convenient way to parameterize a constraint is to set it up in such a way that the standard test statistics produced by the regression can be used without further computation to test the hypothesis. In the preceding example, we could write the regression model as specified in (5-5). Then an equivalent way to test H_0 would be to fit the investment equation with both the real interest rate and the rate of inflation as regressors and to test our theory by simply testing the hypothesis that β_3 equals zero, using the standard t statistic that is routinely computed. When the regression is computed this way, $b_3 = -0.00529$ and the estimated standard error is 0.00287, resulting in a t ratio of -1.844(!). (Exercise: Suppose that the nominal interest rate, rather than the rate of inflation, were included as the extra regressor. What do you think the coefficient and its standard error would be?)

Finally, consider a test of the joint hypothesis,

$$\begin{aligned} \beta_2 + \beta_3 &= 0 \quad \text{(investors consider the real interest rate),} \\ \beta_4 &= 1 \quad \text{(the marginal propensity to invest equals 1),} \\ \beta_5 &= 0 \quad \text{(there is no time trend).} \end{aligned}$$

Then,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}; \quad \mathbf{q} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}; \mathbf{Rb} - \mathbf{q} = \begin{bmatrix} -0.0053 \\ -0.9302 \\ -0.0057 \end{bmatrix}.$$

TABLE 5.2 Estimated Investment Equations (Estimated standard errors in parentheses)

	β_1	β_2	β_3	β_4	β_5
Model (5-4)	−9.135	−0.00860	0.00331	1.930	−0.00566
	(1.366)	(0.00319)	(0.00234)	(0.183)	(0.00149)
	$s = 0.08618$, $R^2 = 0.979753$, $\mathbf{e}'\mathbf{e} = 1.47052$,				
	Est. Cov$[b_2, b_3] = -3.718\text{e}{-6}$				
Model (5-6)	−7.907	−0.00443	0.00443	1.764	−0.00440
	(1.201)	(0.00227)	(0.00227)	(0.161)	(0.00133)
	$s = 0.8670$, $R^2 = 0.979405$, $\mathbf{e}'\mathbf{e} = 1.49578$				

Inserting these values in the formula for the F statistic yields $F = 109.84$. The 5% critical value for $F[3, 198]$ is 2.65. We conclude, therefore, that the data are not consistent with this hypothesis. The result gives no indication as to which of the restrictions is most influential in the rejection of the hypothesis. If the three restrictions are tested one at a time, the t statistics in (5-17) are −1.844, 5.076, and −3.803. Based on the individual test statistics, therefore, we would expect both the second and third hypotheses to be rejected.

5.3.2 TESTS BASED ON THE FIT OF THE REGRESSION

A different approach to hypothesis testing focuses on the fit of the regression. Recall that the least squares coefficient vector $\mathbf{b}$ was chosen to minimize the sum of squared deviations, $\mathbf{e}'\mathbf{e}$. Because R^2 equals $1 - \mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ and $\mathbf{y}'\mathbf{M}^0\mathbf{y}$ is a constant that does not involve $\mathbf{b}$, it follows that if the model contains a constant term, $\mathbf{b}$ is chosen to maximize R^2. One might ask whether choosing some other value for the slopes of the regression leads to a significant loss of fit. For example, in the investment equation (5-4), one might be interested in whether assuming the hypothesis (that investors care only about real interest rates) leads to a substantially worse fit than leaving the model unrestricted. To develop the test statistic, we first examine the computation of the least squares estimator subject to a set of restrictions. We will then construct a test statistic that is based on comparing the R^2's from the two regressions.

5.3.2.a The Restricted Least Squares Estimator

Suppose that we explicitly impose the restrictions of the general linear hypothesis in the regression. The restricted least squares estimator is obtained as the solution to

$$\text{Minimize}_{\mathbf{b}_0} \quad S(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) \quad \text{subject to } \mathbf{R}\mathbf{b}_0 = \mathbf{q}.$$

A Lagrangean function for this problem can be written

$$L^*(\mathbf{b}_0, \boldsymbol{\lambda}) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) + 2\boldsymbol{\lambda}'(\mathbf{R}\mathbf{b}_0 - \mathbf{q}).^2 \tag{5-19}$$

The solutions $\mathbf{b}_*$ and $\boldsymbol{\lambda}_*$ will satisfy the necessary conditions

$$\frac{\partial L^*}{\partial \mathbf{b}_*} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}_*) + 2\mathbf{R}'\boldsymbol{\lambda}_* = \mathbf{0},$$
$$\frac{\partial L^*}{\partial \boldsymbol{\lambda}_*} = 2(\mathbf{R}\mathbf{b}_* - \mathbf{q}) = \mathbf{0}. \tag{5-20}$$

Dividing through by 2 and expanding terms produces the partitioned matrix equation

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda}_* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{q} \end{bmatrix}. \tag{5-21}$$

Assuming that the partitioned matrix in brackets is nonsingular, the restricted least squares estimator is the upper part of the solution

$$\begin{bmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda}_* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{q} \end{bmatrix} = \mathbf{A}^{-1}\mathbf{d}. \tag{5-22}$$

[2]Because λ is not restricted, we can formulate the constraints in terms of 2λ. The convenience of the scaling shows up in (5-20).

If, in addition, $\mathbf{X}'\mathbf{X}$ is nonsingular, then explicit solutions for $\mathbf{b}_*$ and $\boldsymbol{\lambda}_*$ may be obtained by using the formula for the partitioned inverse (A-74),[3]

$$\begin{aligned}\mathbf{b}_* &= \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}) \\ &= \mathbf{b} - \mathbf{C}\mathbf{m}, \\ \lambda_* &= [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}).\end{aligned} \tag{5-23}$$

Greene and Seaks (1991) show that the covariance matrix for $\mathbf{b}_*$ is simply σ^2 times the upper left block of $\mathbf{A}^{-1}$. If $\mathbf{X}'\mathbf{X}$ is nonsingular, an explicit formulation may be obtained:

$$\text{Var}[\mathbf{b}_*|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}. \tag{5-24}$$

Thus,

$$\text{Var}[\mathbf{b}_*|\mathbf{X}] = \text{Var}[\mathbf{b}|\mathbf{X}]\text{—a nonnegative definite matrix.}$$

One way to interpret this reduction in variance is as the value of the information contained in the restrictions. A useful point to note is that $\text{Var}[\mathbf{b}_*|\mathbf{X}]$ is smaller than $\text{Var}[\mathbf{b}|\mathbf{X}]$ even if the restrictions are incorrect.

Note that the explicit solution for $\boldsymbol{\lambda}_*$ involves the discrepancy vector $\mathbf{R}\mathbf{b} - \mathbf{q}$. If the unrestricted least squares estimator satisfies the restriction, the Lagrangean multipliers will equal zero and $\mathbf{b}_*$ will equal $\mathbf{b}$. Of course, this is unlikely. In general, the constrained solution, $\mathbf{b}_*$, is equal to the unconstrained solution, $\mathbf{b}$, minus a term that accounts for the failure of the unrestricted solution to satisfy the constraints.

5.3.2.b The Loss of Fit from Restricted Least Squares

To develop a test based on the restricted least squares estimator, we consider a single coefficient first and then turn to the general case of J linear restrictions. Consider the change in the fit of a multiple regression when a variable z is added to a model that already contains $K - 1$ variables, $\mathbf{x}$. We showed in Section 3.5 (Theorem 3.6) (3-29) that the effect on the fit would be given by

$$R^2_{\mathbf{Xz}} = R^2_{\mathbf{X}} + (1 - R^2_{\mathbf{X}})r^{*2}_{yz}, \tag{5-25}$$

where $R^2_{\mathbf{Xz}}$ is the new R^2 after z is added, $R^2_{\mathbf{X}}$ is the original R^2, and r^*_{yz} is the partial correlation between y and z, controlling for $\mathbf{x}$. So, as we knew, the fit improves (or, at the least, does not deteriorate). In deriving the partial correlation coefficient between y and z in (3-22) we obtained the convenient result

$$r^{*2}_{yz} = \frac{t_z^2}{t_z^2 + (n - K)}, \tag{5-26}$$

where t_z^2 is the square of the t ratio for testing the hypothesis that the coefficient on z is zero in the *multiple* regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{z}$. If we solve (5-25) for r^{*2}_{yz} and (5-26) for t_z^2 and then insert the first solution in the second, then we obtain the result

$$t_z^2 = \frac{(R^2_{\mathbf{Xz}} - R^2_{\mathbf{X}})/1}{(1 - R^2_{\mathbf{Xz}})/(n - K)}. \tag{5-27}$$

[3]The general solution given for $\mathbf{d}_*$ may be usable even if $\mathbf{X}'\mathbf{X}$ is singular. This formulation and a number of related results are given in Greene and Seaks (1991).

We saw at the end of Section 5.4.2 that for a single restriction, such as $\beta_z = 0$,

$$F[1, n - K] = t^2[n - K],$$

which gives us our result. That is, in (5-27), we see that the squared t statistic (i.e., the F statistic) can be computed using the change in the R^2. By interpreting the preceding as the result of *removing* z from the regression, we see that we have proved a result for the case of testing whether a single slope is zero. But the preceding result is general. The test statistic for a single linear restriction is the square of the t ratio in (5-17). By this construction, we see that for a single restriction, F is a measure of the loss of fit that results from imposing that restriction. To obtain this result, we will proceed to the general case of J linear restrictions, which will include one restriction as a special case.

The fit of the restricted least squares coefficients cannot be better than that of the unrestricted solution. Let $\mathbf{e}_*$ equal $\mathbf{y} - \mathbf{X}\mathbf{b}_*$. Then, using a familiar device,

$$\mathbf{e}_* = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}).$$

The new sum of squared deviations is

$$\mathbf{e}_*'\mathbf{e}_* = \mathbf{e}'\mathbf{e} + (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}) \geq \mathbf{e}'\mathbf{e}.$$

(The middle term in the expression involves $\mathbf{X}'\mathbf{e}$, which is zero.) The loss of fit is

$$\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e} = (\mathbf{R}\mathbf{b} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}). \tag{5-28}$$

This expression appears in the numerator of the F statistic in (5-7). Inserting the remaining parts, we obtain

$$F[J, n - K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n - K)}. \tag{5-29}$$

Finally, by dividing both numerator and denominator of F by $\Sigma_i(y_i - \bar{y})^2$, we obtain the general result:

$$F[J, n - K] = \frac{(R^2 - R_*^2)/J}{(1 - R^2)/(n - K)}. \tag{5-30}$$

This form has some intuitive appeal in that the difference in the fits of the two models is directly incorporated in the test statistic. As an example of this approach, consider the joint test that all the slopes in the model are zero. This is the overall F ratio that will be discussed in Section 5.3.2C, where $R_*^2 = 0$.

For imposing a set of **exclusion restrictions** such as $\beta_k = 0$ for one or more coefficients, the obvious approach is simply to omit the variables from the regression and base the test on the sums of squared residuals for the restricted and unrestricted regressions. The F statistic for testing the hypothesis that a subset, say $\boldsymbol{\beta}_2$, of the coefficients are all zero is constructed using $\mathbf{R} = (\mathbf{0}:\mathbf{I})$, $\mathbf{q} = \mathbf{0}$, and $J = K_2 =$ the number of elements in $\boldsymbol{\beta}_2$. The matrix $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is the $K_2 \times K_2$ lower right block of the full inverse matrix. Using our earlier results for partitioned inverses and the results of Section 3.3, we have $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}$ and $\mathbf{R}\mathbf{b} - \mathbf{q} = \mathbf{b}_2$. Inserting these in (5-28) gives the loss of fit that results when we drop a subset of the variables from the regression:

$$\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e} = \mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2.$$

The procedure for computing the appropriate F statistic amounts simply to comparing the sums of squared deviations from the *short* and *long* regressions, which we saw earlier.

5.3.2.c Testing the Significance of the Regression

A question that is usually of interest is whether the regression equation as a whole is significant. This test is a joint test of the hypotheses that *all* the coefficients except the constant term are zero. If all the slopes are zero, then the coefficient of determination, R^2, is zero as well, so we can base a test of this hypothesis on the value of R^2. The central result needed to carry out the test is given in (5-30). This is the special case with $R_*^2 = 0$, so the F statistic, which is usually reported with multiple regression results is

$$F[K - 1, n - K] = \frac{R^2/(K - 1)}{(1 - R^2)/(n - K)}.$$

If the hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ (the part of $\boldsymbol{\beta}$ not including the constant) is true and the disturbances are normally distributed, then this statistic has an F distribution with $K - 1$ and $n - K$ degrees of freedom. Large values of F give evidence against the validity of the hypothesis. Note that a large F is induced by a large value of R^2. The logic of the test is that the F statistic is a measure of the loss of fit (namely, all of R^2) that results when we impose the restriction that all the slopes are zero. If F is large, then the hypothesis is rejected.

Example 5.4F Test for the Earnings Equation

The F ratio for testing the hypothesis that the four slopes in the earnings equation in Example 5.2 are all zero is

$$F[4, 423] = \frac{0.040995/(5 - 1)}{(1 - 0.040995)/(428 - 5)} = 4.521,$$

which is larger than the 95% critical value of 2.39. We conclude that the data are inconsistent with the hypothesis that all the slopes in the earnings equation are zero. We might have expected the preceding result, given the substantial t ratios presented earlier. But this case need not always be true. Examples can be constructed in which the individual coefficients are statistically significant, while jointly they are not. This case can be regarded as pathological, but the opposite one, in which none of the coefficients is significantly different from zero while R^2 is highly significant, is relatively common. The problem is that the interaction among the variables may serve to obscure their individual contribution to the fit of the regression, whereas their joint effect may still be significant.

5.3.2.d Solving Out the Restrictions and a Caution about R^2

In principle, one can usually solve out the restrictions imposed by a linear hypothesis. Algebraically, we would begin by partitioning $\mathbf{R}$ into two groups of columns, one with J and one with $K - J$, so that the first set are linearly independent. (There are many ways to do so; any one will do for the present.) Then, with $\boldsymbol{\beta}$ likewise partitioned and its elements reordered in whatever way is needed, we may write

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{R}_1\boldsymbol{\beta}_1 + \mathbf{R}_2\boldsymbol{\beta}_2 = \mathbf{q}.$$

If the J columns of $\mathbf{R}_1$ are linearly independent, then

$$\boldsymbol{\beta}_1 = \mathbf{R}_1^{-1}[\mathbf{q} - \mathbf{R}_2\boldsymbol{\beta}_2].$$

This suggests that one might estimate the restricted model directly using a transformed equation, rather than use the rather cumbersome restricted estimator shown in (5-23). A simple example illustrates. Consider imposing constant returns to scale on a two input production function,

$$\ln y = \beta_1 + \beta_2 \ln x_1 + \beta_3 \ln x_2 + \varepsilon.$$

The hypothesis of linear homogeneity is $\beta_2 + \beta_3 = 1$ or $\beta_3 = 1 - \beta_2$. Simply building the restriction into the model produces

$$\ln y = \beta_1 + \beta_2 \ln x_1 + (1 - \beta_2) \ln x_2 + \varepsilon$$

or

$$\ln y = \ln x_2 + \beta_1 + \beta_2(\ln x_1 - \ln x_2) + \varepsilon.$$

One can obtain the restricted least squares estimates by linear regression of $(\ln y - \ln x_2)$ on a constant and $(\ln x_1 - \ln x_2)$. However, the test statistic for the hypothesis cannot be computed using the familiar result in (5-30), because the denominators in the two R^2's are different. The statistic in (5-30) could even be negative. The appropriate approach would be to use the equivalent, but appropriate computation based on the sum of squared residuals in (5-29). The general result from this example is that one must be careful in using (5-30) that the dependent variable in the two regressions must be the same.

5.3.3 LAGRANGE MULTIPLIER TESTS

The vector of Lagrange multipliers in the solution for $\mathbf{b}_*$ and $\boldsymbol{\lambda}_*$ in (5-23) is $[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$, that is, a multiple of the least squares discrepancy vector. In principle, a test of the hypothesis that $\boldsymbol{\lambda}_*$ equals zero should be equivalent to a test of the null hypothesis; $\boldsymbol{\lambda}_*$ differs from zero because the restrictions do not hold in the data—that is, because $\mathbf{Rb}$ is not equal to $\mathbf{q}$. A Wald test of the hypothesis that $\boldsymbol{\lambda}_* = \mathbf{0}$ is derived in Section 14.9.1. The chi-squared statistic is computed as

$$W_{LM} = (\mathbf{Rb} - \mathbf{q})' \, [\mathbf{R}\{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}).$$

A feasible version of the statistic is obtained by using s^2 (based on the restricted regression) in place of the unknown σ^2. The large-sample distribution of this Wald statistic would be chi-squared with J degrees of freedom. There is a remarkably simple way to carry out this test. The chi-squared statistic, in this case with J degrees of freedom, can be computed as nR^2 in the regression of $\mathbf{e}_* = \mathbf{y} - \mathbf{Xb}_*$ (the residuals in the constrained regression) on the full set of independent variables as they would appear in the unconstrained regression. For example, for testing the restriction $\boldsymbol{\beta}_2 = 0$ in the model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, we would (1) regress $\mathbf{y}$ on $\mathbf{X}_1$ alone and compute residuals $\mathbf{e}_*$, then (2) compute W_{LM} by regressing $\mathbf{e}_*$ on $(\mathbf{X}_1, \mathbf{X}_2)$ and computing nR^2.

Example 5.5 Production Functions

The data in Appendix Table F5.3 have been used in several studies of production functions.[4] Least squares regression of log output (value added) on a constant and the logs of labor and capital produce the estimates of a Cobb–Douglas production function shown in Table 5.3.

[4]The data are statewide observations on SIC 33, the primary metals industry. They were originally constructed by Hildebrand and Liu (1957) and have subsequently been used by a number of authors, notably Aigner, Lovell, and Schmidt (1977). The 28th data point used in the original study is incomplete; we have used only the remaining 27.

We will construct several hypothesis tests based on these results. A generalization of the Cobb–Douglas model is the *translog* model,[5] which is

$$\ln Y = \beta_1 + \beta_2 \ln L + \beta_3 \ln K + \beta_4(\tfrac{1}{2}\ln^2 L) + \beta_5(\tfrac{1}{2}\ln^2 K) + \beta_6 \ln L \ln K + \varepsilon.$$

As we shall analyze further in Chapter 10, this model differs from the Cobb–Douglas model in that it relaxes the Cobb–Douglas's assumption of a unitary elasticity of substitution. The Cobb–Douglas model is obtained by the restriction $\beta_4 = \beta_5 = \beta_6 = 0$. The results for the two regressions are given in Table 5.3. The *F* statistic for the hypothesis of a Cobb–Douglas model is

$$F[3, 21] = \frac{(0.85163 - 0.67993)/3}{0.67993/21} = 1.768.$$

The critical value from the *F* table is 3.07, so we would not reject the hypothesis that a Cobb–Douglas model is appropriate.

TABLE 5.3 Estimated Production Function

			Translog			***Cobb–Douglas***
Sum of squared residuals			0.67993			0.85163
Standard error of regression			0.17994			0.18837
R-squared			0.95486			0.94346
Model $F[K-1, n-K]$			74.326			200.239
Adjusted *R*-squared			0.94411			0.93875
Number of observations			27			27
Variable	***Coefficient***	***Std.Error***	***t Ratio***	***Coefficient***	***Std.Error***	***t Ratio***
Constant	0.94420	2.911	0.324	1.171	0.3268	3.582
$\ln L$	3.61364	1.548	2.334	0.6030	0.1260	4.787
$\ln K$	−1.89311	1.016	−1.863	0.3757	0.0853	4.402
$\frac{1}{2}\ln^2 L$	−0.96405	0.7074	−1.363			
$\frac{1}{2}\ln^2 K$	0.08529	0.2926	0.291			
$\ln L \times \ln K$	0.31239	0.4389	0.712			

Estimated Covariance Matrix for Translog (Cobb–Douglas) Coefficient Estimates

	Constant	*ln L*	*ln K*	$\frac{1}{2}\ln^2 L$	$\frac{1}{2}\ln^2 K$	*ln L ln K*
Constant	8.472					
	(0.1068)					
$\ln L$	−2.388	2.397				
	(−0.01984)	(0.01586)				
$\ln K$	−0.3313	−1.231	1.033			
	(0.001189)	(−0.00961)	(0.00728)			
$\frac{1}{2}\ln^2 L$	−0.08760	−0.6658	0.5231	0.5004		
$\frac{1}{2}\ln^2 K$	−0.2332	0.03477	0.02637	0.1467	0.08562	
$\ln L \ln K$	0.3635	0.1831	−0.2255	−0.2880	−0.1160	0.1927

[5]Berndt and Christensen (1973). See Example 2.4 and Section 10.3.2 for discussion.

The hypothesis of constant returns to scale is often tested in studies of production. This hypothesis is equivalent to a restriction that the two coefficients of the Cobb–Douglas production function sum to 1. For the preceding data,

$$F[1, 24] = \frac{(0.6030 + 0.3757 - 1)^2}{0.01586 + 0.00728 - 2(0.00961)} = 0.1157,$$

which is substantially less than the 95% critical value of 4.26. We would not reject the hypothesis; the data are consistent with the hypothesis of constant returns to scale. The equivalent test for the translog model would be $\beta_2 + \beta_3 = 1$ and $\beta_4 + \beta_5 + 2\beta_6 = 0$. The F statistic with 2 and 21 degrees of freedom is 1.8991, which is less than the critical value of 3.47. Once again, the hypothesis is not rejected.

In most cases encountered in practice, it is possible to incorporate the restrictions of a hypothesis directly on the regression and estimate a restricted model.[6] For example, to impose the constraint $\beta_2 = 1$ on the Cobb–Douglas model, we would write

$$\ln Y = \beta_1 + 1.0 \ln L + \beta_3 \ln K + \varepsilon,$$

or

$$\ln Y - \ln L = \beta_1 + \beta_3 \ln K + \varepsilon.$$

Thus, the restricted model is estimated by regressing $\ln Y - \ln L$ on a constant and $\ln K$. Some care is needed if this regression is to be used to compute an F statistic. If the F statistic is computed using the sum of squared residuals [see (5-29)], then no problem will arise. If (5-30) is used instead, however, then it may be necessary to account for the restricted regression having a different dependent variable from the unrestricted one. In the preceding regression, the dependent variable in the unrestricted regression is $\ln Y$, whereas in the restricted regression, it is $\ln Y - \ln L$. The R^2 from the restricted regression is only 0.26979, which would imply an F statistic of 285.96, whereas the correct value is 9.935. If we compute the appropriate R_*^2 using the correct denominator, however, then its value is 0.92006 and the correct F value results.

Note that the coefficient on $\ln K$ is negative in the translog model. We might conclude that the estimated output elasticity with respect to capital now has the wrong sign. This conclusion would be incorrect, however. In the translog model, the capital elasticity of output is

$$\frac{\partial \ln Y}{\partial \ln K} = \beta_3 + \beta_5 \ln K + \beta_6 \ln L.$$

If we insert the coefficient estimates and the mean values for $\ln K$ and $\ln L$ (not the logs of the means) of 7.44592 and 5.7637, respectively, then the result is 0.5425, which is quite in line with our expectations and is fairly close to the value of 0.3757 obtained for the Cobb–Douglas model. The estimated standard error for this linear combination of the least squares estimates is computed as the square root of

$$\text{Est. Var}[b_3 + b_5 \overline{\ln K} + b_6 \overline{\ln L}] = \mathbf{w}'(\text{Est. Var}[\mathbf{b}])\mathbf{w},$$

where

$$\mathbf{w} = (0, 0, 1, 0, \overline{\ln K}, \overline{\ln L})'$$

and $\mathbf{b}$ is the full 6×1 least squares coefficient vector. This value is 0.1122, which is reasonably close to the earlier estimate of 0.0853.

Earlier, we used an F test to test the hypothesis that the coefficients on the three second order terms in the translog model were equal to zero, producing the Cobb–Douglas model. To use a Lagrange multiplier test, we use the restricted coefficient vector

$$\mathbf{b}_* = [1.1710, 0.6030, 0.3757, 0.0, 0.0, 0.0]'$$

[6] This case is not true when the restrictions are nonlinear. We consider this issue in Chapter 7.

to compute the residuals in the full regression,

$$e_* = \ln Y - b_{1*} - b_{2*} \ln L - b_{3*} \ln K - b_{4*} \ln^2 L/2 - b_{5*} \ln^2 K/2 - b_{6*} \ln L \ln K.$$

The R^2 in the regression of $\mathbf{e}_*$ on $\mathbf{X}$ is 0.20162, so the chi-squared is 27(0.20162) = 5.444. The critical value from the chi-squared table with 3 degrees of freedom is 7.815, so the null hypothesis is not rejected. Note that the F statistic computed earlier was 1.768. Our large-sample approximation to this would be 5.444/3 = 1.814.

5.4 LARGE-SAMPLE TESTS AND ROBUST INFERENCE

The finite sample distributions of the test statistics, t in (5-13) and F in (5-16), follow from the normality assumption for $\boldsymbol{\varepsilon}$. Without the normality assumption, the exact distributions of these statistics depend on the data and the parameters and are not F, t, and chi-squared. The large-sample results we considered in Section 4.4 suggest that although the usual t and F statistics are still usable, in the more general case without the special assumption of normality, they are viewed as approximations whose quality improves as the sample size increases. By using the results of Section D.3 (on asymptotic distributions) and some large-sample results for the least squares estimator, we can construct a set of usable inference procedures based on already familiar computations.

Assuming the data are well behaved, the *asymptotic* distribution of the least squares coefficient estimator, $\mathbf{b}$, is given by

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right] \quad \text{where } \mathbf{Q} = \text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right). \tag{5-31}$$

The interpretation is that, absent normality of $\boldsymbol{\varepsilon}$, as the sample size, n, grows, the normal distribution becomes an increasingly better approximation to the true, though at this point unknown, distribution of $\mathbf{b}$. As n increases, the distribution of $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta})$ *converges* exactly to a normal distribution, which is how we obtained the preceding finite-sample approximation. This result is based on the central limit theorem and does not require normally distributed disturbances. The second result we will need concerns the estimator of σ^2:

$$\text{plim } s^2 = \sigma^2, \quad \text{where } s^2 = \mathbf{e}'\mathbf{e}/(n - K).$$

With these in place, we can obtain some large-sample results for our test statistics that suggest how to proceed in a finite sample without an assumption of the distribution of the disturbances.

The sample statistic for testing the hypothesis that one of the coefficients, β_k, equals a particular value, β_k^0, is

$$t_k = \frac{\sqrt{n}(b_k - \beta_k^0)}{\sqrt{s^2(\mathbf{X}'\mathbf{X}/n)^{-1}_{kk}}}.$$

(Note that two occurrences of $\sqrt{n}$ cancel to produce our familiar result.) Under the null hypothesis, with normally distributed disturbances, t_k is exactly distributed as t with $n - K$ degrees of freedom. (See Theorem 4.6 and the beginning of this section.) The

exact distribution of this statistic is unknown, however, if $\boldsymbol{\varepsilon}$ is not normally distributed. From the preceding results, we find that the denominator of t_k converges to $\sqrt{\sigma^2 \mathbf{Q}_{kk}^{-1}}$. Hence, if t_k has a limiting distribution, then it is the same as that of the statistic that has this latter quantity in the denominator. (See point 3 of Theorem D.16.) That is, the large-sample distribution of t_k is the same as that of

$$\tau_k = \frac{\sqrt{n}(b_k - \beta_k^0)}{\sqrt{\sigma^2 \mathbf{Q}_{kk}^{-1}}}.$$

But $\tau_k = (b_k - E[b_k])/(\text{Asy. Var}[b_k])^{1/2}$ from the asymptotic normal distribution (under the hypothesis $\beta_k = \beta_k^0$), so it follows that τ_k has a standard normal asymptotic distribution, and this result is the large-sample distribution of our t statistic. Thus, as a large-sample approximation, we will use the standard normal distribution to approximate the true distribution of the test statistic t_k and use the critical values from the standard normal distribution for testing hypotheses.

The result in the preceding paragraph is valid only in large samples. For moderately sized samples, it provides only a suggestion that the t distribution may be a reasonable approximation. The appropriate critical values only *converge* to those from the standard normal, and generally *from above*, although we cannot be sure of this. In the interest of conservatism—that is, in controlling the probability of a Type I error—one should generally use the critical value from the t distribution even in the absence of normality. Consider, for example, using the standard normal critical value of 1.96 for a two-tailed test of a hypothesis based on 25 degrees of freedom. The nominal size of this test is 0.05. The actual size of the test, however, is the true, but unknown, probability that $|t_k| > 1.96$, which is 0.0612 if the $t[25]$ distribution is correct, and some other value if the disturbances are not normally distributed. The end result is that the standard t test retains a large-sample validity. Little can be said about the true size of a test based on the t distribution unless one makes some other equally narrow assumption about $\boldsymbol{\varepsilon}$, but the t distribution is generally used as a reliable approximation.

We will use the same approach to analyze the F statistic for testing a set of J linear restrictions. Step 1 will be to show that with normally distributed disturbances, JF converges to a chi-squared variable as the sample size increases. We will then show that this result is actually independent of the normality of the disturbances; it relies on the central limit theorem. Finally, we consider, as before, the appropriate critical values to use for this test statistic, which only has large-sample validity.

The F statistic for testing the validity of J linear restrictions, $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$, is given in (5-16). With normally distributed disturbances and under the null hypothesis, the exact distribution of this statistic is $F[J, n - K]$. To see how F behaves more generally, divide the numerator and denominator in (5-16) by σ^2 and rearrange the fraction slightly, so

$$F = \frac{(\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\}^{-1}(\mathbf{Rb} - \mathbf{q})}{J(s^2/\sigma^2)}. \tag{5-32}$$

Because plim $s^2 = \sigma^2$, and plim $(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, the denominator of F converges to J and the bracketed term in the numerator will behave the same as $(\sigma^2/n)\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}'$.

(See Theorem D16.3.) Hence, regardless of what this distribution is, if F has a limiting distribution, then it is the same as the limiting distribution of

$$W^* = \frac{1}{J}(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\sigma^2/n)\mathbf{Q}^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$$
$$= \frac{1}{J}(\mathbf{Rb} - \mathbf{q})'\{\text{Asy. Var}[\mathbf{Rb} - \mathbf{q}]\}^{-1}(\mathbf{Rb} - \mathbf{q}). \quad \textbf{(5-33)}$$

This expression is $(1/J)$ times a **Wald statistic**, based on the asymptotic distribution. The large-sample distribution of W^* will be that of $(1/J)$ times a chi-squared with J degrees of freedom. It follows that with normally distributed disturbances, JF converges to a chi-squared variate with J degrees of freedom. The proof is instructive.[7]

THEOREM 5.1 Limiting Distribution of the Wald Statistic
If $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Sigma}]$ *and if* H_0: $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ *is true, then*

$$W = (\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}'\}^{-1}(\mathbf{Rb} - \mathbf{q}) = JF \xrightarrow{d} \chi^2[J].$$

Proof: *Because* $\mathbf{R}$ *is a matrix of constants and* $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$,

$$\sqrt{n}\mathbf{R}(\mathbf{b} - \boldsymbol{\beta}) = \sqrt{n}(\mathbf{Rb} - \mathbf{q}) \xrightarrow{d} N[\mathbf{0}, \mathbf{R}\boldsymbol{\Sigma}\mathbf{R}']. \quad \textbf{(1)}$$

For convenience, write this equation as

$$\mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}]. \quad \textbf{(2)}$$

In Section A.6.11, we define the inverse square root of a positive definite matrix $\mathbf{P}$ *as another matrix, say* $\mathbf{T}$, *such that* $\mathbf{T}^2 = \mathbf{P}^{-1}$, *and denote* $\mathbf{T}$ *as* $\mathbf{P}^{-1/2}$. *Then, by the same reasoning as in (1) and (2),*

$$\text{if } \mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}], \quad \text{then} \quad \mathbf{P}^{-1/2}\mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}^{-1/2}\mathbf{P}\mathbf{P}^{-1/2}] = N[\mathbf{0}, \mathbf{I}]. \quad \textbf{(3)}$$

We now invoke Theorem D.21 for the limiting distribution of a function of a random variable. The sum of squares of uncorrelated (i.e., independent) standard normal variables is distributed as chi-squared. Thus, the limiting distribution of

$$(\mathbf{P}^{-1/2}\mathbf{z})'(\mathbf{P}^{-1/2}\mathbf{z}) = \mathbf{z}'\mathbf{P}^{-1}\mathbf{z} \xrightarrow{d} \chi^2(J). \quad \textbf{(4)}$$

Reassembling the parts from before, we have shown that the limiting distribution of

$$n(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \quad \textbf{(5)}$$

is chi-squared, with J degrees of freedom. Note the similarity of this result to the results of Section B.11.6. Finally, if $\hat{\boldsymbol{\Sigma}}$ *is an appropriate estimator of* $\boldsymbol{\Sigma}$, *such as* $s^2(\mathbf{X}'\mathbf{X}/n)$ *assuming Assumption A4 or the estimators in* (4-37) *or* (4-42), *with*

$$\text{plim } \hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}, \quad \textbf{(6)}$$

then the statistic obtained by replacing $\boldsymbol{\Sigma}$ *by* $\hat{\boldsymbol{\Sigma}}$ *in (5) has the same limiting chi-squared distribution.*

[7] See White (2001, p. 76).

The result in (5-33) is more general than it might appear. It is based generically on Asy.Var[**b**]. We can extend the Wald statistic to use our more robust estimators of Asy.Var[**b**], for examples, the heteroscedasticity robust estimator shown in Section 4.5.2 and the cluster robust estimator in Section 4.5.3 (and other variants such as a time-series correction to be developed in Section 20.5.2).

The appropriate critical values for the F test of the restrictions $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ converge from above to $1/J$ times those for a chi-squared test based on the Wald statistic. For example, for testing $J = 5$ restrictions, the critical value from the chi-squared table for 95% significance is 11.07. The critical values from the F table are $3.33 = 16.65/5$ for $n - K = 10$, $2.60 = 13.00/5$ for $n - K = 25$, $2.40 = 12.00/5$ for $n - K = 50$, $2.31 = 11.55/5$ for $n - K = 100$, and $2.214 = 11.07/5$ for large $n - K$. Thus, with normally distributed disturbances, as n gets large, the F test can be carried out by referring JF to the critical values from the chi-squared table.

The crucial result for our purposes here is that the distribution of the Wald statistic is built up from the asymptotic distribution of **b**, which is normal even without normally distributed disturbances. The implication is that the Wald statistic based on a robust asymptotic covariance matrix for **b** is an appropriate large-sample test statistic. (For linear restrictions, if the disturbances are homoscedastic, then the chi-squared statistic may be computed simply as JF.) This implication relies on the central limit theorem, not on normally distributed disturbances. The critical values from the F table remains a conservative approach that becomes more accurate as the sample size increases. For example, we see Cameron and Miller (2015) recommend basing hypothesis testing on the F distribution even after adjusting the asymptotic covariance matrix for **b** for cluster sampling with a moderate number of clusters.

5.5 TESTING NONLINEAR RESTRICTIONS

The preceding discussion has relied heavily on the linearity of the regression model. When we analyze nonlinear functions of the parameters and nonlinear regression models, most of these exact distributional results no longer hold.

The general problem is that of testing a hypothesis that involves a nonlinear function of the regression coefficients:

$$H_0\colon c(\boldsymbol{\beta}) = q.$$

We shall look first at the case of a single restriction. The more general case, in which $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$ is a set of restrictions, is a simple extension. The counterpart to the test statistic we used earlier would be

$$z = \frac{c(\hat{\boldsymbol{\beta}}) - q}{\text{estimated standard error}},$$

or its square, which in the preceding were distributed as $t[n - K]$ and $F[1, n - K]$, respectively. The discrepancy in the numerator presents no difficulty. Obtaining an estimate of the sampling variance of $c(\hat{\boldsymbol{\beta}}) - q$, however, involves the variance of a nonlinear function of $\hat{\boldsymbol{\beta}}$.

The results we need for this computation are presented in Sections 4.4.4, B.10.3, and D.3.1. A linear Taylor series approximation to $c(\hat{\boldsymbol{\beta}})$ around the true parameter vector $\boldsymbol{\beta}$ is

$$c(\hat{\boldsymbol{\beta}}) \approx c(\boldsymbol{\beta}) + \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{5-34}$$

We must rely on consistency rather than unbiasedness here, because, in general, the expected value of a nonlinear function is not equal to the function of the expected value. If plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, then we are justified in using $c(\hat{\boldsymbol{\beta}})$ as an estimate of $c(\boldsymbol{\beta})$. (The relevant result is the Slutsky theorem.) Assuming that our use of this approximation is appropriate, the variance of the nonlinear function is approximately equal to the variance of the right-hand side, which is, then,

$$\text{Var}[c(\hat{\boldsymbol{\beta}})] \approx \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)' \text{Asy.Var}[\hat{\boldsymbol{\beta}}]\left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right). \tag{5-35}$$

The derivatives in the expression for the variance are functions of the unknown parameters. Because these are being estimated, we use our sample estimates in computing the derivatives and the estimator of the asymptotic variance of $\mathbf{b}$. Finally, we rely on Theorem D.22 in Section D.3.1 and use the standard normal distribution instead of the t distribution for the test statistic. Using $\mathbf{g}(\hat{\boldsymbol{\beta}})$ to estimate $\mathbf{g}(\boldsymbol{\beta}) = \partial c(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, we can now test a hypothesis in the same fashion we did earlier.

Example 5.6 A Long-Run Marginal Propensity to Consume

A consumption function that has different short- and long-run marginal propensities to consume can be written in the form

$$\ln C_t = \alpha + \beta \ln Y_t + \gamma \ln C_{t-1} + \varepsilon_t,$$

which is a **distributed lag** model. In this model, the short-run marginal propensity to consume (MPC) (elasticity, given the variables are in logs) is β, and the long-run MPC is $\delta = \beta/(1 - \gamma)$. Consider testing the hypothesis that $\delta = 1$.

Quarterly data on aggregate U.S. consumption and disposable personal income for the years 1950 to 2000 are given in Appendix Table F5.2. The estimated equation based on these data is

$$\ln C_t = \underset{(0.01055)}{0.003142} + \underset{(0.02873)}{0.07495} \ln Y_t + \underset{(0.02859)}{0.9246} \ln C_{t-1} + e_t, \quad R^2 = 0.999712, \quad s = 0.00874.$$

Estimated standard errors are shown in parentheses. We will also require Est. Asy. Cov[b,c] = −0.0008207. The estimate of the long-run MPC is $d = b/(1 - c) = 0.07495/(1 - 0.9246) = 0.99402$. To compute the estimated variance of d, we will require $g_b = \partial d/\partial b = 1/(1 - c) = 13.2626$ and $g_c = \partial d/\partial c = b/(1 - c)^2 = 13.1834$. The estimated asymptotic variance of d is

$$\begin{aligned}\text{Est. Asy. Var}[d] &= g_b^2 \text{ Est. Asy. Var}[b] + g_c^2 \text{ Est. Asy. Var}[c] + 2g_b g_c \text{ Est. Asy. Cov}[b, c] \\ &= 13.2626^2 \times 0.02873^2 + 13.1834^2 \times 0.02859^2 \\ &\quad + 2(13.2626)(13.1834)(-0.0008207) = 0.0002585.\end{aligned}$$

The square root is 0.016078. To test the hypothesis that the long-run MPC is greater than or equal to 1, we would use

$$z = \frac{0.99403 - 1}{0.016078} = -0.37131.$$

Because we are using a large-sample approximation, we refer to a standard normal table instead of the t distribution. The hypothesis that $\gamma = 1$ is not rejected.

You may have noticed that we could have tested this hypothesis with a linear restriction instead; if $\delta = 1$, then $\beta = 1 - \gamma$, or $\beta + \gamma = 1$. The estimate is $q = b + c - 1 = -0.00045$. The estimated standard error of this linear function is $[0.02873^2 + 0.02859^2 - 2(0.0008207)]^{1/2} = 0.00118$. The t ratio for this test is −0.38135, which is almost the same as before. Because the sample used here is fairly large, this is

to be expected. However, there is nothing in the computations that ensures this outcome. In a smaller sample, we might have obtained a different answer. For example, using only the last 11 years of the data, the t statistics for the two hypotheses are 7.652 and 5.681. The Wald test is not invariant to how the hypothesis is formulated. In a borderline case, we could have reached a different conclusion. This **lack of invariance** does not occur with the likelihood ratio or Lagrange multiplier tests discussed in Chapter 14. On the other hand, both of these tests require an assumption of normality, whereas the Wald statistic does not. This illustrates one of the trade-offs between a more detailed specification and the power of the test procedures that are implied.

The generalization to more than one function of the parameters proceeds along similar lines. Let $\mathbf{c}(\hat{\boldsymbol{\beta}})$ be a set of J functions of the estimated parameter vector and let the $J \times K$ matrix of derivatives of $\mathbf{c}(\hat{\boldsymbol{\beta}})$ be

$$\hat{\mathbf{G}} = \frac{\partial \mathbf{c}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'}. \tag{5-36}$$

The estimate of the asymptotic covariance matrix of these functions is

$$\text{Est. Asy. Var}[\hat{\mathbf{c}}] = \hat{\mathbf{G}}\{\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}]\}\hat{\mathbf{G}}'. \tag{5-37}$$

The jth row of $\hat{\mathbf{G}}$ is the K derivatives of $c_j(\hat{\boldsymbol{\beta}})$ with respect to the K elements of $\hat{\boldsymbol{\beta}}$. For example, the covariance matrix for estimates of the short- and long-run marginal propensities to consume would be obtained using

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1/(1-\gamma) & \beta/(1-\gamma)^2 \end{bmatrix}.$$

The statistic for testing the J hypotheses $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$ is

$$W = (\hat{\mathbf{c}} - \mathbf{q})'\{\text{Est. Asy. Var}[\hat{\mathbf{c}}]\}^{-1}(\hat{\mathbf{c}} - \mathbf{q}). \tag{5-38}$$

In large samples, W has a chi-squared distribution with degrees of freedom equal to the number of restrictions. Note that for a single restriction, this value is the square of the statistic in (5-33).

5.6 CHOOSING BETWEEN NONNESTED MODELS

The classical testing procedures that we have been using have been shown to be most powerful for the types of hypotheses we have considered.[8] Although use of these procedures is clearly desirable, the requirement that we express the hypotheses in the form of restrictions on the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,

$$H_0\text{: } \mathbf{R}\boldsymbol{\beta} = \mathbf{q}$$

versus

$$H_1\text{: } \mathbf{R}\boldsymbol{\beta} \neq \mathbf{q},$$

can be limiting. Two common exceptions are the general problem of determining which of two possible sets of regressors is more appropriate and whether a linear or loglinear

[8] See, for example, Stuart and Ord (1989, Chapter 27).

model is more appropriate for a given analysis. For the present, we are interested in comparing two competing linear models:

$$H_0: \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \tag{5-39a}$$

and

$$H_1: \mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_1. \tag{5-39b}$$

The classical procedures we have considered thus far provide no means of forming a preference for one model or the other. The general problem of testing nonnested hypotheses such as these has attracted an impressive amount of attention in the theoretical literature and has appeared in a wide variety of empirical applications.[9]

5.6.1 TESTING NONNESTED HYPOTHESES

A useful distinction between hypothesis testing, as discussed in the preceding chapters and model selection as considered here, will turn on the asymmetry between the null and alternative hypotheses that is a part of the classical testing procedure.[10] Because, by construction, the classical procedures seek evidence in the sample to refute the *null* hypothesis, how one frames the null can be crucial to the outcome. Fortunately, the Neyman–Pearson methodology provides a prescription; the null is usually cast as the narrowest model in the set under consideration. On the other hand, the classical procedures never reach a sharp conclusion. Unless the significance level of the testing procedure is made so high as to exclude all alternatives, there will always remain the possibility of a Type I error. As such, the null hypothesis is never rejected with certainty, but only with a prespecified degree of confidence. Model selection tests, in contrast, give the competing hypotheses equal standing. There is no natural null hypothesis. However, the end of the process is a firm decision—in testing (5-39a, b), one of the models will be rejected and the other will be retained; the analysis will then proceed in the framework of that one model and not the other. Indeed, it cannot proceed until one of the models is discarded. It is common, for example, in this new setting for the analyst first to test with one model cast as the null, then with the other. Unfortunately, given the way the tests are constructed, it can happen that both or neither model is rejected; in either case, further analysis is clearly warranted. As we shall see, the science is a bit inexact.

The earliest work on nonnested hypothesis testing, notably Cox (1961, 1962), was done in the framework of sample likelihoods and maximum likelihood procedures. Recent developments have been structured around a common pillar labeled the **encompassing principle**.[11] Essentially, the principle directs attention to the question of whether a maintained model can explain the features of its competitors, that is, whether the maintained model encompasses the alternative. Yet a third approach is based on forming a **comprehensive model** that contains both competitors as special cases. When

[9]Surveys on this subject are White (1982a, 1983), Gourieroux and Monfort (1994), McAleer (1995), and Pesaran and Weeks (2001). McAleer's survey tabulates an array of applications, while Gourieroux and Monfort focus on the underlying theory.

[10]See Granger and Pesaran (2000) for discussion.

[11]See Mizon and Richard (1986).

possible, the test between models can be based, essentially, on classical (-like) testing procedures. We will examine tests that exemplify all three approaches.

5.6.2 AN ENCOMPASSING MODEL

The encompassing approach is one in which the ability of one model to explain features of another is tested. Model 0 *encompasses* Model 1 if the features of Model 1 can be explained by Model 0, but the reverse is not true.[12] Because H_0 cannot be written as a restriction on H_1, none of the procedures we have considered thus far is appropriate. One possibility is an artificial nesting of the two models. Let $\overline{\mathbf{X}}$ be the set of variables in $\mathbf{X}$ that are not in $\mathbf{Z}$, define $\overline{\mathbf{Z}}$ likewise with respect to $\mathbf{X}$, and let $\mathbf{W}$ be the variables that the models have in common. Then H_0 and H_1 could be combined in a supermodel:

$$\mathbf{y} = \overline{\mathbf{X}}\,\overline{\boldsymbol{\beta}} + \overline{\mathbf{Z}}\,\overline{\boldsymbol{\gamma}} + \mathbf{W}\,\boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

In principle, H_1 is rejected if it is found that $\overline{\boldsymbol{\gamma}} = \mathbf{0}$ by a conventional F test, whereas H_0 is rejected if it is found that $\overline{\boldsymbol{\beta}} = \mathbf{0}$. There are two problems with this approach. First, $\boldsymbol{\delta}$ remains a mixture of parts of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and it is not established by the F test that either of these parts is zero. Hence, this test does not really distinguish between H_0 and H_1; it distinguishes between H_1 and a hybrid model. Second, this compound model may have an extremely large number of regressors. In a time-series setting, the problem of collinearity may be severe.

Consider an alternative approach. If H_0 is correct, then $\mathbf{y}$ will, apart from the random disturbance $\boldsymbol{\varepsilon}$, be fully explained by $\mathbf{X}$. Suppose we then attempt to estimate $\boldsymbol{\gamma}$ by regression of $\mathbf{y}$ on $\mathbf{Z}$. Whatever set of parameters is estimated by this regression, say, $\mathbf{c}$, if H_0 is correct, then we should estimate exactly the same coefficient vector if we were to regress $\mathbf{X}\boldsymbol{\beta}$ on $\mathbf{Z}$, because $\boldsymbol{\varepsilon}_0$ is random noise under H_0. Because $\boldsymbol{\beta}$ must be estimated, suppose that we use $\mathbf{Xb}$ instead and compute $\mathbf{c}_0$. A test of the proposition that Model 0 encompasses Model 1 would be a test of the hypothesis that $E[\mathbf{c} - \mathbf{c}_0] = \mathbf{0}$. It is straightforward to show that the test can be carried out by using a standard F test to test the hypothesis that $\boldsymbol{\gamma}_1 = \mathbf{0}$ in the augmented regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1,$$

where $\mathbf{Z}_1$ is the variables in $\mathbf{Z}$ that are not in $\mathbf{X}$.[13] (Of course, a line of manipulation reveals that $\overline{\mathbf{Z}}$ and $\mathbf{Z}_1$ are the same, so the tests are also.)

5.6.3 COMPREHENSIVE APPROACH—THE *J* TEST

The ***J* test** proposed by Davidson and MacKinnon (1981) can be shown to be an application of the encompassing principle to the linear regression model.[14] Their suggested alternative to the preceding compound model is

$$\mathbf{y} = (1 - \lambda)\mathbf{X}\boldsymbol{\beta} + \lambda(\mathbf{Z}\boldsymbol{\gamma}) + \boldsymbol{\varepsilon}.$$

In this model, a test of $\lambda = 0$ would be a test against H_1. The problem is that λ cannot be separately estimated in this model; it would amount to a redundant scaling of the

[12]See Aneuryn-Evans and Deaton (1980), Deaton (1982), Dastoor (1983), Gourieroux et al. (1983, 1995), and, especially, Mizon and Richard (1986).

[13]See Davidson and MacKinnon (2004, pp. 671–672).

[14]See Pesaran and Weeks (2001).

regression coefficients. Davidson and MacKinnon's (1984) J test consists of estimating $\boldsymbol{\gamma}$ by a least squares regression of $\mathbf{y}$ on $\mathbf{Z}$ followed by a least squares regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{Z}\hat{\boldsymbol{\gamma}}$, the fitted values in the first regression. A valid test, at least asymptotically, of H_1 is to test H_0: $\lambda = 0$. If H_0 is true, then plim $\hat{\lambda} = 0$. Asymptotically, the ratio $\hat{\lambda}/\text{se}(\hat{\lambda})$ (i.e., the usual t ratio) is distributed as standard normal and may be referred to the standard table to carry out the test. Unfortunately, in testing H_0 versus H_1 and vice versa, all four possibilities (reject both, neither, or either one of the two hypotheses) could occur. This issue, however, is a finite sample problem. Davidson and MacKinnon show that as $n \to \infty$, if H_1 is true, then the probability that $\hat{\lambda}$ will differ significantly from 0 approaches 1.

Example 5.7 J Test for a Consumption Function

Gaver and Geisel (1974) propose two forms of a consumption function:

$$H_0\text{: } C_t = \beta_1 + \beta_2 Y_t + \beta_3 Y_{t-1} + \varepsilon_{0t},$$

and

$$H_1\text{: } C_t = \gamma_1 + \gamma_2 Y_t + \gamma_3 C_{t-1} + \varepsilon_{1t}.$$

The first model states that consumption responds to changes in income over two periods, whereas the second states that the effects of changes in income on consumption persist for many periods. Quarterly data on aggregate U.S. real consumption and real disposable income are given in Appendix Table F5.2. Here we apply the J test to these data and the two proposed specifications. First, the two models are estimated separately (using observations 1950II through 2000IV). The least squares regression of C on a constant, Y, lagged Y, and the fitted values from the second model produces an estimate of λ of 1.0145 with a t ratio of 62.861. Thus, H_0 should be rejected in favor of H_1. But reversing the roles of H_0 and H_1, we obtain an estimate of λ of -10.677 with a t ratio of -7.188. Thus, H_1 is rejected as well.[15]

5.7 A SPECIFICATION TEST

The tests considered so far have evaluated nested models. The presumption is that one of the two models is correct. In Section 5.6, we broadened the range of models considered to allow two nonnested models. It is not assumed that either model is necessarily the true data-generating process; the test attempts to ascertain which of two competing models is closer to the truth. Specification tests fall between these two approaches. The idea of a specification test is to consider a particular null model and alternatives that are not explicitly given in the form of restrictions on the regression equation. A useful way to consider some specification tests is as if the core model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, is the null hypothesis and the alternative is a possibly unstated generalization of that model. Ramsey's (1969) RESET test is one such test which seeks to uncover nonlinearities in the functional form. One (admittedly ambiguous) way to frame the analysis is

$$H_0\text{: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\text{, } H_1\text{: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \text{higher-order powers of } x_k \text{ and other terms} + \boldsymbol{\varepsilon}.$$

A straightforward approach would be to add squares, cubes, and cross-products of the regressors to the equation and test down to H_0 as a restriction on the larger model. Two complications are that this approach might be too specific about the form of the

[15]For related discussion of this possibility, see McAleer, Fisher, and Volker (1982).

alternative hypothesis and, second, with a large number of variables in $\mathbf{X}$, it could become unwieldy. Ramsey's proposed solution is to add powers of $\mathbf{x}_i'\boldsymbol{\beta}$ to the regression using the least squares predictions—typically, one would add the square and, perhaps the cube. This would require a two-step estimation procedure, because in order to add $(\mathbf{x}_i'\mathbf{b})^2$ and $(\mathbf{x}_i'\mathbf{b})^3$, one needs the coefficients. The suggestion, then, is to fit the null model first, using least squares. Then, for the second step, the squares (and cubes) of the predicted values from this first-step regression are added to the equation and it is refit with the additional variables. A (large-sample) Wald test is then used to test the hypothesis of the null model.

As a general strategy, this sort of specification is designed to detect failures of the assumptions of the null model. The obvious virtue of such a test is that it provides much greater generality than a simple test of restrictions such as whether a coefficient is zero. But that generality comes at considerable cost:

1. The test is nonconstructive. It gives no indication what the researcher should do next if the null model is rejected. This is a general feature of specification tests. Rejection of the null model does not imply any particular alternative.
2. Because the alternative hypothesis is unstated, it is unclear what the power of this test is against any specific alternative.
3. For this specific test (perhaps not for some other specification tests we will examine later), because $\mathbf{x}_i'\mathbf{b}$ uses the same $\mathbf{b}$ for every observation, the observations are correlated, while they are assumed to be uncorrelated in the original model. Because of the two-step nature of the estimator, it is not clear what is the appropriate covariance matrix to use for the Wald test. Two other complications emerge for this test. First, it is unclear what the coefficients converge to, assuming they converge to anything. Second, the variance of the difference between $\mathbf{x}_i'\mathbf{b}$ and $\mathbf{x}_i'\boldsymbol{\beta}$ is a function of $\mathbf{x}$, so the second-step regression might be heteroscedastic. The implication is that neither the size nor the power of this test is necessarily what might be expected.

Example 5.8 Size of a RESET Test

To investigate the true size of the RESET test in a particular application, we carried out a Monte Carlo experiment. The results in Table 4.7 give the following estimates of Equation (5-2):

$$\ln \textit{Price} = -8.34237 + 1.31638 \ln \textit{Area} - 0.09623\ \textit{Aspect Ratio} + \varepsilon, \text{ where } sd(\varepsilon) = 1.10435.$$

We take the estimated right-hand side to be our population. We generated 5,000 samples of 430 (the original sample size), by reusing the regression coefficients and generating a new sample of disturbances for each replication. Thus, with each replication, r, we have a new sample of observations on $\ln \textit{Price}_{ir}$ where the regression part is as above reused and a new set of disturbances is generated each time. With each sample, we computed the least squares coefficient, then the predictions. We then recomputed the least squares regression while adding the square and cube of the prediction to the regression. Finally, with each sample, we computed the chi-squared statistic, and rejected the null model if the chi-squared statistic is larger than 5.99, the 95th percentile of the chi-squared distribution with two degrees of freedom. The **nominal size** of this test is 0.05. Thus, in samples of 100, 500, 1,000, and 5,000, we should reject the null model 5, 25, 50, and 250 times. In our experiment, the computed chi-squared exceeded 5.99 8, 31, 65, and 259 times, respectively, which suggests that at least with sufficient replications, the test performs as might be expected. We then investigated the

power of the test by adding 0.1 times the square of ln *Area* to the predictions. It is not possible to deduce the exact power of the RESET test to detect this failure of the null model. In our experiment, with 1,000 replications, the null hypothesis is rejected 321 times. We conclude that the procedure does appear to have the power to detect this failure of the model assumptions.

5.8 MODEL BUILDING—A GENERAL TO SIMPLE STRATEGY

There has been a shift in the general approach to model building. With an eye toward maintaining simplicity, model builders would generally begin with a small specification and gradually build up the model ultimately of interest by adding variables. But, based on the preceding results, we can surmise that just about any criterion that would be used to decide whether to add a variable to a current specification would be tainted by the biases caused by the incomplete specification at the early steps. Omitting variables from the equation seems generally to be the worse of the two errors. Thus, the **simple-to-general** approach to model building has little to recommend it. Researchers are more comfortable beginning their specification searches with large elaborate models involving many variables and perhaps long and complex lag structures. The attractive strategy is then to adopt a **general-to-simple**, downward reduction of the model to the preferred specification. Of course, this must be tempered by two related considerations. In the *kitchen sink* regression, which contains every variable that might conceivably be relevant, the adoption of a fixed probability for the Type I error, say, 5%, ensures that in a big enough model, some variables will appear to be significant, even if by accident. Second, the problems of pretest estimation and **stepwise model building** also pose some risk of ultimately misspecifying the model. To cite one unfortunately common example, the statistics involved often produce unexplainable lag structures in dynamic models with many lags of the dependent or independent variables.

5.8.1 MODEL SELECTION CRITERIA

The preceding discussion suggested some approaches to model selection based on nonnested hypothesis tests. Fit measures and testing procedures based on the sum of squared residuals, such as R^2 and the Cox (1961) test, are useful when interest centers on the within-sample fit or within-sample prediction of the dependent variable. When the model building is directed toward forecasting, within-sample measures are not necessarily optimal. As we have seen, R^2 cannot fall when variables are added to a model, so there is a built-in tendency to overfit the model. This criterion may point us away from the best forecasting model, because adding variables to a model may increase the variance of the forecast error despite the improved fit to the data. With this thought in mind, the **adjusted R^2**,

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2) = 1 - \frac{n-1}{n-K}\left(\frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right), \tag{5-40}$$

has been suggested as a fit measure that appropriately penalizes the loss of degrees of freedom that result from adding variables to the model. Note that $\bar{R}^2$ may fall when a variable is added to a model if the sum of squares does not fall fast enough. (The applicable result appears in Theorem 3.7; $\bar{R}^2$ does not rise when a variable is added to

a model unless the t ratio associated with that variable exceeds one in absolute value.) The adjusted R^2 has been found to be a preferable fit measure for assessing the fit of forecasting models.[16]

The adjusted R^2 penalizes the loss of degrees of freedom that occurs when a model is expanded. There is, however, some question about whether the penalty is sufficiently large to ensure that the criterion will necessarily lead the analyst to the correct model (assuming that it is among the ones considered) as the sample size increases. Two alternative fit measures that have been suggested are the **Akaike Information Criterion**,

$$\text{AIC}(K) = s_y^2(1 - R^2)e^{2K/n} \tag{5-41}$$

and the Schwarz or **Bayesian Information Criterion**,

$$\text{BIC}(K) = s_y^2(1 - R^2)n^{K/n}. \tag{5-42}$$

(There is no degrees of freedom correction in s_y^2.) Both measures improve (decline) as R^2 increases (decreases), but, everything else constant, degrade as the model size increases. Like $\overline{R}^2$, these measures place a premium on achieving a given fit with a smaller number of parameters per observation, K/n. Logs are usually more convenient; the measures reported by most software are

$$\text{AIC}(K) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2K}{n} \tag{5-43}$$

$$\text{BIC}(K) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{K \ln n}{n}. \tag{5-44}$$

Each **prediction criterion** has its virtues, and neither has an obvious advantage over the other.[17] The **Schwarz criterion**, with its heavier penalty for degrees of freedom lost, will lean toward a simpler model. All else given, simplicity does have some appeal.

5.8.2 MODEL SELECTION

The preceding has laid out a number of choices for **model selection**, but, at the same time, has posed some uncomfortable propositions. The pretest estimation aspects of specification search are based on the model builder's knowledge of the truth and the consequences of failing to use that knowledge. While the cautions about blind search for statistical significance are well taken, it does seem optimistic to assume that the correct model is likely to be known with hard certainty at the outset of the analysis. The bias documented in (4-9) is well worth the modeler's attention. But, in practical terms, knowing anything about the magnitude presumes that we know what variables are in $\mathbf{X}_2$, which need not be the case. While we can agree that the model builder will omit income from a demand equation at his peril, we could also have some sympathy for the analyst faced with finding the right specification for his forecasting model among dozens of choices. The tests for nonnested models would seem to free the modeler from having to claim that the specified set of models contain the truth. But, a moment's thought should suggest that the cost of this is the possibly deflated power of these procedures to point

[16]See Diebold (2007), who argues that the simple R^2 has a downward bias as a measure of the out-of-sample, one-step-ahead prediction error variance.

[17]See Diebold (2007).

toward that truth. The *J* test may provide a sharp choice between two alternatives, but it neglects the third possibility that both models are wrong. Vuong's test (see Section 14.6.6) does but, of course, it suffers from the fairly large inconclusive region, which is a symptom of its relatively low power against many alternatives. The upshot of all of this is that there remains much to be accomplished in the area of model selection. Recent commentary has provided suggestions from two perspectives, classical and Bayesian.

5.8.3 CLASSICAL MODEL SELECTION

Hansen (2005) lists four shortcomings of the methodology we have considered here:

1. Parametric vision
2. Assuming a true data-generating process
3. Evaluation based on fit
4. Ignoring model uncertainty

All four of these aspects have framed the analysis of the preceding sections. Hansen's view is that the analysis considered here is too narrow and stands in the way of progress in model discovery.

All the model selection procedures considered here are based on the likelihood function, which requires a specific distributional assumption. Hansen argues for a focus, instead, on semiparametric structures. For regression analysis, this points toward generalized method of moments estimators. Casualties of this reorientation will be distributionally based test statistics such as the Cox and Vuong statistics, and even the AIC and BIC measures, which are transformations of the likelihood function. However, alternatives have been proposed.[18] The second criticism is one we have addressed. The assumed true model can be a straightjacket. Rather (he argues), we should view our specifications as approximations to the underlying true data-generating process—this greatly widens the specification search, to one for a model which provides the best approximation. Of course, that now forces the question of what is best. So far, we have focused on the likelihood function, which in the classical regression can be viewed as an increasing function of R^2. The author argues for a more focused information criterion (FIC) that examines directly the parameters of interest, rather than the fit of the model to the data. Each of these suggestions seeks to improve the process of model selection based on familiar criteria, such as test statistics based on fit measures and on characteristics of the model.

A (perhaps *the*) crucial issue remaining is uncertainty about the model itself. The search for the correct model is likely to have the same kinds of impacts on statistical inference as the search for a specification given the form of the model (see Sections 4.3.2 and 4.3.3). Unfortunately, incorporation of this kind of uncertainty in statistical inference procedures remains an unsolved problem. Hansen suggests one potential route would be the Bayesian model averaging methods discussed next although he does express some skepticism about Bayesian methods in general.

5.8.4 BAYESIAN MODEL AVERAGING

If we have doubts as to which of two models is appropriate, then we might well be convinced to concede that possibly neither one is really the truth. We have painted ourselves into a corner with our left or right approach to testing. The Bayesian approach

[18]For example, by Hong, Preston, and Shum (2000).

to this question would treat it as a problem of comparing the two hypotheses rather than testing for the validity of one over the other. We enter our sampling experiment with a set of prior probabilities about the relative merits of the two hypotheses, which is summarized in a *prior odds ratio*, $P_{01} = \text{Prob}[H_0]/\text{Prob}[H_1]$. After gathering our data, we construct the Bayes factor, which summarizes the weight of the sample evidence in favor of one model or the other. After the data have been analyzed, we have our *posterior odds ratio*, $P_{01}|\text{data} = \text{Bayes factor} \times P_{01}$. The upshot is that ex post, neither model is discarded; we have merely revised our assessment of the comparative likelihood of the two in the face of the sample data. Of course, this still leaves the specification question open. Faced with a choice among models, how can we best use the information we have? Recent work on **Bayesian model averaging** has suggested an answer.[19]

An application by Wright (2003) provides an interesting illustration. Recent advances such as Bayesian VARs have improved the forecasting performance of econometric models. Stock and Watson (2001, 2004) report that striking improvements in predictive performance of international inflation can be obtained by averaging a large number of forecasts from different models and sources. The result is remarkably consistent across subperiods and countries. Two ideas are suggested by this outcome. First, the idea of blending different models is very much in the spirit of Hansen's fourth point. Second, note that the focus of the improvement is not on the fit of the model (point 3), but its predictive ability. Stock and Watson suggested that simple equal-weighted averaging, while one could not readily explain why, seems to bring large improvements. Wright proposed Bayesian model averaging as a means of making the choice of the weights for the average more systematic and of gaining even greater predictive performance.

Leamer (1978) appears to be the first to propose Bayesian model averaging as a means of combining models. The idea has been studied more recently by Min and Zellner (1993) for output growth forecasting, Doppelhofer et al. (2000) for cross-country growth regressions, Koop and Potter (2004) for macroeconomic forecasts, and others. Assume that there are M models to be considered, indexed by $m = 1, \ldots, M$. For simplicity, we will write the mth model in a simple form, $f_m(\mathbf{y}|\mathbf{Z}, \boldsymbol{\theta}_m)$ where $f(.)$ is the density, $\mathbf{y}$ and $\mathbf{Z}$ are the data, and $\boldsymbol{\theta}_m$ is the parameter vector for model m. Assume, as well, that model m^* is the true model, unknown to the analyst. The analyst has priors π_m over the probabilities that model m is the correct model, so π_m is the prior probability that $m = m^*$. The posterior probabilities for the models are

$$\Pi_m = \text{Prob}(m = m^*|\mathbf{y}, \mathbf{Z}) = \frac{P(\mathbf{y}, \mathbf{Z}|m)\pi_m}{\sum_{r=1}^{M} P(\mathbf{y}, \mathbf{Z}|r)\pi_r}, \tag{5-45}$$

where $P(\mathbf{y}, \mathbf{Z}|m)$ is the marginal likelihood for the mth model,

$$P(\mathbf{y}, \mathbf{Z}|m) = \int_{\theta_m} P(\mathbf{y}, \mathbf{Z}|\boldsymbol{\theta}_m, m)P(\boldsymbol{\theta}_m)d\,\boldsymbol{\theta}_m, \tag{5-46}$$

while $P(\mathbf{y}, \mathbf{Z}|\boldsymbol{\theta}_m, m)$ is the conditional (on $\boldsymbol{\theta}_m$) likelihood for the mth model and $P(\boldsymbol{\theta}_m)$ is the analyst's prior over the parameters of the mth model. This provides an alternative set of weights to the $\Pi_m = 1/M$ suggested by Stock and Watson. Let $\hat{\boldsymbol{\theta}}_m$ denote the Bayesian estimate (posterior mean) of the parameters of model m. (See Chapter 16.)

[19]See Hoeting et al. (1999).

Each model provides an appropriate posterior forecast density, $f^*(\mathbf{y}|\mathbf{Z}, \hat{\boldsymbol{\theta}}_m, m)$. The Bayesian model averaged forecast density would then be

$$\bar{f}^* = \sum_{m=1}^{M} f^*(\mathbf{y}|\mathbf{Z}, \hat{\boldsymbol{\theta}}_m, m)\Pi_m. \tag{5-47}$$

A point forecast would be a similarly weighted average of the forecasts from the individual models.

Example 5.9 Bayesian Averaging of Classical Estimates

Many researchers have expressed skepticism of Bayesian methods because of the apparent arbitrariness of the specifications of prior densities over unknown parameters. In the Bayesian model averaging setting, the analyst requires prior densities over not only the model probabilities, π_m, but also the model specific parameters, θ_m. In their application, Doppelhofer, Miller, and Sala-i-Martin (2000) were interested in the appropriate set of regressors to include in a long-term macroeconomic (income) growth equation. With 32 candidates, M for their application was 2^{32} (minus one if the zero regressors model is ignored), or roughly four billion. Forming this many priors would be optimistic in the extreme. The authors proposed a novel method of weighting a large subset (roughly 21 million) of the 2^M possible (classical) least squares regressions. The weights are formed using a Bayesian procedure; however, the estimates that are weighted are the classical least squares estimates. While this saves considerable computational effort, it still requires the computation of millions of least squares coefficient vectors.[20] The end result is a model with 12 independent variables.

5.9 SUMMARY AND CONCLUSIONS

This chapter has focused on the third use of the linear regression model, hypothesis testing. The central result for testing hypotheses is the F statistic. The F ratio can be produced in two equivalent ways: first, by measuring the extent to which the unrestricted least squares estimate differs from what a hypothesis would predict, and second, by measuring the loss of fit that results from assuming that a hypothesis is correct. We then extended the F statistic to more general settings by examining its large-sample properties, which allow us to discard the assumption of normally distributed disturbances and by extending it to nonlinear restrictions.

This is the last of five chapters that we have devoted specifically to the methodology surrounding the most heavily used tool in econometrics, the classical linear regression model. We began in Chapter 2 with a statement of the regression model. Chapter 3 then described computation of the parameters by least squares—a purely algebraic exercise. Chapter 4 reinterpreted least squares as an estimator of an unknown parameter vector and described the finite sample and large-sample characteristics of the sampling distribution of the estimator. Chapter 5 was devoted to building and sharpening the regression model, with statistical results for testing hypotheses about the underlying population. In this chapter, we have examined some broad issues related to model specification and selection of a model among a set of competing alternatives. The concepts considered here are tied very closely to one of the pillars of the paradigm of econometrics; underlying the model is a theoretical construction,

[20] See Sala-i-Martin (1997).

a set of true behavioral relationships that constitute *the model*. It is only on this notion that the concepts of bias and biased estimation and model selection make any sense—"bias" as a concept can only be described with respect to some underlying model against which an estimator can be said to be biased. That is, there must be a yardstick. This concept is a central result in the analysis of specification, where we considered the implications of underfitting (omitting variables) and overfitting (including **superfluous variables**) the model. We concluded this chapter (and our discussion of the classical linear regression model) with an examination of procedures that are used to choose among competing model specifications.

Key Terms and Concepts

- Acceptance region
- Adjusted R^2
- Akaike Information Criterion
- Alternative hypothesis
- Bayesian model averaging
- Bayesian Information Criterion
- Biased estimator
- Comprehensive model
- Consistent
- Distributed lag
- Discrepancy vector
- Encompassing principle
- Exclusion restrictions
- Functionally independent
- General nonlinear hypothesis
- General-to-simple strategy
- Superfluous variables
- *J* test
- Lack of invariance
- Lagrange multiplier tests
- Linear restrictions
- Model selection
- Nested
- Nested models
- Nominal size
- Nonlinear restrictions
- Nonnested
- Nonnested models
- Nonnormality
- Null hypothesis
- One-sided test
- Parameter space
- Power of a test
- Prediction criterion
- Rejection region
- Restricted least squares
- Schwarz criterion
- Simple-to-general
- Size of the test
- Specification test
- Stepwise model building
- *t* ratio
- Testable implications
- Wald criterion
- Wald distance
- Wald statistic
- Wald test

Exercises

1. A multiple regression of y on a constant x_1 and x_2 produces the following results: $\hat{y} = 4 + 0.4x_1 + 0.9x_2$, $R^2 = 8/60$, $\mathbf{e}'\mathbf{e} = 520$, $n = 29$,
$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 29 & 0 & 0 \\ 0 & 50 & 10 \\ 0 & 10 & 80 \end{bmatrix}.$$
Test the hypothesis that the two slopes sum to 1.
2. Using the results in Exercise 1, test the hypothesis that the slope on x_1 is 0 by running the restricted regression and comparing the two sums of squared deviations.
3. The regression model to be analyzed is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ have K_1 and K_2 columns, respectively. The restriction is $\boldsymbol{\beta}_2 = \mathbf{0}$.
   a. Using (5-23), prove that the restricted estimator is simply $[\mathbf{b}_{1*}, \mathbf{0}]$, where $\mathbf{b}_{1*}$ is the least squares coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}_1$.
   b. Prove that if the restriction is $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$ for a nonzero $\boldsymbol{\beta}_2^0$, then the restricted estimator of $\boldsymbol{\beta}_1$ is $\mathbf{b}_{1*} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2^0)$.

4. The expression for the restricted coefficient vector in (5-23) may be written in the form $\mathbf{b}_* = [\mathbf{I} - \mathbf{CR}]\mathbf{b} + \mathbf{w}$, where $\mathbf{w}$ does not involve $\mathbf{b}$. What is $\mathbf{C}$? Show that the covariance matrix of the restricted least squares estimator is

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

and that this matrix may be written as

$$\text{Var}[\mathbf{b}|\mathbf{X}]\{[\text{Var}(\mathbf{b}|\mathbf{X})]^{-1} - \mathbf{R}'[\text{Var}(\mathbf{Rb})|\mathbf{X}]^{-1}\mathbf{R}\}\,\text{Var}[\mathbf{b}|\mathbf{X}].$$

5. Prove the result that the restricted least squares estimator never has a larger covariance matrix than the unrestricted least squares estimator.
6. Prove the result that the R^2 associated with a restricted least squares estimator is never larger than that associated with the unrestricted least squares estimator. Conclude that imposing restrictions never improves the fit of the regression.
7. An alternative way to test the hypothesis $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ is to use a Wald test of the hypothesis that $\boldsymbol{\lambda}_* = \mathbf{0}$, where $\boldsymbol{\lambda}_*$ is defined in (5-23). Prove that

$$\chi^2 = \boldsymbol{\lambda}_*'\{\text{Est. Var}[\boldsymbol{\lambda}_*]\}^{-1}\boldsymbol{\lambda}_* = (n - K)\left[\frac{\mathbf{e}_*'\mathbf{e}_*}{\mathbf{e}'\mathbf{e}} - 1\right].$$

Note that the fraction in brackets is the ratio of two estimators of σ^2. By virtue of (5-28) and the preceding discussion, we know that this ratio is greater than 1. Finally, prove that this test statistic is equivalent to JF, where J is the number of restrictions being tested and F is the conventional F statistic given in (5-16). Formally, the Lagrange multiplier test requires that the variance estimator be based on the restricted sum of squares, not the unrestricted. Then, the test statistic would be $\text{LM} = nJ/[(n - K)/F + J]$.[21]
8. Use the test statistic defined in Exercise 7 to test the hypothesis in Exercise 1.
9. Prove that under the hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the estimator

$$s_*^2 = \frac{(\mathbf{y} - \mathbf{Xb}_*)'(\mathbf{y} - \mathbf{Xb}_*)}{n - K + J},$$

where J is the number of restrictions, is unbiased for σ^2.
10. Show that in the multiple regression of $\mathbf{y}$ on a constant, $\mathbf{x}_1$ and $\mathbf{x}_2$ while imposing the restriction $\beta_1 + \beta_2 = 1$ leads to the regression of $\mathbf{y} - \mathbf{x}_1$ on a constant and $\mathbf{x}_2 - \mathbf{x}_1$.
11. Suppose the true regression model is given by (4-7). The result in (4-9) shows that if $\mathbf{p}_{\mathbf{X}.\mathbf{z}}$ is nonzero and γ is nonzero, then regression of $\mathbf{y}$ on $\mathbf{X}$ alone produces a biased and inconsistent estimator of $\boldsymbol{\beta}$. Suppose the objective is to forecast $\mathbf{y}$, not to estimate the parameters. Consider regression of $\mathbf{y}$ on $\mathbf{X}$ alone to estimate $\boldsymbol{\beta}$ with $\mathbf{b}$ (which is biased). Is the forecast of $\mathbf{y}$ computed using $\mathbf{Xb}$ also biased? Assume that $E[z|\mathbf{X}]$ is a linear function of $\mathbf{X}$. Discuss your findings generally. What are the implications for prediction when variables are omitted from a regression?
12. The log likelihood function for the linear regression model with normally distributed disturbances is shown in (14-39) in Section 14.9.1. Show that at the maximum likelihood estimators of $\mathbf{b}$ for $\boldsymbol{\beta}$ and $\mathbf{e}'\mathbf{e}/n$ for σ^2, the log likelihood is an increasing function of R^2 for the model.

[21] See Godfrey (1988).

13. Show that the model of the alternative hypothesis in Example 5.7 can be written

$$H_1: C_t = \theta_1 + \theta_2 Y_t + \theta_3 Y_{t-1} + \sum_{s=2}^{\infty} \theta_{s+2} Y_{t-s} + \varepsilon_{it} + \sum_{s=1}^{\infty} \lambda_s \varepsilon_{t-s}.$$

As such, it does appear that H_0 is a restriction on H_1. However, because there are an infinite number of constraints, this does not reduce the test to a standard test of restrictions. It does suggest the connections between the two formulations.

Applications

1. The application in Chapter 3 used 15 of the 17,919 observations in Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. (See Appendix Table F3.2.) We will use the full data set for this exercise. The data may be downloaded from the *Journal of Applied Econometrics* data archive at http://www.econ.queensu.ca/jae/12004-v19.7/koop-tobias/. The data file is in two parts. The first file contains the panel of 17,919 observations on variables:

Column 1; *Person id* (ranging from 1 to 2,178),
Column 2; *Education,*
Column 3; *Log of hourly wage,*
Column 4; *Potential experience,*
Column 5; *Time trend.*

Columns 2 through 5 contain time varying variables. The second part of the data set contains time invariant variables for the 2,178 households. These are

Column 1; *Ability,*
Column 2; *Mother's education,*
Column 3; *Father's education,*
Column 4; *Dummy variable for residence in a broken home,*
Column 5; *Number of siblings.*

To create the data set for this exercise, it is necessary to merge these two data files. The *i*th observation in the second file will be replicated T_i times for the set of T_i observations in the first file. The *person id* variable indicates which rows must contain the data from the second file. (How this preparation is carried out will vary from one computer package to another.) (*Note:* We are not attempting to replicate Koop and Tobias's results here—we are only employing their interesting data set.) Let $\mathbf{X}_1$ = [*constant*, *education*, *experience*, *ability*] and let $\mathbf{X}_2$ = [*mother's education*, *father's education*, *broken home*, *number of siblings*].

a. Compute the full regression of ln *wage* on $\mathbf{X}_1$ and $\mathbf{X}_2$ and report all results.
b. Use an F test to test the hypothesis that all coefficients except the constant term are zero.
c. Use an F statistic to test the joint hypothesis that the coefficients on the four household variables in $\mathbf{X}_2$ are zero.
d. Use a Wald test to carry out the test in part c.
e. Use a Lagrange multiplier test to carry out the test in part c.

2. The generalized Cobb–Douglas cost function examined in Application 2 in Chapter 4 is a special case of the translog cost function,

$$\begin{aligned}\ln C = {} & \alpha + \beta \ln Q + \delta_k \ln P_k + \delta_l \ln P_l + \delta_f \ln P_f \\ & + \phi_{kk}[\tfrac{1}{2}(\ln P_k)^2] + \phi_{ll}[\tfrac{1}{2}(\ln P_l)^2] + \phi_{ff}[\tfrac{1}{2}(\ln P_f)^2] \\ & + \phi_{kl}[\ln P_k][\ln P_l] + \phi_{kf}[\ln P_k][\ln P_f] + \phi_{lf}[\ln P_l][\ln P_f] \\ & + \gamma[\tfrac{1}{2}(\ln Q)^2] \\ & + \theta_{Qk}[\ln Q][\ln P_k] + \theta_{Ql}[\ln Q][\ln P_l] + \theta_{Qf}[\ln Q][\ln P_f] + \varepsilon.\end{aligned}$$

The theoretical requirement of linear homogeneity in the factor prices imposes the following restrictions:

$$\begin{aligned}&\delta_k + \delta_l + \delta_f = 1, \qquad \phi_{kk} + \phi_{kl} + \phi_{kf} = 0, \quad \phi_{kl} + \phi_{ll} + \phi_{lf} = 0, \\ &\phi_{kf} + \phi_{lf} + \phi_{ff} = 0, \quad \theta_{QK} + \theta_{Ql} + \theta_{Qf} = 0.\end{aligned}$$

Note that although the underlying theory requires it, the model can be estimated (by least squares) without imposing the linear homogeneity restrictions. [Thus, one could test the underlying theory by testing the validity of these restrictions. See Christensen, Jorgenson, and Lau (1975).] We will repeat this exercise in part b.

A number of additional restrictions were explored in Christensen and Greene's (1976) study. The hypothesis of homotheticity of the production structure would add the additional restrictions

$$\theta_{Qk} = 0, \quad \theta_{Ql} = 0, \quad \theta_{Qf} = 0.$$

Homogeneity of the production structure adds the restriction $\gamma = 0$. The hypothesis that all elasticities of substitution in the production structure are equal to -1 is imposed by the six restrictions $\phi_{ij} = 0$ for all i and j.

We will use the data from the earlier application to test these restrictions. For the purposes of this exercise, denote by $\beta_1, \ldots, \beta_{15}$ the 15 parameters in the cost function above in the order that they appear in the model, starting in the first line and moving left to right and downward.

a. Write out the **R** matrix and **q** vector in (5-8) that are needed to impose the restriction of linear homogeneity in prices.
b. Test the theory of production using all 158 observations. Use an F test to test the restrictions of linear homogeneity. Note, you can use the general form of the F statistic in (5-16) to carry out the test. Christensen and Greene enforced the linear homogeneity restrictions by building them into the model. You can do this by dividing cost and the prices of capital and labor by the price of fuel. Terms with f subscripts fall out of the model, leaving an equation with 10 parameters. Compare the sums of squares for the two models to carry out the test. Of course, the test may be carried out either way and will produce the same result.
c. Test the hypothesis homotheticity of the production structure under the assumption of linear homogeneity in prices.
d. Test the hypothesis of the generalized Cobb–Douglas cost function in Chapter 4 against the more general translog model suggested here, once again (and henceforth) assuming linear homogeneity in the prices.

e. The simple Cobb–Douglas function appears in the first line of the model above. Test the hypothesis of the Cobb–Douglas model against the alternative of the full translog model.
f. Test the hypothesis of the generalized Cobb–Douglas model against the homothetic translog model.
g. Which of the several functional forms suggested here do you conclude is the most appropriate for these data?

3. The gasoline consumption model suggested in part d of Application 1 in Chapter 4 may be written as

$$\ln(G/Pop) = \alpha + \beta_P \ln P_g + \beta_I \ln (Income/Pop) + \gamma_{nc} \ln P_{nc} + \gamma_{uc} \ln P_{uc} + \gamma_{pt} \ln P_{pt} + \tau\, year + \delta_d \ln P_d + \delta_n \ln P_n + \delta_s \ln P_s + \varepsilon.$$

a. Carry out a test of the hypothesis that none of the three aggregate price indices are significant determinants of the demand for gasoline.
b. Consider the hypothesis that the microelasticities are a constant proportion of the elasticity with respect to their corresponding aggregate. Thus, for some positive θ (presumably between 0 and 1), $\gamma_{nc} = \theta\delta_d$, $\gamma_{uc} = \theta\delta_d$, $\gamma_{pt} = \theta\delta_s$. The first two imply the simple linear restriction $\gamma_{nc} = \gamma_{uc}$. By taking ratios, the first (or second) and third imply the nonlinear restriction

$$\frac{\gamma_{nc}}{\gamma_{pt}} = \frac{\delta_d}{\delta_s} \quad \text{or} \quad \gamma_{nc}\delta_s - \gamma_{pt}\delta_d = 0.$$

Describe in detail how you would test the validity of the restriction.
c. Using the gasoline market data in Table F2.2, test the two restrictions suggested here, separately and jointly.

4. The J test in Example 5.7 is carried out using more than 50 years of data. It is optimistic to hope that the underlying structure of the economy did not change in 50 years. Does the result of the test carried out in Example 5.7 persist if it is based on data only from 1980 to 2000? Repeat the computation with this subset of the data.

6

FUNCTIONAL FORM, DIFFERENCE IN DIFFERENCES, AND STRUCTURAL CHANGE

6.1 INTRODUCTION

This chapter will examine a variety of ways that the linear regression model can be adapted for particular situations and specific features of the environment. Section 6.2 begins by using binary variables to accommodate nonlinearities and discrete shifts in the model. Sections 6.3 and 6.4 examine two specific forms of the linear model that are suited for analyzing causal impacts of policy changes, **difference in differences** models and **regression kink** and **regression discontinuity designs**. Section 6.5 broadens the class of models that are linear in the parameters. By using logarithms, quadratic terms, and **interaction terms** (products of variables), the regression model can accommodate a wide variety of functional forms in the data. Section 6.6 examines the issue of specifying and testing for discrete change in the underlying process that generates the data, under the heading of structural change. In a time-series context, this relates to abrupt changes in the economic environment, such as major events in financial markets (e.g., the world financial crisis of 2007–2008) or commodity markets (such as the several upheavals in the oil market). In a cross section, we can modify the regression model to account for discrete differences across groups such as different preference structures or market experiences of men and women.

6.2 USING BINARY VARIABLES

One of the most useful devices in regression analysis is the **binary**, or **dummy variable**. A dummy variable takes the value one for some observations to indicate the presence of an effect or membership in a group and zero for the remaining observations. Binary variables are a convenient means of building discrete shifts of the function into a regression model.

6.2.1 BINARY VARIABLES IN REGRESSION

Dummy variables are usually used in regression equations that also contain other quantitative variables,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \gamma d_i + \varepsilon_i, \tag{6-1}$$

TABLE 6.1 Estimated Earnings Equation

$\ln earnings = \beta_1 + \beta_2 Age + \beta_3 Age^2 + \beta_4 Education + \beta_5 Kids + \varepsilon$

Sum of squared residuals:	599.4582
Standard error of the regression:	1.19044
R^2 based on 428 observations:	0.040995

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age^2	−0.002315	0.000987	−2.345
Education	0.067472	0.025248	2.672
Kids	−0.35119	0.14753	−2.380

where $d_i = 1$ for some condition occurring, and 0 if not.[1] In the earnings equation in Example 5.2, we included a variable *Kids* to indicate whether there were children in the household, under the assumption that for many married women, this fact is a significant consideration in labor supply decisions. The results shown in Example 6.1 appear to be consistent with this hypothesis.

Example 6.1 Dummy Variable in an Earnings Equation

Table 6.1 reproduces the estimated earnings equation in Example 5.2. The variable *Kids* is a dummy variable that equals one if there are children under 18 in the household and zero otherwise. Because this is a semilog equation, the value of −0.35 for the coefficient is an extremely large effect, one which suggests that all other things equal, the earnings of women with children are nearly a third less than those without. This is a large difference, but one that would certainly merit closer scrutiny. Whether this effect results from different labor market effects that influence wages and not hours, or the reverse, remains to be seen. Second, having chosen a nonrandomly selected sample of those with only positive earnings to begin with, it is unclear whether the sampling mechanism has, itself, induced a bias in the estimator of this parameter.

Dummy variables are particularly useful in loglinear regressions. In a model of the form

$$\ln y = \beta_1 + \beta_2 x + \beta_3 d + \varepsilon,$$

the coefficient on the dummy variable, d, indicates a multiplicative shift of the function. The percentage change in $E[y|x, d]$ asociated with the change in d is

$$\begin{aligned}
\%(\Delta E[y|x,d]/\Delta d) &= 100\%\left\{\frac{E[y|x,d=1] - E[y|x,d=0]}{E[y|x,d=0]}\right\} \\
&= 100\%\left\{\frac{\exp(\beta_1 + \beta_2 x + \beta_3)E[\exp(\varepsilon)] - \exp(\beta_1 + \beta_2 x)\,E[\exp(\varepsilon)]}{\exp(\beta_1 + \beta_2 x)E[\exp(\varepsilon)]}\right\} \\
&= 100\%[\exp(\beta_3) - 1]. \qquad \textbf{(6-2)}
\end{aligned}$$

[1] We are assuming at this point (and for the rest of this chapter) that the dummy variable in (6-1) is *exogenous*. That is, the assignment of values of the dummy variable to observations in the sample is unrelated to ε_i. This is consistent with the sort of random assignment to treatment designed in a clinical trial. The case in which d_i is endogenous would occur, for example, when individuals select the value of d_i themselves. Analyses of the effects of program participation, such as job training on wages or agricultural extensions on productivity, would be examples. The endogenous treatment effect model is examined in Section 8.5.

Example 6.2 Value of a Signature

In Example 4.10 we explored the relationship between log of sale price and surface area for 430 sales of Monet paintings. Regression results from the example are shown in Table 6.2. The results suggest a strong relationship between area and price—the coefficient is 1.33372, indicating a highly elastic relationship, and the *t* ratio of 14.70 suggests the relationship is highly significant. A variable (effect) that is clearly left out of the model is the effect of the artist's signature on the sale price. Of the 430 sales in the sample, 77 are for unsigned paintings. The results at the right of Table 6.2 include a dummy variable for whether the painting is signed or not. The results show an extremely strong effect. The regression results imply that

$$E[Price \mid Area, Aspect\ Ratio, Signature) = \exp[-9.64 + 1.35 \ln Area - 0.0\,8\, Aspect\ Ratio + 1.23\, Signature + 0.993^2/2].$$

(See Section 4.8.2.) Computing this result for a painting of the same area and aspect ratio, we find the model predicts that the signature effect would be

$$100\% \times \frac{\Delta E[Price]}{Price} = 100\%[\exp(1.26) - 1] = 252\%.$$

The effect of a signature on an otherwise similar painting is to more than double the price. The estimated standard error for the signature coefficient is 0.1253. Using the delta method, we obtain an estimated standard error for $[\exp(b_3) - 1]$ of the square root of $[\exp(b_3)]^2 \times 0.1253^2$, which is 0.4417. For the percentage difference of 252%, we have an estimated standard error of 44.17%.

Superficially, it is possible that the size effect we observed earlier could be explained by the presence of the signature. If the artist tended on average to sign only the larger paintings, then we would have an explanation for the counterintuitive effect of size. (This would be an example of the effect of multicollinearity of a sort.) For a regression with a continuous variable and a dummy variable, we can easily confirm or refute this proposition. The average size for the 77 sales of unsigned paintings is 1,228.69 square inches. The average size of the other 353 is 940.812 square inches. There does seem to be a substantial systematic difference between signed and unsigned paintings, but it goes in the other direction. We are left with significant findings of both a size and a signature effect in the auction prices of Monet paintings. *Aspect Ratio*, however, appears still to be inconsequential.

TABLE 6.2 Estimated Equations for Log Price

$\ln price = \beta_1 + \beta_2 \ln Area + \beta_3 Aspect\ Ratio + \beta_4 Signature + \varepsilon$

Mean of ln *Price*	0.33274					
Number of observations	430					
Sum of squared residuals	520.765			420.609		
Standard error	1.10435			1.35024		
R-squared	0.33417			0.46223		
Adjusted *R*-squared	0.33105			0.45844		
Variable	***Coefficient***	***Standard Error***	***t Ratio***	***Coefficient***	***Standard Error***	***t Ratio***
Constant	−8.34327	0.67820	−12.30	−9.65443	0.62397	−15.47
ln *Area*	1.31638	0.09205	14.30	1.34379	0.08787	16.22
Aspect ratio	−0.09623	0.15784	−0.61	−0.01966	0.14222	−0.14
Signature	—	—	—	1.26090	0.12519	10.07

Example 6.3 Gender and Time Effects in a Log Wage Equation

Cornwell and Rupert (1988) examined the returns to schooling in a panel data set of 595 heads of households observed in seven years, 1976-1982. The sample data (Appendix Table F8.1) are drawn from years 1976 to 1982 from the "Non-Survey of Economic Opportunity" from the Panel Study of Income Dynamics. A prominent result that appears in different specifications of their regression model is a persistent difference between wages of female and male heads of households. A slightly modified version of their regression model is

$$\ln Wage_{it} = \beta_1 + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \beta_{11} Ed_i + \beta_{12} Fem_i + \sum_{t=1977}^{1982} \gamma_t D_{it} + \varepsilon_{it}.$$

The variables in the model are listed in Example 4.6. (See Appendix Table F8.1 for the data source.)

Least squares estimates of the log wage equation appear at the left side in Table 6.3. Because these data are a panel, it is likely that observations within each group are correlated. The table reports cluster corrected standard errors, based on (4-42). The coefficient on

TABLE 6.3 Estimated Log Wage Equations

	Aggregate Effect			*Individual Fixed Effects*		
Sum of squares	391.056			81.5201		
Residual std. error	0.30708			0.15139		
R-squared	0.55908			0.90808		
Observations	4165			595 × 7		
F[17,577]	1828.50					
	Coefficient	***Clustered Std.Error***	***t Ratio***	***Coefficient***	***Clustered Std.Error***	***t Ratio***
Constant	5.08397	0.12998	39.11	Individual Fixed Effects		
EXP	0.03128	0.00419	7.47	0.10370	0.00691	15.00
EXP^2	−0.00055	0.00009	−5.86	−0.00040	0.00009	−4.43
WKS	0.00394	0.00158	2.50	0.00068	0.00095	0.72
OCC	−0.14116	0.02687	−5.25	−0.01916	0.02033	−0.94
IND	0.05661	0.02343	2.42	0.02076	0.02422	0.86
SOUTH	−0.07180	0.02632	−2.73	0.00309	0.09620	0.03
SMSA	0.15423	0.02349	6.57	−0.04188	0.03133	−1.34
MS	0.09634	0.04301	2.24	−0.02857	0.02887	−0.99
UNION	0.08052	0.02335	3.45	0.02952	0.02689	1.10
ED	0.05499	0.00556	9.88	—	—	—
FEM	−0.36502	0.04829	−7.56	—	—	—
Year(Base = 1976)						
1977	0.07461	0.00601	12.42	—	—	—
1978	0.19611	0.00989	19.82	0.04107	0.01267	3.24
1979	0.28358	0.01016	27.90	0.05170	0.01662	3.11
1980	0.36264	0.00985	36.82	0.05518	0.02132	2.59
1981	0.43695	0.01133	38.58	0.04612	0.02718	1.70
1982	0.52075	0.01211	43.00	0.04650	0.03254	1.43

FEM is −0.36502. Using (6-2), this translates to a roughly 100%[exp(−0.365) − 1] = 31% wage differential. Because the data are a panel, it is quite likely that the disturbances are correlated across the years within a household. Thus, robust standard errors are reported in Table 6.3. The effect of the adjustment is substantial. The conventional standard error for *FEM* based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is 0.02201—less than half the reported value of 0.04829. Note the reported denominator degrees of freedom for the model *F* statistic is 595 − 18 = 577. Given that observations within a unit are not independent, it seems that 4147 would overstate the degrees of freedom. The number of groups of 595 is the natural alternative number of observations. However, if this were the case, then the statistic reported, computed as if there were 4165 observations, would not have an *F* distribution. This remains as an ambiguity in the computation of robust statistics. As we will pursue in Chapter 8, there is yet another ambiguity in this equation. It seems likely unobserved factors that influence ln *Wage* (in ε_{it}) (e.g., ability) might also be influential in the level of education. If so (i.e., if Ed_i is correlated with ε_{it}), then least squares might not be an appropriate method of estimation of the parameters in this model.

It is common for researchers to include a dummy variable in a regression to account for something that applies only to a single observation. For example, in time-series analyses, an occasional study includes a dummy variable that is one only in a single unusual year, such as the year of a major strike or a major policy event. (See, for example, the application to the German money demand function in Section 21.3.5.) It is easy to show (we consider this in the exercises) the very useful implication of this:

A dummy variable that takes the value one only for one observation has the effect of deleting that observation from computation of the least squares slopes and variance estimator (but not from R-squared).

6.2.2 SEVERAL CATEGORIES

When there are several categories, a set of binary variables is necessary. Correcting for seasonal factors in macroeconomic data is a common application. We could write a consumption function for quarterly data as

$$C_t = \beta_1 + \beta_2 x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \varepsilon_t,$$

where x_t is disposable income. Note that only three of the four quarterly dummy variables are included in the model. If the fourth were included, then the four dummy variables would sum to one at every observation, which would replicate the constant term—a case of perfect multicollinearity. This is known as the **dummy variable trap**. To avoid the dummy variable trap, we drop the dummy variable for the fourth quarter. (Depending on the application, it might be preferable to have four separate dummy variables and drop the overall constant.[2]) Any of the four quarters (or 12 months) can be used as the base period.

The preceding is a means of *deseasonalizing* the data. Consider the alternative formulation:

$$C_t = \beta x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \varepsilon_t.$$

[2] See Suits (1984) and Greene and Seaks (1991).

Using the results from Section 3.3 on partitioned regression, we know that the preceding multiple regression is equivalent to first regressing C and x on the four dummy variables and then using the residuals from these regressions in the subsequent regression of deseasonalized consumption on deseasonalized income. Clearly, deseasonalizing in this fashion prior to computing the simple regression of consumption on income produces the same coefficient on income (and the same vector of residuals) as including the set of dummy variables in the regression.

Example 6.4 Genre Effects on Movie Box Office Receipts

Table 4.10 in Example 4.12 presents the results of the regression of log of box office receipts in 2009 for 62 movies on a number of variables including a set of dummy variables for four genres: *Action*, *Comedy*, *Animated*, or *Horror*. The left out category is "any of the remaining 9 genres" in the standard set of 13 that is usually used in models such as this one.[3] The four coefficients are −0.869, −0.016, −0.833, and +0.375, respectively. This suggests that, save for horror movies, these genres typically fare substantially worse at the box office than other types of movies. We note the use of b directly to estimate the percentage change for the category, as we did in Example 6.1 when we interpreted the coefficient of −0.35 on *Kids* as indicative of a 35% change in income. This is an approximation that works well when b is close to zero but deteriorates as it gets far from zero. Thus, the value of −0.869 above does not translate to an 87% difference between *Action* movies and other movies. Using (6-2), we find an estimated difference closer to 100% [exp(−0.869)−1] or about 58%. Likewise, the −0.35 result in Example 6.1 corresponds to an effect of about 29%.

6.2.3 MODELING INDIVIDUAL HETEROGENEITY

In the previous examples, a dummy variable is used to account for a specific event or feature of the observation or the environment, such as whether a painting is signed or not or the season. When the sample consists of repeated observations on a large number of entities, such as the 595 individuals in Example 6.3, a strategy often used to allow for unmeasured (and unnamed) fixed individual characteristics (effects) is to include a full set of dummy variables in the equation, one for each individual. To continue Example 6.3, the extended equation would be

$$\begin{aligned}\ln Wage_{it} = {} & \beta_1 + \sum_{i=1}^{595}\alpha_i A_{it} + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \\ & \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \\ & \beta_{11} Ed_{it} + \beta_{12} Fem_{it} + \sum_{t=1977}^{1982}\gamma_t D_{it} + \varepsilon_{it},\end{aligned}$$

where A_{it} equals one for individual i in every period and zero otherwise. The unobserved effect, α_i, in an earnings model could include factors such as ability, general skill, motivation, and fundamental experience. This model would contain the 12 variables from earlier plus the six time dummy variables for the periods, plus the 595 dummy variables for the individuals. There are some distinctive features of this model to be considered before it can be estimated.

- Because the full set of time dummy variables, D_{it}, $t = 1976, \ldots, 1982$, sums to 1 at every observation, which would replicate the constant term, one of them is dropped—1976 is

[3] Authorities differ a bit on this list. From the MPAA, we have Drama, Romance, Comedy, Action, Fantasy, Adventure, Family, Animated, Thriller, Mystery, Science Fiction, Horror, Crime.

identified as the "base year" in the results in Table 6.3. This avoids a multicollinearity problem known as the dummy variable trap.[4] The same problem will arise with the set of individual dummy variables, A_{it}, $i = 1, \ldots, 595$. The obvious remedy is to drop one of the effects, say the last one. An equivalent strategy that is usually used is to drop the overall constant term, leaving the "fixed effects" form of the model,

$$\begin{aligned}\ln Wage_{it} = &\sum_{i=1}^{595}\alpha_i A_{it} + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \beta_6 Ind_{it} + \\ &\beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \beta_{11} Ed_{it} + \\ &\beta_{12} Fem_{it} + \sum_{t=1977}^{1982}\gamma_t D_{it} + \varepsilon_{it}\end{aligned}$$

(This is a application of Theorem 3.8.) Note that this does not imply that the base year time dummy variable should now be restored. If so, the dummy variable trap would reappear as

$$\sum_{i=1}^{595} A_{it} = \sum_{t=1976}^{1982} D_{it}.$$

In a model that contains a set of fixed individual effects, it is necessary either to drop the overall constant term or one of the effects.

- There is another subtle multicollinearity problem in this model. The variable Fem_{it} does not change within the block of 7 observations for individual i—it is either 1 or 0 in all 7 years for each person. Let the matrix $\mathbf{A}$ be the 4165×595 matrix in which the ith column contains $\mathbf{a}_i$, the dummy variable for individual i. Let **fem** be the 4165×1 vector that contains the variable Fem_{it}; **fem** is the column of the full data matrix that contains FEM_{it}. In the block of seven rows for individual i, the 7 elements of **fem** are all 1 or 0 corresponding to Fem_{it}. Finally, let the 595×1 vector $\mathbf{f}$ equal 1 if individual i is female and 0 if male. Then, it is easy to see that $\mathbf{fem} = \mathbf{Af}$. That is, the column of the data matrix that contains Fem_{it} is a linear combination of the individual dummy variables, again, a multicollinearity problem. This is a general result:

  In a model that contains a full set of N individual effects represented by a set of N dummy variables, any other variable in the model that takes the same value in every period for every individual can be written as a linear combination of those effects.

  This means that the coefficient on Fem_{it} cannot be estimated. The natural remedy is to fix that coefficient at zero—that is, to drop that variable. In fact, the education variable, ED_{it}, has the same characteristic and must also be dropped from the model. This turns out to be a significant disadvantage of this formulation of the model for data such as these. Indeed, in this application, the gender effect was of particular interest. (We will examine the model with individual heterogeneity modeled as fixed effects in greater detail in Chapter 11.)

[4] A second time dummy variable is dropped in the model results on the right-hand side of Table 6.3. This is a result of another dummy variable trap that is specific to this application. The experience variable, *EXP*, is a simple count of the number of years of experience, starting from an individual specific value. For the first individual in the sample, $EXP_{1,t} = 3, \ldots, 9$ while for the second, it is $EXP_{2,t} = 30, \ldots, 36$. With the individual specific constants and the six time dummy variables, it is now possible to reproduce $EXP_{i,t}$ as a linear combination of these two sets of dummy variables. For example, for the first person, $EXP_{1,1} = 3\times A_{1,1}$; $EXP_{1,2} = 3\times A_{1,2} + D_{1,1978}$; $EXP_{1,3} = 3\times A_{1,3} + 2D_{1,1979}$; $EXP_{1,4} = 3\times A_{1,3} + 3D_{1,1980}$ and so on. So, each value EXP_{it} can be produced as a linear combination of A_{it} and one of the D_{it}'s. Dropping a second period dummy variable interrupts this result.

- The model with N individual effects has become very unwieldy. The wage equation now has more than 600 variables in it; later we will analyze a similar data set with more than 7,000 individuals. One might question the practicality of actually doing the computations. This particular application shows the power of the Frisch–Waugh result, Theorem 3.2—the computation of the regression is equally straightforward whether there are a few individuals or millions. To see how this works, write the log wage equation as

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

We are not necessarily interested in the specific constants α_i, but they must appear in the equation to control for the individual unobserved effects. Assume that there are no invariant variables such as FEM_{it} in $\mathbf{x}_{it}$. The mean of the observations for individual i is

$$\bar{y}_i = \frac{1}{7}\sum_{t=1976}^{1982} y_{it} = \alpha_i + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + \bar{\varepsilon}_i.$$

A strategy for estimating $\boldsymbol{\beta}$ without having to worry about α_i is to transform the data using simple deviations from group means:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

This transformed model can be estimated by least squares. All that is necessary is to transform the data beforehand. This computation is automated in all modern software. (Details of the theoretical basis of the computation are considered in Chapter 11.)

> *To compute the least squares estimates of the coefficients in a model that contains N dummy variables for individual fixed effects, the data are transformed to deviations from individual means, then simple least squares is used based on the transformed data. (Time dummy variables are transformed as well.) Standard errors are computed in the ways considered earlier, including robust standard errors for heteroscedasticity. Correcting for clustering within the groups would be natural.*

Notice what becomes of a variable such as FEM when we compute $(x_{it} - \bar{x}_i)$. Because FEM and ED take the same value in every period, the group mean is that value, and the deviations from the means becomes zero at every observation. The regression cannot be computed if $\mathbf{X}$ contains any columns of zeros. Finally, for some purposes, we might be interested in the estimates of the individual effects, α_i. We can show using Theorem 3.2 that the least squares coefficients on A_{it} in the original model would be $a_i = \bar{y}_i - \bar{\mathbf{x}}_i'\mathbf{b}$.

Results of the fixed effects regression are shown at the right in Table 6.3. Accounting for individual effects in this fashion often produces quite substantial changes in the results. Notice that the fit of the model, measured by R^2, improves dramatically. The effect of $UNION$ membership, which was large and significant before has essentially vanished. And, unfortunately, we have lost view of the gender and education effects.

Example 6.5 Sports Economics: Using Dummy Variables for Unobserved Heterogeneity[5]

In 2000, the Texas Rangers major league baseball team signed 24-year-old Alex Rodriguez (A-Rod), who was claimed at the time to be "the best player in baseball," to the largest contract in baseball history (up to that time). It was publicized to be some $25Million/year for

[5] This application is based on Cohen, R. and Wallace, J., "A-Rod: Signing the Best Player in Baseball," *Harvard Business School*, Case 9-203-047, Cambridge, 2003.

10 years, or roughly a quarter of a billion dollars.[6] Treated as a capital budgeting decision, the investment is complicated partly because of the difficulty of valuing the benefits of the acquisition. Benefits would consist mainly of more fans in the stadiums where the team played, more valuable broadcast rights, and increased franchise value. We (and others) consider the first of these. It was projected that A-Rod could help the team win an average of 8 more games per season and would surely be selected as an All-Star every year. How do 8 additional wins translate into a marginal value for the investors? The franchise value and broadcast rights are highly speculative. But there is a received literature on the relationship between team wins and game attendance, which we will use here.[7] The final step will then be to calculate the value of the additional attendance.

Appendix Table F6.5 contains data on attendance, salaries, games won, and several other variables for 30 teams observed from 1985 to 2001. (These are *panel data*. We will examine this subject in greater detail in Chapter 11.) We consider a **dynamic linear regression model**,

$$Attendance_{i,t} = \Sigma_i \alpha_i A_{i,t} + \gamma Attendance_{i,t-1} + \beta_1 Wins_{i,t} + \beta_2 Wins_{i,t-1} + \beta_3 All\ Stars_{i,t} + \varepsilon_{i,t},$$
$$i = 1, \ldots, 30; t = 1985, \ldots, 2001.$$

The previous year's attendance and wins are loyalty effects. The model contains a separate constant term for each team. The effect captured by α_i includes the size of the market and any other unmeasured time constant characteristics of the market.

The team specific dummy variable, $A_{i,t}$ is used to model unit specific **unobserved heterogeneity**. We will revisit this modeling aspect in Chapter 11. The setting is different here in that in the panel data context in Chapter 11, the sampling framework will be with respect to units "*i*" and statistical properties of estimators will refer generally to increases in the number of units. Here, the number of units (teams) is fixed at 30, and asymptotic results would be based on additional years of data.[8]

Table 6.4 presents the regression results for the dynamic model. Results are reported with and without the separate team effects. Standard errors for the estimated coefficients are adjusted for the clustering of the observations by team. The *F* statistic for $H_0: \alpha_i = \alpha$, $i = 1, \ldots, 31$ is computed as

$$F[30,401] = \frac{(23.267 - 20.254)/30}{20.254/401} = 1.988$$

The 95% critical value for $F[30,401]$ is 1.49 so the hypothesis of no separate team effects is rejected. The individual team effects appear to improve the model—note the peculiar negative loyalty effect in the model without the team effects.

In the dynamic equation, the long run equilibrium attendance would be

$$Attendance^* = (\alpha_i + \beta_1 Wins^* + \beta_2 Wins^* + \beta_3 All\ Stars^*)/(1 - \gamma).$$

(See Section 11.11.3.) The marginal value of winning one more game every year would be $(\beta_1 + \beta_2)/(1 - \gamma)$. The effect of winning 8 more games per year and having an additional All-Star on the team every year would be

$$(8(\beta_1 + \beta_2) + \beta_3)/(1 - \gamma) \times 1 \text{ million} = 268{,}270 \text{ additional fans/season.}$$

[6] Though it was widely reported to be a 10-year arrangement, the payout was actually scheduled over more than 20 years, and much of the payment was deferred until the latter years. A realistic present discounted value at the time of the signing would depend heavily on assumptions, but using the 8% standard at the time, would be roughly $160M, not $250M.

[7] See, for example, *The Journal of Sports Economics* and Lemke, Leonard, and Tlhokwane (2009).

[8] There are 30 teams in the data set, but one of the teams changed leagues. This team is treated as two observations.

TABLE 6.4 Estimated Attendance Model

Mean of Attendance	2.22048 Million					
Number of observations	437 (31 Teams)					
	No Team Effects			Team Effects		
Sum of squared residuals	23.267			20.254		
Standard error	0.23207			0.24462		
R-squared	0.74183			0.75176		
Adjusted *R*-squared	0.73076			0.71219		
Variable	***Coefficient***	***Standard Error****	***t Ratio***	***Coefficient***	***Standard Error****	***t Ratio***
$Attendance_{t-1}$	0.70233	0.03507	20.03	0.54914	0.02760	16.76
Wins	0.00992	0.00147	6.75	0.01109	0.00157	7.08
$Wins_{t-1}$	−0.00051	0.00117	−0.43	0.00220	0.00100	2.20
All stars	0.02125	0.01241	1.71	0.01459	0.01402	1.04
Constant	−1.20827	0.87499	−1.38	Individual Team Effects		

*Standard errors clustered at the team level.

In this case, the calculation of monetary value is 268,270 fans times $50 per fan (possibly somewhat high) or about $13.0 million against the cost of roughly $18 to $20 million per season.

6.2.4 SETS OF CATEGORIES

The case in which several sets of dummy variables are needed is much the same as those we have already considered, with one important exception. Consider a model of statewide per capita expenditure on education, y, as a function of statewide per capita income, x. Suppose that we have observations on all $n = 50$ states for $T = 10$ years. A regression model that allows the expected expenditure to change over time as well as across states would be

$$y_{it} = \alpha + \beta x_{it} + \delta_i + \theta_t + \varepsilon_{it}.$$

As before, it is necessary to drop one of the variables in each set of dummy variables to avoid the dummy variable trap. For our example, if a total of 50 state dummies and 10 time dummies is retained, a problem of *perfect multicollinearity* remains; the sums of the 50 state dummies and the 10 time dummies are the same, that is, 1. One of the variables in each of the sets (or the overall constant term and one of the variables in one of the sets) must be omitted.

Example 6.6 Analysis of Covariance

The data in Appendix Table F6.1 were used in a study of efficiency in production of airline services in Greene (2007a). The airline industry has been a favorite subject of study [e.g., Schmidt and Sickles (1984); Sickles, Good, and Johnson (1986)], partly because of interest in this rapidly changing market in a period of deregulation and partly because of an abundance of large, high-quality data sets collected by the (no longer existent) Civil Aeronautics Board. The original data set consisted of 25 firms observed yearly for 15 years (1970 to 1984), a "balanced panel." Several of the firms merged during this period and several others experienced strikes, which reduced the number of complete observations substantially. Omitting these and others

TABLE 6.5 *F* Tests for Firm and Year Effects

Model	*Sum of Squares*	*Restrictions on Full Model*	*F*	*Degrees of Freedom*
Full model	0.17257	0	—	
Time effects only	1.03470	5	65.94	[5, 66]
Firm effects only	0.26815	14	2.61	[14, 66]
No effects	1.27492	19	22.19	[19, 66]

because of missing data on some of the variables left a group of 10 full observations, from which we have selected 6 for the example to follow. We will fit a cost equation of the form

$$\ln C_{i,t} = \beta_1 + \beta_2 \ln Q_{i,t} + \beta_3 \ln^2 Q_{i,t} + \beta_4 \ln P_{fuel,i,t} + \beta_5 \, Load\ Factor_{i,t} + \sum_{t=1}^{14} \theta_t D_{i,t} + \sum_{i=1}^{5} \delta_i F_{i,t} + \varepsilon_{i,t}.$$

The dummy variables are $D_{i,t}$, which is the year variable, and $F_{i,t}$, which is the firm variable. We have dropped the first one in each group. The estimated model for the full specification is

$$\ln C_{i,t} = 12.89 + 0.8866 \ln Q_{i,t} + 0.01261 \ln^2 Q_{i,t} + 0.1281 \ln P_{fuel,i,t} - 0.8855 \, Load\ Factor_{i,t} + \text{time effects} + \text{firm effects} + e_{i,t}.$$

We are interested in whether the firm effects, the time effects, both, or neither are statistically significant. Table 6.5 presents the sums of squares from the four regressions. The *F* statistic for the hypothesis that there are no firm-specific effects is 65.94, which is highly significant. The statistic for the time effects is only 2.61, which is also larger than the critical value of 1.84. In the absence of the year-specific dummy variables, the year-specific effects are probably largely absorbed by the price of fuel.

6.2.5 THRESHOLD EFFECTS AND CATEGORICAL VARIABLES

In most applications, we use dummy variables to account for purely qualitative factors, such as membership in a group, or to represent a particular time period. There are cases, however, in which the dummy variable(s) represents levels of some underlying factor that might have been measured directly if this were possible. For example, education is a case in which we often observe certain thresholds rather than, say, years of education. Suppose, for example, that our interest is in a regression of the form

$$Earnings = \beta_1 + \beta_2 \, Age + Effect\ of\ Education + \varepsilon.$$

The data on education might consist of the highest level of education attained, such as less than high school (*LTHS*), high school (*HS*), college (*C*), post graduate (*PG*). An obviously unsatisfactory way to proceed is to use a variable, *E*, that is 0 for the first group, 1 for the second, 2 for the third, and 3 for the fourth. That would be $Earnings = \beta_1 + \beta_2 \, Age + \beta_3 E + \varepsilon$. The difficulty with this approach is that it assumes that the increment in income at each threshold is the same; β_3 is the difference between income with post graduate study and college and between college and high school.[9]

[9] One might argue that a regression model based on years of education instead of this sort of step function would be likewise problematic. It seems natural that in most cases, the 12th year of education (with graduation) would be far more valuable than the 11th.

This is unlikely and unduly restricts the regression. A more flexible model would use three (or four) binary variables, one for each level of education. Thus, we would write

$$Earnings = \beta_1 + \beta_2\, Age + \delta_B\, HS + \delta_M\, C + \delta_P\, PG + \varepsilon.$$

The correspondence between the coefficients and income for a given age is

$$\begin{aligned}
&\text{Less Than High School: } E[Earnings|Age, LTHS] &&= \beta_1 + \beta_2\, Age,\\
&\text{High School:} \qquad\quad E[Earnings|Age, HS] &&= \beta_1 + \beta_2\, Age + \delta_{HS},\\
&\text{College:} \qquad\quad E[Earnings|Age, C] &&= \beta_1 + \beta_2\, Age + \delta_{C},\\
&\text{Post Graduate:} \qquad E[Earnings|Age, PG] &&= \beta_1 + \beta_2\, Age + \delta_{PG}.
\end{aligned}$$

The differences between, say, δ_{PG} and δ_C and between δ_C and δ_{HS} are of interest. Obviously, these are simple to compute. An alternative way to formulate the equation that reveals these differences directly is to redefine the dummy variables to be 1 if the individual has the level of education, rather than whether the level is the highest obtained. Thus, for someone with post graduate education, all three binary variables are 1, and so on. By defining the variables in this fashion, the regression is now

$$\begin{aligned}
&\textit{Less Than High School: } E[Earnings|Age, LTHS] &&= \beta_1 + \beta_2\, Age,\\
&\textit{High School:} \qquad\quad E[Earnings|Age, HS] &&= \beta_1 + \beta_2\, Age + \delta_{HS},\\
&\textit{College:} \qquad\quad E[Earnings|Age, C] &&= \beta_1 + \beta_2\, Age + \delta_{HS} + \delta_{C},\\
&\textit{Post Graduate:} \qquad E[Earnings|Age, PG] &&= \beta_1 + \beta_2\, Age + \delta_{HS} + \delta_{C} + \delta_{PG}.
\end{aligned}$$

Instead of the difference between post graduate and the base case of less than high school, in this model δ_{PG} is the marginal value of the post graduate education, *after college*.

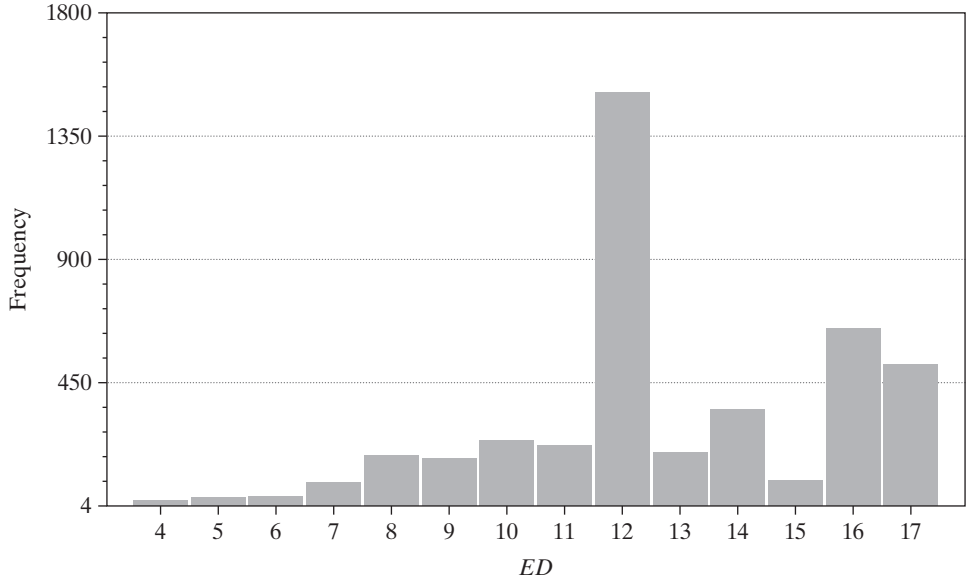
6.2.6 TRANSITION TABLES

When a group of categories appear in the model as a set of dummy variables, as in Example 6.4, each included dummy variable reports the comparison between its category and the "base case." In the movies example, the four reported values each report the comparison to the base category, the nine omitted genres. The comparison of the groups to each other is also a straightforward calculation. In Example 6.4, the reported values for *Action*, *Comedy*, *Animated*, and *Horror* are (−0.869, −0.016, −0.833, +0.375). The implication is, for example, that $E[\ln Revenue|\mathbf{x}]$ is 0.869 less for *Action* movies than the base case. Moreover, based on the same results, the expected log revenue for *Animated* movies is $-0.833 - (-0.869) = +0.036$ greater than for *Action* movies. A standard error for the difference of the two coefficients would be computed using the square root of

$$\begin{aligned}
\text{Asy.Var}[b_{Animated} - b_{Action}] = {}& \text{Asy.Var}[b_{Animated}] + \text{Asy.Var}[b_{Action}]\\
& - 2\text{Asy.Cov}[b_{Animated}, b_{Action}].
\end{aligned}$$

A similar effect could be computed for each pair of outcomes. Hodge and Shankar (2014) propose a useful framework for arranging the effects of a sequence of categories based on this principle. An application to five categories of health outcomes is shown in

Figure 6.1 Education Levels in Log Wage Data.



Contoyannis, Jones, and Rice (2004). The education thresholds example in the previous example is another natural application.

Example 6.7 Education Thresholds in a Log Wage Equation

Figure 6.1 is a histogram for the education levels reported in variable *ED* in the ln *Wage* model of Example 6.3. The model in Table 6.3 constrains the effect of education to be the same 5.5% per year for all values of *ED*. A possible improvement in the specification might be provided by treating the threshold values separately. We have recoded *ED* in these data to be

Less Than High School	$= 1$ if $ED \leq 11$	(22% *of the sample*),
High School	$= 1$ if $ED = 12$	(36% *of the sample*),
College	$= 1$ if $13 \leq ED \leq 16$	(30% *of the sample*),
Post Grad	$= 1$ if $ED = 17$	(12% *of the sample*).

(Admittedly, there might be some misclassification at the margins. It also seems likely that the *Post Grad* category is "top coded"—17 years represents 17 or more.) Table 6.6 reports the respecified regression model. Note, first, the estimated gender effect is almost unchanged. But, the effects of education are rather different. According to these results, the marginal value of high school compared to less than high school is 0.13832, or 14.8%. The estimated marginal value of attending college after high school is $0.29168 - 0.13832 = 0.15336$, 16.57%—this is roughly 4% per year for four years compared to 5.5% estimated earlier. But, again, one might suggest that most of that gain would be a "sheepskin" effect attained in the fourth year by graduating. Hodge and Shankar's "transition matrix" is shown in Table 6.7. (We have omitted the redundant terms and transitions from more education to less which are the negatives of the table entries.)

TABLE 6.6 Estimated log Wage Equations with Education Thresholds

	Threshold Effects			*Education in Years*		
Sum of squared residuals	403.329			391.056		
Standard error of the regression	0.31194			0.30708		
R-squared based on 4165 observations	0.54524			0.55908		
	Coefficient	*Clustered Std.Error*	*t Ratio*	*Coefficient*	*Clustered Std.Error*	*t Ratio*
Constant	5.60883	0.10087	55.61	5.08397	0.12998	39.11
EXP	0.03129	0 .00421	7.44	0.03128	0.00419	7.47
EXP^2	−0.00056	0.00009	−5.97	−0.00055	0.00009	−5.86
WKS	0.00383	0.00157	2.44	0.00394	0.00158	2.50
OCC	−0.16410	0.02683	−6.12	−0.14116	0.02687	−5.25
IND	0.05365	0.02368	2.27	0.05661	0.02343	2.42
SOUTH	−0.07438	0.02704	−2.75	−0.07180	0.02632	−2.73
SMSA	0.16844	0.02368	7.11	0.15423	0.02349	6.57
MS	0.10756	0.04470	2.41	0.09634	0.04301	2.24
UNION	0.07736	0.02405	3.22	0.08052	0.02335	3.45
FEM	−0.35323	0.05005	−7.06	−0.36502	0.04829	−7.56
ED				0.05499	0.00556	9.88
LTHS	0.00000	—	—			
HS	0.13832	0.03351	4.13			
COLLEGE	0.29168	0.04181	6.98			
POSTGRAD	0.40651	0.04896	8.30			
Year(Base = 1976)						
1977	0.07493	0.00608	12.33	0.07461	0.00601	12.42
1978	0.19720	0.00997	19.78	0.19611	0.00989	19.82
1979	0.28472	0.01023	27.83	0.28358	0.01016	27.90
1980	0.36377	0.00997	36.47	0.36264	0.00985	36.82
1981	0.43877	0.01147	38.25	0.43695	0.01133	38.58
1982	0.52357	0.01219	42.94	0.52075	0.01211	43.00

TABLE 6.7 Education Effects in Estimated Log Wage Equation

Effects of switches between categories in education level

Initial Education	*New Education*	*Partial Effect*	*Standard Error*	*t Ratio*
LTHS	*HS*	0.13832	0.03351	4.13
LTHS	*COLLEGE*	0.29168	0.04181	6.98
LTHS	*POSTGRAD*	0.40651	0.04896	8.30
HS	*COLLEGE*	0.15336	0.03047	5.03
HS	*POSTGRAD*	0.26819	0.03875	6.92
COLLEGE	*POSTGRAD*	0.11483	0.03787	3.03

6.3 DIFFERENCE IN DIFFERENCES REGRESSION

Many recent studies have examined the causal effect of a **treatment** on some kind of **response**. Examples include the effect of attending an elite college on lifetime income [Dale and Krueger (2002, 2011)], the effect of cash transfers on child health [Gertler (2004)], the effect of participation in job training programs on income [LaLonde (1986)], the effect on employment of an increase in the minimum wage in one of two neighboring states [Card and Krueger (1994)] and pre- versus post-regime shifts in macroeconomic models [Mankiw (2006)], to name but a few.

6.3.1 TREATMENT EFFECTS

The applications can often be formulated in regression models involving a treatment dummy variable, as in

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta D_i + \varepsilon_i,$$

where the shift parameter, δ (under the right assumptions), measures the causal effect of the treatment or the policy change (conditioned on $\mathbf{x}$) on the sampled individuals. For example, Table 6.6 provides a log wage equation based on a national (U.S.) panel survey. One of the variables is *UNION*, a dummy variable that indicates union membership. Measuring the effect of union membership on wages is a longstanding objective in labor economics—see, for example, Card (2001). Our estimate in Table 6.6 is roughly 0.08, or 8%. It will take a bit of additional specification analysis to conclude that the *UNION* dummy truly does measure the effect of membership in that context.[10]

In the simplest case of a comparison of one group to another, without covariates,

$$y_i = \beta_1 + \delta D_i + \varepsilon_i.$$

Least squares regression of y on D will produce

$$b_1 = (\bar{y} \,|\, D_i = 0),$$

that is, the average outcome of those who did not experience the treatment, and

$$d = (\bar{y} \,|\, D_i = 1) - (\bar{y} \,|\, D_i = 0),$$

the difference in the means of the two groups. Continuing our earlier example, if we measure the *UNION* effect in Table 6.6 without the covariates, we find

$$\ln \mathit{Wage} = 6.673\ (0.023) + 0.00834\ \mathit{UNION}\ (0.028).$$

(Standard errors are in parentheses.) Based on a simple comparison of means, there appears to be a less than 1% impact of union membership. This is in sharp contrast to the 8% reported earlier.

When the analysis is of an intervention that occurs over time to everyone in the sample, such as in Krueger's (1999) analysis of the Tennessee STAR experiment in which school performance measures were observed before and after a policy that dictated a change in class sizes, the treatment dummy variable will be a period indicator, $T_t = 0$ in period 1 and 1 in period 2. The effect in β_2 then measures the change in the outcome variable, for example, school performance, pre- to post-intervention; $b_2 = \bar{y}_1 - \bar{y}_0$.

[10] See, for example, Angrist and Pischke (2009, pp. 221–225.)

The assumption that the treatment group does not change from period 1 to period 2 (or that the treatment group and the control group look the same in all other respects) weakens this analysis. A strategy for strengthening the result is to include in the sample a group of **control observations** that do not receive the treatment. The change in the outcome for the **treatment group** can then be compared to the change for the **control group** under the presumption that the difference is due to the intervention. An intriguing application of this strategy is often used in clinical trials for health interventions to accommodate the **placebo effect**. The placebo effect is a controversial, but apparently tangible outcome in some clinical trials in which subjects "respond" to the treatment even when the treatment is a decoy intervention, such as a sugar or starch pill in a drug trial.[11] A broad template for assessment of the results of such a clinical trial is as follows: The subjects who receive the placebo are the controls. The outcome variable—level of cholesterol, for example—is measured at the baseline for both groups. The treatment group receives the drug, the control group receives the placebo, and the outcome variable is measured pre- and post-treatment. The impact is measured by the difference in differences,

$$E = [(\bar{y}_{exit}|treatment) - (\bar{y}_{baseline}|treatment)] - [(\bar{y}_{exit}|placebo) - (\bar{y}_{baseline}|placebo)].$$

The presumption is that the difference in differences measurement is robust to the placebo effect *if it exists*. If there is no placebo effect, the result is even stronger (assuming there is a result).

A common social science application of treatment effect models is in the evaluation of the effects of discrete changes in policy.[12] A pioneering application is the study of the Manpower Development and Training Act (MDTA) by Ashenfelter and Card (1985) and Card and Krueger (2000). A widely discussed application is Card and Krueger's (1994) analysis of an increase in the minimum wage in New Jersey. The simplest form of the model is one with a pre- and post-treatment observation on a group, where the outcome variable is y, with

$$y_{it} = \beta_1 + \beta_2 T_t + \beta_3 D_i + \delta(T_t \times D_i) + \varepsilon_{it},\ t = 0, 1. \qquad \textbf{(6-3)}$$

In this model, T_t is a dummy variable that is zero in the pre-treatment period and one after the treatment and D_i equals one for those individuals who received the treatment. The change in the outcome variable for the treated individuals will be

$$(y_{i2}|D_i = 1) - (y_{i1}|D_i = 1) = (\beta_1 + \beta_2 + \beta_3 + \delta) - (\beta_1 + \beta_3) = \beta_2 + \delta.$$

For the controls, this is

$$(y_{i2}|D_i = 0) - (y_{i1}|D_i = 0) = (\beta_1 + \beta_2) - \beta_1 = \beta_2.$$

The difference in differences is

$$[(y_{i2}|D_i = 1) - (y_{i1}|D_i = 1)] - [(y_{i2}|D_i = 0) - (y_{i1}|D_i = 0)] = \delta.$$

[11] See Hróbjartsson and Götzsche (2001).

[12] Surveys of literatures on treatment effects, including use of "D-i-D" estimators, are provided by Imbens and Wooldridge (2009), Millimet, Smith, and Vytlacil (2008), Angrist and Pischke (2009), and Lechner (2011).

In the multiple regression of y_{it} on a constant, T, D, and $T \times D$, the least squares estimate of δ will equal the difference in the changes in the means,

$$\begin{aligned} d &= (\bar{y}|\, D = 1, \textit{Period } 2) - (\bar{y}|\, D = 1, \textit{Period } 1) \\ &\quad - (\bar{y}|\, D = 0, \textit{Period } 2) - (\bar{y}|\, D = 0, \textit{Period } 1) \\ &= \Delta\bar{y}|\, \textit{treatment} - \Delta\bar{y}|\, \textit{control}. \end{aligned}$$

The regression is called a difference in differences estimator in reference to this result.

Example 6.8 SAT Scores

Each year, about 1.7 million American high school students take the SAT test. Students who are not satisfied with their performance have the opportunity to retake the test. Some students take an SAT prep course, such as Kaplan or Princeton Review, before the second attempt in the hope that it will help them increase their scores. An econometric investigation might consider whether these courses are effective in increasing scores. The investigation might examine a sample of students who take the SAT test twice, with scores y_{i0} and y_{i1}. The time dummy variable T_t takes value $T_0 = 0$ "before" and $T_1 = 1$ "after." The treatment dummy variable is $D_i = 1$ for those students who take the prep course and 0 for those who do not. The applicable model would be (6-3),

$$\textit{SAT Score}_{i,t} = \beta_1 + \beta_2\, \textit{2ndTest}_t + \beta_3\, \textit{PrepCourse}_i + \delta\, \textit{2ndTest}_t \times \textit{PrepCourse}_i + \varepsilon_{i,t}.$$

The estimate of δ would, in principle, be the treatment, or prep course effect.

This small example illustrates some major complications. First, and probably most important, the setting does not describe a randomized experiment such as the clinical trial suggested earlier would be. The treatment variable, *PrepCourse*, would naturally be taken by those who are persuaded that it would provide a benefit—that is, the treatment variable is not an exogenous variable. Unobserved factors that are likely to contribute to higher test scores (and are embedded in $\varepsilon_{i,t}$) would likely motivate the student to take the prep course as well. This *selection effect* is a compelling confounder of studies of treatment effects when the treatment is voluntary and self selected. Dale and Krueger's (2002, 2011) analysis of the effect of attendance at an elite college provides a detailed analysis of this issue. Second, test performance, like other performance measures, is probably subject to regression to the mean—there is a negative autocorrelation in such measures. In this regression context, an unusually high disturbance in period 0, all else equal, would likely be followed by a low value in period 1. Of course, those who achieve an unusually high test score in period 0 are less likely to return for the second attempt. Together with the selection effect, this produces a very muddled relationship between the outcome and the test preparation that is estimated by least squares. Finally, it is possible that there are other measurable factors (covariates) that might contribute to the test outcome or changes in the outcome. A more complete model might include these covariates. We do note any such variable $x_{i,t}$ would have to vary between the first and second test, else they would simply be absorbed in the constant term.

When the treatment is the result of a policy change or event that occurs completely outside the context of the study, the analysis is often termed a **natural experiment**. Card's (1990) study of a major immigration into Miami in 1979 is an application.

Example 6.9 A Natural Experiment: The Mariel Boatlift

A sharp change in policy can constitute a natural experiment. An example studied by Card (1990) is the Mariel boatlift from Cuba to Miami (May–September 1980), which increased the Miami labor force by 7%. The author examined the impact of this abrupt change in labor market conditions on wages and employment for nonimmigrants. The model compared Miami (the treatment group) to a similar city, Los Angeles (the control group). Let *i* denote an

individual and D denote the "treatment," which for an individual would be equivalent to "lived in the city that experienced the immigration." For an individual in either Miami or Los Angeles, the outcome variable is

$$Y_i = 1 \quad \text{if they are unemployed and 0 if they are employed.}$$

Let c denote the city and let t denote the period, before (1979) or after (1981) the immigration. Then, the unemployment rate in city c at time t is $E[y_{i,0}|c, t]$ if there is no immigration and it is $E[y_{i,1}|c, t]$ if there is the immigration. These rates are assumed to be constants. Then

$$E[y_{i,0}|c, t] = \beta_t + \gamma_c \qquad \text{without the immigration,}$$
$$E[y_{i,1}|c, t] = \beta_t + \gamma_c + \delta \quad \text{with the immigration.}$$

The effect of the immigration on the unemployment rate is measured by δ. The natural experiment is that the immigration occurs in Miami and not in Los Angeles but is not a result of any action by the people in either city. Then,

$$E[y_i|M,79] = \beta_{79} + \gamma_M \quad \textit{and} \quad E[y_i|M, 81] = \beta_{81} + \gamma_M + \delta \quad \textit{for Miami},$$
$$E[y_i|L, 79] = \beta_{79} + \gamma_L \quad \textit{and} \quad E[y_i|L, 81] = \beta_{81} + \gamma_L \qquad \textit{for Los Angeles}.$$

It is assumed that unemployment growth in the two cities would be the same if there were no immigration. If neither city experienced the immigration, the change in the unemployment rate would be

$$E[y_{i,0}|M, 81] - E[y_{i,0}|M, 79] = \beta_{81} - \beta_{79} \qquad \textit{for Miami},$$
$$E[y_{i,0}|L, 81] - E[y_{i,0}|L, 79] = \beta_{81} - \beta_{79} \qquad \textit{for Los Angeles}.$$

If both cities were exposed to migration,

$$E[y_{i,1}|M, 81] - E[y_{i,1}|M, 79] = \beta_{81} - \beta_{79} + \delta \quad \textit{for Miami},$$
$$E[y_{i,1}|L, 81] - E[y_{i,1}|L, 79] = \beta_{81} - \beta_{79} + \delta \quad \textit{for Los Angeles}.$$

Only Miami experienced the immigration (the "treatment"). The difference in differences that quantifies the result of the experiment is

$$\{E[y_{i,1}|M, 81] - E[y_{i,1}|M, 79]\} - \{E[y_{i,0}|L, 81] - E[y_{i,0}|L, 79]\} = \delta.$$

The author examined changes in employment rates and wages in the two cities over several years after the boatlift. The effects were surprisingly modest (essentially nil) given the scale of the experiment in Miami.

Example 6.10 Effect of the Minimum Wage

Card and Krueger's (1994) widely cited analysis of the impact of a change in the minimum wage is similar to Card's analysis of the Mariel Boatlift. In April 1992, New Jersey (NJ) raised its minimum wage from $4.25 to $5.05. The minimum wage in neighboring Pennsylvania (PA) was unchanged. The authors sought to assess the impact of this policy change by examining the change in employment in the two states from February to November, 1992 at fast food restaurants that tended to employ large numbers of people at the minimum wage. Conventional wisdom would suggest that, all else equal, whatever labor market trends were at work in the two states, NJ's would be affected negatively by the abrupt 19% wage increase for minimum wage workers. This certainly qualifies as a natural experiment. NJ restaurants could not opt out of the treatment. The authors were able to obtain data on employment for 331 NJ restaurants and 97 PA restaurants in the first wave. Most of the first wave restaurants provided data for the second wave, 321 and 78, respectively. One possible source of "selection" would be attrition from the sample. Though the numbers are small, the possibility that the second wave sample was substantively composed of firms that were affected by the policy change

TABLE 6.8 Full Time Employment in NJ and PA Restaurants

	PA	*NJ*
First Wave (February)	23.33	20.44
Second Wave (November)	21.17	21.03
Difference	−2.16	0.59
Difference (balanced)	−2.28	0.47

would taint the analysis (e.g., if firms were driven out of business because of the increased labor costs). The authors document at some length the data collection process for the second wave. Results for their experiment are shown in Table 6.8.

The first reported difference uses the full sample of available data. The second uses the "balanced sample" of all stores that reported data in both waves. In both cases, the difference in differences would be

$$\Delta(NJ) - \Delta(\text{PA}) = +2.75 \text{ full time employees.}$$

A superficial analysis of these results suggests that they go in the wrong direction. Employment rose in NJ compared to PA in spite of the increase in the wage. Employment would have been changing in both places due to other economic conditions. The policy effect here might have distorted that trend. But, it is also possible that the trend in the two states was different. It has been assumed throughout so far that it is the same. Card and Krueger (2000) examined this possibility in a followup study. The newer data cast some doubt on the crucial assumption that the trends were the same in the two states.

Card and Krueger (1994) considered the possibility that restaurant specific factors might have influenced their measured outcomes. The implied regression would be

$$y_{it} = \beta_2 T_t + \beta_3 D_i + \delta T_t \times D_i + (\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i) + \varepsilon_{it}, \quad t = 0, 1.$$

Note the individual specific constant term that represents the unobserved heterogeneity and the addition to the regression. In the restaurant study, $\mathbf{x}_i$ was characteristics of the store such as chain store type, ownership, and region—all features that would be the same in both waves. These would be fixed effects. In the difference in differences context, while they might indeed be influential in the outcome levels, it is clear that they will fall out of the differences:

$$\begin{aligned}\Delta E[y_{it}|D_{it} = 0, \mathbf{x}_i] &= \beta_2 \qquad + \Delta(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i),\\ \Delta E[y_{it}|D_{it} = 1, \mathbf{x}_i] &= \beta_2 + \delta + \Delta(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i).\end{aligned}$$

The final term in both cases is zero, which leaves, as before,

$$\Delta E[y_{it}|D_{it} = 1, \mathbf{x}_i] - \Delta E[y_{i\,\mathrm{t}}|D_{it} = 1, \mathbf{x}_i] = \delta.$$

The useful conclusion is that in analyzing differences in differences, time invariant characteristics of the individuals will not affect the conclusions.

The analysis is more complicated if the control variables, $\mathbf{x}_{it}$, do change over time. Then,

$$y_{it} = \beta_2 T_t + \beta_3 D_i + \delta T_t \times D_i + \boldsymbol{\gamma}'\mathbf{x}_{it} + \varepsilon_{it}, \; t = 0, 1.$$

Then,

$$\Delta E[y_{it}|\mathbf{x}_{it}, D_{it} = 1] = \beta_2 + \delta + \boldsymbol{\gamma}'[\Delta\mathbf{x}_{it}|D_{it} = 1]$$
$$\Delta E[y_{it}|\mathbf{x}_{it}, D_{it} = 0] = \beta_2 + \quad \boldsymbol{\gamma}'[\Delta\mathbf{x}_{it}|D_{it} = 0]$$

$$\Delta E[y_{it}|D_{it} = 1, \mathbf{x}_i] - \Delta E[y_{it}|D_{it} = 1, \mathbf{x}_i] = \delta + \boldsymbol{\gamma}'[(\Delta\mathbf{x}_{it}|D_{it} = 1) - (\Delta\mathbf{x}_{it}|D_{it} = 0)].$$

Now, if the effect of D_{it} is measured by the simple difference of means, the result will consist of the causal effect plus an additional term explained by the difference of the changes in the control variables. If individuals have been carefully sampled so that treatment and controls look the same in both periods, then the second effect might be ignorable. If not, then the second part of the regression should become part of the analysis.

6.3.2 EXAMINING THE EFFECTS OF DISCRETE POLICY CHANGES

The differences in differences result provides a convenient methodology for studying the effects of exogenously imposed policy changes. We consider an application from a recent antitrust case.

Example 6.11 Difference in Differences Analysis of a Price Fixing Conspiracy[13]

Roughly 6.5% of all British schoolchildren, and more than 18% of those over 16, attend 2,600 independent fee-paying schools. Of these, roughly 10.5% are "boarders"—the remainder attend on a day basis. Each year from 1997 until June, 2003, a group of 50 of these schools shared information about intended fee increases for boarding and day students. The information was exchanged via a survey known as the "Sevenoaks Survey" (SS). The UK Office of Fair Trading (OFT, Davies (2012)) determined that the conspiracy, which was found to lead to higher fees, was prohibited under the antitrust law, the Competition Act of 1998. The OFT intervention consisted of a modest fine (10,000GBP) on each school, a mandate for the cartel to contribute about 3,000,000GBP to a trust, and prohibition of the Sevenoaks Survey. The OFT investigation was ended in 2006, but for the purposes of the analysis, the intervention is taken to have begun with the 2004/2005 academic year.

The authors of this study investigated the impact of the OFT intervention on the boarding and day fees of the Sevenoaks schools using a difference in differences regression. The pre-intervention period is academic years 2001/02 to 2003/04. The post-intervention period extends to 2011/2012. The sample consisted of the treatment group, the 50 Sevenoaks schools, and 178 schools that were not party to the conspiracy and therefore, not impacted by the treatment. (Not necessarily. More on that below.) The "balanced panel data set" of 12 years times 228 schools, or 2,736 observations, was reduced by missing data to 1,829 for the day fees model and 1,317 for the boarding fees model. Figure 6.2 (Figures 2 and 3 from the study) shows the behavior of the boarding and day fees for the schools for the period of the study.[14] It is difficult to see a difference in the rates of change of the fees. The difference in the levels is obvious, but not yet explained.

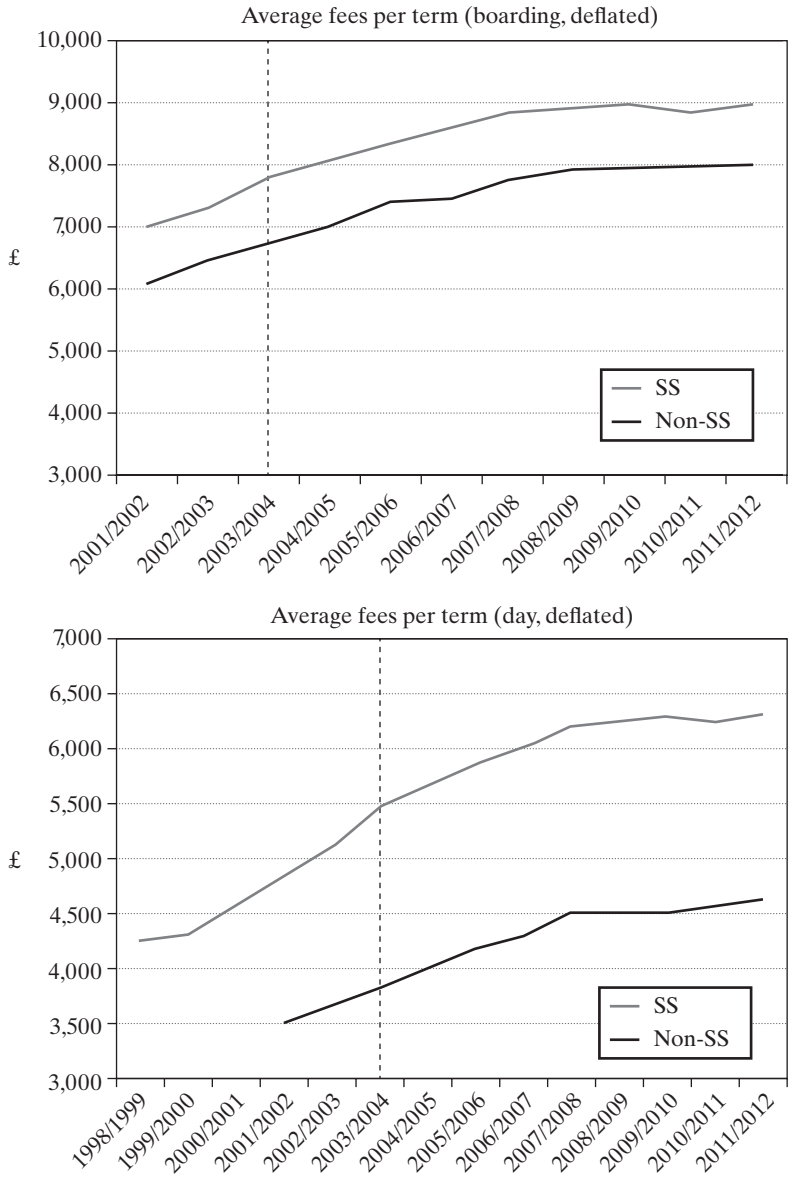
A difference in differences methodology was used to analyze the behavior of the fees. Two key assumptions are noted at the outset.

1. The schools in the control group are not affected by the intervention. This may not be the case. The non-SS schools compete with the SS schools on a price basis. If the pricing behavior of the SS schools is affected by the intervention, that of the non-SS schools may be as well.

[13] This case study is based on UK OFT (2012), Davies (2012) and Pesarisi et al. (2015).

[14] The figures are extracted from the UK OFT (2012) working paper version of the study.

Figure 6.2 Price Increases by Boarding Schools.



2. It must be assumed that the trends and influences that affect the two groups of schools outside the effect of the intervention are the same. (Recall this was an issue in Card and Krueger's analysis of the minimum wage in Example 6.10.)

The linear regression model used to study the behavior of the fees is

$$\ln Fee_{it} = \alpha_i + \beta_1 \%boarder_{it} + \beta_2 \%ranking_{it} + \beta_3 \ln pupils_{it} + \beta_4 year_t + \lambda postintervention_t + \delta SS_{it} \times postintervention_t + \varepsilon_{it}$$

Fee_{it}	= inflation-adjusted day or boarding fees,
%*boarder*	= percentage of the students who are boarders at school *i* in year *t*,
%*ranking*	= percentile ranking of the school in *Financial Times* school rankings,
pupils	= number of students in the school,
year	= linear trend,
postintervention	= dummy variable indicating the period after the intervention,
SS	= dummy variable for Sevenoaks school,
α_i	= school-specific effect, modeled using a school specific dummy variable.

The effect of interest is δ. Several assumptions underlying the data are noted to justify the interpretation of δ as the sought-after causal impact of the intervention.

a. The effect of the intervention is exerted on the fees beginning in 2004/2005.
b. In the absence of the intervention, the regime would have continued on to 2012 as it had in the past.
c. The *Financial Times* ranking variable is a suitable indicator of the quality of the ranked school.
d. As noted earlier, pricing behavior by the control schools was not affected by the intervention.

The regression results are shown in Table 6.9.

The main finding is a decline of 1.5% for day fees and 1.6% for the boarding fees. Figure 6.3 [extracted from the UK OFT (2012) version of the paper] summarizes the estimated cumulative impact of the study. The authors estimated the cumulative savings attributable to the intervention based on the results in Figure 6.3 to be roughly 85 million GBP.

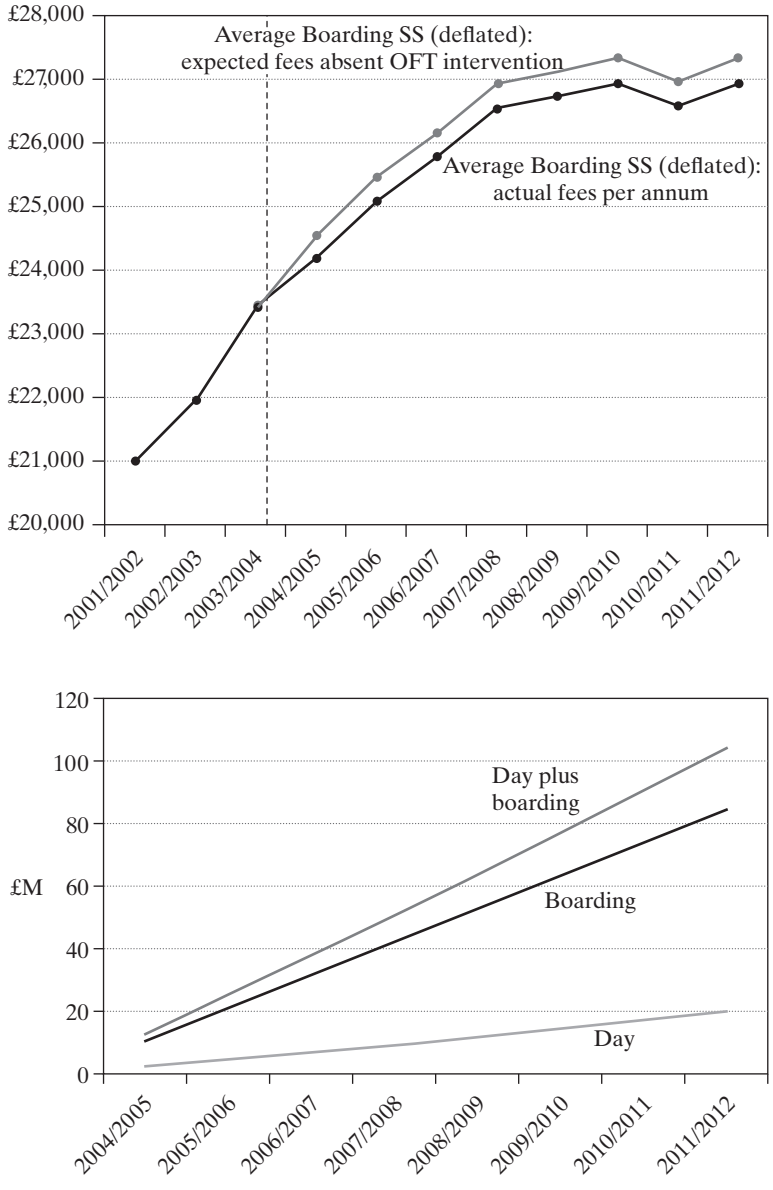
One of the central issues in policy analysis concerns measurement of treatment effects when the treatment results from an individual participation decision. In the clinical trial example given earlier, the control observations (it is assumed) do not know they they are in the control group. The treatment assignment is exogenous to the experiment. In contrast, in Krueger and Dale (1999) study, the assignment to the treatment group, attended the elite college, is completely voluntary and determined by the individual. A crucial aspect of the analysis in this case is to accommodate the almost certain outcome that the treatment dummy might be measuring the latent motivation and initiative of the participants rather than the effect of the program itself. That is the

TABLE 6.9 Estimated Models for Day and Boarding Fees*

	Day Fees	Boarding Feees
% Boarder	0.7730 (0.051)**	0.0367 (0.029)
% Ranking	−0.0147 (0.019)	0.00396 (0.015)
ln Pupils	0.0247 (0.033)	0.0291 (0.021)
Year	0.0698 (0.004)	0.0709 (0.004)
Post-intervention	0.0750 (0.027)	0.0674 (0.022)
Post-intervention and SS	−0.0149 (0.007)	−0.0162 (0.005)
N	1,825	1,311
R^2	0.949	0.957

Source: Pesaresi et al. (2015), Table 1. © Crown copyright 2012
* Model fit by least squares. Estimated individual fixed effects not shown.
** Robust standard errors that account for possible heteroscedasticity and autocorrelation in parentheses.

Figure 6.3 Cumulative Impact of Sevenoaks Intervention.



main appeal of the natural experiment approach—it more closely (possibly exactly) replicates the exogenous treatment assignment of a clinical trial.[15] We will examine some of these cases in Chapters 8 and 19.

[15] See Angrist and Krueger (2001) and Angrist and Pischke (2010) for discussions of this approach.

6.4 USING REGRESSION KINKS AND DISCONTINUITIES TO ANALYZE SOCIAL POLICY

The ideal situation for the analysis of a change in social policy would be a randomized assignment of a sample of individuals to treatment and control groups.[16] There are some notable examples to be found. The Tennessee STAR class size experiment was designed to study the effect of smaller class sizes in the earliest grades on short and long term student performance. [See Mosteller (1995) and Krueger (1999) and, for some criticism, Hanushek (1999, 2002).] A second prominent example is the Oregon Health Insurance Experiment.

> *The Oregon Health Insurance Experiment is a landmark study of the effect of expanding public health insurance on health care use, health outcomes, financial strain, and well-being of low-income adults. It uses an innovative randomized controlled design to evaluate the impact of Medicaid in the United States. Although randomized controlled trials are the gold standard in medical and scientific studies, they are rarely possible in social policy research. In 2008, the state of Oregon drew names by lottery for its Medicaid program for low-income, uninsured adults, generating just such an opportunity. This ongoing analysis represents a collaborative effort between researchers and the state of Oregon to learn about the costs and benefits of expanding public health insurance. (www.nber.org/oregon/)*
>
> *In 2008, a group of uninsured low-income adults in Oregon was selected by lottery to be given the chance to apply for Medicaid. This lottery provides a unique opportunity to gauge the effects of expanding access to public health insurance on the health care use, financial strain, and health of low-income adults using a randomized controlled design. In the year after random assignment, the treatment group selected by the lottery was about 25 percentage points more likely to have insurance than the control group that was not selected. We find that in this first year, the treatment group had substantively and statistically significantly higher health care utilization (including primary and preventive care as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health than the control group. [Finkelstein et al. (2011).]*

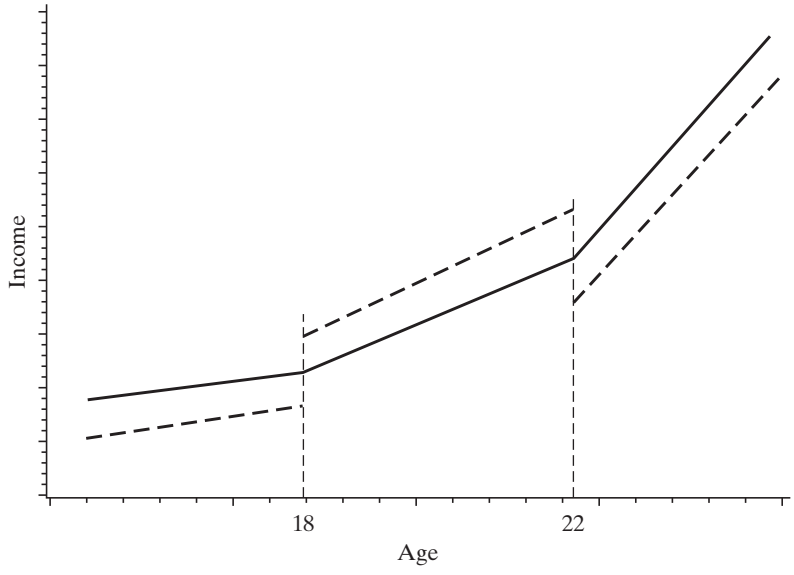
Substantive social science studies such as these, based on random assignment, are rare. The natural experiment approach, such as in Example 6.9, is an appealing alternative when it is feasible. Regression models with kinks and discontinuities have been designed to study the impact of social policy in the absence of randomized assignment.

6.4.1 REGRESSION KINKED DESIGN

A plausible description of the age profile of incomes will show incomes rising throughout but at different rates after some distinct milestones, for example, at age 18, when the typical individual graduates from high school, and at age 22, when he or she graduates from college. The profile of incomes for the typical individual in this population might appear as in Figure 6.4. We could fit such a regression model just by dividing the sample into three subsamples. However, this would neglect the continuity of the proposed function and possibly misspecify the relationship of other variables that might appear in the model. The result would appear more like the dashed figure than the continuous

[16] See Angrist and Pischke (2009).

Figure 6.4 Piecewise Linear Regression.



function we had in mind. Constrained regression can be used to achieve the desired effect. The function we wish to estimate is

$$\begin{aligned} E[income|age] = \alpha^0 + \beta^0\, age \quad & \text{if } age < 18, \\ \alpha^1 + \beta^1\, age \quad & \text{if } age \geq 18 \text{ and } age < 22, \\ \alpha^2 + \beta^2\, age \quad & \text{if } age \geq 22. \end{aligned}$$

Let

$$\begin{aligned} d_1 &= 1 \quad \text{if } age \geq t_1^*, \\ d_2 &= 1 \quad \text{if } age \geq t_2^*, \end{aligned}$$

where $t_1^* = 18$ and $t_2^* = 22$. To combine the three equations, we use

$$income = \beta_1 + \beta_2\, age + \gamma_1 d_1 + \delta_1 d_1\, age + \gamma_2 d_2 + \delta_2 d_2\, age + \varepsilon.$$

This produces the dashed function Figure 6.4. The slopes in the three segments are β_2, $\beta_2 + \delta_1$, and $\beta_2 + \delta_1 + \delta_2$. To make the function *continuous*, we require that the segments join at the thresholds—that is,

$$\begin{aligned} &\beta_1 + \beta_2 t_1^* = (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_1^* \text{ and} \\ &(\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_2^* = (\beta_1 + \gamma_1 + \gamma_2) + (\beta_2 + \delta_1 + \delta_2) t_2^*. \end{aligned}$$

These are linear restrictions on the coefficients. The first one is

$$\gamma_1 + \delta_1 t_1^* = 0 \quad \text{or} \quad \gamma_1 = -\delta_1 t_1^*.$$

Doing likewise for the second, we obtain

$$income = \beta_1 + \beta_2\, age + \delta_1 d_1\, (age - t_1^*) + \delta_2 d_2\, (age - t_2^*) + \varepsilon.$$

Constrained least squares estimates are obtainable by multiple regression, using a constant and the variables

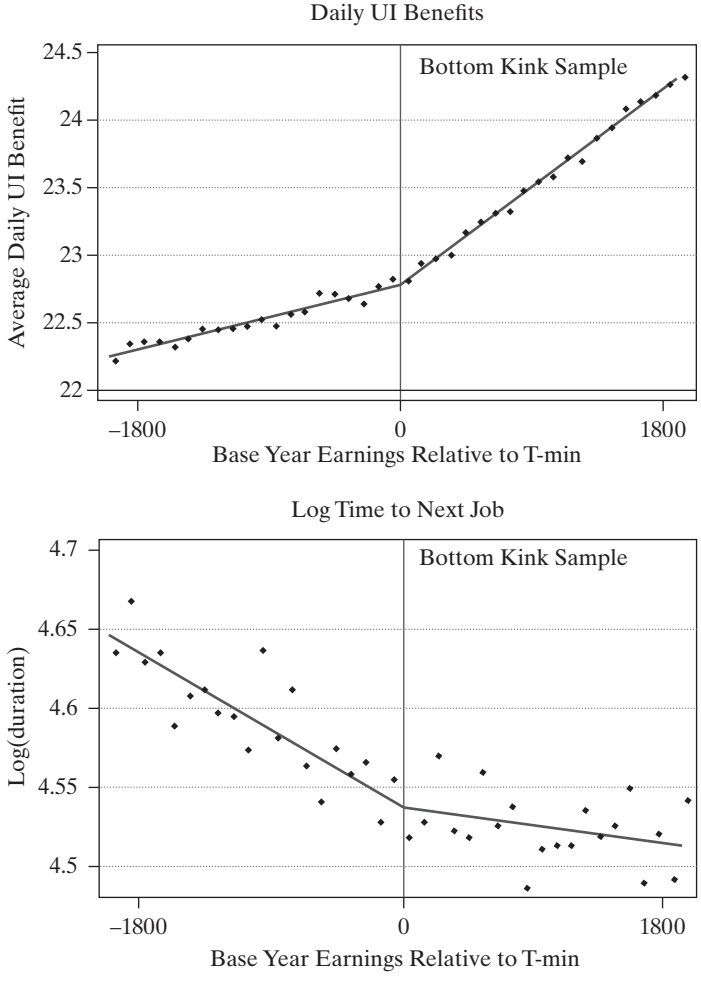
$$x_1 = age,$$
$$x_2 = age - 18 \quad \text{if } age \geq 18 \text{ and 0 othewise,}$$
$$x_3 = age - 22 \quad \text{if } age \geq 22 \text{ and 0 othewise.}$$

We can test the hypothesis that the slope of the function is constant with the joint test of the two restrictions $\delta_1 = 0$ and $\delta_2 = 0$.

Example 6.12 Policy Analysis Using Kinked Regressions

Discontinuities such as those in Figure 6.4 can be used to help identify policy effects. Card, Lee, Pei, and Weber (2012) examined the impact of unemployment insurance (UI) on the duration of joblessness in Austria using a regression kink design. The policy lever, UI, has a sharply defined benefit schedule level tied to base year earnings that can be traced through to its impact on the duration of unemployment. Figure 6.5 [from Card et al. (2012, p. 48)]

Figure 6.5 Regression Kink Design.



suggests the nature of the identification strategy. Simonsen, Skipper, and Skipper (2015) used a similar strategy to examine the effect of a subsidy on the demand for pharmaceuticals in Denmark.

6.4.2 REGRESSION DISCONTINUITY DESIGN

Van der Klaauw (2002) studied financial aid offers that were tied to SAT scores and grade point averages using a regression discontinuity design. The conditions under which the approach can be effective are when (1) the outcome, y, is a continuous variable; (2) the outcome varies smoothly with an assignment variable, A; and (3) treatment is *sharply* assigned based on the value of A, specifically $T = 1(A > A^*)$ where A^* is a fixed threshold or cutoff value. [A **fuzzy design** is based on $\text{Prob}(T = 1|A) = F(A)$. The identification problems with fuzzy design are much more complicated than with sharp design. Readers are referred to Van der Klaauw (2002) for further discussion of fuzzy design.] We assume, then, that

$$y = f(A, T) + \varepsilon.$$

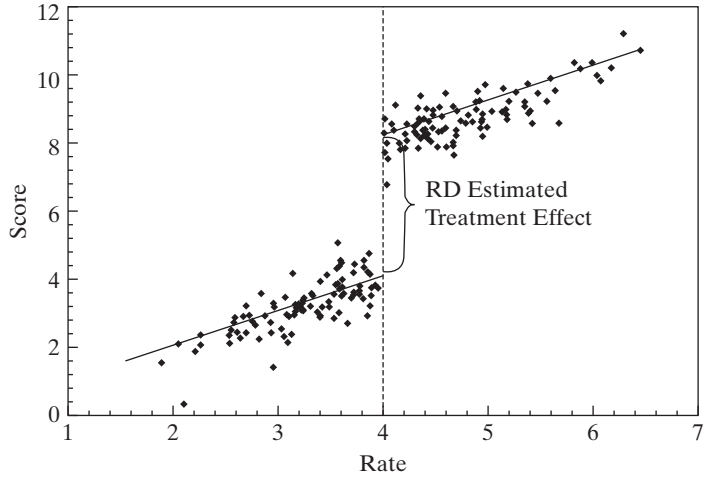
Suppose, for example, the outcome variable is a test score, and that an administrative treatment such as a special education program is funded based on the poverty rates of certain communities. The ideal conditions for a regression discontinuity design based on these assumptions are shown in Figure 6.6. The logic of the calculation is that the points near the threshold value, which have essentially the same stimulus value, constitute a nearly random sample of observations which are segmented by the treatment.

The method requires that $E[\varepsilon|A, T] = E[\varepsilon|A]$—the assignment variable—be exogenous to the experiment. The result in Figure 6.6 is consistent with

$$y = f(A) + \alpha T + \varepsilon,$$

where α will be the treatment effect to be estimated. The specification of $f(A)$ can be problematic; assuming a linear function when something more general is appropriate

Figure 6.6 Regression Discontinuity.



will bias the estimate of α. For this reason, nonparametric methods, such as the LOWESS regression (see Section 12.4), might be attractive. This is likely to enable the analyst to make fuller use of the observations that are more distant from the cutoff point.[17] Identification of the treatment effect begins with the assumption that $f(A)$ is continuous at A^*, so that

$$\lim_{A\uparrow A^*} f(A) = \lim_{A\downarrow A^*} f(A) = f(A^*).$$

Then

$$\lim_{A\downarrow A^*} E[y|A] - \lim_{A\uparrow A^*} E[y|A] = f(A^*) + \alpha + \lim_{A\downarrow A^*} E[\varepsilon|A] - f(A^*) - \lim_{A\uparrow A^*} E[\varepsilon|A]$$
$$= \alpha.$$

With this in place, the treatment effect can be estimated by the difference of the average outcomes for those individuals close to the threshold value, A^*. Details on regression discontinuity design are provided by Trochim (1984, 2000) and Van der Klaauw (2002).

Example 6.13 The Treatment Effect of Compulsory Schooling

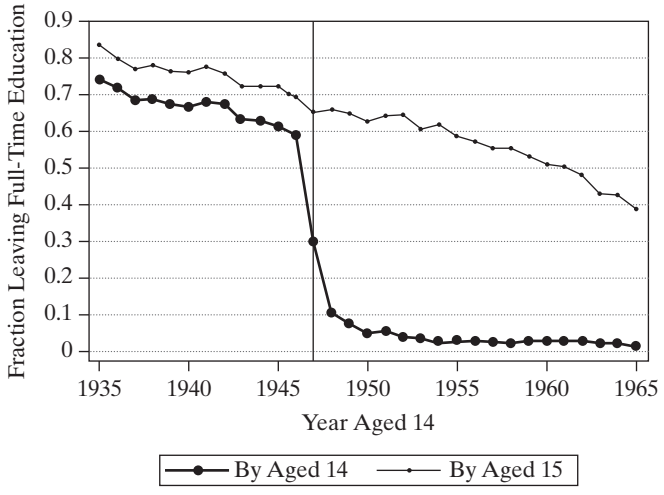
Oreopoulos (2006) examined returns to education in the UK in the context of a discrete change in the national policy on mandatory school attendance. [See, also, Ashenfelter and Krueger (2010b) for a U.S. study.] In 1947, the minimum school-leaving age in Great Britain was changed from 14 to 15 years. In this period, from 1935 to 1960, the exit rate among those old enough in the UK was more than 50%, so the policy change would affect a significant number of students. For those who turned 14 in 1947, the policy would induce a mandatory increase in years of schooling for many students who would otherwise have dropped out. Figure 6.7 (composed from Figures 1 and 6 from the article) shows the quite stark impact of the policy change. (A similar regime change occurred in Northern Ireland in 1957.) A regression of the log of annual earnings that includes a control for birth cohort reveals a distinct break for those born in 1933, that is, those who were affected by the policy change in 1947. The estimated regression produces a return to compulsory schooling of about 7.9% for Great Britain and 11.3% for Northern Ireland. (From Table 2. The figures given are based on least squares regressions. Using instrumental variables produces results of about 14% and 18%, respectively.)

Example 6.14 Interest Elasticity of Mortgage Demand

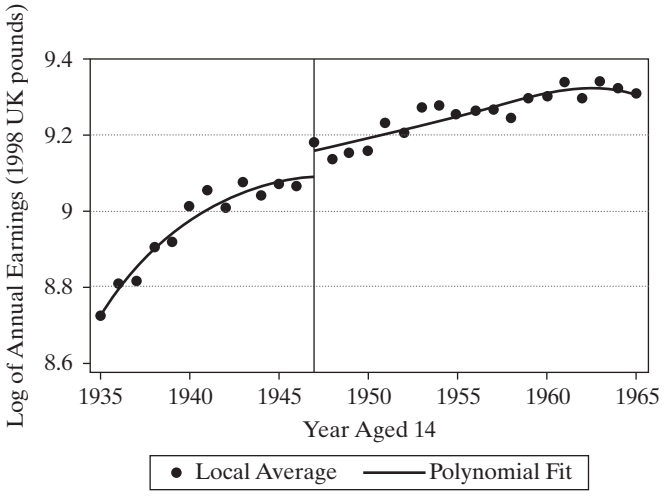
DeFusco and Paciorek (2014, 2016) studied the interest rate elasticity of the demand for mortgages. There is a natural segmentation in this market imposed by the maximum limit on loan sizes eligible for purchase by the Government Sponsored Enterprises (GSEs), Fannie Mae and Freddie Mac. The limits, set by the Federal Housing Finance Agency, vary by housing type and have been adjusted over time. The current loan limit, called the *conforming loan limit (CLL)* for single family homes has been fixed at $417,000 since 2006. A loan that is larger than the CLL is labeled a "jumbo loan." Because the GSEs are able to obtain an implicit subsidy in capital markets, there is a discrete jump in interest rates at the conforming loan limit. The relationship between the mortgage size and the interest rates is key to the specification of the denominator of the elasticity. This foregoing suggests a regression discontinuity approach to the relationship between mortgage rates and loan sizes, such as shown in the left panel

[17] See Van der Klaauw (2002).

Figure 6.7 Regression Discontinuity Design for Returns to Schooling.



Note: The lower line shows the proportion of British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys who report leaving full-time education at or before age 14 from 1935 to 1965. The upper line shows the same, but for age 15. The minimum school leaving age in Great Britain changed in 1947 from 14 to 15.



Note: Local averages are plotted for British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys. The curved line shows the predicted fit from regressing average log annual earnings on a birth cohort quartic polynomial and an indicator for the school leaving age faced at age 14. The school leaving age increased from 14 to 15 in 1947, indicated by the vertical line. Earnings are measured in 1998 UK pounds using the UK retail price index.

Figure 6.8 Regression Discontinuity Design for Mortgage Demand.

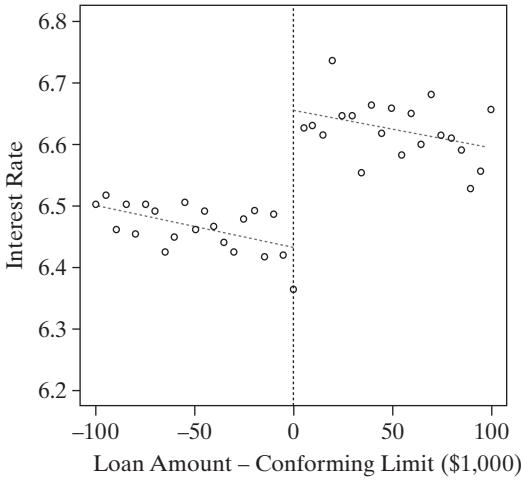


FIG. 2.—Mean Interest Rate Relative to the Conforming Limit, Fixed-Rate Mortgages Only (2006). This figure plots the mean interest rate for fixed rate mortgages originated in 2006 as a function of the loan amount relative to the conforming limit. Each dot represents the mean interest rate within a given $5,000 bin relative to the limit. The dashed lines are predicted values from a regression fit to the binned data allowing for changes in the slope and intercept at the conforming limit. Sample includes all loans in the LPS fixed-rate sample that fall within $100,000 of the conforming limit. See text for details on sample construction.

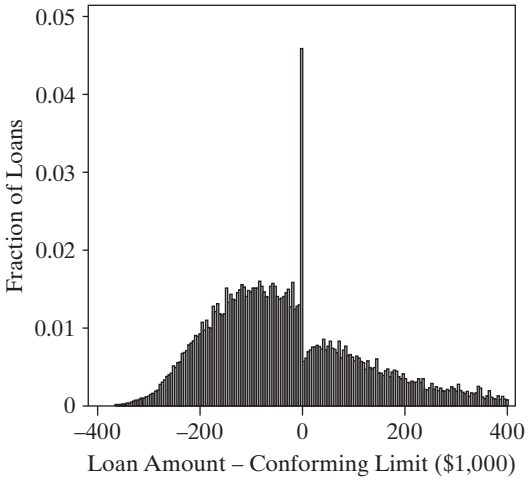


FIG. 3.—Loan Size Distribution Relative to the Conforming Limit. This figure plots the fraction of all loans that are in any given $5,000 bin relative to the conforming limit. Data are pooled across years and each loan is centered at the conforming limit in effect at the date of origination, so that a value of 0 represents a loan at exactly the conforming limit. Sample includes all transactions in the primary DataQuick sample that fall within $400,000 of the conforming limit. See text for details on sample construction.

of Figure 6.8. [Figure 2 in DeFusco and Paciorek (2014).] The semiparametric regression proposed was as follows:

$$r_{i,t} = \alpha_{z(i),t} + \beta J_{i,t} + f^{J=0}(m_{i,t}) + f^{J=1}(m_{i,t}) + s^{LTV}(LTV_{it}) + s^{DTI}(DTI_{i,t}) + s^{FICO}(FICO_{i,t}) + PMI_{i,t} + PP_{i,t} + g(TERM_{i,t}) + \varepsilon_{i,t}.$$

The variables in the specification are:

$r_{i,t}$ = interest rate on loan i originated at time t,
$\alpha_{Z(i),t}$ = fixed effect for zip code and time,
J = dummy variable for jumbo loan ($J=1$) or conforming loan ($J=0$),
$m_{i,t}$ = *size* of the mortgage,
$f^{J=0}$ = $(1-J) \times$ cubic polynomial in the mortgage size,
$f^{J=1}$ = $J \times$ cubic polynomial in the mortgage size,
$LTV_{i,t}$ = loan to value ratio,
$DTI_{i,t}$ = debt to income ratio,
$FICO_{i,t}$ = credit score of borrower,
$PMI_{i,t}$ = dummy variable for whether borrower took out private mortgage insurance,
$PP_{i,t}$ = dummy variable for whether mortgatge has a prepayment penalty,
$TERM_{i,t}$ = control for the length of the mortgage.

A coefficient of interest is β which is the estimate of the jumbo, conforming loan spread. Estimates obtained in this study were roughly 16 basis points. A complication for obtaining the numerator of the elasticity (the response of the mortgage amount) is that the crucial variable *J* is endogenous in the model. This is suggested by the bunching of observations at the CLL that can be seen in the right panel of Figure 6.8. Essentially, individuals who would otherwise take out a jumbo loan near the boundary can take advantage of the lower rate by taking out a slightly smaller mortgage. The implication is that the unobservable characteristics of many individuals who are conforming loan borrowers are those of individuals who are in principle jumbo loan borrowers. The authors consider a semiparametric approach and an instrumental variable approach suggested by Kaufman (2012) (we return to this in Chapter 8) rather than a simple RD approach. (Results are obtained using both approaches.) The instrumental variable used is an indicator related to the appraised home value; the exogeneity of the indicator is argued because home buyers cannot control the appraisal of the home. In the terms developed for IVs in Chapter 8, the instrumental variable is certainly exogenous as it is not controlled by the borrower, and is certainly relevant through the correlation between the appraisal and the size of the mortgage. The main empirical result in the study is an estimate of the interest elasticity of the loan demand, which appears to be measurable at the loan limit. A further complication of the computation is that the increase in the cost of the loan at the loan limit associated with the interest rate increase is not marginal. The increased cost associated the increased interest rate is applied to the entire mortgage, not just the amount by which it exceeds the loan limit. Accounting for that aspect of the computation, the authors obtain estimates of the semi-elasticity ranging from -0.016 to -0.052. They find, for an example, that this suggests an increase in rates from 5% to 6% (a 20% increase) attends a 2% to 3% decrease in demand.

6.5 NONLINEARITY IN THE VARIABLES

It is useful at this point to write the linear regression model in a very general form: Let $\mathbf{z} = z_1, z_2, \ldots, z_L$ be a set of L independent variables; let $f_1, f_2, \ldots, f_K$ be K linearly independent functions of $\mathbf{z}$; let $g(y)$ be an observable function of y; and retain the usual assumptions about the disturbance. The linear regression model may be written

$$\begin{aligned} g(y) &= \beta_1 f_1(\mathbf{z}) + \beta_2 f_2(\mathbf{z}) + \cdots + \beta_K f_K(\mathbf{z}) + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon \\ &= \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon. \end{aligned} \tag{6-4}$$

By using logarithms, exponentials, reciprocals, transcendental functions, polynomials, products, ratios, and so on, this linear model can be tailored to any number of situations.

6.5.1 FUNCTIONAL FORMS

A commonly used form of regression model is the **loglinear model**,

$$\ln y = \ln \alpha + \sum_k \beta_k \ln X_k + \varepsilon = \beta_1 + \sum_k \beta_k x_k + \varepsilon.$$

In this model, the coefficients are elasticities:

$$\left(\frac{\partial y}{\partial X_k}\right)\left(\frac{X_k}{y}\right) = \frac{\partial \ln y}{\partial \ln X_k} = \beta_k. \tag{6-5}$$

In the loglinear equation, measured changes are in proportional or percentage terms; β_k measures the percentage change in y associated with a one percent change in X_k. This removes the units of measurement of the variables from consideration in using the regression model. For example, in Example 6.2, in our analysis of auction prices of Monet paintings, we found an elasticity of price with respect to area of 1.34935. (This is an extremely large value—the value well in excess of 1.0 implies that not only do sale prices rise with area, they rise considerably faster than area.)

An alternative approach sometimes taken is to measure the variables and associated changes in standard deviation units. If the data are standardized before estimation using $x^*_{ik} = (x_{ik} - \bar{x}_k)/s_k$ and likewise for y, then the least squares regression coefficients measure changes in standard deviation units rather than natural units or percentage terms. (Note that the constant term disappears from this regression.) It is not necessary actually to transform the data to produce these results; multiplying each least squares coefficient b_k in the original regression by s_k/s_y produces the same result.

A hybrid of the linear and loglinear models is the **semilog equation**

$$\ln y = \beta_1 + \beta_2 x + \varepsilon. \tag{6-6}$$

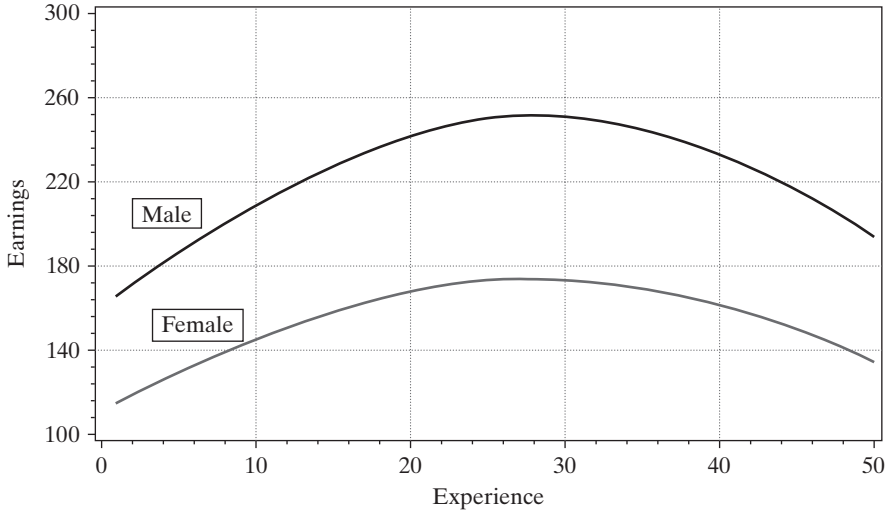
In a semilog equation with a time trend, $d \ln y/dt = \beta_2$ is the average rate of growth of y. The estimated values of 0.0750 and 0.0709 for day fees and boarding fees reported in Table 6.9 suggests that over the full estimation period, after accounting for all other factors, the average rate of growth of the fees was about 7% per year.

The coefficients in the semilog model are partial- or semi-elasticities; in (6-6), β_2 is $\partial \ln y/\partial x$. This is a natural form for models with dummy variables such as the earnings equation in Example 6.1. The coefficient on *Kids* of −0.35 suggests that all else equal, earnings are *approximately* 35% less when there are children in the household.

Example 6.15 Quadratic Regression

The quadratic earnings equation in Example 6.3 shows another use of nonlinearities in the variables. Using the results in Example 6.3, we find that the experience-wage profile appears as in Figure 6.8. This figure suggests an important question in this framework. It is tempting to conclude that Figure 6.8 shows the earnings trajectory of a person as experience accumulates. (The distinctive downturn is probably exaggerated by the use of a quadratic regression rather than a more flexible function.) But that is not what the data provide. The model is based on a cross section, and what it displays is the earnings of different people with different experience levels. How this profile relates to the expected earnings path of one individual is a different, and complicated, question.

Figure 6.9 Experience-Earnings Profile.



6.5.2 INTERACTION EFFECTS

Another useful formulation of the regression model is one with interaction terms. For example, the model for ln *Wage* in Example 6.3 might be extended to allow different partial effects of education for men and women with

$$\ln Wage = \beta_1 ED + \beta_2 FEM + \beta_3 ED \times FEM + \ \ldots \ + \varepsilon.$$

In this model,

$$\frac{\partial E[\ln Wage \,|\, ED, FEM, \ldots]}{\partial ED} = \beta_1 + \beta_3 FEM,$$

which implies that the **marginal effect** of education differs between men and women (assuming that β_3 is not zero).[18] If it is desired to form confidence intervals or test hypotheses about these marginal effects, then the necessary standard error is computed from

$$Var\left(\frac{\partial \hat{E}[\ln Wage \,|\, ED, FEM, \ldots]}{\partial ED}\right) = Var[\hat{\beta}_1] + FEM^2\, Var[\hat{\beta}_3] + 2FEM\, Cov[\hat{\beta}_1, \hat{\beta}_3].$$

(Because *FEM* is a dummy variable, $FEM^2 = FEM$.) The calculation is similar for

$$\begin{aligned}\Delta E[\ln Wage \,|\, ED, FEM, \ldots] &= E[\ln Wage \,|\, ED, FEM = 1, \ldots] - E[\ln Wage \,|\, ED, FEM = 0, \ldots] \\ &= \beta_2 + \beta_3 ED.\end{aligned}$$

[18] See Ai and Norton (2004) and Greene (2010) for further discussion of partial effects in models with interaction terms.

Example 6.16 Partial Effects in a Model with Interactions

We have extended the model in Example 6.3 by adding an interaction term between *FEM* and *ED*. The results for this part of the expanded model are

$$\ln Wage = \ldots + \underset{(0.00588)}{0.05250}\, ED - \underset{(0.15207)}{0.69799}\, FEM + \underset{(0.01055)}{0.02572}\, ED \times FEM + \ldots$$

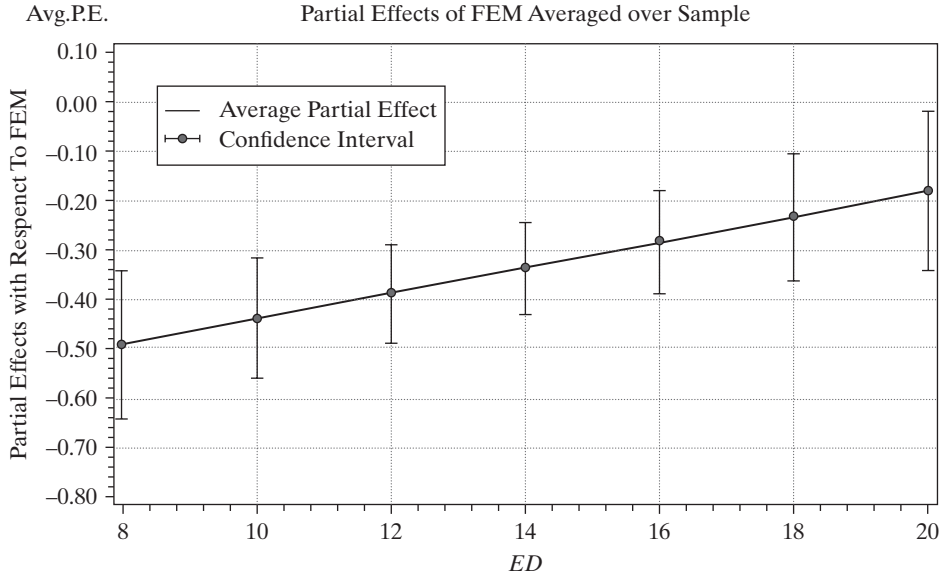
$$Est.Asy.Cov\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 0.0000345423 & & \\ 0.000349259 & 0.0231247 & \\ -0.0000243829 & -0.00152425 & 0.000111355 \end{bmatrix}.$$

The individual coefficients are not informative about the marginal impact of gender or education. The mean value of *ED* in the full sample is 12.8. The partial effect of a year increase in *ED* is 0.05250 (0.00588) for men and 0.05250 + 0.02572 = 0.07823 (0.00986) for women. The gender difference in earnings is $-0.69799 + 0.02572 \times ED$. At the mean value of *ED*, this is −0.36822. The standard error would be $(0.0231247 + 12.8^2(0.000111355) - 2(12.8)(0.00152425)^{1/2} = 0.04846$. A convenient way to summarize the information is a plot of the gender difference for the different values of *ED*, as in Figure 6.10. The figure reveals a richer interpretation of the model produced by the nonlinearity—the gender difference in wages is persistent, but does diminish at higher levels of education.

6.5.3 IDENTIFYING NONLINEARITY

If the functional form is not known a priori, then there are a few approaches that may help to identify any nonlinearity and provide some information about it from the sample. For example, if the suspected nonlinearity is with respect to a single regressor in the equation, then fitting a quadratic or cubic polynomial rather than a linear function may capture some of it. The residuals from a plot of the estimated function can also help to reveal the appropriate functional form.

Figure 6.10 Partial Effects in a Nonlinear Model.



Example 6.17 Functional Form for a Nonlinear Cost Function

In a pioneering study of economies of scale in the U.S. electric power industry, Nerlove (1963) analyzed the production costs of 145 American electricity generating companies. Economies of scale are typically modeled as a characteristic of the production function. Nerlove chose a Cobb–Douglas function to model output as a function of capital, K, labor, L, and fuel, F:

$$Q = \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F} e^{\varepsilon},$$

where Q is output and ε_i embodies the unmeasured differences across firms. The economies of scale parameter is $r = \alpha_K + \alpha_L + \alpha_F$. The value 1.0 indicates constant returns to scale. The production model is loglinear, so assuming that other conditions of the classical regression model are met, the four parameters could be estimated by least squares. But, for a firm that optimizes by choosing its factors of production, the demand for fuel would be $F^* = F^*(Q, P_K, P_L, P_F)$ and likewise for labor and capital. The three factor demands are endogenous and the assumptions of the classical model are violated.

In the regulatory framework in place at the time, state commissions set rates and firms met the demand forthcoming at the regulated prices. Thus, it was argued that output (as well as the factor prices) could be viewed as exogenous to the firm. Based on an argument by Zellner, Kmenta, and Dreze (1966), Nerlove argued that at equilibrium, the *deviation* of costs from the long-run optimum would be independent of output. The firm's objective was cost minimization subject to the constraint of the production function. This can be formulated as a Lagrangean problem,

$$\text{Min}_{K,L,F} P_K K + P_L L + P_F F + \lambda(Q - \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F}).$$

The solution to this minimization problem is the three factor demands and the multiplier (which measures marginal cost). Inserted back into total costs, this produces a loglinear cost function,

$$P_K K + P_L L + P_F F = C(Q, P_K, P_L, P_F) = rAQ^{1/r} P_K^{\alpha_K/r} P_L^{\alpha_L/r} P_F^{\alpha_F/r} e^{\varepsilon/r},$$

or

$$\ln C = \beta_1 + \beta_q \ln Q + \beta_K \ln P_K + \beta_L \ln P_L + \beta_F \ln P_F + u, \qquad \textbf{(6-7)}$$

where $\beta_q = 1/(\alpha_K + \alpha_L + \alpha_F)$ is now the parameter of interest and $\beta_j = \alpha_j/r$, $j = K, L, F$.

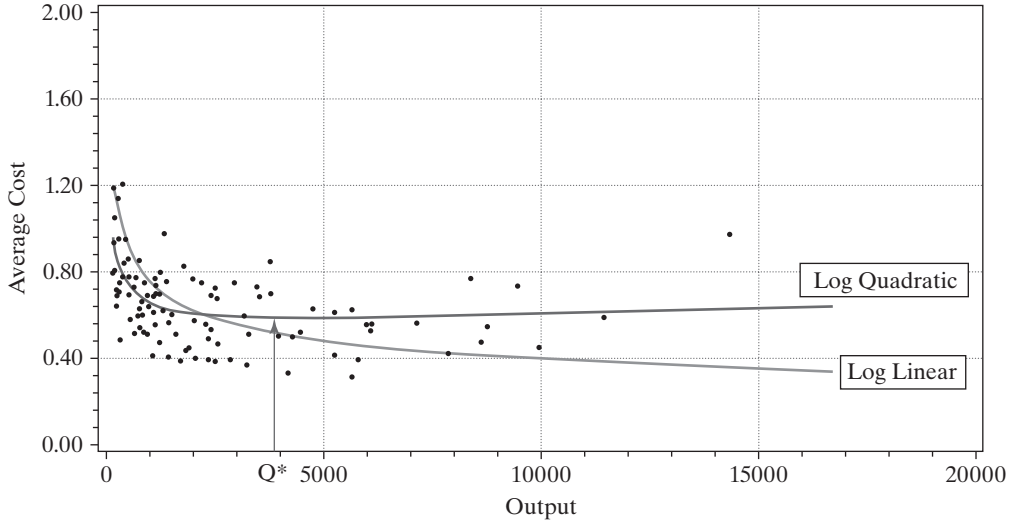
The cost parameters must sum to one; $\beta_K + \beta_L + \beta_F = 1$. This restriction can be imposed by regressing $\ln(C/P_F)$ on a constant, $\ln Q$, $\ln(P_K/P_F)$, and $\ln(P_L/P_F)$. Nerlove's results appear at the left of Table 6.10.[19] The hypothesis of constant returns to scale can be firmly rejected. The t ratio is $(0.721-1)/0.0174 = -16.03$, so we conclude that this estimate is significantly less than 1 or, by implication, r is significantly greater than 1. Note that the coefficient on the capital price is negative. In theory, this should equal α_K/r, which should be positive. Nerlove attributed this to measurement error in the capital price variable. The residuals in a plot of the average costs against the fitted loglinear cost function as in Figure 6.11 suggested that the Cobb-Douglas model was not picking up the increasing average costs at larger outputs, which would suggest diminished economies of scale. An approach used was to expand the cost function to include a quadratic term in log output. This approach corresponds to a more general model. Again, a simple t test strongly suggests that increased generality is called for; $t = 0.051/0.00054 = 9.44$. The output elasticity in this quadratic model is $\beta_q + 2\gamma_{qq} \log Q$. There are economies of scale when this value is less than 1 and constant returns to scale when it equals 1. Using the two values given in the table (0.152 and 0.0052, respectively), we

[19]Nerlove's data appear in Appendix Table F6.2. Figure 6.6 is constructed by computing the fitted log cost values using the means of the logs of the input prices. The plot then uses observations 31–145.

TABLE 6.10 Cobb–Douglas Cost Functions for log (C/P_F) based on 145 observations

	Log-linear			*Log-quadratic*		
Sum of squares	21.637			13.248		
R^2	0.932			0.958		
Variable	*Coefficient*	*Standard Error*	*t Ratio*	*Coefficient*	*Standard Error*	*t Ratio*
Constant	−4.686	0.885	−5.29	−3.764	0.702	−5.36
ln Q	0.721	0.0174	41.4	0.152	0.062	2.45
$\ln^2 Q$	0.000	0.000	—	0.051	0.0054	9.44
ln (P_L/P_F)	0.594	0.205	2.90	0.481	0.161	2.99
ln (P_K/P_F)	−0.0085	0.191	−0.045	0.074	0.150	0.49

Figure 6.11 Estimated Cost Functions.



find that this function does, indeed, produce a U-shaped average cost curve with minimum at ln $Q^* = (1 - 0.152)/(2 \times 0.051) = 8.31$, or $Q = 4079$. This is roughly in the middle of the range of outputs for Nerlove's sample of firms.

6.5.4 INTRINSICALLY LINEAR MODELS

The loglinear model illustrates a nonlinear regression model. The equation is **intrinsically linear**, however. By taking logs of $Y_i = \alpha X_i^{\beta_2} e^{\varepsilon_i}$, we obtain

$$\ln Y_i = \ln \alpha + \beta_2 \ln X_i + \varepsilon_i$$

or

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

Although this equation is linear in most respects, something has changed in that it is no longer linear in α. But, written in terms of β_1, we obtain a fully linear model. That may not be the form of interest, but nothing is lost because β_1 is just $\ln \alpha$. If β_1 can be estimated, then the obvious estimator of α is $\hat{\alpha} = \exp(b_1)$.

This fact leads us to a useful aspect of intrinsically linear models; they have an "invariance property." Using the nonlinear least squares procedure described in the next chapter, we could estimate α and β_2 directly by minimizing the sum of squares function:

$$\text{Minimize with respect to } (\alpha, \beta_2) : S(\alpha, \beta_2) = \sum_{i=1}^{n} (\ln Y_i - \ln \alpha - \beta_2 \ln X_i)^2. \quad \textbf{(6-8)}$$

This is a complicated mathematical problem because of the appearance of the term $\ln \alpha$. However, the equivalent linear least squares problem,

$$\text{Minimize with respect to } (\beta_1, \beta_2) : S(\beta_1, \beta_2) = \sum_{i=1}^{n} ,(y_i - \beta_1 - \beta_2 x_i)^2, \quad \textbf{(6-9)}$$

is simple to solve with the least squares estimator we have used up to this point. The invariance feature that applies is that the two sets of results will be numerically identical; we will get the identical result from estimating α using (6-8) and from using $\exp(\beta_1)$ from (6-9). By exploiting this result, we can broaden the definition of linearity and include some additional cases that might otherwise be quite complex.

DEFINITION 6.1 Intrinsic Linearity
In the linear regression model, if the K parameters $\beta_1, \beta_2, \ldots, \beta_K$ can be written as K one-to-one, possibly nonlinear functions of a set of K underlying parameters $\theta_1, \theta_2, \ldots, \theta_K$, then the model is intrinsically linear in θ.

Example 6.18 Intrinsically Linear Regression

In Section 14.6.4, we will estimate by maximum likelihood the parameters of the model

$$f(y|\beta, x) = \frac{(\beta + x)^{-\rho}}{\Gamma(\rho)} y^{\rho-1} e^{-y/(\beta+x)}.$$

In this model, $E[y|x] = (\beta\rho) + \rho x$, which suggests another way that we might estimate the two parameters. This function is an intrinsically linear regression model, $E[y|x] = \beta_1 + \beta_2 x$, in which $\beta_1 = \beta\rho$ and $\beta_2 = \rho$. We can estimate the parameters by least squares and then retrieve the estimate of β using b_1/b_2. Because this value is a nonlinear function of the estimated parameters, we use the delta method to estimate the standard error. Using the data from that example,[20] the least squares estimates of β_1 and β_2 (with standard errors in parentheses) are −4.1431 (23.734) and 2.4261 (1.5915). The estimated covariance is −36.979. The estimate of β is $-4.1431/2.4261 = -1.708$. We estimate the sampling variance of $\hat{\beta}$ with

$$\begin{aligned}\text{Est. Var}[\hat{\beta}] &= \left(\frac{\partial\hat{\beta}}{\partial b_1}\right)^2 \widehat{Var}[b_1] + \left(\frac{\partial\hat{\beta}}{\partial b_2}\right)^2 \widehat{Var}[b_2] + 2\left(\frac{\partial\hat{\beta}}{\partial b_1}\right)\left(\frac{\partial\hat{\beta}}{\partial b_2}\right)\widehat{Cov}[b_1, b_2] \\ &= 8.689^2.\end{aligned}$$

[20] The data are given in Appendix Table FC.1.

TABLE 6.11 Estimates of the Regression in a Gamma Model: Least Squares versus Maximum Likelihood

	$\boldsymbol{\beta}$		ρ	
	Estimate	***Standard Error***	***Estimate***	***Standard Error***
Least squares	−1.708	8.689	2.426	1.592
Maximum likelihood	−4.719	2.345	3.151	0.794

Table 6.11 compares the least squares and maximum likelihood estimates of the parameters. The lower standard errors for the maximum likelihood estimates result from the inefficient (equal) weighting given to the observations by the least squares procedure. The gamma distribution is highly skewed. In addition, we know from our results in Appendix C that this distribution is an exponential family. We found for the gamma distribution that the sufficient statistics for this density were $\Sigma_i y_i$ and $\Sigma_i \ln y_i$. The least squares estimator does not use the second of these, whereas an efficient estimator will.

The emphasis in intrinsic linearity is on "one to one." If the conditions are met, then the model can be estimated in terms of the functions $\beta_1, \ldots, \beta_K$, and the underlying parameters derived after these are estimated. The one-to-one correspondence is an **identification condition**. If the condition is met, then the underlying parameters of the regression ($\boldsymbol{\theta}$) are said to be **exactly identified** in terms of the parameters of the linear model $\boldsymbol{\beta}$. An excellent example is provided by Kmenta (1986, p. 515).

Example 6.19 CES Production Function

The constant elasticity of substitution production function may be written

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln [\delta K^{-\rho} + (1 - \delta) L^{-\rho}] + \varepsilon. \tag{6-10}$$

A Taylor series approximation to this function around the point $\rho = 0$ is

$$\begin{aligned} \ln y &= \ln \gamma + \nu\delta \ln K + \nu(1 - \delta) \ln L + \rho\nu\delta(1 - \delta) \left\{ -\tfrac{1}{2} [\ln K - \ln L]^2 \right\} + \varepsilon' \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon', \end{aligned} \tag{6-11}$$

where $x_1 = 1$, $x_2 = \ln K$, $x_3 = \ln L$, $x_4 = -\frac{1}{2} \ln^2(K/L)$, and the transformations are

$$\begin{aligned} &\beta_1 = \ln \gamma, \quad \beta_2 = \nu\delta, \quad \beta_3 = \nu(1 - \delta), \quad \beta_4 = \rho\nu\delta(1 - \delta), \\ &\gamma = e^{\beta_1}, \quad \delta = \beta_2/(\beta_2 + \beta_3), \quad \nu = \beta_2 + \beta_3, \quad \rho = \beta_4(\beta_2 + \beta_3)/(\beta_2\beta_3). \end{aligned} \tag{6-12}$$

Estimates of β_1, β_2, β_3, and β_4 can be computed by least squares. The estimates of γ, δ, ν, and ρ obtained by the second row of (6-12) are the same as those we would obtain had we found the nonlinear least squares estimates of (6-11) directly. [As Kmenta shows, however, they are not the same as the nonlinear least squares estimates of (6-10) due to the use of the Taylor series approximation to get to (6-11).] We would use the delta method to construct the estimated asymptotic covariance matrix for the estimates of $\boldsymbol{\theta}' = [\gamma, \delta, \nu, \rho]$. The derivatives matrix is

$$\mathbf{C} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}'} = \begin{bmatrix} e^{\beta_1} & 0 & 0 & 0 \\ 0 & \beta_3/(\beta_2 + \beta_3)^2 & -\beta_2/(\beta_2 + \beta_3)^2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -\beta_3\beta_4/(\beta_2^2\beta_3) & -\beta_2\beta_4/(\beta_2\beta_3^2) & (\beta_2 + \beta_3)/(\beta_2\beta_3) \end{bmatrix}.$$

The estimated covariance matrix for $\hat{\boldsymbol{\theta}}$ is $\hat{\mathbf{C}} \left\{ Asy.Var\left[\hat{\theta} \right] \right\} \hat{\mathbf{C}}'$.

Not all models of the form

$$y_i = \beta_1(\theta)x_{i1} + \beta_2(\theta)x_{i2} + \cdots + \beta_K(\theta)x_{ik} + \varepsilon_i \tag{6-13}$$

are intrinsically linear. Recall that the condition that the functions be one to one (i.e., that the parameters be exactly identified) was required. For example,

$$y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + \beta\gamma x_{i3} + \varepsilon_i$$

is nonlinear. The reason is that if we write it in the form of (6-13), we fail to account for the condition that β_4 equals $\beta_2\beta_3$, which is a **nonlinear restriction**. In this model, the three parameters α, β, and γ are **overidentified** in terms of the four parameters β_1, β_2, β_3, and β_4. Unrestricted least squares estimates of β_2, β_3, and β_4 can be used to obtain two estimates of each of the underlying parameters, and there is no assurance that these will be the same. Models that are not intrinsically linear are treated in Chapter 7.

6.6 STRUCTURAL BREAK AND PARAMETER VARIATION

One of the more common applications of hypothesis testing is in tests of **structural change**.[21] In specifying a regression model, we assume that its assumptions apply to all the observations in the sample. It is straightforward, however, to test the hypothesis that some or all of the regression coefficients are different in different subsets of the data. To analyze an example, we will revisit the data on the U.S. gasoline market that we examined in Examples 2.3 and 4.2. As Figure 4.2 suggests, this market behaved in predictable, unremarkable fashion prior to the oil shock of 1973 and was quite volatile thereafter. The large jumps in price in 1973 and 1980 are clearly visible, as is the much greater variability in consumption. It seems unlikely that the same regression model would apply to both periods.

6.6.1 DIFFERENT PARAMETER VECTORS

The gasoline consumption data span two very different periods. Up to 1973, fuel was plentiful and world prices for gasoline had been stable or falling for at least two decades. The embargo of 1973 marked a transition in this market, marked by shortages, rising prices, and intermittent turmoil. It is possible that the entire relationship described by the regression model changed in 1974. To test this as a hypothesis, we could proceed as follows: Denote the first 21 years of the data in $\mathbf{y}$ and $\mathbf{X}$ as $\mathbf{y}_1$ and $\mathbf{X}_1$ and the remaining years as $\mathbf{y}_2$ and $\mathbf{X}_2$. An unrestricted regression that allows the coefficients to be different in the two periods is

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}. \tag{6-14}$$

Denoting the data matrices as $\mathbf{y}$ and $\mathbf{X}$, we find that the unrestricted least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1'\mathbf{y}_1\\ \mathbf{X}_2'\mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix}, \tag{6-15}$$

[21] This test is often labeled a **Chow test**, in reference to Chow (1960).

which is least squares applied to the two equations separately. Therefore, the total sum of squared residuals from this regression will be the sum of the two residual sums of squares from the two separate regressions:

$$\mathbf{e}'\mathbf{e} = \mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2.$$

The restricted coefficient vector can be obtained by imposing a constraint on least squares. Formally, the restriction $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, where $\mathbf{R} = [\mathbf{I}:-\mathbf{I}]$ and $\mathbf{q} = \mathbf{0}$. The general result given earlier can be applied directly. An easy way to proceed is to build the restriction directly into the model. If the two coefficient vectors are the same, then (6-14) may be written

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix};$$

the restricted estimator can be obtained simply by stacking the data and estimating a single regression. The residual sum of squares from this restricted regression, $\mathbf{e}_*'\mathbf{e}_*$, then forms the basis for the test.

We begin by assuming that the disturbances are homoscedastic, nonautocorrelated, and normally distributed. More general cases are considered in the next section. Under these assumptions, the test statistic is given in (5-29), where J, the number of restrictions, is the number of columns in $\mathbf{X}_2$ and the denominator degrees of freedom is $n_1 + n_2 - 2K$. For this application,

$$F[K, n_1 + n_2 - 2K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}_1'\mathbf{e}_1 - \mathbf{e}_2'\mathbf{e}_2)/K}{(\mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2)/(n_1 + n_2 - 2K)}. \tag{6-16}$$

Example 6.20 Structural Break in the Gasoline Market

Figure 4.2 shows a plot of prices and quantities in the U.S. gasoline market from 1953 to 2004. The first 21 points are the layer at the bottom of the figure and suggest an orderly market. The remainder clearly reflect the subsequent turmoil in this market. We will use the Chow tests described to examine this market. The model we will examine is the one suggested in Example 2.3, with the addition of a time trend:

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln (Income/Pop)_t + \beta_3 \ln PG_t + \beta_4 \ln PNC_t + \beta_5 \ln PUC_t + \beta_6 t + \varepsilon_t.$$

The three prices in the equation are for *G*, new cars and used cars. *Income/Pop* is per capita Income, and *G/Pop* is per capita gasoline consumption. The time trend is computed as t = Year−1952, so in the first period $t = 1$. Regression results for three functional forms are shown in Table 6.12. Using the data for the entire sample, 1953 to 2004, and for the two subperiods, 1953 to 1973 and 1974 to 2004, we obtain the three estimated regressions in the first and last two columns. Using the full set of 52 observations to fit the model, the sum of squares is $\mathbf{e}_*'\mathbf{e}_* = 0.101997$. The *F* statistic for testing the restriction that the coefficients in the two equations are the same is

$$F[6, 40] = \frac{(0.101997 - (0.00202244 + 0.007127899))/6}{(0.00202244 + 0.007127899)/(21 + 31 - 12)} = 67.645.$$

The tabled critical value is 2.336, so, consistent with our expectations, we would reject the hypothesis that the coefficient vectors are the same in the two periods.

TABLE 6.12 Gasoline Consumption Functions

Coefficients	*1953–2004*	*1953–1973*	*1974–2004*
Constant	−26.6787	−22.1647	−15.3238
ln *Income/Pop*	1.6250	0.8482	0.3739
ln *PG*	−0.05392	−0.03227	−0.1240
ln *PNC*	−0.08343	0.6988	−0.001146
ln *PUC*	−0.08467	−0.2905	−0.02167
Year	−0.01393	0.01006	0.004492
R^2	0.9649	0.9975	0.9529
Standard error	0.04709	0.01161	0.01689
Sum of squares	0.101997	0.00202244	0.007127899

6.6.2 ROBUST TESTS OF STRUCTURAL BREAK WITH UNEQUAL VARIANCES

An important assumption made in using the Chow test is that the disturbance variance is the same in both (or all) regressions. In the restricted model, if this is not true, the first n_1 elements of ε have variance σ_1^2, whereas the next n_2 have variance σ_2^2, and so on. The restricted model is, therefore, heteroscedastic, and the results for normally distributed disturbances no longer apply. In several earlier examples, we have gone beyond heteroscedasticity, and based inference on robust specifications that also accommodate clustering and correlation across observations. In both settings, the results behind the F statistic in (6-16) will no longer apply. As analyzed by Schmidt and Sickles (1977), Ohtani and Toyoda (1985), and Toyoda and Ohtani (1986), it is quite likely that the actual probability of a type I error will be larger than the significance level we have chosen. (That is, we shall regard as large an F statistic that is actually less than the *appropriate* but unknown critical value.) Precisely how severe this effect is going to be will depend on the data and the extent to which the variances differ, in ways that are not likely to be obvious.

If the sample size is reasonably large, then we have a test that is valid whether or not the disturbance variances are the same. Suppose that $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are two consistent and asymptotically normally distributed estimators of a parameter based on independent samples, with asymptotic covariance matrices $\mathbf{V}_1$ and $\mathbf{V}_2$. Then, under the null hypothesis that the true parameters are the same,

$$\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2 \text{ has mean } \mathbf{0} \text{ and asymptotic covariance matrix } \mathbf{V}_1 + \mathbf{V}_2.$$

Under the null hypothesis, the Wald statistic,

$$W = (\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2)'(\hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2)^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2), \tag{6-17}$$

has a limiting chi-squared distribution with K degrees of freedom. A test that the difference between the parameters is zero can be based on this statistic.[22] It is straightforward to apply this to our test of common parameter vectors in our regressions. Large values of the statistic lead us to reject the hypothesis.

In a small or moderately sized sample, the Wald test has the unfortunate property that the probability of a type I error is persistently larger than the critical level we

[22] See Andrews and Fair (1988). The true size of this suggested test is uncertain. It depends on the nature of the alternative. If the variances are radically different, the assumed critical values might be somewhat unreliable.

use to carry it out. (That is, we shall too frequently reject the null hypothesis that the parameters are the same in the subsamples.) We should be using a larger critical value. Ohtani and Kobayashi (1986) have devised a "bounds" test that gives a partial remedy for the problem. In general, this test attains its validity in relatively large samples.

Example 6.21 Sample Partitioning by Gender

Example 6.3 considers the labor market experiences of a panel of 595 individuals, each observed 7 times. We have observed persistent differences between men and women in the relationship of log wages to various variables. It might be the case that different models altogether would apply to the two subsamples. We have fit the model in Example 6.3 separately for men and women (omitting *FEM* from the two regressions, of course), and calculated the Wald statistic in (6-17) based on the cluster corrected asymptotic covariance matrices as used in the pooled model as well. The chi-squared statistic with 17 degrees of freedom is 27.587, so the hypothesis of equal parameter vectors is rejected. The sums of squared residuals for the pooled data set for men and for women, respectively, are 416.988, 360.773, and 24.0848; the F statistic is 20.287 with critical value 1.625. This produces the same conclusion.

Example 6.22 The World Health Report

The 2000 version of the World Health Organization's (WHO) *World Health Report* contained a major country-by-country inventory of the world's health care systems. [World Health Organization (2000). See also http://www.who.int/whr/en/.] The book documented years of research and has thousands of pages of material. Among the most controversial and most publicly debated parts of the report was a single chapter that described a comparison of the delivery of health care by 191 countries—nearly all of the world's population. [Evans et al. (2000a,b). See, e.g., Hilts (2000) for reporting in the popular press.] The study examined the efficiency of health care delivery on two measures: the standard one that is widely studied, (disability adjusted) life expectancy (DALE), and an innovative new measure created by the authors that was a composite of five outcomes (COMP) and that accounted for efficiency and fairness in delivery. The regression-style modeling, which was done in the setting of a frontier model (see Section 19.2.4), related health care attainment to two major inputs, education and (per capita) health care expenditure. The residuals were analyzed to obtain the country comparisons.

The data in Appendix Table F6.3 were used by the researchers at the WHO for the study. (They used a panel of data for the years 1993 to 1997. We have extracted the 1997 data for this example.) The WHO data have been used by many researchers in subsequent analyses.[23] The regression model used by the WHO contained DALE or COMP on the left-hand side and health care expenditure, education, and education squared on the right. Greene (2004b) added a number of additional variables such as per capita GDP, a measure of the distribution of income, and World Bank measures of government effectiveness and democratization of the political structure.

Among the controversial aspects of the study was the fact that the model aggregated countries of vastly different characteristics. A second striking aspect of the results, suggested in Hilts (2000) and documented in Greene (2004b), was that, in fact, the "efficient" countries in the study were the 30 relatively wealthy OECD members, while the rest of the world on average fared much more poorly. We will pursue that aspect here with respect to DALE. Analysis of COMP is left as an exercise. Table 6.8 presents estimates of the regression models for DALE for the pooled sample, the OECD countries, and the non-OECD countries, respectively. Superficially, there do not appear to be very large differences across the two subgroups. We first tested the joint significance of the additional variables, income distribution (GINI), per

[23] See, for example, Hollingsworth and Wildman (2002), Gravelle et al. (2002), and Greene (2004b).

TABLE 6.13 Regression Results for Life Expectancy

	All Countries		*OECD*		*Non-OECD*	
Constant	25.237	38.734	42.728	49.328	26.816	41.408
Health exp	0.00629	−0.00180	0.00268	0.00114	0.00955	−0.00178
Education	7.931	7.178	6.177	5.156	7.0433	6.499
Education^2	−0.439	−0.426	−0.385	−0.329	−0.374	−0.372
Gini coeff		−17.333		−5.762		−21.329
Tropic		−3.200		−3.298		−3.144
Pop. Dens.		−0.255e−4		0.000167		−0.425e−4
Public exp		−0.0137		−0.00993		−0.00939
PC GDP		0.000483		0.000108		0.000600
Democracy		1.629		−0.546		1.909
Govt. Eff.		0.748		1.224		0.786
R^2	0.6824	0.7299	0.6483	0.7340	0.6133	0.6651
Std. Err.	6.984	6.565	1.883	1.916	7.366	7.014
Sum of sq.	9121.795	7757.002	92.21064	69.74428	8518.750	7378.598
N	191		30		161	
GDP/Pop	6609.37		18199.07		4449.79	
F test	4.524		0.874		3.311	

capita GDP, and so on. For each group, the *F* statistic is $[(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/7]/[\mathbf{e}'\mathbf{e}/(n - 11)]$. These *F* statistics are shown in the last row of the table. The critical values for *F*[7,180] (all), *F*[7,19] (OECD), and *F*[7,150] (non-OECD) are 2.061, 2.543, and 2.071, respectively. We conclude that the additional explanatory variables are significant contributors to the fit for the non-OECD countries (and for all countries), but not for the OECD countries. Finally, to conduct the structural change test of OECD vs. non-OECD, we computed

$$F[11, 169] = \frac{[7757.002 - (69.74428 + 7378.598)]/11}{(69.74428 + 7378.598)/(191 - 11 - 11)} = 0.637.$$

The 95% critical value for *F*[11,169] is 1.846. So, we do not reject the hypothesis that the regression model is the same for the two groups of countries. The Wald statistic in (6-17) tells a different story. The statistic is 35.221. The 95% critical value from the chi-squared table with 11 degrees of freedom is 19.675. On this basis, we would reject the hypothesis that the two coefficient vectors are the same.

6.6.3 POOLING REGRESSIONS

Extending the homogeneity test to multiple groups or periods should be straightforward. As usual, we begin with independent and identically normally distributed disturbances. Assume there are G groups or periods. (In Example 6.3, we are examining 7 years of observations.) The direct extension of the F statistic in (6-16) would be

$$F[(G - 1)K, \Sigma_{g=1}^{G}(n_g - K)] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \Sigma_{g=1}^{G}\mathbf{e}_g'\mathbf{e}_g)/(G - 1)K}{(\Sigma_{g=1}^{G}\mathbf{e}_g'\mathbf{e}_g)/\Sigma_{g=1}^{G}(n_g - K)}. \quad \textbf{(6-18)}$$

To apply (6-18) to a more general case, begin with the simpler setting of possible heteroscedasticity. Then, we can consider a set of G estimators, $\mathbf{b}_g$, each with associated

asymptotic covariance matrix $\mathbf{V}_g$. A Wald test along the lines of (6-17) can be carried out by testing $H_0: \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 = \mathbf{0}, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_3 = \mathbf{0}, \ldots, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_G = \mathbf{0}$. This can be based on G sets of least squares results. The Wald statistic is

$$W = (\mathbf{Rb})'(\mathbf{R}(Asy.\ Var[\mathbf{b}])\mathbf{R}')^{-1}(\mathbf{Rb}), \tag{6-19}$$

where

$$\mathbf{R} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & -\mathbf{I} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & -\mathbf{I} \end{bmatrix}; \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \cdots \\ \mathbf{b}_G \end{pmatrix}. \tag{6-20}$$

The results in (6-19) and (6-20) are straightforward based on G separate regressions. For example, to test equality of the coefficient vectors for three periods, (6-19) and (6-20) would produce

$$W = [(\mathbf{b}_1 - \mathbf{b}_2)' \quad (\mathbf{b}_1 - \mathbf{b}_3)'] \begin{bmatrix} (\mathbf{V}_1 + \mathbf{V}_2) & \mathbf{V}_1 \\ \mathbf{V}_1 & (\mathbf{V}_1 + \mathbf{V}_3) \end{bmatrix}^{-1} \begin{bmatrix} (\mathbf{b}_1 - \mathbf{b}_2) \\ (\mathbf{b}_1 - \mathbf{b}_3) \end{bmatrix}.$$

The computations are rather more complicated when observations are correlated, as in a panel. In Example 6.3, we are examining seven periods of data but robust calculation of the covariance matrix for the estimates results in correlation across the observations within a group. The implication for current purposes would be that we are not using independent samples for the G estimates of β_g. The following practical strategy for this computation is suggested for the particular application—extensions to other settings should be straightforward. We have seven years of data for individual i, with regression specification

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

For each individual, we construct

$$\widetilde{\mathbf{X}}_i = \begin{bmatrix} \mathbf{x}_{i1}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{x}_{i2}' & \cdots & \mathbf{0}' \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{x}_{i7}' \end{bmatrix} \quad \text{and} \quad \begin{pmatrix} y_{i1} \\ y_{i2} \\ \cdots \\ y_{i7} \end{pmatrix}.$$

Then, the $7K \times 1$ vector of estimated coefficient vectors is computed by least squares,

$$\mathbf{b} = \left[\sum_{i=1}^{595} \widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i \right]^{-1} \sum_{i=1}^{595} \widetilde{\mathbf{X}}_i'\widetilde{y}_i$$

The estimator of the asymptotic covariance matrix of $\mathbf{b}$ is the cluster estimator from (4-41) and (4-42),

$$Est.Asy.Var[\mathbf{b}] = \left[\sum_{i=1}^{595} \widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i \right]^{-1} \left\{ \sum_{i=1}^{595} (\widetilde{\mathbf{X}}_i'\mathbf{e}_i)(\mathbf{e}_i'\widetilde{\mathbf{X}}_i) \right\} \left[\sum_{i=1}^{595} \widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i \right]^{-1}. \tag{6-21}$$

Example 6.23 Pooling in a Log Wage Model

Using the data and model in Example 6.3, the sums of squared residuals are as follows:

1976: 44.3242	1977: 38.7594	1978: 63.9203	1979: 61.4599
1980: 54.9996	1981: 58.6650	1982: 62.9827	Pooled: 513.767

The F statistic based on (6-18) is 14.997. The 95% critical value from the F table with 6×12 and (4165-84) degrees of freedom is 1.293. The large sample approximation for this statistic would be $72(14.997) = 1079.776$ with 72 degrees of freedom. The 95% critical value for the chi-squared distribution with 72 degrees of freedom is 92.808, which is slightly less than 72(1.293). The Wald statistic based on (6-19) using (6-21) to compute the asymptotic covariance matrix is 3068.78 with 72 degrees of freedom. Finally, the Wald statistic based on (6-19) and 7 separate estimates, allowing different variances, is 1478.62. All versions of the test procedure produce the same conclusion. The homogeneity restriction is decisively rejected. We note, this conclusion gives no indication of the nature of the change from year to year.

6.7 SUMMARY AND CONCLUSIONS

This chapter has discussed the functional form of the regression model. We examined the use of dummy variables and other transformations to build nonlinearity into the model to accommodate specific features of the environment, such as the effects of discrete changes in policy. We then considered other nonlinear models in which the parameters of the nonlinear model could be recovered from estimates obtained for a linear regression. The final sections of the chapter described hypothesis tests designed to reveal whether the assumed model had changed during the sample period, or was different for different groups of observations.

Key Terms and Concepts

- Binary variable
- Chow test
- Control group
- Control observations
- Difference in differences
- Dummy variable
- Dummy variable trap
- Dynamic linear regression model
- Exactly identified
- Fuzzy design
- Identification condition
- Interaction terms
- Intrinsically linear
- Loglinear model
- Marginal effect
- Natural experiment
- Nonlinear restriction
- Overidentified
- Placebo effect
- Regression discontinuity design
- Regression kink design
- Response
- Semilog equation
- Structural change
- Treatment
- Treatment group
- Unobserved heterogeneity

Exercises

1. A regression model with $K = 16$ independent variables is fit using a panel of seven years of data. The sums of squares for the seven separate regressions and the pooled regression are shown below. The model with the pooled data allows a separate constant for each year. Test the hypothesis that the same coefficients apply in every year.

	2004	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*All*
Observations	65	55	87	95	103	87	78	570
$\mathbf{e}'\mathbf{e}$	104	88	206	144	199	308	211	1425

2. *Reverse regression.* A method of analyzing statistical data to detect discrimination in the workplace is to fit the regression

$$y = \alpha + \mathbf{x}'\beta + \gamma d + \varepsilon, \tag{1}$$

where y is the wage rate and d is a dummy variable indicating either membership ($d = 1$) or nonmembership ($d = 0$) in the class toward which it is suggested the discrimination is directed. The regressors $\mathbf{x}$ include factors specific to the particular type of job as well as indicators of the qualifications of the individual. The hypothesis of interest is $H_0{:}\gamma \geq 0$ versus $H_1{:}\gamma < 0$. The regression seeks to answer the question, "In a given job, are individuals in the class ($d = 1$) paid less than equally qualified individuals not in the class ($d = 0$)?" Consider an alternative approach. Do individuals in the class in the same job as others, and receiving the same wage, uniformly have higher qualifications? If so, this might also be viewed as a form of discrimination. To analyze this question, Conway and Roberts (1983) suggested the following procedure:

1. Fit (1) by ordinary least squares. Denote the estimates a, $\mathbf{b}$, and c.
2. Compute the set of qualification indices,

$$\mathbf{q} = a\mathbf{i} + \mathbf{X}\mathbf{b}. \tag{2}$$

   Note the omission of $c\mathbf{d}$ from the fitted value.
3. Regress $\mathbf{q}$ on a constant, $\mathbf{y}$ and $\mathbf{d}$. The equation is

$$\mathbf{q} = \alpha_* + \beta_*\mathbf{y} + \gamma_*\mathbf{d} + \varepsilon_*. \tag{3}$$

   The analysis suggests that if $\gamma < 0$, then $\gamma_* > 0$.

a. Prove that the theory notwithstanding, the least squares estimates c and c^* are related by

$$c_* = \frac{(\bar{y}_1 - \bar{y})(1 - R^2)}{(1 - P)(1 - r_{yd}^2)} - c, \tag{4}$$

where

$\bar{y}_1$ = mean of y for observations with $d = 1$,
$\bar{y}$ = mean of y for all observations,
P = mean of d,
R^2 = coefficient of determination for (1),
r_{yd}^2 = squared correlation between y and d.

[*Hint:* The model contains a constant term]. Thus, to simplify the algebra, assume that all variables are measured as deviations from the overall sample means and use a partitioned regression to compute the coefficients in (3). Second, in (2), use the result that based on the least squares results $\mathbf{y} = a\mathbf{i} + \mathbf{X}\mathbf{b} + c\mathbf{d} + \mathbf{e}$, so $\mathbf{q} = \mathbf{y} - c\mathbf{d} - \mathbf{e}$. From here on, we drop the constant term. Thus, in the regression in (3) you are regressing $[\mathbf{y} - c\mathbf{d} - \mathbf{e}]$ on $\mathbf{y}$ and $\mathbf{d}$.

b. Will the sample evidence necessarily be consistent with the theory? [*Hint:* Suppose that $c = 0$.]

A symposium on the Conway and Roberts paper appeared in the *Journal of Business and Economic Statistics* in April 1983.

3. *Reverse regression continued.* This and the next exercise continue the analysis of Exercise 2. In Exercise 2, interest centered on a particular dummy variable in which the regressors were accurately measured. Here we consider the case in which the

crucial regressor in the model is measured with error. The paper by Kamlich and Polachek (1982) is directed toward this issue.

Consider the simple errors in the variables model,

$$y = \alpha + \beta x^* + \varepsilon, \qquad x = x^* + u,$$

where u and ε are uncorrelated and x is the erroneously measured, observed counterpart to x*.

a. Assume that x^*, u, and ε are all normally distributed with means μ^*, 0, and 0, variances σ_*^2, σ_u^2, and σ_ε^2, and zero covariances. Obtain the probability limits of the least squares estimators of α and β.

b. As an alternative, consider regressing x on a constant and y, and then computing the reciprocal of the estimate. Obtain the probability limit of this estimator.

c. Do the "direct" and "reverse" estimators bound the true coefficient?

4. *Reverse regression continued*. Suppose that the model in Exercise 3 is extended to $y = \beta x^* + \gamma d + \varepsilon, x = x^* + u$. For convenience, we drop the constant term. Assume that x^*, ε, and u are independent normally distributed with zero means. Suppose that d is a random variable that takes the values one and zero with probabilities π and $1 - \pi$ in the population and is independent of all other variables in the model. To put this formulation in context, the preceding model (and variants of it) have appeared in the literature on discrimination. We view y as a "wage" variable, x^* as "qualifications," and x as some imperfect measure such as education. The dummy variable, d, is membership ($d = 1$) or nonmembership ($d = 0$) in some protected class. The hypothesis of discrimination turns on $\gamma < 0$ versus $\gamma \geq 0$.

a. What is the probability limit of c, the least squares estimator of γ, in the least squares regression of y on x and d? [*Hints:* The independence of x^* and d is important. Also, plim $\mathbf{d}'\mathbf{d}/n = \text{Var}[d] + E^2[d] = \pi(1 - \pi) + \pi^2 = \pi$. This minor modification does not affect the model substantively, but it greatly simplifies the algebra.] Now suppose that x^* and d are not independent. In particular, suppose that $E[x^* | d = 1] = \mu^1$ and $E[x^* | d = 0] = \mu^0$. Repeat the derivation with this assumption.

b. Consider, instead, a regression of x on y and d. What is the probability limit of the coefficient on d in this regression? Assume that x^* and d are independent.

c. Suppose that x^* and d are not independent, but γ is, in fact, less than zero. Assuming that both preceding equations still hold, what is estimated by $(\bar{y} | d = 1) - (\bar{y} | d = 0)$? What does this quantity estimate if γ does equal zero?

5. *Dummy variable for one observation*. Suppose the data set consists of n observations, $(\mathbf{y}_n, \mathbf{X}_n)$ and an additional observation, $(y_s, \mathbf{x}_s')$. The full data set contains a dummy variable, $\mathbf{d}$, that equals zero save for one (the last) observation. Then, the full data set is

$$(\mathbf{X}_{n,s}, \mathbf{d}_{n,s}) = \begin{bmatrix} \mathbf{X}_n & \mathbf{0} \\ \mathbf{x}_s' & \mathbf{1} \end{bmatrix} \text{ and } \mathbf{y}_{n,s} = \begin{bmatrix} \mathbf{y}_n \\ y_s \end{bmatrix}.$$

It is claimed in the text that in the *full* regression of $\mathbf{y}_{n,s}$ on $(\mathbf{X}_{n,s}, \mathbf{d}_{n,s})$ using all $n+1$ observations, the slopes on $\mathbf{X}_{n,s}$, $\mathbf{b}_{n,s}$, and their estimated standard errors will be the same as those on $\mathbf{X}_n$, $\mathbf{b}_n$ in the *short* regression of $\mathbf{y}_n$ on $\mathbf{X}_n$, and the sum of squared residuals in the full regression will be the same as the sum of squared residuals in

the short regression. That is, the last observation will be ignored. However, the R^2 in the full regression will not be the same as the R^2 in the short regression. Prove these results.

Applications

1. In Application 1 in Chapter 3 and Application 1 in Chapter 5, we examined Koop and Tobias's data on wages, education, ability, and so on. We continue the analysis here. (The source, location and configuration of the data are given in the earlier application.) We consider the model

$$\begin{aligned}\ln Wage = {} & \beta_1 + \beta_2 Educ + \beta_3 Ability + \beta_4 Experience \\ & + \beta_5 Mother's\ education + \beta_6 Father's\ education + \beta_7 Broken\ home \\ & + \beta_8 Siblings + \varepsilon.\end{aligned}$$

   a. Compute the full regression by least squares and report your results. Based on your results, what is the estimate of the marginal value, in $/hour, of an additional year of education, for someone who has 12 years of education when all other variables are at their means and *Broken home* = 0?
   b. We are interested in possible nonlinearities in the effect of education on ln *Wage*. (Koop and Tobias focused on experience. As before, we are not attempting to replicate their results.) A histogram of the education variable shows values from 9 to 20, a spike at 12 years (high school graduation), and a second at 15. Consider aggregating the education variable into a set of dummy variables:

$$\begin{aligned} HS &= 1 \text{ if } Educ \leq 12, 0 \text{ otherwise} \\ Col &= 1 \text{ if } Educ > 12 \text{ and } Educ \leq 16, 0 \text{ otherwise} \\ Grad &= 1 \text{ if } Educ > 16, 0 \text{ otherwise.} \end{aligned}$$

   Replace *Educ* in the model with (*Col, Grad*), making high school (*HS*) the base category, and recompute the model. Report all results. How do the results change? Based on your results, what is the marginal value of a college degree? What is the marginal impact on ln *Wage* of a graduate degree?
   c. The aggregation in part b actually loses quite a bit of information. Another way to introduce nonlinearity in education is through the function itself. Add $Educ^2$ to the equation in part a and recompute the model. Again, report all results. What changes are suggested? Test the hypothesis that the quadratic term in the equation is not needed—that is, that its coefficient is zero. Based on your results, sketch a profile of log wages as a function of education.
   d. One might suspect that the value of education is enhanced by greater ability. We could examine this effect by introducing an interaction of the two variables in the equation. Add the variable

$$Educ\_Ability = Educ \times Ability$$

   to the base model in part a. Now, what is the marginal value of an additional year of education? The sample mean value of ability is 0.052374. Compute a confidence interval for the marginal impact on ln *Wage* of an additional year of education for a person of average ability.

e. Combine the models in c and d. Add both $Educ^2$ and *Educ_Ability* to the base model in part a and reestimate. As before, report all results and describe your findings. If we define *low ability* as less than the mean and *high ability* as greater than the mean, the sample averages are −0.798563 for the 7,864 low-ability individuals in the sample and +0.717891 for the 10,055 high-ability individuals in the sample. Using the formulation in part c, with this new functional form, sketch, describe, and compare the log wage profiles for low- and high-ability individuals.

2. (An extension of Application 1.) Here we consider whether different models as specified in Application 1 would apply for individuals who reside in "Broken homes." Using the results in Section 6.6, test the hypothesis that the same model (not including the *Broken home* dummy variable) applies to both groups of individuals, those with *Broken home* = 0 and with *Broken home* = 1.
3. In Solow's classic (1957) study of technical change in the U.S. economy, he suggests the following aggregate production function: $q(t) = A(t)f[k(t)]$, where $q(t)$ is aggregate output per work hour, $k(t)$ is the aggregate capital labor ratio, and $A(t)$ is the technology index. Solow considered four static models, $q/A = \alpha + \beta \ln k$, $q/A = \alpha - \beta/k$, $\ln(q/A) = \alpha + \beta \ln k$, and $\ln(q/A) = \alpha + \beta/k$. Solow's data for the years 1909 to 1949 are listed in Appendix Table F6.4.
   a. Use these data to estimate the α and β of the four functions listed above. (*Note:* Your results will not quite match Solow's. See the next exercise for resolution of the discrepancy.)
   b. In the aforementioned study, Solow states:

      > A scatter of q / A against k is shown in Chart 4. Considering the amount of a priori doctoring which the raw figures have undergone, the fit is remarkably tight. Except, that is, for the layer of points which are obviously too high. These maverick observations relate to the seven last years of the period, 1943–1949. From the way they lie almost exactly parallel to the main scatter, one is tempted to conclude that in 1943 the aggregate production function simply shifted.

      Compute a scatter diagram of q / A against k and verify the result he notes above.
   c. Estimate the four models you estimated in the previous problem including a dummy variable for the years 1943 to 1949. How do your results change? (*Note:* These results match those reported by Solow, although he did not report the coefficient on the dummy variable.)
   d. Solow went on to surmise that, in fact, the data were fundamentally different in the years before 1943 than during and after. Use a Chow test to examine the difference in the two subperiods using your four functional forms. Note that with the dummy variable, you can do the test by introducing an interaction term between the dummy and whichever function of k appears in the regression. Use an F test to test the hypothesis.

7

NONLINEAR, SEMIPARAMETRIC, AND NONPARAMETRIC REGRESSION MODELS

7.1 INTRODUCTION

Up to this point, our focus has been on the **linear regression model**,

$$y = x_1\beta_1 + x_2\beta_2 + \cdots + \varepsilon. \tag{7-1}$$

Chapters 2 through 5 developed the least squares method of estimating the parameters and obtained the statistical properties of the estimator that provided the tools we used for point and interval estimation, hypothesis testing, and prediction. The modifications suggested in Chapter 6 provided a somewhat more general form of the linear regression model,

$$y = f_1(\mathbf{x})\beta_1 + f_2(\mathbf{x})\beta_2 + \cdots + \varepsilon. \tag{7-2}$$

By the definition we want to use in this chapter, this model is still "linear" because the parameters appear in a linear form. Section 7.2 of this chapter will examine the **nonlinear regression model** [which includes (7-1) and (7-2) as special cases],

$$y = h(x_1, x_2, \ldots, x_P; \beta_1, \beta_2, \ldots, \beta_K) + \varepsilon, \tag{7-3}$$

where the conditional mean function involves P variables and K parameters. This form of the model changes the conditional mean function from $E[y|\mathbf{x}, \boldsymbol{\beta}] = \mathbf{x}'\boldsymbol{\beta}$ to $E[y|\mathbf{x}] = h(\mathbf{x}, \beta)$ for more general functions. This allows a much wider range of functional forms than the linear model can accommodate.[1] This change in the model form will require us to develop an alternative method of estimation, **nonlinear least squares**. We will also examine more closely the interpretation of parameters in nonlinear models. In particular, since $\partial E[y|\mathbf{x}]/\partial\mathbf{x}$ is no longer equal to $\boldsymbol{\beta}$, we will want to examine how $\boldsymbol{\beta}$ should be interpreted.

Linear and nonlinear least squares are used to estimate the parameters of the **conditional mean function**, $E[y|\mathbf{x}]$. As we saw in Example 4.3, other relationships between y and $\mathbf{x}$, such as the **conditional median**, might be of interest. Section 7.3 revisits this idea with an examination of the conditional median function and the least absolute deviations estimator. This section will also relax the restriction that the model coefficients are always the same in the different parts of the distribution

[1] A complete discussion of this subject can be found in Amemiya (1985). Another authoritative treatment is the text by Davidson and MacKinnon (1993).

of y (given $\mathbf{x}$). The LAD estimator estimates the parameters of the conditional median, that is, the 50th percentile function. The **quantile regression model** allows the parameters of the regression to change as we analyze different parts of the conditional distribution.

The model forms considered thus far are semiparametric in nature and less parametric as we move from Section 7.2 to 7.3. The **partially linear regression** examined in Section 7.4 extends (7-1) such that $y = f(z) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon$. The endpoint of this progression is a model in which the relationship between y and x is not forced to conform to a particular parameterized function. Using largely graphical and kernel density methods, we consider in Section 7.5 how to analyze a **nonparametric regression** relationship that essentially imposes little more than $E[y|\mathbf{x}] = h(\mathbf{x})$.

7.2 NONLINEAR REGRESSION MODELS

The general form of the nonlinear regression model is

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i. \tag{7-4}$$

The linear model is obviously a special case. Moreover, some models that appear to be nonlinear, such as

$$y = e^{\beta_1} x_1^{\beta_2} x_2^{\beta_3} e^{\varepsilon},$$

become linear after a transformation, in this case, after taking logarithms. In this chapter, we are interested in models for which there are no such transformations.

Example 7.1 CES Production Function

In Example 6.18, we examined a constant elasticity of substitution production function model,

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln\left[\delta K^{-\rho} + (1 - \delta) L^{-\rho}\right] + \varepsilon. \tag{7-5}$$

No transformation reduces this equation to one that is linear in the parameters. In Example 6.5, a linear Taylor series approximation to this function around the point $\rho = 0$ is used to produce an intrinsically linear equation that can be fit by least squares. The underlying model in (7-5) is nonlinear.

This and the next section will extend the assumptions of the linear regression model to accommodate nonlinear functional forms such as the one in Example 7.1. We will then develop the nonlinear least squares estimator, establish its statistical properties, and then consider how to use the estimator for hypothesis testing and analysis of the model predictions.

7.2.1 ASSUMPTIONS OF THE NONLINEAR REGRESSION MODEL

We shall require a somewhat more formal definition of a nonlinear regression model. Sufficient for our purposes will be the following, which include the linear model as the special case noted earlier. We assume that there is an underlying probability distribution, or data-generating process (DGP) for the observable y_i and a true parameter vector, $\boldsymbol{\beta}$,

which is a characteristic of that DGP. The following are the assumptions of the nonlinear regression model:

NR1. Functional form: The conditional mean function for y_i given $\mathbf{x}_i$ is

$$E[y_i|\mathbf{x}_i] = h(\mathbf{x}_i, \boldsymbol{\beta}), \quad i = 1, \ldots, n,$$

where $h(\mathbf{x}_i, \boldsymbol{\beta})$ is a continuously differentiable function of $\boldsymbol{\beta}$.

NR2. Identifiability of the model parameters: The parameter vector in the model is identified (estimable) if there is no nonzero parameter $\boldsymbol{\beta}^0 \neq \boldsymbol{\beta}$ such that $h(\mathbf{x}_i, \boldsymbol{\beta}^0) = h(\mathbf{x}_i, \boldsymbol{\beta})$ for all $\mathbf{x}_i$. In the linear model, this was the full rank assumption, but the simple absence of "multicollinearity" among the variables in $\mathbf{x}$ is not sufficient to produce this condition in the nonlinear regression model. Example 7.2 illustrates the problem. Full rank will be necessary, but it is not sufficient.

NR3. Zero conditional mean of the disturbance: It follows from Assumption 1 that we may write

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,$$

where $E[\varepsilon_i|h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$. This states that the disturbance at observation i is uncorrelated with the conditional mean function for all observations in the sample. This is not quite the same as assuming that the disturbances and the exogenous variables are uncorrelated, which is the familiar assumption, however. We will want to assume that $\mathbf{x}$ is exogenous in this setting, so added to this assumption will be $\mathrm{E}[\varepsilon|\mathbf{x}] = 0$.

NR4. Homoscedasticity and nonautocorrelation: As in the linear model, we assume conditional homoscedasticity,

$$E[\varepsilon_i^2|h(\mathbf{x}_j, \boldsymbol{\beta}), \quad j = 1, \ldots, n] = \sigma^2, \qquad \text{a finite constant,} \tag{7-6}$$

and nonautocorrelation,

$$E[\varepsilon_i\varepsilon_j|h(\mathbf{x}_i, \boldsymbol{\beta}), h(\mathbf{x}_j, \boldsymbol{\beta}), \quad j = 1, \ldots, n] = 0 \qquad \text{for all } j \neq i.$$

This assumption parallels the specification of the linear model in Chapter 4. As before, we will want to relax these assumptions.

NR5. Data generating process: The DGP for $\mathbf{x}_i$ is assumed to be a well-behaved population such that first and second moments of the data can be assumed to converge to fixed, finite population counterparts. The crucial assumption is that the process generating $\mathbf{x}_i$ is strictly exogenous to that generating ε_i. The data on $\mathbf{x}_i$ are assumed to be "well behaved."

NR6. Underlying probability model: There is a well-defined probability distribution generating ε_i. At this point, we assume only that this process produces a sample of uncorrelated, identically (marginally) distributed random variables ε_i with mean zero and variance σ^2 conditioned on $h(\mathbf{x}_i, \beta)$. Thus, at this point, our statement of the model is **semiparametric**. (See Section 12.3.) We will not be assuming any particular distribution for ε_i. The conditional moment assumptions in 3 and 4 will be sufficient for the results in this chapter.

Example 7.2 Identification in a Translog Demand System

Christensen, Jorgenson, and Lau (1975), proposed the translog **indirect utility function** for a consumer allocating a budget among K commodities,

$$\ln V = \beta_0 + \sum_{k=1}^{K} \beta_k \ln(p_k/M) + \sum_{k=1}^{K}\sum_{j=1}^{K} \gamma_{kj} \ln(p_k/M) \ln(p_j/M),$$

where V is indirect utility, p_k is the price for the kth commodity, and M is income. Utility, direct or indirect, is unobservable, so the utility function is not usable as an empirical model. **Roy's identity** applied to this logarithmic function produces a budget share equation for the kth commodity that is of the form

$$S_k = -\frac{\partial \ln V / \partial \ln p_k}{\partial \ln V / \partial \ln M} = \frac{\beta_k + \sum_{j=1}^{K} \gamma_{kj} \ln(p_j/M)}{\beta_M + \sum_{j=1}^{K} \gamma_{Mj} \ln(p_j/M)}, k = 1, \ldots, K,$$

where $\beta_M = \Sigma_k \beta_k$ and $\gamma_{Mj} = \Sigma_k \gamma_{kj}$. No transformation of the budget share equation produces a linear model. This is an intrinsically nonlinear regression model. (It is also one among a system of equations, an aspect we will ignore for the present.) Although the share equation is stated in terms of observable variables, it remains unusable as an empirical model because of an **identification problem**. If every parameter in the budget share is multiplied by the same constant, then the constant appearing in both numerator and denominator cancels out, and the same value of the function in the equation remains. The indeterminacy is resolved by imposing the normalization $\beta_M = 1$. Note that this sort of identification problem does not arise in the linear model.

7.2.2 THE NONLINEAR LEAST SQUARES ESTIMATOR

The nonlinear least squares estimator is defined as the minimizer of the sum of squares,

$$S(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{2}\sum_{i=1}^{n} [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]^2. \tag{7-7}$$

The first-order conditions for the minimization are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})] \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{7-8}$$

In the linear model, the vector of partial derivatives will equal the regressors, $\mathbf{x}_i$. In what follows, we will identify the derivatives of the conditional mean function with respect to the parameters as the "pseudoregressors," $\mathbf{x}_i^0(\boldsymbol{\beta}) = \mathbf{x}_i^0$. We find that the nonlinear least squares estimator is the solution to

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i^0 \varepsilon_i = \mathbf{0}. \tag{7-9}$$

This is the nonlinear regression counterpart to the least squares normal equations in (3-12). Computation requires an iterative solution. (See Example 7.3.) The method is presented in Section 7.2.6.

Assumptions NR1 and NR3 imply that $E[\varepsilon_i | h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$. In the linear model, it follows, *because of the linearity of the conditional mean*, that ε_i and $\mathbf{x}_i$, are uncorrelated. However, *uncorrelatedness* of ε_i with a particular *nonlinear* function of $\mathbf{x}_i$ (the regression function) does not necessarily imply uncorrelatedness with $\mathbf{x}_i$, itself, nor, for that matter, with other nonlinear functions of $\mathbf{x}_i$. On the other hand, the results we will obtain for the behavior of the estimator in this model are couched not in terms of $\mathbf{x}_i$ but in terms of certain functions of $\mathbf{x}_i$ (the derivatives of the regression function), so, in point of fact, $E[\boldsymbol{\varepsilon} | \mathbf{X}] = \mathbf{0}$ is not even the assumption we need.

The foregoing is not a theoretical fine point. Dynamic models, which are very common in the contemporary literature, would greatly complicate this analysis. If it can be assumed that ε_i is strictly uncorrelated with *any prior information* in the model,

including previous disturbances, then a treatment analogous to that for the linear model would apply. But the convergence results needed to obtain the asymptotic properties of the estimator still have to be strengthened. The dynamic nonlinear regression model is beyond the reach of our treatment here. Strict independence of ε_i and $\mathbf{x}_i$ would be sufficient for uncorrelatedness of ε_i and every function of $\mathbf{x}_i$, but, again, in a dynamic model, this assumption might be questionable. Some commentary on this aspect of the nonlinear regression model may be found in Davidson and MacKinnon (1993, 2004).

If the disturbances in the nonlinear model are normally distributed, then the log of the normal density for the ith observation will be

$$\ln f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = -(1/2)\{\ln 2\pi + \ln \sigma^2 + [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]^2/\sigma^2\}. \tag{7-10}$$

For this special case, we have from item D.2 in Theorem 14.2 (on maximum likelihood estimation), that the derivatives of the log density with respect to the parameters have mean zero. That is,

$$E\left[\frac{\partial \ln f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}}\right] = E\left[\frac{1}{\sigma^2}\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\varepsilon_i\right] = \mathbf{0}, \tag{7-11}$$

so, in the normal case, the derivatives and the disturbances are uncorrelated. Whether this can be assumed to hold in other cases is going to be model specific, but under reasonable conditions, we would assume so.[2]

In the context of the linear model, the **orthogonality condition** $E[\mathbf{x}_i\varepsilon_i] = 0$ produces least squares as a **GMM estimator** for the model. (See Chapter 13.) The orthogonality condition is that the regressors and the disturbance in the model are uncorrelated. In this setting, the same condition applies to the first derivatives of the conditional mean function. The result in (7-11) produces a moment condition which will define the nonlinear least squares estimator as a GMM estimator.

Example 7.3 First-Order Conditions for a Nonlinear Model

The first-order conditions for estimating the parameters of the nonlinear regression model,

$$y_i = \beta_1 + \beta_2 e^{\beta_3 x_i} + \varepsilon_i,$$

by nonlinear least squares [see (7-13)] are

$$\frac{\partial S(\mathbf{b})}{\partial b_1} = -\sum_{i=1}^{n} [y_i - b_1 - b_2 e^{b_3 x_i}] = 0,$$

$$\frac{\partial S(\mathbf{b})}{\partial b_2} = -\sum_{i=1}^{n} [y_i - b_1 - b_2 e^{b_3 x_i}] e^{b_3 x_i} = 0,$$

$$\frac{\partial S(\mathbf{b})}{\partial b_3} = -\sum_{i=1}^{n} [y_i - b_1 - b_2 e^{b_3 x_i}] b_2 x_i e^{b_3 x_i} = 0.$$

These equations do not have an explicit solution.

Conceding the potential for ambiguity, we define a nonlinear regression model at this point as follows:

[2]See Ruud (2000, p. 540).

DEFINITION 7.1 Nonlinear Regression Model
*A **nonlinear regression model** is one for which the first-order conditions for least squares estimation of the parameters are nonlinear functions of the parameters.*

Thus, nonlinearity is defined in terms of the techniques needed to estimate the parameters, not the shape of the regression function. Later we shall broaden our definition to include other techniques besides least squares.

7.2.3 LARGE-SAMPLE PROPERTIES OF THE NONLINEAR LEAST SQUARES ESTIMATOR

Numerous analytical results have been obtained for the nonlinear least squares estimator, such as consistency and asymptotic normality. We cannot be sure that nonlinear least squares is the most efficient estimator, except in the case of normally distributed disturbances. (This conclusion is the same one we drew for the linear model.) But in the semiparametric setting of this chapter, we can ask whether this estimator is optimal in some sense given the information that we do have; the answer turns out to be yes. Some examples that follow will illustrate these points.

It is necessary to make some assumptions about the regressors. The precise requirements are discussed in some detail in Judge et al. (1985), Amemiya (1985), and Davidson and MacKinnon (2004). In the linear regression model, to obtain our asymptotic results, we assume that the sample moment matrix $(1/n)\mathbf{X}'\mathbf{X}$ converges to a positive definite matrix $\mathbf{Q}$. By analogy, we impose the same condition on the derivatives of the regression function, which are called the **pseudoregressors** in the linearized model [defined in (7-29)] when they are computed at the true parameter values. Therefore, for the nonlinear regression model, the analog to (4-19) is

$$\operatorname{plim} \frac{1}{n}\mathbf{X}^{0\prime}\mathbf{X}^{0} = \operatorname{plim} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}\right)\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0'}\right) = \mathbf{Q}^{0}, \tag{7-12}$$

where $\mathbf{Q}^0$ is a positive definite matrix. To establish consistency of $\mathbf{b}$ in the linear model, we required $\operatorname{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. We will use the counterpart to this for the pseudoregressors,

$$\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^0\,\varepsilon_i = \mathbf{0}.$$

This is the orthogonality condition noted earlier in (4-21). In particular, note that orthogonality of the disturbances and the data is not the same condition. Finally, asymptotic normality can be established under general conditions if

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^0\varepsilon_i = \sqrt{n}\,\bar{\mathbf{z}}^0 \xrightarrow{\;d\;} N[\mathbf{0}, \sigma^2\mathbf{Q}^0].$$

With these in hand, the asymptotic properties of the nonlinear least squares estimator are essentially those we have already seen for the linear model, except that in this case we place the derivatives of the linearized function evaluated at $\boldsymbol{\beta}$, $\mathbf{X}^0$, in the role of the regressors.[3]

[3]See Amemiya (1985).

The nonlinear least squares criterion function is

$$S(\mathbf{b}) = \frac{1}{2}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]^2 = \frac{1}{2}\sum_{i=1}^{n} e_i^2, \tag{7-13}$$

where we have inserted what will be the solution value, $\mathbf{b}$. The values of the parameters that minimize (one half of) the sum of squared deviations are the nonlinear least squares estimators. The first-order conditions for a minimum are

$$\mathbf{g}(\mathbf{b}) = -\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]\frac{\partial h(\mathbf{x}_i, \mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}. \tag{7-14}$$

In the linear model of Chapter 3, this produces a set of linear normal equations, (3-12). In this more general case, (7-14) is a set of nonlinear equations that do not have an explicit solution. Note that σ^2 is not relevant to the solution. At the solution,

$$\mathbf{g}(\mathbf{b}) = -\mathbf{X}^{0\prime}\mathbf{e} = \mathbf{0},$$

which is the same as (3-12) for the linear model.

Given our assumptions, we have the following general results:

THEOREM 7.1 Consistency of the Nonlinear Least Squares Estimator
If the following assumptions hold:

a. *The parameter space containing $\boldsymbol{\beta}$ is compact (has no gaps or nonconcave regions).*

b. *For any vector $\boldsymbol{\beta}^0$ in that parameter space, plim (1/n)$S(\boldsymbol{\beta}^0) = q(\boldsymbol{\beta}^0)$, a continuous and differentiable function.*

c. *If $q(\boldsymbol{\beta}^0)$ has a unique minimum at the true parameter vector, $\boldsymbol{\beta}$, then, the nonlinear least squares estimator defined by (7-13) and (7-14) is consistent. We will sketch the proof, then consider why the theorem and the proof differ as they do from the apparently simpler counterpart for the linear model. The proof, notwithstanding the underlying subtleties of the assumptions, is straightforward. The estimator, say, $\mathbf{b}^0$, minimizes (1/n)$S(\boldsymbol{\beta}^0)$. If (1/n)$S(\boldsymbol{\beta}^0)$ is minimized for every n, then it is minimized by $\mathbf{b}^0$ as n increases without bound. We also assumed that the minimizer of $q(\boldsymbol{\beta}^0)$ is uniquely $\boldsymbol{\beta}$. If the minimum value of plim (1/n)$S(\boldsymbol{\beta}^0)$ equals the probability limit of the minimized value of the sum of squares, the theorem is proved. This equality is produced by the continuity in assumption b.*

In the linear model, consistency of the least squares estimator could be established based on plim$(1/n)\mathbf{X}'\mathbf{X} = \mathbf{Q}$ and plim$(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. To follow that approach here, we would use the linearized model and take essentially the same result. The loose end in that argument would be that the linearized model is not the true model and there remains an approximation. For this line of reasoning to be valid, it must also be either

assumed or shown that $\text{plim}(1/n)\mathbf{X}^{0\prime}\boldsymbol{\delta} = \mathbf{0}$ where $\delta_i = h(\mathbf{x}_i, \boldsymbol{\beta})$ minus the Taylor series approximation.[4]

Note that no mention has been made of unbiasedness. The linear least squares estimator in the linear regression model is essentially alone in the estimators considered in this book. It is generally not possible to establish unbiasedness for any other estimator. As we saw earlier, unbiasedness is of fairly limited virtue in any event—we found, for example, that the property would not differentiate an estimator based on a sample of 10 observations from one based on 10,000. Outside the linear case, consistency is the primary requirement of an estimator. Once this is established, we consider questions of efficiency and, in most cases, whether we can rely on asymptotic normality as a basis for statistical inference.

THEOREM 7.2 Asymptotic Normality of the Nonlinear Least Squares Estimator

If the pseudoregressors defined in (7-12) are "well behaved," then

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}(\mathbf{Q}^0)^{-1}\right],$$

where

$$\mathbf{Q}^0 = \text{plim}\frac{1}{n}\mathbf{X}^{0\prime}\mathbf{X}^0.$$

The sample estimator of the asymptotic covariance matrix is

$$Est.Asy.Var[\mathbf{b}] = \hat{\sigma}^2(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}. \qquad \textbf{(7-15)}$$

Asymptotic efficiency of the nonlinear least squares estimator is difficult to establish without a distributional assumption. There is an indirect approach that is one possibility. The assumption of the orthogonality of the pseudoregressors and the true disturbances implies that the nonlinear least squares estimator is a GMM estimator in this context. With the assumptions of homoscedasticity and nonautocorrelation, the optimal weighting matrix is the one that we used, which is to say that in the class of GMM estimators for this model, nonlinear least squares uses the optimal weighting matrix. As such, it is asymptotically efficient in the class of GMM estimators.

The requirement that the matrix in (7-12) converges to a positive definite matrix implies that the columns of the regressor matrix $\mathbf{X}^0$ must be linearly independent. This **identification condition** is analogous to the requirement that the independent variables in the linear model be linearly independent. Nonlinear regression models usually involve several independent variables, and at first blush, it might seem sufficient to examine the data directly if one is concerned with multicollinearity. However, this situation is not the case. Example 7.4 gives an application.

[4]An argument to this effect appears in Mittelhammer et al. (2000, pp. 190–191).

A consistent estimator of σ^2 is based on the residuals,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]^2. \tag{7-16}$$

A degrees of freedom correction, $1/(n - K)$, where K is the number of elements in $\boldsymbol{\beta}$, is not strictly necessary here, because all results are asymptotic in any event. Davidson and MacKinnon (2004) argue that, on average, (7-16) will underestimate σ^2, and one should use the degrees of freedom correction. Most software in current use for this model does, but analysts will want to verify this is the case for the program they are using. With this in mind, the estimator of the asymptotic covariance matrix for the nonlinear least squares estimator is given in (7-15).

Once the nonlinear least squares estimates are in hand, inference and hypothesis tests can proceed in the same fashion as prescribed in Chapter 5. A minor problem can arise in evaluating the fit of the regression in that the familiar measure,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{7-17}$$

is no longer guaranteed to be in the range of 0 to 1. It does, however, provide a useful descriptive measure. An intuitively appealing measure of the fit of the model to the data will be the squared correlation between the fitted and actual values, $h(\mathbf{x}_i,\mathbf{b})$ and y_i. This will differ from R^2, partly because the mean prediction will not equal the mean of the observed values.

7.2.4 ROBUST COVARIANCE MATRIX ESTIMATION

Theorem 7.2 relies on assumption NR4, homoscedasticity and nonautocorrelation. We considered two generalizations in the linear case, heteroscedasticity and autocorrelation due to clustering in the sample. The counterparts for the nonlinear case would be based on the linearized model,

$$\begin{aligned} y_i &= x_i^{0\prime}\boldsymbol{\beta} + [h(\mathbf{x}_i, \boldsymbol{\beta}) - \mathbf{x}_i^{0\prime}\boldsymbol{\beta}] + \varepsilon_i \\ &= \mathbf{x}_i^{0\prime}\boldsymbol{\beta} + u_i. \end{aligned}$$

The counterpart to (4-37) that accommodates unspecified heteroscedasticity would then be

$$Est.Asy.Var[\mathbf{b}] = (\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\left[\sum_{i=1}^{n}\mathbf{x}_i^0\mathbf{x}_i^{0\prime}(y_i - h(\mathbf{x}_i, \mathbf{b}))^2\right](\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}.$$

Likewise, to allow for clustering, the computation would be analogous to (4-41) and (4-42),

$$Est.Asy.Var[\mathbf{b}] = \frac{C}{C-1}(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\left[\sum_{c=1}^{C}\left\{\sum_{i=1}^{N_c}\mathbf{x}_i^0 e_i\right\}\left\{\sum_{i=1}^{N_c}\mathbf{x}_i^0 e_i\right\}'\right](\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}.$$

Note that the residuals are computed as $e_i = y_i - h(\mathbf{x}_i, \mathbf{b})$ using the conditional mean function, not the linearized regression.

7.2.5 HYPOTHESIS TESTING AND PARAMETRIC RESTRICTIONS

In most cases, the sorts of hypotheses one would test in this context will involve fairly simple linear restrictions. The tests can be carried out using the familiar formulas discussed in Chapter 5 and the asymptotic covariance matrix presented earlier. For more involved hypotheses and for nonlinear restrictions, the procedures are a bit less clear-cut. Two principal testing procedures were discussed in Section 5.4: the Wald test, which relies on the consistency and asymptotic normality of the estimator, and the F test, which is appropriate in finite (all) samples, that relies on normally distributed disturbances. In the nonlinear case, we rely on large-sample results, so the Wald statistic will be the primary inference tool. An analog to the F statistic based on the fit of the regression will also be developed later. Finally, **Lagrange multiplier tests** for the general case can be constructed.

The hypothesis to be tested is

$$H_0: \mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}, \tag{7-18}$$

where $\mathbf{c}(\boldsymbol{\beta})$ is a column vector of J continuous functions of the elements of $\boldsymbol{\beta}$. These restrictions may be linear or nonlinear. It is necessary, however, that they be **overidentifying restrictions**. In formal terms, if the original parameter vector has K free elements, then the hypothesis $\mathbf{c}(\boldsymbol{\beta}) - \mathbf{q}$ must impose at least one functional relationship on the parameters. If there is more than one restriction, then they must be functionally independent. These two conditions imply that the $J \times K$ **Jacobian**,

$$\mathbf{R}(\boldsymbol{\beta}) = \partial \mathbf{c}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}', \tag{7-19}$$

must have full row rank and that J, the number of restrictions, must be strictly less than K. This situation is analogous to the linear model, in which $\mathbf{R}(\boldsymbol{\beta})$ would be the matrix of coefficients in the restrictions. (See, as well, Section 5.5, where the methods examined here are applied to the linear model.)

Let $\mathbf{b}$ be the unrestricted, nonlinear least squares estimator, and let $\mathbf{b}_*$ be the estimator obtained when the constraints of the hypothesis are imposed.[5] Which test statistic one uses depends on how difficult the computations are. Unlike the linear model, the various testing procedures vary in complexity. For instance, in our example, the Lagrange multiplier statistic is by far the simplest to compute. Of the methods we will consider, only this test does not require us to compute a nonlinear regression.

The nonlinear analog to the familiar F statistic based on the fit of the regression (i.e., the sum of squared residuals) would be

$$F[J, n - K] = \frac{[S(\mathbf{b}_*) - S(\mathbf{b})]/J}{S(\mathbf{b})/(n - K)}. \tag{7-20}$$

This equation has the appearance of our earlier F ratio in (5-29). In the nonlinear setting, however, neither the numerator nor the denominator has exactly the necessary chi-squared distribution, so the F distribution is only approximate. Note that this F statistic requires that both the restricted and unrestricted models be estimated.

[5]This computational problem may be extremely difficult in its own right, especially if the constraints are nonlinear. We assume that the estimates have been obtained by whatever means are necessary.

The Wald test is based on the distance between $\mathbf{c}(\mathbf{b})$ and $\mathbf{q}$. If the unrestricted estimates fail to satisfy the restrictions, then doubt is cast on the validity of the restrictions. The statistic is

$$\begin{aligned} W &= [\mathbf{c}(\mathbf{b}) - \mathbf{q}]'\{Est.Asy.Var[\mathbf{c}(\mathbf{b}) - \mathbf{q}]\}^{-1}[\mathbf{c}(\mathbf{b}) - \mathbf{q}] \\ &= [\mathbf{c}(\mathbf{b}) - \mathbf{q}]'\{\mathbf{R}(\mathbf{b})\hat{\mathbf{V}}\mathbf{R}'(\mathbf{b})\}^{-1}[\mathbf{c}(\mathbf{b}) - \mathbf{q}], \end{aligned} \tag{7-21}$$

where

$$\hat{\mathbf{V}} = Est.Asy.Var[\mathbf{b}],$$

and $\mathbf{R}(\mathbf{b})$ is evaluated at $\mathbf{b}$, the estimate of $\boldsymbol{\beta}$. Under the null hypothesis, this statistic has a limiting chi-squared distribution with J degrees of freedom. If the restrictions are correct, the Wald statistic and J times the F statistic are asymptotically equivalent. The Wald statistic can be based on the estimated covariance matrix obtained earlier using the unrestricted estimates, which may provide a large savings in computing effort if the restrictions are nonlinear. It should be noted that the small-sample behavior of W can be erratic, and the more conservative F statistic may be preferable if the sample is not large.

The caveat about Wald statistics that applied in the linear case applies here as well. Because it is a pure significance test that does not involve the alternative hypothesis, the Wald statistic is not invariant to how the hypothesis is framed. In cases in which there is more than one equivalent way to specify $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$, W can give different answers depending on which is chosen.

The Lagrange multiplier test is based on the decrease in the sum of squared residuals that would result if the restrictions in the restricted model were released. For the nonlinear regression model, the test has a particularly appealing form.[6] Let $\mathbf{e}_*$ be the vector of residuals $y_i - h(\mathbf{x}_i, \mathbf{b}_*)$ computed using the restricted estimates. Recall that we defined $\mathbf{X}^0$ as an $n \times K$ matrix of derivatives computed at a particular parameter vector in (7-29). Let $\mathbf{X}_*^0$ be this matrix *computed at the restricted estimates*. Then the Lagrange multiplier statistic for the nonlinear regression model is

$$LM = \frac{\mathbf{e}_*'\mathbf{X}_*^0[\mathbf{X}_*^{0'}\mathbf{X}_*^0]^{-1}\mathbf{X}_*^{0'}\mathbf{e}_*}{\mathbf{e}_*'\mathbf{e}_*/n}. \tag{7-22}$$

Under H_0, this statistic has a limiting chi-squared distribution with J degrees of freedom. What is especially appealing about this approach is that it requires only the restricted estimates. This method may provide some savings in computing effort if, as in our example, the restrictions result in a linear model. Note, also, that the Lagrange multiplier statistic is n times the uncentered R^2 in the regression of $\mathbf{e}_*$ on $\mathbf{X}_*^0$. Many Lagrange multiplier statistics are computed in this fashion.

7.2.6 APPLICATIONS

This section will present two applications of estimation and inference for nonlinear regression models. Example 7.4 illustrates a nonlinear consumption function that extends Examples 1.2 and 2.1. The model provides a simple demonstration of estimation and hypothesis testing for a nonlinear model. Example 7.5 analyzes the Box–Cox transformation. This specification is used to provide a more general functional form

[6] This test is derived in Judge et al. (1985). Discussion appears in Mittelhammer et al. (2000).

than the linear regression—it has the linear and loglinear models as special cases. Finally, Example 7.6 in the next section is a lengthy examination of an exponential regression model. In this application, we will explore some of the implications of nonlinear modeling, specifically "interaction effects." We examined interaction effects in Section 6.5.2 in a model of the form

$$y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 xz + \varepsilon.$$

In this case, the interaction effect is $\partial^2 E[y|x, z]/\partial x \partial z = \beta_4$. There is no interaction effect if β_4 equals zero. Example 7.6 considers the (perhaps unintended) implication of the nonlinear model that when $E[y|x, z] = h(x, z, \boldsymbol{\beta})$, there is an interaction effect even if the model is

$$h(x, z, \boldsymbol{\beta}) = h(\beta_1 + \beta_2 x + \beta_3 z).$$

Example 7.4 Analysis of a Nonlinear Consumption Function

The linear model analyzed at the beginning of Chapter 2 is a restricted version of the more general function

$$C = \alpha + \beta Y^{\gamma} + \varepsilon,$$

in which γ equals 1. With this restriction, the model is linear. If γ is free to vary, however, then this version becomes a nonlinear regression. Quarterly data on consumption, real disposable income, and several other variables for the U.S. economy for 1950 to 2000 are listed in Appendix Table F5.2. The restricted linear and unrestricted nonlinear least squares regression results are shown in Table 7.1. The procedures outlined earlier are used to obtain the asymptotic standard errors and an estimate of σ^2. (To make this comparable to s^2 in the linear model, the value includes the degrees of freedom correction.)

In the preceding example, there is no question of collinearity in the data matrix $\mathbf{X} = [\mathbf{i}, \mathbf{y}]$; the variation in Y is obvious on inspection. But, at the final parameter estimates, the R^2 in the regression is 0.998834 and the correlation between the two pseudoregressors $x_2^0 = Y^{\gamma}$ and $x_3^0 = \beta Y^{\gamma} \ln Y$ is 0.999752. The condition number for the normalized matrix of sums of squares and cross products is 208.306. (The condition number is computed by computing the square root of the ratio of the largest to smallest characteristic root of $\mathbf{D}^{-1}\mathbf{X}_0'\mathbf{X}_0\mathbf{D}^{-1}$ where $x_1^0 = 1$ and $\mathbf{D}$ is the diagonal matrix containing the square roots of $\mathbf{x}_k^{0\prime}\mathbf{x}_k^0$ on the diagonal.) Recall that 20 was the benchmark for a problematic data set. By the standards discussed in

TABLE 7.1 Estimated Consumption Functions

	Linear Model		*Nonlinear Model*	
Parameter	***Estimate***	***Standard Error***	***Estimate***	***Standard Error***
α	−80.3547	14.3059	458.7990	22.5014
β	0.9217	0.003872	0.10085	0.01091
γ	1.0000	–	1.24483	0.01205
$\mathbf{e}'\mathbf{e}$	1,536,321.881		504,403.1725	
σ	87.20983		50.0946	
R^2	0.996448		0.998834	
Est.Var[b]	–		0.000119037	
Est.Var[c]	–		0.00014532	
Est.Cov[b,c]	–		−0.000131491	

Sections 4.7.1 and A.6.6, the collinearity problem in this data set is severe. In fact, it appears not to be a problem at all.

For hypothesis testing and confidence intervals, the familiar procedures can be used, with the proviso that all results are only asymptotic. As such, for testing a restriction, the chi-squared statistic rather than the F ratio is likely to be more appropriate. For example, for testing the hypothesis that γ is different from 1, an asymptotic t test, based on the standard normal distribution, is carried out, using

$$z = \frac{1.24483 - 1}{0.01205} = 20.3178.$$

This result is larger than the critical value of 1.96 for the 5% significance level, and we thus reject the linear model in favor of the nonlinear regression. The three procedures for testing hypotheses produce the same conclusion.

$$F[1,204 - 3] = \frac{(1{,}536{,}321.881 - 504{,}403.17)/1}{504{,}403.17/(204 - 3)} = 411.29,$$

$$W = \frac{(1.24483 - 1)^2}{0.01205^2} = 412.805,$$

$$LM = \frac{996{,}103.9}{(1{,}536{,}321.881/204)} = 132.267.$$

For the Lagrange multiplier statistic, the elements in $\mathbf{x}_i{}^*$ are $\mathbf{x}_i{}^* = [1, Y^\gamma, \beta Y^\gamma \ln Y]$. To compute this at the restricted estimates, we use the ordinary least squares estimates for α and β and 1 for γ so that $\mathbf{x}_i{}^* = [1, Y, \beta Y \ln Y]$. The residuals are the least squares residuals computed from the linear regression.

Example 7.5 The Box–Cox Transformation

The **Box–Cox transformation** is used as a device for generalizing the linear model.[7] The transformation is

$$x^{(\lambda)} = (x^\lambda - 1)/\lambda.$$

Special cases of interest are $\lambda = 1$, which produces a linear transformation, $x^{(1)} = x - 1$, and $\lambda = 0$. When λ equals zero, the transformation is, by L'Hôpital's rule,

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{d(x^\lambda - 1)/d\lambda}{1} = \lim_{\lambda \to 0} x^\lambda \times \ln x = \ln x.$$

The regression analysis can be done *conditionally* on λ. For a given value of λ, the model,

$$y = \alpha + \sum_{k=2}^{K} \beta_k x_k^{(\lambda)} + \varepsilon, \qquad \textbf{(7-23)}$$

is a linear regression that can be estimated by least squares. However, if λ in (7-23) is taken to be an unknown parameter, then the regression becomes nonlinear in the parameters.

In principle, each regressor could be transformed by a different value of λ, but, in most applications, this level of generality becomes excessively cumbersome, and λ is assumed to be the same for all the variables in the model.[8] To be defined for all values of λ, x must be strictly positive. In most applications, some of the regressors—for example, a dummy

[7] Box and Cox (1964); Zarembka (1974).

[8] See, for example, Seaks and Layson (1983).

variable—will not be transformed. For such a variable, say v_k, $v_k^{(\lambda)} = v_k$, and the relevant derivatives in (7-24) will be zero. It is also possible to transform y, say, by $y^{(\theta)}$. Transformation of the dependent variable, however, amounts to a specification of the whole model, not just the functional form of the conditional mean. For example, $\theta = 1$ implies a linear equation while $\theta = 0$ implies a logarithmic equation.

Nonlinear least squares is straightforward. In most instances, we can expect to find the least squares value of λ between -2 and 2. Typically, then, λ is estimated by scanning this range for the value that minimizes the sum of squares. Once the optimal value of λ is located, the least squares estimates, the mean squared residual, and this value of λ constitute the nonlinear least squares estimates of the parameters. The optimal value of $\hat{\lambda}$ is an estimate of an unknown parameter. The least squares standard errors will always underestimate the correct asymptotic standard errors if $\hat{\lambda}$ is treated as if it were a known constant.[9] To get the appropriate values, we need the pseudoregressors,

$$\begin{aligned}
\frac{\partial h(.)}{\partial \alpha} &= 1, \\
\frac{\partial h(.)}{\partial \beta_k} &= x_k^{(\lambda)}, \\
\frac{\partial h(.)}{\partial \lambda} &= \sum_{k=1}^{K} \beta_k \frac{\partial x_k^{(\lambda)}}{\partial \lambda} = \sum_{k=1}^{K} \beta_k \left[\frac{1}{\lambda} \left(x_k^{\lambda} \ln x_k - x_k^{(\lambda)} \right) \right].
\end{aligned} \tag{7-24}$$

We can now use (7-15) and (7-16) to estimate the asymptotic covariance matrix of the parameter estimates. Note that $\ln x_k$ appears in $\partial h(.)/\partial \lambda$. If $x_k = 0$, then this matrix cannot be computed.

The coefficients in a nonlinear model are not equal to the slopes (or the elasticities) with respect to the variables. For the Box–Cox model $\ln Y = \alpha + \beta X^{(\lambda)} + \epsilon$,

$$\frac{\partial E[\ln y \,|\, \mathbf{x}]}{\partial \ln x} = x \frac{\partial E[\ln y \,|\, \mathbf{x}]}{\partial x} = \beta x^{\lambda} = \eta.$$

A standard error for this estimator can be obtained using the **delta method**. The derivatives are $\partial \eta / \partial \beta = x^{\lambda} = \eta / \beta$ and $\partial \eta / \partial \lambda = \eta \ln x$. Collecting terms, we obtain

$$Asy.Var[\hat{\eta}] = (\eta/\beta)^2 \left\{ Asy.Var[\hat{\beta}] + (\beta \ln x)^2 \, Asy.Var[\hat{\lambda}] + (2\beta \ln x) \, Asy.Cov[\hat{\beta}, \hat{\lambda}] \right\}.$$

7.2.7 LOGLINEAR MODELS

Loglinear models play a prominent role in statistics. Many derive from a density function of the form $f(y \,|\, \mathbf{x}) = p[y \,|\, \alpha^0 + \mathbf{x}'\boldsymbol{\beta}, \theta]$, where α^0 is a constant term and θ is an additional parameter such that

$$E[y \,|\, \mathbf{x}] = g(\theta) \exp(\alpha^0 + \mathbf{x}'\boldsymbol{\beta}).$$

(Hence the name *loglinear models*). Examples include the Weibull, gamma, lognormal, and exponential models for continuous variables and the Poisson and negative binomial models for counts. We can write $E[y \,|\, \mathbf{x}]$ as $\exp[\ln g(\theta) + \alpha^0 + \mathbf{x}'\boldsymbol{\beta}]$, and then absorb $\ln g(\theta)$ in the constant term in $\ln E[y \,|\, \mathbf{x}] = \alpha + \mathbf{x}'\boldsymbol{\beta}$. The lognormal distribution (see Section B.4.4) is often used to model incomes. For the lognormal random variable,

[9] See Fomby, Hill, and Johnson (1984, pp. 426–431).

$$p[y|\alpha^0 + \mathbf{x}'\boldsymbol{\beta}, \theta] = \frac{\exp[-\frac{1}{2}(\ln y - \alpha^0 - \mathbf{x}'\boldsymbol{\beta})^2/\theta^2]}{\theta y\sqrt{2\pi}}, y > 0,$$

$$E[y|\mathbf{x}] = \exp(\alpha^0 + \mathbf{x}'\boldsymbol{\beta} + \theta^2/2) = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}).$$

The exponential regression model is also consistent with a gamma distribution. The density of a gamma distributed random variable is

$$p[y|\alpha^0 + \mathbf{x}'\boldsymbol{\beta}, \theta] = \frac{\lambda^\theta \exp(-\lambda y)y^{\theta-1}}{\Gamma(\theta)}, y > 0, \theta > 0, \lambda = \exp(-\alpha^0 - x'\beta),$$

$$E[y|\mathbf{x}] = \theta/\lambda = \theta \exp(\alpha^0 + \mathbf{x}'\boldsymbol{\beta}) = \exp(\ln \theta + \alpha^0 + \mathbf{x}'\boldsymbol{\beta}) = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}).$$

The parameter θ determines the shape of the distribution. When $\theta > 2$, the gamma density has the shape of a chi-squared variable (which is a special case). Finally, the Weibull model has a similar form,

$$p[y|\alpha^0 + \mathbf{x}'\boldsymbol{\beta}, \theta] = \theta\lambda^\theta \exp[-(\lambda y)^\theta]y^{\theta-1}, y \geq 0, \theta > 0, \lambda = \exp(-\alpha^0 - \mathbf{x}'\boldsymbol{\beta}),$$

$$E[y|\mathbf{x}] = \Gamma(1 + 1/\theta) \exp(\alpha^0 + \mathbf{x}'\boldsymbol{\beta}) = \exp[\ln \Gamma(1 + 1/\theta) + \alpha^0 + \mathbf{x}'\boldsymbol{\beta}] = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}).$$

In all cases, the maximum likelihood estimator is the most efficient estimator of the parameters. (Maximum likelihood estimation of the parameters of this model is considered in Chapter 14.) However, nonlinear least squares estimation of the model

$$E[y|\mathbf{x}] = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}) + \varepsilon$$

has a virtue in that the nonlinear least squares estimator will be consistent even if the distributional assumption is incorrect—it is *robust* to this type of misspecification since it does not make explicit use of a distributional assumption. However, since the model is nonlinear, the coefficients do not give the magnitudes of the interesting effects in the equation. In particular, for this model,

$$\partial E[y|\mathbf{x}]/\partial x_k = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}) \times \partial(\alpha + \mathbf{x}'\boldsymbol{\beta})/\partial x_k$$
$$= \beta_k \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}).$$

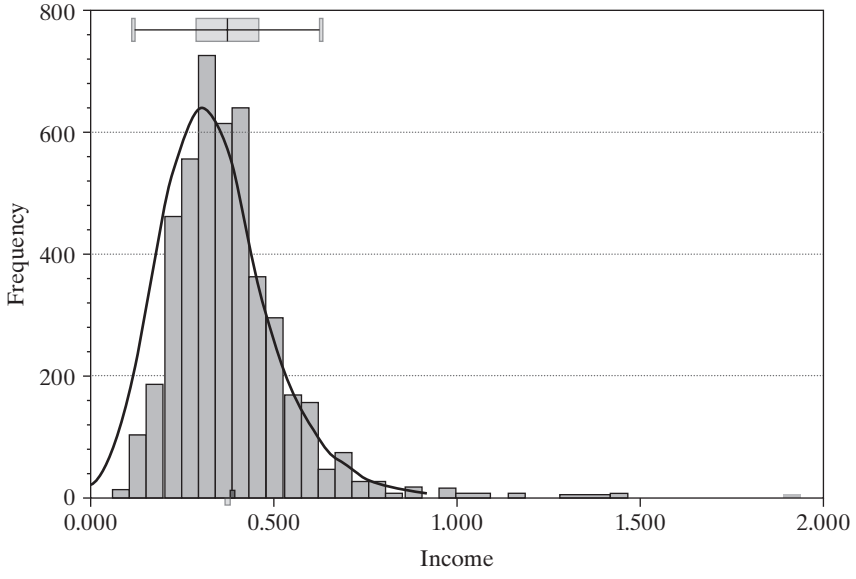
The implication is that the analyst must be careful in interpreting the estimation results, as interest usually focuses on partial effects, not coefficients.

Example 7.6 Interaction Effects in a Loglinear Model for Income

In *Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation*, Riphahn, Wambach, and Million (2003) were interested in counts of physician visits and hospital visits and in the impact that the presence of private insurance had on the utilization counts of interest, that is, whether the data contain evidence of moral hazard. The sample used is an unbalanced panel of 7,293 households, the German Socioeconomic Panel (GSOEP) data set.[10] Among the variables reported in the panel are household income, with numerous

[10]The data are published on the *Journal of Applied Econometrics* data archive Web site, at http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/. The variables in the data file are listed in Appendix Table F7.1. The number of observations in each year varies from one to seven with a total number of 27,326 observations. We will use these data in several examples here and later in the book.

FIGURE 7.1 Histogram and Kernel Density Estimate for Income.



other sociodemographic variables such as age, gender, and education. For this example, we will model the distribution of income using the 1988 wave of the data set, a cross section with 4,483 observations. Two of the individuals in this sample reported zero income, which is incompatible with the underlying models suggested in the development below. Deleting these two observations leaves a sample of 4,481 observations. Figure 7.1 displays a histogram and a kernel density estimator for the household income variable for these observations. Table 7.2 provides descriptive statistics for the exogenous variables used in this application.

We will fit an exponential regression model to the income variable, with

$$Income = \exp(\beta_1 + \beta_2 Age + \beta_3 Age^2 + \beta_4 Education + \beta_5 Female + \beta_6 Female \times Education + \beta_7 Age \times Education) + \varepsilon.$$

As we have constructed the model, the derivative result, $\partial E[y|\mathbf{x}]/\partial x_k = \beta_k \exp(\alpha + \mathbf{x}'\boldsymbol{\beta})$, must be modified because the variables appear either in a quadratic term or as a product with some other variable. Moreover, for the dummy variable, *Female*, we would want to compute the partial effect using

$$\Delta E[y|\mathbf{x}]/\Delta Female = E[y|\mathbf{x}, Female = 1] - E[y|\mathbf{x}, Female = 0].$$

TABLE 7.2 Descriptive Statistics for Variables Used in Nonlinear Regression

Variable	*Mean*	*Std.Dev.*	*Minimum*	*Maximum*
Income	0.344896	0.164054	0.0050	2
Age	43.4452	11.2879	25	64
Educ	11.4167	2.36615	7	18
Female	0.484267	0.499808	0	1

Another consideration is how to compute the partial effects, as sample averages or at the means of the variables. For example, $\partial E[y|\mathbf{x}]/\partial Age = E[y|\mathbf{x}] \times (\beta_2 + 2\beta_3 Age + \beta_7 Educ)$. We will estimate the average partial effects by averaging these values over the sample observations. Table 7.3 presents the nonlinear least squares regression results. Superficially, the pattern of signs and significance might be expected—with the exception of the dummy variable for female.

The average value of *Age* in the sample is 43.4452 and the average value of *Education* is 11.4167. The partial effect of a year of education is estimated to be 0.015736 if it is computed by computing the partial effect for each individual and averaging the results. The partial effect is difficult to interpret without information about the scale of the income variable. Since the average income in the data is about 0.35, these partial effects suggest that an additional year of education is associated with a change in expected income of about 4.5% (i.e., 0.015736/0.35).

The rough calculation of partial effects with respect to *Age* does not reveal the model implications about the relationship between age and expected income. Note, for example, that the coefficient on *Age* is positive while the coefficient on Age^2 is negative. This implies (neglecting the interaction term at the end), that the *Age*—*Income* relationship implied by the model is parabolic. The partial effect is positive at some low values and negative at higher values. To explore this, we have computed the expected *Income* using the model separately for men and women, both with assumed college education (*Educ* = 16) and for the range of ages in the sample, 25 to 64. Figure 7.2 shows the result of this calculation. The upper curve is for men (*Female* = 0) and the lower one is for women. The parabolic shape is as expected; what the figure reveals is the relatively strong effect—*ceteris paribus*, incomes are predicted to rise by about 80% between ages 25 and 48. The figure reveals a second implication of the estimated model that would not be obvious from the regression results. The coefficient on the dummy variable for *Female* is positive, highly significant, and, in isolation, by far the largest effect in the model. This might lead the analyst to conclude that on average, expected incomes in these data are higher for women than men. But Figure 7.2 shows precisely the opposite. The difference is accounted for by the interaction term, *Female* × *Education*. The negative sign on the latter coefficient is suggestive. But the total effect would remain ambiguous without the sort of secondary analysis suggested by the figure.

TABLE 7.3 Estimated Regression Equations

	Nonlinear Least Squares			*Linear Least Squares*	
Variable	***Estimate***	***Std. Error***	***t Ratio***	***Estimate***	***Projection***
Constant	−2.58070	0.17455	14.78	−0.13050	0.10746
Age	0.06020	0.00615	9.79	0.01791	0.00066
Age^2	−0.00084	0.00006082	−13.83	−0.00027	
Education	−0.00616	0.01095	−0.56	−0.00281	0.01860
Female	0.17497	0.05986	2.92	0.07955	0.00075
Female × *Educ*	−0.01476	0.00493	−2.99	−0.00685	
Age × *Educ*	0.00134	0.00024	5.59	0.00055	
$e'e$		106.09825		106.24323	
s		0.15387		0.15410	
R^2		0.12005		0.11880	

FIGURE 7.2 Expected Incomes vs. Age for Men and Women with EDUC = 16.

Finally, in addition to the quadratic term in age, the model contains an interaction term, *Age* × *Education*. The coefficient is positive and highly significant. But, it is not obvious how this should be interpreted. In a linear model,

$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + \beta_4 Education + \beta_5 Female + \beta_6 Female \times Education + \beta_7 Age \times Education + \varepsilon,$$

we would find that $\beta_7 = \partial^2 E[Income|x]/\partial\, Age\, \partial Education$. That is, the "interaction effect" is the change in the partial effect of *Age* associated with a change in *Education* (or vice versa). Of course, if β_7 equals zero, that is, if there is no product term in the model, then there is no interaction effect—the second derivative equals zero. However, this simple interpretation usually does not apply in nonlinear models (i.e., in any nonlinear model). Consider our exponential regression, and suppose that in fact, β_7 is indeed zero. For convenience, let $\mu(x)$ equal the conditional mean function. Then, the partial effect with respect to *Age* is

$$\partial\mu(x)/\partial Age = \mu(x) \times (\beta_2 + 2\beta_3 Age),$$

and

$$\partial^2\mu(x)/\partial Age\partial Educ = \mu(x) \times (\beta_2 + 2\beta_3 Age)\,(\beta_4 + \beta_6 Female), \quad \textbf{(7-25)}$$

which is nonzero even if there is no **interaction term** in the model. The interaction effect in the model that includes the product term, $\beta_7 Age \times Education$, is

$$\partial^2 E[y|x]/\partial Age\partial Educ = \mu(x) \times [\beta_7 + (\beta_2 + 2\beta_3 Age + \beta_7 Educ)\,(\beta_4 + \beta_6 Female + \beta_7 Age)]. \quad \textbf{(7-26)}$$

At least some of what is being called the interaction effect in this model is attributable entirely to the fact the model is nonlinear. To isolate the "functional form effect" from the true "interaction effect," we might subtract (7-25) from (7-26) and then reassemble the components:

$$\partial^2\mu(x)/\partial Age\,\partial Educ = \mu(x)[(\beta_2 + 2\beta_3 Age)(\beta_4 + \beta_6 Female)] + \mu(x)\,\beta_7[1 + Age(\beta_2 + 2\beta_3 Age) + Educ(\beta_4 + \beta_6 Female) + Educ \times Age(\beta_7)]. \quad \textbf{(7-27)}$$

It is clear that the coefficient on the product term bears essentially no relationship to the quantity of interest (assuming it is the change in the partial effects that is of interest). On the other hand, the second term is nonzero if and only if β_7 is nonzero. One might, therefore, identify the second part with the "interaction effect" in the model. Whether a behavioral interpretation could be attached to this is questionable, however. Moreover, that would leave unexplained the functional form effect. The point of this exercise is to suggest that one should proceed with some caution in interpreting interaction effects in nonlinear models. This sort of analysis has a focal point in the literature in Ai and Norton (2004). A number of comments and extensions of the result are to be found, including Greene (2010b).

Section 4.4.5 considered the linear projection as a feature of the joint distribution of y and $\mathbf{x}$. It was noted that, assuming the conditional mean function in the joint distribution is $E[y|\mathbf{x}] = \mu(\mathbf{x})$, then the slopes of linear projection, $\gamma = [E\{\mathbf{x}\mathbf{x}'\}]^{-1}E[\mathbf{x}y]$, might resemble the slopes of $\mu(\mathbf{x})$, $\boldsymbol{\delta} = \partial\mu(\mathbf{x})/\partial\mathbf{x}$ at least for some $\mathbf{x}$. In a loglinear, single-index function model such as the one analyzed here, this would relate to the linear least squares regression of y on $\mathbf{x}$. Table 7.4 reports two sets of least squares regression coefficients. The ones on the right show the regression of *Income* on all of the first- and second-order terms that appear in the conditional mean. This would not be the projection of y on $\mathbf{x}$. At best it might be seen as an approximation to $\mu(\mathbf{x})$. The rightmost coefficients report the projection. Both results suggest superficially that nonlinear least squares and least squares are computing completely different relationships. To uncover the similarity (if there is one), it is useful to consider the partial effects rather than the coefficients. Table 7.4 reports the results of the computations. The average partial effects for the nonlinear regression are obtained by computing the derivatives for each observation and averaging the results. For the linear approximation, the derivatives are linear functions of the variables, so the average partial effects are simply computed at the means of the variables. Finally, the coefficients of the linear projection are immediate estimates of the partial effects. We find, for example, the partial effect of education in the nonlinear model is 0.01574. Although the linear least squares coefficients are very different, if the partial effect for education is computed for the linear approximation the result of 0.01789 is reasonably close, and results from the fact that in the center of the data, the exponential function is passably linear. The linear projection is less effective at reproducing the partial effects. The comparison for the other variables is mixed. The conclusion from Example 4.4 is unchanged. The substantive comparison here would be between the slopes of the nonlinear regression and the slopes of the linear projection. They resemble each other, but not as closely as one might hope.

TABLE 7.4 Estimated Partial Effects

Variable	*Nonlinear Regression*	*Linear Approximation*	*Linear Projection*
Age	0.00095	0.00091	0.00066
Educ	0.01574	0.01789	0.01860
Female	0.00084	0.00135	0.00075

Example 7.7 Generalized Linear Models for the Distribution of Healthcare Costs

Jones, Lomas, and Rice (2014, 2015) examined the distribution of healthcare costs in the UK. Two aspects of the analysis were different from our examinations to this point. First, while nearly all of the development we have considered so far involves regression, that is, the conditional mean (or median) of the distribution of the dependent variable, their interest was in other parts of the distribution, specifically conditional and unconditional tail probabilities for relatively outlying parts of the distribution. Second, the variable under study is nonnegative, highly asymmetric (skewness 13.03), and leptokurtic (kurtosis 363.13—the distribution has a thick right tail). Some values from the estimated survival function (Jones et al., 2015, Table 1) are $S(£500) = 0.8296$, $S(£1{,}000) = 0.5589$, $S(£5{,}000) = 0.1383$, and $S(£10{,}000) = 0.0409$. The skewness and kurtosis values would compare to 0.0 and 3.0, respectively, for the normal distribution. The survival function values for the normal distribution with this mean and standard deviation would be 0.6608, 0.6242, 0.3193, and 0.0732, respectively. The model is constructed with these features of the data in mind. Several methods of fitting the distribution were examined, including a set of nine parametric models. Several of these were special cases of the *generalized beta of the second kind*. The functional forms are *generalized linear models* constructed from a *family* of distributions, such as the normal or exponential, and a link function, $g(\mathbf{x}'\boldsymbol{\beta})$ such that $\text{link}(g(\mathbf{x}'\boldsymbol{\beta})) = \mathbf{x}'\boldsymbol{\beta}$. Thus, if the link function is "ln" (log link), then $g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$. Among the nine special cases examined are

- Gamma family, log link:

$$f(cost\,|\,\mathbf{x}) = \frac{[g(\mathbf{x}'\boldsymbol{\beta})]^{-P}}{\Gamma(P)}\exp[-cost/g(\mathbf{x}'\boldsymbol{\beta})]cost^{P-1},$$

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta});\ E[cost\,|\,\mathbf{x}] = Pg(\mathbf{x}'\boldsymbol{\beta}).$$

- Lognormal family, identity link:

$$f(cost\,|\,\mathbf{x}) = \frac{1}{\sigma cost\sqrt{2\pi}}\exp\left[-\frac{(\ln cost - g(\mathbf{x}'\boldsymbol{\beta}))^2}{2\sigma^2}\right],$$

$$g(\mathbf{x}'\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta};\ E[cost\,|\,\mathbf{x}] = \exp\left[g(\mathbf{x}'\boldsymbol{\beta}) + \tfrac{1}{2}\sigma^2\right].$$

- Finite mixture of two gammas, inverse square root link:

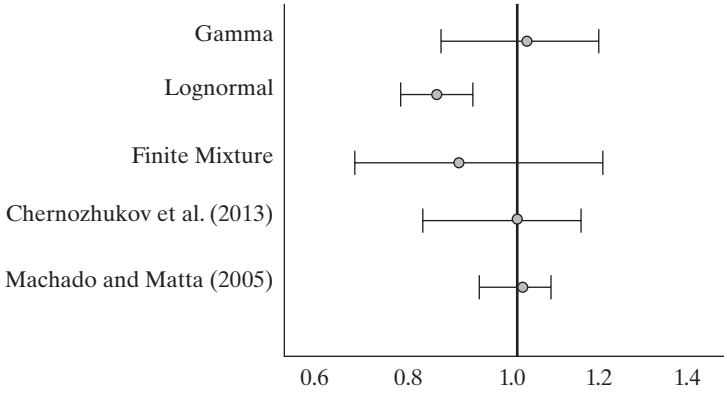
$$f(cost\,|\,\mathbf{x}) = \sum_{j=1}^{2}\alpha_j\frac{[g(\mathbf{x}'\boldsymbol{\beta}_j)]^{-P_j}}{\Gamma(P_j)}\exp[-cost/g(\mathbf{x}'\boldsymbol{\beta}_j)]cost^{P_j-1},\ 0 \le \alpha_j \le 1,\ \sum_{j=1}^{2}\alpha_j = 1,$$

$$g(\mathbf{x}'\boldsymbol{\beta}) = 1/(\mathbf{x}'\boldsymbol{\beta})^2;\ E[cost\,|\,\mathbf{x}] = \alpha_1 P_1 g(\mathbf{x}'\boldsymbol{\beta}_1) + \alpha_2 P_2 g(\mathbf{x}'\boldsymbol{\beta}_2).$$

(The models have been reparameterized here to simplify them and show their similarities.) In each case, there is a conditional mean function. However, the quantity of interest in the study is not the regression function; it is the survival function, $S(cost\,|\,\mathbf{x},k) = \text{Prob}(cost \ge k\,|\,\mathbf{x})$. The measure of a model's performance is its ability to estimate the sample survival rate for values of k; the one of particular interest is the largest, $k = 10{,}000$. The main interest is the marginal rate, $E_{\mathbf{x}}[S(cost\,|\,\mathbf{x},k)] = \int_{\mathbf{x}} S(cost\,|\,\mathbf{x},k)f(\mathbf{x})d\mathbf{x}$. This is estimated by estimating $\boldsymbol{\beta}$ and the ancillary parameters of the specific model, then estimating $S(cost\,|\,k)$ with $(1/n)\Sigma_{i=1}^{n}S(cost_i\,|\,\mathbf{x}_i,k;\hat{\boldsymbol{\beta}})$ The covariates include a set of morbidity characteristics and an interacted cubic function of age and sex. Several semiparametric and nonparametric methods are examined along with the parametric regression–based models. Figure 7.3 shows the bias and variability of the three parametric estimators and two of the proposed semiparametric methods.[11] Overall, none of the 14 methods examined emerges as best overall by a set of fitting criteria that includes bias and variability.

[11]Derived from the results in Figure 4 in Jones et al. (2015).

FIGURE 7.3 Performance of Several Estimators of $S(cost|k)$.



7.2.8 COMPUTING THE NONLINEAR LEAST SQUARES ESTIMATOR

Minimizing the sum of squared residuals for a nonlinear regression is a standard problem in nonlinear optimization that can be solved by a number of methods. (See Section E.3.) The method of Gauss–Newton is often used. This algorithm (and most of the sampling theory results for the asymptotic properties of the estimator) is based on a linear Taylor series approximation to the nonlinear regression function. The iterative estimator is computed by transforming the optimization to a series of linear least squares regressions.

The nonlinear regression model is $y = h(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon$. (To save some notation, we have dropped the observation subscript.) The procedure is based on a linear Taylor series approximation to $h(\mathbf{x}, \boldsymbol{\beta})$ at a particular value for the parameter vector, $\boldsymbol{\beta}^0$,

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h(\mathbf{x}, \boldsymbol{\beta}^0) + \sum_{k=1}^{K} \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0}(\beta_k - \beta_k^0). \tag{7-28}$$

This form of the equation is called the **linearized regression model**. By collecting terms, we obtain

$$h(\mathbf{x}, \beta) \approx \left[h(\mathbf{x}, \boldsymbol{\beta}^0) - \sum_{k=1}^{K} \beta_k^0 \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right) \right] + \sum_{k=1}^{K} \beta_k \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right). \tag{7-29}$$

Let x_k^0 equal the kth partial derivative,[12] $\partial h(\mathbf{x}, \boldsymbol{\beta}^0)/\partial \beta_k^0$. For a given value of $\boldsymbol{\beta}^0$, x_k^0 is a function only of the data, not of the unknown parameters. We now have

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[h^0 - \sum_{k=1}^{K} x_k^0 \beta_k^0 \right] + \sum_{k=1}^{K} x_k^0 \beta_k,$$

which may be written

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h^0 - \mathbf{x}^{0\prime}\boldsymbol{\beta}^0 + \mathbf{x}^{0\prime}\boldsymbol{\beta},$$

[12] You should verify that for the linear regression model, these derivatives are the independent variables.

which implies that

$$y \approx h^0 - \mathbf{x}^{0\prime}\boldsymbol{\beta}^0 + \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon.$$

By placing the known terms on the left-hand side of the equation, we obtain a linear equation,

$$y^0 = y - h^0 + \mathbf{x}^{0\prime}\boldsymbol{\beta}^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0. \tag{7-30}$$

Note that ε^0 contains both the true disturbance, ε, and the error in the first-order Taylor series approximation to the true regression, shown in (7-29). That is,

$$\varepsilon^0 = \varepsilon + \left[h(\mathbf{x}, \boldsymbol{\beta}) - \left(h^0 - \sum_{k=1}^{K} x_k^0 \beta_k^0 + \sum_{k=1}^{K} x_k^0 \beta_k \right) \right]. \tag{7-31}$$

Because all the errors are accounted for, (7-30) is an equality, not an approximation. With a value of $\boldsymbol{\beta}^0$ in hand, we could compute y^0 and $\mathbf{x}^0$ and then estimate the parameters of (7-30) by linear least squares. Whether this estimator is consistent or not remains to be seen.

Example 7.8 Linearized Regression

For the model in Example 7.3, the regressors in the linearized equation would be

$$x_1^0 = \frac{\partial h(.)}{\partial \beta_1^0} = 1,$$
$$x_2^0 = \frac{\partial h(.)}{\partial \beta_2^0} = e^{\beta_3^0 x},$$
$$x_3^0 = \frac{\partial h(.)}{\partial \beta_3^0} = \beta_2^0 x e^{\beta_3^0 x}.$$

With a set of values of the parameters $\boldsymbol{\beta}^0$,

$$y^0 = y - h(x, \beta_1^0, \beta_2^0, \beta_3^0) + \beta_1^0 x_1^0 + \beta_2^0 x_2^0 + \beta_3^0 x_3^0$$

can be linearly regressed on the three pseudoregressors to estimate β_1, β_2, and β_3.

The linearized regression model shown in (7-30) can be estimated by linear least squares. Once a parameter vector is obtained, it can play the role of a new $\boldsymbol{\beta}^0$, and the computation can be done again. The **iteration** can continue until the difference between successive parameter vectors is small enough to assume convergence. One of the main virtues of this method is that at the last iteration the estimate of $(\mathbf{Q}^0)^{-1}$ will, apart from the scale factor $\hat{\sigma}^2/n$, provide the correct estimate of the asymptotic covariance matrix for the parameter estimator.

This iterative solution to the minimization problem is

$$\begin{aligned}
\mathbf{b}_{t+1} &= \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \mathbf{x}_i^{0\prime} \right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \left(y_i - h_i^0 + \mathbf{x}_i^{0\prime} \mathbf{b}_t \right) \right] \\
&= \mathbf{b}_t + \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \mathbf{x}_i^{0\prime} \right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \left(y_i - h_i^0 \right) \right] \\
&= \mathbf{b}_t + (\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1} \mathbf{X}^{0\prime} \mathbf{e}^0 \\
&= \mathbf{b}_t + \boldsymbol{\Delta}_t,
\end{aligned} \tag{7-32}$$

where all terms on the right-hand side are evaluated at $\mathbf{b}_t$ and $\mathbf{e}^0$ is the vector of nonlinear least squares residuals. This algorithm has some intuitive appeal as well. For each iteration, we update the previous parameter estimates by regressing the nonlinear least squares residuals on the derivatives of the regression functions. The process will have converged (i.e., the update will be $\mathbf{0}$) when $\mathbf{X}^{0\prime}\mathbf{e}^0$ is close enough to $\mathbf{0}$. This derivative has a direct counterpart in the normal equations for the linear model, $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

As usual, when using a digital computer, we will not achieve exact convergence with $\mathbf{X}^{0\prime}\mathbf{e}^0$ exactly equal to zero. A useful, scale-free counterpart to the convergence criterion discussed in Section E.3.6 is $\delta = \mathbf{e}^{0\prime}\mathbf{X}^0(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\mathbf{X}^{0\prime}\mathbf{e}^0$. [See (7-22).] We note, finally, that iteration of the linearized regression, although a very effective algorithm for many problems, does not always work. As does Newton's method, this algorithm sometimes "jumps off" to a wildly errant second iterate, after which it may be impossible to compute the residuals for the next iteration. The choice of starting values for the iterations can be crucial. There is art as well as science in the computation of nonlinear least squares estimates.[13] In the absence of information about starting values, a workable strategy is to try the Gauss–Newton iteration first. If it fails, go back to the initial starting values and try one of the more general algorithms, such as BFGS, treating minimization of the sum of squares as an otherwise ordinary optimization problem.

Example 7.9 Nonlinear Least Squares

Example 7.4 considered analysis of a nonlinear consumption function,

$$C = \alpha + \beta Y^{\gamma} + \varepsilon.$$

The linearized regression model is

$$C - (\alpha^0 + \beta^0 Y^{\gamma 0}) + (\alpha^0 1 + \beta^0 Y^{\gamma 0} + \gamma^0 \beta^0 Y^{\gamma 0} \ln Y) = \alpha + \beta(Y^{\gamma 0}) + \gamma(\beta^0 Y^{\gamma 0} \ln Y) + \varepsilon^0.$$

Combining terms, we find that the nonlinear least squares procedure reduces to iterated regression of

$$C^0 = C + \gamma^0 \beta^0 Y^{\gamma 0} \ln Y$$

on

$$\mathbf{x}^0 = \left[\frac{\partial h(.)}{\partial \alpha} \frac{\partial h(.)}{\partial \beta} \frac{\partial h(.)}{\partial \gamma}\right]' = \begin{bmatrix} 1 \\ Y^{\gamma^0} \\ \beta^0 Y^{\gamma^0} \ln Y \end{bmatrix}.$$

Finding the **starting values** for a nonlinear procedure can be difficult. Simply trying a convenient set of values can be unproductive. Unfortunately, there are no good rules for starting values, except that they should be as close to the final values as possible (not particularly helpful). When it is possible, an initial consistent estimator of $\boldsymbol{\beta}$ will be a good starting value. In many cases, however, the only consistent estimator available is the one we are trying to compute by least squares. For better or worse, trial and error is the most frequently used procedure. For the present model, a natural set of values can be obtained because a simple linear model is a special case. Thus, we can start α and β at the linear least squares values that would result in the special case of $\gamma = 1$ and use 1 for the starting value for γ. The **iterations** are begun at the least squares estimates for α and β and 1 for γ.

[13] See McCullough and Vinod (1999).

The solution is reached in eight iterations, after which any further iteration is merely fine tuning the hidden digits (i.e., those that the analyst would not be reporting to their reader; "gradient" is the scale-free convergence measure, δ, noted earlier). Note that the coefficient vector takes a very errant step after the first iteration—the sum of squares becomes huge—but the iterations settle down after that and converge routinely.

Begin NLSQ iterations. Linearized regression.

Iteration = 1; Sum of squares = 1536321.88; Gradient = 996103.930
Iteration = 2; Sum of squares = 0.184780956E + 12; Gradient = 0.184780452E + 12 ($\times 10^{12}$)
Iteration = 3; Sum of squares = 20406917.6; Gradient = 19902415.7
Iteration = 4; Sum of squares = 581703.598; Gradient = 77299.6342
Iteration = 5; Sum of squares = 504403.969; Gradient = 0.752189847
Iteration = 6; Sum of squares = 504403.216; Gradient = 0.526642396E-04
Iteration = 7; Sum of squares = 504403.216; Gradient = 0.511324981E-07
Iteration = 8; Sum of squares = 504403.216; Gradient = 0.606793426E-10

7.3 MEDIAN AND QUANTILE REGRESSION

We maintain the essential assumptions of the linear regression model,

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon,$$

where $E[\varepsilon|\mathbf{x}] = 0$ and $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. If $\varepsilon|\mathbf{x}$ is normally distributed, so that the distribution of $\varepsilon|\mathbf{x}$ is also symmetric, then the median, $\text{Med}[\varepsilon|\mathbf{x}]$, is also zero and $\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. Under these assumptions, least squares remains a natural choice for estimation of $\boldsymbol{\beta}$. But, as we explored in Example 4.3, **least absolute deviations (LAD)** is a possible alternative that might even be preferable in a small sample. Suppose, however, that we depart from the second assumption directly. That is, the statement of the model is

$$\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}.$$

This result suggests a motivation for LAD in its own right, rather than as a robust (to outliers) alternative to least squares.[14] The conditional median of $y_i|\mathbf{x}_i$ might be an interesting function. More generally, other quantiles of the distribution of $y_i|\mathbf{x}_i$ might also be of interest. For example, we might be interested in examining the various quantiles of the distribution of income or spending. Quantile regression (rather than least squares) is used for this purpose. The (linear) quantile regression model can be defined as

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that Prob } [y \leq \mathbf{x}'\boldsymbol{\beta}_q|\mathbf{x}] = q, 0 < q < 1. \qquad \textbf{(7-33)}$$

The **median regression** would be defined for $q = \frac{1}{2}$. Other focal points are the lower and upper quartiles, $q = \frac{1}{4}$ and $q = \frac{3}{4}$, respectively. We will develop the median regression in detail in Section 7.3.1, once again largely as an alternative estimator in the linear regression setting.

The quantile regression model is a richer specification than the linear model that we have studied thus far because the coefficients in (7-33) are indexed by q. The model

[14] In Example 4.3, we considered the possibility that in small samples with possibly thick-tailed disturbance distributions, the LAD estimator might have a smaller variance than least squares.

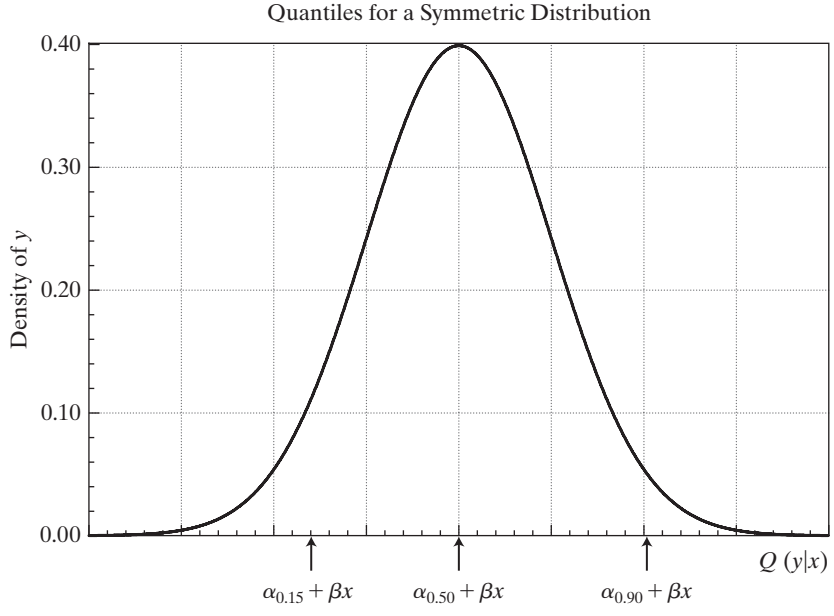
is semiparametric—it requires a much less detailed specification of the distribution of $y|\mathbf{x}$. In the simplest linear model with fixed coefficient vector, $\boldsymbol{\beta}$, the quantiles of $y|\mathbf{x}$ would be defined by variation of the constant term. The implication of the model is shown in Figure 7.4. For a fixed β and conditioned on x, the value of $\alpha_q + \beta x$ such that Prob$(y < \alpha_q + \beta x)$ is shown for $q = 0.15$, 0.5, and 0.9 in Figure 7.4. There is a value of α_q for each quantile. In Section 7.3.2, we will examine the more general specification of the quantile regression model in which the entire coefficient vector plays the role of α_q in Figure 7.4.

7.3.1 LEAST ABSOLUTE DEVIATIONS ESTIMATION

Least squares can be distorted by outlying observations. Recent applications in microeconomics and financial economics involving thick-tailed disturbance distributions, for example, are particularly likely to be affected by precisely these sorts of observations. (Of course, in those applications in finance involving hundreds of thousands of observations, which are becoming commonplace, this discussion is moot.) These applications have led to the proposal of "robust" estimators that are unaffected by outlying observations.[15] In this section, we will examine one of these, the least absolute deviations, or LAD estimator.

That least squares gives such large weight to large deviations from the regression causes the results to be particularly sensitive to small numbers of atypical data points

FIGURE 7.4 Quantile Regression Model.



[15]For some applications, see Taylor (1974), Amemiya (1985, pp. 70–80), Andrews (1974), Koenker and Bassett (1978), Li and Racine (2007), Henderson and Parmeter (2015), and a survey written at a very accessible level by Birkes and Dodge (1993). A somewhat more rigorous treatment is given by Hardle (1990).

when the sample size is small or moderate. The least absolute deviations (LAD) estimator has been suggested as an alternative that remedies (at least to some degree) the problem. The LAD estimator is the solution to the optimization problem,

$$\text{Min}_{\mathbf{b}_0} \sum_{i=1}^{n} |y_i - \mathbf{x}_i'\mathbf{b}_0|.$$

The LAD estimator's history predates least squares (which itself was proposed over 200 years ago). It has seen little use in econometrics, primarily for the same reason that Gauss's method (LS) supplanted LAD at its origination; LS is vastly easier to compute. Moreover, in a more modern vein, its statistical properties are more firmly established than LAD's and samples are usually large enough that the small sample advantage of LAD is not needed.

The LAD estimator is a special case of the quantile regression,

$$\text{Prob}[y_i \leq \mathbf{x}_i'\boldsymbol{\beta}_q] = q.$$

The LAD estimator estimates the *median regression*. That is, it is the solution to the quantile regression when $q = 0.5$. Koenker and Bassett (1978, 1982), Koenker and Hallock (2001), Huber (1967), and Rogers (1993) have analyzed this regression.[16] Their results suggest an estimator for the asymptotic covariance matrix of the quantile regression estimator,

$$Est.Asy.Var[\mathbf{b}_q] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

where $\mathbf{D}$ is a diagonal matrix containing weights,

$$d_i = \left[\frac{q}{f(0)}\right]^2 \text{ if } y_i - \mathbf{x}_i'\boldsymbol{\beta} \text{ is positive and } \left[\frac{1-q}{f(0)}\right]^2 \text{ otherwise,}$$

and $f(0)$ is the true density of the disturbances evaluated at 0.[17] [It remains to obtain an estimate of $f(0)$.] There is a useful symmetry in this result. Suppose that the true density were normal with variance σ^2. Then the preceding would reduce to $\sigma^2(\pi/2)(\mathbf{X}'\mathbf{X})^{-1}$, which is the result we used in Example 4.5. For more general cases, some other empirical estimate of $f(0)$ is going to be required. Nonparametric methods of density estimation are available.[18] But for the small sample situations in which techniques such as this are most desirable (our application below involves 25 observations), nonparametric kernel density estimation of a single ordinate is optimistic; these are, after all, asymptotic results. But asymptotically, as suggested by Example 4.3, the results begin overwhelmingly to favor least squares. For better or

[16]Powell (1984) has extended the LAD estimator to produce a robust estimator for the case in which data on the dependent variable are censored, that is, when negative values of y_i are recorded as zero. See Melenberg and van Soest (1996) for an application. For some related results on other semiparametric approaches to regression, see Butler et al. (1990) and McDonald and White (1993).

[17]Koenker suggests that for independent and identically distributed observations, one should replace d_i with the constant $a = q(1-q)/[f(F^{-1}(q))]^2 = [.50/f(0)]^2$ for the median (LAD) estimator. This reduces the expression to the true asymptotic covariance matrix, $a(\mathbf{X}'\mathbf{X})^{-1}$. The one given is a sample estimator which will behave the same in large samples. (Personal communication with the author.)

[18]See Section 12.4 and, for example, Johnston and DiNardo (1997, pp. 370–375).

worse, a convenient estimator would be a **kernel density estimator** as described in Section 12.4.1. Looking ahead, the computation would be

$$\hat{f}(0) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left[\frac{e_i}{h}\right],$$

where h is the **bandwidth** (to be discussed shortly), $K[.]$ is a weighting, or kernel function, and e_i, $i = 1, \ldots, n$ is the set of residuals. There are no hard and fast rules for choosing h; one popular choice is that used by Stata (2014), $h = .9s/n^{1/5}$. The kernel function is likewise discretionary, though it rarely matters much which one chooses; the logit kernel (see Table 12.2) is a common choice.

The **bootstrap** method of inferring statistical properties is well suited for this application. Since the efficacy of the bootstrap has been established for this purpose, the search for a formula for standard errors of the LAD estimator is not really necessary. The bootstrap estimator for the asymptotic covariance matrix can be computed as follows:

$$Est.Var[\mathbf{b}_{LAD}] = \frac{1}{R}\sum_{r=1}^{R}(\mathbf{b}_{LAD}(r) - \bar{\mathbf{b}}_{LAD})(\mathbf{b}_{LAD}(r) - \bar{\mathbf{b}}_{LAD})',$$

where $\mathbf{b}_{LAD}(r)$ is the rth LAD estimate of $\boldsymbol{\beta}$ based on a sample of n observations, drawn with replacement, from the original data set and $\bar{\mathbf{b}}_{LAD}$ is the mean of the r LAD estimators.

Example 7.10 LAD Estimation of a Cobb–Douglas Production Function

Zellner and Revankar (1970) proposed a generalization of the Cobb–Douglas production function that allows economies of scale to vary with output. Their statewide data on Y = value added (output), K = capital, L = labor, and N = the number of establishments in the transportation industry are given in Appendix Table F7.2. For this application, estimates of the Cobb–Douglas production function,

$$\ln(Y_i/N_i) = \beta_1 + \beta_2 \ln(K_i/N_i) + \beta_3 \ln(L_i/N_i) + \varepsilon_i,$$

are obtained by least squares and LAD. The standardized least squares residuals shown in Figure 7.5 suggest that two observations (Florida and Kentucky) are outliers by the usual construction. The least squares coefficient vectors with and without these two observations are (2.293, 0.279, 0.927) and (2.205, 0.261, 0.879), respectively, which bears out the suggestion that these two points do exert considerable influence. Table 7.5 presents the LAD estimates of the same parameters, with standard errors based on 500 bootstrap replications. The LAD estimates with and without these two observations are identical, so only the former are presented. Using the simple approximation of multiplying the corresponding OLS standard error by $(\pi/2)^{1/2} = 1.2533$ produces a surprisingly close estimate of the bootstrap-estimated standard errors for the two slope parameters (0.102, 0.123) compared with the bootstrap estimates of (0.124, 0.121). The second set of estimated standard errors are based on Koenker's suggested estimator, $0.25/\hat{f}^2(0) = 0.25/1.5467^2 = 0.104502$. The bandwidth and kernel function are those suggested earlier. The results are surprisingly consistent given the small sample size.

7.3.2 QUANTILE REGRESSION MODELS

The quantile regression model is

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that } \text{Prob}[y \leq \mathbf{x}'\boldsymbol{\beta}_q|\mathbf{x}] = q, 0 < q < 1.$$

This is a semiparametric specification. No assumption is made about the distribution of $y|\mathbf{x}$ or about its conditional variance. The fact that q can vary continuously (strictly) between zero and one means that there are an infinite number of possible parameter vectors. It seems reasonable to view the coefficients, which we might write $\boldsymbol{\beta}(q)$ less

FIGURE 7.5 Standardized Residuals for a Production Function.

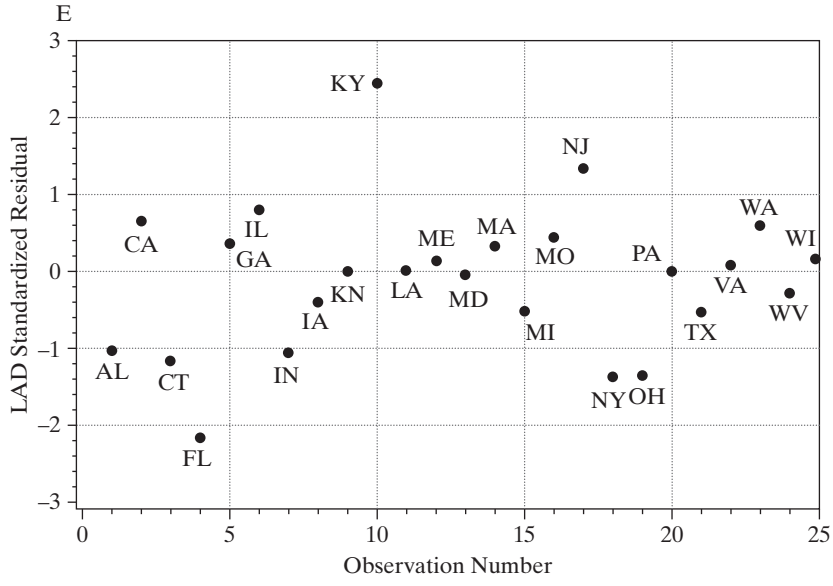


TABLE 7.5 LS and LAD Estimates of a Production Function

	Least Squares			*LAD*				
					Bootstrap		*Kernel Density*	
Coefficient	*Estimate*	*Standard Error*	*t Ratio*	*Estimate*	*Standard Error*	*t Ratio*	*Standard Error*	*t Ratio*
Constant	2.293	0.107	21.396	2.275	0.202	11.246	0.183	12.374
β_k	0.279	0.081	3.458	0.261	0.124	2.099	0.138	1.881
β_l	0.927	0.098	9.431	0.927	0.121	7.637	0.169	5.498
Σe^2	0.7814			0.7984				
$\Sigma \lvert e \rvert$	3.3652			3.2541				

as fixed parameters, as we do in the linear regression model, than loosely as *features* of the distribution of $y|\mathbf{x}$. For example, it is not likely to be meaningful to view $\boldsymbol{\beta}_{49}$ to be discretely different from $\boldsymbol{\beta}_{50}$ or to compute precisely a particular difference such as $\boldsymbol{\beta}_{.5} - \boldsymbol{\beta}_{.3}$. On the other hand, the qualitative difference, or possibly the lack of a difference, between $\boldsymbol{\beta}_{.3}$ and $\boldsymbol{\beta}_{.5}$ as displayed in our following example, may well be an interesting characteristic of the distribution.

The estimator, $\mathbf{b}_q$, of $\boldsymbol{\beta}_q$, for a specific quantile is computed by minimizing the function

$$F_n(\boldsymbol{\beta}_q|\mathbf{y},\mathbf{X}) = \sum_{i:y_i \geq x_i'\boldsymbol{\beta}_q}^{n} q|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q| + \sum_{i:y_i < x_i'\boldsymbol{\beta}_q}^{n} (1-q)|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q|$$
$$= \sum_{i=1}^{n} g(y_i - \mathbf{x}_i'\boldsymbol{\beta}_q|q),$$

where

$$g(e_{i,q}|q) = \begin{cases} qe_{i,q} & \text{if } e_{i,q} \geq 0 \\ (1-q)e_{i,q} & \text{if } e_{i,q} < 0 \end{cases}, e_{i,q} = y_i - \mathbf{x}_i'\boldsymbol{\beta}_q.$$

When $q = 0.5$, the estimator is the least absolute deviations estimator we examined in Example 4.5 and Section 7.3.1. Solving the minimization problem requires an iterative estimator. It can be set up as a linear programming problem.[19]

We cannot use the methods of Chapter 4 to determine the asymptotic covariance matrix of the estimator. But the fact that the estimator is obtained by minimizing a sum does lead to a set of results similar to those we obtained in Section 4.4 for least squares.[20] Assuming that the regressors are well behaved, the quantile regression estimator of $\boldsymbol{\beta}_q$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$Asy.Var.[b_q] = \frac{1}{n}\mathbf{H}^{-1}\,\mathbf{G}\mathbf{H}^{-1},$$

where

$$\mathbf{H} = \text{plim}\,\frac{1}{n}\sum_{i=1}^{n} f_q(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'$$

and

$$\mathbf{G} = \text{plim}\,\frac{q(1-q)}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'.$$

This is the result we had earlier for the LAD estimator, now with quantile q instead of 0.5. As before, computation is complicated by the need to compute the density of ε_q at zero. This will require either an approximation of uncertain quality or a specification of the particular density, which we have hoped to avoid. The usual approach, as before, is to use bootstrapping.

Example 7.11 Quantile Regression for Smoking Behavior

Laporte, Karimova, and Ferguson (2010) employed Becker and Murphy's (1988) model of rational addiction to study the behavior of a sample of Canadian smokers. The rational addiction model is a model of inter-temporal optimization, meaning that, rather than making independent decisions about how much to smoke in each period, the individual plots out an optimal lifetime smoking trajectory, conditional on future values of exogenous variables such as price. The optimal control problem which yields that trajectory incorporates the individual's attitudes to the harm smoking can do to her health and the rate at which she will trade the present against the future. This means that factors like the individual's degree of myopia are built into the trajectory of cigarette consumption which she will follow, and that consumption trajectory is what yields the forward-looking second-order difference equation which characterizes rational addiction behavior.[21]

The proposed empirical model is a dynamic regression,

$$C_t = \alpha + \mathbf{x}_t'\boldsymbol{\beta} + \gamma_1 C_{t+1} + \gamma_0 C_{t-1} + \varepsilon_t.$$

[19]See Koenker and D'Oray (1987) and Koenker (2005).

[20]See Buchinsky (1998).

[21]Laporte et al., p. 1064.

If it is assumed that $\mathbf{x}_t$ is fixed at $\mathbf{x}_*$ and ε_t is fixed at its expected value of zero, then a long run equilibrium consumption occurs where $C_t = C_{t-1} = C^*$ so that

$$C^* = \frac{\alpha + \mathbf{x}_*'\boldsymbol{\beta}}{1 - \gamma_1 - \gamma_0}.$$

(Some restrictions on the coefficients must hold for a finite positive equilibrium to exist. We can see, for example, $\gamma_0 + \gamma_1$ must be less than one.) The long run partial effects are then $\partial C^*/\partial x_{*k} = \beta_k/(1 - \gamma_0 - \gamma_1)$. Various covariates enter the model including, gender, whether smoking is restricted in the workplace, self-assessment of poor diet, price, and whether the individual jumped to zero consumption.

The analysis in the study is done primarily through graphical descriptions of the quantile regressions. Figure 7.6 (Figure 4 from the article) shows the estimates of the coefficient on a gender dummy variable in the model. The center line is the quantile-based coefficient on the dummy variable. The bands show 95% confidence intervals. (The authors do not mention how the standard errors are computed.) The dotted horizontal line shows the least squares estimate of the same coefficient. Note that it coincides with the 50th quantile estimate of this parameter.

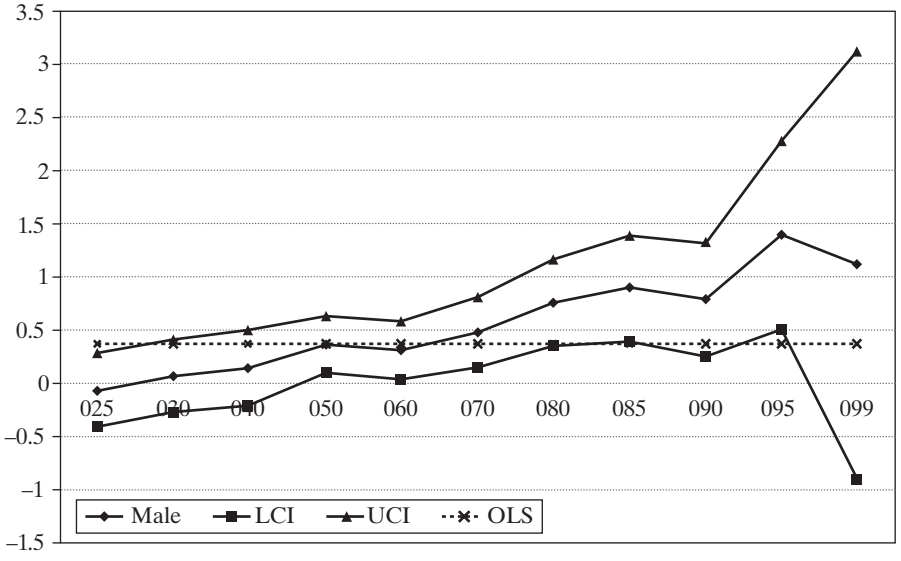
Example 7.12 Income Elasticity of Credit Card Expenditures

Greene (1992, 2007c) analyzed the default behavior and monthly expenditure behavior of a sample (13,444 observations) of credit card users. Among the results of interest in the study was an estimate of the income elasticity of the monthly expenditure. A quantile regression approach might be based on

$$Q[\ln \textit{Spending} \,|\, \mathbf{x}, q] = \beta_{1,q} + \beta_{2,q} \ln \textit{Income} + \beta_{3,q}\, \textit{Age} + \beta_{4,q}\, \textit{Dependents}.$$

The data in Appendix Table F7.3 contain these and numerous other covariates that might explain spending; we have chosen these three for this example only. The 13,444 observations in the

FIGURE 7.6 Male Coefficient in Quantile Regressions.



data set are based on credit card applications. Of the full sample, 10,499 applications were approved and the next 12 months of spending and default behavior were observed.[22] Spending is the average monthly expenditure in the 12 months after the account was initiated. Average monthly income and number of household dependents are among the demographic data in the application. Table 7.6 presents least squares estimates of the coefficients of the conditional mean function as well as full results for several quantiles.[23] Standard errors are shown for the least squares and median ($q = 0.5$) results. The least squares estimate of 1.08344 is slightly and significantly greater than one—the estimated standard error is 0.03212 so the t statistic is $(1 - 1.08344)/0.03212 = 2.60$. This suggests an aspect of consumption behavior that might not be surprising. However, the very large amount of variation over the range of quantiles might not have been expected. We might guess that at the highest levels of spending for any income level, there is (comparably so) some saturation in the response of spending to changes in income.

Figure 7.7 displays the estimates of the income elasticity of expenditure for the range of quantiles from 0.1 to 0.9, with the least squares estimate, which would correspond to the fixed value at all quantiles, shown in the center of the figure. Confidence limits shown in the figure are based on the asymptotic normality of the estimator. They are computed as the estimated income elasticity plus and minus 1.96 times the estimated standard error. Figure 7.8 shows the implied quantile regressions for $q = 0.1$, 0.3, 0.5, 0.7, and 0.9.

TABLE 7.6 Estimated Quantile Regression Models

	Estimated Parameters			
Quantile	***Constant***	***ln Income***	***Age***	***Dependents***
0.1	−6.73560	1.40306	−0.03081	−0.04297
0.2	−4.31504	1.16919	−0.02460	−0.04630
0.3	−3.62455	1.12240	−0.02133	−0.04788
0.4	−2.98830	1.07109	−0.01859	−0.04731
(Median) 0.5	−2.80376	1.07493	−0.01699	−0.04995
Std.Error	(0.24564)	(0.03223)	(0.00157)	(0.01080)
t	−11.41	33.35	−10.79	−4.63
Least Squares	−3.05581	1.08344	−0.01736	−0.04461
Std.Error	(0.23970)	(0.03212)	(0.00135)	(0.01092)
t	−12.75	33.73	−12.88	−4.08
0.6	−2.05467	1.00302	−0.01478	−0.04609
0.7	−1.63875	0.97101	−0.01190	−0.03803
0.8	−0.94031	0.91377	−0.01126	−0.02245
0.9	−0.05218	0.83936	−0.00891	−0.02009

[22]The expenditure data are taken from the credit card records while the income and demographic data are taken from the applications. While it might be tempting to use, for example, Powell's (1986a,b) censored quantile regression estimator to accommodate this large cluster of zeros for the dependent variable, this approach would misspecify the model—the *zeros* represent nonexistent observations, not true zeros and not missing data. A more detailed approach—the one used in the 1992 study—would model separately the presence or absence of the observation on spending and then model spending conditionally on acceptance of the application. We will revisit this issue in Chapter 19 in the context of the sample selection model. The income data are censored at 100,000 and 220 of the observations have expenditures that are filled with $1 or less. We have not "cleaned" the data set for these aspects. The full 10,499 observations have been used as they are in the original data set.

[23]We would note, if (7-33) is the statement of the model, then it does not follow that the conditional mean function is a linear regression. That would be an additional assumption.

FIGURE 7.7 Estimates of Income Elasticity of Expenditure.

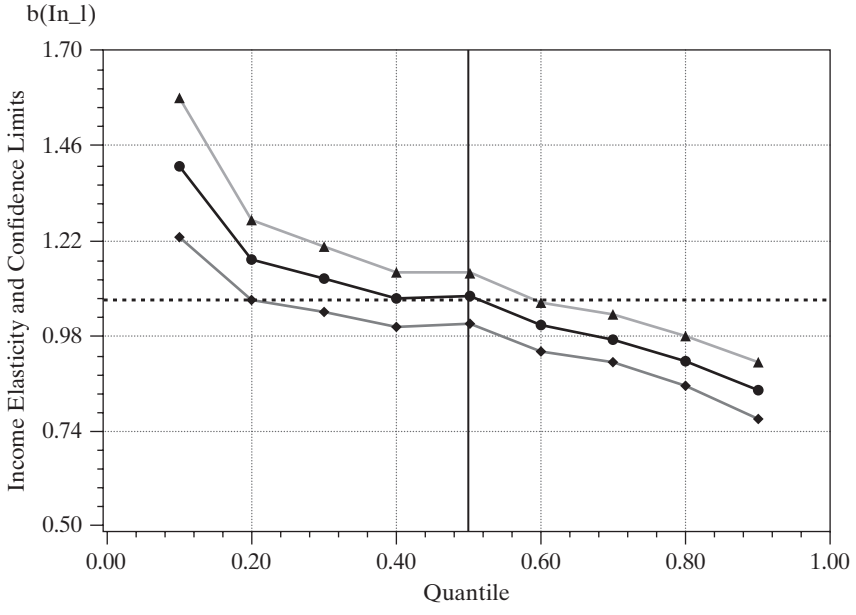
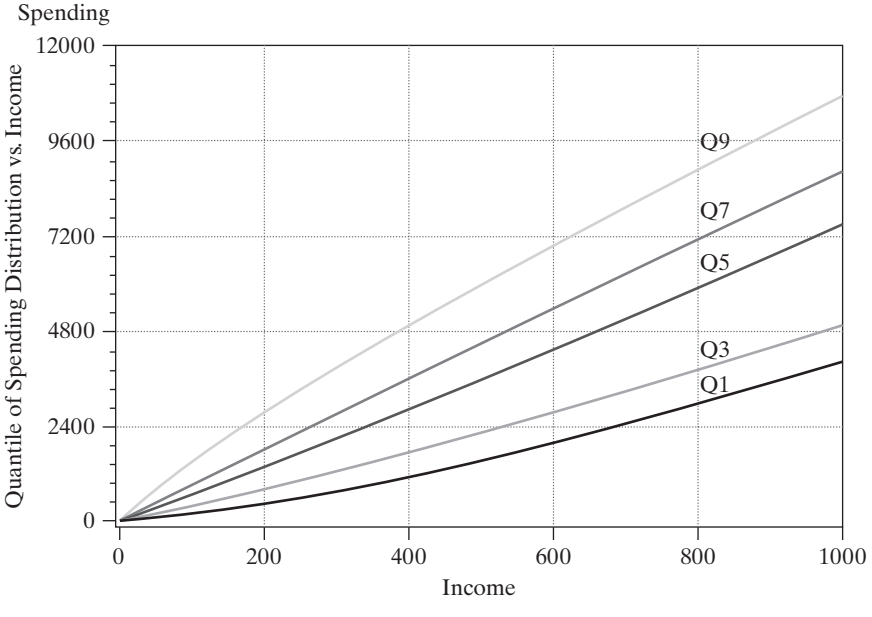


FIGURE 7.8 Quantile Regressions for Spending vs. Income.



7.4 PARTIALLY LINEAR REGRESSION

The proper functional form in the linear regression is an important specification issue. We examined this in detail in Chapter 6. Some approaches, including the use of dummy variables, logs, quadratics, and so on, were considered as a means of capturing nonlinearity. The translog model in particular (Example 2.4) is a well-known approach to approximating an unknown nonlinear function. Even with these approaches, the researcher might still be interested in relaxing the assumption of functional form in the model. The partially linear model is another approach.[24] Consider a regression model in which one variable, x, is of particular interest, and the functional form with respect to x is problematic. Write the model as

$$y_i = f(x_i) + \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where the data are assumed to be well behaved and, save for the functional form, the assumptions of the classical model are met. The function $f(x_i)$ remains unspecified. As stated, estimation by least squares is not feasible until $f(x_i)$ is specified. Suppose the data were such that they consisted of pairs of observations $(y_{j1}, y_{j2}), j = 1, \ldots, n/2$, in which $x_{j1} = x_{j2}$ within every pair. If so, then estimation of $\boldsymbol{\beta}$ could be based on the simple transformed model,

$$y_{j2} - y_{j1} = (\mathbf{z}_{j2} - \mathbf{z}_{j1})'\boldsymbol{\beta} + (\varepsilon_{j2} - \varepsilon_{j1}), \quad j = 1, \ldots, n/2.$$

As long as observations are independent, the constructed disturbances, v_i, still have zero mean, variance now $2\sigma^2$, and remain uncorrelated across pairs, so a classical model applies and least squares is actually optimal. Indeed, with the estimate of $\boldsymbol{\beta}$, say, $\hat{\boldsymbol{\beta}}_d$ in hand, a noisy estimate of $f(x_i)$ could be estimated with $y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_d$ (the estimate contains the estimation error as well as ε_i).[25]

The problem, of course, is that the enabling assumption is heroic. Data would not behave in that fashion unless they were generated experimentally. The logic of the partially linear regression estimator is based on this observation nonetheless. Suppose that the observations are sorted so that $x_1 < x_2 < \cdots < x_n$. Suppose, as well, that this variable is well behaved in the sense that, as the sample size increases, this sorted data vector more completely and uniformly fills the space within which x_i is assumed to vary. Then, intuitively, the difference is "almost" right, and becomes better as the sample size grows.[26] A theory is also developed for a better differencing of groups of two or more observations. The transformed observation is $y_{d,i} = \sum_{m=0}^{M} d_m y_{i-m}$, where $\sum_{m=0}^{M} d_m = 0$ and $\sum_{m=0}^{M} d_m^2 = 1$. (The data are not separated into nonoverlapping groups for this transformation—we merely used that device to motivate the technique.) The pair of weights for $M = 1$ is obviously $\pm\sqrt{0.5}$—this is just a scaling of the simple difference, 1, −1. Yatchew [1998, p. 697)] tabulates *optimal* differencing weights for $M = 1, \ldots, 10$. The values for $M = 2$ are (0.8090, −0.500, −0.3090) and for $M = 3$ are (0.8582, −0.3832, −0.2809, −0.1942). This estimator is shown to be

[24] Analyzed in detail by Yatchew (1998, 2000) and Härdle, Liang, and Gao (2000).

[25] See Estes and Honoré (1995) who suggest this approach (with simple differencing of the data).

[26] Yatchew (1997, 1998) goes more deeply into the underlying theory.

consistent, asymptotically normally distributed, and have asymptotic covariance matrix,[27]

$$Asy.Var[\hat{\boldsymbol{\beta}}_d] = \left(1 + \frac{1}{2M}\right)\frac{\sigma_v^2}{n}E_x[Var[\mathbf{z}|x]].$$

The matrix can be estimated using the sums of squares and cross products of the differenced data. The residual variance is likewise computed with

$$\hat{\sigma}_v^2 = \frac{\sum_{i=M+1}^{n}(y_{d,i} - \mathbf{z}'_{d,i}\hat{\boldsymbol{\beta}}_d)^2}{n - M}.$$

Yatchew suggests that the partial residuals, $y_{d,i} - \mathbf{z}'_{d,i}\hat{\boldsymbol{\beta}}_d$, be smoothed with a kernel density estimator to provide an improved estimator of $f(x_i)$. Manzan and Zeron (2010) present an application of this model to the U.S. gasoline market.

Example 7.13 Partially Linear Translog Cost Function

Yatchew (1998, 2000) applied this technique to an analysis of scale effects in the costs of electricity supply. The cost function, following Nerlove (1963) and Christensen and Greene (1976), was specified to be a translog model (see Example 2.4 and Section 10.3.2) involving labor and capital input prices, other characteristics of the utility, and the variable of interest, the number of customers in the system, *C*. We will carry out a similar analysis using Christensen and Greene's 1970 electricity supply data. The data are given in Appendix Table F4.4. (See Section 10.3.1 for description of the data.) There are 158 observations in the data set, but the last 35 are holding companies that are comprised of combinations of the others. In addition, there are several extremely small New England utilities whose costs are clearly unrepresentative of the best practice in the industry. We have done the analysis using firms 6–123 in the data set. Variables in the data set include Q = output, C = total cost, and PK, PL, and PF = unit cost measures for capital, labor, and fuel, respectively. The parametric model specified is a restricted version of the Christensen and Greene model,

$$c = \beta_1 k + \beta_2 l + \beta_3 q + \beta_4 (q^2/2) + \beta_5 + \varepsilon,$$

where $c = \ln[C/(Q \times PF)]$, $k = \ln(PK/PF)$, $l = \ln(PL/PF)$, and $q = \ln Q$. The partially linear model substitutes $f(q)$ for the last three terms. The division by PF ensures that average cost is homogeneous of degree one in the prices, a theoretical necessity. The estimated equations, with estimated standard errors, are shown here.

$$\begin{array}{lllllllll}
\text{(parametric)} & c & = & -7.32 & + & 0.069k & + & 0.241 - 0.569q + 0.057q^2/2 + \varepsilon, & \\
 & & & (0.333) & & (0.065) & & (0.069)\ (0.042)\ (0.006) & s = 0.13949 \\
\text{(partially linear)} & c_d & = & & & 0.108k_d & + & 0.163l_d + f(q) + v & \\
 & & & & & (0.076) & & (0.081) & s = 0.16529
\end{array}$$

7.5 NONPARAMETRIC REGRESSION

The regression function of a variable y on a single variable x is specified as

$$y = \mu(x) + \varepsilon.$$

No assumptions about distribution, homoscedasticity, serial correlation or, most importantly, functional form are made at the outset; $\mu(x)$ may be quite nonlinear. Because this is the conditional mean, the only substantive restriction would be that

[27] Yatchew (2000, p. 191) denotes this covariance matrix $E[\text{Cov}[\mathbf{z}|\mathbf{x}]]$.

deviations from the conditional mean function are not a function of (correlated with) x. We have already considered several possible strategies for allowing the conditional mean to be nonlinear, including spline functions, polynomials, logs, dummy variables, and so on. But each of these is a "global" specification. The functional form is still the same for all values of x. Here, we are interested in methods that do not assume any particular functional form.

The simplest case to analyze would be one in which several (different) observations on y_i were made with each specific value of x_i. Then, the conditional mean function could be estimated naturally using the simple group means. The approach has two shortcomings, however. Simply connecting the points of means, $(x_i, \bar{y}|x_i)$ does not produce a smooth function. The method would still be assuming something specific about the function between the points, which we seek to avoid. Second, this sort of data arrangement is unlikely to arise except in an experimental situation. Given that data are not likely to be grouped, another possibility is a piecewise regression in which we define "neighborhoods" of points around each x of interest and fit a separate linear or quadratic regression in each neighborhood. This returns us to the problem of continuity that we noted earlier, but the method of splines, discussed in Section 6.3.1, is actually designed specifically for this purpose. Still, unless the number of neighborhoods is quite large, such a function is still likely to be crude.

Smoothing techniques are designed to allow construction of an estimator of the conditional mean function without making strong assumptions about the behavior of the function between the points. They retain the usefulness of the **nearest neighbor** concept but use more elaborate schemes to produce smooth, well-behaved functions. The general class may be defined by a conditional mean estimating function

$$\hat{\mu}(x^*) = \sum_{i=1}^{n} w_i(x^*|x_1, x_2, \ldots, x_n)y_i = \sum_{i=1}^{n} w_i(x^*|\mathbf{x})y_i,$$

where the weights sum to 1. The linear least squares regression line is such an estimator. The predictor is

$$\hat{\mu}(x^*) = a + bx^*,$$

where a and b are the least squares constant and slope. For this function, you can show that

$$w_i(x^*|\mathbf{x}) = \frac{1}{n} + \frac{x^*(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The problem with this particular weighting function, which we seek to avoid here, is that it allows every x_i to be in the neighborhood of x^*, but it does not reduce the weight of any x_i when it is far from x^*. A number of **smoothing functions** have been suggested that are designed to produce a better behaved regression function.[28] We will consider two.

The locally weighted smoothed regression estimator (*loess* or *lowess* depending on your source) is based on explicitly defining a neighborhood of points that is close to x^*. This requires the choice of a bandwidth, h. The **neighborhood** is the set of points for which $|x^* - x_i|$ is small. For example, the set of points that are within the range $x^* \pm h/2$ might constitute the neighborhood. The choice of bandwidth is crucial, as we

[28]See Cleveland (1979) and Schimek (2000).

will explore in the following example, and is also a challenge. There is no single best choice. A common choice is **Silverman's** (1986) **rule of thumb**,

$$h_{Silverman} = \frac{.9[\min(s, IQR)]}{1.349n^{0.2}},$$

where s is the sample standard deviation and IQR is the interquartile range (0.75 quantile minus 0.25 quantile). A suitable weight is then required. Cleveland (1979) recommends the tricube weight,

$$T_i(x^*|\mathbf{x}, h) = \left[1 - \left(\frac{|x_i - x^*|}{h}\right)^3\right]^3.$$

Combining terms, then the weight for the loess smoother is

$$w_i(x^*|\mathbf{x}, h) = 1(x_i \text{ in the neighborhood}) \times T_i(x^*|\mathbf{x}, h).$$

The bandwidth is essential in the results. A wider neighborhood will produce a smoother function, but the wider neighborhood will track the data less closely than a narrower one. A second possibility, similar to the least squares approach, is to allow the neighborhood to be all points but make the weighting function decline smoothly with the distance between x^* and any x_i. A variety of **kernel functions** are used for this purpose. Two common choices are the **logistic kernel**,

$$K(x^*|x_i, h) = \Lambda(v_i)[1 - \Lambda(v_i)] \text{ where } \Lambda(v_i) = \exp(v_i)/[1 + \exp(v_i)], v_i = (x_i - x^*)/h,$$

and the **Epanechnikov kernel**,

$$K(x^*|x_i, h) = 0.75(1 - 0.2v_i^2)/\sqrt{5} \text{ if } |v_i| \leq 5 \text{ and 0 otherwise.}$$

This produces the kernel weighted regression estimator,

$$\hat{\mu}(x^*|\mathbf{x}, h) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left[\frac{x_i - x^*}{h}\right]y_i}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left[\frac{x_i - x^*}{h}\right]},$$

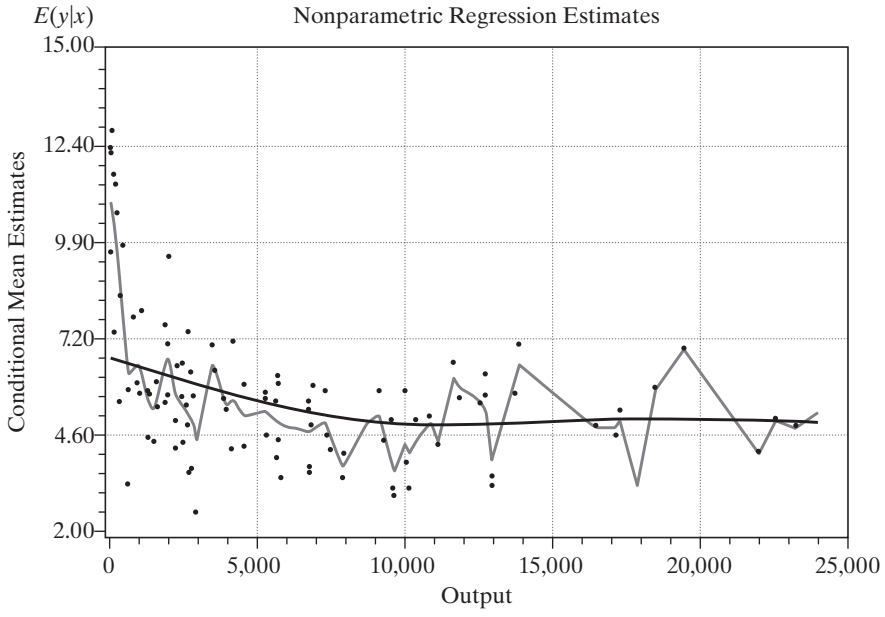
which has become a standard tool in nonparametric analysis.

Example 7.14 A Nonparametric Average Cost Function

In Example 7.13, we fit a partially linear regression for the relationship between average cost and output for electricity supply. Figure 7.9 shows the less ambitious nonparametric regressions of average cost on output. The overall picture is the same as in the earlier example. The kernel function is the logistic density in both cases. The functions in Figure 7.9 use bandwidths of 2,000 and 100. Because 2,000 is a fairly large proportion of the range of variation of output, this function is quite smooth. The other function in Figure 7.9 uses a bandwidth of only 100. The function tracks the data better, but at an obvious cost. The example demonstrates what we and others have noted often. The choice of bandwidth in this exercise is crucial.

Data smoothing is essentially data driven. As with most nonparametric techniques, inference is not part of the analysis—this body of results is largely descriptive. As can be seen in the example, nonparametric regression can reveal interesting characteristics

FIGURE 7.9 Nonparametric Cost Functions.



of the data set. For the econometrician, however, there are a few drawbacks. There is no danger of misspecifying the conditional mean function; however, the great generality of the approach limits the ability to test one's specification or the underlying theory.[29] Most relationships are more complicated than a simple conditional mean of one variable. In Example 7.14, some of the variation in average cost relates to differences in factor prices (particularly fuel) and in load factors. Extensions of the fully nonparametric regression to more than one variable is feasible, but very cumbersome.[30] A promising approach is the partially linear model considered earlier. Henderson and Parmeter (2015) describe extensions of the kernel regression that accommodate multiple regression.

7.6 SUMMARY AND CONCLUSIONS

In this chapter, we extended the regression model to a form that allows nonlinearity in the parameters in the regression function. The results for interpretation, estimation, and hypothesis testing are quite similar to those for the linear model. The two crucial differences between the two models are, first, the more involved estimation procedures needed for the nonlinear model and, second, the ambiguity of the interpretation of the coefficients in the nonlinear model (because the derivatives of the regression are often nonconstant, in contrast to those in the linear model).

[29]See, for example, Blundell, Browning, and Crawford's (2003) extensive study of British expenditure patterns.

[30]See Härdle (1990), Li and Racine (2007), and Henderson and Parmeter (2015).

Key Terms and Concepts

- Bandwidth
- Bootstrap
- Box–Cox transformation
- Conditional mean function
- Conditional median
- Delta method
- Epanechnikov kernel
- GMM estimator
- Identification condition
- Identification problem
- Indirect utility function
- Interaction term
- Iteration
- Jacobian
- Kernel density estimator
- Kernel functions
- Lagrange multiplier test
- Least absolute deviations (LAD)
- Linear regression model
- Linearized regression model
- Logistic kernel
- Median regression
- Nearest neighbor
- Neighborhood
- Nonlinear least squares
- Nonlinear regression model
- Nonparametric regression
- Orthogonality condition
- Overidentifying restrictions
- Partially linear model
- Pseudoregressors
- Quantile regression model
- Roy's identity
- Semiparametric
- Silverman's rule of thumb
- Smoothing function
- Starting values

Exercises

1. Describe how to obtain nonlinear least squares estimates of the parameters of the model $y = \alpha x^{\beta} + \varepsilon$.
2. Verify the following differential equation, which applies to the Box–Cox transformation:

$$\frac{d^i x^{(\lambda)}}{d\lambda^i} = \left(\frac{1}{\lambda}\right)\left[x^{\lambda}(\ln x)^i - \frac{i d^{i-1} x^{(\lambda)}}{d\lambda^{i-1}}\right]. \tag{7-34}$$

Show that the limiting sequence for $\lambda = 0$ is

$$\lim_{\lambda \to 0} \frac{d^i x^{(\lambda)}}{d\lambda^i} = \frac{(\ln x)^{i+1}}{i+1}. \tag{7-35}$$

These results can be used to great advantage in deriving the actual second derivatives of the log-likelihood function for the Box–Cox model.

Applications

1. Using the Box–Cox transformation, we may specify an alternative to the Cobb–Douglas model as

$$\ln Y = \alpha + \beta_k \frac{(K^{\lambda} - 1)}{\lambda} + \beta_l \frac{(L^{\lambda} - 1)}{\lambda} + \varepsilon.$$

Using Zellner and Revankar's data in Appendix Table F7.2, estimate α, β_k, β_l, and λ by using the scanning method suggested in Example 7.5. (Do not forget to scale Y, K, and L by the number of establishments.) Use (7-24), (7-15), and (7-16) to compute the appropriate asymptotic standard errors for your estimates. Compute the two output elasticities, $\partial \ln Y/\partial \ln K$ and $\partial \ln Y/\partial \ln L$, at the sample means of K and L. (*Hint:* $\partial \ln Y/\partial \ln K = K\partial \ln Y/\partial K$.)

2. For the model in Application 1, test the hypothesis that $\lambda = 0$ using a Wald test and a Lagrange multiplier test. Note that the restricted model is the Cobb–Douglas

loglinear model. The LM test statistic is shown in (7-22). To carry out the test, you will need to compute the elements of the fourth column of $\mathbf{X}^0$, the pseudoregressor corresponding to λ is $\partial E[y|x]/\partial\lambda|\,\lambda = 0$. Result (7-35) will be useful.

3. The National Institute of Standards and Technology (NIST) has created a Web site that contains a variety of estimation problems, with data sets, designed to test the accuracy of computer programs. (The URL is http://www.itl.nist.gov/div898/strd/.) One of the five suites of test problems is a set of 27 nonlinear least squares problems, divided into three groups: easy, moderate, and difficult. We have chosen one of them for this application. You might wish to try the others (perhaps to see if the software you are using can solve the problems). This is the Misralc problem (http://www.itl.nist.gov/div898/strd/nls/data/misra1c.shtml). The nonlinear regression model is

$$y_i = h(x, \boldsymbol{\beta}) + \varepsilon$$
$$= \beta_1\left(1 - \frac{1}{\sqrt{1 + 2\beta_2 x_i}}\right) + \varepsilon_i.$$

The data are as follows:

Y	*X*
10.07	77.6
14.73	114.9
17.94	141.1
23.93	190.8
29.61	239.9
35.18	289.0
40.02	332.8
44.82	378.4
50.76	434.8
55.05	477.3
61.01	536.8
66.40	593.1
75.47	689.1
81.78	760.0

For each problem posed, NIST also provides the "certified solution" (i.e., the right answer). For the Misralc problem, the solutions are as follows:

	Estimate	*Estimated Standard Error*
β_1	6.3642725809E+02	4.6638326572E+00
β_2	2.0813627256E–04	1.7728423155E–06
$\mathbf{e}'\mathbf{e}$	4.0966836971E–02	
$s^2 = \mathbf{e}'\mathbf{e}/(n - K)$	5.8428615257E–02	

Finally, NIST provides two sets of starting values for the iterations, generally one set that is "far" from the solution and a second that is "close" to the solution. For this problem, the starting values provided are $\boldsymbol{\beta}^1 = (500, 0.0001)$ and

$\boldsymbol{\beta}^2 = (600, 0.0002)$. The exercise here is to reproduce the NIST results with your software. [For a detailed analysis of the NIST nonlinear least squares benchmarks with several well-known computer programs, see McCullough (1999).]

4. In Example 7.1, the CES function is suggested as a model for production,

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln [\delta K^{-\rho} + (1 - \delta) L^{-\rho}] + \varepsilon. \tag{7-36}$$

Example 6.19 suggested an indirect method of estimating the parameters of this model. The function is linearized around $\rho = 0$, which produces an intrinsically linear approximation to the function,

$$\ln y = \beta_1 + \beta_2 \ln K + \beta_3 \ln L + \beta_4 [1/2(\ln K - \text{Ln } L)^2] + \varepsilon,$$

where $\beta_1 = \ln \gamma$, $\beta_2 = \nu\delta$, $\beta_3 = \nu(1 - \delta)$ and $\beta_4 = \rho\nu\delta(1 - \delta)$. The approximation can be estimated by linear least squares. Estimates of the structural parameters are found by inverting the preceding four equations. An estimator of the asymptotic covariance matrix is suggested using the delta method. The parameters of (7-36) can also be estimated directly using nonlinear least squares and the results given earlier in this chapter.

Christensen and Greene's (1976) data on U.S. electricity generation are given in Appendix Table F4.4. The data file contains 158 observations. Using the first 123, fit the CES production function, using capital and fuel as the two factors of production rather than capital and labor. Compare the results obtained by the two approaches, and comment on why the differences (which are substantial) arise.

The following exercises require specialized software. The relevant techniques are available in several packages that might be in use, such as *SAS, Stata,* or *NLOGIT*. The exercises are suggested as departure points for explorations using a few of the many estimation techniques listed in this chapter.

5. Using the gasoline market data in Appendix Table F2.2, use the partially linear regression method in Section 7.4 to fit an equation of the form

$$\ln(G/Pop) = \beta_1 \ln(Income) + \beta_2 \ln P_{new\ cars} + \beta_3 \ln P_{used\ cars} + g(\ln P_{gasoline}) + \varepsilon.$$

6. To continue the analysis in Application 5, consider a nonparametric regression of *G/Pop* on the price. Using the nonparametric estimation method in Section 7.5, fit the nonparametric estimator using a range of bandwidth values to explore the effect of bandwidth.

8

ENDOGENEITY AND INSTRUMENTAL VARIABLE ESTIMATION

8.1 INTRODUCTION

The assumption that $\mathbf{x}_i$ and ε_i are uncorrelated in the linear regression model,

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon, \tag{8-1}$$

has been crucial in the development thus far. But there are many applications in which this assumption is untenable. Examples include models of treatment effects such as those in Examples 6.8–6.13, models that contain variables that are measured with error, dynamic models involving expectations, and a large variety of common situations that involve variables that are unobserved, or for other reasons are omitted from the equation. Without the assumption that the disturbances and the regressors are uncorrelated, none of the proofs of consistency or unbiasedness of the least squares estimator that were obtained in Chapter 4 will remain valid, so the least squares estimator loses its appeal. This chapter will develop an estimation method that arises in situations such as these.

It is convenient to partition $\mathbf{x}$ in (8-1) into two sets of variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, with the assumption that $\mathbf{x}_1$ is not correlated with ε and $\mathbf{x}_2$ is, or may be (part of the empirical investigation). We are assuming that $\mathbf{x}_1$ is **exogenous** in the model—see Assumption A.3 in the statement of the linear regression model in Section 2.3. It will follow that $\mathbf{x}_2$ is, by this definition, **endogenous** in the model. How does endogeneity arise? Example 8.1 suggests some common settings.

Example 8.1 Models with Endogenous Right-Hand-Side Variables

The following models and settings will appear at various points in this book.

Omitted Variables: In Example 4.2, we examined an equation for gasoline consumption of the form

$$\ln G = \beta_1 + \beta_2 \ln \mathit{Price} + \beta_3 \ln \mathit{Income} + \varepsilon.$$

When income is improperly omitted from this (any) demand equation, the resulting "model" is

$$\ln G = \beta_1 + \beta_2 \ln \mathit{Price} + w,$$

where $w = \beta_3 \ln \mathit{Income} + \varepsilon$. Linear regression of ln$G$ on a constant and ln*Price* does not consistently estimate (β_1, β_2) if ln*Price* is correlated with w. It surely will be in aggregate time-series data. The omitted variable reappears in the equation, in the disturbance, causing **omitted variable bias** in the least squares estimator of the misspecified equation.

Berry, Levinsohn, and Pakes (1995) examined the equilibrium in the U.S. automobile market. The centerpiece of the model is a random utility, multinomial choice model. For consumer i in market t, the utility of brand choice j is $U_{ijt} = U(w_i, p_{jt}, \mathbf{x}_{jt}, \mathbf{f}_{jt} | \boldsymbol{\beta})$, where w_i is individual heterogeneity, p_{jt} is the price, $\mathbf{x}_{jt}$ is a vector of observed attributes, and $\mathbf{f}_{jt}$ is a vector of unobserved features of the brand. Under the assumptions of random utility maximizing, and

aggregating over individuals, the model produces a market share equation, $s_{jt} = s_j(\mathbf{p}_t, \mathbf{X}_t, \mathbf{f}_t | \boldsymbol{\beta})$. Because $\mathbf{f}_t$ is unobserved features that consumers care about (i.e., $\mathbf{f}_{jt}$ influences the market share of brand j), and $\mathbf{f}_{jt}$ is reflected in the price of the brand, p_{jt}, $\mathbf{p}_t$ is endogenous in this choice model that is based on observed market shares.

Endogenous Treatment Effects: Krueger and Dale (1999) and Dale and Krueger (2002, 2011) examined the effect of attendance at an elite college on lifetime earnings. The regression model with a "treatment effect" dummy variable, T, which equals one for those who attended an elite college and zero otherwise, appears as

$$\ln y = \mathbf{x}'\boldsymbol{\beta} + \delta T + \varepsilon.$$

Least squares regression of a measure of earnings, ln y, on $\mathbf{x}$ and T attempts to produce an estimate of δ, the impact of the treatment. It seems inevitable, however, that some unobserved determinants of lifetime earnings, such as ambition, inherent abilities, persistence, and so on would also determine whether the individual had an opportunity to attend an elite college. If so, then the least squares estimator of δ will inappropriately attribute the effect to the treatment, rather than to these underlying factors. Least squares will not consistently estimate δ, ultimately because of the correlation between T and ε.

In order to quantify definitively the impact of attendance at an elite college on the individuals who did so, the researcher would have to conduct an impossible experiment. Individuals in the sample would have to be observed twice, once having attended the elite college and a second time (in a second lifetime) without having done so. Whether comparing individuals who attended elite colleges to other individuals who did not adequately measures the **effect of the treatment on the treated** individuals is the subject of a vast current literature. See, for example, Imbens and Wooldridge (2009) for a survey.

Simultaneous Equations: In an equilibrium model of price and output determination in a market, there would be equations for both supply and demand. For example, a model of output and price determination in a product market might appear,

$$\begin{aligned}
\text{(Demand)} \quad & \textit{Quantity}_D = \alpha_0 + \alpha_1\, \textit{Price} + \alpha_2\, \textit{Income} + \varepsilon_D, \\
\text{(Supply)} \quad & \textit{Quantity}_S = \beta_0 + \beta_1\, \textit{Price} + \beta_2\, \textit{Input Price} + \varepsilon_S, \\
\text{(Equilibrium)} \quad & \textit{Quantity}_D = \textit{Quantity}_S.
\end{aligned}$$

Consider attempting to estimate the parameters of the demand equation by regression of a time series of equilibrium quantities on equilibrium prices and incomes. The equilibrium price is determined by the equation of the two quantities. By imposing the equilibrium condition, we can solve for $\textit{Price} = (\alpha_0 - \beta_0 + \alpha_2\, \textit{Income} - \beta_2\, \textit{Input Price} + \varepsilon_D - \varepsilon_S)/(\beta_1 - \alpha_1)$. The implication is that *Price* is correlated with ε_D—if an external shock causes ε_D to change, that induces a shift in the demand curve and ultimately causes a new equilibrium *Price*. Least squares regression of quantity on *Price* and *Income* does not estimate the parameters of the demand equation consistently. This "feedback" between ε_D and *Price* in this model produces **simultaneous equations bias** in the least squares estimator.

Dynamic Panel Data Models: In Chapter 11, we will examine a dynamic **random effects** model of the form $y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \varepsilon_{it} + u_i$ where u_i contains the time-invariant unobserved features of individual i. Clearly, in this case, the regressor $y_{i,t-1}$ is correlated with the disturbance, $(\varepsilon_{it} + u_i)$—the unobserved heterogeneity is present in y_{it} in every period. In Chapter 13, we will examine a model for municipal expenditure of the form $S_{it} = f(S_{i,t-1}, \ldots) + \varepsilon_{it}$. The disturbances are assumed to be freely correlated across periods, so both $S_{i,t-1}$ and ε_{it} are correlated with $\varepsilon_{i,t-1}$. It follows that they are correlated with each other, which means that this model, even without time-persistent effects, does not satisfy the assumptions of the linear regression model. The regressors and disturbances are correlated.

Omitted Parameter Heterogeneity: Many cross-country studies of economic growth have the following structure (greatly simplified for purposes of this example),

$$\Delta \ln Y_{it} = \alpha_i + \theta_i t + \beta_i \Delta \ln Y_{i,t-1} + \varepsilon_{it},$$

where $\Delta \ln Y_{it}$ is the growth rate of country i in year t.[1] Note that the coefficients in the model are country specific. What does least squares regression of growth rates of income on a time trend and lagged growth rates estimate? Rewrite the growth equation as

$$\begin{aligned}\Delta \ln Y_{it} &= \alpha + \theta t + \beta(\Delta \ln Y_{i,t-1}) + (\alpha_i - \alpha) + (\theta_i - \theta)t + (\beta_i - \beta)(\Delta \ln Y_{i,t-1}) + \varepsilon_{it} \\ &= \alpha + \theta t + \beta(\Delta \ln Y_{i,t-1}) + w_{it}.\end{aligned}$$

We assume that the "average" parameters, α, θ, and β, are meaningful fixed parameters to be estimated. Does the least squares regression of $\Delta \ln Y_{it}$ on a constant, t, and $\Delta \ln Y_{i,t-1}$ estimate these parameters consistently? We might assume that the cross-country variation in the constant terms is purely random, and the time trends, θ_i, are driven by purely exogenous factors. But the differences across countries of the convergence parameters, β_i, are likely at least to be correlated with the growth in incomes in those countries, which will induce a correlation between the lagged income growth and the term $(\beta_i - \beta)$ embedded in w_{it}. If $(\beta_i - \beta)$ is random noise that is uncorrelated with $\Delta \ln Y_{i,t-1}$, then $(\beta_i - \beta)\ \Delta \ln Y_{i,t-1}$ will be also.

Measurement Error: Ashenfelter and Krueger (1994), Ashenfelter and Zimmerman (1997), and Bonjour et al. (2003) examined applications in which an earnings equation,

$$y_{i,t} = f(Education_{i,t},\ \ldots) + \varepsilon_{i,t},$$

is specified for sibling pairs (twins) $t = 1, 2$ for n families. Education is a variable that is inherently unmeasurable; years of schooling is typically the best **proxy variable** available. Consider, in a very simple model, attempting to estimate the parameters of

$$y_{it} = \beta_1 + \beta_2\, Education_{it} + \varepsilon_{it},$$

by a regression of $Earnings_{it}$ on a constant and $Schooling_{it,}$ with

$$Schooling_{it} = Education_{it} + u_{it},$$

where u_{it} is the measurement error. By a simple substitution, we find

$$y_{it} = \beta_1 + \beta_2\, Schooling_{it} + w_{it},$$

where $w_{it} = \varepsilon_{it} - \beta_2 u_{it}$. *Schooling* is clearly correlated with $w_{it} = (\varepsilon_{it} - \beta_2 u_{it})$. The interpretation is that at least some of the variation in *Schooling* is due to variation in the measurement error, u_{it}. Because schooling is correlated with w_{it}, it is endogenous in the earnings equation, and least squares is not a suitable estimator. As we will show later, in cases such as this one, the mismeasurement of a relevant variable causes a particular form of inconsistency, **attenuation bias**, in the estimator of β_2.

Nonrandom Sampling: In a model of the effect of a training program, an employment program, or the labor supply behavior of a particular segment of the labor force, the sample of observations may have voluntarily selected themselves into the observed sample. The Job Training Partnership Act (JTPA) was a job training program intended to provide employment assistance to disadvantaged youth. Anderson et al. (1991) found that for a sample that they examined, the program appeared to be administered most often to the best qualified applicants. In an earnings equation estimated for such a nonrandom sample, the implication is that the disturbances are not truly random. For the application just described, for example, on average, the disturbances are unusually high compared to the

[1] See, for example, Lee, Pesaran, and Smith (1997).

full population. Merely unusually high would not be a problem save for the general finding that the explanation for the nonrandomness is found at least in part in the variables that appear elsewhere in the model. This nonrandomness of the sample translates to a form of omitted variable bias known as **sample selection bias**.

Attrition: We can observe two closely related important cases of nonrandom sampling. In panel data studies of firm performance, the firms still in the sample at the end of the observation period are likely to be a subset of those present at the beginning—those firms that perform badly, "fail," or drop out of the sample. Those that remain are unusual in the same fashion as the previous sample of JTPA participants. In these cases, least squares regression of the performance variable on the covariates (whatever they are) suffers from a form of selection bias known as **survivorship bias**. In this case, the distribution of outcomes, firm performances for the survivors is systematically higher than that for the population of firms as a whole. This produces a phenomenon known as **truncation bias**. In clinical trials and other statistical analyses of health interventions, subjects often drop out of the study for reasons related to the intervention itself—for a quality of life intervention such as a drug treatment for cancer, subjects may leave because they recover and feel uninterested in returning for the exit interview, or they may pass away or become incapacitated and be unable to return. In either case, the statistical analysis is subject to **attrition bias**. The same phenomenon may impact the analysis of panel data in health econometrics studies. For example, Contoyannis, Jones, and Rice (2004) examined self-assessed health outcomes in a long panel data set extracted from the British Household Panel Survey. In each year of the study, a significant number of the observations were absent from the next year's data set, with the result that the sample was winnowed significantly from the beginning to the end of the study.

In all the cases listed in Example 8.1, the term *bias* refers to the result that least squares (or other conventional modifications of least squares) is an inconsistent (persistently biased) estimator of the coefficients of the model of interest. Though the source of the result differs considerably from setting to setting, all ultimately trace back to endogeneity of some or all of the right-hand-side variables and this, in turn, translates to correlation between the regressors and the disturbances. These can be broadly viewed in terms of some specific effects:

- Omitted variables, either observed or unobserved,
- Feedback effects,
- Dynamic effects,
- Endogenous sample design, and so on.

There are three general solutions to the problem of constructing a consistent estimator. In some cases, a more detailed, **structural specification** of the model can be developed. These usually involve specifying additional equations that explain the correlation between $\mathbf{x}_i$ and ε_i in a way that enables estimation of the full set of parameters of interest. We will develop a few of these models in later chapters, including, for example, Chapter 19, where we consider Heckman's (1979) model of sample selection. The second approach, which is becoming increasingly common in contemporary research, is the method of **instrumental variables**. The method of instrumental variables is developed around the following estimation strategy: Suppose that in the model of (8-1), the K variables $\mathbf{x}_i$ may be correlated with ε_i. Suppose as well that there exists a set of L variables $\mathbf{z}_i$, such that $\mathbf{z}_i$ is correlated with $\mathbf{x}_i$, but not with ε_i. We cannot estimate $\boldsymbol{\beta}$ consistently by using the familiar least squares estimator. But the assumed lack of correlation between $\mathbf{z}_i$ and ε_i implies a set of relationships that may allow us construct a consistent estimator

of $\boldsymbol{\beta}$ by using the assumed relationships among $\mathbf{z}_i$, $\mathbf{x}_i$, and ε_i. A third method that builds off the second augments the equation with a constructed exogenous variable (or set of variables), C_i, such that in the presence of the **control function**, $C, \mathbf{x}_{i2}$ is not correlated with ε_i. The best known approach to the sample selection problem turns out to be a control function estimator. The method of two-stage least squares can be construed as another.

This chapter will develop the method of instrumental variables as an extension of the models and estimators that have been considered in Chapters 2–7. Section 8.2 will formalize the model in a way that provides an estimation framework. The method of **instrumental variables (IV)** estimation and **two-stage least squares (2SLS)** is developed in detail in Section 8.3. Two tests of the model specification are considered in Section 8.4. A particular application of the estimation with measurement error is developed in detail in Section 8.5. Section 8.6 will consider nonlinear models and begin the development of the generalized method of moments (GMM) estimator. The IV estimator is a powerful tool that underlies a great deal of contemporary empirical research. A shortcoming, the problem of weak instruments, is considered in Section 8.7. Finally, some observations about instrumental variables and the search for causal effects are presented in Section 8.8.

This chapter will develop the fundamental results for IV estimation. The use of instrumental variables will appear in many applications in the chapters to follow, including multiple equations models in Chapter 10, the panel data methods in Chapter 11, and in the development of the generalized method of moments in Chapter 13.

8.2 ASSUMPTIONS OF THE EXTENDED MODEL

The assumptions of the linear regression model, laid out in Chapters 2 and 4, are:

A.1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$.
A.2. Full rank: The $n \times K$ sample data matrix, $\mathbf{X}$, has full column rank.
A.3. Exogeneity of the independent variables: $E[\varepsilon_i | x_{j1}, x_{j2}, \ldots, x_{jk}] = 0, i, j = 1, \ldots, n$. There is no correlation between the disturbances and the independent variables.
A.4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i, has the same finite variance, σ^2, and is uncorrelated with every other disturbance, ε_j, conditioned on $\mathbf{X}$.
A.5. Stochastic or nonstochastic data: $(x_{i1}, x_{i2}, \ldots, x_{iK}), i = 1, \ldots, n$.
A.6. Normal distribution: The disturbances are normally distributed.

We will maintain the important result that plim $(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}_{\mathbf{xx}}$. The basic assumptions of the regression model have changed, however. First, A.3 (no correlation between $\mathbf{x}$ and ε) is, under our new assumptions,

$$\textbf{A.I3.} \quad E[\varepsilon_i | \mathbf{x}_i] = \eta.$$

We interpret Assumption A.I3 to mean that the regressors now provide information about the expectations of the disturbances. The important implication of A.I3 is that the disturbances and the regressors are now correlated. Assumption A.I3 implies that

$$E[\mathbf{x}_i \varepsilon_i] = \boldsymbol{\gamma} \tag{8-2}$$

for some nonzero $\boldsymbol{\gamma}$. If the data are well behaved, then we can apply Theorem D.5 (Khinchine's theorem) to assert that,

$$\text{plim } (1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\gamma}. \tag{8-3}$$

Notice that the original model results if $\eta = 0$. The implication of (8-3) is that the regressors, $\mathbf{X}$, are no longer exogenous. Assumptions A.4–A.6 will be secondary considerations in the discussion of this chapter. We will develop some essential results with A.4 in place, then turn to robust inference procedures that do not rely on it. As before, we will characterize the essential results based on random sampling from the joint distribution of y and $\mathbf{x}$ (and $\mathbf{z}$). Assumption A.6 is no longer relevant—all results from here forward will be based on asymptotic distributions.

We now assume that there is an additional set of variables, $\mathbf{z} = (z_1, \ldots, z_L)$, that have two essential properties:

1. **Relevance**: They are correlated with the independent variables, $\mathbf{X}$.
2. **Exogeneity**: They are uncorrelated with the disturbance.

We will formalize these notions as we proceed. In the context of our model, variables that have these two properties are instrumental variables. We assume the following:

A.I7. $[\mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$, $i = 1, \ldots, n$, are an i.i.d. sequence of random variables.
A.I8a. $E[x_{ik}^2] = \mathbf{Q}_{\mathbf{xx},kk} < \infty$, a finite constant, $k = 1, \ldots, K$.
A.I8b. $E[z_{il}^2] = \mathbf{Q}_{\mathbf{zz},ll} < \infty$, a finite constant, $l = 1, \ldots, L$.
A.I8c. $E[z_{il}x_{ik}] = \mathbf{Q}_{\mathbf{zx},lk} < \infty$, a finite constant, $l = 1, \ldots, L$, $k = 1, \ldots, K$.
A.I9. $E[\varepsilon_i | \mathbf{z}_i] = 0$.

In later work in time-series models, it will be important to relax assumption A.I7. Finite means of $\mathbf{z}_l$ follows from A.I8b. Using the same analysis as in Section 4.4, we have

$$\text{plim } (1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{\mathbf{zz}}, \text{ a finite, positive definite matrix (\textit{well-behaved data})},$$
$$\text{plim } (1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{zx}}, \text{ a finite, } L \times K \text{ matrix with rank } K \text{ (\textit{relevance})},$$
$$\text{plim } (1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \text{ (\textit{exogeneity})}.$$

In our statement of the regression model, we have assumed thus far the special case of $\eta = 0$; $\boldsymbol{\gamma} = \mathbf{0}$ follows.

For the present, we will assume that $L = K$—there are the same number of instrumental variables as there are right-hand-side variables in the equation. Recall in the introduction and in Example 8.1, we partitioned $\mathbf{x}$ into $\mathbf{x}_1$, a set of K_1 exogenous variables, and $\mathbf{x}_2$, a set of K_2 endogenous variables, on the right-hand side of (8-1). In nearly all cases in practice, the problem of endogeneity is attributable to one or a small number of variables in $\mathbf{x}$. In the Krueger and Dale (1999) study of endogenous treatment effects in Example 8.1, we have a single endogenous variable in the equation, the treatment dummy variable, T. The implication for our formulation here is that in such a case, the K_1 variables $\mathbf{x}_1$ will be K_1 of the variables in $\mathbf{Z}$ and the K_2 remaining variables will be other exogenous variables that are not the same as $\mathbf{x}_2$. The usual interpretation will be that these K_2 variables, $\mathbf{z}_2$, are the instruments for $\mathbf{x}_2$ while the $\mathbf{x}_1$ variables are instruments for themselves. To continue the example, the matrix $\mathbf{Z}$ for the endogenous treatment effects model would contain the K_1 columns of $\mathbf{X}$ and an additional instrumental variable, $\mathbf{z}$, for the treatment dummy variable. In the simultaneous equations model of supply and demand, the endogenous right-hand-side variable is $x_2 = \textit{price}$ while the exogenous variables are (1, *Income*). One might suspect (correctly), that in this model, a set of instrumental variables would be $\mathbf{z} = (1, \textit{Income}, \textit{InputPrice})$. In terms of the underlying relationships among the variables, this intuitive understanding will provide a reliable

guide. For reasons that will be clear shortly, however, it is necessary statistically to treat $\mathbf{Z}$ as the instruments for $\mathbf{X}$ in its entirety.

There is a second subtle point about the use of instrumental variables that will likewise be more evident below. The relevance condition must actually be a statement of conditional correlation. Consider, once again, the treatment effects example, and suppose that z is the instrumental variable in question for the treatment dummy variable T. The relevance condition as stated implies that the correlation between z and $(\mathbf{x},T)$ is nonzero. Formally, what will be required is that the *conditional* correlation of z with $T|\mathbf{x}$ be nonzero. One way to view this is in terms of a projection; the instrumental variable z is relevant if the coefficient on z in the projection of T on $(\mathbf{x}, z)$ is nonzero. Intuitively, z must provide information about the movement of T that is not provided by the $\mathbf{x}$ variables that are already in the model.

8.3 INSTRUMENTAL VARIABLES ESTIMATION

For the general model of Section 8.2, we lose most of the useful results we had for least squares. We will consider the implications for least squares and then construct an alternative estimator for $\boldsymbol{\beta}$ in this extended model.

8.3.1 LEAST SQUARES

The least squares estimator, $\mathbf{b}$, is no longer unbiased,

$$E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\eta \neq \boldsymbol{\beta},$$

so the Gauss–Markov theorem no longer holds. The estimator is also inconsistent,

$$\operatorname{plim}\mathbf{b} = \boldsymbol{\beta} + \operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\operatorname{plim}\left(\frac{\mathbf{X}'\varepsilon}{n}\right) = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma} \neq \boldsymbol{\beta}. \tag{8-4}$$

(The asymptotic distribution is considered in the exercises.) The inconsistency of least squares is not confined to the coefficients on the endogenous variables. To see this, apply (8-4) to the treatment effects example discussed earlier. In that case, all but the last variable in $\mathbf{X}$ are uncorrelated with $\boldsymbol{\varepsilon}$. This means that

$$\operatorname{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \gamma_K \end{pmatrix} = \gamma_K \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

It follows that for this special case, the result in (8-4) is

$$\operatorname{plim}\mathbf{b} = \boldsymbol{\beta} + \gamma_K \times \text{the last column of } \mathbf{Q}_{\mathbf{XX}}^{-1}.$$

There is no reason to expect that any of the elements of the last column of $\mathbf{Q}_{\mathbf{XX}}^{-1}$ will equal zero. The implication is that even though only one of the variables in $\mathbf{X}$ is correlated with $\boldsymbol{\varepsilon}$, all of the elements of $\mathbf{b}$ are inconsistent, not just the estimator of the coefficient

on the endogenous variable. This effect is called **smearing**; the inconsistency due to the endogeneity of the one variable is smeared across all of the least squares estimators.

8.3.2 THE INSTRUMENTAL VARIABLES ESTIMATOR

Because $E[\mathbf{z}_i\varepsilon_i] = 0$ and all terms have finite variances, it follows that $\operatorname{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = 0$. Therefore,

$$\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \left[\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta} + \operatorname{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = \left[\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta}. \quad \textbf{(8-5)}$$

We have assumed that $\mathbf{Z}$ has the same number of variables as $\mathbf{X}$. For example, suppose in our consumption function that $\mathbf{x}_t = [1, Y_t]$ when $\mathbf{z}_t = [1, Y_{t-1}]$. We have also assumed that the rank of $\mathbf{Z}'\mathbf{X}$ is K, so now $\mathbf{Z}'\mathbf{X}$ is a square matrix. It follows that

$$\left[\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1}\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \boldsymbol{\beta},$$

which leads us to the **instrumental variable estimator**,

$$\mathbf{b}_{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \quad \textbf{(8-6)}$$

For a model with a constant term and a single x and instrumental variable z, we have

$$b_{\mathrm{IV}} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} = \frac{Cov(z, y)}{Cov(z, x)}.$$

We have already proved that $\mathbf{b}_{\mathrm{IV}}$ is consistent. We now turn to the **asymptotic distribution**. We will use the same method as in Section 4.4.3. First,

$$\sqrt{n}(\mathbf{b}_{\mathrm{IV}} - \boldsymbol{\beta}) = \left(\frac{\mathbf{Z}'\boldsymbol{X}}{n}\right)^{-1}\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon},$$

which has the same **limiting distribution** as $\mathbf{Q}_{\mathbf{zx}}^{-1}[(1/\sqrt{n})\mathbf{Z}'\boldsymbol{\varepsilon}]$. Our analysis of $(1/\sqrt{n})\mathbf{Z}'\boldsymbol{\varepsilon}$ can be the same as that of $(1/\sqrt{n})\boldsymbol{X}'\boldsymbol{\varepsilon}$ in Section 4.4.3, so it follows that

$$\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon}\right) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\sigma}^2\boldsymbol{Q}_{zz}],$$

and

$$\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon}\right) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{\mathbf{zx}}^{-1}\mathbf{Q}_{\mathbf{zz}}\mathbf{Q}_{\mathbf{xz}}^{-1}].$$

This step completes the derivation for the next theorem.

THEOREM 8.1 Asymptotic Distribution of the Instrumental Variables Estimator

If Assumptions A.1–A5, A.I7, A.I8a–c, and A.I9 all hold for $[y_i, \mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$, *where* $\mathbf{z}$ *is a valid set of* $L = K$ *instrumental variables, then the asymptotic distribution of the instrumental variables estimator* $\mathbf{b}_{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ *is*

$$\mathbf{b}_{\mathrm{IV}} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}_{\mathbf{zx}}^{-1}\mathbf{Q}_{\mathbf{zz}}\mathbf{Q}_{\mathbf{xz}}^{-1}\right]. \tag{8-7}$$

where $\mathbf{Q}_{\mathbf{zx}} = \operatorname{plim}(\mathbf{Z}'\mathbf{X}/n)$ *and* $\mathbf{Q}_{\mathbf{zz}} = \operatorname{plim}(\mathbf{Z}'\mathbf{Z}/n)$. *If Assumption A4 is dropped, then the asymptotic covariance matrix will be the population counterpart to the robust estimators in (8-8h) or (8-8c), below.*

8.3.3 ESTIMATING THE ASYMPTOTIC COVARIANCE MATRIX

To estimate the **asymptotic covariance matrix**, we will require an estimator of σ^2. The natural estimator is

$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b}_{\mathrm{IV}})^2.$$

The correction for degrees of freedom is unnecessary, as all results here are asymptotic, and $\hat{\sigma}^2$ would not be unbiased in any event. Nonetheless, it is standard practice to make the degrees of freedom correction. Using the same approach as in Section 4.4.2 for the regression model, we find that $\hat{\sigma}^2$ is a **consistent estimator** of σ^2. We will estimate $\text{Asy.Var}[\mathbf{b}_{\mathrm{IV}}]$ with

$$\begin{aligned}\text{Est.Asy.Var}[\mathbf{b}_{\mathrm{IV}}] &= \frac{1}{n}\left(\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1} \\ &= \hat{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}.\end{aligned} \tag{8-8}$$

The estimator in (8-8) is based on Assumption A.4, homoscedasticity and nonautocorrelation. By writing the IV estimator as

$$\mathbf{b}_{IV} = \boldsymbol{\beta} + \left[\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right]^{-1}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i$$

we can use the same logic as in (4-35)–(4-37) and (4-40)–(4-42) to construct estimators of the asymptotic covariance matrix that are robust to heteroscedasticity,

$$\begin{aligned}\text{Est.Asy.Var}[\mathbf{b}_{IV}] &= \left[\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'\hat{\varepsilon}_i^2\right]\left[\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{z}_i\right]^{-1} \\ &= n(\mathbf{Z}'\mathbf{X})^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'\hat{\varepsilon}_i^2\right](\mathbf{X}'\mathbf{Z})^{-1},\end{aligned} \tag{8-8h}$$

and to clustering,

$$\text{Est.Asy.Var}[\mathbf{b}_{IV}] = C(\mathbf{Z}'\mathbf{X})^{-1}\left[\left(\frac{C}{C-1}\right)\frac{1}{C}\sum_{c=1}^{C}\left(\sum_{i=1}^{N_c}\mathbf{z}_{ic}\hat{\varepsilon}_{ic}\right)\left(\sum_{i=1}^{N_c}\mathbf{z}_{ic}\hat{\varepsilon}_{ic}\right)'\right](\mathbf{X}'\mathbf{Z})^{-1}, \tag{8-8c}$$

respectively.

8.3.4 MOTIVATING THE INSTRUMENTAL VARIABLES ESTIMATOR

In obtaining the IV estimator, we relied on the solutions to the equations in (8-5), $\text{plim}(\mathbf{Z}'\mathbf{y}/n) = \text{plim}(\mathbf{Z}'\mathbf{X}/n)\boldsymbol{\beta}$ or $\mathbf{Q}_{\mathbf{Zy}} = \mathbf{Q}_{\mathbf{ZX}}\boldsymbol{\beta}$. The IV estimator is obtained by solving this set of K **moment equations**. Because this is a set of K equations in K unknowns, if $\mathbf{Q}_{\mathbf{ZX}}^{-1}$ exists, then there is an exact solution for $\boldsymbol{\beta}$, given in (8-6). The corresponding moment equations if only $\mathbf{X}$ is used would be

$$\text{plim}(\mathbf{X}'\mathbf{y}/n) = \text{plim}(\mathbf{X}'\mathbf{X}/n)\boldsymbol{\beta} + \text{plim}(\mathbf{X}'\boldsymbol{\varepsilon}/n) = \text{plim}(\mathbf{X}'\mathbf{X}/n)\boldsymbol{\beta} + \boldsymbol{\gamma}$$

or

$$\mathbf{Q}_{\mathbf{Xy}} = \mathbf{Q}_{\mathbf{XX}}\boldsymbol{\beta} + \boldsymbol{\gamma},$$

which is, without further restrictions, K equations in $2K$ unknowns. There are insufficient equations to solve this system for either $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$. The further restrictions that would allow estimation of $\boldsymbol{\beta}$ would be $\boldsymbol{\gamma} = \mathbf{0}$; this is precisely the exogeneity assumption A.3. The implication is that the parameter vector $\boldsymbol{\beta}$ is not **identified** in terms of the moments of $\mathbf{X}$ and $\mathbf{y}$ alone—there does not exist a solution. But it is identified in terms of the moments of $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{y}$, plus the K restrictions imposed by the exogeneity assumption, and the relevance assumption that allows computation of $\mathbf{b}_{\text{IV}}$.

By far the most common application of IV estimation involves a single endogenous variable in a multiple regression model,

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots x_{i\,K}\beta_K + \varepsilon_i,$$

with $\text{Cov}(x_K, \varepsilon) \neq 0$. The instrumental variable estimator, based on instrument z, proceeds from two conditions:

- Relevance: $\text{Cov}(z, x_K | x_1, \ldots, x_{K-1}) \neq 0$,
- Exogeneity: $E(\varepsilon | z) = 0$.

In words, the relevance condition requires that the instrument provide explanatory power of the variation of the endogenous variable beyond that provided by the other exogenous variables already in the model. A theoretical basis for the relevance condition would be a projection of x_K on all of the exogenous variables in the model,

$$x_K = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_{K-1}x_{K-1} + \lambda z + u.$$

In this form, the relevance condition will require $\lambda \neq 0$. This can be verified empirically; in a linear regression of x_K on $(x_1, \ldots, x_{K-1}, z)$, one would expect the least squares estimate of λ to be statistically different from zero. The exogeneity condition is not directly testable. It is entirely theoretical. (The Hausman and Wu tests suggested below are only indirect.)

Consider these results in the context of a simplified model,

$$y = \beta x + \delta T + \varepsilon.$$

In order for least squares consistently to estimate δ (and β), it is assumed that movements in T are exogenous to the model, so that covariation of y and T is explainable by the movement of T and not by the movement of ε. When T and ε are correlated and ε varies through some factor not in the equation, the movement of y will appear to be induced by variation in T when it is actually induced by variation in ε which is transmitted through T. If T is exogenous, that is, not correlated with ε, then movements in ε will not "cause" movements in T (we use the term *cause* very loosely here) and will thus not be mistaken for exogenous variation in T. The exogeneity assumption plays precisely this role. What is needed, then, to identify δ is movement in T that is definitely not induced by movement in ε? Enter the instrumental variable, z. If z is an instrumental variable with $\text{Cov}(z, T) \neq 0$ and $\text{Cov}(z, \varepsilon) = 0$, then movement in z provides the variation that we need. If we can consider doing this exercise experimentally, in order to measure the "causal effect" of movement in T, we would change z and then measure the per unit change in y associated with the change in T, knowing that the change in T was induced only by the change in z, not ε. That is, the estimator of δ is $(\Delta y/\Delta z)/(\Delta T/\Delta z)$.

Example 8.2 Instrumental Variable Analysis

Grootendorst (2007) and Deaton (1997) recount what appears to be the earliest application of the method of instrumental variables:

Although IV theory has been developed primarily by economists, the method originated in epidemiology. IV was used to investigate the route of cholera transmission during the London cholera epidemic of 1853–54. A scientist from that era, John Snow, hypothesized that cholera was waterborne. To test this, he could have tested whether those who drank purer water had lower risk of contracting cholera. In other words, he could have assessed the correlation between water purity (x) and cholera incidence (y). Yet, as Deaton (1997) notes, this would not have been convincing: "The people who drank impure water were also more likely to be poor, and to live in an environment contaminated in many ways, not least by the 'poison miasmas' that were then thought to be the cause of cholera." Snow instead identified an instrument that was strongly correlated with water purity yet uncorrelated with other determinants of cholera incidence, both observed and unobserved. This instrument was the identity of the company supplying households with drinking water. At the time, Londoners received drinking water directly from the Thames River. One company, the Lambeth Water Company, drew water at a point in the Thames above the main sewage discharge; another, the Southwark and Vauxhall Company, took water below the discharge. Hence the instrument z was strongly correlated with water purity x. The instrument was also uncorrelated with the unobserved determinants of cholera incidence (y). According to Snow (1855, pp. 74–75), the households served by the two companies were quite similar; indeed: "the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. . . . The experiment, too, is on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge; one group supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity.

A stylized sketch of Snow's experiment is useful for suggesting how the instrumental variable estimator works. The theory states that

$$\text{Cholera Occurrence} = f(\text{Impure Water, Other Factors}).$$

For simplicity, denote the occurrence of cholera in household i with

$$c_i = \alpha + \delta w_i + \varepsilon_i,$$

where c_i represents the presence of cholera, $w_i = 1$ if the household has (measurably) impure water, 0 if not, and δ is the sought after causal effect of the water impurity on the prevalence of cholera. It would seem that one could simply compute $d = (\bar{c}|w = 1) - (\bar{c}|w = 0)$, which would be the result of a regression of c on w, to assess the effect of impure water on the prevalence of cholera. The flaw in this strategy is that a cholera prone environment, u, affects both the water quality, w, and the other factors, ε. Interpret this to say that both Cov(w, u) and Cov(ε, u) are nonzero and therefore, Cov(w, ε) is nonzero. The endogeneity of w in the equation invalidates the regression estimator of δ. The pernicious effect of the common influence, u, works through the unobserved factors, ε. The implication is that $E[c|w] \neq \alpha + \delta w$ because $E[\varepsilon|w] \neq 0$. Rather,

$$E[c|w = 1] = \alpha + \delta + E[\varepsilon|w = 1]$$
$$E[c|w = 0] = \alpha + \cdots + E[\varepsilon|w = 0]$$

so,

$$E[c|w = 1] - E[c|w = 0] = \delta + \{E[\varepsilon|w = 1] - E[\varepsilon|w = 0]\}.$$

It follows that comparing the cholera rates of households with bad water to those with good water, $P[c|w = 1] - P[c|w = 0]$, does not reveal only the impact of the bad water on the prevalence of cholera. It partly reveals the impact of bad water on some other factor in ε that, in turn, impacts the cholera prevalence. Snow's IV approach based on the water supplying company works as follows: Define

$$\text{I} = 1 \text{ if water is supplied by Lambeth,}$$
$$0 \text{ if Southwark and Vauxhall.}$$

To establish the *relevance* of this instrument, Snow argued that

$$E[w|I = 1] \neq E[w|I = 0].$$

Snow's theory was that water supply was the culprit, and Lambeth supplied purer water than Southwark. This can be verified observationally. The instrument is *exogenous* if

$$E[\varepsilon|I = 1] = E[\varepsilon|I = 0].$$

This is the theory of the instrument. Water is supplied randomly to houses. Homeowners do not even know who supplies their water. The assumption is not that the unobserved factor, ε, is unaffected by the water quality. It is that the other factors, not the water quality, are present in equal measure in households supplied by the two different water suppliers. This is Snow's argument that the households supplied by the two water companies are otherwise similar. The assignment is random. To use the instrument, we note $E[c|I] = \delta E[w|I] + E[\varepsilon|I]$, so

$$E[c|I = 1] = \alpha + \delta E[w|I = 1] + E[\varepsilon|I = 1],$$
$$E[c|I = 0] = \alpha + \delta E[w|I = 0] + E[\varepsilon|I = 0].$$

This produces an estimating equation,

$$E[c|I = 1] - E[c|I = 0] = \delta\{E[w|I = 1] - E[w|I = 0]\}$$
$$+ \{E[\varepsilon|I = 1] - E[\varepsilon|I = 0]\}.$$

The second term in braces is zero if I is exogenous, which was assumed. The IV estimator is then

$$\hat{\delta} = \frac{E[c|I = 1] - E[c|I = 0]}{E[w|I = 1] - E[w|I = 0]}.$$

Note that the nonzero denominator results from the relevance condition. We can see that δ is analogous to Cov(c,l)/Cov(w,l), which is (8-6).

To operationalize the estimator, we will use

$$P(c|l=1) = \hat{E}(c|l=1) = \bar{c}_1 = \text{proportion of households supplied by Lambeth that have cholera,}$$
$$P(w|l=1) = \hat{E}(w|l=1) = \bar{w}_1 = \text{proportion of households supplied by Lambeth that have bad water,}$$
$$P(c|l=0) = \hat{E}(c|l=0) = \bar{c}_0 = \text{proportion of households supplied by Vauxhall that have cholera,}$$
$$P(w|l=0) = \hat{E}(w|l=0) = \bar{w}_0 = \text{proportion of households supplied by Vauxhall that have bad water.}$$

To complete this development of Snow's experiment, we can show that the estimator $\hat{\delta}$ is an application of (8-6). Define three dummy variables, $c_i = 1$ if household i suffers from cholera and 0 if not, $w_i = 1$ if household i receives impure water and 0 if not, and $l_i = 1$ if household i receives its water from Lambeth and 0 if from Vauxhall; let **c**, **w**, and **l** denote the column vectors of n observations on the three variables; and let **i** denote a column of ones. For the model $c_i = \alpha + \delta w_i + \varepsilon_i$, we have $\mathbf{Z} = [\mathbf{i}, \mathbf{l}]$, $\mathbf{X} = [\mathbf{i}, \mathbf{w}]$, and $\mathbf{y} = \mathbf{c}$. The estimator is

$$\begin{pmatrix} a \\ d \end{pmatrix} = [\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'\mathbf{y} =$$
$$\begin{bmatrix} \mathbf{i}'\mathbf{i} & \mathbf{i}'\mathbf{w} \\ \mathbf{l}'\mathbf{i} & \mathbf{l}'\mathbf{w} \end{bmatrix}^{-1}\begin{pmatrix} \mathbf{i}'\mathbf{c} \\ \mathbf{l}'\mathbf{c} \end{pmatrix} = \begin{bmatrix} n & n\bar{w} \\ n_1 & n_1\bar{w}_1 \end{bmatrix}^{-1}\begin{pmatrix} n\bar{c} \\ n_1\bar{c}_1 \end{pmatrix} = \frac{1}{nn_1(\bar{w}_1 - \bar{w})}\begin{bmatrix} n_1\bar{w}_1 & -n\bar{w} \\ -n_1 & n \end{bmatrix}\begin{pmatrix} n\bar{c} \\ n_1\bar{c}_1 \end{pmatrix}.$$

Collecting terms, $d = (\bar{c}_1 - \bar{c})/(\bar{w}_1 - \bar{w})$. Because $n = n_0 + n_1$, $\bar{c}_1 = (n_0\bar{c}_1 + n_1\bar{c}_1)/n$ and $\bar{c} = (n_0\bar{c}_0 + n_1\bar{c}_1)/n$, so $\bar{c}_1 - \bar{c} = (n_0/n)(\bar{c}_1 - \bar{c}_0)$. Likewise, $\bar{w}_1 - \bar{w} = (n_0/n)(\bar{w}_1 - \bar{w}_0)$ so $d = (\bar{c}_1 - \bar{c}_0)/(\bar{w}_1 - \bar{w}_0) = \hat{\delta}$. This estimator based on the difference in means is the Wald (1940) estimator.

Example 8.3 Streams as Instruments

In Hoxby (2000), the author was interested in the effect of the amount of school "choice" in a school "market" on educational achievement in the market. The equations of interest were of the form

$$\frac{A_{ikm}}{\ln E_{km}} = \beta_1 C_m + \mathbf{x}'_{ikm}\boldsymbol{\beta}_2 + \bar{\mathbf{x}}'_{.km}\boldsymbol{\beta}_3 + \bar{\mathbf{x}}'_{..m}\boldsymbol{\beta}_4 + \varepsilon_{ikm} + \varepsilon_{km} + \varepsilon_m,$$

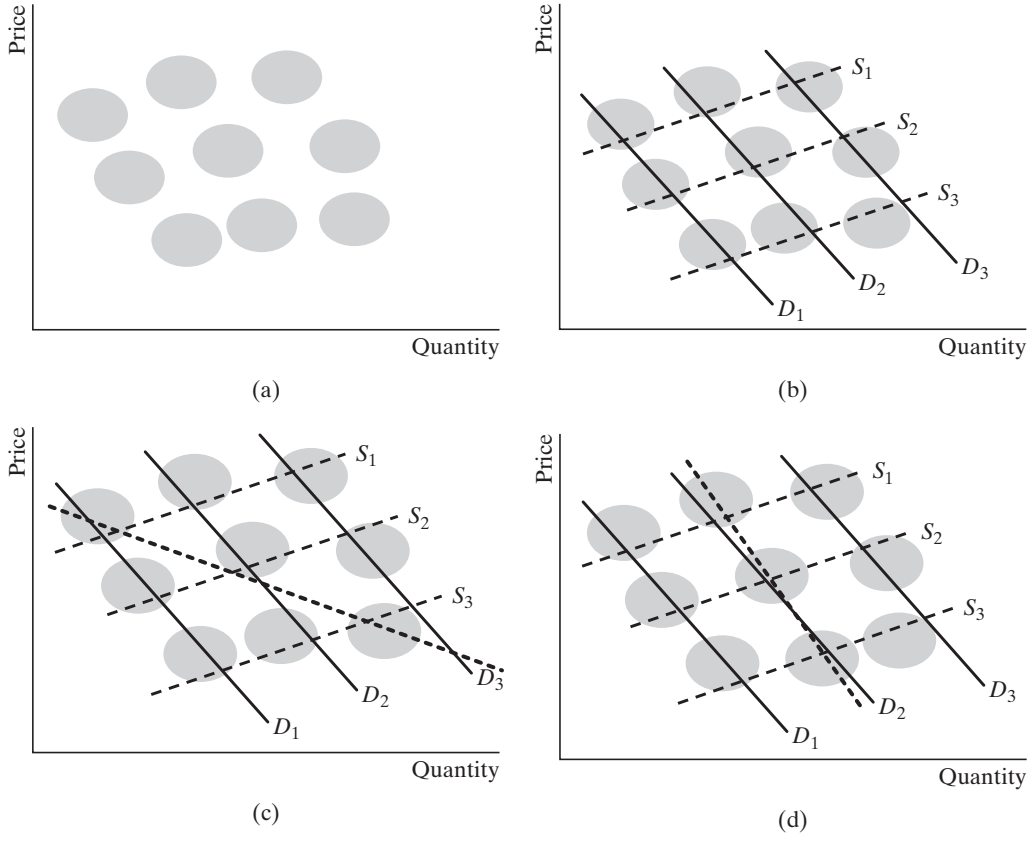
where "*ikm*" denotes household i in district k in market m, A_{ikm} is a measure of achievement, and E_{ikm} is per capita expenditures. The equation contains individual-level data, district means, and market means. The exogenous variables are intended to capture the different sources of heterogeneity at all three levels of aggregation. (The compound disturbance, which we will revisit when we examine panel data specifications in Chapter 10, is intended to allow for random effects at all three levels as well.) Reasoning that the amount of choice available to students, C_m, would be endogenous in this equation, the author sought a valid instrumental variable that would "explain" (be correlated with) C_m but uncorrelated with the disturbances in the equation. In the U.S. market, to a large degree, school district boundaries were set in the late 18th through the 19th centuries and handed down to present-day administrators by historical precedent. In the formative years, the author noted, district boundaries were set in response to natural travel barriers, such as rivers and streams. It follows, as she notes, that "the number of districts in a

given land area is an increasing function of the number of natural barriers"; hence, the number of streams in the physical market area provides the needed instrumental variable.[2] This study is an example of a "natural experiment," as described in Angrist and Pischke (2009).

Example 8.4 Instrumental Variable in Regression

The role of an instrumental variable in identifying parameters in regression models was developed in Working's (1926) classic application, adapted here for our market equilibrium example in Example 8.1. Figure 8.1a displays the observed data for the market equilibria in a market in which there are random disturbances (ε_S, ε_D) and variation in demanders' incomes and input prices faced by suppliers. The market equilibria in Figure 8.1a are scattered about as the aggregates of all these effects. Figure 8.1b suggests the underlying conditions of supply and demand that give rise to these equilibria. Different outcomes in the supply equation corresponding to different values of the input price and different outcomes on the demand side corresponding to different income values produce nine regimes, punctuated

FIGURE 8.1 Identifying a Demand Curve with an Instrumental Variable.



[2]The controversial topic of the study and the unconventional choice of instruments caught the attention of the popular press, for example, http://www.wsj.com/articles/SB113011672134577225 and http://www.thecrimson.com/article/2005/7/8/star-ec-prof-caught-in-academic/, and academic observers including Rothstein (2004).

by the random variation induced by the disturbances. Given the ambiguous mass of points, linear regression of quantity on price (and income) is likely to produce a result such as that shown by the heavy dotted line in Figure 8.1c. The slope of this regression barely resembles the slope of the demand equations. Faced with this prospect, how is it possible to learn about the slope of the demand curve? The experiment needed, shown in Figure 8.1d, would involve two elements: (1) Hold *Income* constant, so we can focus on the demand curve in a particular demand setting. That is the function of multiple regression—Income is included as a conditioning variable in the equation. (2) Now that we have focused on a particular set of demand outcomes (e.g., D2), move the supply curve so that the equilibria now trace out the demand function. That is the function of the changing *InputPrice*, which is the instrumental variable that we need for identification of the demand function(s) for this experiment.

8.4 TWO-STAGE LEAST SQUARES, CONTROL FUNCTIONS, AND LIMITED INFORMATION MAXIMUM LIKELIHOOD

Thus far, we have assumed that the number of instrumental variables in $\mathbf{Z}$ is the same as the number of variables (exogenous plus endogenous) in $\mathbf{X}$. In the typical application, there is one instrument for the single endogenous variable in the equation. The model specification may imply additional instruments. Recall the market equilibrium application considered in Examples 8.1 and 8.4. Suppose this were an agricultural market in which there are two exogenous conditions of supply, *InputPrice* and *Rainfall*. Then, the equations of the model are

$$\text{(Demand)} \quad Quantity_D = \alpha_0 + \alpha_1\, Price + \alpha_2\, Income + \varepsilon_D,$$

$$\text{(Supply)} \quad Quantity_S = \beta_0 + \beta_1\, Price + \beta_2\, Input\ Price + \beta_3\, Rain\ fall + \varepsilon_S,$$

$$\text{(Equilibrium)} \quad Quantity_D = Quantity_S.$$

Given the approach taken in Example 8.4, it would appear that the researcher could simply choose either of the two exogenous variables (instruments) in the supply equation for purpose of identifying the demand equation. Intuition should suggest that simply choosing a subset of the available instrumental variables would waste sample information—it seems inevitable that it will be preferable to use the full matrix $\mathbf{Z}$, even when $L > K$. (In the example above, $\mathbf{z} = (1, Income, InputPrice, Rainfall.)$ The method of two-stage least squares solves the problem of how to use all the information in the sample when $\mathbf{Z}$ contains more variables than are necessary to construct an instrumental variable estimator. We will also examine two other approaches to estimation. *The results developed here also apply to the case in which there is one endogenous variable and one instrument.*

In the model

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \varepsilon,$$

where x_2 is a single variable, and there is a single instrument, z_1, that is relevant and exogenous, then the parameters of the model, $(\boldsymbol{\beta}, \lambda)$, can be estimated using the moments of $(y, \mathbf{x}_1, x_2, z_1)$. The IV estimator in (8-6) shows the one function of the moments that can be used for the estimation. In this case, $(\boldsymbol{\beta}, \lambda)$ are said to be *exactly identified*. There are exactly enough moments for estimation of the parameters. If there were a second exogenous and relevant instrument, say z_2, then we could use z_2 instead of z_1 in (8-6) and obtain a second, different estimator. In this case, the parameters are **overidentified**

in terms of the moments of $(y, \mathbf{x}_1, x_2, z_1, z_2)$. This does not mean that there is now simply a second estimator. If z_1 and z_2 are both exogenous and relevant, then any linear combination of them, $z_* = a_1 z_1 + a_2 z_2$, would also be a valid instrument. More than one IV estimator means an infinite number of possible estimators. Overidentification is qualitatively different from exact identification. The methods examined in this section are usable for overidentified models.

8.4.1 TWO-STAGE LEAST SQUARES

If $\mathbf{Z}$ contains more variables than $\mathbf{X}$, then $\mathbf{Z}'\mathbf{X}$ will be $L \times K$ with rank $K < L$ and will thus not have an inverse—(8-6) is not useable. The crucial result for estimation is plim $(\mathbf{Z}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$. That is, every column of $\mathbf{Z}$ is asymptotically uncorrelated with $\boldsymbol{\varepsilon}$. That also means that every linear combination of the columns of $\mathbf{Z}$ is also uncorrelated with $\boldsymbol{\varepsilon}$, which suggests that one approach would be to choose K linear combinations of the columns of $\mathbf{Z}$. Which to choose? One obvious possibility is simply to choose K variables among the L in $\mathbf{Z}$. Discarding the information contained in the *extra* $L - K$ columns will turn out to be inefficient. A better choice that uses all of the instruments is the projection of the columns of $\mathbf{X}$ in the column space of $\mathbf{Z}$,

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{Z}\mathbf{F}. \tag{8-9}$$

The instruments in this case are linear combinations of the variables (columns) in $\mathbf{Z}$. With this choice of instrumental variables, we have

$$\begin{aligned}\mathbf{b}_{\mathrm{IV}} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.\end{aligned} \tag{8-10}$$

The estimator of the asymptotic covariance matrix will be $\hat{\sigma}^2$ times the bracketed matrix in (8-10). The proofs of consistency and asymptotic normality for this estimator are exactly the same as before, because our proof was generic for any valid set of instruments, and $\hat{\mathbf{X}}$ qualifies.

There are two reasons for using this estimator—one practical, one theoretical. If any column of $\mathbf{X}$ also appears in $\mathbf{Z}$, then that column of $\mathbf{X}$ is reproduced exactly in $\hat{\mathbf{X}}$. This result is important and useful. Consider what is probably the typical application in which the regression contains K variables, only one of which, say, the kth, is correlated with the disturbances. We have one or more instrumental variables in hand, as well as the other $K - 1$ variables that certainly qualify as instrumental variables in their own right. Then what we would use is $\mathbf{Z} = [\mathbf{X}_{(k)}, \mathbf{z}_1, \mathbf{z}_2, \ldots]$, where we indicate omission of the kth variable by (k) in the subscript. Another useful interpretation of $\hat{\mathbf{X}}$ is that each column is the set of fitted values when the corresponding column of $\mathbf{X}$ is regressed on all the columns of $\mathbf{Z}$. The coefficients for $\mathbf{x}_k$ are in the kth column of $\mathbf{F}$ in (8-9). It also makes clear why each $\mathbf{x}_k$ that appears in $\mathbf{Z}$ is perfectly replicated. Every $\mathbf{x}_k$ provides a perfect predictor for itself, without any help from the remaining variables in $\mathbf{Z}$. In the example, then, every column of $\mathbf{X}$ except the one that is omitted from $\mathbf{X}_{(k)}$ is replicated exactly, whereas the one that *is* omitted is replaced in $\hat{\mathbf{X}}$ by the predicted values in the regression of this variable on all the $\mathbf{z}$'s including the other $\mathbf{x}$ variables.

Of all the different linear combinations of $\mathbf{Z}$ that we might choose, $\hat{\mathbf{X}}$ is the most efficient in the sense that the asymptotic covariance matrix of an IV estimator based on a linear combination $\mathbf{Z}\mathbf{F}$ is smaller when $\mathbf{F} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ than with any other $\mathbf{F}$ that

uses all L columns of $\mathbf{Z}$; *a fortiori*, this result eliminates linear combinations obtained by dropping any columns of $\mathbf{Z}$.[3]

We close this section with some practical considerations in the use of the instrumental variables estimator. By just multiplying out the matrices in the expression, you can show that

$$\begin{aligned}\mathbf{b}_{\mathbf{IV}} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= (\mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}\end{aligned} \tag{8-11}$$

because $\mathbf{I} - \mathbf{M}_{\mathbf{Z}}$ is idempotent. Thus, when (*and only when*) $\hat{\mathbf{X}}$ is the set of instruments, the IV estimator is computed by least squares regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$. This conclusion suggests that $\mathbf{b}_{\mathrm{IV}}$ can be computed in two steps, first by computing $\hat{\mathbf{X}}$, then by the least squares regression. For this reason, this is called the two-stage least squares (2SLS) estimator. One should be careful of this approach, however, in the computation of the asymptotic covariance matrix; $\hat{\sigma}^2$ should not be based on $\hat{\mathbf{X}}$. The estimator

$$s_{\mathrm{IV}}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mathbf{x}}_i'\mathbf{b}_{\mathrm{IV}})^2}{n}$$

is inconsistent for σ^2, with or without a correction for degrees of freedom. (The appropriate calculation is built into modern software.)

An obvious question is where one is likely to find a suitable set of instrumental variables. The recent literature on natural experiments focuses on local policy changes such as the Mariel Boatlift (Example 6.9) or global policy changes that apply to the entire economy such as mandatory schooling (Example 6.13), or natural outcomes such as occurrences of streams (Example 8.3) or birthdays [Angrist and Krueger (1992)]. In many time-series settings, lagged values of the variables in the model provide natural candidates. In other cases, the answer is less than obvious and sometimes involves some creativity as in Examples 8.9 and 8.11. Unfortunately, there usually is not much choice in the selection of instrumental variables. The choice of $\mathbf{Z}$ is often ad hoc.

Example 8.5 Instrumental Variable Estimation of a Labor Supply Equation

Cornwell and Rupert (1988) analyzed the returns to schooling in a panel data set of 595 observations on heads of households. The sample data are drawn from years 1976 to 1982 from the "Non-Survey of Economic Opportunity" from the Panel Study of Income Dynamics. The estimating equation is

$$\begin{aligned}\ln Wage_{it} = {} & \alpha_1 + \alpha_2 Exp_{it} + \alpha_3 Exp_{it}^2 + \alpha_4 Wks_{it} + \alpha_5 Occ_{it} + \alpha_6 Ind_{it} + \alpha_7 South_{it} + \\ & \alpha_8 SMSA_{it} + \alpha_9 MS_{it} + \alpha_{10} Union_{it} + \alpha_{11} Ed_i + \alpha_{12} Fem_i + \alpha_{13} Blk_i + \varepsilon_{it}.\end{aligned}$$

(The variables are described in Example 4.6.) The main interest of the study, beyond comparing various estimation methods, is α_{11}, the return to education. The equation suggested is a **reduced form equation**; it contains all the variables in the model but does not specify the underlying structural relationships. In contrast, the three-equation model specified at the beginning of this section is a **structural equation system**. The reduced form for this model would consist of separate regressions of *Price* and *Quantity* on (1, *Income, InputPrice, Rainfall*). We will return to the idea of reduced forms in the setting of simultaneous equations models in Chapter 10. For the present, the implication for the suggested model is that this

[3]See Brundy and Jorgenson (1971) and Wooldridge (2010, pp. 103–104).

market equilibrium equation represents the outcome of the interplay of supply and demand in a labor market. Arguably, the supply side of this market might consist of a household labor supply equation such as

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it}.$$

(One might prefer a different set of right-hand-side variables in this structural equation.) Structural equations are more difficult to specify than reduced forms. If the number of weeks worked and the accepted wage offer are determined jointly, then $lnWage_{it}$ and u_{it} in this equation are correlated. We consider two instrumental variable estimators based on

$$\mathbf{z}_1 = [1, Ind_{it}, Ed_i, Union_{it}, Fem_i]$$

and

$$\mathbf{z}_2 = [1, Ind_{it}, Ed_i, Union_{it}, Fem_i, SMSA_{it}].$$

We begin by examining the relevance condition. In the regression of ln *Wage* on $\mathbf{z}_1$, the *t* ratio on *Ind* is +6.02. In the regression of ln *Wage* on $\mathbf{z}_2$, the Wald statistic for the joint test that the coefficients on *Ind* and *SMSA* are both zero is +240.932. In both cases, the hypothesis is rejected, and we conclude that the instruments are, indeed, relevant. Table 8.1 presents the three sets of estimates. The least squares estimates are computed using the standard results in Chapters 3 and 4. One noteworthy result is the very small coefficient on the log wage variable. The second set of results is the instrumental variable estimates. Note that, here, the single instrument is IND_{it}. As might be expected, the log wage coefficient becomes considerably larger. The other coefficients are, perhaps, contradictory. One might have different expectations about all three coefficients. The third set of coefficients are the two-stage least squares estimates based on the larger set of instrumental variables. In this case, *SMSA* and *Ind* are both used as instrumental variables.

8.4.2 A CONTROL FUNCTION APPROACH

A control function is a constructed variable that is added to a model to "control for" the correlation between an endogenous variable and the unobservable elements. In the presence of the control function, the endogenous variable becomes exogenous. Control functions appear in the estimators for several of the nonlinear models we will consider later in the book. For the linear model we are studying here, the approach provides a

TABLE 8.1 Estimated Labor Supply Equation

	OLS		*IV with* $\mathbf{Z}_1$		*IV with* $\mathbf{Z}_2$		*Control Function*	
Variable	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
Constant	44.7665	1.2153	18.8987	13.0590	30.7044	4.9997	30.7044	4.9100
ln *Wage*	0.7326	0.1972	5.1828	2.2454	3.1518	0.8572	3.1518	0.8418
Education	−0.1532	0.03206	−0.4600	0.1578	−0.3200	0.0661	−0.3200	0.0649
Union	−1.9960	0.1701	−2.3602	0.2567	−2.1940	0.1860	−2.1940	0.1826
Female	−1.3498	0.2642	0.6957	1.0650	−0.2378	0.4679	−0.2378	0.4594
$\hat{u}$							−2.5594	0.8659
$\hat{\sigma}$[a]	1.0301		5.3195		5.1110		5.0187	

[a]Square root of sum of squared residuals/*n*.

useful view of the IV estimator. For the model underlying the preceding example, we have a structural equation,

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it},$$

and the projection (based on $\mathbf{z}_2$),

$$\ln Wage = \gamma_1 + \gamma_2 Ind_{it} + \gamma_3 Ed_i + \gamma_4 Union_{it} + \gamma_5 Fem_i + \gamma_6 SMSA_{it} + u_{it}.$$

The ultimate source of the endogeneity of ln *Wage* in the structural equation for *Wks* is the correlation of the unobservable variables, u and ε. If u were observable—we'll call this observed counterpart $\hat{u}$—then the parameters in the augmented equation,

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \rho\hat{u} + \widetilde{\varepsilon}_{it},$$

could be estimated consistently by least squares. In the presence of $\hat{u}$, ln *Wage* is uncorrelated with the unobservable in this equation—$\hat{u}$ would be the control function that we seek.

To formalize the approach, write the main equation as

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \varepsilon, \tag{8-12}$$

where x_2 is the endogenous variable, so $E[x_2\varepsilon] \neq 0$. The instruments, including $\mathbf{x}_1$, are in $\mathbf{z}$. The projection of x_2 on $\mathbf{z}$ is

$$x_2 = \mathbf{z}'\boldsymbol{\pi} + u, \tag{8-13}$$

with $E[\mathbf{z}u] = 0$. We can also form the projection of ε on u,

$$\varepsilon = \rho u + w, \tag{8-14}$$

where $\rho = \sigma_{uw}/\sigma_w^2$. By construction, u and w are uncorrelated. Finally, insert (8-14) in (8-12) so that

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \rho u + w. \tag{8-15}$$

This is the control function form we had earlier. The loose end, as before, is that in order to proceed, we must observe u. We cannot observe u directly, but we can estimate it using (8-13), the "reduced form" equation for x_2—this is the equation we used to check the relevance of the instrument(s) earlier. We can estimate u as the residual in (8-13), then in the second step, estimate $(\boldsymbol{\beta}, \lambda, \rho)$ by simple least squares. The estimating equation is

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \rho(x_2 - \mathbf{z}'\mathbf{p}) + \widetilde{w}. \tag{8-16}$$

(The constructed disturbance $\widetilde{w}$ contains both w and the estimation error, $\mathbf{z}'\mathbf{p} - \mathbf{z}'\boldsymbol{\pi}$.) The estimated residual is a control function. The control function estimates with estimated standard errors for the model in Example 8.5 are shown in the two rightmost columns in Table 8.1.

This approach would not seem to provide much economy over 2SLS. It still requires two steps (essentially the same two steps). Surprisingly, as you can see in Table 8.1, it is actually identical to 2SLS, at least for the coefficients. (The proof of this result is pursued in the exercises.) The standard errors, however, are different. The general outcome is that control function estimators, because they contain constructed variables, require an

adjustment of the standard errors. (We will examine several applications, notably Heckman's sample selection model in Chapter 19.) Correction of the standard errors associated with control function estimators often requires elaborate post-estimation calculations (though some of them are built-in procedures in modern software).[4] The calculation for 2SLS, however, is surprisingly simple. The difference between the CF standard errors and the appropriate 2SLS standard errors is a simple scaling.[5] The 2SLS difference is the estimator of σ. Because the coefficients on $\mathbf{x}$ are identical to 2SLS, the sum of squared residuals for the CF estimator is smaller than that for the 2SLS estimator. (See Theorem 3.5.) The values are shown in the last row of Table 8.1. It follows that the only correction needed is to rescale the CF covariance matrix by $(\hat{\sigma}_{CF}/\hat{\sigma}_{2SLS})^2 = (5.1110/5.0187)^2$.

8.4.3 LIMITED INFORMATION MAXIMUM LIKELIHOOD[6]

We have considered estimation of the two equation model,

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it},$$
$$\ln Wage_{it} = \gamma_1 + \gamma_2 Ind_{it} + \gamma_3 Ed_i + \gamma_3 Union_{it} + \gamma_4 Fem_i + \gamma_5 SMSA_{it} + u_i,$$

using 2SLS. In generic form, the equations are

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \varepsilon,$$
$$x_2 = \mathbf{z}'\boldsymbol{\gamma} + u.$$

The control function estimator is always identical to 2SLS. They use exactly the same information contained in the moments and the two conditions, relevance and exogeneity. If we add to this system an assumption that (ε, u) have a bivariate normal density, then we can construct another estimator, the limited information maximum likelihood estimator. The estimator is formed from the joint density of the two variables, $(y, x_2|\mathbf{x}_1, \mathbf{z})$. We can write this as $f(\varepsilon, u|\mathbf{x}_1, \mathbf{z})\text{abs}|\mathbf{J}|$ where $\mathbf{J}$ is the Jacobian of the transformation from (ε, u) to (y, x_2),[7] $\text{abs}|\mathbf{J}| = 1, \varepsilon = (y - \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda)$, and $u = (x_2 - \mathbf{z}'\boldsymbol{\gamma})$. The joint normal distribution with correlation ρ can be written $f(\varepsilon, u|\mathbf{x}_1, \mathbf{z}) = f(\varepsilon|u, \mathbf{x}_1, \mathbf{z})f(u|\mathbf{x}_1, \mathbf{z})$, where $u \sim N[0, \sigma_u^2]$ and $\varepsilon|u \sim N[(\rho\sigma_\varepsilon/\sigma_u)u, (1 - \rho^2)\sigma_\varepsilon^2]$. (See Appendix B.9.) For convenience, write the second of these as $N[\tau u, \sigma_w^2]$. Then, the log of the joint density for an observation in the sample will be

$$\ln f_i = \ln f(\varepsilon_i|u_i) + \ln f(u_i) = -(1/2)\ln \sigma_w^2 - (1/2)\{[y_i - \mathbf{x}_1'\boldsymbol{\beta} - x_{2i}\lambda - \tau(x_{2i} - \mathbf{z}_i'\gamma)]/\sigma_w\}^2 \quad \textbf{(8-17)}$$
$$- (1/2)\ln \sigma_u^2 - (1/2)\{[x_{2i} - \mathbf{z}_i'\gamma]/\sigma_u\}^2.$$

[4] See, for example, Wooldridge (2010, Appendix 6A and Chapter 12).

[5] You can see this in the results. The ratio of any two of the IV standard errors is the same as the ratio for the CF standard errors. For example, for *ED* and *Union*, 0.0661/0.1860 = 0.0649/0.1826.

[6] Maximum likelihood estimation is developed in detail in Chapter 14. The term *Limited Information* refers to the focus on only one structural equation in what might be a larger system of equations, such as those considered in Section 10.4.

[7] $\mathbf{J} = \begin{bmatrix} \partial\varepsilon/\partial y & \partial\varepsilon/\partial x_2 \\ \partial u/\partial y & \partial u/\partial x_2 \end{bmatrix} = \begin{bmatrix} 1 & \lambda \\ 0 & 1 \end{bmatrix}$, so $\text{abs}|\mathbf{J}| = 1$.

TABLE 8.2 Estimated Labor Supply Equation

	2SLS		*LIML*	
Variable	***Estimated Parameter***	***Standard Error***[a]	***Estimated Parameter***	***Standard Error***[a]
Constant	30.7044	8.25041	30.6392	5.05118
ln *Wage*	3.15182	1.41058	3.16303	0.87325
Education	−0.31997	0.11453	−0.32074	0.06755
Union	−2.19398	0.30507	−2.19490	0.19697
Female	−0.23784	0.79781	−0.23269	0.46572
σ_w	5.01870[b]		5.01865	0.03339
Constant			5.71303	0.03316
Ind			0.08364	0.01284
Education			0.06560	0.00232
Union			0.05853	0.01448
Female			−0.46930	0.02158
SMSA			0.18225	0.01289
σ_u			0.38408	0.00384
τ			−2.57121	0.90334

[a] Standard errors are clustered at the individual level using (8-8c).
[b] Based on mean squared residual.

The log likelihood to be maximized is $\Sigma_i \ln f_i$.[8] Table 8.2 compares the 2SLS and LIML estimates for the model of Example 8.5 using instruments $\mathbf{z}_2$. The LIML estimates are only slightly different from the 2SLS results, but have substantially smaller standard errors. We can view this as the payoff to the narrower specification, that is, the additional normality assumption (though one should be careful about drawing a conclusion about the efficiency of an estimator based on one set of results). There is yet another approach to estimation. The LIML estimator could be computed in two steps, by computing the estimates of $\boldsymbol{\gamma}$ and σ_u first (by least squares estimation of the second equation), then maximizing the log likelihood over $(\boldsymbol{\beta}, \lambda, \tau, \sigma_w)$. This would be identical to the control function estimator—$(\boldsymbol{\beta}, \lambda, \tau)$ would be estimated by regressing y on $(\mathbf{x}_1, x_2, \hat{u})$, then σ_w would be estimated using the residuals. (Note that this would not estimate σ_ε. That would be done by using only the coefficients on $\mathbf{x}_1$ and x_2 to compute the residuals.)

8.5 ENDOGENOUS DUMMY VARIABLES: ESTIMATING TREATMENT EFFECTS

The leading recent application of models of sample selection and endogeneity is the evaluation of "treatment effects." The central focus is on analysis of the effect of participation in a treatment, C, on an outcome variable, y—examples include job training

[8] The parameter estimates would be computed by minimizing (8-17)using one of the methods described in Appendix E. If the equation is overidentified, the least variance ratio estimator described in Section 10.4.4 is an alternative estimation approach. The two approaches will produce the same results.

programs[9] and education.[10] Imbens and Wooldridge (2009, pp. 22–23) cite a number of labor market applications. Recent, more narrow, examples include Munkin and Trivedi's (2007) analysis of the effect of dental insurance and Jones and Rice's (2011) survey that notes a variety of techniques and applications in health economics. A simple starting point, useful for framing ideas, is the linear regression model with a "treatment dummy variable,"

$$y = \mathbf{x}'\boldsymbol{\beta} + \delta C + \varepsilon.$$

The analysis turns on whether it is possible to estimate the "treatment effect" (here, δ), and under what assumptions is δ a meaningful quantity that we are interested in measuring.

Empirical measurement of treatment effects, such as the impact of going to college or participating in a job training or agricultural extension program, presents a large variety of econometric complications. The natural, ultimate objective of an analysis of a treatment or intervention would be *the effect of treatment on the treated.* For example, what is the effect of a college education on the lifetime income of someone who goes to college? Measuring this effect econometrically encounters at least two compelling complications:

Endogeneity of the treatment: The analyst risks attributing to the treatment causal effects that should be attributed to factors that motivate both the treatment and the outcome. In our example, the individual who goes to college might well have succeeded (more) in life than his or her counterpart who did not go to college even if the individual did not attend college. Example 6.8 suggests another case in which some of the students who take the SAT a second time in hopes of improving their scores also take a test preparation course ($C = 1$),

$$\Delta SAT = (SAT_1 - SAT_0) = \mathbf{x}'\boldsymbol{\beta} + \delta C + \varepsilon.$$

The complication here would be whether it is appropriate to attach a causal interpretation to δ.

Missing counterfactual: The preceding thought experiment is not actually the effect we wish to measure. In order to measure the impact of college attendance on lifetime earnings in a pure sense, we would have to run an individual's lifetime twice, once with college attendance and once without (and with all other conditions as they were). Any individual is observed in only one of the two states, so the pure measurement is impossible. The SAT example has the same nature − the experiment can only be run once, either with $C = 1$ or with $C = 0$.

Accommodating these two problems forms the focal point of this enormous and still growing literature. Rubin's causal model (1974, 1978) provides a useful framework for the analysis. Every individual in a population has a potential outcome, y, and can be exposed to the treatment, C. We will denote by C the binary indicator of whether or not the individual receives the treatment. Thus, the potential outcomes are $y|(C = 1) = y_1$ and $y|(C = 0) = y_0$. We can combine these in

$$y = Cy_1 + (1 - C)y_0 = y_0 + C(y_1 - y_0).$$

[9]See LaLonde (1986), Business Week (2009), Example 8.6.

[10]For example, test scores, Angrist and Lavy (1999), Van der Klaauw (2002).

The *average treatment effect*, averaged across the entire population, is

$$ATE = E[y_1 - y_0].$$

The compelling complication is that the individual will exist in only one of the two states, so it is not possible to estimate ATE without further assumptions. More specifically, what the researcher would prefer to see is the average treatment effect on the treated,

$$ATET = E[y_1 - y_0 | C = 1],$$

and note that the second term is now the missing counterfactual.[11]

One of the major themes of the recent research is to devise robust methods of estimation that do not rely heavily on fragile assumptions such as identification by functional form (e.g., relying on bivariate normality) and identification by exclusion restrictions (e.g., relying on basic instrumental variable estimators). This is a challenging exercise—we will rely heavily on these assumptions in much of the rest of this book. For purposes of the general specification, we will denote by $\mathbf{x}$ the exogenous information that will be brought to bear on this estimation problem. The vector $\mathbf{x}$ may (usually will) be a set of variables that will appear in a regression model, but it is useful to think more generally than that and consider $\mathbf{x}$ rather to be an information set. Certain minimal assumptions are necessary to make any headway at all. The following appear at different points in the analysis.

Conditional independence: Receiving the treatment, C, does not depend on the outcome variable once the effect of $\mathbf{x}$ on the outcome is accounted for. In particular, $(y_0, y_1) | \mathbf{x}$ is independent of C. Completely random assignment to the treatment would certainly imply this. If assignment is completely random, then we could omit the effect of $\mathbf{x}$ in this assumption. A narrower case would be assignment based completely on observable criteria ($\mathbf{x}$), which would be "selection on observables" (as opposed to "selection on unobservables which is the foundation of models of "sample selection"). This assumption is extended for regression approaches with the **conditional mean independence assumption**: $E[y_0 | \mathbf{x}, C] = E[y_0 | \mathbf{x}]$ and $E[y_1 | \mathbf{x}, C] = E[y_1 | \mathbf{x}]$. This states that the outcome in the untreated state does not affect the participation. The assumption is also labeled *ignorability of the treatment*. As its name implies (and as is clear from the definitions), under ignorability, $ATE = ATET$.

Distribution of potential outcomes: The model that is used for the outcomes is the same for treated and nontreated, $f(y | \mathbf{x}, C = 1) = f(y | \mathbf{x}, C = 0)$. In a regression context, this would mean that the same regression applies in both states and that the disturbance is uncorrelated with T, or that T is exogenous. This is a very strong assumption that we will relax later.

[11] Imbens and Angrist (1994) define a still narrower margin, the "local average treatment effect," or *LATE*. *LATE* is defined with respect to a specific binary instrumental variable. Unlike *ATET*, the *LATE* is defined for a subpopulation related to the instrumental variable and differs with the definition of the instrument. Broadly, the *LATE* narrows the relevant subpopulation to those induced to participate by the variation of the instrument. This specification extends the function of the IV to make it part of the specification of the model to the extent that the object of estimation (*LATE*) is defined by the IV, not independently of it, as in the usual case.

Stable unit treatment value assumption (SUTVA): The treatment of individual i does not affect the outcome of any other individual, j. Without this assumption, which observations are subject to treatment becomes ambiguous. Pure random sampling of observations in a data set would be sufficient for statistical purposes.

Overlap assumption: For any value of $\mathbf{x}$, $0 < \text{Prob}(C = 1|\mathbf{x}) < 1$. The strict inequality in this assumption means that for any $\mathbf{x}$, the population will contain a mix of treated and nontreated individuals. The usefulness of the overlap assumption is that with it, we can expect to find, for any treated individual, an individual who looks like the treated individual, but is not treated. This assumption will be useful for regression approaches.

The following sections will describe three major tools used in the analysis of treatment effects: instrumental variable regression, regression analysis with control functions, and propensity score matching. A fourth, regression discontinuity design, was discussed in Section 6.4.2. As noted, this is a huge and rapidly growing literature. For example, Imbens and Wooldridge's (2009) survey paper runs to 85 pages and includes nearly 300 references, most of them since 2000 (likewise, Wooldridge (2010, Chapter 21)). Our purpose here is to provide some of the vocabulary and a superficial introduction to methods. The survey papers by Imbens and Wooldridge (2009) and Jones and Rice (2010) provide greater detail. The conference volume by Millment, Smith, and Vytlacil (2008) contains many theoretical contributions and empirical applications.[12] A *Journal of Business and Economic Statistics* symposium [Angrist (2001)] raised many of the important questions on whether and how it is possible to measure treatment effects.

Example 8.6 German Labor Market Interventions

"Germany long had the highest ratio of unfilled jobs to unemployed people in Europe. Then, in 2003, Berlin launched the so-called Hartz reforms, ending generous unemployment benefits that went on indefinitely. Now payouts for most recipients drop sharply after a year, spurring people to look for work. From 12.7% in 2005, unemployment fell to 7.1% last November. Even now, after a year of recession, Germany's jobless rate has risen to just 8.6%.

At the same time, lawmakers introduced various programs intended to make it easier for people to learn new skills. One initiative instructed the Federal Labor Agency, which had traditionally pushed the long-term unemployed into government-funded make-work positions, to cooperate more closely with private employers to create jobs. That program last year paid Dutch staffing agency Randstad to teach 15,000 Germans information technology, business English, and other skills. And at a Daimler truck factory in Wörth, 55 miles west of Stuttgart, several dozen short-term employees at risk of being laid off got government help to continue working for the company as mechanic trainees.

Under a second initiative, Berlin pays part of the wages of workers hired from the ranks of the jobless. Such payments make employers more willing to take on the costs of training new workers. That extra training, in turn, helps those workers keep their jobs after the aid expires, a study by the government-funded Institute for Employment Research found. Café Nenninger in the city of Kassel, for instance, used the program to train an unemployed single mother. Co-owner Verena Nenninger says she was willing to take a chance on her in part because the government picked up about a third of her salary the first year. 'It was very helpful, because you never know what's going to happen,' Nenninger says." [*Business Week* (2009)]

[12]In the initial essay in the volume, Goldberger (2008) reproduces Goldberger (1972), in which the author explores the endogeneity issue in detail with specific reference to the Head Start program of the 1960s.

Example 8.7 Treatment Effects on Earnings

LaLonde (1986) analyzed the results of a labor market experiment, The National Supported Work Demonstration, in which a group of disadvantaged workers lacking basic job skills were given work experience and counseling in a sheltered environment. Qualified applicants were assigned to training positions randomly. The treatment group received the benefits of the program. Those in the control group "were left to fend for themselves."[13] The training period was 1976–1977; the outcome of interest for the sample examined here was post-training 1978 earnings. We will attempt to replicate some of the received results based on these data in Example 8.10.

Example 8.8 The Oregon Health Insurance Experiment

The Oregon Health Insurance Experiment is a landmark study of the effect of expanding public health insurance on health care use, health outcomes, financial strain, and well-being of low-income adults. It uses an innovative randomized controlled design to evaluate the impact of Medicaid in the United States. Although randomized controlled trials are the gold standard in medical and scientific studies, they are rarely possible in social policy research. In 2008, the state of Oregon drew names by lottery for its Medicaid program for low-income, uninsured adults, generating just such an opportunity. This ongoing analysis represents a collaborative effort between researchers and the state of Oregon to learn about the costs and benefits of expanding public health insurance. (www.nber.org/oregon/) (Further details appear in Chapter 6.)

Example 8.9 The Effect of Counseling on Financial Management

Smith, Hochberg, and Greene (2014) examined the impact of a financial management skills program on later credit outcomes such as credit scores, debt, and delinquencies of a sample of home purchasers. From the abstract of the study:

> *. . . . [D]evelopments in mortgage products and drastic changes in the housing market have made the realization of becoming a homeowner more challenging. Fortunately, homeownership counseling is available to help navigate prospective homebuyers in their quest. But the effectiveness of such counseling over time continues to be contemplated. Previous studies have made important strides in our understanding of the value of homeownership counseling, but more work is needed. More specifically, homeownership education and counseling have never been rigorously evaluated through a randomized field experiment.*
>
> *This study is based on a long-term (five-year) effort undertaken by the Federal Reserve Bank of Philadelphia on the effectiveness of pre-purchase homeownership and financial management skills counseling. . . . [T]he study employs an experimental design, with study participants randomly assigned to a control or a treatment group. Participants completed a baseline survey and were tracked for four years after receiving initial assistance by means of an annual survey, which also tracks participants' life changes over time. To assist in the analysis, additional information was obtained annually to track changes in the participants' creditworthiness. The study considers the influence of counseling on credit scores, total debt, and delinquencies in payments.*

8.5.1 REGRESSION ANALYSIS OF TREATMENT EFFECTS

An earnings equation that purports to account for the value of a college education is

$$\ln \text{Earnings}_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta C_i + \varepsilon_i,$$

[13]The demonstration was run in numerous cities in the mid-1970s. See LaLonde (1986, pp. 605–609) for details on the NSW experiments.

where C_i is a dummy variable indicating whether or not the individual attended college. The same format has been used in any number of other analyses of programs, experiments, and treatments. The question is: Does δ measure the value of a college education (assuming that the rest of the regression model is correctly specified)? The answer is no if the typical individual who chooses to go to college would have relatively high earnings whether or not he or she went to college. The problem is one of self-selection. If our observation is correct, then least squares estimates of δ will actually overestimate the treatment effect—it will likely pick up the college effect as well as effects explainable by the other latent factors (that are not in $\mathbf{x}$). The same observation applies to estimates of the treatment effects in other settings in which the individuals themselves decide whether or not they will receive the treatment.

8.5.2 INSTRUMENTAL VARIABLES

The starting point to the formulation of the earnings equation would be the familiar RCM,

$$y = \mu_0 + C(\mu_1 - \mu_0) + \varepsilon_0 + C(\varepsilon_1 - \varepsilon_0),$$

where $\mu_j = E[y_j]$. Suppose, first, that $\varepsilon_1 = \varepsilon_0$, so the final term falls out of the equation. [Though the assumption is unmotivated, we note that no sample will contain direct observations on $(\varepsilon_1 - \varepsilon_0)$—no individual will be in both states—so the assumption is a reasonable normalization.] There is no presumption at this point that ε_j is uncorrelated with $\mathbf{x}$. Suppose, as well, that there exist instrumental variables, $\mathbf{z}$, that contain at least one variable that is not in $\mathbf{x}$, such that the linear projection of ε_0 on $\mathbf{x}$ and $\mathbf{z}$, $Proj(\varepsilon_0 | \mathbf{x}, \mathbf{z})$, equals $Proj(\varepsilon_0 | \mathbf{x})$. That is, $\mathbf{z}$ is *exogenous*. (See Section 4.4.5 and (4-34) for definition of the linear projection. It will be convenient to assume that $\mathbf{x}$ and $\mathbf{z}$ have no variables in common.) The linear projection is $Proj(\varepsilon_0 | \mathbf{x}) = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}$. Then,

$$y = (\mu_0 + \gamma_0) + \delta C + \mathbf{x}'\boldsymbol{\gamma} + w_0,$$

where $w_0 = \varepsilon_0 - (\gamma_0 + \mathbf{x}'\boldsymbol{\gamma})$. By construction, w_0 and $\mathbf{x}$ are uncorrelated. There is also no assumption that C is uncorrelated with w_0 since we have assumed that C is correlated with ε_0 at the outset. The setup would seem now to lend itself to a familiar IV approach. However, we have yet to certify $\mathbf{z}$ as a proper instrument. We assumed $\mathbf{z}$ is exogenous. We assume it is *relevant*, still using the projections, with $Proj(C | \mathbf{x}, \mathbf{z}) \neq Proj(C | \mathbf{x})$. This would be the counterpart to the relevance condition in Assumption 1 in Section 8.2. The model is, then,

$$y = \lambda_0 + \delta C + \mathbf{x}'\boldsymbol{\gamma} + w_0.$$

The parameters of this model can, in principle, be estimated by 2SLS. In the notation of Section 6.3, $\mathbf{X}_i = [1, C_i, \mathbf{x}_i']$ and $\mathbf{Z}_i = [1, \mathbf{z}_i', \mathbf{x}_i']$. Consistency and asymptotic normality of the 2SLS estimator are based on the usual results. See Theorem 8.1. Because we have not assumed anything about $\text{Var}[w_0 | \mathbf{x}]$, efficiency is unclear. Consistency is the objective, however, and inference can be based on heteroscedasticity robust estimators of the asymptotic covariance matrix of the 2SLS estimator, as in (8-8h) or (8-8c).

The relevance assumption holds that in the projection of C on $\mathbf{x}$ and $\mathbf{z}$,

$$C = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z + w_c = \mathbf{f}'\boldsymbol{\gamma}_c + w_c,$$

$\boldsymbol{\gamma}_z$ is not zero. Strictly, the projection works. However, because C is a binary variable, w_c equals either $-\mathbf{f}'\boldsymbol{\gamma}_c$ or $1 - \mathbf{f}'\boldsymbol{\gamma}_c$, so the lack of correlation between w_c and $\mathbf{f}$ (specifically $\mathbf{z}$) is a result of the construction of the linear projection, not necessarily a characteristic of the underlying design of the real-world counterpart to the variables in the model (though one would expect $\mathbf{z}$ to have been chosen with this in mind). One might observe that the understanding of the functioning of the instrument is that its variation makes participation more (or less) likely. As such, the relevance of the instrument is to the probability of participation. A more convincing specification that is consistent with this observation, albeit one less general, can replace the relevance assumption with a formal parametric specification of the conditional probability that C equals 1, $Prob(C = 1|\mathbf{x}, \mathbf{z}) = F(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta}) \neq Prob(C = 1|\mathbf{x})$. We also replace projections with expected values in the exogeneity assumption; $Proj(\varepsilon_0|\mathbf{x}, \mathbf{z}) = Proj(\varepsilon_0|\mathbf{x})$ will now be $E(\varepsilon_0|\mathbf{x}, \mathbf{z}) = Proj(\varepsilon_0|\mathbf{x}) = (\gamma_0 + \mathbf{x}'\boldsymbol{\gamma})$. This suggests an instrument of the form $F(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta}) = Prob(C = 1|\mathbf{x}, \mathbf{z})$, a known function—the usual choice would be a probit model (see Section 17.2)—$\Phi(\theta_0 + \mathbf{x}'\boldsymbol{\theta}_x + \mathbf{z}'\boldsymbol{\theta}_z)$ where $\Phi(t)$ is the standard normal CDF. To reiterate, the conditional probability is correlated with $C|\mathbf{x}$ but not correlated with $w_0|\mathbf{x}$. With this additional assumption, a natural instrument in the form of $\hat{F}(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta}) = \Phi(\hat{\theta}_0 + \mathbf{x}'\hat{\boldsymbol{\theta}}_x + \mathbf{z}'\hat{\boldsymbol{\theta}}_z)$ (estimated by maximum likelihood) can be used. The advantages of this approach are internally consistent specification of the treatment dummy variable and some gain in efficiency of the estimator that follows from the narrower assumptions.

This approach creates an additional issue that is not present in the previous linear approach. The approach suggested here would succeed even if there were no variables in $\mathbf{z}$. The IV estimator is $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ where the rows of $\mathbf{Z}$ and $\mathbf{X}$ are $[1, \hat{\Phi}, \mathbf{x}']$ and $[1, C, \mathbf{x}']$. As long as $\hat{\Phi}$ is not a linear function of $\mathbf{x}$ (and is both relevant and exogenous), then the parameters will be identified by this IV estimator. Because $\hat{\Phi}$ is nonlinear, it could meet these requirements even without any variables in $\mathbf{z}$. The parameters in this instance are identified by the nonlinear functional form of the probability model. Typically, the probability is at least reasonably highly (linearly) correlated with the variables in the model, so possibly severe problems of multicollinearity are likely to appear. But, more to the point, the entire logic of the instrumental variable approach is based on an exogenous source of variation that is correlated with the endogenous variable and not with the disturbance. The nonlinear terms in the probability model do not persuasively pass that test. Thus, the typical application does, indeed, ensure that there are excluded (from the main equation) variables in $\mathbf{z}$.[14]

Finally, note that because $\hat{F}(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta})$ is not a linear function of $\mathbf{x}$ and $\mathbf{z}$, this IV estimator is not two-stage least squares. That is, $\mathbf{y}$ is not regressed on $(1, \hat{\Phi}, \mathbf{x})$ to estimate $\lambda_0, \delta, \boldsymbol{\gamma}$. Rather, the estimator is in (8-6), $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$. Because no assumption has been made about the disturbance variance, the robust covariance matrix estimator in (8-8h) should be used.

[14]As an example, Scott, Schurer, Jensen, and Sivey (2009) state, "Although the model is formally identified by its nonlinear functional form, as long as the full rank condition of the data matrix is ensured (Heckman, 1978; Wilde, 2000), we introduce exclusion restrictions to aid identification of the causal parameter . . . The row vector I_{ij} captures the variables included in the PIP participation Equation (5) but excluded from the outcome Equation (4)." ("The Effects of an Incentive Program on Quality of Care in Diabetes Management," *Health Economics*, 19, 2009, pp. 1091–1108, Section 4.2.)

8.5.3 A CONTROL FUNCTION ESTIMATOR

The list of assumptions and implications that produced the second IV estimator above was:

Rubin Causal Model	$y = Cy_1 + (1 - C)y_0$ $= \mu_0 + C(\mu_1 - \mu_0) + \varepsilon_0 + C(\varepsilon_1 - \varepsilon_0),$
Nonignorability of the Treatment	$\text{Cov}(C, \varepsilon_0) \neq 0,$
Normalization	$\varepsilon_1 - \varepsilon_0 = 0,$
Exogeneity and Linearity	$Proj(\varepsilon_0 \mid \mathbf{x}, \mathbf{z}) = E[\varepsilon_0 \mid \mathbf{x}, \mathbf{z}] = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma},$ no assumption is made about $\text{Var}[\varepsilon_0 \mid \mathbf{x}],$
Relevance of the Instruments	$Prob(C = 1 \mid \mathbf{x}, \mathbf{z}) = F(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta}) \neq Prob(C = 1 \mid \mathbf{x}),$
Reduced Form	$y = \lambda_0 + \delta C + \mathbf{x}'\boldsymbol{\gamma} + w_0,\ Cov(\mathbf{x}, w_0) = 0$ is implied,
Endogenous Treatment Dummy Variable	$Cov(C, w_0) \neq 0,$
Probit Model for Prob(C = 1|x, z)	$C* = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z + w_c, w_c \sim N[0, 1^2],$ $C = 1$ if $C* > 0$ and $C = 0$ if $C* \leq 0,$ $Prob(C = 1 \mid \mathbf{x}, \mathbf{z}) = \Phi(\theta_0 + \mathbf{x}'\boldsymbol{\theta}_x + \mathbf{z}'\boldsymbol{\theta}_z).$

The source of the endogeneity of the treatment dummy variable is now more explicit. Because neither **x** nor **z** is correlated with w_0, the source is the correlation of w_c and w_0. As in all such cases, the ultimate source of the endogeneity is the covariation among the unobservables in the model.

The foregoing is sufficient to produce a consistent instrumental variable estimator. We now pursue whether with the same data and assumptions, there is a regression-based estimator. Based on the assumptions, we find that

$$E[y \mid C = 1, \mathbf{x}, \mathbf{z}] = \lambda_0 + \delta + \mathbf{x}'\boldsymbol{\gamma} + E[w_0 \mid C = 1, \mathbf{x}, \mathbf{z}],$$
$$E[y \mid C = 0, \mathbf{x}, \mathbf{z}] = \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + E[w_0 \mid C = 0, \mathbf{x}, \mathbf{z}].$$

Because we have not specified the last term, the model is incomplete. Suppose the model is fully parameterized with (w_0, w_c) bivariate normally distributed with means 0, variances σ^2 and 1 and covariance $\rho\sigma$. Under these assumptions, the functional form of the conditional mean is known,

$$\begin{aligned} E[y \mid C = 1, \mathbf{x}, \mathbf{z}] &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta + E[w_0 \mid C = 1, \mathbf{x}, \mathbf{z}] \\ &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta + E[w_0 \mid w_c > (-\gamma_0 - \mathbf{x}'\boldsymbol{\gamma}_x - \mathbf{z}'\boldsymbol{\gamma}_z)] \\ &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta + \rho\sigma\left[\frac{\phi(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)}{\Phi(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)}\right]. \end{aligned}$$

The counterpart for $C = 0$ would be

$$E[y \mid C = 0, \mathbf{x}, \mathbf{z}] = \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \rho\sigma\left[\frac{-\phi(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)}{[1 - \Phi(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)]}\right].$$

By using the symmetry of the normal distribution, $\phi(t) = \phi(-t)$ and $\Phi(t) = 1 - \Phi(-t)$, we can combine these into a single regression,

$$\begin{aligned} E[y \mid C\, \mathbf{x}_i, \mathbf{z}_i] &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta C + \rho\sigma\left[\frac{(2C - 1)\phi[(2C - 1)(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)]}{\Phi[(2C - 1)(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)]}\right] \\ &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta C + \tau G(C, \mathbf{x}, \mathbf{z}: \boldsymbol{\theta}). \end{aligned}$$

(See Theorem 19.5.) The result is a feature of the bivariate normal distribution. There are two approaches that could be taken. The conditional mean function is a nonlinear regression that can be estimated by nonlinear least squares. The bivariate normality assumption carries an implicit assumption of homoscedasticity, so there is no need for a heteroscedasticity robust estimator for the covariance matrix. Nonlinear least squares might be quite cumbersome. A simpler, two-step "control function" approach would be to fit the probit model as before, then compute the bracketed term and add it as an additional term. The estimating equation is

$$y = \lambda_0 + \delta C + \mathbf{x}'\boldsymbol{\gamma} + \boldsymbol{\tau}\hat{G} + h,$$

where $h = y - E[y|C, \mathbf{x}, \mathbf{z}]$. This can be estimated by linear least squares. As with other control function estimators, the asymptotic covariance matrix for the estimator must be adjusted for the constructed regressor. [See Heckman (1979) for results related to this model.] The result of Murphy and Topel (2002) can be used to obtain the correction. Bootstrapping can be used as well. [This turns out to be identical to Heckman's (1979) "sample selection" model developed in Section 19.5.2. A covariance matrix for the two-step estimator as well as a full information maximum likelihood estimator are developed there.]

The precision and compactness of this result has been purchased by adding the bivariate normality assumption. It has also been made much simpler with the still unmotivated assumption, $\varepsilon_1 - \varepsilon_0 = 0$. A distributional assumption can be substituted for the normalization. Wooldridge (2010, pp. 945–948) assumes that $[w_c, (\varepsilon_1 - \varepsilon_0)]$ are bivariate normally distributed, and obtains another control function estimator, again based on properties of the bivariate normal distribution.

8.5.4 PROPENSITY SCORE MATCHING

If the treatment assignment is completely ignorable, then, as noted, estimation of the treatment effects is greatly simplified. Suppose, as well, that there are observable variables that influence both the outcome and the treatment assignment. Suppose it is possible to obtain pairs of individuals matched by a common $\mathbf{x}_i$, one with $C_i = 0$, the other with $C_i = 1$. If done with a sufficient number of pairs so as to average over the population of $\mathbf{x}_i$s, then a *matching estimator*, the average value of $(y_i|C_i = 1) - (y_i|C_i = 0)$, would estimate $E[y_1 - y_0]$, which is what we seek. Of course, it is optimistic to hope to find a large sample of such matched pairs, both because the sample overall is finite and because there may be many regressors, and the "cells" in the distribution of $\mathbf{x}_i$ are likely to be thinly populated. This will be worse when the regressors are continuous, for example, with a family income variable. Rosenbaum and Rubin (1983) and others[15] suggested, instead, matching on the propensity score, $F(\mathbf{x}_i) = \text{Prob}(C_i = 1|\mathbf{x}_i)$. Individuals with similar propensity scores are paired and the average treatment effect is then estimated by the differences in outcomes. Various strategies are suggested by the authors for obtaining the necessary subsamples and for verifying the conditions under which the procedures will be valid.[16] We will examine and try to replicate a well-known application in Example 8.10.

[15]Other important references in this literature are Becker and Ichino (1999), Dehejia and Wahba (1999), LaLonde (1986), Heckman, Ichimura, and Todd (1997, 1998), Robins and Rotnitzky (1995), Heckman, Ichimura, Smith, and Todd (1998), Heckman, LaLonde, and Smith (1999), Heckman, Tobias, and Vytlacil (2003), Hirano, Imbens, and Ridder (2003), and Heckman and Vytlacil (2000).

[16]See, for example, Becker and Ichino (2002).

Example 8.10 Treatment Effects on Earnings

LaLonde (1986) analyzed the results of a labor market experiment, The National Supported Work Demonstration, in which a group of disadvantaged workers lacking basic job skills were given work experience and counseling in a sheltered environment. Qualified applicants were assigned to training positions randomly. The treatment group received the benefits of the program. Those in the control group "were left to fend for themselves." The training period was 1976–1977; the outcome of interest for the sample examined here was posttraining 1978 earnings.

LaLonde reports a large variety of estimates of the treatment effect, for different subgroups and using different estimation methods. Nonparametric estimates for the group in our sample are roughly \$900 for the income increment in the posttraining year. (See LaLonde, p. 609.) Similar results are reported from a two-step regression-based estimator similar to the control function estimator in Section 8.5.3. (See LaLonde's footnote to Table 6, p. 616.)

LaLonde's data are fairly well traveled, having been used in replications and extensions in, for example, Dehejia and Wahba (1999), Becker and Ichino (2002), Stata (2006), Dehejia (2005), Smith and Todd (2005), and Wooldridge (2010). We have reestimated the matching estimates reported in Becker and Ichino along with several side computations including the estimators developed in Sections 8.5.2 and 8.5.3. The data in the file used there (and here) contain 2,490 control observations and 185 treatment observations on the following variables:

$$
\begin{aligned}
t &= \text{treatment dummy variable,}\\
age &= \text{age in years,}\\
educ &= \text{education in years,}\\
marr &= \text{dummy variable for married,}\\
black &= \text{dummy variable for black,}\\
hisp &= \text{dummy variable for Hispanic,}\\
nodegree &= \text{dummy for no degree (not used),}\\
re74 &= \text{real earnings in 1974,}\\
re75 &= \text{real earnings in 1975,}\\
re78 &= \text{real earnings in 1978.}
\end{aligned}
$$

Transformed variables added to the equation are

$$
\begin{aligned}
age^2 &= age \text{ squared,}\\
educ^2 &= educ \text{ squared,}\\
re74^2 &= re74 \text{ squared,}\\
re75^2 &= re75 \text{ squared,}\\
blacku\ 74 &= black \text{ times } 1(re74 = 0).
\end{aligned}
$$

We also scaled all earnings variables by 10,000 before beginning the analysis. (See Appendix Table F19.3. The data are downloaded from the Website http://users.nber.org/~rdehejia/nswdata2.html. The two specific subsamples are in http://www.nber.org/~rdehejia//nsw_control.txt, and http://www.nber.org/~rdehejia/nsw_treated.txt.) (We note that Becker and Ichino report they were unable to replicate Dehejia and Wahba's results, although they could come reasonably close. We, in turn, were not able to replicate either set of results, though we, likewise, obtained quite similar results. See Table 8.3.)

To begin, Figure 8.2 describes the re78 data for the treatment group in the upper panel and the controls in the lower. Any regression- (or sample means–) based analysis of the differences of the two distributions will reflect the fact that the mean of the controls is far larger than that of the treatment group. The re74 and re75 data appear similar, so estimators that account

for the observable past values should be able to isolate the difference attributable to the treatment, if there is a difference.

Table 8.3 lists the results obtained with the regression-based methods and matching based on the propensity scores. The specification for the regression-based approaches is

TABLE 8.3 Estimates of Average Treatment Effect on the Treated

Simple difference in means: $\overline{re78}_1 - \overline{re78}_0 = 6{,}349 - 21{,}553 = -15{,}204$[a]

Estimator	δ	*Standard Error (Method)*
Regression Based		
Simple OLS	859[a]	765[a] (Robust Standard Error)
2SLS	2,021	1,690 (Robust Standard Error)
IV Using predicted probabilities	2,145	1,131 (Robust Standard Error)
2 Step Control Function	2,273	1,012 (100 Bootstrap Replications)
		1,249 (Heckman Two Step)
Propensity Score Matching[c]		
Matching	1,571	669 (25 Bootstrap Replications)
Becker and Ichino	1,537[b]	1,016[b] (100 Bootstrap Replications)

[a] See Wooldridge (2010, p. 929, Table 21.1).
[b] See Becker and Ichino (2002, p. 374) based on Kernel Matching and common support. Number of controls = 1,157 (1,155 here).
[c] Becker and Ichino employed the `pscore` and `attk` routines in *Stata*. Results here used `LOGIT` and `PSMATCH` in *NLOGIT6*.

TABLE 8.4 Empirical Distribution of Propensity Scores

Percent	*Lower*	*Upper*
0–5	0.000591	0.000783
5–10	0.000787	0.001061
10–15	0.001065	0.001377
15–20	0.001378	0.001748
20–25	0.001760	0.002321
25–30	0.002340	0.002956
30–35	0.002974	0.004057
35–40	0.004059	0.005272
40–45	0.005278	0.007486
45–50	0.007557	0.010451
50–55	0.010563	0.014643
55–60	0.014686	0.022462
60–65	0.022621	0.035060
65–70	0.035075	0.051415
70–75	0.051415	0.076188
75–80	0.076376	0.134189
80–95	0.134238	0.320638
85–90	0.321233	0.616002
90–95	0.624407	0.949418
95–100	0.949418	0.974835

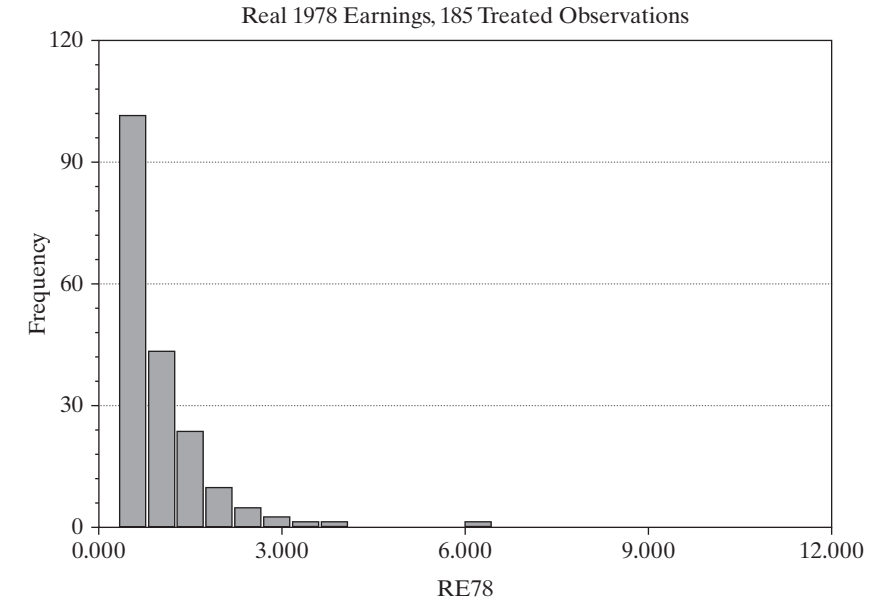
Sample size = 1,347
Average score = 0.137238
Std. Dev score = 0.274079

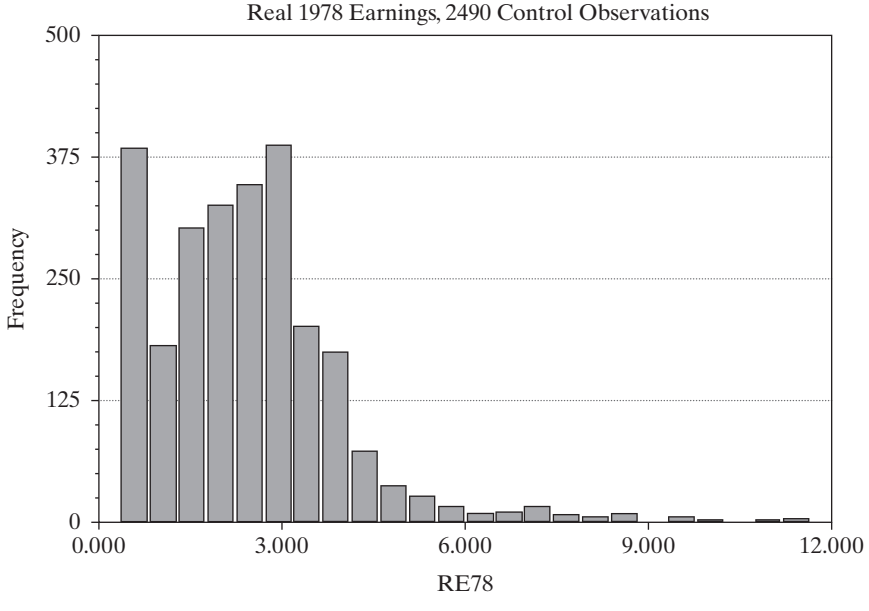
	Lower	Upper	# Obs
1	0.000591	0.098016	1041
2	0.098016	0.195440	63
3	0.195440	0.390289	65
4	0.390289	0.585138	36
5	0.585138	0.779986	32
6	0.779986	0.877411	17
7	0.877411	0.926123	7
8	0.926123	0.974835	86

$$re78 = \lambda_0 + y_1 age + y_2 educ + y_3 black + y_4 hisp + y_5 marr + y_6 re74 + y_7 re75 + \delta T + w_0.$$

The additional variables in **z** are (age^2, $educ^2$, $re74^2$, $re75^2$, *blacku*74). [Note, for consistency with Becker and Ichino, *nodegree* was not used. The specification of **x** in the regression equation follows Wooldridge (2010).] As anticipated, the simple difference in means is

FIGURE 8.2 Real 1978 Earnings, Treated Versus Controls.





uninformative. The regression-based estimates are quite consistent; the estimate of ATT is roughly $2,100. The propensity score method focuses only on the observable differences in the observations (including, crucially, *re*74 and *re*75) and produces an estimate of about $1,550.

The propensity score matching analysis proceeded as follows: A logit model in which the included variables were a *constant, age, age*^{2}, *education, education*^{2}, *marr, black, hisp, re74, re75, re742, re752*, and *blacku74* was computed for the treatment assignment. The fitted probabilities are used for the propensity scores. By means of an iterative search, the range of propensity scores was partitioned into eight regions within which, by a simple *F* test, the mean scores of the treatments and controls were not statistically different. The partitioning is shown in Table 8.4. The 1,347 observations are all the treated observations and the 1,162 control observations are those whose propensity scores fell within the range of the scores for the treated observations.

Within each interval, each treated observation is paired with a small number of the nearest control observations. We found the average difference between treated observation and control to equal $1,574.35. Becker and Ichino reported $1,537.94.

8.6 HYPOTHESIS TESTS

There are several tests to be carried out in this model.

8.6.1 TESTING RESTRICTIONS

For testing linear restrictions in H_0: $\boldsymbol{R\beta} = \boldsymbol{q}$, *the Wald statistic based on whatever form of* Asy.Var$[\mathbf{b}_{IV}]$ *has been computed will be the usual choice. The test statistic, based on the unrestricted estimator, will be*

$$\chi^2[J] = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}\,\text{Est.Asy.Var}(\hat{\boldsymbol{\beta}})\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}). \tag{8-18}$$

For testing the simple hypothesis that a coefficient equals zero, this is the square of the usual *t* ratio that is always reported with the estimated coefficient. The *t* ratio, itself, can be used instead, though the implication is that the large sample critical value, 1.96 for 95%, for example, would be used rather than the *t* distribution.

For the 2SLS estimator based on least squares regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$ an asymptotic *F* statistic can be computed as follows:

$$F[J, n - K] = \frac{\left\{\sum_{i=1}^{n}(y_i - \hat{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_{\text{Restricted}})^2 - \sum_{i=1}^{n}(y_i - \hat{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_{\text{Unrestricted}})^2\right\}/J}{\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{Unrestricted}})^2/(n - K)}. \tag{8-19}$$

[See Wooldridge (2010, p. 105).] As in the regression model [see (5-14) and (5-15)], an approximation to the *F* statistic will be the chi-squared statistic, *JF*. Unlike the earlier case, however, *J* times the statistic in (8-19) is not equal to the result in (8-18) even if the denominator is rescaled by $(n-K)/n$. They are different approximations. The *F* statistic is computed using both resticted and unresticted estimators.

A third approach to testing the hypothesis of the restrictions can be based on the Lagrange multiplier principle. The moment equation for the 2SLS estimator is

$$\bar{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i\hat{\varepsilon}_i = \mathbf{0}.$$

(Note that the residuals are computed using the original $\mathbf{x}$, not the prediction.) The mean vector $\bar{\mathbf{g}}$ will equal $\mathbf{0}$ when it is computed using $\hat{\boldsymbol{\beta}}_{\text{Unrestricted}}$ to compute the residuals. It will generally not equal zero if $\hat{\boldsymbol{\beta}}_{\text{Restricted}}$ is used instead. We consider using a Wald test to test the hypothesis that $\text{E}[\bar{\mathbf{g}}] = 0$. The asymptotic variance of $\bar{\mathbf{g}}$ will be estimated using $1/n$ times the matrix in (8-8), (8-8h) or (8-8c), whichever is appropriate. The Wald statistic will be

$$\chi^2[J] = \left[\sum_{i=1}^{n} \hat{\mathbf{x}}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{Restricted}})\right]' \left(\frac{1}{n}\text{Est.Asy.Var}\left[\hat{\boldsymbol{\beta}}_{\text{Restricted}}\right]\right)^{-1} \left[\sum_{i=1}^{n} \hat{\mathbf{x}}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{Restricted}})\right].$$

A convenient way to carry out this test is the approximation $\chi^2[J] = nR^2$ where the R^2 is the uncentered R^2 in the least squares regression of $\hat{\boldsymbol{\varepsilon}}$ on $\hat{\mathbf{X}}$.

8.6.2 SPECIFICATION TESTS

There are two aspects of the model that we would be interested in verifying if possible, rather than assuming them at the outset. First, it will emerge in the derivation in Section 8.4.1 that of the two estimators considered here, least squares and instrumental variables, the first is unambiguously more efficient (i.e., has a smaller variance around its mean). The IV estimator is robust; it is consistent whether or not $\text{plim}(\mathbf{X}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$. However, if IV is not needed, that is, if $\boldsymbol{\gamma} = \mathbf{0}$, then least squares would be a better estimator by virtue of its smaller variance.[17] For this reason, and possibly in the interest of a test of the theoretical specification of the model, a test that reveals information about the bias of least squares will be useful. Second, the use of two-stage least squares with $L > K$, that is, with "additional" instruments, entails $L - K$ restrictions on the relationships among the variables in the model. As might be apparent from the derivation thus far, when there are K variables in $\mathbf{X}$, some of which may be endogenous, then there must be at least K variables in $\mathbf{Z}$ in order to identify the parameters of the model, that is, to obtain consistent estimators of the parameters using the information in the sample. When there is an excess of instruments, one is actually imposing additional, arguably superfluous restrictions on the process generating the data. Consider, once again, the agricultural market example at the end of Section 8.3.4. In that structure, it is certainly safe to assume that *Rainfall* is an exogenous event that is uncorrelated with the disturbances in the demand equation. But, it is conceivable that the interplay of the markets involved might be such that the *InputPrice* is correlated with the shocks in the demand equation. In the market for biofuels, corn is both an input in the market supply and an output in other markets. In treating *InputPrice* as exogenous in that example, we would be imposing the assumption that *InputPrice* is uncorrelated with ε_D, at least by some measure unnecessarily because the parameters of the demand equation can be estimated without this assumption. This section will describe two specification tests that consider these aspects of the IV estimator.

[17] It is possible that even if least squares is inconsistent, it might still be more precise. If LS is only slightly biased but has a much smaller variance than IV, then by the expected squared error criterion, variance plus squared bias, least squares might still prove the preferred estimator. This turns out to be nearly impossible to verify empirically.

8.6.3 TESTING FOR ENDOGENEITY: THE HAUSMAN AND WU SPECIFICATION TESTS

If the regressors in the model are not correlated with the disturbances and are not measured with error, then there would be some benefit to using the least squares (LS) estimator rather than the IV estimator. Consider a comparison of the two covariance matrices *under the hypothesis that both estimators are consistent, that is, assuming* $\text{plim}\,(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$ and assuming A.4 (Section 8.2). The difference between the asymptotic covariance matrices of the two estimators is

$$\begin{aligned}\text{Asy.Var}[\mathbf{b}_{\text{IV}}] - \text{Asy.Var}[\mathbf{b}_{\text{LS}}] &= \frac{\sigma^2}{n}\text{plim}\left(\frac{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} - \frac{\sigma^2}{n}\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\\ &= \frac{\sigma^2}{n}\text{plim}\, n[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]\\ &= \frac{\sigma^2}{n}\text{plim}\, n\{[\mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{X}]^{-1} - [\mathbf{X}'\mathbf{X}]^{-1}\}\\ &= \frac{\sigma^2}{n}\text{plim}\, n\{[\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}]^{-1} - [\mathbf{X}'\mathbf{X}]^{-1}\}.\end{aligned} \tag{8-20}$$

The matrix in braces is nonnegative definite, which establishes that least squares is more efficient than IV. Our interest in the difference between these two estimators goes beyond the question of efficiency. The null hypothesis of interest will be specifically whether $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. Seeking the covariance between $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ through $(1/n)\mathbf{X}'\mathbf{e}$ is fruitless, of course, because $(1/n)\mathbf{X}'\mathbf{e} = \mathbf{0}$. In a seminal paper, Hausman (1978) developed an alternative testing strategy. The logic of Hausman's approach is as follows. Under the null hypothesis, we have two consistent estimators of $\boldsymbol{\beta}$, $\mathbf{b}_{\text{LS}}$ and $\mathbf{b}_{\text{IV}}$. Under the alternative hypothesis, only one of these, $\mathbf{b}_{\text{IV}}$, is consistent. The suggestion, then, is to examine $\mathbf{d} = \mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}}$. Under the null hypothesis, $\text{plim}\,\mathbf{d} = \mathbf{0}$, whereas under the alternative, $\text{plim}\,\mathbf{d} \neq \mathbf{0}$. We will test this hypothesis with a Wald statistic,

$$\begin{aligned}H &= \mathbf{d}'\{\text{Est.Asy.Var}[\mathbf{d}]\}^{-1}\mathbf{d}\\ &= (\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}})'\{\text{Est.Asy.Var}[\mathbf{b}_{\text{IV}}] - \text{Est.Asy.Var}[\mathbf{b}_{\text{LS}}]\}^{-1}(\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}})\\ &= (\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}})'\hat{\mathbf{H}}^{-1}(\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}}),\end{aligned}$$

where $\hat{\mathbf{H}}^{-1}$ is the estimator of the covariance matrix in (8-20). Under the null hypothesis, we have two different, but consistent, estimators of σ^2. If we use s^2 as the common estimator, then the statistic will be

$$H = \frac{\mathbf{d}'[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{d}}{s^2}.$$

It is tempting to invoke our results for the full rank quadratic form in a normal vector and conclude the degrees of freedom for this chi-squared statistic is K. However, the rank of $[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]$ is only $K^* = K - K_0$, where K_0 is the number of exogenous variables in $\mathbf{X}$ (and the ordinary inverse will not exist), so K^* is the degrees of freedom for the test. The Wald test requires a generalized inverse [see Hausman and Taylor (1981)], so it is going to be a bit cumbersome. An alternative **variable addition test** approach devised by Wu (1973) and Durbin (1954) is simpler. An F or Wald statistic with

K^* and $n - K - K^*$ degrees of freedom can be used to test the joint significance of the elements of $\boldsymbol{\gamma}$ in the augmented regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{X}}^*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*, \tag{8-21}$$

where $\hat{\mathbf{X}}^*$ are the fitted values in regressions of the variables in $\mathbf{X}^*$ on $\mathbf{Z}$. This result is equivalent to the Hausman test for this model.[18]

Example 8.5 Labor Supply Model (Continued)

For the labor supply equation estimated in Example 8.5, we used the Wu (variable addition) test to examine the endogeneity of the ln *Wage* variable. For the first step, ln $Wage_{it}$ is regressed on $\mathbf{z}_{1,it}$. The predicted value from this equation is then added to the least squares regression of Wks_{it} on $\mathbf{x}_{it}$. The results of this regression are

$$\begin{aligned} Wks_{it} = {} & 18.8987 + 0.6938 \ln Wage_{it} - 0.4600\, Ed_i - 2.3602 Union_{it} \\ & (12.3284)\ (0.1980) \qquad\quad (0.1490) \qquad (0.2423) \\ & + 0.6958\, Fem_i + 4.4891\, fitted \ln Wage_{it} + u_{it}, \\ & \quad (1.0054) \qquad (2.1290), \end{aligned}$$

where the estimated standard errors are in parentheses. The *t* ratio on the fitted log wage coefficient is 2.108, which is larger than the critical value from the standard normal table of 1.96. Therefore, the hypothesis of exogeneity of the log *Wage* variable is rejected. If $\mathbf{z}_{2,it}$ is used instead, the *t* ratio on the predicted value is 2.96, which produces the same conclusion.

The control function estimator based on (8-16),

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \rho(x_2 - \mathbf{z}'\mathbf{p}) + \widetilde{w},$$

resembles the estimating equation in (8-21). It is actually equivalent. If the residual in (8-16) is replaced by the prediction, $\mathbf{z}'\mathbf{p}$, the identical least squares results are obtained save for the coefficient on the residual, which changes sign. The results in the preceding example would thus be identical save for the sign of the coefficient on the prediction of ln *Wage*, which would be negative. The implication (as happens in many applications) is that the control function estimator provides a simple constructive test for endogeneity that is the same as the Hausman–Wu test. A test of the significance of the coefficient on the control function is equivalent to the Hausman test.

8.6.4 A TEST FOR OVERIDENTIFICATION

The motivation for choosing the IV estimator is not efficiency. The estimator is constructed to be consistent; efficiency is a secondary consideration. In Chapter 13, we will revisit the issue of efficient method of moments estimation. The observation that 2SLS represents the most efficient use of all L instruments establishes only the efficiency of the estimator in the class of estimators that use K linear combinations of the columns of $\mathbf{Z}$. The IV estimator is developed around the **orthogonality conditions**,

$$E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}. \tag{8-22}$$

The sample counterpart to this is the **moment equation**,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i = \mathbf{0}. \tag{8-23}$$

[18] Algebraic derivations of this result can be found in the articles and in Davidson and MacKinnon (2004, Section 8.7).

The solution, when $L = K$, is $\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$, as we have seen. If $L > K$, then there is no single solution, and we arrived at 2SLS as a strategy. Estimation is still based on (8-23). However, the sample counterpart is now L equations in K unknowns and (8-23) has no solution. Nonetheless, under the hypothesis of the model, (8-22) remains true. We can consider the additional restrictions as a hypothesis that might or might not be supported by the sample evidence. The excess of moment equations provides a way to test the overidentification of the model. The test will be based on (8-23), which, when evaluated at $\mathbf{b}_{\mathrm{IV}}$, will not equal zero when $L > K$, though the hypothesis in (8-22) might still be true.

The test statistic will be a Wald statistic. (See Section 5.4.) The sample statistic, based on (8-23) and the IV estimator, is

$$\overline{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e_{\mathrm{IV},i} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\mathbf{b}_{\mathrm{IV}}).$$

The Wald statistic is

$$\chi^2[L - K] = \overline{\mathbf{m}}'[\mathrm{Var}(\overline{\mathbf{m}})]^{-1}\overline{\mathbf{m}}.$$

To complete the construction, we require an estimator of the variance. There are two ways to proceed. Under the assumption of the model,

$$\mathrm{Var}[\overline{\mathbf{m}}] = \frac{\sigma^2}{n^2}\mathbf{Z}'\mathbf{Z},$$

which can be estimated easily using the sample estimator of σ^2. Alternatively, we might base the estimator on (8-22), which would imply that an appropriate estimator would be

$$\mathit{Est.}\mathrm{Var}[\overline{\mathbf{m}}] = \frac{1}{n^2}\sum_{i=1}(\mathbf{z}_i e_{\mathrm{IV},i})(\mathbf{z}_i e_{\mathrm{IV},i})' = \frac{1}{n^2}\sum_{i=1} e_{\mathrm{IV},i}^2\mathbf{z}_i\mathbf{z}_i'.$$

These two estimators will be numerically different in a finite sample, but under the assumptions that we have made so far, both (multiplied by n) will converge to the same matrix, so the choice is immaterial. Current practice favors the second. The Wald statistic is, then,

$$LM = n\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e_{\mathrm{IV},i}\right)'\left[\frac{1}{n}\sum_{i=1}^{n} e_{\mathrm{IV},i}^2\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e_{\mathrm{IV},i}\right).$$

A remaining detail is the number of degrees of freedom. The test can only detect the failure of $L - K$ moment equations, so that is the rank of the quadratic form; the limiting distribution of the statistic is chi squared with $L - K$ degrees of freedom. If the equation is exactly identified, then $(1/\mathrm{n})\mathbf{Z}'\mathbf{e}_{\mathrm{IV}}$ will be exactly zero. As we saw in testing linear restrictions in Section 8.5.1, there is a convenient way to compute the LM statistic. The chi-squared statistic can be computed as n times the uncentered R^2 in the linear regression of $\mathbf{e}_{\mathrm{IV}}$ on $\mathbf{Z}$ that would be

$$LM = \frac{\mathbf{e}_{IV}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e}_{IV}}{\mathbf{e}_{IV}'\mathbf{e}_{IV}}.$$

Example 8.11 Overidentification of the Labor Supply Equation

In Example 8.5, we computed 2SLS estimates of the parameters of an equation for weeks worked. The estimator is based on

$$\mathbf{x} = [1, \ln \textit{Wage}, \textit{Education}, \textit{Union}, \textit{Female}]$$

and

$$\mathbf{z} = [1, \text{Ind}, \text{Education}, \text{Union}, \text{Female}, \text{SMSA}].$$

There is one overidentifying restriction. The sample moment based on the 2SLS results in Table 8.1 is

$$(1/4165)\mathbf{Z}'\mathbf{e}_{2\,\text{SLS}} = [0, .03476, 0, 0, 0, -.01543]'.$$

The chi-squared statistic is 1.09399 with one degree of freedom. If the first suggested variance estimator is used, the statistic is 1.05241. Both are well under the 95 percent critical value of 3.84, so the hypothesis of overidentification is not rejected. Table 8.5 displays the 2SLS estimates based on the two instruments separately and the estimates based on both.

We note a final implication of the test. One might conclude, based on the underlying theory of the model, that the overidentification test relates to one particular instrumental variable and not another. For example, in our market equilibrium example with two instruments for the demand equation, *Rainfall* and *InputPrice*, rainfall is obviously exogenous, so a rejection of the overidentification restriction would eliminate *InputPrice* as a valid instrument. However, this conclusion would be inappropriate; the test suggests only that one or more of the elements in (8-22) are nonzero. It does not suggest which elements in particular these are.

8.7 WEAK INSTRUMENTS AND LIML

Our analysis thus far has focused on the "identification" condition for IV estimation, that is, the "exogeneity assumption," A.I9, which produces

$$\text{plim}\,(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}. \tag{8-24}$$

Taking the "relevance" assumption,

$$\text{plim}\,(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{ZX}}, \text{ a finite, nonzero, } L \times K \text{ matrix with rank } K, \tag{8-25}$$

TABLE 8.5 2SLS Estimates of the Labor Supply Equation

	IND		*SMSA*		*IND and SMSA*	
Variable	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
Constant	18.8987	20.26604	33.0018	9.10852	30.7044	8.25041
LWAGE	5.18285	3.47416	2.75658	1.56100	3.15182	1.41058
ED	−0.46000	0.24352	−0.29272	0.12414	−0.31997	0.11453
UNION	−2.36016	0.43069	−2.16164	0.30395	−2.19398	0.30507
FEM	0.69567	1.66754	−0.41950	0.85547	−0.23784	0.79781
$\hat{\sigma}$	5.32268		5.08719		5.11405	

as given produces a consistent IV estimator. In absolute terms, with (8-24) in place, (8-25) is sufficient to assert consistency. As such, researchers have focused on *exogeneity* as the defining problem to be solved in constructing the IV estimator. A growing literature has argued that greater attention needs to be given to the relevance condition. While, strictly speaking, (8-25) is indeed sufficient for the asymptotic results we have claimed, the common case of "weak instruments," in which (8-25) is only barely true has attracted considerable scrutiny. In practical terms, instruments are "weak" when they are only slightly correlated with the right-hand-side variables, $\mathbf{X}$; that is, $(1/n)\mathbf{Z}'\mathbf{X}$ is *close* to zero. Researchers have begun to examine these cases, finding in some an explanation for perverse and contradictory empirical results.[19]

Superficially, the problem of weak instruments shows up in the asymptotic covariance matrix of the IV estimator,

$$\text{Asy.Var}[\mathbf{b}_{\text{IV}}] = \frac{\sigma_\varepsilon^2}{n}\left[\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1},$$

which will be "large" when the instruments are weak, and, other things equal, larger the weaker they are. However, the problems run deeper than that. Nelson and Startz (1990a,b) and Hahn and Hausman (2003) list two implications: (i) The 2SLS estimator is badly biased toward the ordinary least squares estimator, which is known to be inconsistent, and (ii) the standard first-order asymptotics (such as those we have used in the preceding) will not give an accurate framework for statistical inference. Thus, the problem is worse than simply lack of precision. There is also at least some evidence that the issue goes well beyond "small sample problems."[20]

Current research offers several prescriptions for detecting weakness in instrumental variables. For a single endogenous variable ($\mathbf{x}$ that is correlated with $\boldsymbol{\varepsilon}$), the standard approach is based on the first-step OLS regression of 2SLS. The conventional F statistic for testing the hypothesis that all the coefficients in the regression

$$x_i = \mathbf{z}_i'\boldsymbol{\pi} + u_i$$

are zero is used to test the "hypothesis" that the instruments are weak. An F statistic less than 10 signals the problem.[21] When there are more than one endogenous variables in the model, testing each one separately using this test is not sufficient, because collinearity among the variables could impact the result but would not show up in either test. Shea (1997) proposes a four-step multivariate procedure that can be used. Godfrey (1999) derived a surprisingly simple alternative method of doing the computation. For endogenous variable k, the Godfrey statistic is the ratio of the estimated variances of the two estimators, OLS and 2SLS,

$$R_k^2 = \frac{v_k(OLS)/\mathbf{e}'\mathbf{e}(\text{OLS})}{v_k(2SLS)/\mathbf{e}'\mathbf{e}(\text{2SLS})},$$

[19]Important references are Nelson and Startz (1990a,b), Staiger and Stock (1997), Stock, Wright, and Yogo (2002), Hahn and Hausman (2002, 2003), Kleibergen (2002), Stock and Yogo (2005), and Hausman, Stock, and Yogo (2005).

[20]See Bound, Jaeger, and Baker (1995).

[21]See Nelson and Startz (1990b), Staiger and Stock (1997), and Stock and Watson (2007, Chapter 12) for motivation of this specific test.

where v_k (OLS) is the kth diagonal element of $[\mathbf{e}'\mathbf{e}(\text{OLS})/(n - K)](\mathbf{X}'\mathbf{X})^{-1}$ and v_k (2SLS) is defined likewise. With the scalings, the statistic reduces to

$$R_k^2 = \frac{(\mathbf{X}'\mathbf{X})^{kk}}{(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{kk}},$$

where the superscript indicates the element of the inverse matrix. The F statistic can then be based on this measure, $F = [R_k^2/(L - 1)]/[(1 - R_k^2)/(n - L)]$ assuming that $\mathbf{Z}$ contains a constant term.

It is worth noting that the test for weak instruments is not a specification test, nor is it a constructive test for building the model. Rather, it is a strategy for helping the researcher avoid basing inference on unreliable statistics whose properties are not well represented by the familiar asymptotic results, for example, distributions under assumed null model specifications. Several extensions are of interest. Other statistical procedures are proposed in Hahn and Hausman (2002) and Kleibergen (2002).

The stark results of this section call the IV estimator into question. In a fairly narrow circumstance, an alternative estimator is the "moment"-free LIML estimator discussed in Section 8.4.3. Another, perhaps somewhat unappealing, approach is to retreat to least squares. The OLS estimator is not without virtue. The asymptotic variance of the OLS estimator,

$$\text{Asy.Var}[\mathbf{b}_{\text{LS}}] = (\sigma^2/n)\mathbf{Q}_{\mathbf{XX}}^{-1},$$

is unambiguously smaller than the asymptotic variance of the IV estimator,

$$\text{Asy.Var}[\mathbf{b}_{\text{IV}}] = (\sigma^2/n)(\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}})^{-1}.$$

(The proof is considered in the exercises.) Given the preceding results, it could be far smaller. The OLS estimator is inconsistent, however,

$$\text{plim } \mathbf{b}_{\text{LS}} - \boldsymbol{\beta} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma},$$

[see (8-4)]. By a mean squared error comparison, it is unclear whether the OLS estimator with

$$M(\mathbf{b}_{\text{LS}}|\boldsymbol{\beta}) = (\sigma^2/n)\mathbf{Q}_{\mathbf{XX}}^{-1} + \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Q}_{\mathbf{XX}}^{-1},$$

or the IV estimator, with

$$M(\mathbf{b}_{\text{IV}}|\boldsymbol{\beta}) = (\boldsymbol{\sigma}^2/n)(\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}})^{-1},$$

is more precise. The natural recourse in the face of weak instruments is to drop the endogenous variable from the model or improve the instrument set. Each of these is a specification issue. Strictly in terms of estimation strategy within the framework of the data and specification in hand, there is scope for OLS to be the preferred strategy.

8.8 MEASUREMENT ERROR

Thus far, it has been assumed (at least implicitly) that the data used to estimate the parameters of our models are true measurements on their theoretical counterparts. In practice, this situation happens only in the best of circumstances. All sorts of measurement problems creep into the data that must be used in our analyses. Even

carefully constructed survey data do not always conform exactly to the variables the analysts have in mind for their regressions. Aggregate statistics such as GDP are only estimates of their theoretical counterparts, and some variables, such as depreciation, the services of capital, and "the interest rate," do not even exist in an agreed-upon theory. At worst, there may be no physical measure corresponding to the variable in our model; intelligence, education, and permanent income are but a few examples. Nonetheless, they all have appeared in very precisely defined regression models.

8.8.1 LEAST SQUARES ATTENUATION

In this section, we examine some of the received results on regression analysis with badly measured data. The biases introduced by measurement error can be rather severe. There are almost no known finite-sample results for the models of measurement error; nearly all the results that have been developed are asymptotic.[22] The following presentation will use a few simple asymptotic results for the classical regression model.

The simplest case to analyze is that of a regression model with a single regressor and no constant term. Although this case is admittedly unrealistic, it illustrates the essential concepts, and we shall generalize it presently. Assume that the model,

$$y^* = \beta x^* + \varepsilon, \tag{8-26}$$

conforms to all the assumptions of the classical normal regression model. If data on y^* and x^* were available, then β would be estimable by least squares. Suppose, however, that the observed data are only imperfectly measured versions of y^* and x^*. In the context of an example, suppose that y^* is ln(output/labor) and x^* is ln(capital/labor). Neither factor input can be measured with precision, so the observed y and x contain errors of measurement. We assume that

$$y = y^* + v \quad \text{with } v \sim N[0, \sigma_v^2], \tag{8-27a}$$

$$x = x^* + u \quad \text{with } u \sim N[0, \sigma_u^2]. \tag{8-27b}$$

Assume, as well, that u and v are independent of each other and of y^* and x^*. (As we shall see, adding these restrictions is not sufficient to rescue a bad situation.)

As a first step, insert (8-27a) into (8-26), assuming for the moment that only y^* is measured with error,

$$y = \beta x^* + \varepsilon + v = \beta x^* + \varepsilon'.$$

This result still conforms to the assumptions of the classical regression model. As long as the regressor is measured properly, measurement error on the dependent variable can be absorbed in the disturbance of the regression and ignored. To save some cumbersome notation, therefore, we shall henceforth assume that the measurement error problems concern only the independent variables in the model.

Consider, then, the regression of y on the observed x. By substituting (8-27b) into (8-26), we obtain

$$y = \beta x + [\varepsilon - \beta u] = \beta x + w. \tag{8-28}$$

[22] See, for example, Imbens and Hyslop (2001).

Because x equals $x^* + u$, the regressor in (8-28) is correlated with the disturbance,

$$\text{Cov}[x, w] = \text{Cov}[x^* + u, \varepsilon - \beta u] = -\beta\sigma_u^2. \tag{8-29}$$

This result violates one of the central assumptions of the classical model, so we can expect the least squares estimator,

$$b = \frac{(1/n)\sum_{i=1}^{n} x_i y_i}{(1/n)\sum_{i=1}^{n} x_i^2},$$

to be inconsistent. To find the probability limits, insert (8-26) and (8-27b) and use the Slutsky theorem,

$$\text{plim } b = \frac{\text{plim}(1/n)\sum_{i=1}^{n}(x_i^* + u_i)(\beta x_i^* + \varepsilon_i)}{\text{plim}(1/n)\sum_{i=1}^{n}(x_i^* + u_i)^2}.$$

Because x^*, ε, and u are mutually independent, this equation reduces to

$$\text{plim } b = \frac{\beta Q^*}{Q^* + \sigma_u^2} = \frac{\beta}{1 + \sigma_u^2/Q^*}, \tag{8-30}$$

where $Q^* = \text{plim}(1/n)\sum_i x_i^{*2}$. As long as σ_u^2 is positive, b is inconsistent, with a persistent bias toward zero. Clearly, the greater the variability in the measurement error, the worse the bias. The effect of biasing the coefficient toward zero is called **attenuation**.

In a multiple regression model, matters only get worse. Suppose, to begin, we assume that $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$, allowing every observation on every variable to be measured with error. The extension of the earlier result is

$$\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) = \mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}, \quad \text{and} \quad \text{plim}\left(\frac{\mathbf{X}'\mathbf{y}}{n}\right) = \mathbf{Q}^*\boldsymbol{\beta}.$$

Hence,

$$\text{plim } \mathbf{b} = [\mathbf{Q}^* + \Sigma_{uu}]^{-1}\mathbf{Q}^*\boldsymbol{\beta} = \boldsymbol{\beta} - [\boldsymbol{Q}^* + \boldsymbol{\Sigma}_{uu}]^{-1}\boldsymbol{\Sigma}_{uu}\boldsymbol{\beta}. \tag{8-31}$$

This probability limit is a mixture of all the parameters in the model. In the same fashion as before, bringing in outside information could lead to identification. The amount of information necessary is extremely large, however, and this approach is not particularly promising.

It is common for only a single variable to be measured with error. One might speculate that the problems would be isolated to the single coefficient. Unfortunately, this situation is not the case. For a single bad variable—assume that it is the first—the matrix $\boldsymbol{\Sigma}_{uu}$ is of the form

$$\boldsymbol{\Sigma}_{uu} = \begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

It can be shown that for this special case,

$$\text{plim } b_1 = \frac{\beta_1}{1 + \sigma_u^2 q^{*11}}, \tag{8-32a}$$

[note the similarity of this result to (8-30)], and, for $k \neq 1$,

$$\text{plim } b_k = \beta_k - \beta_1\left[\frac{\sigma_u^2 q^{*k1}}{1 + \sigma_u^2 q^{*11}}\right], \tag{8-32b}$$

where q^{*k1} is the $(k,1)$th element in $(\mathbf{Q}^*)^{-1}$.[23] This result depends on several unknowns and cannot be estimated. The coefficient on the badly measured variable is still biased toward zero. The other coefficients are all biased as well, although in unknown directions. A badly measured variable contaminates all the least squares estimates.[24] If more than one variable is measured with error, there is very little that can be said.[25] Although expressions can be derived for the biases in a few of these cases, they generally depend on numerous parameters whose signs and magnitudes are unknown and, presumably, unknowable.

8.8.2 INSTRUMENTAL VARIABLES ESTIMATION

An alternative set of results for estimation in this model (and numerous others) is built around the method of instrumental variables. Consider once again the errors in variables model in (8-26) and (8-27a,b). The parameters, β, σ_ε^2, q^*, and σ_u^2 are not identified in terms of the moments of x and y. Suppose, however, that there exists a variable z such that z is correlated with x^* but not with u. For example, in surveys of families, income is notoriously badly reported, partly deliberately and partly because respondents often neglect some minor sources. Suppose, however, that one could determine the total amount of checks written by the head(s) of the household. It is quite likely that this z would be highly correlated with income, but perhaps not significantly correlated with the errors of measurement. If $\text{Cov}[x^*, z]$ is not zero, then the parameters of the model become estimable, as

$$\text{plim } \frac{(1/n)\sum_i y_i z_i}{(1/n)\sum_i x_i z_i} = \frac{\beta\,\text{Cov}[x^*, z]}{\text{Cov}[x^*, z]} = \beta. \tag{8-33}$$

The special case when the instrumental variable is binary produces a useful result. If z_i is a dummy variable such that $\bar{x}_{|z=1} - \bar{x}_{|z=0}$ is not zero—that is, the instrument is relevant (see Section 8.2), then the estimator in (8-33) is

$$\hat{\beta} = \frac{\bar{y}_{|z=1} - \bar{y}_{|z=0}}{\bar{x}_{|z=1} - \bar{x}_{|z=0}}.$$

A proof of the result is given in Example 8.2.[26] This is called the Wald (1940) estimator.

For the general case, $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$, suppose that there exists a matrix of variables $\mathbf{Z}$ that is not correlated with the disturbances or the measurement error,

[23] Use (A-66) to invert $[\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}] = [\mathbf{Q}^* + (\sigma_u\mathbf{e}_1)(\sigma_u\mathbf{e}_1)']$, where $\mathbf{e}_1$ is the first column of a $K \times K$ identity matrix. The remaining results are then straightforward.

[24] This point is important to remember when the presence of measurement error is suspected.

[25] Some firm analytic results have been obtained by Levi (1973), Theil (1961), Klepper and Leamer (1983), Garber and Klepper (1980), Griliches (1986), and Cragg (1997).

[26] The proof in Example 8.2 is given for a dependent variable that is also binary. However, the proof is generic, and extends without modification to this case.

but is correlated with regressors, $\mathbf{X}$. Then the instrumental variables estimator, based on $\mathbf{Z}$, $\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$, is consistent and asymptotically normally distributed with asymptotic covariance matrix that is estimated with

$$\text{Est.Asy.Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2[\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{Z}'\mathbf{Z}][\mathbf{X}'\mathbf{Z}]^{-1}. \quad \textbf{(8-34)}$$

For more general cases, Theorem 8.1 and the results in Section 8.3 apply.

8.8.3 PROXY VARIABLES

In some situations, a variable in a model simply has no observable counterpart. Education, intelligence, ability, and like factors are perhaps the most common examples. In this instance, unless there is some observable indicator for the variable, the model will have to be treated in the framework of missing variables. Usually, however, such an indicator can be obtained; for the factors just given, years of schooling and test scores of various sorts are familiar examples. The usual treatment of such variables is in the measurement error framework. If, for example,

$$\text{income} = \beta_1 + \beta_2 \,\text{education} + \varepsilon,$$

and

$$\text{years of schooling} = \text{education} + u,$$

then the model of Section 8.8.1 applies. The only difference here is that the true variable in the model is "latent." No amount of improvement in reporting or measurement would bring the proxy closer to the variable for which it is proxying.

The preceding is a pessimistic assessment, perhaps more so than necessary. Consider a **structural model**,

$$\textit{Earnings} = \beta_1 + \beta_2\, \textit{Experience} + \beta_3\, \textit{Industry} + \beta_4\, \textit{Ability} + \varepsilon.$$

Ability is unobserved, but suppose that an indicator, say, *IQ*, is. If we suppose that *IQ* is related to *Ability* through a relationship such as

$$IQ = \alpha_1 + \alpha_2\, \textit{Ability} + v,$$

then we may solve the second equation for *Ability* and insert it in the first to obtain the **reduced form equation**,

$$\textit{Earnings} = (\beta_1 - \beta_4\alpha_1/\alpha_2) + \beta_2\, \textit{Experience} + \beta_3\, \textit{Industry} + (\beta_4/\alpha_2)IQ + (\varepsilon - v\beta_4/\alpha_2).$$

This equation is intrinsically linear and can be estimated by least squares. We do not have consistent estimators of β_1 and β_4, but we do have them for the coefficients of interest, β_2 and β_3. This would appear to solve the problem. We should note the essential ingredients; we require that the **indicator**, *IQ*, not be related to the other variables in the model, and we also require that v not be correlated with any of the variables. (A perhaps obvious additional requirement is that the proxy not provide information in the regression that would not be provided by the missing variable if it were observed. In the context of the example, this would require that $E[\textit{Earnings} \,|\, \textit{Experience}, \textit{Industry}, \textit{Ability}, IQ] = E[\textit{Earnings} \,|\, \textit{Experience}, \textit{Industry}, \textit{Ability}]$.) In this instance, some of the parameters of the structural model are identified in terms of observable data. Note, though, that *IQ* is not a proxy variable; it is an

indicator of the latent variable, *Ability*. This form of modeling has figured prominently in the education and educational psychology literature. Consider in the preceding small model how one might proceed with not just a single indicator, but say with a battery of test scores, all of which are indicators of the same latent ability variable.

It is to be emphasized that a proxy variable is not an instrument (or the reverse). Thus, in the instrumental variables framework, it is implied that we do not regress $\mathbf{y}$ on $\mathbf{Z}$ to obtain the estimates. To take an extreme example, suppose that the full model was

$$\begin{aligned}\mathbf{y} &= \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon},\\ \mathbf{X} &= \mathbf{X}^* + \mathbf{U},\\ \mathbf{Z} &= \mathbf{X}^* + \mathbf{W}.\end{aligned}$$

That is, we happen to have two badly measured estimates of $\mathbf{X}^*$. The parameters of this model can be estimated without difficulty if $\mathbf{W}$ is uncorrelated with $\mathbf{U}$ and $\mathbf{X}^*$, *but not by regressing* $\mathbf{y}$ *on* $\mathbf{Z}$. The instrumental variables technique is called for.

When the model contains a variable such as education or ability, the question that naturally arises is, If interest centers on the other coefficients in the model, why not just discard the problem variable?[27] This method produces the familiar problem of an omitted variable, compounded by the least squares estimator in the full model being inconsistent anyway. Which estimator is worse? McCallum (1972) and Wickens (1972) show that the asymptotic bias (actually, degree of inconsistency) is worse if the proxy is omitted, even if it is a bad one (has a high proportion of measurement error). This proposition neglects, however, the precision of the estimates. Aigner (1974) analyzed this aspect of the problem and found, as might be expected, that it could go either way. He concluded, however, that "there is evidence to broadly support use of the proxy."

Example 8.12 Income and Education in a Study of Twins

The traditional model used in labor economics to study the effect of education on income is an equation of the form

$$y_i = \beta_1 + \beta_2\,\text{age}_i + \beta_3\,\text{age}_i^2 + \beta_4\,\text{education}_i + \mathbf{x}_i'\boldsymbol{\beta}_5 + \varepsilon_i,$$

where y_i is typically a wage or yearly income (perhaps in log form) and $\mathbf{x}_i$ contains other variables, such as an indicator for sex, region of the country, and industry. The literature contains discussion of many possible problems in estimation of such an equation by least squares using measured data. Two of them are of interest here:

1. Although "education" is the variable that appears in the equation, the data available to researchers usually include only "years of schooling." This variable is a proxy for education, so an equation fit in this form will be tainted by this problem of measurement error. Perhaps surprisingly so, researchers also find that reported data on years of schooling are themselves subject to error, so there is a second source of measurement error. For the present, we will not consider the first (much more difficult) problem.
2. Other variables, such as "ability"—we denote these μ_i—will also affect income and are surely correlated with education. If the earnings equation is estimated in the form shown above, then the estimates will be further biased by the absence of this "omitted variable." For reasons we will explore in Chapter 19, this bias has been called the **selectivity effect** in recent studies.

[27] This discussion applies to the measurement error and latent variable problems equally.

Simple cross-section studies will be considerably hampered by these problems. But, in a study of twins, Ashenfelter and Krueger (1994) analyzed a data set that allowed them, with a few simple assumptions, to ameliorate these problems.[28]

Annual "twins festivals" are held at many places in the United States. The largest is held in Twinsburg, Ohio. The authors interviewed about 500 individuals over the age of 18 at the August 1991 festival. Using pairs of twins as their observations enabled them to modify their model as follows: Let (y_{ij}, A_{ij}) denote the earnings and age for twin j, $j = 1, 2$, for pair i. For the education variable, only self-reported "schooling" data, S_{ij}, are available. The authors approached the measurement problem in the schooling variable, S_{ij}, by asking each twin how much schooling he or she had and how much schooling his or her sibling had. Denote reported schooling *by* sibling *m of* sibling j by $S_{ij}(m)$. So, the self-reported years of schooling of twin 1 is $S_{i1}(1)$. When asked how much schooling twin 1 has, twin 2 reports $S_{i1}(2)$. The measurement error model for the schooling variable is

$$S_{ij}(m) = S_{ij} + u_{ij}(m), \quad j, m = 1, 2, \quad \text{where } S_{ij} = \text{"true' schooling for twin } j \text{ of pair } i.$$

We assume that the two sources of measurement error, $u_{ij}(m)$, are uncorrelated and they and S_{ij} have zero means. Now, consider a simple bivariate model such as the one in (8-26),

$$y_{ij} = \beta S_{ij} + \varepsilon_{ij}.$$

As we saw earlier, a least squares estimate of β using the reported data will be attenuated,

$$\text{plim } b = \frac{\beta \times \text{Var}[S_{ij}]}{\text{Var}[S_{ij}] + \text{Var}[u_{ij}(j)]} = \beta q.$$

(Because there is no natural distinction between twin 1 and twin 2, the assumption that the variances of the two measurement errors are equal is innocuous.) The factor q is sometimes called the **reliability ratio**. In this simple model, if the reliability ratio were known, then β could be consistently estimated. In fact, the construction of this model allows just that. Because the two measurement errors are uncorrelated,

$$\begin{aligned}\text{Corr}[S_{i1}(1), S_{i1}(2)] &= \text{Corr}[S_{i2}(1), S_{i2}(2)] \\ &= \frac{\text{Var}[S_{i1}]}{\{\{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(1)]\} \times \{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(2)]\}\}^{1/2}} = q.\end{aligned}$$

In words, the correlation between the two reported education attainments measures the reliability ratio. The authors obtained values of 0.920 and 0.877 for 298 pairs of identical twins and 0.869 and 0.951 for 92 pairs of fraternal twins, thus providing a quick assessment of the extent of measurement error in their schooling data.

The earnings equation is a multiple regression, so this result is useful for an overall assessment of the problem, but the numerical values are not sufficient to undo the overall biases in the least squares regression coefficients. An instrumental variables estimator was used for that purpose. The estimating equation for $y_{ij} = \ln \mathit{Wage}_{ij}$ with the least squares (OLS) and instrumental variable (IV) estimates is as follows:

$y_{ij} =$	$\beta_1 +$	$\beta_2 \text{age}_i +$	$\beta_3 \text{age}_i^2 +$	$\beta_4 S_{ij}(j) +$	$\beta_5 S_{im}(m) +$	$\beta_6 \text{sex}_i +$	$\beta_7 \text{race}_i + \varepsilon_{ij}$
LS		(0.088)	(−0.087)	(0.084)		(0.204)	(−0.410)
IV		(0.088)	(−0.087)	(0.116)	(−0.037)	(0.206)	(−0.428).

[28]Other studies of twins and siblings include Bound, Chorkas, Haskel, Hawkes, and Spector (2003). Ashenfelter and Rouse (1998), Ashenfelter and Zimmerman (1997), Behrman and Rosengweig (1999), Isacsson (1999), Miller, Mulvey, and Martin (1995), Rouse (1999), and Taubman (1976).

In the equation, S_{ij} (j) is the person's report of his or her own years of schooling and S_{im} (m) is the sibling's report of the sibling's own years of schooling. The problem variable is schooling. To obtain a consistent estimator, the method of instrumental variables was used, using each sibling's report of the other sibling's years of schooling as a pair of instrumental variables. The estimates reported by the authors are shown below the equation. (The constant term was not reported, and for reasons not given, the second schooling variable was not included in the equation when estimated by LS.) This preliminary set of results is presented to give a comparison to other results in the literature. The age, schooling, and gender effects are comparable with other received results, whereas the effect of race is vastly different, −40%, here compared with a typical value of +9% in other studies. The effect of using the instrumental variable estimator on the estimates of β_4 is of particular interest. Recall that the reliability ratio was estimated at about 0.9, which suggests that the IV estimate would be roughly 11% higher (1/0.9). Because this result is a multiple regression, that estimate is only a crude guide. The estimated effect shown above is closer to 38%.

The authors also used a different estimation approach. Recall the issue of selection bias caused by unmeasured effects. The authors reformulated their model as

$$y_{ij} = \beta_1 + \beta_2 \text{age}_i + \beta_3 \text{age}_i^2 + \beta_4 S_{ij}(j) + \beta_6 \text{sex}_i + \beta_7 \text{race}_i + \mu_i + \varepsilon_{ij}.$$

Unmeasured latent effects, such as "ability," are contained in μ_i. Because μ_i is not observable but is, it is assumed, correlated with other variables in the equation, the least squares regression of y_{ij} on the other variables produces a biased set of coefficient estimates.[29] The difference between the two earnings equations is

$$y_{i1} - y_{i2} = \beta_4[S_{i1}(1) - S_{i2}(2)] + \varepsilon_{i1} - \varepsilon_{i2}.$$

This equation removes the latent effect but, it turns out, worsens the measurement error problem. As before, β_4 can be estimated by instrumental variables. There are two instrumental variables available, $S_{i2}(1)$ and $S_{i1}(2)$. (It is not clear in the paper whether the authors used the two separately or the difference of the two.) The least squares estimate is 0.092, which is comparable to the earlier estimate. The instrumental variable estimate is 0.167, which is nearly 82% higher. The two reported standard errors are 0.024 and 0.043, respectively. With these figures, it is possible to carry out Hausman's test,

$$H = \frac{(0.167 - 0.092)^2}{0.043^2 - 0.024^2} = 4.418.$$

The 95% critical value from the chi-squared distribution with one degree of freedom is 3.84, so the hypothesis that the LS estimator is consistent would be rejected. The square root of H, 2.102, would be treated as a value from the standard normal distribution, from which the critical value would be 1.96. The authors reported a t statistic for this regression of 1.97.

8.9 NONLINEAR INSTRUMENTAL VARIABLES ESTIMATION

In Section 8.2, we extended the linear regression model to allow for the possibility that the regressors might be correlated with the disturbances. The same problem can arise in nonlinear models. The consumption function estimated in Example 7.4 is almost surely

[29] This is a "fixed effects model"—see Section 11.4. The assumption that the latent effect, *ability*, is common between the twins and fully accounted for is a controversial assumption that ability is accounted for by *nature* rather than *nurture*. A search of the Internet on the subject of the "nature versus nurture debate" will turn up millions of citations. We will not visit the subject here.

a case in point. In this section, we will extend the method of instrumental variables to nonlinear regression models.

In the nonlinear model,

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,$$

the covariates $\mathbf{x}_i$ may be correlated with the disturbances. We would expect this effect to be transmitted to the pseudoregressors, $\mathbf{x}_i^0 = \partial h(\mathbf{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$. If so, then the results that we derived for the linearized regression would no longer hold. Suppose that there is a set of variables $[\mathbf{z}_1, \ldots, \mathbf{z}_L]$ such that

$$\operatorname{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \tag{8-35}$$

and

$$\operatorname{plim}(1/n)\mathbf{Z}'\mathbf{X}^0 = \mathbf{Q}_{\mathbf{zx}}^0 \neq \mathbf{0},$$

where $\mathbf{X}^0$ is the matrix of pseudoregressors in the linearized regression, evaluated at the true parameter values. If the analysis that we used for the linear model in Section 8.3 can be applied to this set of variables, then we will be able to construct a consistent estimator for $\boldsymbol{\beta}$ using the instrumental variables. As a first step, we will attempt to replicate the approach that we used for the linear model. The linearized regression model is given in (7-30),

$$\mathbf{y} = \mathbf{h}(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \approx \mathbf{h}^0 + \mathbf{X}^0(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \boldsymbol{\varepsilon},$$

or

$$\mathbf{y}^0 \approx \mathbf{X}^0\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y}^0 = \mathbf{y} - \mathbf{h}^0 + \mathbf{X}^0\boldsymbol{\beta}^0.$$

For the moment, we neglect the approximation error in linearizing the model. In (8-35), we have assumed that

$$\operatorname{plim}(1/n)\mathbf{Z}'\mathbf{y}^0 = \operatorname{plim}(1/n)\mathbf{Z}'\mathbf{X}^0\boldsymbol{\beta}. \tag{8-36}$$

Suppose, as we assumed before, that there are the same number of instrumental variables as there are parameters, that is, columns in $\mathbf{X}^0$. (*Note:* This number need not be the number of variables.) Then the "estimator" used before is suggested,

$$\mathbf{b}_{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X}^0)^{-1}\mathbf{Z}'\mathbf{y}^0. \tag{8-37}$$

The logic is sound, but there is a problem with this estimator. The unknown parameter vector $\boldsymbol{\beta}$ appears on both sides of (8-36). We might consider the approach we used for our first solution to the nonlinear regression model, that is, with some initial estimator in hand, iterate back and forth between the instrumental variables regression and recomputing the pseudoregressors until the process converges to the fixed point that we seek. Once again, the logic is sound, and in principle, this method does produce the estimator we seek.

If we add to our preceding assumptions

$$\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{\mathbf{zz}}],$$

then we will be able to use the same form of the asymptotic distribution for this estimator that we did for the linear case. Before doing so, we must fill in some gaps in the preceding. First, despite its intuitive appeal, the suggested procedure for finding the estimator is very unlikely to be a good algorithm for locating the estimates. Second, we do not wish to limit ourselves to the case in which we have the same number of instrumental variables as parameters. So, we will consider the problem in general terms. The estimation criterion for nonlinear instrumental variables is a quadratic form,

$$\begin{aligned}\text{Min}_{\boldsymbol{\beta}}\, S(\boldsymbol{\beta}) &= \tfrac{1}{2}\{[\mathbf{y} - \mathbf{h}(\mathbf{X}, \boldsymbol{\beta})]'\mathbf{Z}\}(\mathbf{Z}'\mathbf{Z})^{-1}\{\mathbf{Z}'[\mathbf{y} - \mathbf{h}(\mathbf{X}, \boldsymbol{\beta})]\} \\ &= \tfrac{1}{2}\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}).^{30}\end{aligned} \tag{8-38}$$

The first-order conditions for minimization of this weighted sum of squares are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \mathbf{0}. \tag{8-39}$$

This result is the same one we had for the linear model with $\mathbf{X}^0$ in the role of $\mathbf{X}$. This problem, however, is highly nonlinear in most cases, and the repeated least squares approach is unlikely to be effective. But it is a straightforward minimization problem in the frameworks of Appendix E, and instead, we can just treat estimation here as a problem in nonlinear optimization.

We have approached the formulation of this instrumental variables estimator more or less strategically. However, there is a more structured approach. The orthogonality condition,

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0},$$

defines a GMM estimator. With the homoscedasticity and nonautocorrelation assumption, the resultant **minimum distance estimator** produces precisely the criterion function suggested above. We will revisit this estimator in this context in Chapter 13.

With well-behaved *pseudoregressors* and instrumental variables, we have the general result for the nonlinear instrumental variables estimator; this result is discussed at length in Davidson and MacKinnon (2004).

THEOREM 8.2 Asymptotic Distribution of the Nonlinear Instrumental Variables Estimator

With well-behaved instrumental variables and pseudoregressors,

$$\mathbf{b}_{\text{IV}} \overset{a}{\sim} N[\boldsymbol{\beta}, (\sigma^2/n)(\mathbf{Q}^0_{\mathbf{xz}}(\mathbf{Q}_{\mathbf{zz}})^{-1}\mathbf{Q}^0_{\mathbf{zx}})^{-1}].$$

We estimate the asymptotic covariance matrix with

$$\text{Est.Asy.Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2[\hat{\mathbf{X}}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{X}}^0]^{-1},$$

where $\hat{\mathbf{X}}^0$ *is* $\mathbf{X}^0$ *computed using* $\mathbf{b}_{\text{IV}}$.

[30] Perhaps the more natural point to begin the minimization would be $S^0(\boldsymbol{\beta}) = [\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{Z}][\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta})]$. We have bypassed this step because the criterion in (8-38) and the estimator in (8-39) will turn out (following and in Chapter 13) to be a simple yet more efficient GMM estimator.

As a final observation, note that the 2SLS interpretation of the instrumental variables estimator for the linear model still applies here, with respect to the IV estimator. That is, at the final estimates, the first-order conditions (normal equations) imply that

$$\mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}^{0}\boldsymbol{\beta},$$

which says that the estimates satisfy the normal equations for a linear regression of $\mathbf{y}$ (not $\mathbf{y}^0$) on the predictions obtained by regressing the columns of $\mathbf{X}^0$ on $\mathbf{Z}$. The interpretation is not quite the same here, because to compute the predictions of $\mathbf{X}^0$, we must have the estimate of $\boldsymbol{\beta}$ in hand. Thus, this two-stage least squares approach does not show how to compute $\mathbf{b}_{\text{IV}}$; it shows a characteristic of $\mathbf{b}_{\text{IV}}$.

Example 8.13 Instrumental Variables Estimates of the Consumption Function

The consumption function in Example 7.4 was estimated by nonlinear least squares without accounting for the nature of the data that would certainly induce correlation between $\mathbf{X}^0$ and $\boldsymbol{\varepsilon}$. As done earlier, we will reestimate this model using the technique of instrumental variables. For this application, we will use the one-period lagged value of consumption and one- and two-period lagged values of income as instrumental variables. Table 8.6 reports the nonlinear least squares and instrumental variables estimates. Because we are using two periods of lagged values, two observations are lost. Thus, the least squares estimates are not the same as those reported earlier.

The instrumental variable estimates differ considerably from the least squares estimates. The differences can be deceiving, however. Recall that the MPC in the model is $\beta\gamma Y^{\gamma-1}$. The 2000.4 value for *DPI* that we examined earlier was 6634.9. At this value, the instrumental variables and least squares estimates of the MPC are 1.1543 with an estimated standard error of 0.01234 and 1.08406 with an estimated standard error of 0.008694, respectively. These values do differ a bit, but less than the quite large differences in the parameters might have led one to expect. We do note that the IV estimate is considerably greater than the estimate in the linear model, 0.9217 (and greater than one, which seems a bit implausible).

8.10 NATURAL EXPERIMENTS AND THE SEARCH FOR CAUSAL EFFECTS

Econometrics and statistics have historically been taught, understood, and operated under the credo that "correlation is not causation." But, much of the still-growing field of microeconometrics and some of what we have done in this chapter have been advanced as "causal modeling."[31] In the contemporary literature on treatment effects

TABLE 8.6 Nonlinear Least Squares and Instrumental Variable Estimates

	Instrumental Variables		*Least Squares*	
Parameter	***Estimate***	***Standard Error***	***Estimate***	***Standard Error***
α	627.031	26.6063	468.215	22.788
β	0.040291	0.006050	0.0971598	0.01064
γ	1.34738	0.016816	1.24892	0.1220
σ	57.1681	–	49.87998	–
$\mathbf{e}'\mathbf{e}$	650,369.805	–	495,114.490	–

[31]See, for example, Chapter 2 of Cameron and Trivedi (2005), which is entitled "Causal and Noncausal Models" and, especially, Angrist, Imbens, and Rubin (1996), Angrist and Krueger (2001), and Angrist and Pischke (2009, 2010).

and program evaluation, the point of the econometric exercise really is to establish more than mere statistical association—in short, the answer to the question "Does the program *work*?" requests an econometric response more committed than "the data seem to be consistent with that hypothesis." A cautious approach to econometric modeling has nonetheless continued to base its view of "causality" essentially on statistical grounds.[32]

An example of the sort of causal model considered here is an equation such as Krueger and Dale's (1999) model for earnings attainment and elite college attendance,

$$\ln Earnings = \mathbf{x}'\boldsymbol{\beta} + \delta T + \varepsilon,$$

in which δ is the "causal effect" of attendance at an elite college. In this model, T cannot vary autonomously, outside the model. Variation in T is determined partly by the same hidden influences that determine lifetime earnings. Though a causal effect can be attributed to T, measurement of that effect, δ, cannot be done with multiple linear regression. The technique of linear instrumental variables estimation has evolved as a mechanism for disentangling causal influences. As does least squares regression, the method of instrumental variables must be defended against the possibility that the underlying statistical relationships uncovered could be due to "something else." But, when the instrument is the outcome of a "natural experiment," true exogeneity can be claimed. It is this purity of the result that has fueled the enthusiasm of the most strident advocates of this style of investigation. The power of the method lends an inevitability and stability to the findings. This has produced a willingness of contemporary researchers to step beyond their cautious roots.[33] Example 8.14 describes a controversial contribution to this literature. On the basis of a natural experiment, the authors identify a cause-and-effect relationship that would have been viewed as beyond the reach of regression modeling under earlier paradigms.[34]

Example 8.14 Does Television Watching Cause Autism?

The following is the abstract of economists Waldman, Nicholson, and Adilov's (2008) study of autism.[35]

> An extensive literature in medicine investigates the health consequences of early childhood television watching. However, this literature does not address the issue of reverse causation, i.e., does early childhood television watching cause specific health outcomes or do children more likely to have these health outcomes watch more television? This paper uses a natural experiment to investigate the health consequences of early childhood television watching and so is not subject to questions concerning reverse causation. Specifically, we use repeated cross-sectional data from 1972 through 1992 on county-level mental retardation rates, county-level autism rates, and county-level children's cable-television subscription rates to investigate how early childhood television watching affects the prevalence of mental retardation and autism. We find a strong negative correlation

[32]See, among many recent commentaries on this line of inquiry, Heckman and Vytlacil (2007).

[33]See, e.g., Angrist and Pischke (2009, 2010). In reply, Keane (2010, p. 48) opines "What has always bothered me about the 'experimentalist' school is the false sense of certainty it conveys. The basic idea is that if we have a 'really good instrument,' we can come up with 'convincing' estimates of 'causal effects' that are not 'too sensitive to assumptions."

[34]See the symposium in the Spring 2010 *Journal of Economic Perspectives*, Angrist and Pischke (2010), Leamer (2010), Sims (2010), Keane (2010), Stock (2010), and Nevo and Whinston (2010).

[35]Extracts from Waldman, M., Nicholson, S. and Adilov, N., "Positive and Negative Mental Health Consequences of Early Childhood Television Watching," Working Paper w17786, National Bureau of Economic Research, Cambridge, 2012.

> between average county-level cable subscription rates when a birth cohort is below three and subsequent mental retardation diagnosis rates, but a strong positive correlation between the same cable subscription rates and subsequent autism diagnosis rates. Our results thus suggest that early childhood television watching has important positive and negative health consequences.

The authors continue (at page 19),

> "We next examine the role of precipitation on autism diagnoses. One possibility concerning the autism results in Table 5 is that the positive coefficients on the main cable variable may not be due to early childhood television watching being a trigger for autism but rather to some other factor positively associated with precipitation being a trigger. That is, Waldman et al. (2008)[36] find a positive correlation between the precipitation a cohort experiences prior to age three and the cohort's subsequent autism diagnosis rate, where the interpretation put forth in that paper is that there is an environmental trigger for autism positively correlated with precipitation that drives up the autism diagnosis rate when precipitation prior to age three is high. Possibilities include any potential trigger positively associated with indoor activity such as early childhood television watching, which is the focus here, vitamin D deficiency which could be more common when children are indoors more and not exposed to the sun, and any indoor chemical where exposure will be higher when the child spends more time indoors. So one possibility concerning the results in Table 5 is that cable and precipitation are positively correlated and early childhood television watching is not a trigger for autism. In this scenario the positive and statistically significant cable coefficients found in the table would not be due to the positive correlation between cable and early childhood television watching, but rather to one of these other factors being the trigger and the positive coefficients arise because cable, through a correlation with precipitation, is also correlated with this unknown 'other' trigger."

They conclude (on p 30): "We believe our results are sufficiently suggestive of early childhood television watching decreasing mental retardation and increasing autism that clinical studies focused on the health effects of early childhood television watching are warranted. Only a clinical study can show definitively the health effects of early childhood television watching."

The authors add (at page 3), "Although consistent with the hypothesis that early childhood television watching is an important trigger for autism, our first main finding is also consistent with another possibility. Specifically, because precipitation is likely correlated with young children spending more time indoors generally, not just young children watching more television, our first main finding could be due to any indoor toxin. *Therefore, we also employ a second instrumental variable or natural experiment, that is correlated with early childhood television watching but unlikely to be substantially correlated with time spent indoors*." (Emphasis added.) They conclude (on pp. 39–40): "Using the results found in Table 3's pooled cross-sectional analysis of California, Oregon, and Washington's county-level autism rates, we find that if early childhood television watching is the sole trigger driving the positive correlation between autism and precipitation then thirty-eight percent of autism diagnoses are due to the incremental television watching due to precipitation."

---

[36]Waldman, M., S. Nicholson, N. Adilov, and J. Williams, "Autism Prevalence and Precipitation Rates in California, Oregon, and Washington Counties," *Archives of Pediatrics & Adolescent Medicine*, 162, 2008, pp. 1026–1034.

Waldman, Nicholson, and Adilov's (2008)[37] study provoked an intense and widespread response among academics, autism researchers, and the public. Whitehouse (2007), writing in the *Wall Street Journal*, surveyed some of the discussion, which touches upon the methodological implications of the search for "causal effects" in econometric research. The author lamented that the power of techniques involving instrumental variables and natural experiments to uncover causal relationships had emboldened economists to venture into areas far from their traditional expertise, such as the causes of autism [Waldman et al. (2008)].[38]

Example 8.15 Is Season of Birth a Valid Instrument?

Buckles and Hungerman (BH, 2008) list more than 20 studies of long-term economic outcomes that use season of birth as an instrumental variable, beginning with one of the earliest and best-known papers in the "natural experiments" literature, Angrist and Krueger (1991). The assertion of the validity of season of birth as a proper instrument is that family background is unrelated to season of birth, but it is demonstrably related to long-term outcomes such as income and education. The assertion justifies using dummy variables for season of birth as instrumental variables in outcome equations. If, on the other hand, season of birth is correlated with family background, then it will "fail the exclusion restriction in most IV settings where it has been used" (BH, page 2). According to the authors, the randomness of quarter of birth over the population[39] has been taken as a given, without scientific investigation of the claim. Using data from live birth certificates and census data, BH found a numerically modest, but statistically significant relationship between birth dates and family background. They found "women giving birth in the winter look different from other women; they are younger, less educated, and less likely to be married . . . . The fraction of children born to women without a high school degree is about 10% higher (2 percentage points) in January than in May . . . We also document a 10% decline in the fraction of children born to teenagers from January to May." Precisely why there should be such a relationship remains uncertain. Researchers differ (of course) on the numerical implications of BH's finding.[40] But, the methodological implication of their finding is consistent with the observation in Whitehouse's article, that bad instruments can produce misleading results.

8.11 SUMMARY AND CONCLUSIONS

The instrumental variable (IV) estimator, in various forms, is among the most fundamental tools in econometrics. Broadly interpreted, it encompasses most of the estimation methods that we will examine in this book. This chapter has developed the basic results for IV estimation of linear models. The essential departure point is the exogeneity and relevance assumptions that define an instrumental variable. We then analyzed linear IV estimation in the form of the two-stage least squares estimator. With only a few special exceptions related to simultaneous equations models with two variables, almost no finite-sample properties have been established for the IV estimator. (We temper that,

[37]Published as NBER working paper 12632 in 2006.

[38]Whitehouse criticizes the use of proxy variables, e.g., Waltman's use of rainfall patterns for TV viewing. As we have examined in this chapter, an instrumental variable is not a proxy and this mischaracterizes the technique. It remains true, as emphasized by some prominent researchers quoted in the article, that a bad instrument can produce misleading results.

[39]See, for example, Kleibergen (2002).

[40]See Lahart (2009).

however, with the results in Section 8.7 on weak instruments, where we saw evidence that whatever the finite-sample properties of the IV estimator might be, under some well-discernible circumstances, these properties are not attractive.) We then examined the asymptotic properties of the IV estimator for linear and nonlinear regression models. Finally, some cautionary notes about using IV estimators when the instruments are only weakly relevant in the model are examined in Section 8.7.

Key Terms and Concepts

- AttenuationAsymptotic covariance matrix
- Asymptotic distribution
- Attenuation bias
- Attrition bias
- Attrition
- Consistent estimator
- Effect of the treatment on the treated
- Endogenous treatment effect
- Endogenous
- Exogenous
- Identification
- Indicator
- Instrumental variable estimator
- Instrumental variables (IV)
- Limiting distribution
- Minimum distance estimator
- Moment equations
- Natural experiment
- Nonrandom sampling
- Omitted parameter heterogeneity
- Omitted variable bias
- Omitted variables
- Orthogonality conditions
- Overidentification
- Panel data
- Proxy variable
- Random effects
- Reduced form equation
- Relevance
- Reliability ratio
- Sample selection bias
- Selectivity effect
- Simultaneous equations bias
- Simultaneous equations
- Smearing
- Structural equation system
- Structural model
- Structural specification
- Survivorship bias
- Truncation bias
- Two-stage least squares (2SLS)
- Variable addition test
- Weak instruments

Exercises

1. In the discussion of the instrumental variable estimator, we showed that the least squares estimator, $\mathbf{b}_{\text{LS}}$, is biased and inconsistent. Nonetheless, $\mathbf{b}_{\text{LS}}$ does estimate something—plim $\mathbf{b} = \boldsymbol{\theta} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\boldsymbol{\gamma}$. Derive the asymptotic covariance matrix of $\mathbf{b}_{\text{LS}}$ and show that $\mathbf{b}_{\text{LS}}$ is asymptotically normally distributed.
2. For the measurement error model in (8-26) and (8-27), prove that when only x is measured with error, the squared correlation between y and x is less than that between y^* and x^*. (Note the assumption that $y* = y$.) Does the same hold true if y^* is also measured with error?
3. Derive the results in (8-32a) and (8-32b) for the measurement error model. Note the hint in Footnote 4 in Section 8.5.1 that suggests you use result (A-66) when you need to invert

$$[\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}] = [\mathbf{Q}^* + (\sigma_u\mathbf{e}_1)(\sigma_u\mathbf{e}_1)'].$$

4. At the end of Section 8.7, it is suggested that the OLS estimator could have a smaller mean squared error than the 2SLS estimator. Using (8-4), the results of Exercise 1, and Theorem 8.1, show that the result will be true if

$$\mathbf{Q}_{\mathbf{XX}} - \mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}} \gg \frac{1}{(\sigma^2/n) + \boldsymbol{\gamma}'\mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma}}\boldsymbol{\gamma}\boldsymbol{\gamma}'.$$

How can you verify that this is at least possible? The right-hand side is a rank one, nonnegative definite matrix. What can be said about the left-hand side?

5. Consider the linear model, $y_i = \alpha + \beta x_i + \varepsilon_i$, in which $\text{Cov}[x_i, \varepsilon_i] = \gamma \neq 0$. Let z be an exogenous, relevant instrumental variable for this model. Assume, as well, that z is binary—it takes only values 1 and 0. Show the algebraic forms of the LS estimator and the IV estimator for both α and β.
6. This is easy to show. In the expression for $\hat{\mathbf{X}}$, if the kth column in $\mathbf{X}$ is one of the columns in $\mathbf{Z}$, say the lth, then the kth column in $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ will be the lth column of an $L \times L$ identity matrix. This result means that the kth column in $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ will be the lth column in $\mathbf{Z}$, which is the kth column in $\mathbf{X}$.
7. Prove that the control function approach in (8-16) produces the same estimates as 2SLS.
8. Prove that in the control function estimator in (8-16), you can use the predictions, $\mathbf{z}'\mathbf{p}$, instead of the residuals to obtain the same results apart from the sign on the control function itself, which will be reversed.

Applications

1. In Example 8.5, we have suggested a model of a labor market. From the "reduced form" equation given first, you can see the full set of variables that appears in the model—that is the "endogenous variables," ln $Wage_{it}$, and Wks_{it}, and all other exogenous variables. The labor supply equation suggested next contains these two variables and three of the exogenous variables. From these facts, you can deduce what variables would appear in a labor "demand" equation for ln $Wage_{it}$. Assume (for purpose of our example) that ln $Wage_{it}$ is determined by Wks_{it} and the remaining appropriate exogenous variables. (We should emphasize that this exercise is purely to illustrate the computations—the structure here would not provide a theoretically sound model for labor market equilibrium.)
   a. What is the labor demand equation implied?
   b. Estimate the parameters of this equation by OLS and by 2SLS and compare the results. (Ignore the panel nature of the data set. Just pool the data.)
   c. Are the instruments used in this equation relevant? How do you know?

9

THE GENERALIZED REGRESSION MODEL AND HETEROSCEDASTICITY

9.1 INTRODUCTION

In this and the next several chapters, we will extend the multiple regression model to disturbances that violate Assumption A.4 of the classical regression model. The **generalized linear regression model** is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ E[\boldsymbol{\varepsilon}|\mathbf{X}] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] &= \sigma^2\boldsymbol{\Omega} = \boldsymbol{\Sigma}, \end{aligned} \tag{9-1}$$

where $\boldsymbol{\Omega}$ is a positive definite matrix. The covariance matrix is written in the form $\sigma^2\boldsymbol{\Omega}$ at several points so that we can obtain the classical model, $\sigma^2\mathbf{I}$ as a convenient special case.

The two leading cases are **heteroscedasticity** and **autocorrelation**. Disturbances are heteroscedastic when they have different variances. Heteroscedasticity arises in numerous applications, in both cross-section and time-series data. Volatile high-frequency time-series data, such as daily observations in financial markets, are heteroscedastic. Heteroscedasticity appears in cross-section data where the scale of the dependent variable and the explanatory power of the model tend to vary across observations. Microeconomic data, such as expenditure surveys, are typical. Even after accounting for firm size, we expect to observe greater variation in the profits of large firms than in those of small ones. The variance of profits might also depend on product diversification, research and development expenditure, and industry characteristics and therefore might also vary across firms of similar sizes. When analyzing family spending patterns, we find that there is greater variation in expenditure on certain commodity groups among high-income families than low ones due to the greater discretion allowed by higher incomes.

The disturbances are still assumed to be uncorrelated across observations, so $\sigma^2\boldsymbol{\Omega}$ would be

$$\sigma^2\boldsymbol{\Omega} = \sigma^2\begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

Autocorrelation is usually found in time-series data. Economic time series often display a *memory* in that variation around the regression function is not independent from one period

to the next. The seasonally adjusted price and quantity series published by government agencies are examples. Time-series data are usually homoscedastic, so $\sigma^2\mathbf{\Omega}$ might be

$$\sigma^2\mathbf{\Omega} = \sigma^2\begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}.$$

The values that appear off the diagonal depend on the model used for the disturbance. In most cases, consistent with the notion of a fading memory, the values decline as we move away from the diagonal.

A number of other cases considered later will fit in this framework. **Panel data**, consisting of cross sections observed at several points in time, may exhibit both heteroscedasticity and autocorrelation. In the *random effects model*, $y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it}$, with $E[\varepsilon_{it}|\mathbf{x}_{it}] = E[u_i|\mathbf{x}_{it}] = 0$, the implication is that

$$\sigma^2\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Gamma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma} & \cdots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Gamma} \end{bmatrix} \text{ where } \mathbf{\Gamma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \cdots & \sigma_u^2 \\ & & \vdots & \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix}.$$

The specification exhibits autocorrelation. We shall consider it in Chapter 11. Models of spatial autocorrelation, examined in Chapter 11, and multiple equation regression models, considered in Chapter 10, are also forms of the generalized regression model.

This chapter presents some general results for this extended model. We will focus on the model of heteroscedasticity in this chapter and in Chapter 14. A general model of autocorrelation appears in Chapter 20. Chapters 10 and 11 examine in detail other specific types of generalized regression models. We first consider the consequences for the least squares estimator of the more general form of the regression model. This will include devising an appropriate estimation strategy, still based on least squares. We will then examine alternative estimation approaches that can make better use of the characteristics of the model.

9.2 ROBUST LEAST SQUARES ESTIMATION AND INFERENCE

The generalized regression model in (9-1) drops assumption A.4. If $\mathbf{\Omega} \neq \mathbf{I}$, then the disturbances may be heteroscedastic or autocorrelated or both. The least squares estimator is

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i. \tag{9-2}$$

The covariance matrix of the estimator based on (9-1) and (9-2) would be

$$\begin{aligned} Var[\mathbf{b}|\mathbf{X}] &= \frac{1}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\sigma^2\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}\mathbf{x}_i\mathbf{x}_j'}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \\ &= \frac{1}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'(\sigma^2\mathbf{\Omega})\mathbf{X}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}. \end{aligned} \tag{9-3}$$

Based on (9-3), we see that $s^2(\mathbf{X}'\mathbf{X})^{-1}$ would not be the appropriate estimator for the asymptotic covariance matrix for the least squares estimator, $\mathbf{b}$. In Section 4.5, we considered a strategy for estimation of the appropriate covariance matrix, without making explicit assumptions about the form of $\mathbf{\Omega}$, for two cases, heteroscedasticity and *clustering* (which resembles the random effects model suggested in the Introduction). We will add some detail to that discussion for the heteroscedasticity case. Clustering is revisited in Chapter 11.

The matrix $(\mathbf{X}'\mathbf{X}/n)$ is readily computable using the sample data. The complication is the center matrix that involves the unknown $\sigma^2\mathbf{\Omega}$. For estimation purposes, σ^2 is not a separate unknown parameter. We can arbitrarily scale the unknown $\mathbf{\Omega}$, say, by κ, and σ^2 by $1/\kappa$ and obtain the same product. We will remove the indeterminacy by assuming that trace $(\mathbf{\Omega}) = n$, as it is when $\mathbf{\Omega} = \mathbf{I}$. Let $\mathbf{\Sigma} = \sigma^2\mathbf{\Omega}$. It might seem that to estimate $(1/n)\mathbf{X}'\mathbf{\Sigma}\mathbf{X}$, an estimator of $\mathbf{\Sigma}$, which contains $n(n + 1)/2$ unknown parameters, is required. But fortunately (because with only n observations, this would be hopeless), this observation is not quite right. What is required is an estimator of the $K(K + 1)/2$ unknown elements in the center matrix $\mathbf{Q}_* = \text{plim}\dfrac{\mathbf{X}'(\sigma^2\mathbf{\Omega})\mathbf{X}}{n} = \text{plim}\dfrac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{x}_i\mathbf{x}_j'$. The point is that $\mathbf{Q}_*$ is a matrix of sums of squares and cross products that involves σ_{ij} *and* the rows of $\mathbf{X}$. The least squares estimator $\mathbf{b}$ is a consistent estimator of $\boldsymbol{\beta}$, which implies that the least squares residuals e_i are "pointwise" consistent estimators of their population counterparts ε_i. The general approach, then, will be to use $\mathbf{X}$ and $\mathbf{e}$ to devise an estimator of $\mathbf{Q}_*$ for the heteroscedasticity case, $\sigma_{ij} = 0$ when $i \neq j$.

We seek an estimator of $\mathbf{Q}_* = \text{plim}(1/n)\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'$. White (1980, 2001) shows that, under very general conditions, the estimator

$$\mathbf{S}_0 = \frac{1}{n}\sum_{i=1}^{n}e_i^2\mathbf{x}_i\mathbf{x}_i' \tag{9-4}$$

has plim $\mathbf{S}_0 = \mathbf{Q}_*$.[1] The end result is that the **White heteroscedasticity consistent estimator**

$$\begin{aligned}\text{Est.Asy.Var}[\mathbf{b}] &= \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}e_i^2\mathbf{x}_i\mathbf{x}_i'\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\\ &= n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{S}_0(\mathbf{X}'\mathbf{X})^{-1}\end{aligned} \tag{9-5}$$

can be used to estimate the asymptotic covariance matrix of $\mathbf{b}$. This result implies that without actually specifying the type of heteroscedasticity, we can still make appropriate inferences based on the least squares estimator. This implication is especially useful if we are unsure of the precise nature of the heteroscedasticity (which is probably most of the time).

A number of studies have sought to improve on the White estimator for least squares.[2] The asymptotic properties of the estimator are unambiguous, but its usefulness in small samples is open to question. The possible problems stem from the general result that the squared residuals tend to underestimate the squares of the true disturbances.

[1] See also Eicker (1967), Horn, Horn, and Duncan (1975), and MacKinnon and White (1985).

[2] See, for example, MacKinnon and White (1985) and Messer and White (1984).

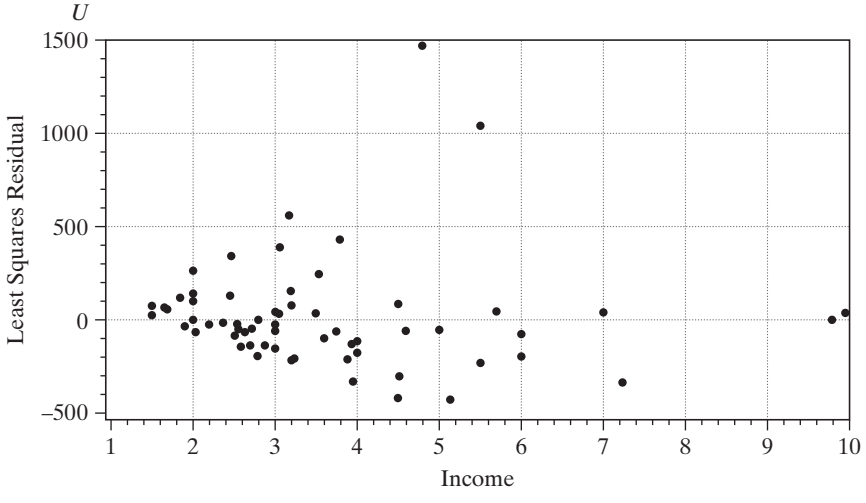
[That is why we use $1/(n - K)$ rather than $1/n$ in computing s^2.] The end result is that in small samples, at least as suggested by some Monte Carlo studies,[3] the White estimator is a bit too optimistic; the matrix is a bit too small, so asymptotic t ratios are a little too large. Davidson and MacKinnon (1993) suggest a number of fixes, which include: (1) scaling up the end result by a factor $n/(n - K)$ and (2) using the squared residual scaled by its true variance, e_i^2/m_{ii}, instead of e_i^2, where $m_{ii} = 1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$.[4] (See Exercise 9.6.b.) On the basis of their study, Davidson and MacKinnon strongly advocate one or the other correction. Their admonition "One should *never* use [the White estimator] because [(2)] *always* performs better" seems a bit strong, but the point is well taken. The use of sharp asymptotic results in small samples can be problematic. The last two rows of Table 9.1 show the recomputed standard errors with these two modifications.

Example 9.1 Heteroscedastic Regression and the White Estimator

The data in Appendix Table F7.3 give monthly credit card expenditure, for 13,444 individuals. A subsample of 100 observations used here is given in Appendix Table F9.1. The estimates are based on the 72 of these 100 observations for which expenditure is positive. Linear regression of monthly expenditure on a constant, age, income and its square, and a dummy variable for home ownership produces the residuals plotted in Figure 9.1. The pattern of the residuals is characteristic of a regression with heteroscedasticity.

Using White's estimator for the regression produces the results in the row labeled "White S. E." in Table 9.1. The adjustment of the least squares results is fairly large, but the Davidson and MacKinnon corrections to White are, even in this sample of only 72 observations, quite modest. The two income coefficients are individually and jointly statistically significant based on the

FIGURE 9.1 Plot of Residuals against Income.



[3] For example, MacKinnon and White (1985).

[4] This is the standardized residual in (4-69). The authors also suggest a third correction, e_i^2/m_{ii}^2, as an approximation to an estimator based on the "jackknife" technique, but their advocacy of this estimator is much weaker than that of the other two. Note that both $n/(n - K)$ and m_{ii} converge to 1 (quickly). The Davidson and MacKinnon results are strictly small sample considerations.

TABLE 9.1 Least Squares Regression Results

	Constant	*Age*	*OwnRent*	*Income*	*Income²*
Sample mean		31.28	0.36	3.369	
Coefficient	−237.15	−3.0818	27.941	234.35	−14.997
Standard error	199.35	5.5147	82.922	80.366	7.4693
t ratio	−1.19	−0.5590	0.337	2.916	−2.0080
White S.E.	212.99	3.3017	92.188	88.866	6.9446
D. and M. (1)	220.79	3.4227	95.566	92.122	7.1991
D. and M. (2)	221.09	3.4477	95.672	92.084	7.1995

$R^2 = 0.243578$, $s = 284.7508$, R^2 without Income and Income$^2 = 0.06393$.

Mean expenditure = \$262.53, Income is × \$10,000
Tests for heteroscedasticity: White = 14.239, Breusch–Pagan = 49.061, Koenker–Bassett = 7.241.

individual *t* ratios and $F(2, 67) = [(0.244 - 0.064)/2]/[0.756/(72 - 5)] = 7.976$. The 1% critical value is 4.94. (Using the internal digits, the value is 7.956.)

The differences in the estimated standard errors seem fairly minor given the extreme heteroscedasticity. One surprise is the decline in the standard error of the age coefficient. The *F* test is no longer available for testing the joint significance of the two income coefficients because it relies on homoscedasticity. A Wald test, however, may be used in any event. The chi-squared test is based on

$$W = (\mathbf{Rb})'[\mathbf{R}(\text{Est.Asy.Var}[\mathbf{b}])\mathbf{R}']^{-1}(\mathbf{Rb}) \quad \text{where } \mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and the estimated asymptotic covariance matrix is the White estimator. The *F* statistic based on least squares is 7.976. The Wald statistic based on the White estimator is 20.604; the 95% critical value for the chi-squared distribution with two degrees of freedom is 5.99, so the conclusion is unchanged.

9.3 PROPERTIES OF LEAST SQUARES AND INSTRUMENTAL VARIABLES

The essential results for the classical model with $E[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$ and $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}$ are developed in Chapters 2 through 6. The least squares estimator

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \tag{9-6}$$

is best linear unbiased (BLU), consistent and asymptotically normally distributed, and if the disturbances are normally distributed, asymptotically efficient. We now consider which of these properties continue to hold in the model of (9-1). To summarize, the least squares estimator retains only some of its desirable properties in this model. It remains unbiased, consistent, and asymptotically normally distributed. It will, however, no longer be efficient and the usual inference procedures based on the *t* and *F* distributions are no longer appropriate.

9.3.1 FINITE-SAMPLE PROPERTIES OF LEAST SQUARES

By taking expectations on both sides of (9-6), we find that if $E[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$, then

$$E[\mathbf{b}] = E_{\mathbf{X}}[E[\mathbf{b}|\mathbf{X}]] = \boldsymbol{\beta} \tag{9-7}$$

and

$$\begin{aligned}\text{Var}[\mathbf{b}|\mathbf{X}] &= E[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'|\mathbf{X}]\\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}]\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\boldsymbol{\Omega})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\\ &= \frac{\sigma^2}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}.\end{aligned} \tag{9-8}$$

Because the variance of the least squares estimator is not $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, statistical inference based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ may be misleading. There is usually no way to know whether $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is larger or smaller than the true variance of $\mathbf{b}$ in (9-8). Without Assumption A.4, the familiar inference procedures based on the F and t distributions will no longer be appropriate even if A.6 (normality of $\boldsymbol{\varepsilon}$) is maintained.

THEOREM 9.1 Finite-Sample Properties of b in the Generalized Regression Model

If the regressors and disturbances are uncorrelated, then the least squares estimator is unbiased in the generalized regression model. With nonstochastic regressors, or conditional on $\mathbf{X}$*, the sampling variance of the least squares estimator is given by (9-8). If the regressors are stochastic, then the unconditional variance is* $E_{\mathbf{X}}[\text{Var}[\mathbf{b}|\mathbf{X}]]$*. From (9-6),* $\mathbf{b}$ *is a linear function of* $\boldsymbol{\varepsilon}$*. Therefore, if* $\boldsymbol{\varepsilon}$ *is normally distributed, then* $\mathbf{b}|\mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}]$.

9.3.2 ASYMPTOTIC PROPERTIES OF LEAST SQUARES

If $\text{Var}[\mathbf{b}|\mathbf{X}]$ converges to zero, then $\mathbf{b}$ is mean square consistent.[5] With well-behaved regressors, $(\mathbf{X}'\mathbf{X}/n)^{-1}$ will converge to a constant matrix, and σ^2/n will converge to zero. But $(\sigma^2/n)(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n)$ need not converge to zero. By writing this product as

$$\frac{\sigma^2}{n}\left(\frac{\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}}{n}\right) = \left(\frac{\sigma^2}{n}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}\mathbf{x}_i\mathbf{x}_j'}{n}\right) \tag{9-9}$$

we see that the matrix is a sum of n^2 terms, divided by n. Thus, the product is a scalar that is $\text{O}(1/n)$ times a matrix that is $\text{O}(n)$ (at least at this juncture) which is $\text{O}(1)$. So, it does appear that if the product in (9-9) does converge, it might converge to a matrix of nonzero constants. In this case, the covariance matrix of the least squares estimator would not converge to zero, and consistency would be difficult to establish. We will examine in some detail the conditions under which the matrix in (9-9) converges to a constant matrix. If it does, then because σ^2/n does vanish, least squares is consistent as well as unbiased.

[5] The argument based on the linear projection in Section 4.4.5 cannot be applied here because, unless $\boldsymbol{\Omega} = \mathbf{I}$, $(\mathbf{X}, \mathbf{y})$ cannot be treated as a random sample from a joint distribution.

Consistency will depend on both $\mathbf{X}$ and $\boldsymbol{\Omega}$. A formula that separates the two components is as follows:[6]

1. The smallest characteristic root of $\mathbf{X}'\mathbf{X}$ increases without bound as $n \to \infty$, which implies that $\text{plim}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$. If the regressors satisfy the Grenander conditions in Table 4.2, then they will meet this requirement.
2. The largest characteristic root of $\boldsymbol{\Omega}$ is finite for all n. For the heteroscedastic model, the variances are the characteristic roots, which requires them to be finite. For models with autocorrelation, the requirements are that the elements of $\boldsymbol{\Omega}$ be finite and that the off-diagonal elements not be too large relative to the diagonal elements. We will examine this condition in Chapter 20.

The least squares estimator is asymptotically normally distributed if the limiting distribution of

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\varepsilon} \tag{9-10}$$

is normal. If $\text{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, then the limiting distribution of the right-hand side is the same as that of

$$\mathbf{v}_{n,LS} = \mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i, \tag{9-11}$$

where $\mathbf{x}_i'$ is a row of $\mathbf{X}$. The question now is whether a central limit theorem can be applied directly to $\mathbf{v}$. If the disturbances are merely heteroscedastic and still uncorrelated, then the answer is generally yes. In fact, we already showed this result in Section 4.4.2 when we invoked the Lindeberg–Feller central limit theorem (D.19) or the Lyapounov theorem (D.20). The theorems allow unequal variances in the sum. *The proof of asymptotic normality* in *Section 4.4.2 is general enough to include this model without modification.* As long as $\mathbf{X}$ is well behaved and the diagonal elements of $\boldsymbol{\Omega}$ are finite and well behaved, the least squares estimator is asymptotically normally distributed, with the covariance matrix given in (9-8). *In the heteroscedastic case, if the variances of ε_i are finite and are not dominated by any single term, so that the conditions of the Lindeberg–Feller central limit theorem apply to $\mathbf{v}_{n,LS}$ in (9-11), then the least squares estimator is asymptotically normally distributed with covariance matrix*

$$\text{Asy.Var}[\mathbf{b}] = \frac{\sigma^2}{n}\mathbf{Q}^{-1}\,\text{plim}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\right)\mathbf{Q}^{-1}. \tag{9-12}$$

For the most general case, asymptotic normality is much more difficult to establish because the sums in (9-11) are not necessarily sums of independent or even uncorrelated random variables. Nonetheless, Amemiya (1985) and Anderson (1971) have established the asymptotic normality of $\mathbf{b}$ in a model of autocorrelated disturbances general enough to include most of the settings we are likely to meet in practice. We will revisit this issue in Chapter 20 when we examine time-series modeling. We can conclude that, except in particularly unfavorable cases, we have the following theorem.

[6] Amemiya (1985, p. 184).

THEOREM 9.2 Asymptotic Properties of b in the Generalized Regression Model

If $\mathbf{Q} = \text{plim}(\mathbf{X}'\mathbf{X}/n)$ *and* $\text{plim}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n)$ *are both finite positive definite matrices, then* $\mathbf{b}$ *is consistent for* $\boldsymbol{\beta}$*. Under the assumed conditions,* $\text{plim}\ \mathbf{b} = \boldsymbol{\beta}$*. If the regressors are sufficiently well behaved and the off-diagonal terms in* $\boldsymbol{\Omega}$ *diminish sufficiently rapidly, then the least squares estimator is asymptotically normally distributed with mean* $\boldsymbol{\beta}$ *and asymptotic covariance matrix given in (9-12).*

9.3.3 HETEROSCEDASTICITY AND Var[b|X]

In the presence of heteroscedasticity, the least squares estimator $\mathbf{b}$ is still unbiased, consistent, and asymptotically normally distributed. The asymptotic covariance matrix is given in (9-12). For this case, with well-behaved regressors,

$$\text{Asy.Var}[\mathbf{b}|\mathbf{X}] = \frac{\sigma^2}{n}\mathbf{Q}^{-1}\left(\text{plim}\frac{1}{n}\sum_{i=1}^{n}\omega_i\mathbf{x}_i\mathbf{x}_i'\right)\mathbf{Q}^{-1}.$$

The mean square consistency of $\mathbf{b}$ depends on the limiting behavior of the matrix

$$\mathbf{Q}_n^* = \frac{1}{n}\sum_{i=1}^{n}\omega_i\mathbf{x}_i\mathbf{x}_i'.$$

If $\mathbf{Q}_n^*$ converges to a positive definite matrix, then as $n \to \infty$, $\mathbf{b}$ will converge to $\boldsymbol{\beta}$ in mean square. Under most circumstances, if ω_i is finite for all i, then we would expect this result to be true. Note that $\mathbf{Q}_n^*$ is a weighted sum of the squares and cross products of $\mathbf{x}$ with weights ω_i/n, which sum to 1. We have already assumed that another weighted sum, $\mathbf{X}'\mathbf{X}/n$, in which the weights are $1/n$, converges to a positive definite matrix $\mathbf{Q}$, so it would be surprising if $\mathbf{Q}_n^*$ did not converge as well. In general, then, we would expect that

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\mathbf{Q}^*\mathbf{Q}^{-1}\right], \quad \text{with } \mathbf{Q}^* = \text{plim}\ \mathbf{Q}_n^*.$$

The conventionally estimated covariance matrix for the least squares estimator $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is inappropriate; the appropriate matrix is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$. It is unlikely that these two would coincide, so the usual estimators of the standard errors are likely to be erroneous. In this section, we consider how erroneous the conventional estimator is likely to be. It is easy to show that if $\mathbf{b}$ is consistent for $\boldsymbol{\beta}$, then $\text{plim}\ s^2 = \text{plim}\ \mathbf{e}'\mathbf{e}/(n-K) = \sigma^2$, assuming $\text{tr}(\boldsymbol{\Omega}) = n$. The normalization $\text{tr}(\boldsymbol{\Omega}) = n$ implies that $\sigma^2 = \bar{\sigma}^2 = (1/n)\Sigma_i\sigma_i^2$ and $\omega_i = \sigma_i^2/\bar{\sigma}^2$. Therefore, the least squares estimator, s^2, converges to $\text{plim}\ \bar{\sigma}^2$, that is, the probability limit of the average variance of the disturbances.

The difference between the conventional estimator and the appropriate (true) covariance matrix for $\mathbf{b}$ is

$$\text{Est.Var}[\mathbf{b}|\mathbf{X}] - \text{Var}[\mathbf{b}|\mathbf{X}] = s^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}. \quad \textbf{(9-13)}$$

In a large sample (so that $s^2 \approx \sigma^2$), this difference is approximately equal to

$$\mathbf{D} = \frac{\sigma^2}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left[\frac{\mathbf{X}'\mathbf{X}}{n} - \frac{\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}}{n}\right]\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}. \quad \textbf{(9-14)}$$

The difference between the two matrices hinges on the bracketed matrix,

$$\boldsymbol{\Delta} = \Sigma_{i=1}^{n}(1/n)\mathbf{x}_i\mathbf{x}_i' - \Sigma_{i=1}^{n}(\omega_i/n)\mathbf{x}_i\mathbf{x}_i' = (1/n)\Sigma_{i=1}^{n}(1-\omega_i)\mathbf{x}_i\mathbf{x}_i', \tag{9-15}$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. These are two weighted averages of the matrices $\mathbf{x}_i\mathbf{x}_i'$ using weights 1 for the first term and ω_i for the second. The scaling $\text{tr}(\boldsymbol{\Omega}) = n$ implies that $\Sigma_i(\omega_i/n) = 1$. Whether the weighted average based on ω_i/n differs much from the one using $1/n$ depends on the weights. If the weights are related to the values in $\mathbf{x}_i$, then the difference can be considerable. If the weights are uncorrelated with $\mathbf{x}_i\mathbf{x}_i'$, however, then the weighted average will tend to equal the unweighted average.

Therefore, the comparison rests on whether the heteroscedasticity is related to any of x_k or $x_j \times x_k$. The conclusion is that, in general: *If the heteroscedasticity is not correlated with the variables in the model, then at least in large samples, the ordinary least squares computations, although not the optimal way to use the data, will not be misleading*.

9.3.4 INSTRUMENTAL VARIABLE ESTIMATION

Chapter 8 considered cases in which the regressors, $\mathbf{X}$, are correlated with the disturbances, $\boldsymbol{\varepsilon}$. The instrumental variables (IV) estimator developed there enjoys a kind of robustness that least squares lacks in that it achieves consistency whether or not $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are correlated, while $\mathbf{b}$ is neither unbiased nor consistent. However, efficiency was not a consideration in constructing the IV estimator. We will reconsider the IV estimator here, but because it is inefficient to begin with, there is little to say about the implications of (9-1) for the efficiency of the estimator. As such, the relevant question for us to consider here would be, essentially, does IV still work in the generalized regression model. Consistency and asymptotic normality will be the useful properties.

The IV/2SLS estimator is

$$\begin{aligned}\mathbf{b}_{\text{IV}} &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\\ &= [\hat{\mathbf{X}}'\mathbf{X}]^{-1}\hat{\mathbf{X}}'\mathbf{y}\\ &= \boldsymbol{\beta} + [\hat{\mathbf{X}}'\mathbf{X}]^{-1}\hat{\mathbf{X}}'\boldsymbol{\varepsilon},\end{aligned} \tag{9-16}$$

where $\mathbf{X}$ is the set of K regressors and $\mathbf{Z}$ is a set of $L \geq K$ instrumental variables. We now consider the extension of Theorem 9.2 to the IV estimator when $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\boldsymbol{\Omega}$. Suppose that $\mathbf{X}$ and $\mathbf{Z}$ are well behaved as assumed in Section 8.2. That is, $(1/n)\mathbf{Z}'\mathbf{Z}$, $(1/n)\mathbf{X}'\mathbf{X}$, and $(1/n)\mathbf{Z}'\mathbf{X}$ all converge to finite nonzero matrices. For convenience let

$$\begin{aligned}\mathbf{Q}_{\mathbf{XX.Z}} &= \text{plim}\left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\\ &= [\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}}]^{-1}\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}.\end{aligned}$$

If $\mathbf{Z}$ is a valid set of instrumental variables, that is, if the second term in (9-16) vanishes asymptotically, then

$$\text{plim}\,\mathbf{b}_{\text{IV}} = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{XX.Z}}\,\text{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta}.$$

The large sample behavior of $\mathbf{b}_{\text{IV}}$ depends on the behavior of

$$\mathbf{v}_{n,\mathrm{IV}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i.$$

This result is exactly the one we analyzed in Section 4.4.2. If the sampling distribution of $\mathbf{v}_n$ converges to a normal distribution, then we will be able to construct the asymptotic distribution for $\mathbf{b}_{\mathrm{IV}}$. This set of conditions is the same that was necessary for $\mathbf{X}$ when we considered $\mathbf{b}$ above, with $\mathbf{Z}$ in place of $\mathbf{X}$. We will rely on the results of Anderson (1971) or Amemiya (1985) that, under very general conditions,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i \xrightarrow{d} N\left[\mathbf{0}, \sigma^2 \operatorname{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)\right].$$

With the other results already in hand, we now have the following.

THEOREM 9.3 Asymptotic Properties of the IV Estimator in the Generalized Regression Model

If the regressors and the instrumental variables are well behaved in the fashions discussed above, then $\mathbf{b}_{\mathrm{IV}}$ *is consistent and asymptotically normally distributed with*

$$\mathbf{b}_{\mathrm{IV}} \overset{a}{\sim} N[\boldsymbol{\beta}, \mathbf{V}_{\mathrm{IV}}],$$

where

$$\mathbf{V}_{\mathrm{IV}} = \frac{\sigma^2}{n}(\mathbf{Q}_{\mathbf{XX.Z}}) \operatorname{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)(\mathbf{Q}'_{\mathbf{XX.Z}}).$$

9.4 EFFICIENT ESTIMATION BY GENERALIZED LEAST SQUARES

Efficient estimation of $\boldsymbol{\beta}$ in the generalized regression model requires knowledge of $\boldsymbol{\Omega}$. To begin, it is useful to consider cases in which $\boldsymbol{\Omega}$ is a known, symmetric, positive definite matrix. This assumption will occasionally be true, though in most models $\boldsymbol{\Omega}$ will contain unknown parameters that must also be estimated. We shall examine this case in Section 9.4.2.

9.4.1 GENERALIZED LEAST SQUARES (GLS)

Because $\boldsymbol{\Omega}$ is a positive definite symmetric matrix, it can be factored into

$$\boldsymbol{\Omega} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}',$$

where the columns of $\mathbf{C}$ are the characteristic vectors of $\boldsymbol{\Omega}$ and the characteristic roots of $\boldsymbol{\Omega}$ are arrayed in the diagonal matrix, $\boldsymbol{\Lambda}$. Let $\boldsymbol{\Lambda}^{1/2}$ be the diagonal matrix with ith diagonal element, $\sqrt{\lambda_i}$, and let $\mathbf{T} = \mathbf{C}\boldsymbol{\Lambda}^{1/2}$. Then, $\boldsymbol{\Omega} = \mathbf{T}\mathbf{T}'$. Also, let $\mathbf{P}' = \mathbf{C}\boldsymbol{\Lambda}^{-1/2}$, so $\boldsymbol{\Omega}^{-1} = \mathbf{P}'\mathbf{P}$. Premultiply the model in (9-1) by $\mathbf{P}$ to obtain

$$\mathbf{P}\mathbf{y} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon}$$

or

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_*. \tag{9-17}$$

The conditional variance of $\boldsymbol{\varepsilon}_*$ is

$$E[\boldsymbol{\varepsilon}_*\boldsymbol{\varepsilon}_*'|\mathbf{X}_*] = \mathbf{P}\sigma^2\boldsymbol{\Omega}\mathbf{P}' = \sigma^2\mathbf{I},$$

so the classical regression model applies to this transformed model. Because $\boldsymbol{\Omega}$ is assumed to be known, $\mathbf{y}_*$ and $\mathbf{X}_*$ are observed data. In the classical model, ordinary least squares is efficient; hence,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* \\ &= (\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{y} \\ &= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}\end{aligned}$$

is the **efficient estimator** of $\boldsymbol{\beta}$. This estimator is the **generalized least squares (GLS)** or Aitken (1935) estimator of $\boldsymbol{\beta}$. This estimator is in contrast to the **ordinary least squares (OLS)** estimator, which uses a *weighting matrix*, $\mathbf{I}$, instead of $\boldsymbol{\Omega}^{-1}$. By appealing to the classical regression model in (9-17), we have the following theorem, which includes the generalized regression model analogs to our results of Chapter 4:

THEOREM 9.4 Properties of the Generalized Least Squares Estimator

If $E[\boldsymbol{\varepsilon}_*|\mathbf{X}_*] = \mathbf{0}$, *then*

$$E[\hat{\boldsymbol{\beta}}|\mathbf{X}_*] = E[(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_*|\mathbf{X}_*] = \boldsymbol{\beta} + E[(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\boldsymbol{\varepsilon}_*|\mathbf{X}_*] = \boldsymbol{\beta}.$$

The GLS estimator $\hat{\boldsymbol{\beta}}$ *is unbiased. This result is equivalent to* $E[\mathbf{P}\boldsymbol{\varepsilon}|\mathbf{P}\mathbf{X}] = \mathbf{0}$, *but because* $\mathbf{P}$ *is a matrix of known constants, we return to the familiar requirement* $E[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$. *The requirement that the regressors and disturbances be uncorrelated is unchanged.*

The GLS estimator is consistent if $\text{plim}(1/n)\mathbf{X}_*'\mathbf{X}_* = \mathbf{Q}_*$, *where* $\mathbf{Q}_*$ *is a finite positive definite matrix. Making the substitution, we see that this implies*

$$\text{plim}[(1/n)\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1} = \mathbf{Q}_*^{-1}. \tag{9-18}$$

We require the transformed data $\mathbf{X}_* = \mathbf{P}\mathbf{X}$, *not the original data* $\mathbf{X}$, *to be well behaved.*[7] *Under the assumption in (9-1), the following hold:*

The GLS estimator is asymptotically normally distributed, with mean $\boldsymbol{\beta}$ *and sampling variance*

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}_*] = \sigma^2(\mathbf{X}_*'\mathbf{X}_*)^{-1} = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}. \tag{9-19}$$

The GLS estimator $\hat{\boldsymbol{\beta}}$ *is the minimum variance linear unbiased estimator in the generalized regression model. This statement follows by applying the Gauss–Markov theorem to the model in (9-17). The result in Theorem 9.5 is* ***Aitken's theorem*** *(1935), and* $\hat{\boldsymbol{\beta}}$ *is sometimes called the Aitken estimator. This broad result includes the Gauss–Markov theorem as a special case when* $\boldsymbol{\Omega} = \mathbf{I}$.

[7] Once again, to allow a time trend, we could weaken this assumption a bit.

For testing hypotheses, we can apply the full set of results in Chapter 5 to the transformed model in (9-17). For testing the J linear restrictions, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the appropriate statistic is

$$F[J, n - K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})}{J} = \frac{(\hat{\boldsymbol{\varepsilon}}_c'\hat{\boldsymbol{\varepsilon}}_c - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/J}{\hat{\sigma}^2},$$

where the residual vector is

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}$$

and

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n - K} = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - K}. \tag{9-20}$$

The constrained GLS residuals, $\hat{\boldsymbol{\varepsilon}}_c = \mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}}_c$, are based on

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}} - [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}).^{8}$$

To summarize, all the results for the classical model, including the usual inference procedures, apply to the transformed model in (9-17).

There is no precise counterpart to R^2 in the generalized regression model. Alternatives have been proposed, but care must be taken when using them. For example, one choice is the R^2 in the transformed regression, (9-17). But this regression need not have a constant term, so the R^2 is not bounded by zero and one. Even if there is a constant term, the transformed regression is a computational device, not the model of interest. That a good (or bad) fit is obtained in the model in (9-17) may be of no interest; the dependent variable in that model, y_*, is different from the one in the model as originally specified. The usual R^2 often suggests that the fit of the model is improved by a correction for heteroscedasticity and degraded by a correction for autocorrelation, but both changes can often be attributed to the computation of y_*. A more appealing fit measure might be based on the residuals from the original model once the GLS estimator is in hand, such as

$$R_G^2 = 1 - \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Like the earlier contender, however, this measure is not bounded in the unit interval. In addition, this measure cannot be reliably used to compare models. The generalized least squares estimator minimizes the **generalized sum of squares**

$$\boldsymbol{\varepsilon}_*'\boldsymbol{\varepsilon}_* = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

not $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. As such, there is no assurance, for example, that dropping a variable from the model will result in a decrease in R_G^2, as it will in R^2. Other goodness-of-fit measures, designed primarily to be a function of the sum of squared residuals (raw or weighted by $\boldsymbol{\Omega}^{-1}$) and to be bounded by zero and one, have been proposed.[9] Unfortunately, they all suffer from at least one of the previously noted shortcomings. The R^2-like measures in

[8] Note that this estimator is the constrained OLS estimator using the transformed data. [See (5-23).]

[9] See Judge et al. (1985, p. 32) and Buse (1973).

this setting are purely descriptive. That being the case, the squared sample correlation between the actual and predicted values, $r^2_{y,\hat{y}} = corr^2(y, \hat{y}) = corr^2(y, \mathbf{x}'\hat{\beta})$, would likely be a useful descriptor. Note, though, that this is not a proportion of variation explained, as is R^2; it is a measure of the agreement of the model predictions with the actual data.

9.4.2 FEASIBLE GENERALIZED LEAST SQUARES (FGLS)

To use the results of Section 9.4.1, $\boldsymbol{\Omega}$ must be known. If $\boldsymbol{\Omega}$ contains unknown parameters that must be estimated, then generalized least squares is not feasible. But with an unrestricted $\boldsymbol{\Omega}$, there are $n(n + 1)/2$ additional parameters in $\sigma^2\boldsymbol{\Omega}$. This number is far too many to estimate with n observations. Obviously, some structure must be imposed on the model if we are to proceed.

The typical problem involves a small set of parameters $\boldsymbol{\alpha}$ such that $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\alpha})$. For example, a commonly used formula in time-series settings is

$$\boldsymbol{\Omega}(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2} \\ & & & & \vdots & \\ \rho^{n-1} & \rho^{n-2} & & \cdots & & 1 \end{bmatrix},$$

which involves only one additional unknown parameter. A model of heteroscedasticity that also has only one new parameter is

$$\sigma_i^2 = \sigma^2 z_i^{\theta} \tag{9-21}$$

for some exogenous variable z. Suppose, then, that $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$. (We consider later how such an estimator might be obtained.) To make GLS estimation feasible, we shall use $\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}(\hat{\boldsymbol{\alpha}})$ instead of the true $\boldsymbol{\Omega}$. The issue we consider here is whether using $\boldsymbol{\Omega}(\hat{\boldsymbol{\alpha}})$ requires us to change any of the results of Section 9.4.1.

It would seem that if plim $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$, then using $\hat{\boldsymbol{\Omega}}$ is asymptotically equivalent to using the true $\boldsymbol{\Omega}$.[10] Let the **feasible generalized least squares** estimator be denoted

$$\hat{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}.$$

Conditions that imply that $\hat{\hat{\boldsymbol{\beta}}}$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}$ are

$$\text{plim}\left[\left(\frac{1}{n}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right) - \left(\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)\right] = \mathbf{0} \tag{9-22}$$

and

$$\text{plim}\left[\left(\frac{1}{\sqrt{n}}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{\varepsilon}\right) - \left(\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}\right)\right] = \mathbf{0}. \tag{9-23}$$

The first of these equations states that if the weighted sum of squares matrix based on the true $\boldsymbol{\Omega}$ converges to a positive definite matrix, then the one based on $\hat{\boldsymbol{\Omega}}$ converges to the same matrix. We are assuming that this is true. In the second condition, if the

[10] This equation is sometimes denoted plim $\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}$. Because $\boldsymbol{\Omega}$ is $n \times n$, it cannot have a probability limit. We use this term to indicate convergence element by element.

transformed regressors are well behaved, then the right-hand-side sum will have a limiting normal distribution. This condition is exactly the one we used in Chapter 4 to obtain the asymptotic distribution of the least squares estimator; here we are using the same results for $\mathbf{X}_*$ and $\boldsymbol{\varepsilon}_*$. Therefore, (9-23) requires the same condition to hold when $\boldsymbol{\Omega}$ is replaced with $\hat{\boldsymbol{\Omega}}$.[11]

These conditions, in principle, must be verified on a case-by-case basis. Fortunately, in most familiar settings, they are met. If we assume that they are, then the FGLS estimator based on $\hat{\boldsymbol{\alpha}}$ has the same **asymptotic properties** as the GLS estimator. This result is extremely useful. Note, especially, the following theorem.

THEOREM 9.5 Efficiency of the FGLS Estimator
An asymptotically efficient FGLS estimator does not require that we have an efficient estimator of $\boldsymbol{\alpha}$; *only a consistent one is required to achieve full efficiency for the FGLS estimator.*

Except for the simplest cases, the **finite-sample properties** and exact distributions of FGLS estimators are unknown. The asymptotic efficiency of FGLS estimators may not carry over to small samples because of the variability introduced by the estimated $\boldsymbol{\Omega}$. Some analyses for the case of heteroscedasticity are given by Taylor (1977). A model of autocorrelation is analyzed by Griliches and Rao (1969). In both studies, the authors find that, over a broad range of parameters, FGLS is more efficient than least squares. But if the departure from the classical assumptions is not too severe, then least squares may be more efficient than FGLS in a small sample.

9.5 HETEROSCEDASTICITY AND WEIGHTED LEAST SQUARES

In the heteroscedastic regression model,

$$\text{Var}[\varepsilon_i|\mathbf{X}] = \sigma_i^2 = \sigma^2\omega_i, \quad i = 1, \ldots, n.$$

This form is an arbitrary scaling which allows us to use a normalization, $\text{trace}(\boldsymbol{\Omega}) = \Sigma_i\omega_i = n$. This makes the classical regression with homoscedastic disturbances a simple special case with $\omega_i = 1, i = 1, \ldots, n$. Intuitively, one might then think of the ωs as weights that are scaled in such a way as to reflect only the variety in the disturbance variances. The scale factor σ^2 then provides the overall scaling of the disturbance process.

We will examine the heteroscedastic regression model, first in general terms, then with some specific forms of the disturbance covariance matrix. Specification tests for heteroscedasticity are considered in Section 9.6. Section 9.6 considers generalized (weighted) least squares, which requires knowledge at least of the form of $\boldsymbol{\Omega}$. Finally, two common applications are examined in Section 9.7.

[11] The condition actually requires only that if the right-hand-side sum has *any* limiting distribution, then the left-hand one has the same one. Conceivably, this distribution might not be the normal distribution, but that seems unlikely except in a specially constructed, theoretical case.

9.5.1 WEIGHTED LEAST SQUARES

The GLS estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}. \tag{9-24}$$

In the most general case, $\text{Var}[\varepsilon_i|\mathbf{X}] = \sigma_i^2 = \sigma^2\omega_i$, $\boldsymbol{\Omega}^{-1}$ is a diagonal matrix whose ith diagonal element is $1/\omega_i$. The GLS estimator is obtained by regressing

$$\mathbf{Py} = \begin{bmatrix} y_1/\sqrt{\omega_1} \\ y_2/\sqrt{\omega_2} \\ \vdots \\ y_n/\sqrt{\omega_n} \end{bmatrix} \quad \text{on} \quad \mathbf{PX} = \begin{bmatrix} \mathbf{x}_1'/\sqrt{\omega_1} \\ \mathbf{x}_2'/\sqrt{\omega_2} \\ \vdots \\ \mathbf{x}_n'/\sqrt{\omega_n} \end{bmatrix}.$$

Applying ordinary least squares to the transformed model, we obtain the **weighted least squares** estimator.

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} w_i\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{n} w_i\mathbf{x}_i y_i\right], \tag{9-25}$$

where $w_i = 1/\omega_i$.[12] The logic of the computation is that observations with smaller variances receive a larger weight in the computations of the sums and therefore have greater influence in the estimates obtained.

9.5.2 WEIGHTED LEAST SQUARES WITH KNOWN Ω

A common specification is that the variance is proportional to one of the regressors or its square. Our earlier example of family expenditures is one in which the relevant variable is usually income. Similarly, in studies of firm profits, the dominant variable is typically assumed to be firm size. If

$$\sigma_i^2 = \sigma^2 x_{ik}^2,$$

then the transformed regression model for GLS is

$$\frac{y}{x_k} = \beta_k + \beta_1\left(\frac{x_1}{x_k}\right) + \beta_2\left(\frac{x_2}{x_k}\right) + \cdots + \frac{\varepsilon}{x_k}. \tag{9-26}$$

If the variance is proportional to x_k instead of x_k^2, then the weight applied to each observation is $1/\sqrt{x_k}$ instead of $1/x_k$.

In (9-26), the coefficient on x_k becomes the constant term. But if the variance is proportional to any power of x_k other than two, then the transformed model will no longer contain a constant, and we encounter the problem of interpreting R^2 mentioned earlier. For example, no conclusion should be drawn if the R^2 in the regression of y/z on $1/z$ and x/z is higher than in the regression of y on a constant and x for any z, including x. The good fit of the weighted regression might be due to the presence of $1/z$ on both sides of the equality.

It is rarely possible to be certain about the nature of the heteroscedasticity in a regression model. In one respect, this problem is only minor. The weighted least squares estimator

[12] The weights are often denoted $w_i = 1/\sigma_i^2$. This expression is consistent with the equivalent $\hat{\boldsymbol{\beta}} = [\mathbf{X}'(\sigma^2\boldsymbol{\Omega})^{-1}\mathbf{X}']^{-1}\mathbf{X}'(\sigma^2\boldsymbol{\Omega})^{-1}\mathbf{y}$. The σ^2s cancel, leaving the expression given previously.

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{n} w_i \mathbf{x}_i y_i\right]$$

is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances. But using the wrong set of weights has two other consequences that may be less benign. First, the improperly weighted least squares estimator is inefficient. This point might be moot if the correct weights are unknown, but the GLS standard errors will also be incorrect. The asymptotic covariance matrix of the estimator

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{9-27}$$

is

$$\text{Asy.Var}[\hat{\boldsymbol{\beta}}] = \sigma^2[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\Omega}\mathbf{V}^{-1}\mathbf{X}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}. \tag{9-28}$$

This result may or may not resemble the usual estimator, which would be the matrix in brackets, and underscores the usefulness of the White estimator in (9-5).

The standard approach in the literature is to use OLS with the White estimator or some variant for the asymptotic covariance matrix. One could argue both flaws and virtues in this approach. In its favor, **robustness to unknown heteroscedasticity** is a compelling virtue. In the clear presence of heteroscedasticity, however, least squares can be inefficient. The question becomes whether using the wrong weights is better than using no weights at all. There are several layers to the question. If we use one of the models mentioned earlier—Harvey's, for example, is a versatile and flexible candidate—then we may use the wrong set of weights and, in addition, estimation of the variance parameters introduces a new source of variation into the slope estimators for the model. However, the weights we use might well be better than none. A heteroscedasticity robust estimator for weighted least squares can be formed by combining (9-27) with the White estimator. The weighted least squares estimator in (9-27) is consistent with any set of weights $\mathbf{V} = \text{diag}[v_1, v_2, \ldots, v_n]$. Its asymptotic covariance matrix can be estimated with

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\left[\sum_{i=1}^{n}\left(\frac{e_i^2}{v_i^2}\right)\mathbf{x}_i\mathbf{x}_i'\right](\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{9-29}$$

Any consistent estimator can be used to form the residuals. The weighted least squares estimator is a natural candidate.

9.5.3 ESTIMATION WHEN Ω CONTAINS UNKNOWN PARAMETERS

The general form of the heteroscedastic regression model has too many parameters to estimate by ordinary methods. Typically, the model is restricted by formulating $\sigma^2\boldsymbol{\Omega}$ as a function of a few parameters, as in $\sigma_i^2 = \sigma^2 x_i^{\alpha}$ or $\sigma_i^2 = \sigma^2(\mathbf{x}_i'\boldsymbol{\alpha})^2$. Write this as $\boldsymbol{\Omega}(\boldsymbol{\alpha})$. FGLS based on a consistent estimator of $\boldsymbol{\Omega}(\boldsymbol{\alpha})$ (meaning a consistent estimator of $\boldsymbol{\alpha}$) is asymptotically equivalent to full GLS. The new problem is that we must first find consistent estimators of the unknown parameters in $\boldsymbol{\Omega}(\boldsymbol{\alpha})$. Two methods are typically used, two-step GLS and maximum likelihood. We consider the two-step estimator here and the maximum likelihood estimator in Chapter 14.

For the heteroscedastic model, the GLS estimator is

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n}\left(\frac{1}{\sigma_i^2}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{n}\left(\frac{1}{\sigma_i^2}\right)\mathbf{x}_i y_i\right]. \tag{9-30}$$

The **two-step estimators** are computed by first obtaining estimates $\hat{\sigma}_i^2$, usually using some function of the ordinary least squares residuals. Then, $\hat{\hat{\boldsymbol{\beta}}}$ uses (9-30) and $\hat{\sigma}_i^2$. The ordinary least squares estimator of $\boldsymbol{\beta}$, although inefficient, is still consistent. As such, statistics computed using the ordinary least squares residuals, $e_i = (y_i - \mathbf{x}_i'\mathbf{b})$, will have the same asymptotic properties as those computed using the true disturbances, $\varepsilon_i = (y_i - \mathbf{x}_i'\boldsymbol{\beta})$. This result suggests a regression approach for the true disturbances and variables $\mathbf{z}_i$ that may or may not coincide with $\mathbf{x}_i$. Now $E[\varepsilon_i^2|\mathbf{z}_i] = \sigma_i^2$, so $\varepsilon_i^2 = \sigma_i^2 + v_i$, where v_i is just the difference between ε_i^2 and its conditional expectation. Because ε_i is unobservable, we would use the least squares residual, for which $e_i = \varepsilon_i - \mathbf{x}_i'(\mathbf{b} - \boldsymbol{\beta}) = \varepsilon_i + u_i$. Then, $e_i^2 = \varepsilon_i^2 + u_i^2 + 2\varepsilon_i u_i$. But, in large samples, as $\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$, terms in u_i will become negligible, so that at least approximately,[13]

$$e_i^2 = \sigma_i^2 + v_i^*. \tag{9-31}$$

The procedure suggested is to treat the variance function as a regression and use the squares or some other functions of the least squares residuals as the dependent variable.[14] For example, if $\sigma_i^2 = \mathbf{z}_i'\boldsymbol{\alpha}$, then a consistent estimator of $\boldsymbol{\alpha}$ will be the least squares slopes, $\mathbf{a}$, in the "model,"

$$e_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} + v_i^*.$$

In this model, v_i^* is both heteroscedastic and autocorrelated, so $\mathbf{a}$ is consistent but inefficient. But consistency is all that is required for asymptotically efficient estimation of $\boldsymbol{\beta}$ using $\boldsymbol{\Omega}(\hat{\boldsymbol{\alpha}})$. It remains to be settled whether improving the estimator of $\boldsymbol{\alpha}$ in this and the other models we will consider would improve the small sample properties of the two-step estimator of $\boldsymbol{\beta}$.[15]

The two-step estimator may be iterated by recomputing the residuals after computing the FGLS estimates and then reentering the computation. The asymptotic properties of the iterated estimator are the same as those of the two-step estimator, however. In some cases, this sort of iteration will produce the maximum likelihood estimator at convergence. Yet none of the estimators based on regression of squared residuals on other variables satisfy the requirement. Thus, iteration in this context provides little additional benefit, if any.

9.6 TESTING FOR HETEROSCEDASTICITY

Tests for heteroscedasticity are based on the following strategy. Ordinary least squares is a consistent estimator of $\boldsymbol{\beta}$ even in the presence of heteroscedasticity. As such, the ordinary least squares residuals will mimic, albeit imperfectly because of sampling variability, the heteroscedasticity of the true disturbances. Therefore, tests designed to detect heteroscedasticity will, in general, be applied to the ordinary least squares residuals.

[13] See Amemiya (1985) and Harvey (1976) for formal analyses.

[14] See, for example, Jobson and Fuller (1980).

[15] Fomby, Hill, and Johnson (1984, pp. 177–186) and Amemiya (1985, pp. 203–207; 1977) examine this model.

9.6.1 WHITE'S GENERAL TEST

To formulate the available tests, it is necessary to specify, at least in rough terms, the nature of the heteroscedasticity. White's (1980) test proposes a general hypothesis of the form

$$H_0 : \sigma_i^2 = E[\varepsilon_i^2 | \mathbf{x}_i] = \sigma^2 \text{ for all } i,$$
$$H_1 : \textit{Not } H_0.$$

A simple operational version of his test is carried out by obtaining nR^2 in the regression of the squared OLS residuals, e_i^2, on a constant and all unique variables contained in $\mathbf{x}$ and $\mathbf{x} \otimes \mathbf{x}$. The statistic has a limiting chi-squared distribution with $P - 1$ degrees of freedom, where P is the number of regressors in the equation, including the constant. An equivalent approach is to use an F test to test the hypothesis that $\boldsymbol{\gamma}_1 = \mathbf{0}$ and $\boldsymbol{\gamma}_2 = \mathbf{0}$ in the regression

$$e_i^2 = \gamma_0 + \mathbf{x}_i'\boldsymbol{\gamma}_1 + (\mathbf{x}_i \otimes \mathbf{x}_i)'\boldsymbol{\gamma}_2 + v_i^*.$$

[As before, $(\mathbf{x}_i \otimes \mathbf{x}_i)$ contains only the unique components.] The **White test** is extremely general. To carry it out, we need not make any specific assumptions about the nature of the heteroscedasticity.

9.6.2 THE LAGRANGE MULTIPLIER TEST

Breusch and Pagan (1979) and Godfrey (1988) present a **Lagrange multiplier test** of the hypothesis that $\sigma_i^2 = \sigma^2 f(\alpha_0 + \boldsymbol{\alpha}'\mathbf{z}_i)$, where $\mathbf{z}_i$ is a vector of independent variables. The disturbance is homoscedastic if $\boldsymbol{\alpha} = \mathbf{0}$. The test can be carried out with a simple regression:

$$\text{LM} = \tfrac{1}{2} \times \text{explained sum of squared residuals in the regression of } e_i^2/(\mathbf{e}'\mathbf{e}/n) \text{ on } (1, \mathbf{z}_i). \quad \textbf{(9-32)}$$

For computational purposes, let $\mathbf{Z}$ be the $n \times P$ matrix of observations on $(1, \mathbf{z}_i)$, and let $\mathbf{g}$ be the vector of observations of $g_i = e_i^2/(\mathbf{e}'\mathbf{e}/n) - 1$. Then $\text{LM} = (1/2)[\mathbf{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{g}]$. Under the null hypothesis of homoscedasticity, LM has a limiting chi-squared distribution with $P - 1$ degrees of freedom.

It has been argued that the **Breusch–Pagan Lagrange multiplier test** is sensitive to the assumption of normality. Koenker (1981) and Koenker and Bassett (1982) suggest that the computation of LM be based on a more **robust estimator** of the variance of ε_i^2,

$$V = \frac{1}{n}\sum_{i=1}^{n}\left[e_i^2 - \frac{\mathbf{e}'\mathbf{e}}{n}\right]^2.$$

Let $\mathbf{u}$ equal $(e_1^2, e_2^2, \ldots, e_n^2)$ and $\mathbf{i}$ be an $n \times 1$ column of 1s. Then $\bar{u} = \mathbf{e}'\mathbf{e}/n$. With this change, the computation becomes $\text{LM} = (1/V)(\mathbf{u} - \bar{u}\,\mathbf{i})'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{u} - \bar{u}\,\mathbf{i})$. Under normality, this modified statistic will have the same limiting distribution as the Breusch–Pagan statistic, but absent normality, there is some evidence that it provides a more powerful test. Waldman (1983) has shown that if the variables in $\mathbf{z}_i$ are the same as those used for the White test described earlier, then the two tests are algebraically the same.

Example 9.2 Testing for Heteroscedasticity

We use the suggested diagnostics to test for heteroscedasticity in the credit card expenditure data in Example 9.2.

1. **White's Test:** There are 15 variables in $(\mathbf{x}, \mathbf{x} \otimes \mathbf{x})$, including the constant term. But because $\text{OwnRent}^2 = \text{OwnRent}$ and $\text{Income} \times \text{Income} = \text{Income}^2$, which is also in the equation, only 13 of the 15 are unique. Regression of the squared least squares residuals on these 13 variables produces $R^2 = 0.199013$. The chi-squared statistic is therefore $72(0.199013) = 14.329$. The 95% critical value of chi-squared with 12 degrees of freedom is 21.03, so despite what might seem to be obvious in Figure 9.1, the hypothesis of homoscedasticity is not rejected by this test.
2. **Breusch–Pagan Test:** This test requires a specific alternative hypothesis. For this purpose, we specify the test based on $\mathbf{z} = [1, \text{Income}, \text{Income}^2]$. Using the least squares residuals, we compute $g_i = e_i^2/(\mathbf{e}'\mathbf{e}/72) - 1$; then $\text{LM} = \frac{1}{2}\mathbf{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{g}$. The computation produces $\text{LM} = 41.920$. The critical value for the chi-squared distribution with two degrees of freedom is 5.99, so the hypothesis of homoscedasticity is rejected. The Koenker and Bassett variant of this statistic is only 6.187, which is still significant but much smaller than the LM statistic. The wide difference between these two statistics suggests that the assumption of normality is erroneous. If the Breusch and Pagan test is based on (1, x), the chi squared statistic is 49.061 with 4 degrees of freedom, while the Koenker and Bassett version is 7.241. The same conclusions are reached.

9.7 TWO APPLICATIONS

This section will present two common applications of the heteroscedastic regression model, Harvey's model of **multiplicative heteroscedasticity** and a model of **groupwise heteroscedasticity** that extends to the disturbance variance some concepts that are usually associated with variation in the regression function.

9.7.1 MULTIPLICATIVE HETEROSCEDASTICITY

Harvey's (1976) model of multiplicative heteroscedasticity is a very flexible, general model that includes most of the useful formulations as special cases. The general formulation is

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}_i'\boldsymbol{\alpha}).$$

A model with heteroscedasticity of the form $\sigma_i^2 = \sigma^2 \prod_{m=1}^{M} z_{im}^{\alpha_m}$ results if the logs of the variables are placed in $\mathbf{z}_i$. The groupwise heteroscedasticity model described in Example 9.4 is produced by making $\mathbf{z}_i$ a set of group dummy variables (one must be omitted). In this case, σ^2 is the disturbance variance for the base group whereas for the other groups, $\sigma_g^2 = \sigma^2 \exp(\alpha_g)$.

Example 9.3 Multiplicative Heteroscedasticity

In Example 6.6, we fit a cost function for the U.S. airline industry of the form

$$\ln C_{it} = \beta_1 + \beta_2 \ln Q_{it} + \beta_3(\ln Q_{it})^2 + \beta_4 \ln P_{fuel,i,t} + \beta_5 \, Loadfactor_{i,t} + \varepsilon_{i,t}$$

where $C_{i,t}$ is total cost, $Q_{i,t}$ is output, and $P_{fuel,i,t}$ is the price of fuel, and the 90 observations in the data set are for six firms observed for 15 years. (The model also included dummy variables

for firm and year, which we will omit for simplicity.) We now consider a revised model in which the load factor appears in the variance of $\varepsilon_{i,t}$ rather than in the regression function. The model is

$$\sigma_{i,t}^2 = \sigma^2 \exp(\gamma\, Loadfactor_{i,t})$$
$$= \exp(\gamma_1 + \gamma_2\, Loadfactor_{i,t}).$$

The constant in the implied regression is $\gamma_1 = \ln \sigma^2$. Figure 9.2 shows a plot of the least squares residuals against *Loadfactor* for the 90 observations. The figure does suggest the presence of heteroscedasticity. (The dashed lines are placed to highlight the effect.) We computed the LM statistic using (9-32). The chi-squared statistic is 2.959. This is smaller than the critical value of 3.84 for one degree of freedom, so on this basis, the null hypothesis of homoscedasticity with respect to the load factor is not rejected.

To begin, we use OLS to estimate the parameters of the cost function and the set of residuals, $e_{i,t}$. Regression of $\log(e_{it}^2)$ on a constant and the load factor provides estimates of γ_1 and γ_2, denoted c_1 and c_2. The results are shown in Table 9.2. As Harvey notes, $\exp(c_1)$ does not necessarily estimate σ^2 consistently—for normally distributed disturbances, it is low by a factor of 1.2704. However, as seen in (9-24), the estimate of σ^2 (biased or otherwise) is not needed to compute the FGLS estimator. Weights $w_{i,t} = \exp(-c_1 - c_2\, Loadfactor_{i,t})$ are computed using these estimates, then weighted least squares using (9-25) is used to obtain the FGLS estimates of $\boldsymbol{\beta}$. The results of the computations are shown in Table 9.2.

We might consider iterating the procedure. Using the results of FGLS at step 2, we can recompute the residuals, then recompute c_1 and c_2 and the weights, and then reenter the iteration. The process converges when the estimate of c_2 stabilizes. This requires seven iterations. The results are shown in Table 9.2. As noted earlier, iteration does not produce any gains here. The second step estimator is already fully efficient. Moreover, this does not produce the MLE, either. That would be obtained by regressing $[e_{i,t}^2/\exp(c_1 + c_2\, Loadfactor_{i,t}) - 1]$ on the constant and load factor at each iteration to obtain the new estimates. We will revisit this in Chapter 14.

FIGURE 9.2 Plot of Residuals against Load Factor.

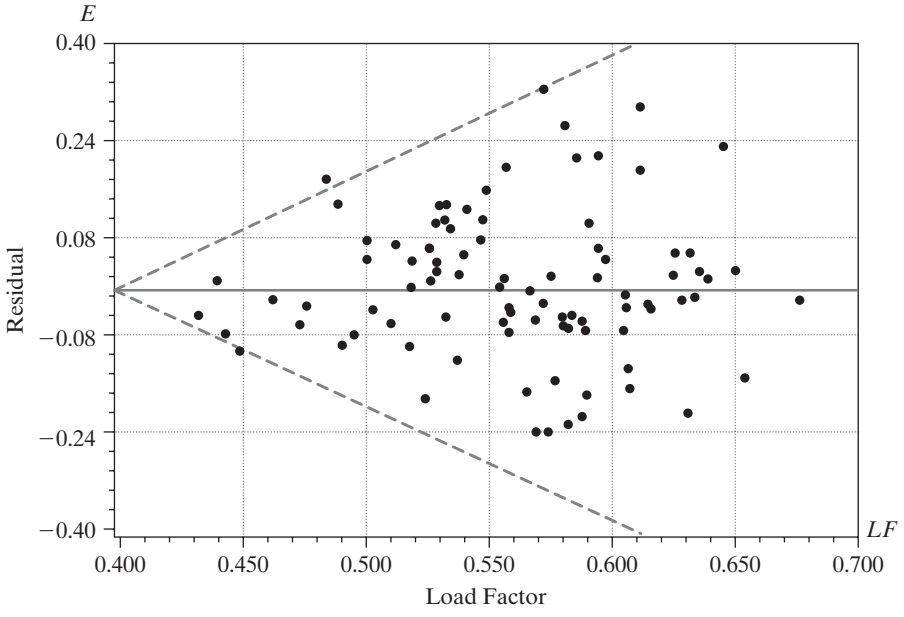


TABLE 9.2 Multiplicative Heteroscedasticity Model

	Constant	ln *Q*	$\ln^2 Q$	ln P_f	R^2	*Sum of Squares*
OLS	9.1382	0.92615	0.029145	0.41006		
	0.24507[a]	0.032306	0.012304	0.018807	0.9861674[b]	1.577479[c]
	0.22595[d]	0.030128	0.011346	0.017524		
Two step	9.2463	0.92136	0.024450	0.40352	0.986119	1.612938
	0.21896	0.033028	0.011412	0.016974		
Iterated[e]	9.2774	0.91609	0.021643	0.40174	0.986071	1.645693
	0.20977	0.032993	0.011017	0.016332		

[a]Conventional OLS standard errors
[b]Squared correlation between actual and fitted values
[c]Sum of squared residuals
[d]White robust standard errors
[e]Values of c_2 by iteration: 8.254344, 11.622473, 11.705029, 11.710618, 11.711012, 11.711040, 11.711042

9.7.2 GROUPWISE HETEROSCEDASTICITY

A groupwise heteroscedastic regression has the structural equations

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n$$
$$E[\varepsilon_i|\mathbf{x}_i] = 0.$$

The n observations are grouped into G groups, each with n_g observations. The slope vector is the same in all groups, but within group g,

$$\text{Var}[\varepsilon_{ig}|\mathbf{x}_{ig}] = \sigma_g^2, \quad i = 1, \ldots, n_g.$$

If the variances are known, then the GLS estimator is

$$\hat{\boldsymbol{\beta}} = \left[\sum_{g=1}^{G}\left(\frac{1}{\sigma_g^2}\right)\mathbf{X}_g'\mathbf{X}_g\right]^{-1}\left[\sum_{g=1}^{G}\left(\frac{1}{\sigma_g^2}\right)\mathbf{X}_g'\mathbf{y}_g\right]. \tag{9-33}$$

Because $\mathbf{X}_g'\mathbf{y}_g = \mathbf{X}_g'\mathbf{X}_g\mathbf{b}_g$, where $\mathbf{b}_g$ is the OLS estimator in the gth subset of observations,

$$\hat{\boldsymbol{\beta}} = \left[\sum_{g=1}^{G}\left(\frac{1}{\sigma_g^2}\right)\mathbf{X}_g'\mathbf{X}_g\right]^{-1}\left[\sum_{g=1}^{G}\left(\frac{1}{\sigma_g^2}\right)\mathbf{X}_g'\mathbf{X}_g\mathbf{b}_g\right] = \left[\sum_{g=1}^{G}\mathbf{V}_g\right]^{-1}\left[\sum_{g=1}^{G}\mathbf{V}_g\mathbf{b}_g\right] = \sum_{g=1}^{G}\mathbf{W}_g\mathbf{b}_g.$$

This result is a matrix weighted average of the G least squares estimators. The weighting matrices are $\mathbf{W}_g = \left[\sum_{g=1}^{G}(\text{Var}[\mathbf{b}_g|\mathbf{X}_g])^{-1}\right]^{-1}(\text{Var}[\mathbf{b}_g|\mathbf{X}_g])^{-1}$. The estimator with the smaller covariance matrix therefore receives the larger weight. [If $\mathbf{X}_g$ is the same in every group, then the matrix $\mathbf{W}_g$ reduces to the simple, $w_g\mathbf{I} = \left(h_g/\sum_g h_g\right)\mathbf{I}$ where $h_g = 1/\sigma_g^2$.]

The preceding is a useful construction of the estimator, but it relies on an algebraic result that might be unusable. If the number of observations in any group is smaller than the number of regressors, then the group-specific OLS estimator cannot be computed.

But, as can be seen in (9-33), that is not what is needed to proceed; what is needed are the weights. As always, pooled least squares is a consistent estimator, which means that using the group-specific subvectors of the OLS residuals,

$$\hat{\sigma}_g^2 = \frac{\mathbf{e}_g'\mathbf{e}_g}{n_g}, \tag{9-34}$$

provides the needed estimator for the group-specific disturbance variance. Thereafter, (9-33) is the estimator and the inverse matrix in that expression gives the estimator of the asymptotic covariance matrix.

Continuing this line of reasoning, one might consider iterating the estimator by returning to (9-34) with the two-step FGLS estimator, recomputing the weights, then returning to (9-33) to recompute the slope vector. This can be continued until convergence. It can be shown that so long as (9-34) is used without a degrees of freedom correction, then if this does converge, it will do so at the maximum likelihood estimator (with normally distributed disturbances).[16]

For testing the homoscedasticity assumption, both White's test and the LM test are straightforward. The variables thought to enter the conditional variance are simply a set of $G - 1$ group dummy variables, not including one of them (to avoid the dummy variable trap), which we'll denote $\mathbf{Z}^*$. Because the columns of $\mathbf{Z}^*$ are binary and orthogonal, to carry out White's test, we need only regress the squared least squares residuals on a constant and $\mathbf{Z}^*$ and compute NR^2 where $N = \sum_g n_g$. The LM test is also straightforward. For purposes of this application of the LM test, it will prove convenient to replace the overall constant in $\mathbf{Z}$ in (9-32) with the remaining group dummy variable. Because the column space of the full set of dummy variables is the same as that of a constant and $G - 1$ of them, all results that follow will be identical. In (9-32), the vector $\mathbf{g}$ will now be G subvectors where each subvector is the n_g elements of $[(e_{ig}^2/\hat{\sigma}^2) - 1]$, and $\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/N$. By multiplying it out, we find that $\mathbf{g}'\mathbf{Z}$ is the G vector with elements $n_g[(\hat{\sigma}_g^2/\hat{\sigma}^2) - 1]$, while $(\mathbf{Z}'\mathbf{Z})^{-1}$ is the $G \times G$ matrix with diagonal elements $1/n_g$. It follows that

$$LM = \frac{1}{2}\mathbf{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{g} = \frac{1}{2}\sum_{g=1}^{G} n_g\left(\frac{\hat{\sigma}_g^2}{\hat{\sigma}^2} - 1\right)^2. \tag{9-35}$$

Both statistics have limiting chi-squared distributions with $G - 1$ degrees of freedom under the null hypothesis of homoscedasticity. (There are only $G - 1$ degrees of freedom because the hypothesis imposes $G - 1$ restrictions, that the G variances are all equal to each other. Implicitly, one of the variances is free and the other $G - 1$ equal that one.)

Example 9.4 Groupwise Heteroscedasticity

Baltagi and Griffin (1983) is a study of gasoline usage in 18 of the 30 OECD countries. The model analyzed in the paper is

$$\ln(\textit{Gasoline usage/car})_{i,t} = \beta_1 + \beta_2 \ln(\textit{Per capita income})_{i,t} + \beta_3 \ln \textit{Price}_{i,t} + \beta_4 \ln(\textit{Cars per capita})_{i,t} + \varepsilon_{i,t},$$

[16] See Oberhofer and Kmenta (1974).

where i = country and $t = 1960, \ldots, 1978$. This is a balanced panel (see Section 11.2) with $19(18) = 342$ observations in total. The data are given in Appendix Table F9.2.

Figure 9.3 displays the OLS residuals using the least squares estimates of the model above with the addition of 18 country dummy variables (1 to 18) (and without the overall constant). (The country dummy variables are used so that the country-specific residuals will have mean zero.) The F statistic for testing the null hypothesis that all the constants are equal is

$$F[(G-1),(\Sigma_{g=1}^{G} n_g - K - G)] = \frac{(\mathbf{e}_0'\mathbf{e}_0 - \mathbf{e}_1'\mathbf{e}_1)/(G-1)}{\mathbf{e}_1'\mathbf{e}_1/(\Sigma_{g=1}^{G} n_g - K - G)} = \frac{(14.90436 - 2.73649)/17}{2.73649/(342-3-18)} = 83.960798,$$

where $\mathbf{e}_0$ is the vector of residuals in the regression with a single constant term and $\mathbf{e}_1$ is the regression with country-specific constan t terms. The critical value from the F table with 17 and 321 degrees of freedom is 1.655. The regression results are given in Table 9.3. Figure 9.3 does convincingly suggest the presence of groupwise heteroscedasticity. The White and LM statistics are $342(0.38365) = 131.21$ and 279.588, respectively. The critical value from the chi-squared distribution with 17 degrees of freedom is 27.587. So, we reject the hypothesis of homoscedasticity and proceed to fit the model by feasible GLS. The two-step estimates are shown in Table 9.3. The FGLS estimator is computed by using weighted least squares, where the weights are $1/\hat{\sigma}_g^2$ for each observation in country g. Comparing the White standard errors to the two-step estimators, we see that in this instance, there is a substantial gain to using feasible generalized least squares.

FIGURE 9.3 Plot of OLS Residuals by Country.

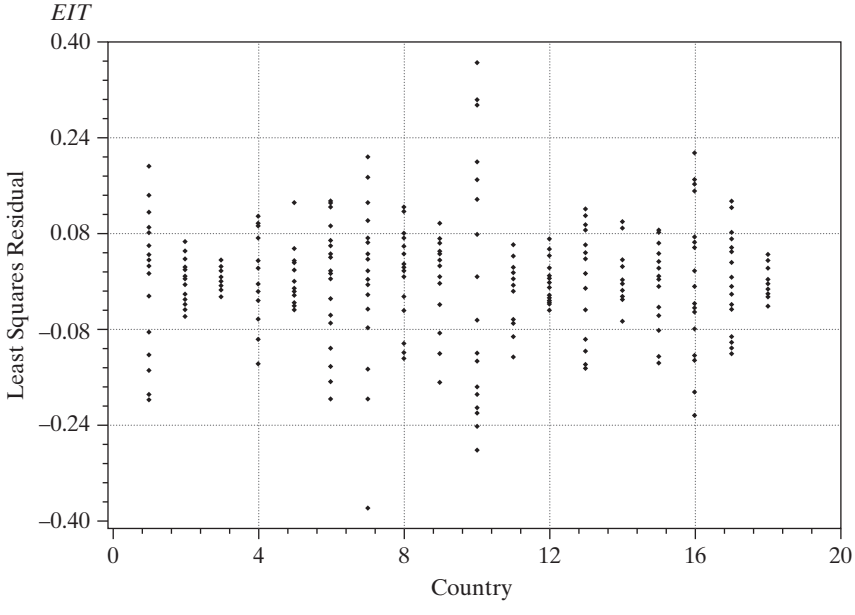


TABLE 9.3 Estimated Gasoline Consumption Equations

	OLS			*FGLS*	
	Coefficient	***Std. Error***	***White Std. Err.***	***Coefficient***	***Std. Error***
ln *Income*	0.66225	0.07339	0.07277	0.57507	0.02927
ln *Price*	−0.32170	0.04410	0.05381	−0.27967	0.03519
ln *Cars/Cap*.	−0.64048	0.02968	0.03876	−0.56540	0.01613
Country 1	2.28586	0.22832	0.22608	2.43707	0.11308
Country 2	2.16555	0.21290	0.20983	2.31699	0.10225
Country 3	3.04184	0.21864	0.22479	3.20652	0.11663
Country 4	2.38946	0.20809	0.20783	2.54707	0.10250
Country 5	2.20477	0.21647	0.21087	2.33862	0.10101
Country 6	2.14987	0.21788	0.21846	2.30066	0.10893
Country 7	2.33711	0.21488	0.21801	2.57209	0.11206
Country 8	2.59233	0.24369	0.23470	2.72376	0.11384
Country 9	2.23255	0.23954	0.22973	2.34805	0.10795
Country 10	2.37593	0.21184	0.22643	2.58988	0.11821
Country 11	2.23479	0.21417	0.21311	2.39619	0.10478
Country 12	2.21670	0.20304	0.20300	2.38486	0.09950
Country 13	1.68178	0.16246	0.17133	1.90306	0.08146
Country 14	3.02634	0.39451	0.39180	3.07825	0.20407
Country 15	2.40250	0.22909	0.23280	2.56490	0.11895
Country 16	2.50999	0.23566	0.26168	2.82345	0.13326
Country 17	2.34545	0.22728	0.22322	2.48214	0.10955
Country 18	3.05525	0.21960	0.22705	3.21519	0.11917

9.8 SUMMARY AND CONCLUSIONS

This chapter has introduced a major extension of the classical linear model. By allowing for heteroscedasticity and autocorrelation in the disturbances, we expand the range of models to a large array of frameworks. We will explore these in the next several chapters. The formal concepts introduced in this chapter include how this extension affects the properties of the least squares estimator, how an appropriate estimator of the asymptotic covariance matrix of the least squares estimator can be computed in this extended modeling framework, and, finally, how to use the information about the variances and covariances of the disturbances to obtain an estimator that is more efficient than ordinary least squares.

We have analyzed in detail one form of the generalized regression model, the model of heteroscedasticity. We first considered least squares estimation. The primary result for least squares estimation is that it retains its consistency and asymptotic normality, but some correction to the estimated asymptotic covariance matrix may be needed for appropriate inference. The White estimator is the standard approach for this computation. After examining two general tests for heteroscedasticity, we then narrowed the model to some specific parametric forms, and considered weighted (generalized) least squares for efficient estimation and maximum likelihood estimation. If the form of the heteroscedasticity is known but involves unknown parameters, then it remains uncertain

whether FGLS corrections are better than OLS. Asymptotically, the comparison is clear, but in small or moderately sized samples, the additional variation incorporated by the estimated variance parameters may offset the gains to GLS.

Key Terms and Concepts

- Asymptotic properties
- Autocorrelation
- Breusch–Pagan Lagrange multiplier test
- Efficient estimator
- Feasible generalized least squares (FGLS)
- Finite-sample properties
- Generalized least squares (GLS)
- Generalized linear regression model
- Generalized sum of squares
- Groupwise heteroscedasticity
- Heteroscedasticity
- Lagrange multiplier test
- Multiplicative heteroscedasticity
- Ordinary least squares (OLS)
- Panel data
- Robust estimator
- Robustness to unknown heteroscedasticity
- Two-step estimator
- Weighted least squares (WLS)
- White heteroscedasticity robust estimator
- White test
- Aitken's theorem

Exercises

1. What is the covariance matrix, $\text{Cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}} - \mathbf{b}]$, of the GLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$ and the difference between it and the OLS estimator, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$? The result plays a pivotal role in the development of specification tests in Hausman (1978).
2. This and the next two exercises are based on the test statistic usually used to test a set of J linear restrictions in the generalized regression model,

$$F[J, n - K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})/J}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n - K)},$$

   where $\hat{\boldsymbol{\beta}}$ is the GLS estimator. Show that if $\boldsymbol{\Omega}$ is known, if the disturbances are normally distributed and if the null hypothesis, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, is true, then this statistic is exactly distributed as F with J and $n - K$ degrees of freedom. What assumptions about the regressors are needed to reach this conclusion? Need they be nonstochastic?
3. Now suppose that the disturbances are not normally distributed, although $\boldsymbol{\Omega}$ is still known. Show that the limiting distribution of the previous statistic is $(1/J)$ times a chi-squared variable with J degrees of freedom. (*Hint*: The denominator converges to σ^2.) Conclude that, in the generalized regression model, the limiting distribution of the Wald statistic,

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'\{\mathbf{R}(\text{Est. Var}[\hat{\boldsymbol{\beta}}])\mathbf{R}'\}^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}),$$

   is chi-squared with J degrees of freedom, regardless of the distribution of the disturbances, as long as the data are otherwise well behaved. Note that in a finite sample, the true distribution may be approximated with an $F[J, n - K]$ distribution. It is a bit ambiguous, however, to interpret this fact as implying that the statistic

is asymptotically distributed as F with J and $n - K$ degrees of freedom, because the limiting distribution used to obtain our result is the chi-squared, not the F. In this instance, the $F[J, n - K]$ is a random variable that tends asymptotically to the chi-squared variate.

4. Finally, suppose that $\mathbf{\Omega}$ must be estimated, but that assumptions (9-22) and (9-23) are met by the estimator. What changes are required in the development of the previous problem?
5. In the generalized regression model, if the K columns of $\mathbf{X}$ are characteristic vectors of $\mathbf{\Omega}$, then ordinary least squares and generalized least squares are identical. (The result is actually a bit broader; $\mathbf{X}$ may be any linear combination of exactly K characteristic vectors. This result is Kruskal's theorem.)
   a. Prove the result directly using matrix algebra.
   b. Prove that if $\mathbf{X}$ contains a constant term and if the remaining columns are in deviation form (so that the column sum is zero), then the model of Exercise 8 is one of these cases. (The seemingly unrelated regressions model with identical regressor matrices, discussed in Chapter 10, is another.)
6. In the generalized regression model, suppose that $\mathbf{\Omega}$ is known.
   a. What is the covariance matrix of the OLS and GLS estimators of $\boldsymbol{\beta}$?
   b. What is the covariance matrix of the OLS residual vector $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$?
   c. What is the covariance matrix of the GLS residual vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$?
   d. What is the covariance matrix of the OLS and GLS residual vectors?
7. Suppose that y has the pdf $f(y|\mathbf{x}) = (1/\mathbf{x}'\boldsymbol{\beta})e^{-y/(\mathbf{x}'\boldsymbol{\beta})}$, $y > 0$. Then $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ and $\text{Var}[y|\mathbf{x}] = (\mathbf{x}'\boldsymbol{\beta})^2$. For this model, prove that GLS and MLE are the same, even though this distribution involves the same parameters in the conditional mean function and the disturbance variance.
8. Suppose that the regression model is $y = \mu + \varepsilon$, where ε has a zero mean, constant variance, and equal correlation, ρ, across observations. Then $\text{Cov}[\varepsilon_i, \varepsilon_j] = \sigma^2\rho$ if $i \neq j$. Prove that the least squares estimator of μ is inconsistent. Find the characteristic roots of $\mathbf{\Omega}$ and show that Condition 2 before (9-10) is violated.
9. Suppose that the regression model is $y_i = \mu + \varepsilon_i$, where

$$E[\varepsilon_i|x_i] = 0, \text{Cov}[\varepsilon_i, \varepsilon_j|x_i, x_j] = 0 \quad \text{for } i = j, \text{ but } \text{Var}[\varepsilon_i|x_i] = \sigma^2 x_i^2, x_i > 0.$$

   a. Given a sample of observations on y_i and x_i, what is the most efficient estimator of μ? What is its variance?
   b. What is the OLS estimator of μ, and what is the variance of the OLS estimator?
   c. Prove that the estimator in part a is at least as efficient as the estimator in part b.
10. For the model in Exercise 9, what is the probability limit of $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$? Note that s^2 is the least squares estimator of the residual variance. It is also n times the conventional estimator of the variance of the OLS estimator,

$$\text{Est.Var}[\bar{y}] = s^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{s^2}{n}.$$

How does this equation compare with the true value you found in part b of Exercise 9? Does the conventional estimator produce the correct estimator of the true asymptotic variance of the least squares estimator?

11. For the model in Exercise 9, suppose that ε is normally distributed, with mean zero and variance $\sigma^2[1 + (\gamma x)^2]$. Show that σ^2 and γ^2 can be consistently estimated by a regression of the least squares residuals on a constant and x^2. Is this estimator efficient?
12. Two samples of 50 observations each produce the following moment matrices. (In each case, $\mathbf{X}$ is a constant and one variable.)

	Sample 1	Sample 2
$\mathbf{X}'\mathbf{X}$	$\begin{bmatrix} 50 & 300 \\ 300 & 2100 \end{bmatrix}$	$\begin{bmatrix} 50 & 300 \\ 300 & 2100 \end{bmatrix}$
$\mathbf{y}'\mathbf{X}$	[300 2000]	[300 2200]
$\mathbf{y}'\mathbf{y}$	[2100]	[2800]

    a. Compute the least squares regression coefficients and the residual variances s^2 for each data set. Compute the R^2s for each regression.
    b. Compute the OLS estimate of the coefficient vector assuming that the coefficients and disturbance variance are the same in the two regressions. Also compute the estimate of the asymptotic covariance matrix of the estimate.
    c. Test the hypothesis that the variances in the two regressions are the same without assuming that the coefficients are the same in the two regressions.
    d. Compute the two-step FGLS estimator of the coefficients in the regressions, assuming that the constant and slope are the same in both regressions. Compute the estimate of the covariance matrix and compare it with the result of part b.
13. Suppose that in the groupwise heteroscedasticity model of Section 9.7.2, $\mathbf{X}_i$ is the same for all i. What is the generalized least squares estimator of $\boldsymbol{\beta}$? How would you compute the estimator if it were necessary to estimate σ_i^2?
14. The model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \beta + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

satisfies the groupwise heteroscedastic regression model of Section 9.7.2 All variables have zero means. The following sample second-moment matrix is obtained from a sample of 20 observations:

$$\begin{array}{c} \\ y_1 \\ y_2 \\ x_1 \\ x_2 \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & y_2 & x_1 & x_2 \end{array} \\ \begin{bmatrix} 20 & 6 & 4 & 3 \\ 6 & 10 & 3 & 6 \\ 4 & 3 & 5 & 2 \\ 3 & 6 & 2 & 10 \end{bmatrix} \end{array}.$$

    a. Compute the two separate OLS estimates of β, their sampling variances, the estimates of σ_1^2 and σ_2^2, and the R^2s in the two regressions.
    b. Carry out the Lagrange multiplier test of the hypothesis that $\sigma_1^2 = \sigma_2^2$.
    c. Compute the two-step FGLS estimate of β and an estimate of its sampling variance. Test the hypothesis that β equals 1.
    d. Compute the maximum likelihood estimates of β, σ_1^2, and σ_2^2 by iterating the FGLS estimates to convergence.

15. The following table presents a hypothetical panel of data:

	$i = 1$		$i = 2$		$i = 3$	
t	y	x	y	x	y	x
1	30.27	24.31	38.71	28.35	37.03	21.16
2	35.59	28.47	29.74	27.38	43.82	26.76
3	17.90	23.74	11.29	12.74	37.12	22.21
4	44.90	25.44	26.17	21.08	24.34	19.02
5	37.58	20.80	5.85	14.02	26.15	18.64
6	23.15	10.55	29.01	20.43	26.01	18.97
7	30.53	18.40	30.38	28.13	29.64	21.35
8	39.90	25.40	36.03	21.78	30.25	21.34
9	20.44	13.57	37.90	25.65	25.41	15.86
10	36.85	25.60	33.90	11.66	26.04	13.28

a. Estimate the groupwise heteroscedastic model of Section 9.7.2. Include an estimate of the asymptotic variance of the slope estimator. Use a two-step procedure, basing the FGLS estimator at the second step on residuals from the pooled least squares regression.
b. Carry out the Lagrange multiplier tests of the hypothesis that the variances are all equal.

Applications

1. This application is based on the following data set.

***50 Observations on* y:**								
−1.42	2.75	2.10	−5.08	1.49	1.00	0.16	−1.11	1.66
−0.26	−4.87	5.94	2.21	−6.87	0.90	1.61	2.11	−3.82
−0.62	7.01	26.14	7.39	0.79	1.93	1.97	−23.17	−2.52
−1.26	−0.15	3.41	−5.45	1.31	1.52	2.04	3.00	6.31
5. 51	−15.22	−1.47	−1.48	6.66	1.78	2.62	−5.16	−4.71
−0.35	−0.48	1.24	0.69	1.91				
***50 Observations on* x_1:**								
−1.65	1.48	0.77	0.67	0.68	0.23	−0.40	−1.13	0.15
−0.63	0.34	0.35	0.79	0.77	−1.04	0.28	0.58	−0.41
−1.78	1.25	0.22	1.25	−0.12	0.66	1.06	−0.66	−1.18
−0.80	−1.32	0.16	1.06	−0.60	0.79	0.86	2.04	−0.51
0. 02	0.33	−1.99	0.70	−0.17	0.33	0.48	1.90	−0.18
−0.18	−1.62	0.39	0.17	1.02				
***50 Observations on* x_2:**								
−0.67	0.70	0.32	2.88	−0.19	−1.28	−2.72	−0.70	−1.55
−0.74	−1.87	1.56	0.37	−2.07	1.20	0.26	−1.34	−2.10
0. 61	2.32	4.38	2.16	1.51	0.30	−0.17	7.82	−1.15
1. 77	2.92	−1.94	2.09	1.50	−0.46	0.19	−0.39	1.54
1. 87	−3.45	−0.88	−1.53	1.42	−2.70	1.77	−1.89	−1.85
2. 01	1.26	−2.02	1.91	−2.23				

a. Compute the OLS regression of y on a constant, x_1, and x_2. Be sure to compute the conventional estimator of the asymptotic covariance matrix of the OLS estimator as well.
b. Compute the White estimator of the appropriate asymptotic covariance matrix for the OLS estimates.
c. Test for the presence of heteroscedasticity using White's general test. Do your results suggest the nature of the heteroscedasticity?
d. Use the Breusch-Pagan (1980) and Godfrey (1988) Lagrange multiplier test to test for heteroscedasticity.
e. Reestimate the parameters using a two-step FGLS estimator. Use Harvey's formulation, $\text{Var}[\varepsilon_i | x_{i1}, x_{i2}] = \sigma^2 \exp(\gamma_1 x_{i1} + \gamma_2 x_{i2})$.

2. (We look ahead to our use of maximum likelihood to estimate the models discussed in this chapter in Chapter 14.) In Example 9.3, we computed an iterated FGLS estimator using the airline data and the model $\text{Var}[\varepsilon_{it} | \mathit{Loadfactor}_{i,t}] = \exp(\gamma_1 + \gamma_2 \mathit{Loadfactor}_{i,t})$. The weights computed at each iteration were computed by estimating (γ_1, γ_2) by least squares regression of $\ln \hat{\varepsilon}_{i,t}^2$ on a constant and *Loadfactor*. The maximum likelihood estimator would proceed along similar lines, however the weights would be computed by regression of $[\hat{\varepsilon}_{i,t}^2 / \hat{\sigma}_{i,t}^2 - 1]$ on a constant and $\mathit{Loadfactor}_{i,t}$ instead. Use this alternative procedure to estimate the model. Do you get different results?

10
SYSTEMS OF REGRESSION EQUATIONS

10.1 INTRODUCTION

There are many settings in which the single-equation models of the previous chapters apply to a group of related variables. In these contexts, we will want to consider the several models jointly. Here are example:

1. **Set of Regression Equations.** Munnell's (1990) model for output by the 48 contiguous states in the U.S., m, at time t is

$$\ln GSP_{mt} = \beta_{1m} + \beta_{2m} \ln pc_{mt} + \beta_{3m} \ln hwy_{mt} + \beta_{4m} \ln water_{mt} + \beta_{5m} \ln util_{mt} + \beta_{6m} \ln emp_{mt} + \beta_{7m} unemp_{mt} + \varepsilon_{mt},$$

   where the variables are labor and public capital. Taken one state at a time, this provides a set of 48 linear regression models. The application develops a model in which the observations are correlated across time (t,s) within a state. It would be natural as well for observations at a point in time to be correlated across states (m,n), at least for some states. An important question is whether it is valid to assume that the coefficient vector is the same for all states in the sample.
2. **Identical Regressors.** The capital asset pricing model of finance specifies that, for a given security,

$$r_{it} - r_{ft} = \alpha_i + \beta_i(r_{mt} - r_{ft}) + \varepsilon_{it},$$

   where r_{it} is the return over period t on security i, r_{ft} is the return on a risk-free security, r_{mt} is the market return, and β_i is the security's beta coefficient. The disturbances are obviously correlated across securities. The knowledge that the return on security i exceeds the risk-free rate by a given amount provides some information about the excess return of security j, at least for some j's. It may be useful to estimate the equations jointly rather than ignore this connection. The fact that the right-hand side, $[constant, r_{mt} - r_{ft}]$, is the same for all i makes this model an interesting special case of the more general set of regressions.
3. **Dynamic Linear Equations.** Pesaran and Smith (1995) proposed a dynamic model for wage determination in 38 UK industries. The central equation is of the form

$$y_{mt} = \alpha_m + \mathbf{x}'_{mt}\boldsymbol{\beta}_m + \gamma_m y_{m,t-1} + \varepsilon_{mt}.$$

   Nair-Reichert and Weinhold's (2001) cross-country analysis of growth in developing countries takes the same form. In both cases, each group (industry, country) could be analyzed separately. However, the connections across groups and the interesting question of "poolability"—that is, whether it is valid to assume identical

coefficients—is a central part of the analysis. The lagged dependent variable in the model produces a substantial complication.

4. **System of Demand Equations.** In a model of production, the optimization conditions of economic theory imply that, if a firm faces a set of factor prices $\mathbf{p}$, then its set of cost-minimizing factor demands for producing output Q will be a set of M equations of the form $x_m = f_m(Q, \mathbf{p})$. The empirical model is

$$\begin{aligned} x_1 &= f_1(Q, \mathbf{p}|\boldsymbol{\theta}) + \varepsilon_1, \\ x_2 &= f_2(Q, \mathbf{p}|\boldsymbol{\theta}) + \varepsilon_2, \\ &\cdots \\ x_M &= f_M(Q, \mathbf{p}|\boldsymbol{\theta}) + \varepsilon_M, \end{aligned}$$

where $\boldsymbol{\theta}$ is a vector of parameters that are part of the technology and ε_m represents errors in optimization. Once again, the disturbances should be correlated. In addition, the same parameters of the production technology will enter all the demand equations, so the set of equations has cross-equation restrictions. Estimating the equations separately will waste the information that the same set of parameters appears in all the equations.

5. **Vector Autoregression.** A useful formulation that appears in many macroeconomics applications is the vector autoregression, or VAR. In Chapter 13, we will examine a model of Swedish municipal government fiscal activities in the form

$$\begin{aligned} S_{m,t} &= \alpha_1 + \gamma_{11} S_{m,t-1} + \gamma_{12} R_{m,t-1} + \gamma_{13} G_{m,t-1} + \varepsilon_{S,m,t}, \\ R_{m,t} &= \alpha_2 + \gamma_{21} S_{m,t-1} + \gamma_{22} R_{m,t-1} + \gamma_{23} G_{m,t-1} + \varepsilon_{R,m,t}, \\ G_{m,t} &= \alpha_3 + \gamma_{31} S_{m,t-1} + \gamma_{32} R_{m,t-1} + \gamma_{33} G_{m,t-1} + \varepsilon_{G,m,t}, \end{aligned}$$

where S, R, and G are spending, tax revenues, and grants, respectively, for municipalities m in period t. VARs without restrictions are similar to Example 2 above. The dynamic equations can be used to trace the influences of shocks in a system as they exert their influence through time.

6. **Linear panel data model.** In Chapter 11, we will examine models for **panel data** – $t = 1, \ldots, T$ repeated observations on individuals m, of the form

$$y_{mt} = \mathbf{x}_{mt}'\boldsymbol{\beta} + \varepsilon_{mt}.$$

In Example 11.1, we consider a wage equation,

$$\ln Wage_{mt} = \beta_1 + \beta_2 Experience_{mt} + \ldots + \mathbf{x}_{mt}'\boldsymbol{\beta} + \varepsilon_{mt}.$$

For some purposes, it is useful to consider this model as a set of T regression equations, one for each period. Specification of the model focuses on correlations of the unobservables in ε_{mt} across periods and with dynamic behavior of $\ln Wage_{mt}$.

7. **Simultaneous Equations System.** A common form of a model for equilibrium in a market would be

$$\begin{aligned} Q_{Demand} &= \alpha_1 + \alpha_2 Price + \alpha_3 Income + \mathbf{d}'\boldsymbol{\alpha} + \varepsilon_{Demand}, \\ Q_{Supply} &= \beta_1 + \beta_2 Price + \beta_3 FactorPrice + \mathbf{s}'\boldsymbol{\beta} + \varepsilon_{Supply}, \\ Q_{Equilibrium} &= Q_{Demand} = Q_{Supply}, \end{aligned}$$

where **d** and **s** are exogenous variables that influence the equilibrium through their impact on the demand and supply curves, respectively. This model differs from those suggested thus far because the implication of the third equation is that *Price* is not exogenous in the equation system. The equations of this model fit into the endogenous variables framework developed in Chapter 8. The multiple equations framework developed in this chapter provides additional results for estimating "simultaneous equations models" such as this.

This chapter will develop the essential theory for sets of related regression equations. Section 10.2 examines the general model in which each equation has its own set of parameters and examines efficient estimation techniques and the special case in which the coefficients are the same in all equations. Production and consumer demand models are special cases of the general model in which the equations obey an *adding-up constraint* that has implications for specification and estimation. Such demand systems are examined in Section 10.3. This section examines an application of the seemingly unrelated regressions model that illustrates the interesting features of empirical demand studies. The seemingly unrelated regressions model is also extended to the translog specification, which forms the platform for many microeconomic studies of production and cost. Finally, Section 10.4 combines the results of Chapter 8 on models with endogenous variables with the development in this chapter of multiple equation systems. In this section, we will develop **simultaneous equations models**. The supply and demand model suggested in Example 6 above, of equilibrium in which price and quantity in a market are jointly determined, is an application.

10.2 THE SEEMINGLY UNRELATED REGRESSIONS MODEL

All the examples suggested in the Introduction have a common structure, which we may write as

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1,\\ \mathbf{y}_2 &= \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2,\\ &\cdots\\ \mathbf{y}_M &= \mathbf{X}_M\boldsymbol{\beta}_M + \boldsymbol{\varepsilon}_M.\end{aligned}$$

There are M equations and T observations in the sample.[1] The **seemingly unrelated regressions (SUR)** model is

$$\mathbf{y}_m = \mathbf{X}_m\boldsymbol{\beta}_m + \boldsymbol{\varepsilon}_m, \quad m = 1, \ldots, M. \tag{10-1}$$

The equations are labeled "seemingly unrelated" because they are linked by the possible correlation of the unobserved disturbances, ε_{mt} and ε_{nt}.[2] By stacking the sets of observations, we obtain

[1]The use of T is not meant to imply any connection to time series. For instance, in the fourth example, above, the data might be cross sectional.

[2]See Zellner (1962) who coined the term.

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_M \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_M \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{10-2}$$

The $MT \times 1$ vector of disturbances is

$$\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2', \ \ldots, \boldsymbol{\varepsilon}_M']'.$$

We assume strict exogeneity of $\mathbf{X}_i$,

$$E[\boldsymbol{\varepsilon} | \mathbf{X}_1, \mathbf{X}_2, \ \ldots, \mathbf{X}_M] = \mathbf{0},$$

and homoscedasticity and nonautocorrelation within each equation,

$$E[\boldsymbol{\varepsilon}_m \boldsymbol{\varepsilon}_m' | \mathbf{X}_1, \mathbf{X}_2, \ \ldots, \mathbf{X}_M] = \sigma_{mm}\mathbf{I}_T.$$

The strict exogeneity assumption is a bit stronger than necessary for present purposes. We could allow more generality by assuming only $E[\boldsymbol{\varepsilon}_m | \mathbf{X}_m] = \mathbf{0}$—that is, allowing the disturbances in equation n to be correlated with the regressors in equation m but not equation n. But that extension would not arise naturally in an application. A total of T observations are to be used in estimating the parameters of the M equations. Each equation involves K_m regressors, for a total of $K = \sum_{m=1}^{M} K_m$ in (10-2). We will require $T > K_m$ (so that, if desired, we could fit each equation separately). The data are assumed to be well behaved,as described in Section 4.4.1, so we shall not treat the issue separately here. For the present, we also assume that disturbances are not correlated across periods (or individuals) but may be correlated across equations (at a point in time or for a given individual). Therefore,

$$E[\varepsilon_{mt}\varepsilon_{ns} | \mathbf{X}_1, \mathbf{X}_2, \ \ldots, \mathbf{X}_M] = \sigma_{mn}, \quad \text{if } t = s \text{ and 0 if } t \neq s.$$

The disturbance formulation for the entire model is

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}_1, \mathbf{X}_2, \ \ldots, \mathbf{X}_M] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11}\mathbf{I} & \sigma_{12}\mathbf{I} & \cdots & \sigma_{1M}\mathbf{I} \\ \sigma_{21}\mathbf{I} & \sigma_{22}\mathbf{I} & \cdots & \sigma_{2M}\mathbf{I} \\ & \vdots & & \\ \sigma_{M1}\mathbf{I} & \sigma_{M2}\mathbf{I} & \cdots & \sigma_{MM}\mathbf{I} \end{bmatrix} \tag{10-3}$$

$$= \boldsymbol{\Sigma} \otimes \mathbf{I},$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ & \vdots & & \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix}$$

is the $M \times M$ covariance matrix of the disturbances for the tth observation, $\boldsymbol{\varepsilon}_t$.

The SUR model thus far assumes that each equation obeys the assumptions of the linear model of Chapter 2—no heteroscedasticity or autocorrelation (within or across equations). Bartels and Fiebig (1992), Bartels and Aigner (1991), Mandy and Martins-Filho (1993), and Kumbhakar (1996) suggested extensions that involved heteroscedasticity within each equation. Autocorrelation of the disturbances of regression models is usually

not the focus of the investigation, though Munnell's application to aggregate statewide data might be a natural application.[3] (It might also be a natural candidate for the "spatial autoregression" model of Section 11.7.) All of these extensions are left for more advanced treatments and specific applications.

10.2.1 ORDINARY LEAST SQUARES AND ROBUST INFERENCE

For purposes of developing effective estimation methods, there are two ways to visualize the arrangement of the data. Consider the model in Example 10.2, which examines a cost function for electricity generation. The three equations are

$$\begin{aligned}\ln(C/P_f) &= \alpha_1 + \alpha_2 \ln Q + \alpha_3 \ln(P_k/P_f) + \alpha_4 \ln(P_l/P_f) + \varepsilon_c,\\ s_k &= \beta_1 + \varepsilon_k,\\ s_l &= \gamma_1 + \varepsilon_l,\end{aligned}$$

where C is total cost, P_k, P_l, and P_f are unit prices for capital, labor, and fuel, Q is output, and s_k and s_l are cost shares for capital and labor. (The fourth equation, for s_f, is obtained from $s_k + s_l + s_f = 1$.) There are $T = 145$ observations for each of the $M = 3$ equations. The data may be *stacked by equations* as in the following,

$$\begin{bmatrix}\mathbf{ln(C/P_f)}\\ \mathbf{s_k}\\ \mathbf{s_l}\end{bmatrix} = \begin{bmatrix}\mathbf{i} & \mathbf{ln\ Q} & \mathbf{ln(P_k/P_f)} & \mathbf{ln(P_l/P_f)} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{i} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{i}\end{bmatrix}\begin{pmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4\\ \beta_1\\ \gamma_1\end{pmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_c\\ \boldsymbol{\varepsilon}_k\\ \boldsymbol{\varepsilon}_l\end{bmatrix}, \quad \textbf{(10-4)}$$

$$\begin{bmatrix}\mathbf{y}_c\\ \mathbf{y}_k\\ \mathbf{y}_l\end{bmatrix} = \begin{bmatrix}\mathbf{X}_c & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{X}_l\end{bmatrix}\begin{pmatrix}\boldsymbol{\beta}_c\\ \boldsymbol{\beta}_k\\ \boldsymbol{\beta}_l\end{pmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_c\\ \boldsymbol{\varepsilon}_k\\ \boldsymbol{\varepsilon}_l\end{bmatrix}.$$

Each block of data in the bracketed matrices contains the T observations for equation m. The covariance matrix for the $MT \times 1$ vector of disturbances appears in (10-3). The data may instead be *stacked by observations* by reordering the rows to obtain

$$\mathbf{y} = \begin{bmatrix}\begin{pmatrix}\ln(C/P_f)\\ s_k\\ s_l\end{pmatrix} i = firm\ 1\\ \cdots\\ \begin{pmatrix}\ln(C/P_f)\\ s_k\\ s_l\end{pmatrix} i = firm\ T\end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix}\begin{pmatrix}\varepsilon_c\\ \varepsilon_k\\ s_l\end{pmatrix} i = firm\ 1\\ \cdots\\ \begin{pmatrix}\varepsilon_c\\ \varepsilon_k\\ \varepsilon_l\end{pmatrix} i = firm\ T\end{bmatrix} \text{ and } \mathbf{X} \text{ likewise.} \quad \textbf{(10-5)}$$

[3] Dynamic SUR models are proposed by Anderson, and Blundell (1982). Other applications are examined in Kiviet, Phillips, and Schipp (1995), DesChamps (1998), and Wooldridge (2010, p. 194). The VAR models are an important group of applications, but they come from a different analytical framework. Related results may be found in Guilkey and Schmidt (1973), Guilkey (1974), Berndt and Savin (1977), Moschino and Moro (1994), McLaren (1996), and Holt (1998).

By this arrangement,

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\, \mathbf{X}] = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix} = \mathbf{I} \otimes \boldsymbol{\Sigma}. \tag{10-6}$$

The arrangement in (10-4) will be more convenient for formulating the applications, as in Example 10.4. The format in (10-5) will be more convenient for formulating the estimator and examining its properties.

From (10-2), we can see that with no restrictions on $\boldsymbol{\beta}$, ordinary least squares estimation of $\boldsymbol{\beta}$ will be equation by equation OLS,

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \Rightarrow \mathbf{b}_m = (\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y}_m.$$

Therefore,

$$\mathbf{b}_{\mathrm{m}} = \boldsymbol{\beta}_m + (\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\boldsymbol{\varepsilon}_m.$$

Because this is a simple regression model with homoscedastic and nonautocorrelated disturbances, the familiar estimator of the asymptotic covariance matrix for $(\mathbf{b}_m, \mathbf{b}_n)$ is

$$\hat{\mathbf{V}} = \text{Est.Asy.Cov}[\mathbf{b}_m, \mathbf{b}_n] = s_{mn}(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{X}_n\,(\mathbf{X}_n'\mathbf{X}_n)^{-1}, \tag{10-7}$$

where $s_{mn} = \mathbf{e}_m'\mathbf{e}_n/T$. There is a small ambiguity about the degrees of freedom in s_{mn}. For the diagonal elements, $(T - K_m)$ would be appropriate. One suggestion for the off-diagonal elements that seems natural, but does not produce an unbiased estimator, is $[(T - K_m)(T - K_n)]^{1/2}$.[4]

For inference purposes, Equation (10-7) relies on the two assumptions of homoscedasticity and nonautocorrelation. We can see in (10-6) what features are accommodated and what are not. The estimator does allow a form of heteroscedasticity across equations, in that $\sigma_{mm} \neq \sigma_{nn}$ when $m \neq n$. This is not a real generality, however. For example, in the cost-share equation, it allows the variance of the cost disturbance to be different from the share disturbance, but that would be expected. It does assume that observations are homoscedastic within each equation, in that $\mathrm{E}[\boldsymbol{\varepsilon}_m\boldsymbol{\varepsilon}_m' \,|\, \mathbf{X}] = \sigma_{mm}\mathbf{I}$. It allows observations to be correlated across equations, in that $\sigma_{mn} \neq 0$, but it does not allow observations at different times (or different firms in our example) to be correlated. So, the estimator thus far is not generally robust. Robustness to autocorrelation would be the case of lesser interest, save for the panel data models considered in the next chapter. An extension to more general heteroscedasticity might be attractive. We can allow the diagonal matrices in (10-6) to vary arbitrarily or to depend on $\mathbf{X}_m$. The common $\boldsymbol{\Sigma}$ in (10-6) would be replaced with $\boldsymbol{\Sigma}_m$. The estimator in (10-7) would be replaced by

$$\hat{\mathbf{V}}_{Robust} = \text{Est.Asy.Var}[\mathbf{b}] = \left(\sum_{t=1}^{T}\mathbf{X}_t'\mathbf{X}_t\right)^{-1}\left(\sum_{t=1}^{T}(\mathbf{X}_t'\mathbf{e}_t)(\mathbf{e}_t'\mathbf{X}_t)\right)\left(\sum_{t=1}^{T}\mathbf{X}_t'\mathbf{X}_t\right)^{-1}. \tag{10-8}$$

[4]See Srivastava and Giles (1987).

Note $\mathbf{X}_t$ is M rows and $\Sigma_{m=1}^{M} K_m$ columns corresponding to the tth observations for all M equations, while $\mathbf{e}_t$ is an $M \times 1$ vector of OLS residuals based on (10-5). For example, in (10-5), $\mathbf{X}_1$ is the 3×6 matrix,

$$\mathbf{X}_1 = \begin{bmatrix} 1 & \ln Q & \ln(P_k/P_f) & \ln(P_l/P_f) & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{firm\ 1}.$$

Then, (10-8) would be a multiple equation version of the White estimator for arbitrary heteroscedasticity shown in Section 9.4.4.

For testing hypotheses, either within or across equations, of the form H_0: $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, we can use the Wald statistic,

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}\hat{\mathbf{V}}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}),$$

which has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. For simple hypotheses involving one coefficient, such as H_0: $\beta_k = 0$, we would report the square root of W as the "asymptotic t ratio," $z_k = \hat{\beta}_k/\text{Asy.S.E.}(\hat{\beta}_k)$ where the asymptotic standard error would be the square root of the diagonal element of $\hat{\mathbf{V}}$. This would have a standard normal distribution in large samples under the null hypothesis.

10.2.2 GENERALIZED LEAST SQUARES

Each equation is, by itself, a linear regression. Therefore, the parameters could be estimated consistently, if not efficiently, one equation at a time, by ordinary least squares. The **generalized regression model** applies to the stacked model in (10-2). In (10-3), where the $\mathbf{I}$ matrix is $T \times T$, the $MT \times MT$ covariance matrix for all of the disturbances is $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}$ and

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}.^{5} \tag{10-9}$$

The efficient estimator is generalized least squares.[6] The GLS estimator is

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = [\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{X}]^{-1}\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}.$$

Denote the mnth element of $\boldsymbol{\Sigma}^{-1}$ by σ^{mn}. Expanding the **Kronecker products** produces

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \sigma^{11}\mathbf{X}_1'\mathbf{X}_1 & \sigma^{12}\mathbf{X}_1'\mathbf{X}_2 & \cdots & \sigma^{1M}\mathbf{X}_1'\mathbf{X}_M \\ \sigma^{21}\mathbf{X}_2'\mathbf{X}_1 & \sigma^{22}\mathbf{X}_2'\mathbf{X}_2 & \cdots & \sigma^{2M}\mathbf{X}_2'\mathbf{X}_M \\ & & \vdots & \\ \sigma^{M1}\mathbf{X}_M'\mathbf{X}_1 & \sigma^{M2}\mathbf{X}_M'\mathbf{X}_2 & \cdots & \sigma^{MM}\mathbf{X}_M'\mathbf{X}_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{m=1}^{M} \sigma^{1m}\mathbf{X}_1'\mathbf{y}_m \\ \sum_{m=1}^{M} \sigma^{2m}\mathbf{X}_2'\mathbf{y}_m \\ \vdots \\ \sum_{m=1}^{M} \sigma^{Mm}\mathbf{X}_M'\mathbf{y}_m \end{bmatrix}. \tag{10-10}$$

[5] See Appendix Section A.5.5.

[6] See Zellner (1962).

The asymptotic covariance matrix for the GLS estimator is the bracketed inverse matrix in (10-10).[7] All the results of Chapter 9 for the generalized regression model extend to this model.

This estimator is obviously different from ordinary least squares. At this point, however, the equations are linked only by their disturbances—hence the name *seemingly unrelated* regressions model—so it is interesting to ask just how much efficiency is gained by using generalized least squares instead of ordinary least squares. Zellner (1962) and Dwivedi and Srivastava (1978) have noted two important special cases:

1. If the equations are *actually* unrelated—that is, if $\sigma_{mn} = 0$ for $m \neq n$—then there is obviously no payoff to GLS estimation of the full set of equations. Indeed, full GLS is equation by equation OLS.[8]
2. If the equations have **identical explanatory variables**—that is, if $\mathbf{X}_m = \mathbf{X}_n = \mathbf{X}$—then generalized least squares (GLS) is identical to equation by equation ordinary least squares (OLS). This case is common, notably in the capital asset pricing model in empirical finance (see the chapter Introduction) and in VAR models. A proof is considered in the exercises. This general result is lost if there are any restrictions on $\boldsymbol{\beta}$, either within or across equations. (The application in Example 10.2 is one of these cases.) The $\mathbf{X}$ matrices are identical, but there are cross-equation restrictions on the parameters, for example, in (10-4), $\beta_1 = \alpha_3$ and $\gamma_1 = \alpha_4$. Also, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ for this case is given by the large inverse matrix in brackets in (10-10), which would be estimated by Est.Asy.Cov $[\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\beta}}_n] = \hat{\sigma}_{mn}(\mathbf{X}'\mathbf{X})^{-1}$, $m, n = 1, \ldots, M$, where $\hat{\sigma}_{mn} = \mathbf{e}_m'\mathbf{e}_n/T$. For the full set of estimators, Est.Asy.Cov$[\hat{\boldsymbol{\beta}}] = \hat{\boldsymbol{\Sigma}} \otimes (\mathbf{X}'\mathbf{X})^{-1}$.

In the more general case, with unrestricted correlation of the disturbances and different regressors in the equations, the extent to which GLS provides an improvement over OLS is complicated and depends on the data. Two propositions that apply generally are as follows:

1. The greater the correlation of the disturbances, the greater the efficiency gain obtained by using GLS.
2. The less correlation there is between the $\mathbf{X}$ matrices, the greater the gain in efficiency in using GLS.[9]

10.2.3 FEASIBLE GENERALIZED LEAST SQUARES

The computation in (10-10) assumes that $\boldsymbol{\Sigma}$ is known, which, as usual, is unlikely to be the case. FGLS estimators based on the OLS residuals may be used.[10] A first step to estimate the elements of $\boldsymbol{\Sigma}$ uses

$$\hat{\sigma}_{mn} = s_{mn} = \mathbf{e}_m'\mathbf{e}_n/T. \qquad \textbf{(10-11)}$$

[7] A robust covariance matrix along the lines of (10-8) could be constructed. However, note that the structure of $\boldsymbol{\Sigma} = \mathrm{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t']$ has been used explicitly to construct the GLS estimator. The greater generality would be accommodated by assuming that $\mathrm{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'|\mathbf{X}_t] = \boldsymbol{\Sigma}_t$ is not restricted, again, a form of heteroscedasticity robust covariance matrix. This extension is not standard in applications, however. [See Wooldridge (2010, pp. 173–176) for further development.]

[8] See also Kruskal (1968), Baltagi (1989), and Bartels and Fiebig (1992) for other cases where OLS equals GLS.

[9] See Binkley (1982) and Binkley and Nelson (1988).

[10] See Zellner (1962) and Zellner and Huang (1962). The FGLS estimator for this model is also labeled *Zellner's efficient estimator*, or ZEF, in reference to Zellner (1962), where it was introduced.

With

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & s_{22} & \cdots & s_{2M} \\ & \vdots & & \\ s_{M1} & s_{M2} & \cdots & s_{MM} \end{bmatrix} \tag{10-12}$$

in hand, FGLS can proceed as usual.

The FGLS estimator requires inversion of the matrix $\mathbf{S}$ where the mnth element is given by (10-11). This matrix is $M \times M$. It is computed from the least squares residuals using

$$\mathbf{S} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t' = \frac{1}{T}\mathbf{E}'\mathbf{E}, \tag{10-13}$$

where $\mathbf{e}_t'$ is a $1 \times M$ vector containing all M residuals for the M equations at time t, placed as the tth row of the $T \times M$ matrix of residuals, $\mathbf{E}$. The rank of this matrix cannot be larger than T. Note what happens if $M > T$. In this case, the $M \times M$ matrix has rank T, which is less than M, so it must be singular, and the FGLS estimator cannot be computed. In Example 10.1, we aggregate the 48 states into $M = 9$ regions. It would not be possible to fit a full model for the $M = 48$ states with only $T = 17$ observations. The data set is too short to obtain a positive definite estimate of $\mathbf{\Sigma}$.

10.2.4 TESTING HYPOTHESES

For testing a hypothesis about $\boldsymbol{\beta}$, a statistic analogous to the F ratio in multiple regression analysis is

$$F[J, MT - K] = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})/J}{\hat{\boldsymbol{\varepsilon}}'\mathbf{\Omega}^{-1}\hat{\boldsymbol{\varepsilon}}/(MT - K)}. \tag{10-14}$$

The computation uses the the unknown $\mathbf{\Omega}$. If we insert the estimator $\hat{\mathbf{\Omega}}$ based on (10-11) and use the result that the denominator in (10-14) converges to one in T (M is fixed) then, in large samples, the statistic will behave the same as

$$\hat{F} = \frac{1}{J}\left(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q}\right)\left\{\mathbf{R}\,\text{Est.Asy.Var.}\left[\hat{\hat{\boldsymbol{\beta}}}\right]\mathbf{R}'\right\}^{-1}\left(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q}\right). \tag{10-15}$$

This can be referred to the standard F table. Because it uses the estimated $\mathbf{\Sigma}$, even with normally distributed disturbances, the F distribution is only valid approximately. In general, the statistic $F[J, n]$ converges to $1/J$ times a chi-squared $[J]$ as $n \to \infty$. Therefore, an alternative test statistic that has a limiting chi-squared distribution with J degrees of freedom when the null hypothesis is true is

$$J\hat{F} = \left(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q}\right)'\left\{\mathbf{R}\,\text{Est.Asy.Var.}\left[\hat{\hat{\boldsymbol{\beta}}}\right]\mathbf{R}'\right\}^{-1}\left(\mathbf{R}\hat{\hat{\boldsymbol{\beta}}} - \mathbf{q}\right). \tag{10-16}$$

This is a Wald statistic that measures the distance between $\mathbf{R}\hat{\hat{\boldsymbol{\beta}}}$ and $\mathbf{q}$.

One hypothesis of particular interest is the **homogeneity or pooling restriction** of equal coefficient vectors in (10-2). The pooling restriction is that $\boldsymbol{\beta}_m = \boldsymbol{\beta}_M$, $i = 1, \ldots, M - 1$. Consistent with (10-15) and (10-16), we would form the hypothesis as

$$\mathbf{R}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & -\mathbf{I} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & -\mathbf{I} \\ & & \cdots & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \cdots \\ \boldsymbol{\beta}_M \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 - \boldsymbol{\beta}_M \\ \boldsymbol{\beta}_2 - \boldsymbol{\beta}_M \\ \cdots \\ \boldsymbol{\beta}_{M-1} - \boldsymbol{\beta}_M \end{pmatrix} = \mathbf{0}. \qquad \textbf{(10-17)}$$

This specifies a total of $(M - 1)K$ restrictions on the $MK \times 1$ parameter vector. Denote the estimated asymptotic covariance for $\left(\hat{\hat{\boldsymbol{\beta}}}_m, \hat{\hat{\boldsymbol{\beta}}}_n\right)$ as $\hat{\mathbf{V}}_{mn}$. The matrix in braces in (10-16) would have the typical $K \times K$ block,

$$\left\{\mathbf{R}\,\text{Est.Asy.Var.}\left[\hat{\hat{\boldsymbol{\beta}}}\right]\mathbf{R}'\right\}_{mn} = \hat{\mathbf{V}}_{mn} - \hat{\mathbf{V}}_{mM} - \hat{\mathbf{V}}_{Mn} + \hat{\mathbf{V}}_{MM}.$$

It is also of interest to assess statistically whether the off-diagonal elements of $\boldsymbol{\Sigma}$ are zero. If so, then the efficient estimator for the full parameter vector, absent within group heteroscedasticity or autocorrelation, is equation-by-equation ordinary least squares. There is no standard test for the general case of the SUR model unless the additional assumption of normality of the disturbances is imposed in (10-1) and (10-2). With normally distributed disturbances, the standard trio of tests, Wald, **likelihood ratio**, and Lagrange multiplier, can be used. The Wald test is likely to be quite cumbersome. The likelihood ratio statistic for testing the null hypothesis that the matrix $\boldsymbol{\Sigma}$ in (10-3) is a diagonal matrix against the alternative that Σ is simply an unrestricted positive definite matrix would be

$$\lambda_{LR} = T[\ln|\mathbf{S}_0| - \ln|\mathbf{S}_1|], \qquad \textbf{(10-18)}$$

where $\mathbf{S}_1$ is the residual covariance matrix defined in (10-12) (without a degrees of freedom correction). The residuals are computed using maximum likelihood estimates of the parameters, not FGLS.[11] Under the null hypothesis, the model would be efficiently estimated by individual equation OLS, so

$$\ln|\mathbf{S}_0| = \sum_{m=1}^{M} \ln(\mathbf{e}_m'\mathbf{e}_m/T).$$

The statistic would be used for a chi-squared test with $M(M - 1)/2$ degrees of freedom. The Lagrange multiplier statistic developed by Breusch and Pagan (1980) is

$$\lambda_{LM} = T\sum_{m=2}^{M}\sum_{n=1}^{m-1} r_{mn}^2, \qquad \textbf{(10-19)}$$

based on the sample correlation matrix of the M sets of T OLS residuals. This has the same large sample distribution under the null hypothesis as the likelihood ratio statistic, but is obviously easier to compute, as it only requires the OLS residuals. Alternative approaches that have been suggested, such as the LR test in (10-18), are based on the "excess variation," $(\hat{\boldsymbol{\Sigma}}_0 - \hat{\boldsymbol{\Sigma}}_1)$.[12]

[11] In the SUR model of this chapter, the MLE for normally distributed disturbances can be computed by iterating the FGLS procedure, back and forth between (10-10) and (10-12), until the estimates are no longer changing.

[12] See, for example, Johnson and Wichern (2005, p. 424).

10.2.5 THE POOLED MODEL

If the variables in $\mathbf{X}_m$ are all the same and the coefficient vectors in (10-2) are assumed all to be equal, then the **pooled model**,

$$y_{mt} = \mathbf{x}_{mt}'\boldsymbol{\beta} + \varepsilon_{mt},$$

results. Collecting the T observations for group m, we obtain

$$\mathbf{y}_m = \mathbf{X}_m\boldsymbol{\beta} + \boldsymbol{\varepsilon}_m.$$

For all M groups,

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_M\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_M\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\\ \vdots\\ \boldsymbol{\varepsilon}_M\end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{10-20}$$

where

$$\begin{aligned} E[\boldsymbol{\varepsilon}_i|\mathbf{X}] &= \mathbf{0},\\ E[\boldsymbol{\varepsilon}_m\boldsymbol{\varepsilon}_n'|\mathbf{X}] &= \sigma_{mn}\mathbf{I}, \end{aligned} \tag{10-21}$$

or

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\Sigma} \otimes \mathbf{I}.$$

The generalized least squares estimator under this assumption is

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [\mathbf{X}'(\boldsymbol{\Sigma} \otimes \mathbf{I})^{-1}\mathbf{X}]^{-1}[\mathbf{X}'(\boldsymbol{\Sigma} \otimes \mathbf{I})^{-1}\mathbf{y}]\\ &= \left[\sum_{m=1}^{M}\sum_{n=1}^{M}\sigma^{mn}\mathbf{X}_m'\mathbf{X}_n\right]^{-1}\left[\sum_{m=1}^{M}\sum_{n=1}^{M}\sigma^{mn}\mathbf{X}_m'\mathbf{y}_n\right]. \end{aligned} \tag{10-22}$$

The FGLS estimator can be computed using (10-11), where $\mathbf{e}_m$ would be a subvector of the pooled OLS residual vector using all MT observations.

Example 10.1 A Regional Production Model for Public Capital

Munnell (1990) proposed a model of productivity of public capital at the state level. The central equation of the analysis that we will extend here is a Cobb–Douglas production function,

$$\begin{aligned} \ln gsp_{mt} = {} & \alpha_m + \beta_{1m}\ln pc_{mt} + \beta_{2m}\ln hwy_{mt} + \beta_{3m}\ln water_{mt} + \beta_{4m}\ln util_{mt}\\ & + \beta_{5m}\ln emp_{mt} + \beta_{6m}\, unemp_{mt} + \varepsilon_{mt}, \end{aligned}$$

where

gsp	=	gross state product,
pc	=	private capital,
hwy	=	highway capital,
water	=	water utility capital,
util	=	utility capital,
emp	=	employment (labor),
unemp	=	unemployment rate.

The data, measured for the 48 contiguous states in the U.S. (excluding Alaska and Hawaii) and years 1970–1986 are given in Appendix Table F10.1. We will aggregate the data for the 48 states into nine regions consisting of the following groups of states (the state codes appear in the data file):

Gulf States = GF = AL, FL, LA, MS,
Southwest = SW = AZ, NV, NM, TX, UT,
West Central = WC = CA, OR, WA,
Mountain = MT = CO, ID, MT, ND, SD, WY,
Northeast = NE = CT, ME, MA, NH, RI, VT,
Mid Atlantic = MA = DE, MD, NJ, NY, PA, VA,
South = SO = GA, NC, SC, TN, WV, AR,
Midwest = MW = IL, IN, KY, MI, MN, OH, WI,
Central = CN = IA, KS, MO, NE, OK.

This defines a nine-equation model. Note that with only 17 observations per state, it is not possible to fit the unrestricted 48-equation model. This would be a case of the short rank problem noted at the end of Section 10.2.2. The calculations for the data setup are are described in Application 1 at the end of this chapter, where the reader is invited to replicate the computations and fill in the omitted parts of Table 10.3.

We initially estimated the nine equations of the regional productivity model separately by OLS. The OLS estimates are shown in Table 10.1. (For brevity, the estimated standard errors are not shown.)

The correlation matrix for the OLS residuals is shown in Table 10.2.

TABLE 10.1 Estimates of Seemingly Unrelated Regression Equations

Region		α	β_1	β_2	β_3	β_4	β_5	β_6	R^2
GF	OLS	11.570	0.002	−2.028	0.101	1.358	0.805	−0.007	0.997
	FGLS	12.310	−0.201	−1.886	0.178	1.190	0.953	−0.003	
SW	OLS	3.028	0.164	−0.075	−0.169	0.637	0.362	−0.017	0.998
	FGLS	4.083	0.077	−0.131	−0.136	0.522	0.539	−0.156	
WC	OLS	3.590	0.295	0.174	−0.226	−0.215	0.917	−0.008	0.994
	FGLS	1.960	0.170	0.132	−0.347	0.895	1.070	−0.006	
MT	OLS	6.378	−0.153	−0.123	0.306	−0.533	1.344	0.005	0.999
	FGLS	3.463	−0.115	0.180	0.262	−0.330	1.079	−0.002	
NE	OLS	−13.730	−0.020	0.661	−0.969	−0.107	3.380	0.034	0.985
	FGLS	−12.294	−0.118	0.934	−0.557	−0.290	2.494	0.020	
MA	OLS	−22.855	−0.378	3.348	−0.264	−1.778	2.637	0.026	0.986
	FGLS	−18.616	−0.311	3.060	−0.109	−1.659	2.186	0.018	
SO	OLS	3.922	0.043	−0.773	−0.035	0.137	1.665	0.008	0.994
	FGLS	3.162	−0.063	−0.641	−0.081	0.281	1.620	0.008	
MW	OLS	−9.111	0.233	1.604	0.717	−0.356	−0.259	−0.034	0.989
	FGLS	−9.258	0.096	1.612	0.694	−0.340	−0.062	−0.031	
CN	OLS	−5.621	0.386	1.267	0.546	−0.108	−0.475	−0.313	0.995
	FGLS	−3.405	0.295	0.934	0.539	0.003	−0.321	−0.030	

TABLE 10.2 Correlations of OLS Residuals

	GF	*SW*	*WC*	*MT*	*NE*	*MA*	*SO*	*MW*	*CN*
GF	1								
SW	0.173	1							
WC	0.447	0.697	1						
MT	−0.547	−0.290	−0.537	1					
NE	0.525	0.489	0.343	−0.241	1				
MA	0.425	0.132	0.130	−0.322	0.259	1			
SO	0.763	0.314	0.505	−0.351	0.783	0.388	1		
MW	0.167	0.565	0.574	−0.058	0.269	−0.037	0.366	1	
CN	0.325	0.119	0.037	0.091	0.200	0.713	0.350	0.298	1

The correlations are large enough to suggest that there is substantial correlation of the disturbances across regions. The LM statistic in (10-19) for testing the hypothesis that the covariance matrix of the disturbances is diagonal equals 103.1 with 8(9)/2 = 36 degrees of freedom. The critical value from the chi-squared table is 50.998, so the null hypothesis that $\sigma_{mn} = 0$ (or $\rho_{mn} = 0$) for all $m \neq n$, that is, that the seemingly unrelated regressions are actually unrelated, is rejected on this basis. Table 10.1 also presents the FGLS estimates of the model parameters. These are computed in two steps, with the first-step OLS results producing the estimate of $\boldsymbol{\Sigma}$ for FGLS. The correlations in Table 10.2 suggest that there is likely to be considerable benefit to using FGLS in terms of efficiency of the estimator. The individual equation OLS estimators are consistent, but they neglect the cross-equation correlation and heteroscedasticity. A comparison of some of the estimates for the main capital and labor coefficients appears in Table 10.3. The estimates themselves are comparable. But the estimated standard errors for the FGLS coefficients are roughly half as large as the corresponding OLS values. This suggests a large gain in efficiency from using GLS rather than OLS.

The pooling restriction is formulated as

$$H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_M,$$
$$H_1: \text{Not } H_0.$$

TABLE 10.3 Comparison of OLS and FGLS Estimates*

	$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_5$	
Region	***OLS***	***FGLS***	***OLS***	***FGLS***
GF	0.002 (0.301)	−0.201 (0.142)	0.805 (0.159)	0.953 (0.085)
SW	0.164 (0.166)	0.077 (0.086)	0.362 (0.165)	0.539 (0.085)
WC	0.295 (0.205)	0.170 (0.092)	0.917 (0.377)	1.070 (0.171)
MT	−0.153 (0.084)	−0.115 (0.048)	1.344 (0.188)	1.079 (0.105)
NE	−0.020 (0.286)	−0.118 (0.131)	3.380 (1.164)	2.494 (0.479)
MA	−0.378 (0.167)	−0.311 (0.081)	2.673 (1.032)	2.186 (0.448)
SO	0.043 (0.279)	−0.063 (0.104)	1.665 (0.414)	1.620 (0.185)
MW	0.233 (0.206)	0.096 (0.102)	−0.259 (0.303)	−0.062 (0.173)
CN	0.386 (0.211)	0.295 (0.090)	−0.475 (0.259)	−0.321 (0.169)
Pooled	0.260 (0.017)	0.254 (0.006)	0.330 (0.030)	0.343 (0.001)

*Estimates of Capital (β_1) and Labor (β_5) coefficients. Estimated standard errors in parentheses.

The **R** matrix for this hypothesis is shown in (10-17). The test statistic is in (10-16). For our model with nine equations and seven parameters in each, the null hypothesis imposes (9-1)7 = 56 restrictions. The computed test statistic is 6092.5, which is far larger than the critical value from the chi-squared table, 74.468. So, the hypothesis of homogeneity is rejected. Part of the pooled estimator is shown in Table 10.3. The benefit of the restrictions on the estimator can be seen in the much smaller standard errors in every case compared to the separate estimators. If the hypothesis that all the coefficient vectors were the same were true, the payoff to using that information would be obvious. Because the hypothesis is rejected, that benefit is less clear, as now the pooled estimator does not consistently estimate any of the individual coefficient vectors.

10.3 SYSTEMS OF DEMAND EQUATIONS: SINGULAR SYSTEMS

Many of the applications of the seemingly unrelated regression model have estimated **systems of demand equations**, either commodity demands, factor demands, or factor share equations in studies of production. Each is merely a particular application of the model of Section 10.2. But some special problems arise in these settings. First, the parameters of the systems are usually constrained across the equations. This usually takes the form of parameter equality constraints across the equations, such as the symmetry assumption in production and cost models—see (10-32) and (10-33).[13] A second feature of many of these models is that the disturbance covariance matrix $\mathbf{\Sigma}$ is singular, which would seem to preclude GLS (or FGLS).

10.3.1 COBB–DOUGLAS COST FUNCTION

Consider a **Cobb–Douglas** production function,

$$Q = \alpha_0 \prod_{m=1}^{M} x_m^{\alpha_m}.$$

Profit maximization with an exogenously determined output price calls for the firm to maximize output for a given cost level C (or minimize costs for a given output Q). The Lagrangean for the maximization problem is

$$\Lambda = \alpha_0 \prod_{m=1}^{M} x_m^{\alpha_m} + \lambda(C - \mathbf{p}'\mathbf{x}),$$

where $\mathbf{p}$ is the vector of M factor prices. The necessary conditions for maximizing this function are

$$\frac{\partial \Lambda}{\partial x_m} = \frac{\alpha_m Q}{x_m} - \lambda p_m = 0 \quad \text{and} \quad \frac{\partial \Lambda}{\partial \lambda} = C - \mathbf{p}'\mathbf{x} = 0.$$

The joint solution provides $x_m(Q, \mathbf{p})$ and $\lambda(Q, \mathbf{p})$. The total cost of production is then

$$\sum_{m=1}^{M} p_m x_m = \sum_{m=1}^{M} \frac{\alpha_m Q}{\lambda}.$$

The cost share allocated to the mth factor is

[13] See Silver and Ali (1989) for a discussion of testing symmetry restrictions.

$$\frac{p_m x_m}{\sum_{m=1}^{M} p_m x_m} = \frac{\alpha_m}{\sum_{m=1}^{M} \alpha_m} = \beta_m. \tag{10-23}$$

The full model is[14]

$$\begin{aligned} \ln C &= \beta_0 + \beta_q \ln Q + \sum_{m=1}^{M} \beta_m \ln p_m + \varepsilon_c, \\ s_m &= \beta_m + \varepsilon_m, m = 1, \ldots, M. \end{aligned} \tag{10-24}$$

Algebraically, $\sum_{m=1}^{M} \beta_m = 1$ and $\sum_{m=1}^{M} s_m = 1$. (This is the cost function analysis begun in Example 6.17. We will return to that application below.) The cost shares will also sum identically to one in the data. It therefore follows that $\sum_{m=1}^{M} \varepsilon_m = 0$ at every data point so the system is singular. For the moment, ignore the cost function. Let the $M \times 1$ disturbance vector from the shares be $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_M]'$. Because $\boldsymbol{\varepsilon}'\mathbf{i} = 0$, where $\mathbf{i}$ is a column of 1s, it follows that $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{i}] = \boldsymbol{\Sigma}\mathbf{i} = \mathbf{0}$, which implies that $\boldsymbol{\Sigma}$ is singular. Therefore, the methods of the previous sections cannot be used here. (You should verify that the *sample* covariance matrix of the OLS residuals will also be singular.)

The solution to the singularity problem appears to be to drop one of the equations, estimate the remainder, and solve for the last parameter from the other $M - 1$. The constraint $\sum_{m=1}^{M} \beta_m = 1$ states that the cost function must be homogeneous of degree one in the prices. If we impose the constraint

$$\beta_M = 1 - \beta_1 - \beta_2 - \cdots - \beta_{M-1}, \tag{10-25}$$

then the system is reduced to a nonsingular one,

$$\begin{aligned} \ln\left(\frac{C}{p_M}\right) &= \beta_0 + \beta_q \ln Q + \sum_{m=1}^{M-1} \beta_m \ln\left(\frac{p_m}{p_M}\right) + \varepsilon_c, \\ s_m &= \beta_m + \varepsilon_m, \quad m = 1, \ldots, M - 1. \end{aligned}$$

This system provides estimates of β_0, β_q, and $\beta_1, \ldots, \beta_{M-1}$. The last parameter is estimated using (10-25). It is immaterial which factor is chosen as the numeraire; FGLS will be **invariant** to which factor is chosen.

Example 10.2 Cobb–Douglas Cost Function

Nerlove's (1963) study of the electric power industry that we examined in Example 6.6 provides an application of the Cobb–Douglas cost function model. His ordinary least squares estimates of the parameters were listed in Example 6.6. Among the results are (unfortunately) a negative capital coefficient in three of the six regressions. Nerlove also found that the simple Cobb–Douglas model did not adequately account for the relationship between output and average cost. Christensen and Greene (1976) further analyzed the Nerlove data and augmented the data set with cost share data to estimate the complete **demand system**. Appendix Table F6.2 lists Nerlove's 145 observations with Christensen and Greene's cost share data. Cost is the total cost of generation in millions of dollars, output is in millions of kilowatt-hours, the capital price is an index of construction costs, the wage rate is in dollars per hour for production and maintenance, the fuel price is an index of the cost per BTU of fuel purchased by the firms, and the data reflect the 1955 costs of production. The regression estimates are given in Table 10.4.

[14]We leave as an exercise the derivation of β_0, which is a mixture of all the parameters, and β_q, which equals $1/\Sigma_m \alpha_m$.

Least squares estimates of the Cobb–Douglas cost function are given in the first column. The coefficient on capital is negative. Because $\beta_m = \beta_q \partial \ln Q / \partial \ln x_m$—that is, a positive multiple of the output elasticity of the mth factor—this finding is troubling. The third column presents the constrained FGLS estimates. To obtain the constrained estimator, we set up the model in the form of the pooled SUR estimator in (10-20),

$$\mathbf{y} = \begin{bmatrix} \ln(C/P_f) \\ s_k \\ s_l \end{bmatrix} = \begin{bmatrix} \mathbf{i} & \ln Q & \ln(P_k/P_f) & \ln(P_l/P_f) \\ \mathbf{0} & \mathbf{0} & \mathbf{i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{i} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_q \\ \beta_k \\ \beta_l \end{pmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_c \\ \boldsymbol{\varepsilon}_k \\ \boldsymbol{\varepsilon}_l \end{bmatrix}.$$

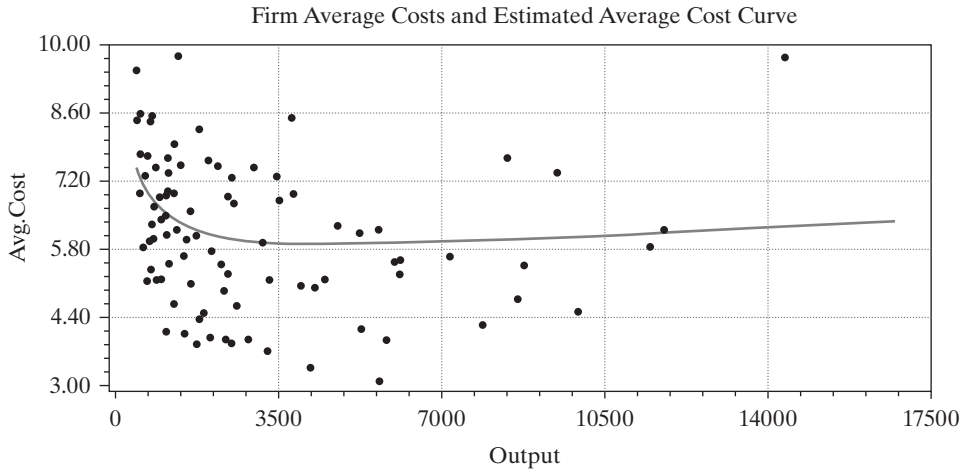
Note this formulation imposes the restrictions $\beta_1 = \alpha_3$ and $\gamma_1 = \alpha_4$ on (10-4). There are $3(145) = 435$ observations in the data matrices. The estimator is then FGLS, as shown in (10-22). An additional column is added for the log quadratic model. Two things to note are the dramatically smaller standard errors and the now positive (and reasonable) estimate of the capital coefficient. The estimates of economies of scale in the basic Cobb–Douglas model are $1/\beta_q = 1.39$ (column 1) and 1.31 (column 3), which suggest some increasing returns to scale. Nerlove, however, had found evidence that at extremely large firm sizes, economies of scale diminished and eventually disappeared. To account for this (essentially a classical U-shaped average cost curve), he appended a quadratic term in log output in the cost function. The single equation and FGLS estimates are given in the second and fourth sets of results.

The quadratic output term gives the average cost function the expected U-shape. We can determine the point where average cost reaches its minimum by equating $\partial \ln C / \partial \ln Q$ to 1. This is $Q* = \exp[(1 - \beta_q)/(2\beta_{qq})]$. Using the FGLS estimates, this value is $Q* = 4{,}669$. (Application 5 considers using the delta method to construct a confidence interval for Q^*.) About 85% of the firms in the sample had output less than this, so by these estimates, most firms in the sample had not yet exhausted the available economies of scale. Figure 10.1 shows predicted and actual average costs for the sample. (To obtain a reasonable scale, the smallest one third of the firms are omitted from the figure.) Predicted average costs are computed at the sample averages of the input prices. The figure does reveal that that beyond a quite small scale, the economies of scale, while perhaps statistically significant, are economically quite small.

TABLE 10.4 Cost Function Estimates (Estimated standard errors in parentheses)

		Ordinary Least Squares		*Constrained Feasible GLS*	
Constant	β_0	−4.686	−3.764	−7.069	−5.707
		(0.885)	(0.702)	(0.107)	(0.165)
ln *Output*	β_q	0.721	0.153	0.766	0.239
		(0.0174)	(0.0618)	(0.0154)	(0.0587)
$\ln^2$ *Output*	β_{qq}		0.0505		0.0451
			(0.0054)		(0.00508)
ln $P_{capital}$	β_k	−0.0085	0.0739	0.424	0.425
		(0.191)	(0.150)	(0.00946)	(0.00943)
ln P_{labor}	β_l	0.594	0.481	0.106	0.106
		(0.205)	(0.161)	(0.00386)	(0.00380)
ln P_{fuel}	β_f	0.414	0.445	0.470	0.470
		(0.0989)	(0.0777)	(0.0101)	(0.0100)

FIGURE 10.1 Predicted Average Costs.



10.3.2 FLEXIBLE FUNCTIONAL FORMS: THE TRANSLOG COST FUNCTION

The classic paper by Arrow et al. (1961) called into question the inherent restriction of the popular Cobb–Douglas model that all elasticities of factor substitution are equal to one. Researchers have since developed numerous **flexible functions** that allow substitution to be unrestricted.[15] Similar strands of literature have appeared in the analysis of commodity demands.[16] In this section, we examine in detail a specific model of production.

Suppose that production is characterized by a production function, $Q = f(\mathbf{x})$. The solution to the problem of minimizing the cost of producing a specified output rate given a set of factor prices produces the cost-minimizing set of factor demands $x_m* = x_m(Q, \mathbf{p})$. The total cost of production is given by the cost function,

$$C = \sum_{m=1}^{M} p_m x_m(Q, \mathbf{p}) = C(Q, \mathbf{p}). \tag{10-26}$$

If there are **constant returns to scale**, then it can be shown that $C = Qc(\mathbf{p})$ or $C/Q = c(\mathbf{p})$, where $c(\mathbf{p})$ is the per unit or average cost function.[17] The cost-minimizing factor demands are obtained by applying **Shephard's lemma (1970)**, which states that if $C(Q, \mathbf{p})$ gives the minimum total cost of production, then the cost-minimizing set of factor demands is given by

$$x_m^* = \frac{\partial C(Q, \mathbf{p})}{\partial p_m}. \tag{10-27}$$

[15]See, in particular, Berndt and Christensen (1973).

[16]See, for example, Christensen, Jorgenson, and Lau (1975) and two surveys, Deaton and Muellbauer (1980) and Deaton (1983). Berndt (1990) contains many useful results.

[17]The Cobb–Douglas function of the previous section gives an illustration. The restriction of constant returns to scale is $\beta_q = 1$, which is equivalent to $C = Qc(\mathbf{p})$. Nerlove's more general version of the cost function allows nonconstant returns to scale. See Christensen and Greene (1976) and Diewert (1974) for some of the formalities of the cost function and its relationship to the structure of production.

Alternatively, by differentiating logarithmically, we obtain the cost-minimizing factor cost shares,

$$s_m^* = \frac{\partial \ln C(Q, \mathbf{p})}{\partial \ln p_m} = \frac{p_m}{C}\frac{\partial C(Q, \mathbf{p})}{\partial p_m} = \frac{p_m x_m^*}{C}. \tag{10-28}$$

With constant returns to scale, $\ln C(Q, \mathbf{p}) = \ln Q + \ln c(\mathbf{p})$, so

$$s_m^* = \frac{\partial \ln c(\mathbf{p})}{\partial \ln p_m}. \tag{10-29}$$

In many empirical studies, the objects of estimation are the elasticities of factor substitution and the own price elasticities of demand, which are given by

$$\theta_{mn} = \frac{c(\partial^2 c/\partial p_m \partial p_n)}{(\partial c/\partial p_m)(\partial c/\partial p_n)}$$

and

$$\eta_m = s_m \theta_{mm}.$$

By suitably parameterizing the cost function (10-26) and the cost shares (10-29), we obtain an M or $M + 1$ equation econometric model that can be used to estimate these quantities.

The transcendental logarithmic or **translog function** is the most frequently used flexible function in empirical work.[18] By expanding $\ln c(\mathbf{p})$ in a second-order Taylor series about the point $\ln(\mathbf{p}) = \mathbf{0}$, we obtain

$$\ln c \approx \beta_0 + \sum_{m=1}^{M}\left(\frac{\partial \ln c}{\partial \ln p_m}\right)\log p_m + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M}\left(\frac{\partial^2 \ln c}{\partial \ln p_m\, \partial \ln p_n}\right)\ln p_m \ln p_n, \tag{10-30}$$

where all derivatives are evaluated at the expansion point. If we treat these derivatives as the coefficients, then the cost function becomes

$$\begin{aligned}\ln c = \beta_0 + \beta_1 \ln p_1 + \cdots + \beta_M \ln p_M + \delta_{11}\left(\frac{1}{2}\ln^2 p_1\right) + \delta_{12}\ln p_1 \ln p_2 \\ + \delta_{22}\left(\frac{1}{2}\ln^2 p_2\right) + \cdots + \delta_{MM}\left(\frac{1}{2}\ln^2 p_M\right).\end{aligned} \tag{10-31}$$

This is the translog cost function. If δ_{mn} equals zero, then it reduces to the Cobb–Douglas function in Section 10.3.1. The cost shares are given by

$$\begin{aligned}
s_1 &= \frac{\partial \ln c}{\partial \ln p_1} = \beta_1 + \delta_{11}\ln p_1 + \delta_{12}\ln p_2 + \cdots + \delta_{1M}\ln p_M, \\
s_2 &= \frac{\partial \ln c}{\partial \ln p_2} = \beta_2 + \delta_{21}\ln p_1 + \delta_{22}\ln p_2 + \cdots + \delta_{2M}\ln p_M, \\
&\vdots \\
s_M &= \frac{\partial \ln c}{\partial \ln p_M} = \beta_M + \delta_{M1}\ln p_1 + \delta_{M2}\ln p_2 + \cdots + \delta_{MM}\ln p_M.
\end{aligned} \tag{10-32}$$

[18]The function was proposed in a series of papers by Berndt, Christensen, Jorgenson, and Lau, including Berndt and Christensen (1973) and Christensen et al. (1975).

The theory implies a number of restrictions on the parameters. The matrix of second derivatives must be symmetric (by Young's theorem for continuous functions). The cost function must be linearly homogeneous in the factor prices. This implies $\Sigma_{m=1}^{M}(\partial \ln c(\mathbf{p})/\partial \ln p_m) = 1$. This implies the adding-up restriction, $\Sigma_{m=1}^{M} s_m = 1$. Together, these imply the following set of cross-equation restrictions:

$$\begin{aligned} \delta_{mn} &= \delta_{nm} && \text{(symmetry)}, \\ \sum_{m=1}^{M} \beta_m &= 1 && \text{(linear homogeneity)}, \\ \sum_{m=1}^{M} \delta_{mn} &= \sum_{n=1}^{M} \delta_{mn} = 0. \end{aligned} \qquad \textbf{(10-33)}$$

The system of **share equations** in (10-32) produces a seemingly unrelated regressions model that can be used to estimate the parameters of the model.[19] To make the model operational, we must impose the restrictions in (10-33) and solve the problem of **singularity of the disturbance covariance matrix** of the share equations. The first is accomplished by dividing the first $M - 1$ prices by the Mth, thus eliminating the last term in each row and column of the parameter matrix. As in the Cobb–Douglas model, we obtain a nonsingular system by dropping the Mth share equation. For the translog cost function, the elasticities of substitution are particularly simple to compute once the parameters have been estimated,

$$\theta_{mn} = \frac{\delta_{mn} + s_m s_n}{s_m s_n}, \quad \theta_{mm} = \frac{\delta_{mm} + s_m(s_m - 1)}{s_m^2}. \qquad \textbf{(10-34)}$$

These elasticities will differ at every data point. It is common to compute them at some central point such as the means of the data.[20] The factor-specific demand elasticities are then computed using $\eta_m = s_m \theta_{mm}$.

Example 10.3 A Cost Function for U.S. Manufacturing

A number of studies using the translog methodology have used a four-factor model, with capital K, labor L, energy E, and materials M, the factors of production. Among the studies to employ this methodology was Berndt and Wood's (1975) estimation of a translog cost function for the U.S. manufacturing sector. The three factor shares used to estimate the model are

$$\begin{aligned} s_K &= \beta_K + \delta_{KK} \ln\left(\frac{p_K}{p_M}\right) + \delta_{KL} \ln\left(\frac{p_L}{p_M}\right) + \delta_{KE} \ln\left(\frac{p_E}{p_M}\right), \\ s_L &= \beta_L + \delta_{KL} \ln\left(\frac{p_K}{p_M}\right) + \delta_{LL} \ln\left(\frac{p_L}{p_M}\right) + \delta_{LE} \ln\left(\frac{p_E}{p_M}\right), \\ s_E &= \beta_E + \delta_{KE} \ln\left(\frac{p_K}{p_M}\right) + \delta_{LE} \ln\left(\frac{p_L}{p_M}\right) + \delta_{EE} \ln\left(\frac{p_E}{p_M}\right). \end{aligned}$$

[19]The system of factor share equations estimates all of the parameters in the model except for the overall constant term, β_0. The cost function can be omitted from the model. Without the assumption of constant returns to scale, however, the cost function will contain parameters of interest that do not appear in the share equations. In this case, one would want to include it in the equation system. See Christensen and Greene (1976) for an application.

[20]They will also be highly nonlinear functions of the parameters and the data. A method of computing asymptotic standard errors for the estimated elasticities is presented in Anderson and Thursby (1986). Krinsky and Robb (1986, 1990, 1991). (See also Section 15.3.) proposed their method as an alternative approach to this computation.

Berndt and Wood's data are reproduced in Appendix Table F10.2. Constrained FGLS estimates of the parameters presented in Table 10.4 were obtained by constructing the pooled regression in (10-20) with data matrices

$$\mathbf{y} = \begin{bmatrix} \mathbf{s_K} \\ \mathbf{s_L} \\ \mathbf{s_E} \end{bmatrix}, \tag{10-35}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \mathbf{0} & \ln \mathbf{P_K/P_M} & \ln \mathbf{P_L/P_M} & \ln \mathbf{P_E/P_M} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \mathbf{0} & \mathbf{0} & \ln \mathbf{P_K/P_M} & \mathbf{0} & \ln \mathbf{P_L/P_M} & \ln \mathbf{P_K/P_M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{i} & \mathbf{0} & \mathbf{0} & \ln \mathbf{P_K/P_M} & \mathbf{0} & \ln \mathbf{P_L/P_M} & \ln \mathbf{P_E/P_M} \end{bmatrix},$$

$\boldsymbol{\beta}' = (\beta_K, \beta_L, \beta_E, \delta_{KK}, \delta_{KL}, \delta_{KE}, \delta_{LL}, \delta_{LE}, \delta_{EE})$.

Estimates are then obtained by iterating the two-step procedure in (10-11) and (10-22).[21] The parameters not estimated directly in (10-35) are computed using (10-33). The implied estimates of the elasticities of substitution and demand elasticities for 1959 (the central year in the data) are given in TABLE 10.5 using the fitted cost shares and the estimated parameters in (10-34). The departure from the Cobb–Douglas model with unit elasticities is substantial. For example, the results suggest almost no substitutability between energy and labor and some complementarity between capital and energy.

The underlying theory requires that the cost function satisfy three regularity conditions, homogeneity of degree one in the input prices, monotonicity in the prices, and quasiconcavity. The first of these is imposed by (10-33), which we built into the model. The second is obtained if all of the fitted cost shares are positive, which we have verified at every observation. The third requires that the matrix,

$$\mathbf{F}_t = \boldsymbol{\Delta} - diag(\mathbf{s}_t) + \mathbf{s}_t\mathbf{s}_t',$$

TABLE 10.5 Parameter Estimates for Aggregate Translog Cost Function (Standard errors in parentheses)

	Constant	*Capital*	*Labor*	*Energy*	*Materials*
Capital	0.05689	0.02949	−0.00005	−0.01067	−0.01877*
	(0.00135)	(0.00580)	(0.00385)	(0.00339)	(0.00971)
Labor	0.25344		0.07543	−0.00476	−0.07063*
	(0.00223)		(0.00676)	(0.00234)	(0.01060)
Energy	0.04441			0.01835	−0.00294*
	(0.00085)			(0.00499)	(0.00800)
Materials	0.64526*				0.09232*
	(0.00330)				(0.02247)

*Derived using (10-33).

[21]The estimates do not match those reported by Berndt and Wood. To purge their data of possible correlation with the disturbances, they first regressed the prices on 10 exogenous macroeconomic variables, such as U.S. population, government purchases of labor services, real exports of durable goods and U.S. tangible capital stock, and then based their analysis on the fitted values. The estimates given here are, in general, quite close to theirs. For example, their estimates of the constants in Table 10.5 are 0.60564, 0.2539, 0.0442, and 0.6455. Berndt and Wood's estimate of θ_{EL} for 1959 is 0.64 compared to ours in Table 10.5 of 0.60564.

TABLE 10.6 Estimated Elasticities

	Capital	*Labor*	*Energy*	*Materials*
		Cost Shares for 1959		
Fitted	0.05640	0.27452	0.04389	0.62519
Actual	0.06185	0.27303	0.04563	0.61948
		Implied Elasticities of Substitution, 1959		
Capital	−7.4612			
Labor	0.99691	−1.64179		
Energy	−3.31133	0.60533	−12.2566	
Materials	0.46779	0.58848	0.89334	−0.36331
		Implied Own Price Elasticities		
	−0.420799	−0.45070	−0.53793	−0.22714

be negative semidefinite, where $\boldsymbol{\Delta}$ is the symmetric matrix of coefficients on the quadratic terms in Table 10.5 and $\mathbf{s}_t$ is the vector of factor shares. This condition can be checked at each observation by verifying that the characteristic roots of $\mathbf{F}_t$ are all nonpositive. For the 1959 data, the four characteristic roots are (0, −0.00152, −0.06277, −0.23514). The results for the other years are similar. The estimated cost function satisfies the theoretical regularity conditions.

10.4 SIMULTANEOUS EQUATIONS MODELS

The seemingly unrelated regression model,

$$y_{mt} = \mathbf{x}_{mt}'\boldsymbol{\beta}_m + \varepsilon_{mt},$$

derives from a set of regression equations that are related through the disturbances. The regressors, $\mathbf{x}_{mt}$, are *exogenous* and can vary for reasons that are not explained within the model. Thus, the coefficients are directly interpretable as partial or causal effects and can be estimated by least squares or other methods that are based on the conditional mean functions, $E[y_{mt}|\mathbf{x}_{mt}] = \mathbf{x}_{mt}'\boldsymbol{\beta}$. In the market equilibrium model suggested in the Introduction,

$$\begin{aligned} Q_{Demand} &= \alpha_1 + \alpha_2 Price + \alpha_3 Income + \mathbf{d}'\boldsymbol{\alpha} + \varepsilon_{Demand}, \\ Q_{Supply} &= \beta_1 + \beta_2 Price + \beta_3 FactorPrice + \mathbf{s}'\boldsymbol{\beta} + \varepsilon_{Supply}, \\ Q_{Equilibrium} &= Q_{Demand} = Q_{Supply}, \end{aligned}$$

neither of the two market equations is a conditional mean. The partial equilibrium experiment of changing the equilibrium price and inducing a change in the equilibrium quantity in the hope of eliciting an estimate of the demand elasticity, α_2 (or supply elasticity, β_2), makes no sense. The model is of the joint determination of quantity *and* price. Price changes when the market equilibrium changes, but that is induced by changes in other factors, such as changes in incomes or other variables that affect the supply function. Nonetheless, the elasticities of demand and supply, α_2 and β_2, are of interest, and do have a causal interpretation in the context of the model. This section considers the theory and methods that apply for estimation and analysis of systems of interdependent equations.

As we saw in Example 8.4, least squares regression of observed equilibrium quantities on price and the other factors will compute an ambiguous mixture of the supply and demand functions. The result follows from the *endogeneity* of *Price* in either equation. Simultaneous equations models arise in settings such as this one, in which the set of equations are interdependent. Simultaneous equations models will fit in the framework developed in Chapter 8, where we considered equations in which some of the right-hand-side variables are endogenous—that is, correlated with the disturbances. The substantive difference at this point is the source of the endogeneity. In our treatments in Chapter 8, endogeneity arose, for example, in the models of omitted variables, measurement error, or endogenous treatment effects, essentially as an unintended deviation from the assumptions of the linear regression model. In the simultaneous equations framework, endogeneity is a fundamental part of the specification. This section will consider the issues of specification and estimation in systems of simultaneous equations. We begin in Section 10.4.1 with a development of a general framework for the analysis and a statement of some fundamental issues. Section 10.4.2 presents the simultaneous equations model as an extension of the seemingly unrelated regressions model in Section 10.2. The ultimate objective of the analysis will be to learn about the model coefficients. The issue of whether this is even possible is considered in Section 10.4.3, where we develop the issue of identification. Once the identification question is settled, methods of estimation and inference are presented in Sections 10.4.4 and 10.4.5.

Example 10.4. Reverse Causality and Endogeneity in Health

As we examined in Chapter 8, endogeneity arises from several possible sources. The case considered in this chapter is simultaneity, sometimes labeled *reverse causality.* Consider a familiar modeling framework in health economics, the "health production function" (see Grossman (1972)), in which we might model health outcomes as

$$\textit{Health} = f(\textit{Income}, \textit{Education}, \textit{Health Care}, \textit{Age}, \ \ldots, \varepsilon_H = \text{other factors}).$$

It is at least debatable whether this can be treated as a regression. For any individual, arguably, lower incomes are associated with lower results for health. But which way does the "causation run?" It may also be that variation in health is a driver of variation in income. A natural companion might appear

$$\textit{Income} = g(\textit{Health}, \textit{Education}, \ \ldots, \varepsilon_I = \text{labor market factors}).$$

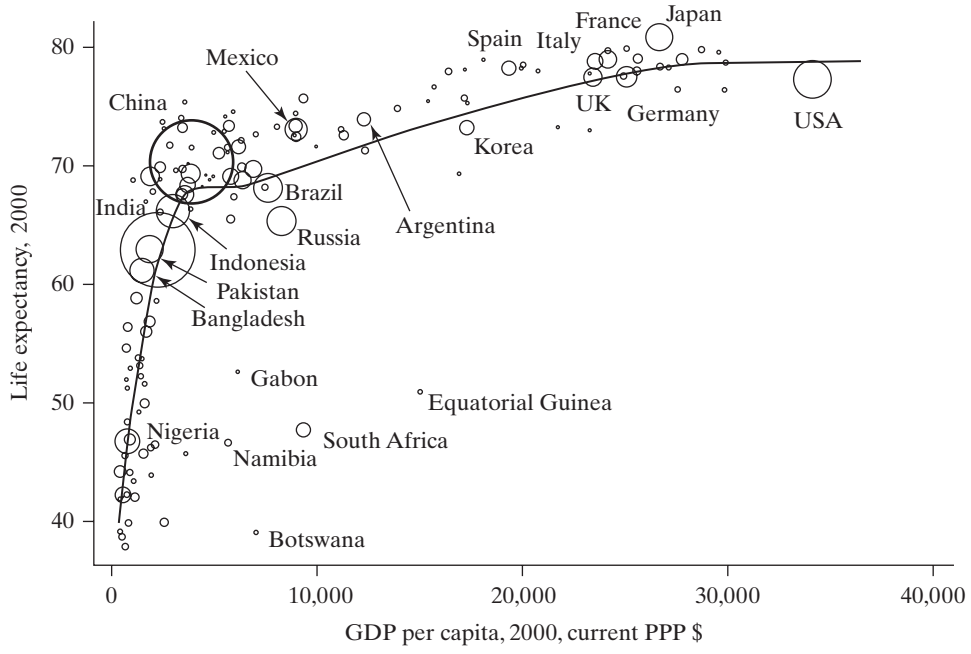
The causal effect of income on health could, in principle, be examined through the experiment of varying income, assuming that external factors such as labor market conditions could be driving the change in income. But, in the second equation, we could likewise be interested in how variation in health outcomes affect incomes. The idea is similarly complicated at the aggregate level. Deaton's (2003) updated version of the "Preston Curve" (1978) in Figure 10.2 suggests covariation between health (life expectancy) and income (per capita GDP) for a group of countries. Which variable is driving which is part of a longstanding discussion.

10.4.1 SYSTEMS OF EQUATIONS

Consider a simplified version of the equilibrium model,

$$\begin{aligned}
\text{demand equation:} \quad & q_{d,t} = \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t}, \\
\text{supply equation:} \quad & q_{s,t} = \beta_1 p_t + \varepsilon_{s,t}, \\
\text{equilibrium condition:} \quad & q_{d,t} = q_{s,t} = q_t.
\end{aligned}$$

FIGURE 10.2 Updated Preston Curve.



These equations are **structural equations** in that they are derived from theory and each purports to describe a particular aspect of the economy. Because the model is one of the joint determination of price and quantity, they are labeled **jointly dependent** or **endogenous** variables. Income, x, is assumed to be determined outside of the model, which makes it **exogenous**. The disturbances are added to the usual textbook description to obtain an **econometric model**. All three equations are needed to determine the equilibrium price and quantity, so the system is **interdependent**. Finally, because an equilibrium solution for price and quantity in terms of income and the disturbances is, indeed, implied (unless α_1 equals β_1), the system is said to be a **complete system of equations**. As a general rule, it is not possible to estimate all the parameters of incomplete systems. (It may be possible to estimate some of them, as will turn out to be the case with this example).

Suppose that interest centers on estimating the demand elasticity α_1. For simplicity, assume that ε_d and ε_s are well behaved, classical disturbances with

$$\begin{aligned} E[\varepsilon_{d,t}|x_t] &= E[\varepsilon_{s,t}|x_t] = 0, \\ E[\varepsilon_{d,t}^2|x_t] &= \sigma_d^2, \\ E[\varepsilon_{s,t}^2|x_t] &= \sigma_s^2, \\ E[\varepsilon_{d,t}\varepsilon_{s,t}|x_t] &= 0. \end{aligned}$$

All variables are mutually uncorrelated with observations at different time periods. Price, quantity, and income are measured in logarithms in deviations from their sample

means. Solving the equations for p and q in terms of x, ε_d, and ε_s produces the **reduced form** of the model,

$$\begin{aligned} p &= \frac{\alpha_2 x}{\beta_1 - \alpha_1} + \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} = \pi_1 x + v_1, \\ q &= \frac{\beta_1 \alpha_2 x}{\beta_1 - \alpha_1} + \frac{\beta_1 \varepsilon_d - \alpha_1 \varepsilon_s}{\beta_1 - \alpha_1} = \pi_2 x + v_2. \end{aligned} \tag{10-36}$$

(Note the role of the "completeness" requirement that α_1 not equal β_1. This means that the two lines are not parallel.) It follows that $\text{Cov}[p, \varepsilon_d] = \sigma_d^2/(\beta_1 - \alpha_1)$ and $\text{Cov}[p, \varepsilon_s] = -\sigma_s^2/(\beta_1 - \alpha_1)$ so neither the demand nor the supply equation satisfies the assumptions of the classical regression model. The price elasticity of demand cannot be consistently estimated by least squares regression of q on x and p. This result is characteristic of simultaneous equations models. Because the endogenous variables are all correlated with the disturbances, the least squares estimators of the parameters of equations with endogenous variables on the right-hand side are inconsistent.[22]

Suppose that we have a sample of T observations on p, q, and x such that

$$\text{plim}(1/T)\mathbf{x}'\mathbf{x} = \sigma_x^2.$$

Because least squares is inconsistent, we might instead use an **instrumental variable estimator**. (See Section 8.3.) The only variable in the system that is not correlated with the disturbances is x. Consider, then, the IV estimator, $\hat{\beta}_1 = (\mathbf{x}'\mathbf{p})^{-1}\mathbf{x}'\mathbf{q}$. This estimator has

$$\text{plim}\ \hat{\beta}_1 = \text{plim}\ \frac{\mathbf{x}'\mathbf{q}/T}{\mathbf{x}'\mathbf{p}/T} = \frac{\sigma_x^2 \beta_1 \alpha_2/(\beta_1 - \alpha_1)}{\sigma_x^2 \alpha_2/(\beta_1 - \alpha_1)} = \beta_1.$$

Evidently, the parameter of the supply curve can be estimated by using an instrumental variable estimator. In the least squares regression of $\mathbf{p}$ on $\mathbf{x}$, the predicted values are $\hat{\mathbf{p}} = (\mathbf{x}'\mathbf{p}/\mathbf{x}'\mathbf{x})\mathbf{x}$. It follows that in the instrumental variable regression the instrument is $\hat{\mathbf{p}}$. That is,

$$\hat{\beta}_1 = \frac{\hat{\mathbf{p}}'\mathbf{q}}{\hat{\mathbf{p}}'\mathbf{p}}.$$

Because $\hat{\mathbf{p}}'\mathbf{p} = \hat{\mathbf{p}}'\hat{\mathbf{p}}$, $\hat{\beta}_1$ is also the slope in a regression of q on these predicted values. This interpretation defines the two-stage least squares estimator.

It would seem natural to use a similar device to estimate the parameters of the demand equation, but unfortunately, we have already used all of the information in the sample. Not only does least squares fail to estimate the demand equation consistently, but without some further assumptions, the sample contains no other information that can be used. This example illustrates the **problem of identification** alluded to in the introduction to this section.

[22] This failure of least squares is sometimes labeled *simultaneous equations bias*.

10.4.2 A GENERAL NOTATION FOR LINEAR SIMULTANEOUS EQUATIONS MODELS[23]

The **structural form** of the model is

$$\begin{aligned}
&\gamma_{11}y_{t1} + \gamma_{21}y_{t2} + \cdots + \gamma_{M1}y_{tM} + \beta_{11}x_{t1} + \cdots + \beta_{K1}x_{tK} = \varepsilon_{t1},\\
&\gamma_{12}y_{t1} + \gamma_{22}y_{t2} + \cdots + \gamma_{M2}y_{tM} + \beta_{12}x_{t1} + \cdots + \beta_{K2}x_{tK} = \varepsilon_{t2},\\
&\qquad\qquad\qquad\qquad\qquad\vdots\\
&\gamma_{1M}y_{t1} + \gamma_{2M}y_{t2} + \cdots + \gamma_{MM}y_{tM} + \beta_{1M}x_{t1} + \cdots + \beta_{KM}x_{tK} = \varepsilon_{tM}.
\end{aligned} \tag{10-37}$$

There are M equations and M endogenous variables, denoted $y_1, \ldots, y_M$. There are K exogenous variables, $x_1, \ldots, x_K$, that may include predetermined values of $y_1, \ldots, y_M$ as well.[24] The first element of $\mathbf{x}_t$ will usually be the constant, 1. Finally, $\varepsilon_{t1}, \ldots, \varepsilon_{tM}$ are the **structural disturbances**. The subscript t will be used to index observations, $t = 1, \ldots, T$.

In matrix terms, the system may be written

$$[y_1 \quad y_2 \quad \cdots \quad y_M]_t \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1M} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2M} \\ & & \vdots & \\ \gamma_{M1} & \gamma_{M2} & \cdots & \gamma_{MM} \end{bmatrix}$$

$$+ [x_1 \quad x_2 \quad \cdots \quad x_K]_t \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ & & \vdots & \\ \beta_{K1} & \beta_{K2} & \cdots & \beta_{KM} \end{bmatrix} = [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_M]_t,$$

or

$$\mathbf{y}_t'\boldsymbol{\Gamma} + \mathbf{x}_t'\mathbf{B} = \boldsymbol{\varepsilon}_t'.$$

Each column of the parameter matrices is the vector of coefficients in a particular equation. The underlying theory will imply a number of restrictions on $\boldsymbol{\Gamma}$ and $\mathbf{B}$. One of the variables in each equation is labeled the *dependent* variable so that its coefficient in the model will be 1. Thus, there will be at least one "1" in each column of $\boldsymbol{\Gamma}$. This **normalization** is not a substantive restriction. The relationship defined for a given equation will be unchanged if every coefficient in the equation is multiplied by the same constant. Choosing a dependent variable simply removes this indeterminacy. If there are any identities, then the corresponding columns of $\boldsymbol{\Gamma}$ and $\mathbf{B}$ will be completely known, and there will be no disturbance for that equation. Because not all variables appear in all equations, some of the parameters will be zero. The theory may also impose other types of restrictions on the parameter matrices.

If $\boldsymbol{\Gamma}$ is an upper triangular matrix, then the system is said to be a **triangular system**. In this case, the model is of the form

[23]We will be restricting our attention to linear models. Nonlinear systems bring forth numerous complications that are beyond the scope of this text. Gallant (1987), Gallant and Holly (1980), Gallant and White (1988), Davidson and MacKinnon (2004), and Wooldridge (2010) provide further discussion.

[24]For the present, it is convenient to ignore the special nature of lagged endogenous variables and treat them the same as strictly exogenous variables.

$$\begin{aligned} y_{t1} &= f_1(\mathbf{x}_t) + \varepsilon_{t1}, \\ y_{t2} &= f_2(y_{t1}, \mathbf{x}_t) + \varepsilon_{t2}, \\ &\vdots \\ y_{tM} &= f_M(y_{t1}, y_{t2}, \ldots, y_{t,M-1}, \mathbf{x}_t) + \varepsilon_{tM}. \end{aligned}$$

The joint determination of the variables is a **recursive model**. The first is completely determined by the exogenous factors. Then, given the first, the second is likewise determined, and so on.

The solution of the system of equations that determines $\mathbf{y}_t$ in terms of $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_t$ is the reduced form of the model,

$$\begin{aligned} \mathbf{y}_t' &= [x_1 \quad x_2 \quad \cdots \quad x_K]_t \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1M} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2M} \\ & & \vdots & \\ \pi_{K1} & \pi_{K2} & \cdots & \pi_{KM} \end{bmatrix} + [v_1 \quad \cdots \quad v_M]_t \\ &= -\mathbf{x}_t'\mathbf{B}\boldsymbol{\Gamma}^{-1} + \boldsymbol{\varepsilon}_t'\boldsymbol{\Gamma}^{-1} \\ &= \mathbf{x}_t'\boldsymbol{\Pi} + \mathbf{v}_t'. \end{aligned}$$

For this solution to exist, the model must satisfy the **completeness condition** for simultaneous equations systems: $\boldsymbol{\Gamma}$ must be nonsingular.

Example 10.5 Structure and Reduced Form in a Small Macroeconomic Model

Consider the model

$$\begin{aligned} \text{consumption: } c_t &= \alpha_0 + \alpha_1 y_t + \alpha_2 c_{t-1} + \varepsilon_{t,c}, \\ \text{investment: } i_t &= \beta_0 + \beta_1 r_t + \beta_2 (y_t - y_{t-1}) + \varepsilon_{t,i}, \\ \text{demand: } y_t &= c_t + i_t + g_t. \end{aligned}$$

The model contains an autoregressive consumption function based on output, y_t, and one lagged value, an investment equation based on interest, r_t, and the growth in output, and an equilibrium condition. The model determines the values of the three endogenous variables c_t, i_t, and y_t. This model is a **dynamic model**. In addition to the exogenous variables r_t and government spending, g_t, it contains two **predetermined variables**, c_{t-1} and y_{t-1}. These are obviously not exogenous, but with regard to the current values of the endogenous variables, they may be regarded as having already been determined. The deciding factor is whether or not they are uncorrelated with the current disturbances, which we might assume. The reduced form of this model is

$$c_t = [\alpha_0(1-\beta_2) + \beta_0\alpha_1 + \alpha_1\beta_1 r_t + \alpha_1 g_t + \alpha_2(1-\beta_2)c_{t-1} - \alpha_1\beta_2 y_{t-1} + (1-\beta_2)\varepsilon_{t,c} + \alpha_1\varepsilon_{t,i}]/\Lambda,$$

$$i_t = [\alpha_0\beta_2 + \beta_0(1-\alpha_1) + \beta_1(1-\alpha_1)r_t + \beta_2 g_t + \alpha_2\beta_2 c_{t-1} - \beta_2(1-\alpha_1)y_{t-1} + \beta_2\varepsilon_{t,c} + (1-\alpha_1)\varepsilon_{t,i}]/\Lambda,$$

$$y_t = [\alpha_0 + \beta_0 + \beta_1 r_t + g_t + \alpha_2 c_{t-1} - \beta_2 y_{t-1} + \varepsilon_{t,c} + \varepsilon_{t,i}]/\Lambda,$$

where $\Lambda = 1 - \alpha_1 - \beta_2$. The completeness condition is that $\alpha_1 + \beta_2$ not equal one. Note that the reduced form preserves the equilibrium condition, $y_t = c_t + i_t + g_t$. Denote $\mathbf{y}' = [c, i, y]$, $\mathbf{x}' = [1, r, g, c_{-1}, y_{-1}]$ and

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -\alpha_1 & -\beta_2 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -\alpha_0 & -\beta_0 & 0 \\ 0 & -\beta_1 & 0 \\ 0 & 0 & -1 \\ -\alpha_2 & 0 & 0 \\ 0 & \beta_2 & 0 \end{bmatrix}, \quad \boldsymbol{\Gamma}^{-1} = \frac{1}{\Lambda}\begin{bmatrix} 1-\beta_2 & \beta_2 & 1 \\ \alpha_1 & 1-\alpha_1 & 1 \\ \alpha_1 & \beta_2 & 1 \end{bmatrix}.$$

Then, the reduced form coefficient matrix is

$$\boldsymbol{\Pi}' = \frac{1}{\Lambda}\begin{bmatrix} \alpha_0(1-\beta_2)+\beta_0\alpha_1 & \alpha_1\beta_1 & \alpha_1 & \alpha_2(1-\beta_2) & -\beta_2\alpha_1 \\ \alpha_0\beta_2+\beta_0(1-\alpha_1) & \beta_1(1-\alpha_1) & \beta_2 & \alpha_2\beta_2 & -\beta_2(1-\alpha_1) \\ \alpha_0+\beta_0 & \beta_1 & 1 & \alpha_2 & -\beta_2 \end{bmatrix}.$$

There is an ambiguity in the interpretation of coefficients in a simultaneous equations model. The effects in the structural form of the model would be labeled "causal," in that they are derived directly from the underlying theory. However, in order to trace through the effects of autonomous changes in the variables in the model, it is necessary to work through the reduced form. For example, the interest rate does not appear in the consumption function. But that does not imply that changes in r_t would not "cause" changes in consumption, because changes in r_t change investment, which impacts demand which, in turn, does appear in the consumption function. Thus, we can see from the reduced form that $\Delta c_t/\Delta r_t = \alpha_1\beta_1/\Lambda$. Similarly, the "experiment," $\Delta c_t/\Delta y_t$ is meaningless without first determining what caused the change in y_t. If the change were induced by a change in the interest rate, we would find $(\Delta c_t/\Delta r_t)/(\Delta y_t/\Delta r_t) = (\alpha_1\beta_1/\Lambda)/(\beta_1/\Lambda) = \alpha_1$.

The structural disturbances are assumed to be randomly drawn from an M-variate distribution with

$$E[\boldsymbol{\varepsilon}_t|\mathbf{x}_t] = \mathbf{0} \quad \text{and} \quad E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'|\mathbf{x}_t] = \boldsymbol{\Sigma}.$$

For the present, we assume that

$$E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s'|\mathbf{x}_t, \mathbf{x}_s] = \mathbf{0}, \quad \forall t, s.$$

It will occasionally be useful to assume that $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution, but we shall postpone this assumption until it becomes necessary. It may be convenient to retain the identities without disturbances as separate equations. If so, then one way to proceed with the stochastic specification is to place rows and columns of zeros in the appropriate places in $\boldsymbol{\Sigma}$. It follows that the **reduced-form disturbances**, $\mathbf{v}_t' = \boldsymbol{\varepsilon}_t'\boldsymbol{\Gamma}^{-1}$, have

$$E[\mathbf{v}_t|\mathbf{x}_t] = (\boldsymbol{\Gamma}^{-1})'\mathbf{0} = \mathbf{0},$$
$$E[\mathbf{v}_t\mathbf{v}_t'|\mathbf{x}_t] = (\boldsymbol{\Gamma}^{-1})'\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Omega}.$$

This implies that

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Omega}\boldsymbol{\Gamma}.$$

The preceding formulation describes the model as it applies to an observation $[\mathbf{y}', \mathbf{x}', \boldsymbol{\varepsilon}']_t$ at a particular point in time or in a cross section. In a sample of data, each joint observation will be one row in a data matrix,

$$[\mathbf{Y} \quad \mathbf{X} \quad \mathbf{E}] = \begin{bmatrix} \mathbf{y}_1' & \mathbf{x}_1' & \boldsymbol{\varepsilon}_1' \\ \mathbf{y}_2' & \mathbf{x}_2' & \boldsymbol{\varepsilon}_2' \\ & \vdots & \\ \mathbf{y}_T' & \mathbf{x}_T' & \boldsymbol{\varepsilon}_T' \end{bmatrix}.$$

In terms of the full set of T observations, the structure is

$$\mathbf{Y}\boldsymbol{\Gamma} + \mathbf{X}\mathbf{B} = \mathbf{E},$$

with

$$E[\mathbf{E}|\mathbf{X}] = \mathbf{0} \quad \text{and} \quad E[(1/T)\mathbf{E}'\mathbf{E}|\mathbf{X}] = \boldsymbol{\Sigma}.$$

Under general conditions, we can strengthen this to $\text{plim}[(1/T)\mathbf{E}'\mathbf{E}] = \boldsymbol{\Sigma}$. For convenience in what follows, we will denote a statistic consistently estimating a quantity, such as this one, with

$$(1/T)\mathbf{E}'\mathbf{E} \rightarrow \boldsymbol{\Sigma}.$$

An important assumption is

$$(1/T)\mathbf{X}'\mathbf{X} \rightarrow \mathbf{Q}, \text{ a finite positive definite matrix.} \tag{10-38}$$

We also assume that

$$(1/T)\mathbf{X}'\mathbf{E} \rightarrow \mathbf{0}. \tag{10-39}$$

This assumption is what distinguishes the predetermined variables from the endogenous variables. The reduced form is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Pi} + \mathbf{V}, \quad \text{where } \mathbf{V} = \mathbf{E}\boldsymbol{\Gamma}^{-1}.$$

Combining the earlier results, we have

$$\frac{1}{T}\begin{bmatrix} \mathbf{Y}' \\ \mathbf{X}' \\ \mathbf{V}' \end{bmatrix}[\mathbf{Y} \quad \mathbf{X} \quad \mathbf{V}] \rightarrow \begin{bmatrix} \boldsymbol{\Pi}'\mathbf{Q}\boldsymbol{\Pi} + \boldsymbol{\Omega} & \boldsymbol{\Pi}'\mathbf{Q} & \boldsymbol{\Omega} \\ \mathbf{Q}\boldsymbol{\Pi} & \mathbf{Q} & \mathbf{0}' \\ \boldsymbol{\Omega} & \mathbf{0} & \boldsymbol{\Omega} \end{bmatrix}. \tag{10-40}$$

10.4.3 THE IDENTIFICATION PROBLEM

Solving the **identification** problem precedes estimation. We have in hand a certain amount of information to use for inference about the underlying structure consisting of the sample data and theoretical restrictions on the model such as what variables do and do not appear in each of the equations. The issue is whether the information is sufficient to produce estimates of the parameters of the specified model. The case of measurement error that we examined in Section 8.5 is about identification. The sample regression coefficient, b, converges to a function of two underlying parameters, β and σ_u^2; $b = \mathbf{x}'\mathbf{y}/\mathbf{x}'\mathbf{x} \rightarrow \beta/[1 + \sigma_u^2/Q]$, where $(\mathbf{x}'\mathbf{x}/T) \rightarrow Q$. With no further information about σ_u^2, we cannot infer a unique β from the sample information, b and Q—there are different pairs of β and σ_u^2 that are consistent with the same information (b,Q). If there were some nonsample information available, such as $Q = \sigma_u^2$, then there would be a unique solution for β, in particular, $b \rightarrow \beta/2$.

Identification is a theoretical exercise. It arises in all econometric settings in which the parameters of a model are to be deduced from the combination of sample information and nonsample (theoretical) information. The crucial issue is whether it is possible to deduce the values of structural parameters uniquely from the sample information and **nonsample information** provided by theory, mainly restrictions on parameter values. The issue of identification is the subject of a lengthy literature including Working (1927), Bekker and Wansbeek (2001), and continuing through the contemporary discussion of natural experiments [Section 8.8 and Angrist and Pischke (2010), with commentary], instrumental variable estimation in general, and "identification strategies."

The structural model consists of the equation system

$$\mathbf{y}'\boldsymbol{\Gamma} + \mathbf{x}'\mathbf{B} = \boldsymbol{\varepsilon}'.$$

Each column in $\boldsymbol{\Gamma}$ and $\mathbf{B}$ are the parameters of a specific equation in the system. The information consists of the sample information, $(\mathbf{Y}, \mathbf{X})$, and other nonsample information in the form of restrictions on parameter matrices. The sample data provide sample moments, $\mathbf{X}'\mathbf{X}/T$, $\mathbf{X}'\mathbf{Y}/T$, and $\mathbf{Y}'\mathbf{Y}/T$. For purposes of identification, suppose we could observe as large a sample as desired. Then, based on our sample information, we could observe [from (10-40)]

$$\begin{aligned}
&(1/T)\mathbf{X}'\mathbf{X} \rightarrow \mathbf{Q},\\
&(1/T)\mathbf{X}'\mathbf{Y} = (1/T)\mathbf{X}'(\mathbf{X}\boldsymbol{\Pi} + \mathbf{V}) \rightarrow \mathbf{Q}\boldsymbol{\Pi},\\
&(1/T)\mathbf{Y}'\mathbf{Y} = (1/T)(\mathbf{X}\boldsymbol{\Pi} + \mathbf{V})'(\mathbf{X}\boldsymbol{\Pi} + \mathbf{V}) \rightarrow \boldsymbol{\Pi}'\mathbf{Q}\boldsymbol{\Pi} + \boldsymbol{\Omega}.
\end{aligned}$$

Therefore, $\boldsymbol{\Pi}$, the matrix of reduced-form coefficients, is observable

$$[(1/T)\mathbf{X}'\mathbf{X}]^{-1}[(1/T)\mathbf{X}'\mathbf{Y}] \rightarrow \boldsymbol{\Pi}.$$

This estimator is simply the equation-by-equation least squares regression of $\mathbf{Y}$ on $\mathbf{X}$. Because $\boldsymbol{\Pi}$ is observable, $\boldsymbol{\Omega}$ is also,

$$[(1/T)\mathbf{Y}'\mathbf{Y}] - [(1/T)\mathbf{Y}'\mathbf{X}][(1/T)\mathbf{X}'\mathbf{X}]^{-1}[(1/T)\mathbf{X}'\mathbf{Y}] \rightarrow \boldsymbol{\Omega}.$$

This result is the matrix of least squares residual variances and covariances. Therefore,

$\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$ can be estimated consistently by least squares regression of $\mathbf{Y}$ on $\mathbf{X}$.

The information in hand, therefore, consists of $\boldsymbol{\Pi}$, $\boldsymbol{\Omega}$, and whatever other nonsample information we have about the structure. The question is whether we can deduce $(\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\Sigma})$ from $(\boldsymbol{\Pi}, \boldsymbol{\Omega})$. A simple counting exercise immediately reveals that the answer is no—there are M^2 parameters in $\boldsymbol{\Gamma}$, $M(M + 1)/2$ in $\boldsymbol{\Sigma}$ and KM in $\mathbf{B}$, to be deduced. The sample data contain KM elements in $\boldsymbol{\Pi}$ and $M(M + 1)/2$ elements in $\boldsymbol{\Omega}$. By simply counting equations and unknowns, we find that our data are insufficient by M^2 pieces of information. We have (in principle) used the sample information already, so these M^2 additional restrictions are going to be provided by the theory of the model. The M^2 additional **restrictions** come in the form of normalizations—one coefficient in each equation equals one—most commonly **exclusion restrictions**, which set coefficients to zero and other relationships among the parameters, such as linear relationships, or specific values attached to coefficients. In some instances, restrictions on $\boldsymbol{\Sigma}$, such as assuming that certain disturbances are uncorrelated, will provide additional information. A small example will help fix ideas.

Example 10.6 Identification of a Supply and Demand Model

Consider a market in which q is quantity of Q, p is price, and z is the price of Z, a related good. We assume that z enters both the supply and demand equations. For example, Z might be a crop that is purchased by consumers and that will be grown by farmers instead of Q if its price rises enough relative to p. Thus, we would expect $\alpha_2 > 0$ and $\beta_2 < 0$. So,

$$\begin{aligned} q_d &= \alpha_0 + \alpha_1 p + \alpha_2 z + \varepsilon_d \qquad (\textit{demand}),\\ q_s &= \beta_0 + \beta_1 p + \beta_2 z + \varepsilon_s \qquad (\textit{supply}),\\ q_d &= q_s = q \qquad\qquad\qquad (\textit{equilibrium}). \end{aligned}$$

The reduced form is

$$q = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1\beta_2 - \alpha_2\beta_1}{\alpha_1 - \beta_1} z + \frac{\alpha_1\varepsilon_s - \beta_1\varepsilon_d}{\alpha_1 - \beta_1} = \pi_{11} + \pi_{21} z + \nu_q,$$

$$p = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2 - \alpha_2}{\alpha_1 - \beta_1} z + \frac{\varepsilon_s - \varepsilon_d}{\alpha_1 - \beta_1} = \pi_{12} + \pi_{22} z + \nu_p.$$

With only four reduced-form coefficients and six structural parameters, that there will not be a complete solution for all six structural parameters in terms of the four reduced form parameters. This model is unidentified. There is insufficient information in the sample and the theory to deduce the structural parameters.

Suppose, though, that it is known that $\beta_2 = 0$ (farmers do not substitute the alternative crop for this one). Then the solution for β_1 is π_{21}/π_{22}. After a bit of manipulation, we also obtain $\beta_0 = \pi_{11} - \pi_{12}\pi_{21}/\pi_{22}$. The exclusion restriction identifies the supply parameters; $\beta_2 = 0$ excludes z from the supply equation. But this step is as far as we can go. With this restriction, the model becomes partially identified. Some, but not all, of the parameters can be estimated.

Now, suppose that income x, rather than z, appears in the demand equation. The revised model is

$$\begin{aligned} q &= \alpha_0 + \alpha_1 p + \alpha_2 x + \varepsilon_1,\\ q &= \beta_0 + \beta_1 p + \beta_2 z + \varepsilon_2. \end{aligned}$$

Note that one variable is now excluded from each equation. The structure is now

$$[q \quad p]\begin{bmatrix} 1 & 1 \\ -\alpha_1 & -\beta_1 \end{bmatrix} + [1 \; x \; z]\begin{bmatrix} -\alpha_0 & -\beta_0 \\ -\alpha_2 & 0 \\ 0 & -\beta_2 \end{bmatrix} = [\varepsilon_1 \quad \varepsilon_2].$$

The reduced form is

$$[q \quad p] = [1 \; x \; z]\begin{bmatrix} (\alpha_1\beta_0 - \alpha_0\beta_1)/\Lambda & (\beta_0 - \alpha_0)/\Lambda \\ -\alpha_2\beta_1/\Lambda & -\alpha_2/\Lambda \\ \alpha_1\beta_2/\Lambda & \beta_2/\Lambda \end{bmatrix} + [\nu_1 \quad \nu_2],$$

where $\Lambda = (\alpha_1 - \beta_1)$. The unique solutions for the structural parameters in terms of the reduced-form parameters are now

$$\begin{aligned} \alpha_0 &= \pi_{11} - \pi_{12}\left(\frac{\pi_{31}}{\pi_{32}}\right), & \beta_0 &= \pi_{11} - \pi_{12}\left(\frac{\pi_{21}}{\pi_{22}}\right),\\ \alpha_1 &= \frac{\pi_{31}}{\pi_{32}}, & \beta_1 &= \frac{\pi_{21}}{\pi_{22}},\\ \alpha_2 &= \pi_{22}\left(\frac{\pi_{21}}{\pi_{22}} - \frac{\pi_{31}}{\pi_{32}}\right), & \beta_2 &= \pi_{32}\left(\frac{\pi_{31}}{\pi_{32}} - \frac{\pi_{21}}{\pi_{22}}\right). \end{aligned}$$

With this formulation, all of the parameters are identified. This is an example of an exactly identified model. An additional variation is worth a look. Suppose that a second variable, *w* (weather), appears in the supply equation,

$$q = \alpha_0 + \alpha_1 p + \alpha_2 x + \varepsilon_1,$$
$$q = \beta_0 + \beta_1 p + \beta_2 z + \beta_3 w + \varepsilon_2.$$

You can easily verify that, the reduced form matrix is the same as the previous one, save for an additional row that contains $[\alpha_1\beta_3/\Lambda, \beta_3/\Lambda]$. This implies that there is now a second solution for α_1, π_{41}/π_{42}. The two solutions, this and π_{31}/π_{32}, will be different. This model is overidentified. There is more information in the sample and theory than is needed to deduce the structural parameters.

Some equation systems are identified and others are not. The formal mathematical conditions under which an equation in a system is identified turns on two results known as the rank and order conditions. The *order condition* is a simple counting rule. It requires that the number of exogenous variables that appear elsewhere in the equation system must be at least as large as the number of endogenous variables in the equation. (Other specific restrictions on the parameters will be included in this count—note that an "exclusion restriction" is a type of linear restriction.) We used this rule when we constructed the IV estimator in Chapter 8. In that setting, we required our model to be at least *identified* by requiring that the number of instrumental variables not contained in $\mathbf{X}$ be at least as large as the number of endogenous variables. The correspondence of that single equation application with the condition defined here is that the rest of the equation system is the source of the instrumental variables. One simple order condition for identification of an equation system is that each equation contain "its own" exogenous variable that does not appear elsewhere in the system.

The **order condition** is necessary for identification; the **rank condition** is sufficient. The equation system in (10-37) in structural form is $\mathbf{y}'\boldsymbol{\Gamma} = -\mathbf{x}'\mathbf{B} + \boldsymbol{\varepsilon}'$. The reduced form is $\mathbf{y}' = \mathbf{x}'(-\mathbf{B}\,\boldsymbol{\Gamma}^{-1}) + \boldsymbol{\varepsilon}'\boldsymbol{\Gamma}^{-1} = \mathbf{x}'\boldsymbol{\Pi} + \mathbf{v}'$. The way we are going to deduce the parameters in $(\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\Sigma})$ is from the reduced form parameters $(\boldsymbol{\Pi}, \boldsymbol{\Omega})$. For the jth equation, the solution is contained in $\boldsymbol{\Pi}\boldsymbol{\Gamma}_j = -\mathbf{B}_j$, where $\boldsymbol{\Gamma}_j$ contains all the coefficients in the jth equation that multiply endogenous variables. One of these coefficients will equal one, usually some will equal zero, and the remainder are the nonzero coefficients on endogenous variables in the equation, $\mathbf{Y}_j$ [these are denoted γ_j in (10-41) following]. Likewise, $\mathbf{B}_j$ contains the coefficients in equation j on all exogenous variables in the model—some of these will be zero and the remainder will multiply variables in $\mathbf{X}_j$, the exogenous variables that appear in this equation [these are denoted $\boldsymbol{\beta}_j$ in (10-41) following]. The empirical counterpart will be $\mathbf{P}\mathbf{c}_j = \mathbf{b}_j$, where $\mathbf{P}$ is the estimated reduced form, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, and $\mathbf{c}_j$ and $\mathbf{b}_j$ will be the estimates of the jth columns of $\boldsymbol{\Gamma}$ and $\mathbf{B}$. The rank condition ensures that there is a solution to this set of equations. In practical terms, the rank condition is difficult to establish in large equation systems. Practitioners typically take it as a given. In small systems, such as the two-equation systems that dominate contemporary research, it is trivial, as we examine in the next example. We have already used the rank condition in Chapter 8, where it played a role in the relevance condition for instrumental variable estimation. In particular, note after the statement of the assumptions for instrumental variable estimation, we assumed $\text{plim}(1/T)\mathbf{Z}'\mathbf{X}$ is a matrix with rank K. (This condition is often labeled the *rank condition* in contemporary applications. It not identical, but it is sufficient for the condition mentioned here.)

Example 10.7 The Rank Condition and a Two-Equation Model

The following two-equation recursive model provides what is arguably the platform for much of contemporary econometric analysis. The main equation of interest is

$$y = \gamma f + \beta x + \varepsilon.$$

The variable f is endogenous (it is correlated with ε.); x is exogenous (it is uncorrelated with ε). The analyst has in hand an instrument for f, z. The instrument, z, is relevant, in that in the auxiliary equation,

$$f = \lambda x + \delta z + w,$$

δ is not zero. The exogeneity assumption is $E[\varepsilon z] = E[wz] = 0$. Note that the source of the endogeneity of f is the assumed correlation of w and ε. For purposes of the exercise, assume that $E[xz] = 0$ and the data satisfy $\mathbf{x}'\mathbf{z} = 0$—this actually loses no generality. In this two-equation model, the second equation is already in reduced form; x and z are both exogenous. It follows that λ and δ are estimable by least squares. The estimating equations for (γ, β) are

$$\mathbf{P}\boldsymbol{\gamma}_1 = \begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{z} \\ \mathbf{z}'\mathbf{x} & \mathbf{z}'\mathbf{z} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x}'\mathbf{y} & \mathbf{x}'\mathbf{f} \\ \mathbf{z}'\mathbf{y} & \mathbf{z}'\mathbf{f} \end{bmatrix} \begin{pmatrix} 1 \\ -\gamma \end{pmatrix} = \begin{bmatrix} \mathbf{x}'\mathbf{y}/\mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{f}/\mathbf{x}'\mathbf{x} \\ \mathbf{z}'\mathbf{y}/\mathbf{z}'\mathbf{z} & \mathbf{z}'\mathbf{f}/\mathbf{z}'\mathbf{z} \end{bmatrix} \begin{pmatrix} 1 \\ -\gamma \end{pmatrix} = \boldsymbol{\beta}_{\mathrm{j}} = \begin{pmatrix} \beta \\ 0 \end{pmatrix}.$$

The solutions are $\gamma = (\mathbf{z}'\mathbf{y}/\mathbf{z}'\mathbf{f})$ and $\beta = (\mathbf{x}'\mathbf{y}/\mathbf{x}'\mathbf{x} - (\mathbf{z}'\mathbf{y}/\mathbf{z}'\mathbf{f})\mathbf{x}'\mathbf{f}/\mathbf{x}'\mathbf{x})$. Because $\mathbf{x}'\mathbf{x}$ cannot equal zero, the solution depends on $(\mathbf{z}'\mathbf{f}/\mathbf{z}'\mathbf{z})$ not equal to zero—formally that this part of the reduced form coefficient matrix have rank $M = 1$, which would be the rank condition. Note that the solution for γ is the instrumental variable estimator, with z as instrument for f. (The simplicity of this solution turns on the assumption that $\mathbf{x}'\mathbf{z} = 0$. The algebra gets a bit more complicated without it, but the conclusion is the same.)

The rank condition is based on the exclusion restrictions in the model—whether the exclusion restrictions provide enough information to identify the coefficients in the jth equation. Formally, the idea can be developed thusly. With the jth equation written as in (10-41), we call $\mathbf{X}_j$ the *included exogenous variables*. The remaining excluded exogenous variables are denoted $\mathbf{X}_j^*$. The M_j variables $\mathbf{Y}_j$ in (10-41) are the included endogenous variables. With this distinction, we can write the M_j reduced forms for $\mathbf{Y}_j$ as $\overline{\boldsymbol{\Pi}}_j = \begin{bmatrix} \boldsymbol{\Pi}_j \\ \boldsymbol{\Pi}_j^* \end{bmatrix}$. The rank condition (which we state without proof) is that the rank of the lower part of the $M_j \times (K_j + K_j^*)$ matrix, $\overline{\boldsymbol{\Pi}}_j$, equal M_j. In the preceding example, in the first equation, $\mathbf{Y}_j$ is f, $M_j = 1$, $\mathbf{X}_j$ is x, $\mathbf{X}_j^*$ is z, and $\overline{\boldsymbol{\Pi}}_j$ is estimated by the regression of f on x and z; $\boldsymbol{\Pi}_j$ is the coefficient on x and $\boldsymbol{\Pi}_j^*$ is the coefficient on z. The rank condition we noted earlier is that what is estimated by $\mathbf{z}'\mathbf{f}/\mathbf{z}'\mathbf{z}$, which would correspond to $\boldsymbol{\Pi}_j^*$ not equal zero, meaning that it has rank 1.

Casual statements of the rank condition based on an IV regression of a variable $\mathbf{y}_{IV}$ on $(M_j + K_j)$ endogenous and exogeneous variables in $\mathbf{X}_{IV}$, using $K_j + K_j^*$ exogenous and instrumental variables in $\mathbf{Z}_{IV}$ (in the most familiar cases, $M_j = K_j^* = 1$), state that the rank requirement is that $(\mathbf{Z}_{IV}'\mathbf{X}_{IV}/T)$ be nonsingular. In the notation we are using here, $\mathbf{Z}_{IV}$ would be $\mathbf{X} = (\mathbf{X}_j, \mathbf{X}_j^*)$ and $\mathbf{X}_{iv}$ would be $(\mathbf{X}_j, \mathbf{Y}_j)$. This nonsingularity would correspond to full rank of $\text{plim}(\mathbf{X}'\mathbf{X}/T)$ times $\text{plim}[(\mathbf{X}'\mathbf{X}^*/T, \mathbf{X}'\mathbf{Y}_j/T)]$ because $\text{plim}(\mathbf{X}'\mathbf{X}/T) = \mathbf{Q}$, which is nonsingular [see (10-40)]. The first K_j columns of this matrix are the last K_j columns of an identity matrix, which have rank K_j. The last M_j columns are estimates of $\mathbf{Q}\overline{\boldsymbol{\Pi}}_j$, which we require to have rank M_j, so the requirement is that $\overline{\boldsymbol{\Pi}}_j$ have rank M_j. But, if $K_j^* \geq M_j$ (the order condition), then all that is needed is $\text{rank}(\boldsymbol{\Pi}_j^*) = M_j$, so, in practical terms, the casual statement is correct. It is stronger than necessary; the formal mathematical condition is only that the lower half of the matrix must have rank M_j, but the practical result is much easier to visualize.

It is also easy to verify that the rank condition requires that the predictions of $\mathbf{Y}_j$ using $(\mathbf{X}_j, \mathbf{X}_j^*)\overline{\boldsymbol{\Pi}}_j$ be linearly independent. Continuing this line of thinking, if we use 2SLS, the rank condition requires that the predicted values of the included endogenous variables not be collinear, which makes sense.

10.4.4 SINGLE EQUATION ESTIMATION AND INFERENCE

For purposes of estimation and inference, we write the model in the way that the researcher would typically formulate it,

$$\begin{aligned}\mathbf{y}_j &= \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Y}_j\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j \\ &= \mathbf{Z}_j\boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j,\end{aligned} \tag{10-41}$$

where $\mathbf{y}_j$ is the "dependent variable" in the equation, $\mathbf{X}_j$ is the set of exogenous variables that appear in the jth equation—note that this is not all the variables in the model—and $\mathbf{Z}_j = (\mathbf{X}_j, \mathbf{Y}_j)$. The full set of exogenous variables in the model, including $\mathbf{X}_j$ and variables that appear elsewhere in the model (including a constant term if any equation includes one), is denoted $\mathbf{X}$. For example, in the supply/demand model in Example 10.6, the full set of exogenous variables is $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \mathbf{z})$, while $\mathbf{X}_{Demand} = (\mathbf{1}, \mathbf{x})$ and $\mathbf{X}_{Supply} = (\mathbf{1}, \mathbf{z})$. Finally, $\mathbf{Y}_j$ is the endogenous variables that appear on the right-hand side of the jth equation. Once again, this is likely to be a subset of the endogenous variables in the full model. In Example 10.6, $\mathbf{Y}_j = (\mathbf{price})$ in both cases.

There are two approaches to estimation and inference for simultaneous equations models. **Limited information estimators** are constructed for each equation individually. The approach is analogous to estimation of the seemingly unrelated regressions model in Section 10.2 by least squares, one equation at a time. **Full information estimators** are used to estimate all equations simultaneously. The counterpart for the seemingly unrelated regressions model is the feasible generalized least squares estimator discussed in Section 10.2.3. The major difference to be accommodated at this point is the endogeneity of $\mathbf{Y}_j$ in (10-41).

The equation in (10-41) is precisely the model developed in Chapter 8. Least squares will generally be unsuitable as it is inconsistent due to the correlation between $\mathbf{Y}_j$ and $\boldsymbol{\varepsilon}_j$. The usual approach will be two-stage least squares as developed in Sections 8.3.2 through 8.3.4. The only difference between the case considered here and that in Chapter 8 is the source of the instrumental variables. In our general model in Chapter 8, the source of the instruments remained somewhat ambiguous; the overall rule was "outside the model." In this setting, the instruments come from elsewhere in the model—that is, "not in the jth equation." For estimating the linear simultaneous equations model, the most common estimator is

$$\begin{aligned}\hat{\boldsymbol{\delta}}_{j,\,2\,\mathrm{SLS}} &= [\hat{\mathbf{Z}}_j'\hat{\mathbf{Z}}_j]^{-1}\hat{\mathbf{Z}}_j'\mathbf{y}_j \\ &= [(\mathbf{Z}_j'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}_j)]^{-1}(\mathbf{Z}_j'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_j,\end{aligned} \tag{10-42}$$

where all columns of $\hat{\mathbf{Z}}_j'$ are obtained as predictions in a regression of the corresponding column of $\mathbf{Z}_j$ on $\mathbf{X}$. This equation also results in a useful simplification of the estimated asymptotic covariance matrix,

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\delta}}_{j,\,2\,SLS}] = \hat{\sigma}_{jj}(\hat{\mathbf{Z}}_j'\hat{\mathbf{Z}}_j)^{-1}.$$

It is important to note that σ_{jj} is estimated by

$$\hat{\sigma}_{jj} = \frac{(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_j)'(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_j)}{T}, \tag{10-43}$$

using the original data, not $\hat{\mathbf{Z}}_j$.

Note the role of the order condition for identification in the two-stage least squares estimator. Formally, the order condition requires that the number of exogenous variables that appear elsewhere in the model (not in this equation) be at least as large as the number of endogenous variables that appear in this equation. The implication will be that we are going to predict $\mathbf{Z}_j = (\mathbf{X}_j, \mathbf{Y}_j)$ using $\mathbf{X} = (\mathbf{X}_j, \mathbf{X}_j^*)$. In order for these predictions to be linearly independent, there must be at least as many variables used to compute the predictions as there are variables being predicted. Comparing $(\mathbf{X}_j, \mathbf{Y}_j)$ to $(\mathbf{X}_j, \mathbf{X}_j^*)$, we see that there must be at least as many variables in $\mathbf{X}_j^*$ as there are in $\mathbf{Y}_j$, which is the order condition. The practical rule of thumb that every equation have at least one variable in it that does not appear in any other equation will guarantee this outcome.

Two-stage least squares is used nearly universally in estimation of linear simultaneous equation models—for precisely the reasons outlined in Chapter 8. However, some applications (and some theoretical treatments) have suggested that the **limited information maximum likelihood (LIML) estimator** based on the normal distribution may have better properties. The technique has also found recent use in the analysis of weak instruments. A result that emerges from the derivation is that the LIML estimator has the same asymptotic distribution as the 2SLS estimator, and the latter does not rely on an assumption of normality. This raises the question why one would use the LIML technique given the availability of the more robust (and computationally simpler) alternative. Small sample results are sparse, but they would favor 2SLS as well.[25] One significant virtue of LIML is its invariance to the normalization of the equation. Consider an example in a system of equations,

$$y_1 = y_2\gamma_2 + y_3\gamma_3 + x_1\beta_1 + x_2\beta_2 + \varepsilon_1.$$

An equivalent equation would be

$$\begin{aligned} y_2 &= y_1(1/\gamma_2) + y_3(-\gamma_3/\gamma_2) + x_1(-\beta_1/\gamma_2) + x_2(-\beta_2/\gamma_2) + \varepsilon_1(-1/\gamma_2) \\ &= y_1\widetilde{\gamma}_1 + y_3\widetilde{\gamma}_3 + x_1\widetilde{\beta}_1 + x_2\widetilde{\beta}_2 + \widetilde{\varepsilon}_1. \end{aligned}$$

The parameters of the second equation can be manipulated to produce those of the first. But, as you can easily verify, the 2SLS estimator is not invariant to the normalization of the equation—2SLS would produce numerically different answers. LIML would give the same numerical solutions to both estimation problems suggested earlier. A second virtue is LIML's better performance in the presence of weak instruments.

The LIML, or **least variance ratio** estimator, can be computed as follows.[26] Let

$$\mathbf{W}_j^0 = \mathbf{E}_j^{0\prime}\mathbf{E}_j^0, \tag{10-44}$$

where

$$\mathbf{Y}_j^0 = [\mathbf{y}_j, \mathbf{Y}_j],$$

and

$$\mathbf{E}_j^0 = \mathbf{M}_j\mathbf{Y}_j^0 = [\mathbf{I} - \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j']\mathbf{Y}_j^0. \tag{10-45}$$

[25] See Phillips (1983).

[26] The LIML estimator was derived by Anderson and Rubin (1949, 1950). [See, also, Johnston (1984).] The much simpler and equally efficient two-stage least squares estimator remains the estimator of choice.

Each column of $\mathbf{E}_j^0$ is a set of least squares residuals in the regression of the corresponding column of $\mathbf{Y}_j^0$ on $\mathbf{X}_j$, that is, only the exogenous variables that appear in the jth equation. Thus, $\mathbf{W}_j^0$ is the matrix of sums of squares and cross products of these residuals. Define

$$\mathbf{W}_j^1 = \mathbf{E}_j^{1\prime}\mathbf{E}_j^1 = \mathbf{Y}_j^{0\prime}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}_j^0. \tag{10-46}$$

That is, $\mathbf{W}_j^1$ is defined like $\mathbf{W}_j^0$ except that the regressions are on all the x's in the model, not just the ones in the jth equation. Let

$$\lambda_1 = \text{smallest characteristic root of } (\mathbf{W}_j^1)^{-1}\mathbf{W}_j^0. \tag{10-47}$$

This matrix is asymmetric, but all its roots are real and greater than or equal to 1. [Depending on the available software, it may be more convenient to obtain the identical smallest root of the symmetric matrix $\mathbf{D} = (\mathbf{W}_j^1)^{-1/2}\mathbf{W}_j^0(\mathbf{W}_j^1)^{-1/2}$.] Now partition $\mathbf{W}_j^0$ into $\mathbf{W}_j^0 = \begin{bmatrix} w_{jj}^0 & \mathbf{w}_j^{0\prime} \\ \mathbf{w}_j^0 & \mathbf{W}_{jj}^0 \end{bmatrix}$ corresponding to $[\mathbf{y}_j, \mathbf{Y}_j]$, and partition $\mathbf{W}_j^1$ likewise. Then, with these parts in hand,

$$\hat{\boldsymbol{\gamma}}_{j,\,\text{LIML}} = [\mathbf{W}_{jj}^0 - \lambda_1\mathbf{W}_{jj}^1]^{-1}(\mathbf{w}_j^0 - \lambda_1\mathbf{w}_j^1) \tag{10-48}$$

and

$$\hat{\boldsymbol{\beta}}_{j,\,\text{LIML}} = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'(\mathbf{y}_j - \mathbf{Y}_j\hat{\boldsymbol{\gamma}}_{j,\,\text{LIML}}).$$

Note that $\boldsymbol{\beta}_j$ is estimated by a simple least squares regression. [See (3-18).] The asymptotic covariance matrix for the LIML estimator is identical to that for the 2SLS estimator.

Example 10.8 Simultaneity in Health Production

Example 7.1 analyzed the incomes of a subsample of Riphahn, Wambach, and Million's (2003) data on health outcomes in the German Socioeconomic Panel. Here we continue Example 10.4 and consider a Grossman (1972) style model for health and incomes. Our two-equation model is

$$\begin{aligned} \textit{Health Satisfaction} &= \alpha_1 + \gamma_1\,\textit{ln Income} + \alpha_2\,\textit{Female} + \alpha_3\,\textit{Working} + \alpha_4\,\textit{Public} + \alpha_5\,\textit{Add On} \\ &\quad + \alpha_6\,\textit{Age} + \varepsilon_H, \\ \textit{ln Income} &= \beta_1 + \gamma_2\,\textit{Health Satisfaction} + \beta_2\,\textit{Female} + \beta_3\,\textit{Education} + \beta_4\,\textit{Married} \\ &\quad + \beta_5\textit{HHKids} + \beta_6\textit{Age} + \varepsilon_I. \end{aligned}$$

For purposes of this application, we avoid panel data considerations by examining only the 1994 wave (cross section) of the data, which contains 3,377 observations. The health outcome variable is *Self Assessed Health Satisfaction* (HSAT). Whether this variable actually corresponds to a commonly defined objective measure of health outcomes is debateable. We will treat it as such. Second, the variable is a scale variable, coded in this data set 0 to 10. [In more recent versions of the GSOEP data, and in the British (BHPS) and Australian (HILDA) counterparts, it is coded 0 to 4.] We would ordinarily treat such a variable as a discrete ordered outcome, as we do in Examples 18.14 and 18.15. We will treat it as if it were continuous in this example, and recognize that there is likely to be some distortion in the measured effects that we are interested in. *Female*, *Working*, *Married*, and *HHkids* are dummy variables, the last indicating whether there are children living in the household. *Education* and *Age* are in years. *Public* and *AddOn* are dummy variables that indicate whether the individual takes up the public health insurance and, if so, whether he or she also takes up the additional

AddOn insurance, which covers some additional costs. Table 10.7 presents OLS and 2SLS estimates of the parameters of the two-equation model. The differences are striking. In the health outcome equation, the OLS coefficient on *ln Income* is quite large (0.42) and highly significant ($t = 5.17$). However, the effect almost doubles in the 2SLS results. The strong negative effect of having the public health insurance might make one wonder if the insurance takeup is endogenous in the same fashion as *ln Income*. (In the original study from which these data were borrowed, the authors were interested in whether takeup of the add on insurance had an impact on usage of the health care system (number of doctor visits). The 2SLS estimates of the *ln Income* equation are also distinctive. Now, the extremely small effect of health estimated by OLS (0.020) becomes the dominant effect, with marital status, in the 2SLS results.

Both equations are overidentified—each has three excluded exogenous variables. Regression of the 2SLS residuals from the HSAT equation on all seven exogenous variables (and the constant) gives an R^2 of 0.0005916, so the chi-squared test of the overidentifying restrictions is 3,337(.0005916) = 1.998. With two degrees of freedom, the critical value is 5.99, so the restrictions would not be rejected. For the *ln Income* equation, the R^2 in the regression of the residuals on all of the exogenous variables is 0.000426, so the test statistic is 1.438, which is not significant. On this basis, we conclude that the specification of the model is adequate.

TABLE 10.7 Estimated Health Production Model (absolute *t* ratios in parentheses)

	Health Equation			***ln Income Equation***		
	OLS	***2SLS***	***LIML***	***OLS***	***2SLS***	***LIML***
Constant	8.903	9.201	9.202	−1.817	−5.379	−5.506
	(40.67)	(30.31)	(30.28)	(30.81)	(8.65)	(8.46)
ln *Income*	0.418	0.710	0.712			
	(5.17)	(3.20)	(3.20)			
Health				0.020	0.497	0.514
				(5.83)	(6.12)	(6.04)
Female	−0.211	−0.218	−0.218	−0.011	0.126	0.131
	(2.76)	(2.85)	(2.85)	(0.70)	(2.78)	(2.79)
Working	0.339	0.259	0.259			
	(3.76)	(2.43)	(2.43)			
Public	−0.472	−0.391	−0.391			
	(4.10)	(3.05)	(3.04)			
Add On	0.204	0.140	0.139			
	(0.80)	(0.54)	(0.54)			
Education				0.055	0.017	0.016
				(17.00)	(1.65)	(1.46)
Married				0.352	0.263	0.260
				(18.11)	(5.08)	(4.86)
Age	−0.038	−0.039	−0.039	−0.002	0.017	0.018
	(11.55)	(11.60)	(11.60)	(2.58)	(4.53)	(4.51)
HHKids				−0.062	−0.061	−0.061
				(3.45)	(1.32)	(1.28)

10.4.5 SYSTEM METHODS OF ESTIMATION

We may formulate the full system of equations as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_M \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_M \end{bmatrix} \tag{10-49}$$

or

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where

$$E[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}, \quad \text{and} \quad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} \otimes \mathbf{I}. \tag{10-50}$$

[See (10-3).] The least squares estimator,

$$\mathbf{d} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y},$$

is equation-by-equation ordinary least squares and is inconsistent. But even if ordinary least squares were consistent, we know from our results for the seemingly unrelated regressions model that it would be inefficient compared with an estimator that makes use of the cross-equation correlations of the disturbances. For the first issue, we turn once again to an IV estimator. For the second, as we did Section 10.2.1, we use a generalized least squares approach. Thus, assuming that the matrix of instrumental variables, $\overline{\mathbf{W}}$, satisfies the requirements for an IV estimator, a consistent though inefficient estimator would be

$$\hat{\boldsymbol{\delta}}_{\text{IV}} = (\overline{\mathbf{W}}'\mathbf{Z})^{-1}\overline{\mathbf{W}}'\mathbf{y}. \tag{10-51}$$

Analogous to the seemingly unrelated regressions model, a more efficient estimator would be based on the generalized least squares principle,

$$\hat{\boldsymbol{\delta}}_{\text{IV, GLS}} = [\overline{\mathbf{W}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{Z}]^{-1}\,\overline{\mathbf{W}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}, \tag{10-52}$$

or, where $\mathbf{W}_j$ is the set of instrumental variables for the jth equation,

$$\hat{\boldsymbol{\delta}}_{\text{IV, GLS}} = \begin{bmatrix} \sigma^{11}\mathbf{W}_1'\mathbf{Z}_1 & \sigma^{12}\mathbf{W}_1'\mathbf{Z}_2 & \cdots & \sigma^{1M}W_1'Z_M \\ \sigma^{21}\mathbf{W}_2'\mathbf{Z}_1 & \sigma^{22}\mathbf{W}_2'\mathbf{Z}_2 & \cdots & \sigma^{2M}W_2'Z_M \\ & & \vdots & \\ \sigma^{M1}\mathbf{W}_M'\mathbf{Z}_1 & \sigma^{M2}\mathbf{W}_M'\mathbf{Z}_2 & \cdots & \sigma^{MM}W_M'Z_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{n=1}^{M}\sigma^{1n}\mathbf{W}_1'\mathbf{y}_n \\ \sum_{n=1}^{M}\sigma^{2n}\mathbf{W}_2'\mathbf{y}_n \\ \vdots \\ \sum_{n=1}^{M}\sigma^{Mn}\mathbf{W}_M'\mathbf{y}_n \end{bmatrix}.$$

Three IV techniques are generally used for joint estimation of the entire system of equations: three-stage least squares, GMM, and **full information maximum likelihood (FIML)**. In the small minority of applications that use a system estimator, 3SLS is usually the estimator of choice. For dynamic models, GMM is sometimes preferred. The FIML estimator is generally of theoretical interest, as it brings no advantage over 3SLS, but is much more complicated to compute.

Consider the IV estimator formed from

$$\overline{\mathbf{W}} = \hat{\mathbf{Z}} = \text{diag}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_1, \ \ldots, \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_M] = \begin{bmatrix} \hat{\mathbf{Z}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Z}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \hat{\mathbf{Z}}_M \end{bmatrix}.$$

The IV estimator,

$$\hat{\boldsymbol{\delta}}_{\text{IV}} = [\hat{\mathbf{Z}}'\mathbf{Z}]^{-1}\hat{\mathbf{Z}}'\mathbf{y},$$

is simply equation-by-equation 2SLS. We have already established the consistency of 2SLS. By analogy to the seemingly unrelated regressions model of Section 10.2, however, we would expect this estimator to be less efficient than a GLS estimator. A natural candidate would be

$$\hat{\boldsymbol{\delta}}_{\text{3SLS}} = [\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{Z}]^{-1}\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}.$$

For this estimator to be a valid IV estimator, we must establish that

$$\text{plim}\ \frac{1}{T}\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\boldsymbol{\varepsilon} = \mathbf{0},$$

which is M sets of equations, each one of the form

$$\text{plim}\ \frac{1}{T}\sum_{j=1}^{M}\sigma^{ij}\hat{\mathbf{Z}}_i'\boldsymbol{\varepsilon}_j = \mathbf{0}.$$

Each is the sum of vectors, all of which converge to zero, as we saw in the development of the 2SLS estimator. The second requirement, that

$$\text{plim}\ \frac{1}{T}\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{Z} \neq \mathbf{0},$$

and that the matrix be nonsingular, can be established along the lines of its counterpart for 2SLS. Identification of every equation by the rank condition is sufficient.

Once again, using the idempotency of $\mathbf{I} - \mathbf{M}$, we may also interpret this estimator as a GLS estimator of the form

$$\hat{\boldsymbol{\delta}}_{\text{3SLS}} = [\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})]\mathbf{y}. \tag{10-53}$$

The appropriate asymptotic covariance matrix for the estimator is

$$\text{Asy.Var}[\hat{\boldsymbol{\delta}}_{\text{3SLS}}] = (\overline{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\overline{\mathbf{Z}})^{-1}, \tag{10-54}$$

where $\overline{\mathbf{Z}} = \text{diag}[\mathbf{X}\boldsymbol{\Pi}_j, \mathbf{X}_j]$. This matrix would be estimated with the bracketed inverse matrix in (10-53).

Using sample data, we find that $\overline{\mathbf{Z}}$ may be estimated with $\hat{\mathbf{Z}}$. The remaining difficulty is to obtain an estimate of $\boldsymbol{\Sigma}$. In estimation of the seemingly unrelated regressions model, for efficient estimation, any consistent estimator of $\boldsymbol{\Sigma}$ will do. The designers of the 3SLS method, Zellner and Theil (1962), suggest the natural choice arising out of the two-stage least estimates. The **three-stage least squares (3SLS) estimator** is thus defined as follows:

1. Estimate $\boldsymbol{\Pi}$ by ordinary least squares and compute $\hat{\mathbf{Y}}_m$ for each equation.
2. Compute $\hat{\boldsymbol{\delta}}_{m,\,\text{2SLS}}$ for each equation; then

$$\hat{\sigma}_{mn} = \frac{(\mathbf{y}_m - \mathbf{Z}_m\hat{\boldsymbol{\delta}}_m)(\mathbf{y}_n - \mathbf{Z}_n\hat{\boldsymbol{\delta}}_n)}{T}. \tag{10-55}$$

3. Compute the GLS estimator according to (10-53) and an estimate of the asymptotic covariance matrix according to (10-54) using $\hat{\mathbf{Z}}$ and $\hat{\boldsymbol{\Sigma}}$.

By showing that the 3SLS estimator satisfies the requirements for an IV estimator, we have established its consistency. The question of asymptotic efficiency remains. It can be shown that of all IV estimators that use only the sample information embodied in the system, 3SLS is asymptotically efficient.

Example 10.9 Klein's Model I

A widely used example of a simultaneous equations model of the economy is Klein's (1950) Model I. The model may be written

$$\begin{aligned}
C_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3(W_t^p + W_t^g) + \varepsilon_{1t} && \text{(consumption)},\\
I_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} && \text{(investment)},\\
W_t^p &= \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t} && \text{(private wages)},\\
X_t &= C_t + I_t + G_t && \text{(equilibrium demand)},\\
P_t &= X_t - T_t - W_t^p && \text{(private profits)},\\
K_t &= K_{t-1} + I_t && \text{(capital stock)}.
\end{aligned}$$

The endogenous variables are each on the left-hand side of an equation and are labeled on the right. The exogenous variables are G_t = government nonwage spending, T_t = indirect business taxes plus net exports, W_t^g = government wage bill, A_t = time trend measured as years from 1931, and the constant term. There are also three predetermined variables: the lagged values of the capital stock, private profits, and total demand. The model contains three **behavioral equations**, an **equilibrium condition**, and two accounting identities. This model provides an excellent example of a small, dynamic model of the economy. It has also been widely used as a test ground for simultaneous equations estimators. Klein estimated the parameters using yearly aggregate data for the U.S. for 1921 to 1941. The data are listed in Appendix Table F10.3. Table 10.8 presents limited and full information estimates for Klein's Model I based on the original data.

It might seem, in light of the entire discussion, that one of the structural estimators described previously should always be preferred to ordinary least squares, which alone among the estimators considered here is inconsistent. Unfortunately, the issue is not so clear. First, it is often found that the OLS estimator is surprisingly close to the structural estimator. It can be shown that, at least in some cases, OLS has a smaller variance about its mean than does 2SLS about its mean, leading to the possibility that OLS might be more precise in a mean-squared-error sense. But this result must be tempered by the finding that the OLS standard errors are, in all likelihood, not useful for inference purposes. Obviously, this discussion is relevant only to finite samples. Asymptotically, 2SLS must dominate OLS, and in a correctly specified model, any full information estimator (3SLS) must dominate any limited information one (2SLS). The finite sample properties are of crucial importance. Most of what we know is asymptotic properties, but most applications are based on rather small or moderately sized samples.

Although the system methods of estimation are asymptotically better, they have two problems. First, any specification error in the structure of the model will be propagated throughout the system by 3SLS. The limited information estimators will, by and large,

TABLE 10.8 Estimates of Klein's Model I (Estimated asymptotic standard errors in parentheses)

	2SLS				**3SLS**			
C	16.6	0.017	0.216	0.810	16.4	0.125	0.163	0.790
	(1.32)	(0.118)	(0.107)	(0.040)	(1.30)	(0.108)	(0.100)	(0.038)
I	20.3	0.150	0.616	−0.158	28.2	−0.013	0.756	−0.195
	(7.54)	(0.173)	(0.162)	(0.036)	(6.79)	(0.162)	(0.153)	(0.033)
W^P	1.50	0.439	0.147	0.130	1.80	0.400	0.181	0.150
	(1.15)	(0.036)	(0.039)	(0.029)	(1.12)	(0.032)	(0.034)	(0.028)
	LIML				**OLS**			
C	17.1	−0.222	0.396	0.823	16.2	0.193	0.090	0.796
	(1.84)	(0.202)	(0.174)	(0.055)	(1.30)	(0.091)	(0.091)	(0.040)
I	22.6	0.075	0.680	−0.168	10.1	0.480	0.333	−0.112
	(9.24)	(0.219)	(0.203)	(0.044)	(5.47)	(0.097)	(0.101)	(0.027)
W^P	1.53	0.434	0.151	0.132	1.50	0.439	0.146	0.130
	(2.40)	(0.137)	(0.135)	(0.065)	(1.27)	(0.032)	(0.037)	(0.032)

confine a problem to the particular equation in which it appears. Second, in the same fashion as the SUR model, the finite-sample variation of the estimated covariance matrix is transmitted throughout the system. Thus, the finite-sample variance of 3SLS may well be as large as or larger than that of 2SLS.[27]

10.5 SUMMARY AND CONCLUSIONS

This chapter has surveyed the specification and estimation of multiple equations models. The SUR model is an application of the generalized regression model introduced in Chapter 9. The advantage of the SUR formulation is the rich variety of behavioral models that fit into this framework. We began with estimation and inference with the SUR model, treating it essentially as a generalized regression. The major difference between this set of results and the single-equation model in Chapter 9 is practical. While the SUR model is, in principle, a single equation GR model with an elaborate covariance structure, special problems arise when we explicitly recognize its intrinsic nature as a set of equations linked by their disturbances. The major result for estimation at this step is the feasible GLS estimator. In spite of its apparent complexity, we can estimate the SUR model by a straightforward two-step GLS approach that is similar to the one we used for models with heteroscedasticity in Chapter 9. We also extended the SUR model to autocorrelation and heteroscedasticity. Once again, the multiple equation nature of the model complicates these applications. Section 10.4 presented a common application of the seemingly unrelated regressions model, the estimation of demand systems. One of the signature features of this literature is the seamless transition from the theoretical models of optimization of consumers and producers to the sets of empirical demand equations derived from Roy's identity for consumers and Shephard's lemma for producers.

[27]See Cragg (1967) and the many related studies listed by Judge et al. (1985, pp. 646–653).

The multiple equations models surveyed in this chapter involve most of the issues that arise in analysis of linear equations in econometrics. Before one embarks on the process of estimation, it is necessary to establish that the sample data actually contain sufficient information to provide estimates of the parameters in question. This is the question of identification. Identification involves both the statistical properties of estimators and the role of theory in the specification of the model. Once identification is established, there are numerous methods of estimation. We considered three single-equation techniques, least squares, instrumental variables, and maximum likelihood. Fully efficient use of the sample data will require joint estimation of all the equations in the system. Once again, there are several techniques-these are extensions of the single-equation methods including three-stage least squares-and full information maximum likelihood. In both frameworks, this is one of those benign situations in which the computationally simplest estimator is generally the most efficient one.

Key Terms and Concepts

- Behavioral equation
- Cobb–Douglas model
- Complete system of equations
- Completeness condition
- Constant returns to scale
- Demand system
- Dynamic model
- Econometric model
- Equilibrium condition
- Exclusion restrictions
- Exogenous
- Flexible functional
- Full information estimator
- Full information maximum likelihood (FIML)
- Generalized regression model
- Homogeneity restriction
- Identical explanatory variables
- Identification
- Instrumental variable estimator
- Interdependent
- Invariance
- Jointly dependent
- Kronecker product
- Least variance ratio
- Likelihood ratio test
- Limited information estimator
- Limited information maximum likelihood (LIML) estimator
- Nonsample information
- Normalization
- Order condition
- Pooled model
- Predetermined variable
- Problem of identification
- Rank condition
- Reduced form
- Reduced-form disturbance
- Restrictions
- Seemingly unrelated regressions (SUR)
- Share equations
- Shephard's lemma
- Singular disturbance covariance matrix
- Simultaneous equations bias
- Structural disturbance
- Structural equation
- Structural form
- Systems of demand equations
- Three-stage least squares (3SLS) estimator
- Translog function
- Triangular system

Exercises

1. A sample of 100 observations produces the following sample data:

$$\bar{y}_1 = 1, \bar{y}_2 = 2, \mathbf{y}_1'\mathbf{y}_1 = 150, \mathbf{y}_2'\mathbf{y}_2 = 550, \mathbf{y}_1'\mathbf{y}_2 = 260.$$

The underlying seemingly unrelated regressions model is

$$y_1 = \mu + \varepsilon_1,$$
$$y_2 = \mu + \varepsilon_2.$$

a. Compute the OLS estimate of μ, and estimate the sampling variance of this estimator.
b. Compute the FGLS estimate of μ and the sampling variance of the estimator.

2. Consider estimation of the following two-equation model:

$$y_1 = \beta_1 + \varepsilon_1,$$
$$y_2 = \beta_2 x + \varepsilon_2.$$

A sample of 50 observations produces the following moment matrix:

$$\begin{array}{c} \\ 1 \\ y_1 \\ y_2 \\ x \end{array} \begin{array}{c} \begin{array}{cccc} 1 & y_1 & y_2 & x \end{array} \\ \begin{bmatrix} 50 & & & \\ 150 & 500 & & \\ 50 & 40 & 90 & \\ 100 & 60 & 50 & 100 \end{bmatrix} \end{array}.$$

a. Write the explicit formula for the GLS estimator of $[\beta_1, \beta_2]$. What is the asymptotic covariance matrix of the estimator?
b. Derive the OLS estimator and its sampling variance in this model.
c. Obtain the OLS estimates of β_1 and β_2, and estimate the sampling covariance matrix of the two estimates. Use n instead of $(n - 1)$ as the divisor to compute the estimates of the disturbance variances.
d. Compute the FGLS estimates of β_1 and β_2 and the estimated sampling covariance matrix.
e. Test the hypothesis that $\beta_2 = 1$.

3. The model

$$y_1 = \beta_1 x_1 + \varepsilon_1,$$
$$y_2 = \beta_2 x_2 + \varepsilon_2$$

satisfies all the assumptions of the seemingly unrelated regressions model. All variables have zero means. The following sample second-moment matrix is obtained from a sample of 20 observations:

$$\begin{array}{c} \\ y_1 \\ y_2 \\ x_1 \\ x_2 \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & y_2 & x_1 & x_2 \end{array} \\ \begin{bmatrix} 20 & 6 & 4 & 3 \\ 6 & 10 & 3 & 6 \\ 4 & 3 & 5 & 2 \\ 3 & 6 & 2 & 10 \end{bmatrix} \end{array}.$$

a. Compute the FGLS estimates of β_1 and β_2.
b. Test the hypothesis that $\beta_1 = \beta_2$.
c. Compute the maximum likelihood estimates of the model parameters.
d. Use the likelihood ratio test to test the hypothesis in part b.

4. Prove that in the model

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1,$$
$$\mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2,$$

generalized least squares is equivalent to equation-by-equation ordinary least squares if $\mathbf{X}_1 = \mathbf{X}_2$. The general case is considered in Exercise 14.

5. Consider the two-equation system

$$y_1 = \beta_1 x_1 + \varepsilon_1,$$
$$y_2 = \beta_2 x_2 + \beta_3 x_3 + \varepsilon_2.$$

Assume that the disturbance variances and covariance are known. Now suppose that the analyst of this model applies GLS but erroneously omits x_3 from the second equation. What effect does this specification error have on the consistency of the estimator of β_1?

6. Consider the system

$$y_1 = \alpha_1 + \beta x + \varepsilon_1,$$
$$y_2 = \alpha_2 + \varepsilon_2.$$

The disturbances are freely correlated. Prove that GLS applied to the system leads to the OLS estimates of α_1 and α_2 but to a mixture of the least squares slopes in the regressions of y_1 and y_2 on x as the estimator of β. What is the mixture? To simplify the algebra, assume (with no loss of generality) that $\bar{x} = 0$.

7. For the model

$$y_1 = \alpha_1 + \beta x + \varepsilon_1,$$
$$y_2 = \alpha_2 + \varepsilon_2,$$
$$y_3 = \alpha_3 + \varepsilon_3,$$

assume that $y_{i2} + y_{i3} = 1$ at every observation. Prove that the sample covariance matrix of the least squares residuals from the three equations will be singular, thereby precluding computation of the FGLS estimator. How could you proceed in this case?

8. Consider the following two-equation model:

$$y_1 = \gamma_1 y_2 + \beta_{11} x_1 + \beta_{21} x_2 + \beta_{31} x_3 + \varepsilon_1,$$
$$y_2 = \gamma_2 y_1 + \beta_{12} x_1 + \beta_{22} x_2 + \beta_{32} x_3 + \varepsilon_2.$$

a. Verify that, as stated, neither equation is identified.
b. Establish whether or not the following restrictions are sufficient to identify (or partially identify) the model:
   (1) $\beta_{21} = \beta_{32} = 0$,
   (2) $\beta_{12} = \beta_{22} = 0$,
   (3) $\gamma_1 = 0$,
   (4) $\gamma_1 = \gamma_2$ and $\beta_{32} = 0$,
   (5) $\sigma_{12} = 0$ and $\beta_{31} = 0$,
   (6) $\gamma_1 = 0$ and $\sigma_{12} = 0$,
   (7) $\beta_{21} + \beta_{22} = 1$,
   (8) $\sigma_{12} = 0, \beta_{21} = \beta_{22} = \beta_{31} = \beta_{32} = 0$,
   (9) $\sigma_{12} = 0, \beta_{11} = \beta_{21} = \beta_{22} = \beta_{31} = \beta_{32} = 0$.

9. Obtain the reduced form for the model in Exercise 8 under each of the assumptions made in parts a and in parts b(1) and b(9).
10. The following model is specified:

$$y_1 = \gamma_1 y_2 + \beta_{11} x_1 + \varepsilon_1,$$
$$y_2 = \gamma_2 y_1 + \beta_{22} x_2 + \beta_{32} x_3 + \varepsilon_2.$$

All variables are measured as deviations from their means. The sample of 25 observations produces the following matrix of sums of squares and cross products:

$$\begin{array}{c} \\ y_1 \\ y_2 \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccccc} y_1 & y_2 & x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 20 & 6 & 4 & 3 & 5 \\ 6 & 10 & 3 & 6 & 7 \\ 4 & 3 & 5 & 2 & 3 \\ 3 & 6 & 2 & 10 & 8 \\ 5 & 7 & 3 & 8 & 15 \end{bmatrix} \end{array}.$$

a. Estimate the two equations by OLS.
b. Estimate the parameters of the two equations by 2SLS. Also estimate the asymptotic covariance matrix of the 2SLS estimates.
c. Obtain the LIML estimates of the parameters of the first equation.
d. Estimate the two equations by 3SLS.
e. Estimate the reduced form coefficient matrix by OLS and indirectly by using your structural estimates from part b.

11. For the model

$$y_1 = \gamma_1 y_2 + \beta_{11} x_1 + \beta_{21} x_2 + \varepsilon_1,$$
$$y_2 = \gamma_2 y_1 + \beta_{32} x_3 + \beta_{42} x_4 + \varepsilon_2$$

show that there are two restrictions on the reduced form coefficients. Describe a procedure for estimating the model while incorporating the restrictions.

12. Prove that

$$\text{plim} \frac{\mathbf{Y}_m' \boldsymbol{\varepsilon}_m}{T} = \underline{\boldsymbol{\omega}}_m - \boldsymbol{\Omega}_{mm} \boldsymbol{\gamma}_m.$$

13. Prove that an underidentified equation cannot be estimated by 2SLS.
14. Prove the general result in point 2 in Section 10.2.2, if the $\mathbf{X}$ matrices in (10-1) are identical, then full GLS is equation-by-equation OLS. *Hints:* If all the $\mathbf{X}_m$ matrices are identical, then the inverse matrix in (10-10) is $[\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X}]^{-1}$. Also, $\mathbf{X}_m'\mathbf{y}_m = \mathbf{X}'\mathbf{y}_m = \mathbf{X}'\mathbf{X}\mathbf{b}_m$. Use these results to show that for the first equation,

$$\hat{\boldsymbol{\beta}}_1 = \sum_{n=1}^{M} \sigma_{1n} \sum_{l=1}^{M} \sigma^{nl} \mathbf{b}_l = \mathbf{b}_1 \left(\sum_{n=1}^{M} \sigma_{1n} \sigma^{n1} \right) + \mathbf{b}_2 \left(\sum_{n=1}^{M} \sigma_{1n} \sigma^{n2} \right) + \cdots$$
$$+ \mathbf{b}_M \left(\sum_{n=1}^{M} \sigma_{1n} \sigma^{nM} \right),$$

and likewise for the others.

Applications

Some of these applications will require econometric software for the computations. The calculations are standard, and are available as commands in, for example, *Stata, SAS, E-Views* or *LIMDEP*, or as existing programs in *R*.

1. **Statewide aggregate production function.** Continuing Example 10.1, data on output, the capital stocks, and employment are aggregated by summing the values for the individual states (before taking logarithms). The unemployment rate for each region, m, at time t is determined by a weighted average of the unemployment rates for the states in the region, where the weights are

$$w_{nt} = emp_{nt} / \sum_{j=1}^{M_m} emp_{jt},$$

where M_m is the number of states in region m. Then, the unemployment rate for region m at time t is the following average of the unemployment rates of the states (n) in region (m) at time t:

$$unemp_{mt} = \Sigma_j w_{nt}(j) unemp_{nt}(j).$$

2. Continuing the analysis of Section 10.3.2, we find that a translog cost function for one output and three factor inputs that does not impose constant returns to scale is

$$\begin{aligned}\ln C = {} & \alpha + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln p_3 + \delta_{11} \tfrac{1}{2} \ln^2 p_1 + \delta_{12} \ln p_1 \ln p_2 \\ & + \delta_{13} \ln p_1 \ln p_3 + \delta_{22} \tfrac{1}{2} \ln^2 p_2 + \delta_{23} \ln p_2 \ln p_3 + \delta_{33} \tfrac{1}{2} \ln^2 p_3 \\ & + \gamma_{q1} \ln Q \ln p_1 + \gamma_{q2} \ln Q \ln p_2 + \gamma_{q3} \ln Q \ln p_3 \\ & + \beta_q \ln Q + \beta_{qq} \tfrac{1}{2} \ln^2 Q + \varepsilon_c.\end{aligned}$$

The factor share equations are

$$\begin{aligned}S_1 &= \beta_1 + \delta_{11} \ln p_1 + \delta_{12} \ln p_2 + \delta_{13} \ln p_3 + \gamma_{q1} \ln Q + \varepsilon_1, \\ S_2 &= \beta_2 + \delta_{12} \ln p_1 + \delta_{22} \ln p_2 + \delta_{23} \ln p_3 + \gamma_{q2} \ln Q + \varepsilon_2, \\ S_3 &= \beta_3 + \delta_{13} \ln p_1 + \delta_{23} \ln p_2 + \delta_{33} \ln p_3 + \gamma_{q3} \ln Q + \varepsilon_3.\end{aligned}$$

[See Christensen and Greene (1976) for analysis of this model.]

a. The three factor shares must add identically to 1. What restrictions does this requirement place on the model parameters?
b. Show how the adding-up condition in (10-33) can be imposed directly on the model by specifying the translog model in (C/p_3), (p_1/p_3), and (p_2/p_3) and dropping the third share equation. (See Example 10.3.) Notice that this reduces the number of free parameters in the model to 10.
c. Continuing part b, the model as specified with the symmetry and equality restrictions has 15 parameters. By imposing the constraints, you reduce this number to 10 in the estimating equations. How would you obtain estimates of the parameters not estimated directly?
d. Estimate each of the three equations you obtained in part b by ordinary least squares. Do the estimates appear to satisfy the cross-equation equality and symmetry restrictions implied by the theory?

e. Using the data in Section 10.3.1, estimate the full system of three equations (cost and the two independent shares), imposing the symmetry and cross-equation equality constraints.
f. Using your parameter estimates, compute the estimates of the elasticities in (10-34) at the means of the variables.
g. Use a likelihood ratio statistic to test the joint hypothesis that $\gamma_{qi} = 0, i = 1, 2, 3$. [*Hint:* Just drop the relevant variables from the model.]

3. The Grunfeld investment data in Appendix Table 10.4 constitute a classic data set that has been used for decades to develop and demonstrate estimators for seemingly unrelated regressions.[28] Although somewhat dated at this juncture, they remain an ideal application of the techniques presented in this chapter. The data consist of time series of 20 yearly observations on 10 firms. The three variables are

I_{it} = gross investment,
F_{it} = market value of the firm at the end of the previous year,
C_{it} = value of the stock of plant and equipment at the end of the previous year.

The main equation in the studies noted is

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it}.$$

a. Fit the 10 equations separately by ordinary least squares and report your results.
b. Use a Wald (Chow) test to test the "aggregation" restriction that the 10 coefficient vectors are the same.
c. Use the seemingly unrelated regressions (FGLS) estimator to reestimate the parameters of the model, once again, allowing the coefficients to differ across the 10 equations. Now, use the pooled model and, again, FGLS, to estimate the constrained equation with equal parameter vectors, and test the aggregation hypothesis.
d. Using the OLS residuals from the separate regressions, use the LM statistic in (10-17) to test for the presence of cross-equation correlation.
e. An alternative specification to the model in part c that focuses on the variances rather than the means is a groupwise heteroscedasticity model. For the current application, you can fit this model using (10-20), (10-21), and (10-22), while imposing the much simpler model with $\sigma_{ij} = 0$ when $i \neq j$. Do the results of the pooled model differ in the two cases considered, simple OLS and groupwise heteroscedasticity?

4. The data in Appendix Table F5.2 may be used to estimate a small macroeconomic model. Use these data to estimate the model in Example 10.5. Estimate the parameters of the two equations by two-stage and three-stage least squares.
5. Using the cost function estimates in Example 10.2, we obtained an estimate of the efficient scale, $Q* = \exp[(1 - \beta_q)/(2\beta_{qq})]$. We can use the delta method in Section 4.5.4 to compute an asymptotic standard error for the estimator of Q^* and a confidence interval. The estimators of the two parameters are $b_q = 0.23860$ and $b_{qq} = 0.04506$. The estimates of the asymptotic covariance matrix are

[28] See Grunfeld (1958), Grunfeld and Griliches (1960), Boot and de Witt (1960), and Kleiber and Zeileis (2010).

$v_q = 0.00344554$, $v_{qq} = 0.0000258021$, $c_{q,qq} = -0.000291067$. Use these results to form a 95% confidence interval for Q^*. (*Hint:* $\partial Q^*/\partial b_j = Q^* \partial \ln Q^*/\partial b_j$.)

6. Using the estimated health outcomes model in Example 10.8, determine the expected values of *ln Income* and *Health Satisfaction* for a person with the following characteristics: *Female* = 1, *Working* = 1, *Public* = 1, *AddOn* = 0, *Education* = 14, *Married* = 1, *HHKids* = 1, *Age* = 35. Now, repeat the calculation with the same person but with *Age* = 36. Likewise, with *Female* = 0 (and *Age* = 35). Note, the sample range of *Income* is 0 – 3.0, with sample mean approximately 0.4. The income data are in 10,000DM units (pre-Euro). In both cases, note how the health satisfaction outcome changes when the exogenous variable (*Age* or *Female*) changes (by one unit).

11
MODELS FOR PANEL DATA

11.1 INTRODUCTION

Data sets that combine time series and cross sections are common in economics. The published statistics of the OECD contain numerous series of economic aggregates observed yearly for many countries. The Penn World Tables [CIC (2010)] is a data bank that contains national income data on 167 countries for more than 60 years. Recently constructed **longitudinal data sets** contain observations on thousands of individuals or families, each observed at several points in time. Other empirical studies have examined time-series data on sets of firms, states, countries, or industries simultaneously. These data sets provide rich sources of information about the economy. The analysis of panel data allows the model builder to learn about economic processes while accounting for both heterogeneity across individuals, firms, countries, and so on and for dynamic effects that are not visible in cross sections. Modeling in this context often calls for complex stochastic specifications. In this chapter, we will survey the most commonly used techniques for time-series—cross-section (e.g., cross-country) and panel (e.g., longitudinal)—data. The methods considered here provide extensions to most of the models we have examined in the preceding chapters. Section 11.2 describes the specific features of panel data. Most of this analysis is focused on individual data, rather than cross-country aggregates. We will examine some aspects of aggregate data modeling in Section 11.10. Sections 11.3, 11.4, and 11.5 consider in turn the three main approaches to regression analysis with panel data, pooled regression, the fixed effects model, and the random effects model. Section 11.6 considers robust estimation of covariance matrices for the panel data estimators, including a general treatment of cluster effects. Sections 11.7 through 11.10 examine some specific applications and extensions of panel data methods. Spatial autocorrelation is discussed in Section 11.7. In Section 11.8, we consider sources of endogeneity in the random effects model, including a model of the sort considered in Chapter 8 with an endogenous right-hand-side variable and then two approaches to dynamic models. Section 11.9 builds the fixed and random effects models into nonlinear regression models. Finally, Section 11.10 examines random parameter models. The random parameters approach is an extension of the fixed and random effects model in which the heterogeneity that the FE and RE models build into the constant terms is extended to other parameters as well.

Panel data methods are used throughout the remainder of this book. We will develop several extensions of the fixed and random effects models in Chapter 14 on maximum likelihood methods, and in Chapter 15 where we will continue the development of random parameter models that is begun in Section 11.10. Chapter 14 will also present methods for handling discrete distributions of random parameters under the heading of

latent class models. In Chapter 21, we will return to the models of nonstationary panel data that are suggested in Section 11.8.4. The fixed and random effects approaches will be used throughout the applications of discrete and limited dependent variables models in microeconometrics in Chapters 17, 18, and 19.

11.2 PANEL DATA MODELING

Many recent studies have analyzed panel, or longitudinal, data sets. Two very famous ones are the *National Longitudinal Survey of Labor Market Experience* (NLS, www.bls.gov/nls/nlsdoc.htm) and the *Michigan Panel Study of Income Dynamics* (PSID, http://psidonline.isr.umich.edu/). In these data sets, very large cross sections, consisting of thousands of microunits, are followed through time, but the number of periods is often quite small. The PSID, for example, is a study of roughly 6,000 families and 15,000 individuals who have been interviewed periodically from 1968 to the present. In contrast, the *European Community Household Panel* (ECHP, http://ec.europa.eu/eurostat/web/microdata/european-community-household-panel) ran for a total of eight years (waves). An ongoing study in the United Kingdom is the *Understanding Society* survey (www.understandingsociety.ac.uk/about) that grew out of the *British Household Panel Survey* (BHPS). This survey that was begun in 1991 with about 5,000 households has expanded to over 40,000 participants. Many very rich data sets have recently been developed in the area of health care and health economics, including the *German Socioeconomic Panel* (GSOEP, www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/GSOEP.aspx), AHRQ's *Medical Expenditure Panel Survey* (MEPS, www.meps.ahrq.gov/), and the *Household Income and Labour Dynamics in Australia* (HILDA, www.melbourneinstitute.com/hilda/). Constructing long, evenly spaced time series in contexts such as these would be prohibitively expensive, but for the purposes for which these data are typically used, it is unnecessary. Time effects are often viewed as transitions or discrete changes of state. The Current Population Survey (CPS, www.census.gov/cps/), for example, is a monthly survey of about 50,000 households that interviews households monthly for four months, waits for eight months, then reinterviews. This two-wave, **rotating panel** format allows analysis of short-term changes as well as a more general analysis of the U.S. national labor market. They are typically modeled as specific to the period in which they occur and are not carried across periods within a cross-sectional unit.[1] Panel data sets are more oriented toward cross-section analyses; they are wide but typically short. Heterogeneity across units is an integral part—indeed, often the central focus—of the analysis. [See, e.g., Jones and Schurer (2011).]

The analysis of panel or longitudinal data is the subject of one of the most active and innovative bodies of literature in econometrics,[2] partly because panel data provide such a rich environment for the development of estimation techniques and theoretical results. In more practical terms, however, researchers have been able to use time-series cross-sectional data to examine issues that could not be studied in either cross-sectional

[1]Formal time-series modeling for panel data is briefly examined in Section 21.4.

[2]A compendium of the earliest literature is Maddala (1993). Book-length surveys on the econometrics of panel data include Hsiao (2003), Dielman (1989), Matyas and Sevestre (1996), Raj and Baltagi (1992), Nerlove (2002), Arellano (2003), and Baltagi (2001, 2013, 2015). There are also lengthy surveys devoted to specific topics, such as limited dependent variable models [Hsiao, Lahiri, Lee, and Pesaran (1999)], discrete choice models [Greene (2015)] and semiparametric methods [Lee (1998)].

or time-series settings alone. Recent applications have allowed researchers to study the impact of health policy changes[3] and, more generally, the dynamics of labor market behavior. In principle, the methods of Chapters 6 and 21 can be applied to longitudinal data sets. In the typical panel, however, there are a large number of cross-sectional units and only a few periods. Thus, the time-series methods discussed there may be somewhat problematic. Recent work has generally concentrated on models better suited to these short and wide data sets. The techniques are focused on cross-sectional variation, or heterogeneity. In this chapter, we shall examine in detail the most widely used models and look briefly at some extensions.

11.2.1 GENERAL MODELING FRAMEWORK FOR ANALYZING PANEL DATA

The fundamental advantage of a panel data set over a cross section is that it will allow the researcher great flexibility in modeling differences in behavior across individuals. The basic framework for this discussion is a regression model of the form

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}. \end{aligned} \tag{11-1}$$

There are K regressors in $\mathbf{x}_{it}$, not including a constant term. The **heterogeneity**, or **individual effect**, is $\mathbf{z}_i'\boldsymbol{\alpha}$ where $\mathbf{z}_i$ contains a constant term and a set of individual or group-specific variables, which may be observed, such as race, sex, location, and so on; or unobserved, such as family specific characteristics, individual heterogeneity in skill or preferences, and so on, all of which are taken to be constant over time t. As it stands, this model is a classical regression model. If $\mathbf{z}_i$ is observed for all individuals, then the entire model can be treated as an ordinary linear model and fit by least squares. The complications arise when c_i is unobserved, which will be the case in most applications. Consider, for example, analyses of the effect of education and experience on earnings from which "ability" will always be a missing and unobservable variable. In health care studies, for example, of usage of the health care system, "health" and "health care" will be unobservable factors in the analysis.

The main objective of the analysis will be consistent and efficient estimation of the **partial effects**,

$$\boldsymbol{\beta} = \partial E[y_{it}|\mathbf{x}_{it}]/\partial \mathbf{x}_{it}.$$

Whether this is possible depends on the assumptions about the unobserved effects. We begin with a **strict exogeneity** assumption for the independent variables,

$$E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, c_i] = E[\varepsilon_{it}|\mathbf{X}_i, c_i] = 0.$$

This implies the current disturbance is uncorrelated with the independent variables in every period, past, present, and future. A looser assumption of contemporaneous exogeneity is sometimes useful. If

$$E[y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i] = E[y_{it}|\mathbf{x}_{it}, c_i] = \mathbf{x}_{it}'\boldsymbol{\beta} + c_i,$$

then

$$E[\varepsilon_{it}|\mathbf{x}_{it}, c_i] = 0.$$

[3]For example, Riphahn et al.'s (2003) analysis of reforms in German public health insurance regulations.

The regression model with this assumption restricts influences of $\mathbf{x}$ on $E[y|\mathbf{x}, c]$ to the current period. In this form, we can see that we have ruled out dynamic models such as

$$y_{it} = \mathbf{w}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + c_i + \varepsilon_{it}$$

because as long as γ is nonzero, covariation between ε_{it} and $\mathbf{x}_{it} = (\mathbf{w}_{it}, y_{i,t-1})$ is transmitted through c_i in $y_{i,t-1}$. We will return to dynamic specifications in Section 11.8.3. In some settings (such as the static fixed effects model in Section 11.4), strict exogeneity is stronger than necessary. It is, however, a natural assumption. It will prove convenient to start there, and loosen the assumption in specific cases where it would be useful.

The crucial aspect of the model concerns the heterogeneity. A convenient assumption is mean independence,

$$E[c_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots] = \alpha.$$

If the unobserved variable(s) are uncorrelated with the included variables, then, as we shall see, they may be included in the disturbance of the model. This is the assumption that underlies the random effects model, as we will explore later. It is, however, a particularly strong assumption—it would be unlikely in the labor market and health care examples mentioned previously. The alternative would be

$$E[c_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots,] = h(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots) = h(\mathbf{X}_i)$$

for some unspecified, but nonconstant function of $\mathbf{X}_i$. This formulation is more general, but at the same time, considerably more complicated, the more so because estimation may require yet further assumptions about the nature of the regression function.

11.2.2 MODEL STRUCTURES

We will examine a variety of different models for panel data. Broadly, they can be arranged as follows:

1. **Pooled Regression:** If $\mathbf{z}_i$ contains only a constant term, then ordinary least squares provides consistent and efficient estimates of the common α and the slope vector $\boldsymbol{\beta}$.
2. **Fixed Effects:** If $\mathbf{z}_i$ is unobserved, but correlated with $\mathbf{x}_{it}$, then the least squares estimator of $\boldsymbol{\beta}$ is biased and inconsistent as a consequence of an omitted variable. However, in this instance, the model

   $$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it},$$

   where $\alpha_i = \mathbf{z}_i'\boldsymbol{\alpha}$, embodies all the observable effects and specifies an estimable conditional mean. This **fixed effects** approach takes α_i to be a group-specific constant term in the regression model. It should be noted that the term "fixed" as used here signifies the correlation of c_i and $\mathbf{x}_{it}$, not that c_i is nonstochastic.
3. **Random Effects:** If the unobserved individual heterogeneity, however formulated, is uncorrelated with $\mathbf{x}_{it}$, then the model may be formulated as

   $$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + E[\mathbf{z}_i'\boldsymbol{\alpha}] + \{\mathbf{z}_i'\boldsymbol{\alpha} - E[\mathbf{z}_i'\boldsymbol{\alpha}]\} + \varepsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + u_i + \varepsilon_{it}, \end{aligned}$$

   that is, as a linear regression model with a compound disturbance that may be consistently, albeit inefficiently, estimated by least squares. This random effects

approach specifies that u_i is a group-specific random element, similar to ε_{it} except that for each group, there is but a single draw that enters the regression identically in each period. Again, the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not. We will examine this basic formulation, then consider an extension to a dynamic model.

4. **Random Parameters:** The random effects model can be viewed as a regression model with a random constant term. With a sufficiently rich data set, we may extend this idea to a model in which the other coefficients vary randomly across individuals as well. The extension of the model might appear as

$$y_{it} = \mathbf{x}_{it}'(\boldsymbol{\beta} + \mathbf{u}_i) + (\alpha + u_i) + \varepsilon_{it},$$

where $\mathbf{u}_i$ is a random vector that induces the variation of the parameters across individuals. This random parameters model has recently enjoyed widespread attention in several fields. It represents a natural extension in which researchers broaden the amount of heterogeneity across individuals while retaining some commonalities—the parameter vectors still share a common mean. Some recent applications have extended this yet another step by allowing the mean value of the parameter distribution to be person specific, as in

$$y_{it} = \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{u}_i) + (\alpha + u_i) + \varepsilon_{it},$$

where $\mathbf{z}_i$ is a set of observable, person-specific variables, and $\boldsymbol{\Delta}$ is a matrix of parameters to be estimated. As we will examine in Chapter 17, this **hierarchical model** is extremely versatile.

11.2.3 EXTENSIONS

The short list of model types provided earlier only begins to suggest the variety of applications of panel data methods in econometrics. We will begin in this chapter to study some of the formulations and uses of linear models. The random and fixed effects models and random parameters models have also been widely used in models of censoring, binary, and other discrete choices, and models for event counts. We will examine all of these in the chapters to follow. In some cases, such as the models for count data in Chapter 18, the extension of random and fixed effects models is straightforward, if somewhat more complicated computationally. In others, such as in binary choice models in Chapter 17 and censoring models in Chapter 19, these panel data models have been used, but not before overcoming some significant methodological and computational obstacles.

11.2.4 BALANCED AND UNBALANCED PANELS

By way of preface to the analysis to follow, we note an important aspect of panel data analysis. As suggested by the preceding discussion, a panel data set will consist of n sets of observations on individuals to be denoted $i = 1, \ldots, n$. If each individual in the data set is observed the same number of times, usually denoted T, the data set is a **balanced panel**. An **unbalanced panel** data set is one in which individuals may be observed different numbers of times. We will denote this T_i. A **fixed panel** is one in which the same set of individuals is observed for the duration of the study. The data sets we will examine in this chapter, while not all balanced, are fixed.

A rotating panel is one in which the cast of individuals changes from one period to the next. For example, Gonzalez and Maloney (1999) examined self-employment decisions in Mexico using the National Urban Employment Survey. This is a quarterly data set drawn from 1987 to 1993 in which individuals are interviewed five times. Each quarter, one-fifth of the individuals is rotated out of the data set. The U.S. Census Bureau's SIPP data (Survey of Income and Program Participation, www.census.gov/programs-surveys/sipp/data.html) is another rotating panel. Some discussion and numerous references may be found in Baltagi (2013)

Example 11.1 A Rotating Panel: The Survey of Income and Program Participation (SIPP) Data

From the Census Bureau's home site for this data set:

The SIPP survey design is a continuous series of national panels, with sample size ranging from approximately 14,000 to 52,000 interviewed households. The duration of each panel ranges from $2\frac{1}{2}$ years to 4 years. The SIPP sample is a multistage-stratified sample of the U.S. civilian non-institutionalized population. From 1984 to 1993, a new panel of households was introduced each year in February. A 4-year panel was implemented in April 1996; however, a 3-year panel that was started in February 2000 was canceled after 8 months due to budget restrictions. Consequently, a 3-year panel was introduced in February 2001. The $2\frac{1}{2}$ year 2004 SIPP Panel was started in February 2004 and was the first SIPP panel to use the 2000 decennial-based redesign of the sample. The 2014 panel, starting in February 2014, is the first SIPP panel to use the 2010 decennial as the basis for its sample.

11.2.5 ATTRITION AND UNBALANCED PANELS

Unbalanced panels arise in part because of nonrandom attrition from the sample. Individuals may appear for only a subset of the waves. In general, if the attrition is systematically related to the outcome variable in the model being studied, then it may induce conditions of nonrandom sampling bias—sometimes called *sample selection*. The nature of the bias is unclear, but sample selection bias as a general aspect of econometric analysis is well documented. [An example would be attrition of subjects from a medical clinical trial for reasons related to the efficacy (or lack of) of the drug under study.] Verbeek and Nijman (1992) proposed a nonconstructive test for attrition in panel data models—the test results detect the condition but do not imply a strategy if the hypothesis of no nonrandom attrition is rejected. Wooldridge (2002 and 2010, pp. 837–844) describes an *inverse probability weighting* (IPW) approach for correcting for nonrandom attrition.

Example 11.2 Attrition and Inverse Probability Weighting in a Model for Health

Contoyannis, Jones, and Rice (2004) employed an ordered probit model to study self-assessed health in the first eight waves of the BHPS.[4] The sample exhibited some attrition as shown in Table 11.1 (from their Table V). (Although the sample size does decline after each wave, the remainder at each wave is not necessarily a subset of the previous wave. Some individuals returned to the sample. A subsample of observations for which attrition at each wave was an *absorbing state*—they did not return—was analyzed separately. This group is used for IPW-2 in the results below.) To examine the issue of nonrandom attrition, the authors first employed Nijman and Verbeek's tests. This entails adding three variables to the model:

[4]See Chapter 18 and Greene and Hensher (2010).

TABLE 11.1 Attrition from BHPS

Wave	*Individuals*	*Survival*	*Exited*	*Attrition*
1	10,256	—	—	—
2	8,957	87.33%	1299	12.67%
3	8,162	79.58%	795	8.88%
4	7,825	76.30%	337	4.13%
5	7,430	72.45%	395	5.05%
6	7,238	70.57%	192	2.58%
7	7,102	69.25%	136	1.88%
8	6,839	66.68%	263	3.70%

$NEXT\ WAVE_{it}$ = 1 if individual i is in the sample in wave $t + 1$,
$ALL\ WAVE_{it}$ = 1 if individual i is in the sample for all waves,
NUMBER OF WAVES = the number of waves for which the individual is present.

The results at this step included those in Table 11.2 (extracted from their Table IX). Curiously, at this step, the authors found strong evidence of nonrandom attrition in the subsample of men in the sample, but not in that for women. The authors then employed an inverse probability weighting approach to "correct" for the possibility of nonrandom attrition. They employed two procedures. First, for each individual in the sample, construct $\mathbf{d}_i = (d_{i1}, \ldots, d_{iT})$ where $d_{it} = 1$ if individual i is present in wave t. By construction, $d_{i1} = 1$ for everyone. A vector of covariates observed at the baseline that is thought to be relevant to attrition in each period is designated $\mathbf{z}_{i1}$. This includes ln Income, marital status, age, race, education, household size, and health status, and some indicators of morbidity. For each period, a probit model is fit for $\text{Prob}(d_{it} = 1 \,|\, \mathbf{z}_{i1})$ and fitted probabilities, $\hat{p}_{it}$ are computed. (Note: $\hat{p}_{i1} = 1$.) With these fitted probabilities in hand, the model is estimated by maximizing the criterion function, in their case, the log-likelihood function, $\ln L = \Sigma_i \Sigma_t (d_{it}/\hat{p}_{it}) \ln L_{it}$. (For the models examined in this chapter, the log-likelihood term would be the negative of a squared residuals to maximize the negative of the sum or squares.) These results are labeled IPW-1 in Table 11.3. For the second method, the sample is restricted to the subset for which attrition was permanent. For each period, the list of variables is expanded to include z_{i1} and $z_{i,t-1}$. The predicted probabilities at each, computed using the probit model, are denoted $\hat{\pi}_{is}$. Finally, to account for the fact that the sample at each wave is based on selection from the previous wave (so that $d_{it} = \Pi_{s \le t} d_{is}$) the probabilities are likewise adjusted: $\hat{p}_{it} = \Pi_{s=1}^{t} \hat{\pi}_{is}$. The results below show the influence of the sample treatment on one of the estimated coefficients in the full model.

TABLE 11.2 Tests for Attrition Bias

	Men		*Women*	
	β	***t Ratio***	***β***	***t Ratio***
NEXT WAVE	0.199	5.67	0.060	1.77
ALL WAVES	0.139	4.46	0.071	2.45
NUMBER OF WAVES	0.031	3.54	0.016	1.88

TABLE 11.3 Estimated Coefficients* on ln Income in Ordered Probit Models (Standard errors in Parentheses)

	Balanced Sample	*Unbalanced*	*IPW-1*	*IPW-2*
	NT = 19,460	NT = 24,371	NT = 24,370	NT = 23,211
Men	0.036 (0.022)	0.035 (0.019)	0.035 (0.020)	0.043 (0.021)
Women	0.029 (0.021)	0.033 (0.018)	0.021 (0.019)	0.018 (0.020)

*Coefficient on ln Income in Dynamic Ordered Probit Model. (Extracted from Table X and Table XI.)

Example 11.3 Attrition and Sample Selection in an Earnings Model for Physicians

Cheng and Trivedi (2015) approached the attrition question from a nonrandom sample selection perspective in their panel data study of Australian physicians' earnings. The starting point is a "missing at random" (MAR) interpretation of attrition. If individuals exit the sample for reasons that are unrelated to the variable under study—specifically, unrelated to the unobservables in the equation being used to model that variable—then attrition has no direct implications for the estimation of the parameters of the model.

Table 11.4 (derived from Table I in the article) shows that about one-third of the initial sample in their four-wave panel ultimately exited the sample. (Some individuals did return. The table shows the net effect.)

The model is a structural system,

$$\textit{Attrition}: \quad A_{it}^{*} = \mathbf{z}_{it}'\boldsymbol{\gamma} + u_{it}; \qquad A_{it} = 1 \text{ if } A_{it}^{*} > 0,$$

$$\ln \textit{Wages}: \quad y_{it}^{*} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{f}_i'\boldsymbol{\delta} + \alpha_i + \varepsilon_{it}; \; y_{it} = y_{it}^{*} \text{ if } A_{it} = 0, \text{ unobserved otherwise,}$$

where $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are time-varying exogenous variables, $\mathbf{f}_i$ is time-invariant, possibly endogenous variables, and α_i is a fixed effect. This setup is an application of Heckman's (1979) sample selection framework. (See Section 19.5.) The implication of the observation mechanism for the observed data is

$$\begin{aligned} E[y_{it} | \mathbf{x}_{it}, \mathbf{f}_i, \alpha_i, A_{it} = 0] &= \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{f}_i'\boldsymbol{\delta} + \alpha_i + E[\varepsilon_{it} | u_{it} \leq -\mathbf{z}_{it}'\boldsymbol{\gamma}] \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{f}_i'\boldsymbol{\delta} + \alpha_i + \theta\, \lambda(\mathbf{z}_{it}'\boldsymbol{\gamma}). \end{aligned}$$

[In this reduced form of the model, θ is not (yet) a structural parameter. A nonzero value of this coefficient implies the presence of the attrition (selection) effect. The effect is generic until some structure is placed on the joint observation and attrition mechanism.] If ε_{it} and u_{it} are correlated, then $[\theta\lambda(\mathbf{z}_{it}'\boldsymbol{\gamma})]$ will be nonzero. Regression of y_{it} on $\mathbf{x}_{it}$, $\mathbf{f}_i$, and whatever device is used to control for the fixed effects will be affected by the missing *selection effect*, $\lambda_{it} = \lambda(\mathbf{z}_{it}'\boldsymbol{\gamma})$. If this omitted variable is correlated with $(\mathbf{x}_{it}, \mathbf{f}_i, \alpha_i)$, then the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are likely to be distorted. A partial solution is obtained by using first differences in the

TABLE 11.4 Attrition from the Medicine in Australia Balancing Employment and Life Data

	General Practitioners			*Specialists*		
Year	***N***	***Attrition****	***Survival***	***N***	***Attrition***	***Survival***
1	3906	840	100.0%	4596	926	100.0%
2	3066	242	78.5%	3670	303	79.9%
3	2824	270	72.3%	3367	299	73.3%
4	2554	—	65.4%	3068	—	66.8%

* Net attrition takes place after the indicated year.

regression. First differences will eliminate the time-invariant components of the regression, $(\mathbf{f}_i, \alpha_i)$, but will not solve the selection problem unless the attrition mechanism is also time invariant, which is not assumed. This nonzero correlation will be the *attrition effect*.

If there is attrition bias (in the estimator that ignores attrition), then the sample should become progressively less random as the observation period progresses. This suggests a possible indirect test for attrition bias. The full *unbalanced* sample contains a *balanced* subsample of individuals who are present for all waves of the panel. (Individuals who left and rejoined the panel would be bypassed for purposes of this exercise.) Under the MAR assumption, estimation of $\boldsymbol{\beta}$ based on the unbalanced full sample and the balanced subsample should produce the same results (aside from some sampling variability). This suggests one might employ a Hausman style test. (See Section 11.5.6.) The authors employed a more direct strategy. A narrow assumption that $(\varepsilon_{it}, u_{it})$ are bivariate normally distributed with zero means, variances, σ^2 and 1, and correlation ρ (a variance for u_{it} is not identified) produces

$$\theta_t \, \lambda(\mathbf{z}_{it}'\boldsymbol{\gamma}_t) = \theta_t \frac{\phi(-\mathbf{z}_{it}'\boldsymbol{\gamma}_t)}{\Phi(-\mathbf{z}_{it}'\boldsymbol{\gamma}_t)}.$$

Estimates of the coefficients in this "control function" regression are computed for each of waves 2–4 and added to the first difference regression,

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \sum_{t=2}^{4} \theta_t \hat{\lambda}_{it} + w_{it},$$

which is then estimated using least squares. Standard errors are computed using bootstrapping. Under the joint normality assumption, this control function estimator is robust, in that if there is an attrition effect (nonzero ρ), the effect is accounted for while if $\rho = 0$, the original estimator (within or first differences) will be consistent on its own. A second approach that loosens the bivariate normality assumption is based on a copula model (Section 12.2.2) that is estimated by maximum likelihood.

Table 11.5 below (derived from Tables III and IV in the paper) summarizes the results. The bivariate normal model strongly suggests the presence of the attrition effect, though the impact on the main estimation result is relatively modest. But the results for the copula are quite different. The effect is found to be significant only for the specialists. The impact on the hours coefficient is quite large for this group as well.

TABLE 11.5 Earnings Models and Tests for Attrition Bias

	General Practitioners	*Specialists*
Fixed Effects Hours Coefficient		
Unbalanced*	0.460 (0.027) [7776]	0.287 (0.022) [8904]
Balanced	0.407 (0.038) [3464]	0.356 (0.029) [4204]
First Differences Hours Coefficient		
Unbalanced	0.428 (0.042) [4106]	0.174 (0.038) [4291]
Balanced	0.387 (0.055) [2598]	0.244 (0.053) [3153]
Bivariate Normal Hazards Attrition Model		
Hours coefficient	0.422 (0.041) [4043]	0.180 (0.035) [4875]
Wald Statistic (3 df)	42.47	38.65
p Value	0.000	0.000
Frank Copula Attrition Model		
Marginals	Probit, Student's *t*	Logit, logistic
Hours coefficient	0.315 (0.043) [5166]	0.104 (0.026) [6109]
Wald Statistic (1 df)	1.862	7535.119
p Value	0.172	0.000

* Standard errors in parentheses. Sample size in brackets.

Unbalanced panels may arise for systematic reasons that induce problems that look like sample selection issues. But the attrition from a panel data set may also be completely ignorable, that is, due to issues that are out of the view of the analyst. In such cases, it is reasonable simply to treat the unbalanced nature of the data as a characteristic of the random sampling. Almost none of the useful theory that we will examine here relies on an assumption that the panel is balanced. The development to follow is structured so that the distinction between balanced and unbalanced panels, beyond the attrition issue, will entail little more than a trivial change in notation—where for convenience we write T suggesting a balanced panel, merely changing T to T_i generalizes the results. We will note specifically when this is not the case, such as in Breusch and Pagan's (1980) LM statistic.

11.2.6 WELL-BEHAVED PANEL DATA

The asymptotic properties of the estimators in the classical regression model were established in Section 4.4 under the following assumptions:

A.1. ***Linearity:*** $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$.
A.2. ***Full rank:*** The $n \times K$ sample data matrix, $\mathbf{X}$ has full column rank for every $n > K$.
A.3. ***Strict exogeneity of the independent variables:*** $E[\varepsilon_i | x_{j1}, x_{j2}, \ldots, x_{jK}] = 0, i, j = 1, \ldots, n$.
A.4. ***Homoscedasticity and nonautocorrelation:*** $E[\varepsilon_i \varepsilon_j | \mathbf{X}] = \sigma_\varepsilon^2$ if $i = j$ and 0 otherwise.

The following are the crucial results needed: For consistency of $\mathbf{b}$, we need

$$\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \text{plim}\ \overline{\mathbf{Q}}_n = \mathbf{Q}, \text{ a positive definite matrix,}$$
$$\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \text{plim}\ \overline{\mathbf{w}}_n = E[\overline{\mathbf{w}}_n] = \mathbf{0}.$$

(For consistency of s^2, we added a fairly weak assumption about the moments of the disturbances.) To establish asymptotic normality, we required consistency and

$$\sqrt{n}\,\overline{\mathbf{w}}_n \xrightarrow{d} N[0, \sigma^2\mathbf{Q}].$$

With these in place, the desired characteristics are then established by the methods of Sections 4.4.1 and 4.4.2.

Exceptions to the assumptions are likely to arise in a **panel data** set. The sample will consist of multiple observations on each of many observational units. For example, a study might consist of a set of observations made at different points in time on a large number of families. In this case, the $\mathbf{x}$'s will surely be correlated across observations, at least within observational units. They might even be the same for all the observations on a single family.

The panel data set could be treated as follows. Assume for the moment that the data consist of a fixed number of observations, say T, on a set of n families, so that the total number of rows in $\mathbf{X}$ is $N = nT$. The matrix $\overline{\mathbf{Q}}_n$, in which n is all the observations in the sample, is

$$\overline{\mathbf{Q}}_n = \frac{1}{n}\sum_i \frac{1}{T}\mathbf{X}_i'\mathbf{X}_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_i.$$

We then view the set of observations on the ith unit as if they were a single observation and apply our convergence arguments to the number of units increasing without bound. The point is that the conditions that are needed to establish convergence will apply with respect to the number of observational units. The number of observations taken for each observation unit might be fixed and could be quite small.

This chapter will contain relatively little development of the properties of estimators as was done in Chapter 4. We will rely on earlier results in Chapters 4, 8, and 9 and focus instead on a variety of models and specifications.

11.3 THE POOLED REGRESSION MODEL

We begin the analysis by assuming the simplest version of the model, the **pooled model**,

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, i = 1, \ldots, n, t = 1, \ldots, T_i, \quad \textbf{(11-2)}$$

$$E[\varepsilon_{it}, |\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}] = 0,$$
$$E[\varepsilon_{it}\varepsilon_{js}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}] = \sigma_\varepsilon^2 \text{ if } i = j \text{ and } t = s \text{ and } = 0 \text{ if } i \neq j \text{ or } t \neq s.$$

In this form, if the remaining assumptions of the classical model are met (zero conditional mean of ε_{it}, homoscedasticity, uncorrelatedness across observations, i and strict exogeneity of $\mathbf{x}_{it}$), then no further analysis beyond the results of Chapter 4 is needed. Ordinary least squares is the efficient estimator and inference can reliably proceed along the lines developed in Chapter 5.

11.3.1 LEAST SQUARES ESTIMATION OF THE POOLED MODEL

The crux of the panel data analysis in this chapter is that the assumptions underlying ordinary least squares estimation of the pooled model are unlikely to be met. The question, then, is what can be expected of the estimator when the heterogeneity does differ across individuals? The fixed effects case is obvious. As we will examine later, omitting (or ignoring) the heterogeneity when the fixed effects model is appropriate renders the least squares estimator inconsistent—sometimes wildly so. In the random effects case, in which the true model is

$$y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$

where $E[c_i|\mathbf{X}_i] = \alpha$, we can write the model

$$\begin{aligned} y_{it} &= \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + (c_i - E[c_i|\mathbf{X}_i]) \\ &= \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + u_i \\ &= \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + w_{it}. \end{aligned}$$

In this form, we can see that the unobserved heterogeneity induces autocorrelation; $E[w_{it}w_{is}] = \sigma_u^2$ when $t \neq s$. As we explored in Chapter 9—we will revisit it in Chapter 20—the ordinary least squares estimator in the generalized regression model may be consistent, but the conventional estimator of its asymptotic variance is likely to underestimate the true variance of the estimator.

11.3.2 ROBUST COVARIANCE MATRIX ESTIMATION AND BOOTSTRAPPING

Suppose we consider the model more generally. Stack the T_i observations for individual i in a single equation,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{w}_i,$$

where $\boldsymbol{\beta}$ now includes the constant term. In this setting, there may be heteroscedasticity across individuals. However, in a panel data set, the more substantive effect is cross-observation correlation, or autocorrelation. In a longitudinal data set, the group of observations may all pertain to the same individual, so any latent effects left out of the model will carry across all periods. Suppose, then, we assume that the disturbance vector consists of ε_{it} plus these omitted components. Then,

$$\begin{aligned}\text{Var}[\mathbf{w}_i|\mathbf{X}_i] &= \sigma_\varepsilon^2\mathbf{I}_{T_i} + \boldsymbol{\Sigma}_i \\ &= \boldsymbol{\Omega}_i.\end{aligned}$$

(The subscript i on $\boldsymbol{\Omega}_i$ does not necessarily indicate a different variance for each i. The designation is necessary because the matrix is $T_i \times T_i$.) The ordinary least squares estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= \left[\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{y}_i \\ &= \left[\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{X}_i'(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{w}_i) \\ &= \boldsymbol{\beta} + \left[\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{w}_i.\end{aligned}$$

Consistency can be established along the lines developed in Chapter 4. The true asymptotic covariance matrix would take the form we saw for the generalized regression model in (9-8),

$$\begin{aligned}\text{Asy.Var}[\mathbf{b}] &= \frac{1}{n}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{w}_i\mathbf{w}_i'\mathbf{X}_i\right]\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1} \\ &= \frac{1}{n}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Omega}_i\mathbf{X}_i\right]\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}.\end{aligned}$$

This result provides the counterpart to (9-12). As before, the center matrix must be estimated. In the same fashion as the White estimator, we can estimate this matrix with

$$\text{Est.Asy.Var}[\mathbf{b}] = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\hat{\mathbf{w}}_i\hat{\mathbf{w}}_i'\mathbf{X}_i\right]\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}, \tag{11-3}$$

where $\hat{\mathbf{w}}_i'$ is the vector of T_i residuals for individual i. In fact, the logic of the White estimator *does* carry over to this estimator. Note, however, this is not quite the same as (9-5). It is quite likely that the more important issue for appropriate estimation of the asymptotic covariance matrix is the correlation across observations, not heteroscedasticity. As such, it is likely that the White estimator in (9-5) is not the

solution to the inference problem here. Example 11.4 shows this effect at work. This is the "cluster" robust estimator developed in Section 4.5.3.

Bootstrapping offers another approach to estimating an appropriate covariance matrix for the estimator. We used this approach earlier in a cross-section setting in Example 4.6 where we devised an estimator for the LAD estimator. Here, we will take the group or cluster as the unit of observation. For example, in the data in Example 11.4, there are 595 groups of 7 observations, so the block of 7 observations is the unit of observation. To compute the **block bootstrap** estimator, we use the following procedure. For each of R repetitions, draw random samples of $N = 595$ blocks with replacement. (Each time, some blocks are drawn more than once and others are not drawn.) After the R repetitions, compute the empirical variance of the R replicates. The estimator is

$$\text{Est.Asy.Var}[\mathbf{b}] = \frac{1}{R}\sum_{r=1}^{R}(\mathbf{b}_r - \bar{\mathbf{b}})(\mathbf{b}_r - \bar{\mathbf{b}})'.$$

Example 11.4 Wage Equation

Cornwell and Rupert (1988) analyzed the returns to schooling in a balanced panel of 595 observations on heads of households. The sample data are drawn from years 1976–1982 from the "Non-Survey of Economic Opportunity" from the Panel Study of Income Dynamics. Our estimating equation is a modified version of the one in the paper (without the time fixed effects);

$$\begin{aligned}\ln Wage_{it} = {} & \beta_1 + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} \\ & + \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} \\ & + \beta_{10} Union_{it} + \beta_{11} Ed_i + \beta_{12} Fem_i + \beta_{13} Blk_i + \varepsilon_{it}\end{aligned}$$

where the variables in the model are

Exp = years of full-time work experience,
Wks = weeks worked,
Occ = 1 if the individual has a blue-collar occupation, 0 if not,
Ind = 1 if the individual works in a manufacturing industry, 0 if not,
South = 1 if the individual resides in the south, 0 if not,
SMSA = 1 if the individual resides in an SMSA, 0 if not,
MS = 1 if the individual is married, 0 if not
Union = 1 if the individual's wage is set by a union contract, 0 if not
Ed = years of education
Fem = 1 if the individual is female, 0 if not,
Blk = 1 if the individual is black, 0 if not.

See Appendix Table F8.1 for the data source. Note that *Ed*, *Fem*, and *Blk* are **time invariant**. The main interest of the study, beyond comparing various estimation methods, is β_{11}, the return to education. Table 11.6 reports the least squares estimates based on the full sample of 4,165 observations. [The authors do not report OLS estimates. However, they do report linear least squares estimates of the fixed effects model, which are simple least squares using deviations from individual means. (See Section 11.4.)] The conventional OLS standard errors are given in the second column of results. The third column gives the robust standard errors computed using (11-3). For these data, the computation is

$$\text{Est.Asy.Var}[b] = \left[\sum_{i=1}^{595}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{595}\left(\sum_{t=1}^{7}\mathbf{x}_{it}e_{it}\right)\left(\sum_{t=1}^{7}\mathbf{x}_{it}e_{it}\right)'\right]\left[\sum_{i=1}^{595}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}.$$

TABLE 11.6 Wage Equation Estimated by OLS

Variable	*Least Squares Estimate*	*Standard Error*	*Clustered Std. Error*	*Bootstrapped Std. Error*	*White Hetero. Robust Std. Error*
Constant	5.25112	0.07129	0.12355	0.11171	0.07435
Exp	0.00401	0.00216	0.00408	0.00434	0.00216
ExpSq	−0.00067	0.00005	0.00009	0.00010	0.00005
Wks	0.00422	0.00108	0.00154	0.00164	0.00114
Occ	−0.14001	0.01466	0.02724	0.02555	0.01494
Ind	0.04679	0.01179	0.02366	0.02153	0.01199
South	−0.05564	0.01253	0.02616	0.02414	0.01274
SMSA	0.15167	0.01207	0.02410	0.02323	0.01208
MS	0.04845	0.02057	0.04094	0.03749	0.02049
Union	0.09263	0.01280	0.02367	0.02553	0.01233
Ed	0.05670	0.00261	0.00556	0.00483	0.00273
Fem	−0.36779	0.02510	0.04557	0.04460	0.02310
Blk	−0.16694	0.02204	0.04433	0.05221	0.02075

The robust standard errors are generally about twice the uncorrected ones. In contrast, the White robust standard errors are almost the same as the uncorrected ones. This suggests that for this model, ignoring the within-group correlations does, indeed, substantially affect the inferences one would draw. The block bootstrap standard errors based on 100 replications are shown in the last column. As expected, the block bootstrap results are quite similar to the two-step residual-based method.

11.3.3 CLUSTERING AND STRATIFICATION

Many recent studies have analyzed survey data sets, such as the Current Population Survey (CPS). Survey data are often drawn in clusters, partly to reduce costs. For example, interviewers might visit all the families in a particular block. In other cases, effects that resemble the common random effects in panel data treatments might arise naturally in the sampling setting. Consider, for example, a study of student test scores across several states. Common effects could arise at many levels in such a data set. Education curriculum or funding policies in a state could cause a "state effect"; there could be school district effects, school effects within districts, and even teacher effects within a particular school. Each of these is likely to induce correlation across observations that resembles the random (or fixed) effects we have identified. One might be reluctant to assume that a tightly structured model such as the simple random effects specification is at work. But, as we saw in Example 11.1, ignoring common effects can lead to serious inference errors.

Moulton (1986, 1990) examined the bias of the conventional least squares estimator of Asy.Var[**b**], $s^2(\mathbf{X}'\mathbf{X})^{-1}$. The calculation is complicated because the comparison ultimately depends on the group sizes, the data themselves, and the within-group cross-observation correlation of the common effects. For a simple case,

$$y_{i,g} = \beta_1 + x_{i,g}\beta_2 + u_{i,g} + w_g,$$

a broad, approximate result is the Moulton factor,

$$\frac{\text{Cluster Corrected Variance}}{\text{OLS Uncorrected Variance}} \approx [1 + (n_g - 1)r_x r_u],$$

where n_g is the group size, r_x is the cross-observation correlation (within a group) of $x_{i,g}$ and r_u is the "intraclass correlation," $\sigma_w^2/(\sigma_w^2 + \sigma_u^2)$. The Moulton bias factor suggests that the conventional standard error is biased downward, potentially quite substantially if n_g is large. It is worth noting the Moulton bias might create the impression that the correction of the standard errors *always* increases the standard errors. Algebraically, this is not true—a counterexample appears in Example 4.5. The Moulton result suggests a correction to the OLS standard errors. However, using it would require several approximations of unknown size (based on there being more than one regressor, variable cluster sizes, and needing an estimator for r_u). The robust estimator suggested in Section 11.3.2 will be a preferable approach.

A refinement to (11-3) is sometimes employed to account for small-sample effects when the number of clusters is likely to be a significant proportion of a finite total, such as the number of school districts in a state. A degrees of freedom correction as shown in (11-4) is often employed for this purpose. The robust covariance matrix estimator would be

$$\begin{aligned}\text{Est.Asy.Var}[\mathbf{b}] &= \left[\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right]^{-1}\left[\frac{G}{G-1}\sum_{g=1}^{G}\left(\sum_{i=1}^{n_g}\mathbf{x}_{ig}\hat{w}_{ig}\right)\left(\sum_{i=1}^{n_g}\mathbf{x}_{ig}\hat{w}_{ig}\right)'\right]\left[\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right]^{-1} \\ &= \left[\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right]^{-1}\left[\frac{G}{G-1}\sum_{g=1}^{G}(\mathbf{X}_g'\hat{\mathbf{w}}_g)(\hat{\mathbf{w}}_g'\mathbf{X}_g)\right]\left[\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right]^{-1}, \qquad \textbf{(11-4)}\end{aligned}$$

where G is the number of clusters in the sample and each cluster consists of n_g, $g = 1, \ldots, G$ observations. [Note that this matrix is simply $G/(G-1)$ times the matrix in (11-3).] A further correction (without obvious formal motivation) sometimes employed is a degrees of freedom correction, $[(\Sigma_g n_g) - 1]/[(\Sigma_g n_g) - K]$.

Many further refinements for more complex samples—consider the test scores example—have been suggested. For a detailed analysis, see Cameron and Trivedi (2005, Chapter 24) and Cameron and Miller (2015). Several aspects of the computation are discussed in Wooldridge (2010, Chapter 20) as well. An important question arises concerning the use of asymptotic distributional results in cases in which the number of clusters might be relatively small. Angrist and Lavy (2002) find that the clustering correction after pooled OLS, as we have done in Example 11.3, is not as helpful as might be hoped for (though our correction with 595 clusters each of size 7 would be "safe" by these standards). But, the difficulty might arise, at least in part, from the use of OLS in the presence of the common effects. Kezde (2001) and Bertrand, Dufflo, and Mullainathan (2002) find more encouraging results when the correction is applied after estimation of the fixed effects regression. Yet another complication arises when the groups are very large and the number of groups is relatively small, for example, when the panel consists of many large samples from a subset (or even all) of the U.S. states. Since the asymptotic theory we have used to this point assumes the opposite, the results will be less reliable in this case. Donald and Lang (2007) find that this case gravitates toward analysis of group means rather than the individual data. Wooldridge (2003) provides results that help explain this finding. Finally, there is a natural question as to whether the correction

is even called for if one has used a random effects, generalized least squares procedure (see Section 11.5) to do the estimation at the first step. If the data-generating mechanism were strictly consistent with the random effects model, the answer would clearly be negative. Under the view that the random effects specification is only an approximation to the correlation across observations in a cluster, then there would remain residual correlation that would be accommodated by the correction in (11-4) (or some GLS counterpart). (This would call the specific random effects correction in Section 11.5 into question, however.) A similar argument would motivate the correction after fitting the fixed effects model as well. We will pursue these possibilities in Section 11.6.4 after we develop the fixed and random effects estimator in detail.

11.3.4 ROBUST ESTIMATION USING GROUP MEANS

The pooled regression model can also be estimated using the sample means of the data. The implied regression model is obtained by premultiplying each group by $(1/T)\mathbf{i}'$ where $\mathbf{i}'$ is a row vector of ones,

$$(1/T)\mathbf{i}'\mathbf{y}_i = (1/T)\mathbf{i}'\mathbf{X}_i\boldsymbol{\beta} + (1/T)\mathbf{i}'w_i$$

or

$$\bar{y}_{i.} = \bar{\mathbf{x}}_i'\boldsymbol{\beta} + \bar{w}_i.$$

In the transformed linear regression, the disturbances continue to have zero conditional means but heteroscedastic variances $\sigma_i^2 = (1/T^2)\mathbf{i}'\boldsymbol{\Omega}_i\mathbf{i}$. With $\boldsymbol{\Omega}_i$ unspecified, this is a heteroscedastic regression for which we would use the White estimator for appropriate inference. Why might one want to use this estimator when the full data set is available? If the classical assumptions are met, then it is straightforward to show that the asymptotic covariance matrix for the group means estimator is unambiguously larger, and the answer would be that there is no benefit. But failure of the classical assumptions is what brought us to this point, and then the issue is less clear-cut. In the presence of unstructured cluster effects the efficiency of least squares can be considerably diminished, as we saw in the preceding example. The loss of information that occurs through the averaging might be relatively small, though in principle the disaggregated data should still be better.

We emphasize that using **group means** does not solve the problem that is addressed by the fixed effects estimator. Consider the general model,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + c_i\boldsymbol{i} + \mathbf{w}_i,$$

where as before, c_i is the latent effect. If the mean independence assumption, $E[c_i|\mathbf{X}_i] = \alpha$, is not met, then the effect will be transmitted to the group means as well. In this case, $E[c_i|\mathbf{X}_i] = h(\mathbf{X}_i)$. A common specification is Mundlak's (1978), where we employ the projection of c_i on the group means (see Section 4.4.5),

$$c_i|\mathbf{X}_i = \bar{\mathbf{x}}_i'.\boldsymbol{\gamma} + v_i.$$

Then,

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + \bar{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + [\varepsilon_{it} + v_i] \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + \bar{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + u_{it}, \end{aligned}$$

where, by construction, $Cov[u_{it}, \bar{\mathbf{x}}_i] = 0$. Taking means as before,

$$\begin{aligned}\bar{y}_{i.} &= \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \bar{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + \bar{u}_{i.} \\ &= \bar{\mathbf{x}}_{i.}'(\boldsymbol{\beta} + \boldsymbol{\gamma}) + \bar{u}_{i.}.\end{aligned}$$

The implication is that the group means estimator estimates not $\boldsymbol{\beta}$, but $\boldsymbol{\beta} + \boldsymbol{\gamma}$. Averaging the observations in the group collects the entire set of effects, observed and latent, in the group means.

One consideration that remains, which, unfortunately, we cannot resolve analytically, is the possibility of measurement error. If the regressors are measured with error, then, as we examined in Section 8.7, the least squares estimator is inconsistent and, as a consequence, efficiency is a moot point. In the panel data setting, if the measurement error is random, then using group means would work in the direction of averaging it out—indeed, in this instance, assuming the benchmark case $\mathbf{x}_{itk} = \mathbf{x}_{itk}^{*} + u_{itk}$, one could show that the group means estimator would be consistent as $T \to \infty$ while the OLS estimator would not.

Example 11.5 Robust Estimators of the Wage Equation

Table 11.7 shows the group means estimates of the wage equation shown in Example 11.4 with the original least squares estimates. In both cases, a robust estimator is used for the covariance matrix of the estimator. It appears that similar results are obtained with the means.

11.3.5 ESTIMATION WITH FIRST DIFFERENCES

First differencing is another approach to estimation. Here, the intent would explicitly be to transform latent heterogeneity out of the model. The base case would be

$$y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$

TABLE 11.7 Wage Equation Estimated by OLS

Coefficient	*OLS Estimated Coefficient*	*Cluster Robust Standard Error*	*Group Means Estimates*	*White Robust Standard Error*
Constant	5.25112	0.12330	5.12143	0.20425
Exp	0.04010	0.00408	0.03190	0.00478
Exp^2	−0.00067	0.00009	−0.00057	0.00010
Wks	0.00422	0.00154	0.00919	0.00360
Occ	−0.14001	0.02724	−0.16762	0.03382
Ind	0.04679	0.02366	0.05792	0.02554
South	−0.05564	0.02616	−0.05705	0.02597
SMSA	0.15167	0.02410	0.17578	0.02576
MS	0.04845	0.04094	0.11478	0.04770
Union	0.09263	0.02367	0.10907	0.02923
Ed	0.05670	0.00556	0.05144	0.00555
Fem	−0.36779	0.04557	−0.31706	0.05473
Blk	−0.16694	0.04433	−0.15780	0.04501

which implies the first differences equation,

$$\Delta y_{it} = \Delta c_i + (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \Delta \varepsilon_{it},$$

or

$$\begin{aligned}\Delta y_{it} &= (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \varepsilon_{it} - \varepsilon_{i,t-1} \\ &= (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + u_{it}.\end{aligned}$$

The advantage of the **first difference** approach is that it removes the latent heterogeneity from the model whether the fixed or random effects model is appropriate. The disadvantage is that the differencing also removes any time-invariant variables from the model. In our example, we had three, *Ed*, *Fem*, and *Blk*. If the time-invariant variables in the model are of no interest, then this is a robust approach that can estimate the parameters of the time-varying variables consistently. Of course, this is not helpful for the application in the example because the impact of *Ed* on ln *Wage* was the primary object of the analysis. Note, as well, that the differencing procedure trades the cross-observation correlation in c_i for a moving average (MA) disturbance, $u_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$.[5] The new disturbance, $u_{i,t}$, is autocorrelated, though across only one period. Nonetheless, in order to proceed, it would have to be true that $\Delta \mathbf{x}_t$ is uncorrelated with $\Delta \varepsilon_t$. Strict exogeneity of $\mathbf{x}_{it}$ is sufficient, but in the absence of that assumption, such as if only $\text{Cov}(\varepsilon_{it}, \mathbf{x}_{it}) = 0$ has been assumed, then it is conceivable that $\Delta \mathbf{x}_t$ and $\Delta \varepsilon_t$ could be correlated. The presence of a lagged value of y_{it} in the original equation would be such a case. Procedures are available for using two-step feasible GLS for an MA disturbance (see Chapter 20). Alternatively, this model is a natural candidate for OLS with the Newey–West robust covariance estimator because the right number of lags (one) is known. (See Section 20.5.2.)

As a general observation, with a variety of approaches available, the first difference estimator does not have much to recommend it, save for one very important application. Many studies involve two period panels, a before and an after treatment. In these cases, as often as not, the phenomenon of interest may well specifically be the change in the outcome variable—the "treatment effect." Consider the model

$$y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \theta S_{it} + \varepsilon_{it},$$

where $t = 1, 2$ and $S_{it} = 0$ in period 1 and 1 in period 2; S_{it} indicates a treatment that takes place between the two observations. The treatment effect would be

$$E[\Delta y_i | (\Delta \mathbf{x}_i = 0)] = \theta,$$

which is precisely the constant term in the first difference regression,

$$\Delta y_i = \theta + (\Delta \mathbf{x}_i)'\boldsymbol{\beta} + u_i.$$

We examined cases like these in detail in Section 6.3.

11.3.6 THE WITHIN- AND BETWEEN-GROUPS ESTIMATORS

The pooled regression model is

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}. \qquad \textbf{(11-5a)}$$

[5] If the original disturbance, ε_{it} were a random walk, $\varepsilon_{i,t} = \varepsilon_{i,t-1} + u_{it}$, then the disturbance in the first differenced equation would be homoscedastic and nonautocorrelated. This would be a narrow assumption that might apply in a particular situation. This would not seem to be a natural specification for the model in Example 11.4, for example.

In terms of the group means,

$$\bar{y}_{i.} = \alpha + \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \bar{\varepsilon}_{i.}, \tag{11-5b}$$

while in terms of deviations from the group means,

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_{i.}.$$

For convenience later, write this as

$$\ddot{y} = \ddot{\mathbf{x}}_{it}'\boldsymbol{\beta} + \ddot{\varepsilon}_{it}. \tag{11-5c}$$

[We are assuming there are no time-invariant variables in $\mathbf{x}_{it}$, such as *Ed* in Example 11.4. These would become all zeros in (11-5c).] All three are classical regression models, and in principle, all three could be estimated, at least consistently if not efficiently, by ordinary least squares. [Note that (11-5b) defines only n observations, the group means.] Consider then the matrices of sums of squares and cross products that would be used in each case, where we focus only on estimation of $\boldsymbol{\beta}$. In (11-5a), the moments would accumulate variation about the overall means, $\bar{\bar{y}}$ and $\bar{\bar{\mathbf{x}}}$, and we would use the total sums of squares and cross products,

$$\mathbf{S}_{xx}^{total} = \sum_{i=\mathbf{1}}^{n}\sum_{t=\mathbf{1}}^{T}(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})' \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \sum_{i=\mathbf{1}}^{n}\sum_{t=\mathbf{1}}^{T}(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})(y_{it} - \bar{\bar{y}}). \tag{11-6}$$

For (11-5c), because the data are in deviations already, the means of $(y_{it} - \bar{y}_{i.})$ and $(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})$ are zero. The moment matrices are **within-groups** (i.e., variation around group means) sums of squares and cross products,

$$\mathbf{S}_{xx}^{within} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})' \quad \text{and} \quad \mathbf{S}_{xy}^{within} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(y_{it} - \bar{y}_{i.}).$$

Finally, for (11-5b), the mean of group means is the overall mean. The moment matrices are the **between-groups** sums of squares and cross products—that is, the variation of the group means around the overall means,

$$\mathbf{S}_{xx}^{between} = \sum_{i=1}^{n} T(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})' \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \sum_{i=1}^{n} T(\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})(\bar{y}_{i.} - \bar{\bar{y}}).$$

It is easy to verify that

$$\mathbf{S}_{xx}^{total} = \mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between} \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}.$$

Therefore, there are three possible least squares estimators of $\boldsymbol{\beta}$ corresponding to the decomposition. The least squares estimator is

$$\mathbf{b}^{total} = \left[\mathbf{S}_{xx}^{total}\right]^{-1}\mathbf{S}_{xy}^{total} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1}\left[\mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}\right]. \tag{11-7}$$

The within-groups estimator is

$$\mathbf{b}^{within} = \left[\mathbf{S}_{xx}^{within}\right]^{-1}\mathbf{S}_{xy}^{within}. \tag{11-8}$$

This is the dummy variable estimator developed in Section 11.4. An alternative estimator would be the between-groups estimator,

$$\mathbf{b}^{between} = \left[\mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xy}^{between}. \tag{11-9}$$

This is the **group means estimator**. This least squares estimator of (11-5b) is based on the n sets of groups means. (Note that we are assuming that n is at least as large as K.) From the preceding expressions (and familiar previous results),

$$\mathbf{S}_{xy}^{within} = \mathbf{S}_{xx}^{within}\mathbf{b}^{within} \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \mathbf{S}_{xx}^{between}\mathbf{b}^{between}.$$

Inserting these in (11-7), we see that the least squares estimator is a **matrix weighted average** of the within- and between-groups estimators:

$$\mathbf{b}^{total} = \mathbf{F}^{within}\mathbf{b}^{within} + \mathbf{F}^{between}\mathbf{b}^{between}, \tag{11-10}$$

where

$$\mathbf{F}^{within} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xx}^{within} = \mathbf{I} - \mathbf{F}^{between}.$$

The form of this result resembles the Bayesian estimator in the classical model discussed in Chapter 16. The resemblance is more than passing; it can be shown[6] that

$$\mathbf{F}^{within} = \{[\text{Asy.Var}(\mathbf{b}^{within})]^{-1} + [\text{Asy.Var}(\mathbf{b}^{between})]^{-1}\}^{-1}[\text{Asy.Var}(\mathbf{b}^{within})]^{-1},$$

which is essentially the same mixing result we have for the Bayesian estimator. In the weighted average, the estimator with the smaller variance receives the greater weight.

Example 11.6 Analysis of Covariance and the World Health Organization (WHO) Data

The decomposition of the total variation in Section 11.3.6 extends to the linear regression model the familiar *analysis of variance*, or ANOVA, that is often used to decompose the variation in a variable in a clustered or stratified sample, or in a panel data set. One of the useful features of panel data analysis as we are doing here is the ability to analyze the between-groups variation (heterogeneity) to learn about the main regression relationships and the within-groups variation to learn about dynamic effects.

The WHO data used in Example 6.22 is an unbalanced panel data set—we used only one year of the data in Example 6.22. Of the 191 countries in the sample, 140 are observed in the full five years, one is observed four times, and 50 are observed only once. The original WHO studies (2000a, 2000b) analyzed these data using the fixed effects model developed in the next section. The estimator is that in (11-8). It is easy to see that groups with one observation will fall out of the computation, because if $T_i = 1$, then the observation equals the group mean. These data have been used by many researchers in similar panel data analyses.[7] Gravelle et al. (2002a) have strongly criticized these analyses, arguing that the WHO data are much more like a cross section than a panel data set.

From Example 6.22, the model used by the researchers at WHO was

$$\ln DALE_{it} = \alpha_i + \beta_1 \ln \textit{Health Expenditure}_{it} + \beta_2 \ln \textit{Education}_{it} + \beta_3 \ln^2 \textit{Education}_{it} + \varepsilon_{it}.$$

Additional models were estimated using WHO's composite measure of health care attainment, *COMP*. The analysis of variance for a variable x_{it} is based on the decomposition

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}(x_{it} - \bar{\bar{x}})^2 = \sum_{i=1}^{n}\sum_{t=1}^{T_i}(x_{it} - \bar{x}_{i.})^2 + \sum_{t=1}^{n}T_i(\bar{x}_{i.} - \bar{\bar{x}})^2.$$

[6]See, for example, Judge et al. (1985).

[7]See, e.g., Greene (2004c) and several references.

TABLE 11.8 Analysis of Variance for WHO Data on Health Care Attainment

Variable	*Within-Groups Variation (%)*	*Between-Groups Variation (%)*
COMP	5.635	94.635
DALE	0.150	99.850
Expenditure	0.635	99.365
Education	0.177	99.823

Dividing both sides of the equation by the left-hand side produces the decomposition

$$1 = \textit{Within-groups proportion} + \textit{Between-groups proportion}.$$

The first term on the right-hand side is the within-group variation that differentiates a panel data set from a cross section (or simply multiple observations on the same variable). Table 11.8 lists the decomposition of the variation in the variables used in the WHO studies.

The results suggest the reasons for the authors' concern about the data. For all but COMP, virtually all the variation in the data is between groups—that is cross-sectional variation. As the authors argue, these data are only slightly different from a cross section.

11.4 THE FIXED EFFECTS MODEL

The fixed effects model arises from the assumption that the omitted effects, c_i, in the regression model of (11-1),

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, i = 1, \ldots, n, t = 1, \ldots, T_i, \\ E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}] &= 0, \\ E[\varepsilon_{it}\varepsilon_{js}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}] &= \sigma_\varepsilon^2 \text{ if } i = j \text{ and } t = s \text{ and } = 0 \text{ if } i \neq j \text{ or } t \neq s, \end{aligned} \tag{11-11}$$

can be arbitrarily correlated with the included variables. In a generic form,

$$E[c_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}] = E[c_i|\mathbf{X}_i] = h(\mathbf{X}_i). \tag{11-12}$$

We also assume that $\text{Var}[c_i|\mathbf{X}_i]$ is constant and all observations c_i and c_j are independent. We emphasize it is (11-12) that signifies the fixed effects model, not that any variable is fixed in this context and random elsewhere. The formulation implies that the heterogeneity across groups is captured in the constant term.[8] In (11-1), $\mathbf{z}_i = (1)$ and

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

Each α_i can be treated as an unknown parameter to be estimated.

11.4.1 LEAST SQUARES ESTIMATION

Let $\mathbf{y}_i$ and $\mathbf{X}_i$ be the T observations for the ith unit, let $\mathbf{i}$ be a $T \times 1$ column of ones, and let $\boldsymbol{\varepsilon}_i$ be the associated $T \times 1$ vector of disturbances.[9] Then,

$$\mathbf{y}_i = \mathbf{i}\alpha_i + \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i.$$

[8]It is also possible to allow the slopes to vary across i. A study on the topic is Cornwell and Schmidt (1984). We will examine this case in Section 11.4.6.

[9]The assumption of a fixed group size, T, at this point is purely for convenience. As noted in Section 11.2.4, the unbalanced case is a minor variation.

Collecting these terms gives

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_n\end{bmatrix} = \begin{bmatrix}\mathbf{i} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{i} & \cdots & \mathbf{0}\\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{i}\end{bmatrix}\begin{pmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_n\end{pmatrix} + \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_n\end{bmatrix}\begin{pmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_K\end{pmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\\ \vdots\\ \boldsymbol{\varepsilon}_n\end{bmatrix}$$

or

$$\mathbf{y} = [\mathbf{d}_1 \quad \mathbf{d}_2, \ \ldots, \mathbf{d}_n \quad \mathbf{X}]\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix} + \boldsymbol{\varepsilon},$$

where $\mathbf{d}_i$ is a dummy variable indicating the ith unit. Let the $nT \times n$ matrix $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ \ldots, \mathbf{d}_n]$. Then, assembling all nT rows gives

$$\mathbf{y} = \mathbf{D}\,\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{11-13}$$

This model is occasionally referred to as the **least squares dummy variable (LSDV) model** (although the "least squares" part of the name refers to the technique usually used to estimate it, not to the model itself).

This model is a classical regression model, so no new results are needed to analyze it. If n is small enough, then the model can be estimated by ordinary least squares with K regressors in $\mathbf{X}$ and n columns in $\mathbf{D}$, as a multiple regression with $K + n$ parameters. Of course, if n is thousands, as is typical, then treating (11-13) as an ordinary regression will be extremely cumbersome. But, by using familiar results for a partitioned regression, we can reduce the size of the computation.[10] We write the least squares estimator of $\boldsymbol{\beta}$ as

$$\mathbf{b}_{\text{LSDV}} = [\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{y}] = \mathbf{b}^{within}, \tag{11-14}$$

where

$$\mathbf{M}_\mathbf{D} = \mathbf{I}_{nT} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'.$$

Because $\mathbf{M}_\mathbf{D}$ is symmetric and idempotent, $\mathbf{b}_{\text{LSDV}} = [(\mathbf{X}'\mathbf{M}_\mathbf{D})(\mathbf{M}_\mathbf{D}\mathbf{X})]^{-1}[(\mathbf{X}'\mathbf{M}_\mathbf{D})(\mathbf{M}_\mathbf{D}\mathbf{y})]$. This amounts to a least squares regression using the transformed data $\mathbf{M}_\mathbf{D}\mathbf{X} = \ddot{\mathbf{X}}$ and $\mathbf{M}_\mathbf{D}\mathbf{y} = \ddot{\mathbf{y}}$. The structure of $\mathbf{D}$ is particularly convenient; its columns are orthogonal, so

$$\mathbf{M}_\mathbf{D} = \begin{bmatrix}\mathbf{M}^0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{M}^0 & \mathbf{0} & \cdots & \mathbf{0}\\ & & \cdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M}^0\end{bmatrix}.$$

Each matrix on the diagonal is

$$\mathbf{M}^0 = \mathbf{I}_T - \frac{1}{T}\mathbf{i}\mathbf{i}'. \tag{11-15}$$

Premultiplying any $T \times 1$ *vector* $\mathbf{z}_i$ by $\mathbf{M}^0$ creates $\mathbf{M}^0\mathbf{z}_i = \mathbf{z}_i - \bar{z}\mathbf{i}$. (Note that the mean is taken over only the T observations for unit i.) Therefore, the least squares regression of $\mathbf{M}_\mathbf{D}\mathbf{y} = \ddot{\mathbf{y}}$ on $\mathbf{M}_\mathbf{D}\mathbf{X} = \ddot{\mathbf{X}}$ is equivalent to a regression of $[y_{it} - \bar{y}_{i.}] = \ddot{y}_{it}$ on

[10] See Theorem 3.2.

$[\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}] = \ddot{\mathbf{x}}_{it}$, where $\bar{y}_{i.}$ and $\bar{\mathbf{x}}_{i.}$ are the scalar and $K \times 1$ vector of means of y_{it} and $\mathbf{x}_{it}$ over the T observations for group i.[11]

In terms of the within transformed data, then,

$$\begin{aligned}\mathbf{b}_{LSDV} &= \left[\sum_{i=1}^{n}(\mathbf{M}^0\mathbf{X}_i)'(\mathbf{M}^0\mathbf{X}_i)\right]^{-1}\left[\sum_{i=1}^{n}(\mathbf{M}^0\mathbf{X}_i)'(\mathbf{M}^0\mathbf{y}_i)\right] \\ &= \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i\right] \\ &= (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\ddot{\mathbf{y}}.\end{aligned} \tag{11-16a}$$

The dummy variable coefficients can be recovered from the other normal equation in the partitioned regression,

$$\mathbf{D}'\mathbf{D}\mathbf{a} + \mathbf{D}'\mathbf{X}\mathbf{b}_{\text{LSDV}} = \mathbf{D}'\mathbf{y}$$

or

$$\mathbf{a} = [\mathbf{D}'\mathbf{D}]^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{X}\mathbf{b}_{\text{LSDV}}).$$

This implies that for each i,

$$a_i = \bar{y}_{i.} - \mathbf{x}_{i.}'\mathbf{b}_{\text{LSDV}}. \tag{11-16b}$$

The appropriate estimator of the asymptotic covariance matrix for $\mathbf{b}$ is

$$\text{Est.Asy.Var.}[\mathbf{b}_{\text{LSDV}}] = s^2\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}. \tag{11-17}$$

Based on (11-14) and (11-16), the disturbance variance estimator is

$$\begin{aligned}s^2 = \frac{(\ddot{\mathbf{y}} - \ddot{\mathbf{X}}\mathbf{b})'(\ddot{\mathbf{y}} - \ddot{\mathbf{X}}\mathbf{b}_{\text{LSDV}})}{nT - n - K} &= \frac{\sum_{i=1}^{n}(\ddot{\mathbf{y}}_i - \ddot{\mathbf{X}}_i\mathbf{b}_{\text{LSDV}})'(\ddot{\mathbf{y}}_i - \ddot{\mathbf{X}}_i\mathbf{b}_{\text{LSDV}})}{nT - n - K} \\ &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b}_{\text{LSDV}} - a_i)^2}{nT - n - K}.\end{aligned} \tag{11-18}$$

The itth residual used in this computation is

$$\begin{aligned}e_{it} = y_{it} - \mathbf{x}_{it}'\mathbf{b}_{\text{LSDV}} - a_i &= y_{it} - \mathbf{x}_{it}'\mathbf{b}_{\text{LSDV}} - (\bar{y}_{i.} - \bar{\mathbf{x}}_{i.}'\mathbf{b}_{\text{LSDV}}) \\ &= (y_{it} - \bar{y}_{i.}) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\,\mathbf{b}_{\text{LSDV}}.\end{aligned}$$

Thus, the numerator in s^2 is exactly the sum of squared residuals using the least squares slopes and the data in group mean deviation form. But, done in this fashion, one might then use $nT - K$ instead of $nT - n - K$ for the denominator in computing s^2, so a correction would be necessary.[12] For the individual effects,

[11] An interesting special case arises if $T = 2$. In the two-period case, you can show—we leave it as an exercise—that this least squares regression is done with nT first difference observations, by regressing observation $(y_{i2} - y_{i1})$ (and its negative) on $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$ (and its negative).

[12] The maximum likelihood estimator of σ^2 for the fixed effects model with normally distributed disturbances is $\Sigma_i\Sigma_t e_{it}^2/nT$, with no degrees of freedom correction. This is a case in which the MLE is biased, given (11-18) which gives the unbiased estimator. This bias in the MLE for a fixed effects model is an example (actually, the first example) of the incidental parameters problem. [See Neyman and Scott (1948) and Lancaster (2000).] With a bit of manipulation it is clear that although the estimator is biased, if T increases asymptotically, then the bias eventually diminishes to zero. This is the signature feature of estimators that are affected by the incidental parameters problem.

$$\text{Asy.Var}[a_i] = \frac{\sigma_\varepsilon^2}{T} + \bar{\mathbf{x}}_{i.}'\{\text{Asy.Var}[\mathbf{b}]\}\bar{\mathbf{x}}_{i.}, \tag{11-19}$$

so a simple estimator based on s^2 can be computed.

With increasing n, the asymptotic variance of a_i declines to a lower bound of σ_ε^2/T *which does not converge to zero*. The constant term estimators in the fixed effects model are not consistent estimators of α_i. They are not inconsistent because they gravitate toward the wrong parameter. They are so because their asymptotic variances do not converge to zero, even as the sample size grows. It is easy to see why this is the case. We see that each a_i is estimated using only T observations—assume n were infinite, so that $\boldsymbol{\beta}$ were known. Because T is not assumed to be increasing, we have the surprising result. The constant terms are inconsistent unless $T \to \infty$, which is not part of the model.

We note a major shortcoming of the fixed effects approach. Any **time-invariant** variables in $\mathbf{x}_{it}$ will mimic the individual specific constant term. Consider the application of Example 11.3. We could write the fixed effects formulation as

$$\ln Wage_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + [\beta_{10}Ed_i + \beta_{11}Fem_i + \beta_{12}Blk_i + c_i] + \varepsilon_{it}.$$

The fixed effects formulation of the model will absorb the last four terms in the regression in α_i. The coefficients on the time-invariant variables cannot be estimated. For any $\mathbf{x}_k$ that is time invariant, every observation is the group mean, so $\mathbf{M_D}\mathbf{x}_k = \ddot{\mathbf{x}}_k = 0$ so the corresponding column of $\ddot{\mathbf{X}}$ becomes a column of zeros and $(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}$ will not exist.

11.4.2 A ROBUST COVARIANCE MATRIX FOR $\mathbf{b}_{\text{LSDV}}$

The LSDV estimator is computed as

$$\mathbf{b}_{\text{LSDV}} = \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i\right] = \boldsymbol{\beta} + \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]. \tag{11-20}$$

The asymptotic covariance matrix for the estimator derives from

$$\text{Var}[(\mathbf{b}_{\text{LSDV}} - \boldsymbol{\beta})|\mathbf{X}] = \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1} E\left\{\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]'|\mathbf{X}\right\}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}.$$

The center matrix is a double sum over $i,j = 1, \ldots, n$, but terms with $i \neq j$ are independent and have expectation zero, so the matrix is

$$E\left\{\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right]'|\mathbf{X}\right\} = E\left\{\left[\sum_{i=1}^{n}(\ddot{\mathbf{X}}_i\ddot{\boldsymbol{\varepsilon}}_i)(\ddot{\boldsymbol{\varepsilon}}_i'\ddot{\mathbf{X}}_i)\right]|\mathbf{X}\right\}.$$

Each term in the sum is $(\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i)(\ddot{\boldsymbol{\varepsilon}}_i'\ddot{\mathbf{X}}_i) = (\mathbf{X}_i'\mathbf{M}^0\mathbf{M}^0\boldsymbol{\varepsilon}_i)(\boldsymbol{\varepsilon}_i'\mathbf{M}^0\mathbf{M}^0\mathbf{X}_i')$. But $\mathbf{M}^0$ is idempotent, so $\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\varepsilon}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\varepsilon}_i$, and we have assumed that $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}] = \sigma_\varepsilon{}^2\mathbf{I}$. Collecting the terms,

$$\begin{aligned}
\text{Var}[(\mathbf{b}_{\text{LSDV}} - \boldsymbol{\beta})|\mathbf{X}] &= \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1} E\left\{\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\ddot{\mathbf{X}}_i\right]|\mathbf{X}\right\}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1} \\
&= \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left\{\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'(\sigma_\varepsilon^2\mathbf{I})\ddot{\mathbf{X}}_i\right]\right\}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1} \\
&= \sigma_\varepsilon^2\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1},
\end{aligned}$$

which produces the estimator in (11-17). If the disturbances in (11-11) are heteroscedastic and/or autocorrelated, then $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}] \neq \sigma_\varepsilon{}^2\mathbf{I}$. A robust counterpart to (11-4) would be

$$\text{Est.Asy.Var}[\mathbf{b}_{\text{LSDV}}] = \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left\{\left[\sum_{i=1}^{n}(\ddot{\mathbf{X}}_i'\mathbf{e}_i)(\mathbf{e}_i'\ddot{\mathbf{X}}_i)\right]\right\}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}, \quad \textbf{(11-21)}$$

where e_{it} is the residual shown after (11-18). Note that using $\ddot{e}_{it}$ in this calculation gives exactly the same result because $\bar{e}_{i.} = 0$. [13]

11.4.3 TESTING THE SIGNIFICANCE OF THE GROUP EFFECTS

The t ratio for a_i can be used for a test of the hypothesis that α_i equals zero. This hypothesis about one specific group, however, is typically not useful for testing in this regression context. If we are interested in differences across groups, then we can test the hypothesis that the constant terms are all equal with an F test. Under the null hypothesis of equality, the efficient estimator is pooled least squares. The F ratio used for this test is

$$F(n-1, nT-n-K) = \frac{(R^2_{LSDV} - R^2_{Pooled})/(n-1)}{(1-R^2_{LSDV})/(nT-n-K)},$$

where *LSDV* indicates the dummy variable model and *Pooled* indicates the pooled or restricted model with only a single overall constant term. Alternatively, the model may have been estimated with an overall constant and $n-1$ dummy variables instead. All other results (i.e., the least squares slopes, s^2, R^2) will be unchanged, but rather than estimate α_i, each dummy variable coefficient will now be an estimate of $\alpha_i - \alpha_1$ where group "1" is the omitted group. The F test that the coefficients on these $n-1$ dummy variables are zero is identical to the one above. It is important to keep in mind, however, that although the statistical results are the same, the interpretation of the dummy variable coefficients in the two formulations is different.[14]

Example 11.7 Fixed Effects Estimates of a Wage Equation

We continue Example 11.4 by computing the fixed effects estimates of the wage equation, now

$$\begin{aligned}\ln Wage_{it} = {} & \alpha_i + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} \\ & + \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} \\ & + \beta_{10} Union_{it} + 0 \times Ed_i + 0 \times Fem_i + 0 \times Blk_i + \varepsilon_{it}.\end{aligned}$$

Because *Ed, Fem*, and *Blk* are time invariant, their coefficients will not be estimable, and will be set to zero. The OLS and fixed effects estimates are presented in Table 11.9. Each is accompanied by the conventional standard errors and the robust standard errors. We note, first, the rather large change in the parameters that occurs when the fixed effects specification is used. Even some statistically significant coefficients in the least squares results change sign in the fixed effects results. Likewise, the robust standard errors are characteristically much larger than the conventional counterparts. The fixed effects standard errors increased

[13] See Arellano (1987) and Arellano and Bover (1995).

[14] The F statistic can also be based on the sum of squared residuals rather than the R^2s, [See (5-29) and (5-30).] In this connection, we note that the software package Stata contains two estimators for the fixed effects linear regression, `areg` and `xtreg`. In computing the former, Stata uses $\Sigma_i\Sigma_t(y_{it} - \bar{\bar{y}})^2$ as the denominator, as it would in computing the counterpart for the constrained regression. But `xtreg` (which is the procedure typically used) uses $\Sigma_i\Sigma_t(y_{it} - \bar{y}_i)^2$, which is smaller. The R^2 produced by `xtreg` will be smaller, as will be the F statistic, possibly substantially so.

TABLE 11.9 Wage Equation Estimated by OLS and LSDV

	Pooled OLS			*Fixed Effects LSDV*		
Variable	*Least Squares Estimate*	*Standard Error*	*Clustered Std. Error*	*Fixed Effects Estimates*	*Standard Error*	*Robust Std. Error*
R^2	0.42861			0.90724		
Constant	5.25112	0.07129	0.12355	—	—	—
Exp	0.00401	0.00216	0.00408	0.11321	0.00247	0.00438
ExpSq	−0.00067	0.00005	0.00009	−0.00042	0.00006	0.00009
Wks	0.00422	0.00108	0.00154	0.00084	0.00060	0.00094
Occ	−0.14001	0.01466	0.02724	−0.02148	0.01379	0.02053
Ind	0.04679	0.01179	0.02366	0.01921	0.01545	0.02451
South	−0.05564	0.01253	0.02616	−0.00186	0.03431	0.09650
SMSA	0.15167	0.01207	0.02410	−0.04247	0.01944	0.03186
MS	0.04845	0.02057	0.04094	−0.02973	0.01899	0.02904
Union	0.09263	0.01280	0.02367	0.03278	0.01493	0.02709
Ed	0.05670	0.00261	0.00556	—	—	—
Fem	−0.36779	0.02510	0.04557	—	—	—
Blk	−0.16694	0.02204	0.04433	—	—	—

more than might have been expected, given that heteroscedasticity is not a major issue, but a source of autocorrelation is in the equation (as the fixed effects). The large changes suggest that there may yet be some additional, unstructured correlation remaining in ε_{it}. The test for the presence of the fixed effects is based on

$$F = [(0.90724 - 0.42861)/594]/[(1 - 0.90724)/(4165 - 595 - 9)] = 30.933.$$

The critical value from the F table would be less than 1.3, so the hypothesis of homogeneity is rejected.

11.4.4 FIXED TIME AND GROUP EFFECTS

The least squares dummy variable approach can be extended to include a time-specific effect as well. One way to formulate the extended model is simply to add the time effect, as in

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \delta_t + \varepsilon_{it}. \quad \textbf{(11-22)}$$

This model is obtained from the preceding one by the inclusion of an additional $T - 1$ dummy variables. (One of the time effects must be dropped to avoid perfect collinearity—the group effects and time effects both sum to one.) If the number of variables is too large to handle by ordinary regression, then this model can also be estimated by using the partitioned regression. There is an asymmetry in this formulation, however, because each of the group effects is a group-specific intercept, whereas the time effects are **contrasts**—that is, comparisons to a base period (the one that is excluded). A symmetric form of the model is

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mu + \alpha_i + \delta_t + \varepsilon_{it}, \quad \textbf{(11-23)}$$

where a full n and T effects are included, but the restrictions

$$\sum_i \alpha_i = \sum_t \delta_t = 0$$

are imposed. Least squares estimates of the slopes in this model are obtained by regression of

$$y_{*it} = y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{\bar{y}}$$

on

$$\mathbf{x}_{*it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{.t} + \bar{\bar{\mathbf{x}}}, \tag{11-24}$$

where the period-specific and overall means are

$$\bar{y}_{.t} = \frac{1}{n}\sum_{i=1}^{n} y_{it} \qquad \text{and} \qquad \bar{\bar{y}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it},$$

and likewise for $\bar{\mathbf{x}}_{.t}$ and $\bar{\bar{\mathbf{x}}}$. The overall constant and the dummy variable coefficients can then be recovered from the normal equations as

$$\begin{aligned}
\hat{\mu} &= m = \bar{\bar{y}} - \bar{\bar{\mathbf{x}}}'\mathbf{b}, \\
\hat{\alpha}_i &= a_i = (\bar{y}_{i.} - \bar{\bar{y}}) - (\bar{\mathbf{x}}_{i.} - \bar{\bar{\mathbf{x}}})'\mathbf{b}, \\
\hat{\delta}_t &= d_t = (\bar{y}_{.t} - \bar{\bar{y}}) - (\bar{\mathbf{x}}_{.t} - \bar{\bar{\mathbf{x}}})'\mathbf{b}.
\end{aligned} \tag{11-25}$$

The estimator of the asymptotic covariance matrix for $\mathbf{b}$ is computed using the sums of squares and cross products of $\mathbf{x}_{*it}$ computed in (11-24) and

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b} - m - a_i - d_t)^2}{nT - (n-1) - (T-1) - K - 1}. \tag{11-26}$$

The algebra of the two-way fixed effects estimator is rather complex—see, for example, Baltagi (2014). It is not obvious from the presentation so far, but the template result in (11-24) is incorrect if the panel is unbalanced. Unfortunately, for the unwary, the result does not fail in a way that would make the mistake obvious; if the panel is unbalanced, (11-24) simply leads to the wrong answer, but one that could look right. A numerical example is shown in Example 11.8. The conclusion for the practitioner is that (11-24) should only be used with balanced panels, but the augmented one-way estimator can be used in all cases.

Example 11.8 Two-Way Fixed Effects with Unbalanced Panel Data

The following experiment is done with the Cornwell and Rupert data used in Examples 11.4, 11.5, and 11.7. There are 595 individuals and 7 periods. Each group is 7 observations. Based on the balanced panel using all 595 individuals, in the fixed effects regression of ln *Wage* on just *Wks*, both methods give the answer $b = 0.00095$. If the first 300 groups are shortened by dropping the last 3 years of data, the unbalanced panel now has 300 groups with $T = 4$ and 295 with $T = 7$. For the same regression, the one-way estimate with time dummy variables is 0.00050 but the template result in (11-24) (which is incorrect) gives 0.00283.

11.4.5 REINTERPRETING THE WITHIN ESTIMATOR: INSTRUMENTAL VARIABLES AND CONTROL FUNCTIONS

The fixed effects model, in basic form, is

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + (c_i + \varepsilon_{it}).$$

We once again first consider least squares estimation. As we have already noted, for this case, $\mathbf{b}_{\text{OLS}}$ is inconsistent because of the correlation between $\mathbf{x}_{it}$ and c_i. Therefore, in the absence of the dummy variables, $\mathbf{x}_{it}$ is endogenous in this model. We used the within estimator in Section 11.4.1 instead of least squares to remedy the problem. The LSDV estimator is

$$\mathbf{b}_{\text{LSDV}} = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\ddot{\mathbf{y}}.$$

The LSDV estimator is computed by regressing $\mathbf{y}$ transformed to deviations from group means on the same transformation of $\mathbf{X}$; that is, $\mathbf{M_D y}$ on $\mathbf{M_D X}$. But, because $\mathbf{M_D}$ is idempotent, we may also write $\mathbf{b}_{LSDV} = (\ddot{\mathbf{X}}'\mathbf{X})^{-1}\ddot{\mathbf{X}}'\mathbf{y}$. In this form, $\ddot{\mathbf{X}}$ appears to be a set of instrumental variables, precisely in the form of (8-6). We have already demonstrated the consistency of the estimator, though it remains to verify the exogeneity and relevance conditions. These are both straightforward to verify. For the exogeneity condition, let $\mathbf{c}$ denote the full set of common effects. By construction, $(1/nT)\ddot{\mathbf{X}}'\mathbf{c} = \mathbf{0}$. We have assumed at the outset that $\text{plim}(1/nT)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. We need $\text{plim}(1/nT)\mathbf{X}'\mathbf{M_D}\boldsymbol{\varepsilon} = \text{plim}(1/nT)\mathbf{X}'(\mathbf{M_D}\boldsymbol{\varepsilon})$. If $\mathbf{X}$ is uncorrelated with ε, it will be uncorrelated with $\boldsymbol{\varepsilon}$ in deviations from its group means. For the relevance condition, all that will be needed is full rank of $(1/nT)\ddot{\mathbf{X}}'\mathbf{X}$, which is equivalent to $(1/nT)(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})$. This matrix will have full rank so long as no variables in $\mathbf{X}$ are time invariant—note that $(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}$ is used to compute $\mathbf{b}_{\text{LSDV}}$. The conclusion is that the data in group mean deviations form, that is, $\ddot{\mathbf{X}}$, are valid instrumental variables for estimation of the fixed effects model. This useful result will reappear when we examine Hausman and Taylor's model in Section 11.8.2.

We continue to assume that there are no time-invariant variables in $\mathbf{X}$. The matrix of group means is obtained as $\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{X} = \mathbf{P_D X} = (\mathbf{I} - \mathbf{M_D})\mathbf{X}$. [See (11-14)–(11-17).] Consider, then, least squares regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{P_D X}$, that is, on $\mathbf{X}$ and the group means, $\overline{\mathbf{X}}$. Using the partitioned regression formulation [Theorem 3.2 and (3-19)], we find this estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathbf{b}_{\text{Mundlak}} &= (\mathbf{X}'\mathbf{M_{PX}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M_{PX}}\mathbf{y}\\ &= \{\mathbf{X}'[\mathbf{I} - \overline{\mathbf{X}}(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}']\mathbf{X}\}^{-1} \times \{\mathbf{X}'[\mathbf{I} - \overline{\mathbf{X}}(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}']\mathbf{y}\}.\end{aligned}$$

This simplifies considerably. Recall $\overline{\mathbf{X}} = \mathbf{P_D X}$ and $\mathbf{P_D}$ is idempotent. We expand the first matrix in braces.

$$\begin{aligned}\{\mathbf{X}'[\mathbf{I} - (\mathbf{P_D X})[(\mathbf{P_D X})'(\mathbf{P_D X})]^{-1}(\mathbf{P_D X})']\mathbf{X}\} &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{P_D X}[\mathbf{X}'\mathbf{P}_{\mathbf{D}}'\mathbf{P_D X}]^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{D}}'\mathbf{X}\\ &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{P_D X}\\ &= \mathbf{X}'[\mathbf{I} - \mathbf{P_D}]\mathbf{X}\\ &= \mathbf{X}'\mathbf{M_D X}.\end{aligned}$$

The same result will emerge for the second term, which implies that the coefficients on $\mathbf{X}$ in the regression of $\mathbf{y}$ on $(\mathbf{X}, \overline{\mathbf{X}})$ is the within estimator, $\mathbf{b}_{\text{LSDV}}$. So, the group means qualify as a control function, as defined in Section 8.4.2. This useful insight makes the Mundlak approach a very useful method of dealing with fixed effects in regression, and by extension, in many other settings that appear in the literature.

11.4.6 PARAMETER HETEROGENEITY

With a small change in notation, the common effects model in (11-1) becomes

$$\begin{aligned} y_{it} &= c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \\ &= (\alpha + u_i) + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \\ &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \end{aligned}$$

where $E[u_i] = 0$ and $E[\alpha_i] = \alpha$. The heterogeneity affects the constant term. We can extend the model to allow other parameters to be heterogeneous as well. In the labor market model examined in Example 11.3, an extension in which the partial effect of weeks worked depends on both market and individual characteristics, might appear as

$$\begin{aligned} \ln Wage_{it} = {} & \alpha_{i1} + \alpha_{i2}\, Wks_{it} + \beta_2\, Exp_{it} + \beta_3\, Exp_{it}^2 + \beta_5\, Occ_{it} \\ & + \beta_6\, Ind_{it} + \beta_7\, South_{it} + \beta_8\, SMSA_{it} + \beta_9\, MS_{it} \\ & + \beta_{10}\, Union_{it} + \beta_{11}\, Ed_i + \beta_{12}\, Fem_i + \beta_{13}\, Blk_i + \varepsilon_{it} \end{aligned}$$

$$\boldsymbol{\alpha}_i = \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} = \boldsymbol{\alpha} + \mathbf{u}_i.$$

Another interesting case is a random trend model, $y_{it} = \alpha_{i1} + \alpha_{i2}t + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$. As before, the difference between the random and fixed effects models is whether $E[\mathbf{u}_i|\mathbf{X}_i]$ is zero or not. For the present, we will allow this to be nonzero—a fixed effects form of the model.

The preceding developments have been concerned with a strategy for estimation and inference about $\boldsymbol{\beta}$ in the presence of $\mathbf{u}_i$. In this fixed effects setting, the dummy variable approach of Section 11.4.1 can be extended essentially with only a small change in notation. First, let's generalize the model slightly,

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\alpha}_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

In the basic common effects model, $\mathbf{z}_{it}' = (1)$; in the random trend model, $\mathbf{z}_{it}' = (1,t)$; in the suggested extension of the labor market model, $\mathbf{z}_{it}' = (1, Wks_{it})$, with $E[\mathbf{u}_i|\mathbf{X}_i,\mathbf{Z}_i] \neq \mathbf{0}$ (fixed effects) and $E[\mathbf{u}_i\mathbf{u}_i'|\mathbf{X}_i, \mathbf{Z}_i] = \boldsymbol{\Sigma}$, a constant, positive definite matrix. The strict exogeneity assumption now is $E[\varepsilon_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT}, \mathbf{u}_i] = 0$. For the present, we assume ε_{it} is homoscedastic and nonautocorrelated, so $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{u}_i] = \sigma_\varepsilon{}^2\mathbf{I}$. We can approach estimation of $\boldsymbol{\beta}$ the same way we did in Section 11.4.1. Recall the LSDV estimator is based on

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{11-27}$$

where $\mathbf{D}$ is the $nT \times n$ matrix of individual specific dummy variables. The estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned} \mathbf{b}_{\text{LSDV}} &= (\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{D}\mathbf{y} = \left[\sum_{i=1}^n \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left[\sum_{i=1}^n \ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i\right], \\ \mathbf{a}_{\text{LSDV}} &= (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{X}\mathrm{b}_{\text{LSDV}}) = \bar{\mathbf{y}} - \bar{\mathbf{X}}\mathbf{b}_{\text{LSDV}}, \\ \mathbf{M}_\mathbf{D} &= \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'. \end{aligned}$$

The special structure of $\mathbf{D}$—the columns are orthogonal—allows the calculations to be done with two convenient steps: (1) compute $\mathbf{b}_{\text{LSDV}}$ by regression of $(y_{it} - \bar{y}_{i.})$ on $(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})$; (2) compute a_i as $(1/T)\Sigma_t(y_{it} - \mathbf{x}_{it}'\mathbf{b}_{\text{LSDV}})$.

No new results are needed to develop the fixed effects estimator in the extended model. In this specification, we have simply respecified $\mathbf{D}$ to contain two or more sets of N columns. For the time trend case, for example, define an $nT \times 1$ column vector of time trends, $\mathbf{t}^{*\prime} = (1, 2, \ldots, T, 1, 2, \ldots, T, \ldots, 1, 2, \ldots, T)$. Then, $\mathbf{D}$ has $2N$ columns, $\{[\mathbf{d}_1, \ldots, \mathbf{d}_n], [\mathbf{d}_1 \circ \mathbf{t}^*, \mathbf{d}_2 \circ \mathbf{t}^* \ldots, \mathbf{d}_n \circ \mathbf{t}^*]\}$. This is an $nT \times 2n$ matrix of dummy variables and interactions of the dummy variables with the time trend. (The operation $\mathbf{d}_i \circ \mathbf{t}^*$ is the Hadamard product—element by element multiplication—of $\mathbf{d}_i$ and $\mathbf{t}^*$.) With $\mathbf{D}$ redefined this way, the results in Section 11.4.1 can be applied as before. For example, for the random trends model, $\ddot{\mathbf{X}}_i$ is obtained by "detrending" the columns of $\mathbf{X}_i$. Define $\mathbf{Z}_i$ to be the $T \times 2$ matrix $(\mathbf{1},\mathbf{t})$. Then, for individual i, the block of data in $\ddot{\mathbf{X}}_i$ is $[\mathbf{I} - \mathbf{Z}_i(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i']\mathbf{X}_i$ and $\mathbf{b}_{\text{LSDV}}$ is computed using (11-20). (Note that this requires that T be at least $J + 1$ where J is the number of variables in $\mathbf{Z}$. In the simpler fixed effects case, we require at least two observations in group i. Here, in the random trend model, that would be three observations.) In computing s^2, the appropriate degrees of freedom will be $(n(T - J) - K)$. The asymptotic covariance matrices in (11-17) and (11-21) are computed as before.[15] For each group, $\mathbf{a}_i = (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{\text{LSDV}})$. The natural estimator of $\alpha = E[\alpha_i]$ would be $\bar{\mathbf{a}} = \frac{1}{n}\Sigma_{i=1}^n \mathbf{a}_i$. The asymptotic variance matrix for $\bar{\mathbf{a}}$ can be estimated with

$$\text{Est.Asy.Var}[\bar{\mathbf{a}}] = (1/n^2)\Sigma_i \mathbf{f}_i\mathbf{f}_i' \text{ where } \mathbf{f}_i = [(\mathbf{a}_i - \bar{\mathbf{a}}) - \mathbf{C}\mathbf{A}^{-1}\ddot{\mathbf{X}}_i'\mathbf{e}_i], \mathbf{A} = \tfrac{1}{n}\Sigma_{i=1}^n \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i \text{ and } \mathbf{C} = \tfrac{1}{n}\Sigma_{i=1}^n (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\mathbf{X}_i.$$

[See Wooldridge (2010, p. 381).]

Example 11.9 Heterogeneity in Time Trends in an Aggregate Production Function

We extend Munnell's (1990) proposed model of productivity of public capital at the state level that was estimated in Example 10.1. The central equation of the analysis that we will extend here is a Cobb–Douglas production function,

$$\ln gsp_{it} = \alpha_{i1} + \alpha_{i2}t + \beta_1 \ln pc_{it} + \beta_2 \ln hwy_{it} + \beta_3 \ln water_{it} + \beta_4 \ln util_{it} + \beta_5 \ln emp_{it} + \beta_6 unemp_{it} + \varepsilon_{it},$$

where
- *gsp* = gross state product,
- *pc* = private capital,
- *hwy* = highway capital,
- *water* = water utility capital,
- *util* = utility capital,
- *emp* = employment (labor),
- *unemp* = unemployment rate.

The data, measured for the lower 48 U.S. states (excluding Alaska and Hawaii) and years 1970–1986, are given in Appendix Table F10.1. Table 11.10 reports estimates of the several

[15]The random trends model is a special case that can be handled by differences rather than the partitioned regression method used here. In $y_{it} = \alpha_{i1} + \alpha_{i2}t + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$, $(y_{it} - y_{i,t-1}) = \Delta y_{it} = \alpha_{i2} + (\Delta\mathbf{x}_{it})'\boldsymbol{\beta} + \Delta\varepsilon_{it}$. The time trend becomes the common effect. This can be treated as a fixed effects model. Or, taking a second difference, $\Delta y_{it} - \Delta y_{i,t-1} = \Delta^2 y_{it}$ removes α_{i2} and leaves a linear regression, $\Delta^2 y_{it} = \Delta^2\mathbf{x}_{it}'\boldsymbol{\beta} + \Delta^2\varepsilon_{it}$. Details are given in Wooldridge (2010, pp. 375–377).

TABLE 11.10 Estimates of Fixed Effects Statewide Production Functions

	Constant	*Trend*	ln *PC*	ln *Hwy*	ln *Water*	ln *Util*	ln *Emp*	*Unemp*	R^2
Pooled Model	1.91618	0.00108	0.30669	0.06708	0.11607	0.01054	0.54838	−0.00812	0.99307
Std.Error	0.05287	0.00072	0.01163	0.01633	0.01246	0.01241	0.01555	0.00149	
Robust S.E.	0.21420	0.00162	0.05136	0.05881	0.03481	0.04138	0.06825	0.00341	
Fixed Effects		0.00625	0.13751	0.08529	0.02966	−0.09807	0.75870	−0.00732	0.99888
Std.Error		0.00080	0.02814	0.03010	0.01574	0.01760	0.02915	0.00098	
Robust S.E.		0.00179	0.08248	0.08878	0.04147	0.05672	0.08744	0.00226	
Random Trend			0.11228	−0.01120	−0.03181	−0.08828	0.71207	−0.00793	0.99953
Std.Error	5.41942	0.01108	0.02645	0.03900	0.0166	0.02339	0.03105	0.00083	
Robust S.E	0.54657	0.00207	0.04189	0.07554	0.03110	0.04655	0.04803	0.00123	
Difference			−0.09104	−0.20767	−0.00094	0.12013	0.94832	−0.00374	
Std.Error			0.02324	0.10785	0.03582	0.05647	0.05312	0.00080	
Newey–West(2)			0.03746	0.14590	0.04343	0.07181	0.06576	0.00095	
Robust S.E.			0.04129	0.14358	0.03635	0.07552	0.06920	0.00112	

fixed effects models. The pooled estimator is computed using simple least squares for all 816 observations. The standard errors use $s^2(\mathbf{X}'\mathbf{X})^{-1}$. The robust standard errors are based on (11-4). For the two fixed effects models, the standard errors are based on (11–17) and (11–21). Finally, for the difference estimator, two sets of robust standard errors are computed. The Newey–West estimator assumes that ε_{it} in the model is homoscedastic so that $\Delta^2\varepsilon_{it} = \varepsilon_{it} - 2\varepsilon_{i,t-1} + \varepsilon_{i,t-2}$. The robust standard errors are based, once again, on (11-4). Note that two observations have been dropped from each state with the second difference estimator. The patterns of the standard errors are predictable. They all rise substantially with the correction for clustering, in spite of the presence of the fixed effects. The effect is quite substantial, with most of the standard errors rising by a factor of 2 to 4. The Newey–West correction (see Section 20.5.2) of the difference estimators seems mostly to cover the effect of the autocorrelation. The *F* test for the hypothesis that neither the constant nor the trend are heterogeneous is F[94,816-96-6] = [(0.99953 − 0.99307)/94]/[(1 − 0.99953)/(816 − 96 − 6)] = 104.40. The critical value from the *F* table is 1.273, so the hypothesis of homogeneity is rejected. The differences in the estimated parameters across the specifications are also quite pronounced. The difference between the random trend and difference estimators is striking, given that these are two different estimation approaches to the same model.

11.5 RANDOM EFFECTS

The fixed effects model allows the unobserved individual effects to be correlated with the included variables. We then modeled the differences between units as parametric shifts of the regression function. This model might be viewed as applying only to the cross-sectional units in the study, not to additional ones outside the sample. For example, an intercountry comparison may well include the full set of countries for which it is reasonable to assume that the model is constant. Example 6.5 is based on a panel consisting of data on 31 baseball teams. Save for rare discrete changes in the league, these 31 units will always be the entire population. If the individual effects are strictly uncorrelated with the regressors, then it might be appropriate to model the individual specific constant terms as randomly distributed across cross-sectional units. This view would be appropriate if we believed that sampled cross-sectional units were drawn from a large population. It would certainly be the case for the longitudinal data sets listed in the introduction to this chapter and for the labor market data we have used in several examples in this chapter.[16]

The payoff to this form is that it greatly reduces the number of parameters to be estimated. The cost is the possibility of inconsistent estimators, if the assumption is inappropriate.

Consider, then, a reformulation of the model,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it}, \qquad \textbf{(11-28)}$$

where there are K regressors including a constant and now the single constant term is the mean of the unobserved heterogeneity, $E[\mathbf{z}_i'\boldsymbol{\alpha}]$. The component u_i is the random heterogeneity specific to the ith observation and is constant through time; recall from Section 11.2.1, $u_i = \{\mathbf{z}_i'\boldsymbol{\alpha} - E[\mathbf{z}_i'\boldsymbol{\alpha}]\}$. For example, in an analysis of families, we can view

[16]This distinction is not hard and fast; it is purely heuristic. We shall return to this issue later. See Mundlak (1978) for a methodological discussion of the distinction between fixed and random effects.

u_i as the collection of factors, $\mathbf{z}_i'\boldsymbol{\alpha}$, not in the regression that are specific to that family. We continue to assume strict exogeneity:

$$\begin{aligned} E[\varepsilon_{it}|\mathbf{X}_i] &= E[u_i|\mathbf{X}_i] = 0, \\ E[\varepsilon_{it}^2|\mathbf{X}_i] &= \sigma_\varepsilon^2, \\ E[u_i^2|\mathbf{X}_i] &= \sigma_u^2, \\ E[\varepsilon_{it}u_j|\mathbf{X}_i] &= 0 \quad \text{for all } i, t, \text{ and } j, \\ E[\varepsilon_{it}\varepsilon_{js}|\mathbf{X}_i] &= 0 \quad \text{if } t \neq s \text{ or } i \neq j, \\ E[u_iu_j|\mathbf{X}_i, \mathbf{X}_j] &= 0 \quad \text{if } i \neq j. \end{aligned} \tag{11-29}$$

As before, it is useful to view the formulation of the model in blocks of T observations for group i, $\mathbf{y}_i$, $\mathbf{X}_i$, $u_i\mathbf{i}$, and $\boldsymbol{\varepsilon}_i$. For these T observations, let

$$\eta_{it} = \varepsilon_{it} + u_i$$

and

$$\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}]'.$$

In view of this form of $\boldsymbol{\eta}_{it}$, we have what is often called an **error components model**. For this model,

$$\begin{aligned} E[\eta_{it}^2|\mathbf{X}_i] &= \sigma_\varepsilon^2 + \sigma_u^2, \\ E[\eta_{it}\eta_{is}|\mathbf{X}_i] &= \sigma_u^2, \quad t \neq s, \\ E[\eta_{it}\eta_{js}|\mathbf{X}_i] &= 0 \qquad \text{for all } t \text{ and } s, \text{ if } i \neq j. \end{aligned} \tag{11-30}$$

For the T observations for unit i, let $\boldsymbol{\Sigma} = E[\boldsymbol{\eta}_i, \boldsymbol{\eta}_i'|\mathbf{X}]$. Then

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ & \cdots & & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2\,\mathbf{I}_T + \sigma_u^2\mathbf{i}_T\mathbf{i}_T', \tag{11-31}$$

where $\mathbf{i}_T$ is a $T \times 1$ column vector of 1s. Because observations i and j are independent, the disturbance covariance matrix for the full nT observations is

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ & & & \vdots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}. \tag{11-32}$$

11.5.1 LEAST SQUARES ESTIMATION

The model defined by (11-28),

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it},$$

with the strict exogeneity assumptions in (11-29) and the covariance matrix detailed in (11-31) and (11-32), is a generalized regression model that fits into the framework we developed in

Chapter 9. The disturbances are autocorrelated in that observations are correlated across time within a group, though not across groups. All the implications of Section 9.2 would apply here. In particular, the parameters of the random effects model can be estimated consistently, though not efficiently, by ordinary least squares (OLS). An appropriate robust asymptotic covariance matrix for the OLS estimator would be given by (11-3).

There are other consistent estimators available as well. By taking deviations from group means, we obtain

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i.$$

This implies that (assuming there are no time-invariant regressors in $\mathbf{x}_{it}$), the LSDV estimator of (11-14) is a consistent estimator of $\boldsymbol{\beta}$. An estimator based on first differences,

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \varepsilon_{it} - \varepsilon_{i,t-1}.$$

(The LSDV and first differences estimators are robust to whether the correct specification is actually a random or a fixed effects model.) As is OLS, LSDV is inefficient because, as we will show in Section 11.5.2, there is an efficient GLS estimator that is not equal to $\mathbf{b}_{\text{LSDV}}$. The group means (between groups) regression model,

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + u_i + \bar{\varepsilon}_i, i = 1, \ldots, n,$$

provides a fourth method of consistently estimating the coefficients $\boldsymbol{\beta}$. None of these is the preferred estimator in this setting because the GLS estimator will be more efficient than any of them. However, as we saw in Chapters 9 and 10, many generalized regression models are estimated in two steps, with the first step being a robust least squares regression that is used to produce a first round estimate of the variance parameters of the model. That would be the case here as well. To suggest where this logic will lead in Section 11.5.3, note that for the four cases noted, the sum of squared residuals can produce the following consistent estimators of functions of the variances:

$$\begin{aligned}
&\text{(Pooled)} && \text{plim } [\mathbf{e}_{\text{pooled}}'\mathbf{e}_{\text{pooled}}/(nT)] = \sigma_u^2 + \sigma_\varepsilon^2,\\
&\text{(LSDV)} && \text{plim } [\mathbf{e}_{\text{LSDV}}'\mathbf{e}_{\text{LSDV}}/(n(T-1)-K)] = \sigma_\varepsilon^2,\\
&\text{(Differences)} && \text{plim } [\mathbf{e}_{\text{FD}}'\mathbf{e}_{FD}/(n(T-1))] = 2\sigma_\varepsilon^2,\\
&\text{(Means)} && \text{plim } [\mathbf{e}_{\text{means}}'\mathbf{e}_{\text{means}}/(nT)] = \sigma_u^2 + \sigma_\varepsilon^2/T.
\end{aligned}$$

Baltagi (2001) suggests yet another method of moments estimator that could be based on the pooled OLS results. Based on (11-31), $\text{Cov}(\varepsilon_{it},\varepsilon_{is}) = \sigma_u^2$ within group i for $t \neq s$. There are $T(T-1)/2$ pairs of residuals that can be used, so for each group, we could use $(1/(T(T-1)/2))\Sigma_s\Sigma_t\, e_{it}\, e_{is}$ to estimate σ_u^2. Because we have n groups that each provide an estimator, we can average the n implied estimators, to obtain

$$\text{(OLS)} \quad \text{plim } \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{t=2}^{T}\sum_{s=1}^{t-1}e_{it}e_{is}}{T(T-1)/2} = \sigma_u^2.$$

Different pairs of these estimators (and other candidates not shown here) could provide a two-equation method of moments estimator of $(\sigma_u^2, \sigma_\varepsilon^2)$. (Note that the last of these is using a covariance to estimate a variance. Unfortunately, unlike the others, this could be negative in a finite sample.) With these in mind, we will now develop an efficient generalized least squares estimator.

11.5.2 GENERALIZED LEAST SQUARES

The generalized least squares estimator of the slope parameters is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = \left(\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right).$$

To compute this estimator as we did in Chapter 9 by transforming the data and using ordinary least squares with the transformed data, we will require $\boldsymbol{\Omega}^{-1/2} = [\mathbf{I}_n \otimes \boldsymbol{\Sigma}]^{-1/2} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}^{-1/2}$. We need only find $\boldsymbol{\Sigma}^{-1/2}$, which is

$$\boldsymbol{\Sigma}^{-1/2} = \left[\mathbf{I}_T - \frac{\theta_i}{T}\mathbf{i}_T\mathbf{i}_T'\right], \tag{11-33}$$

where

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

The transformation of $\mathbf{y}_i$ and $\mathbf{X}_i$ for GLS is therefore

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{y}_i = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} y_{i1} - \theta\bar{y}_{i.} \\ y_{i2} - \theta\bar{y}_{i.} \\ \vdots \\ y_{i\,\mathrm{T}} - \theta\bar{y}_{i.} \end{bmatrix}, \tag{11-34}$$

and likewise for the rows of $\mathbf{X}_i$. For the data set as a whole, then, generalized least squares is computed by the regression of these partial deviations of y_{it} on the same transformations of $\mathbf{x}_{it}$. Note the similarity of this procedure to the computation in the LSDV model, which uses $\theta = 1$ in (11-15).

It can be shown that the GLS estimator is, like the pooled OLS estimator, a matrix weighted average of the within- and between-units estimators,

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{F}}^{within}\mathbf{b}^{within} + (\mathbf{I} - \hat{\mathbf{F}}^{within})\mathbf{b}^{between},$$

where now,

$$\hat{\mathbf{F}}^{within} = [\mathbf{S}_{xx}^{within} + \lambda\mathbf{S}_{xx}^{between}]^{-1}\mathbf{S}_{xx}^{within},$$

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2} = (1 - \theta)^2.$$

To the extent that λ differs from one, we see that the inefficiency of ordinary least squares will follow from an inefficient weighting of the two estimators. Compared with generalized least squares, ordinary least squares places too much weight on the between-units variation. It includes all of it in the variation in $\mathbf{X}$, rather than apportioning some of it to random variation across groups attributable to the variation in u_i across units.

Unbalanced panels complicate the random effects model a bit. The matrix $\boldsymbol{\Omega}$ in (11-32) is no longer $\mathbf{I}_n \otimes \boldsymbol{\Sigma}$ because the diagonal blocks in $\boldsymbol{\Omega}$ are of different sizes. In (11-33), the ith diagonal block in $\boldsymbol{\Omega}^{-1/2}$ is

$$\boldsymbol{\Sigma}_i^{-1/2} = \frac{1}{\sigma_\varepsilon}\left[\mathbf{I}_{T_i} - \frac{\theta_i}{T_i}\mathbf{i}_{T_i}\mathbf{i}_{T_i}'\right],\ \theta_i = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T_i\sigma_u^2}}.$$

In principle, estimation is still straightforward, because the source of the groupwise heteroscedasticity is only the unequal group sizes. Thus, for GLS, or FGLS with estimated variance components, it is necessary only to use the group-specific θ_i in the transformation in (11-34).

11.5.3 FEASIBLE GENERALIZED LEAST SQUARES ESTIMATION OF THE RANDOM EFFECTS MODEL WHEN $\boldsymbol{\Sigma}$ IS UNKNOWN

If the variance components are known, generalized least squares can be computed as shown earlier. Of course, this is unlikely, so as usual, we must first estimate the disturbance variances and then use an FGLS procedure. A heuristic approach to estimation of the variance components is as follows:

and

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \varepsilon_{it} + u_i \\ \bar{y}_{i.} &= \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \alpha + \bar{\varepsilon}_{i.} + u_i. \end{aligned} \tag{11-35}$$

Therefore, taking deviations from the group means removes the heterogeneity,

$$y_{it} - \bar{y}_{i.} = [\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}]'\boldsymbol{\beta} + [\varepsilon_{it} - \bar{\varepsilon}_{i.}]. \tag{11-36}$$

Because

$$E\left[\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_{i.})^2\right] = (T-1)\sigma_\varepsilon^2,$$

if $\boldsymbol{\beta}$ were observed, then an unbiased estimator of σ_ε^2 based on T observations in group i would be

$$\hat{\sigma}_\varepsilon^2(i) = \frac{\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_{i.})^2}{T-1}. \tag{11-37}$$

Because $\boldsymbol{\beta}$ must be estimated—the LSDV estimator is consistent, indeed, unbiased in general—we make the degrees of freedom correction and use the LSDV residuals in

$$s_e^2(i) = \frac{\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{T-K-1} \tag{11-38}$$

(Note that based on the LSDV estimates, $\bar{e}_{i.}$ is actually zero. We will carry it through nonetheless to maintain the analogy to (11-35) where $\bar{\varepsilon}_{i.}$ is not zero but is an estimator of $E[\varepsilon_{it}] = 0$.) We have n such estimators, so we average them to obtain

$$\bar{s}_e^2 = \frac{1}{n}\sum_{i=1}^{n} s_e^2(i) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{T-K-1}\right] = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{nT - nK - n} \tag{11-39a}$$

The degrees of freedom correction in $\bar{s}_e^2$ is excessive because it assumes that α and $\boldsymbol{\beta}$ are reestimated for each i. The estimated parameters are the n means $\bar{y}_{i\cdot}$ and the K slopes. Therefore, we propose the unbiased estimator[17]

$$\hat{\sigma}_\varepsilon^2 = s_{LSDV}^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i.})^2}{nT - n - K}.$$

[17]A formal proof of this proposition may be found in Maddala (1971) or in Judge et al. (1985, p. 551).

This is the variance estimator in the fixed effects model in (11-18), appropriately corrected for degrees of freedom. It remains to estimate σ_u^2. Return to the original model specification in (11-35). In spite of the correlation across observations, this is a classical regression model in which the ordinary least squares slopes and variance estimators are both consistent and, in most cases, unbiased. Therefore, using the ordinary least squares residuals from the model with only a single overall constant, we have

$$\text{plim } s^2_{Pooled} = \text{plim } \frac{\mathbf{e}'\mathbf{e}}{nT - K - 1} = \sigma_\varepsilon^2 + \sigma_u^2. \qquad \textbf{(11-39b)}$$

This provides the two estimators needed for the variance components; the second would be $\hat{\sigma}_u^2 = s^2_{Pooled} - s^2_{\text{LSDV}}$. As noted in Section 11.5.1, there are a variety of pairs of variance estimators that can be used to obtain estimates of ${\sigma_\varepsilon}^2$ and ${\sigma_u}^2$.[18] The estimators based on s^2_{LSDV} and s^2_{Pooled} are common choices. Alternatively, let $[\mathbf{b}, a]$ be any consistent estimator of $[\boldsymbol{\beta}, \alpha]$ in (11-35), such as the ordinary least squares estimator. Then, s^2_{Pooled} provides a consistent estimator of $m_{ee} = \sigma_\varepsilon^2 + \sigma_u^2$. The mean squared residuals using a regression based only on the n group means in (11-35) provides a consistent estimator of $m_{**} = \sigma_u^2 + (\sigma_\varepsilon^2/T)$, so we can use

$$\hat{\sigma}_\varepsilon^2 = \frac{T}{T-1}(m_{ee} - m_{**}),$$
$$\hat{\sigma}_u^2 = \frac{T}{T-1}m_{**} - \frac{1}{T-1}m_{ee} = \omega m_{**} + (1 - \omega)m_{ee},$$

where $\omega > 1$. A possible complication is that the estimator of ${\sigma_u}^2$ can be negative in any of these cases. This happens fairly frequently in practice, and various ad hoc solutions are typically tried. (The first approach is often to try out different pairs of moments. Unfortunately, typically, one failure is followed by another. It would seem that this failure of the estimation strategy should suggest to the analyst that there is a problem with the specification to begin with. A last solution in the face of a persistently negative estimator is to set ${\sigma_u}^2$ to the value the data are suggesting, zero, and revert to least squares.)

11.5.4 ROBUST INFERENCE AND FEASIBLE GENERALIZED LEAST SQUARES

The feasible GLS estimator based on (11-28) and (11-31) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}X)^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}) = \left(\sum_{i=1}^n \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^n \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{y}_i\right).$$

There is a subscript i on $\boldsymbol{\Sigma}_i$ because of the consideration of unbalanced panels discussed at the end of Section 11.5.2. If the panel is unbalanced, a minor adjustment is needed because $\boldsymbol{\Sigma}_i$ is $T_i \times T_i$ and because of the specific computation of θ_i. The feasible GLS estimator is then

$$\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{\beta} + \left(\sum_{i=1}^n \mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^n \mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\varepsilon}_i\right). \qquad \textbf{(11-40)}$$

[18]See, for example, Wallace and Hussain (1969), Maddala (1971), Fuller and Battese (1974), and Amemiya (1971). This is a point on which modern software varies. Generally, programs begin with (11-39a) and (11-39b) to estimate the variance components. Others resort to different strategies based on, for example, the group means estimator. The unfortunate implication for the unwary is that different programs can systematically produce different results using the same model and the same data. The practitioner is strongly advised to consult the program documentation for resolution.

This form suggests a way to accommodate failure of the random effects assumption in (11-28). Following the approach used in the earlier applications, the estimator would be

$$\text{Est.Asy.Var}[\hat{\hat{\boldsymbol{\beta}}}] = \left(\sum_{i=1}^{n}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{e}_i)(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{e}_i)'\right)\left(\sum_{i=1}^{n}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}. \tag{11-41}$$

With this estimator in hand, inference would be based on Wald statistics rather than F statistics.

There is a loose end in the proposal just made. If assumption (11-28) fails, then what are the properties of the generalized least squares estimator based on $\boldsymbol{\Sigma}$ in (11-31)? The FGLS estimator remains consistent and asymptotically normally distributed—consider that OLS is also a consistent estimator that uses the wrong covariance matrix. And (11-41) would provide an appropriate estimator to use for statistical inference about $\boldsymbol{\beta}$. However, in this case, (11-31) is the wrong starting point for FGLS estimation.

If the random effects assumption is not appropriate, then a more general starting point is

$$\mathbf{y}_i = \alpha\mathbf{i} + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,\ E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i] = \boldsymbol{\Sigma},$$

which returns us to the pooled regression model in Section 11.3.1. An appealing approach based on that would base feasible GLS on (11-32) and, assuming n is reasonably large and T is relatively small, would use $\hat{\boldsymbol{\Sigma}} = \dfrac{1}{n}\sum_{i=1}^{n}\mathbf{e}_{OLS,i}\mathbf{e}_{OLS,i}'$. Then, feasible GLS would be based on (11-40). One serious complication is how to accommodate an unbalanced panel. With the random effects formulation, the covariances in $\boldsymbol{\Sigma}$ are identical, so positioning of the observations in the matrix is arbitrary. This is not so with an unbalanced panel. We will see in the example below, in this more general case, a distinct pattern in the locations of the cells in the matrix emerges. It is unclear what should be done with the unfilled cells in $\boldsymbol{\Sigma}$.

11.5.5 TESTING FOR RANDOM EFFECTS

Breusch and Pagan (1980) have devised a **Lagrange multiplier test** for the random effects model based on the OLS residuals.[19] For

$$H_0\colon \sigma_u^2 = 0,$$
$$H_1\colon \sigma_u^2 > 0,$$

the test statistic is

$$LM = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}\left[\sum_{t=1}^{T}e_{it}\right]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}e_{it}^2} - 1\right]^2 = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}(T\bar{e}_{i.})^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}e_{it}^2} - 1\right]^2. \tag{11-42}$$

Under the null hypothesis, the limiting distribution of LM is chi-squared with one degree of freedom. (The computation for an unbalanced panel replaces the multiple by $[(\Sigma_{i=1}^{n}T_i)^2]/[2\Sigma_{i=1}^{n}T_i(T_i-1)]$ and replaces T with T_i in the summations.) The LM

[19]Thus far, we have focused strictly on generalized least squares and moments-based consistent estimation of the variance components. The LM test is based on maximum likelihood estimation, instead. See Maddala (1971) and Baltagi (2013) for this approach to estimation.

statistic is based on normally distributed disturbances. Wooldridge (2010) proposed a statistic that is more robust to the distribution, $z = \frac{\sum_{i=1}^{n}(\sum_{t=2}^{T_i}\sum_{s=1}^{t-1}e_{it}e_{is})}{\sqrt{\sum_{i=1}^{n}(\sum_{t=2}^{T_i}\sum_{s=1}^{t-1}e_{it}e_{is})^2}}$, which converges to $N[0,1]$ in all cases, or z^2 which has a limiting chi-squared distribution with one degree of freedom. The inner double sums in the statistic sum the below diagonal terms in $\mathbf{e}_i\mathbf{e}_i'$ which is one-half the sum of all the terms minus the diagonals, $\mathbf{e}_i'\mathbf{e}_i$. The ith term in the sum is $\frac{1}{2}[(\sum_{t=1}^{T}e_{it})^2 - (\sum_{t=1}^{T}e_{it}^2)] = f_i$. By manipulating this result, we find that $z^2 = (n\bar{f}^2/s_f^2)$ (where s_f^2 is computed around the assumed $E[f_i] = 0$), which would be the standard test statistic for the hypothesis that $E[f_i] = 0$. This makes sense, because f_i is essentially composed of the difference between two estimators of σ_ε^2.[20] With some tedious manipulation, we can show that the LM statistic is also a multiple of $n\bar{f}^2$.

Example 11.10 Test for Random Effects

We are interested in comparing the random and fixed effects estimators in the Cornwell and Rupert wage interested equation,

$$\begin{aligned}\ln Wage_{it} = {} & \beta_1 + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} \\ & + \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} \\ & + \beta_{10} Union_{it} + \beta_{11} Ed_i + \beta_{12} Fem_i + \beta_{13} Blk_i + c_i + \varepsilon_{it}.\end{aligned}$$

The least squares estimates appear in Table 11.6 in Example 11.4. We will test for the presence of random effects. The computations in the two statistics are simpler than it might appear at first. The LM statistic is

$$LM = \frac{nT}{2(T-1)}\left[\frac{\mathbf{e}'\mathbf{D}\mathbf{D}'\mathbf{e}}{\mathbf{e}'\mathbf{e}} - 1\right]^2,$$

where $\mathbf{D}$ is the matrix of individual dummy variables in (11-13). To compute z^2, we compute

$$\mathbf{f} = \frac{1}{2}(\mathbf{D}'\mathbf{e} \circ \mathbf{D}'\mathbf{e} - \mathbf{D}'(\mathbf{e} \circ \mathbf{e})),$$

($\circ$ is the Hadamard product—element by element multiplication) then $z^2 = \mathbf{i}'\mathbf{f}/\mathbf{f}'\mathbf{f}$. The results for the two statistics are $LM = 3497.02$ and $z^2 = 179.66$. These far exceed the 95% critical value for the chi-squared distribution with one degree of freedom, 3.84. At this point, we conclude that the classical regression model without the heterogeneity term is inappropriate for these data. The result of the test is to reject the null hypothesis in favor of the random effects model. But it is best to reserve judgment on that because there is another competing specification that might induce these same results, the fixed effects model. We will examine this possibility in the subsequent examples.

With the variance estimators in hand, FGLS can be used to estimate the parameters of the model. All of our earlier results for FGLS estimators apply here. In particular, all that is needed for efficient estimation of the model parameters are consistent estimators of the variance components, and there are several.[21]

[20] Wooldridge notes that z can be negative, suggesting a negative estimate of σ_u^2. This counterintuitive result arises, once again (see Section 11.5.1), from using a covariance estimator to estimate a variance. However, with some additional manipulation, we find that the numerator of z is actually $(nT/2)[\hat{\sigma}_\varepsilon^2(\text{based on } \bar{e}_i) - \hat{\sigma}_\varepsilon^2(\text{based on } e_{it})]$ so the outcome is not so contradictory as it might appear—since the statistic has a standard normal distribution, the negative result should occur half of the time. The test is not actually based on the covariance; it is based on the difference of two estimators of the same variance (under the null hypothesis). The numerator of the LM statistic, $\mathbf{e}'\mathbf{D}\mathbf{D}'\mathbf{e} - \mathbf{e}'\mathbf{e}$, is the same as that of z, though it is squared to produce the test statistic.

[21] See Hsiao (2003), Baltagi (2005), Nerlove (2002), Berzeg (1979), and Maddala and Mount (1973).

Example 11.11 Estimates of the Random Effects Model

In the previous example, we found the total sum of squares for the least squares estimator was 506.766. The fixed effects (LSDV) estimates for this model appear in Table 11.10. The sum of squares is 82.26732. Therefore, the moment estimators of the variance parameters are

$$\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_u^2 = \frac{506.766}{4165 - 13} = 0.122053$$

and

$$\hat{\sigma}_\varepsilon^2 = \frac{82.26732}{4165 - 595 - 9} = 0.0231023.$$

The implied estimator of σ_u^2 is 0.098951. (No problem of negative variance components has emerged. Note that the three time-invariant variables have not been used in computing the fixed effects estimator to estimate σ_ε^2.) The estimate of θ for FGLS is

$$\hat{\theta} = 1 - \sqrt{\frac{0.0231023}{0.0231023 + 7(0.098951)}} = 0.820343.$$

FGLS estimates are computed by regressing the partial differences of ln $Wage_{it}$ on the partial differences of the constant and the 12 regressors, using this estimate of θ in (11-33). The full GLS estimates are obtained by estimating $\boldsymbol{\Sigma}$ using the OLS residuals. The estimate of $\boldsymbol{\Sigma}$ is listed below with the other estimates. Thus, $\hat{\boldsymbol{\Sigma}} = \frac{1}{595}\sum_{i=1}^{595}\mathbf{e}_i\mathbf{e}_i'$. The estimate of $\boldsymbol{\Omega} = \sigma_\varepsilon^2\mathbf{I} + \sigma_U^2\mathbf{ii}'$. Estimates of the parameters using the OLS and random effects estimators appear in Table 11.11. The similarity of the estimates is to be expected given that, under the hypothesis of the model, all three estimators are consistent.

The random effects specification is a substantive restriction on the stochastic part of the regression. The assumption that the disturbances are equally correlated across periods regardless of how far apart the periods are may be a particularly strong assumption, particularly if the time dimension of the panel is relatively long. The force of the restrictions can be seen in the covariance matrices shown below. In the random effects model, the cross period correlation is $\sigma_u^2/(\sigma_\varepsilon^2 + \sigma_u^2)$ which we have estimated as 0.9004 for all periods. But, the first column of the estimate of $\boldsymbol{\Sigma}$ suggests quite a different pattern; the cross period covariances diminish substantially with the separation in time. If an AR(1) pattern is assumed, $\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + v_{i,t}$ then the implied estimate of ρ would be $r = 0.1108/0.1418 = 0.7818$. The next two periods appear consistent with the pattern, ρ^2 then ρ^3. The first-order autoregression might be a reasonable candidate for the model. At the same time, the diagonal elements of $\hat{\boldsymbol{\Sigma}}$ do not strongly suggest much heteroscedasticity across periods.

None of the desirable properties of the estimators in the random effects model rely on T going to infinity.[22] Indeed, T is likely to be quite small. The estimator of σ_ε^2 is equal to an average of n estimators, each based on the T observations for unit i. [See (11-39a).] Each component in this average is, in principle, consistent. That is, its variance is of order $1/T$ or smaller. Because T is small, this variance may be relatively large. But each term provides some information about the parameter. The average over the n cross-sectional units has a variance of order $1/(nT)$, which will go to zero if n increases, even if we regard T as fixed. The conclusion to draw is that nothing in this treatment relies on T growing large. Although it can be shown that some consistency results will follow for T increasing, the typical panel data set is based on data sets for which it does not make sense to

[22] See Nickell (1981).

TABLE 11.11 Wage Equation Estimated by GLS

Variable	Least Squares Estimate	Clustered Std. Error	Random Effects Ests.	Standard Error	Generalized Least Squares	Standard Error
Constant	5.25112	0.12355	4.04144	0.08330	5.31019	0.07948
Exp	0.04010	0.00408	0.08748	0.00225	0.04478	0.00388
ExpSq	−0.00067	0.00009	−0.00076	0.00005	−0.00071	0.00009
Wks	0.00422	0.00154	0.00096	0.00059	0.00071	0.00055
Occ	−0.14001	0.02724	−0.04322	0.01299	−0.03842	0.01265
Ind	0.04679	0.02366	0.00378	0.01373	0.02671	0.01340
South	−0.05564	0.02616	−0.00825	0.02246	−0.06089	0.02129
SMSA	0.15167	0.02410	−0.02840	0.01616	0.06737	0.01669
MS	0.04845	0.04094	−0.07090	0.01793	−0.02610	0.02020
Union	0.09263	0.02367	0.05835	0.01350	0.03544	0.01316
Ed	0.05670	0.00556	0.10707	0.00511	0.06507	0.00429
Fem	−0.36779	0.04557	−0.30938	0.04554	−0.39606	0.03889
Blk	−0.16694	0.04433	−0.21950	0.05252	−0.15154	0.04262

GLS Estimated Covariance Matrix of ε_i

1	2	3	4	5	6	7
0.1418						
0.1108	0.1036					
0.0821	0.0748	0.1135				
0.0583	0.0579	0.0845	0.1046			
0.0368	0.0418	0.0714	0.0817	0.1008		
0.0152	0.0250	0.0627	0.0799	0.0957	0.1246	
−0.0056	0.0099	0.0585	0.0822	0.1024	0.1259	0.1629

Estimated Covariance Matrix for ε_i Based on Random Effects Model

1	2	3	4	5	6	7
0.1221						
0.0989	0.1221					
0.0989	0.0989	0.1221				
0.0989	0.0989	0.0989	0.1221			
0.0989	0.0989	0.0989	0.0989	0.1221		
0.0989	0.0989	0.0989	0.0989	0.0989	0.1221	
0.0989	0.0989	0.0989	0.0989	0.0989	0.0989	0.1221

assume that T increases without bound or, in some cases, at all.[23] As a general proposition, it is necessary to take some care in devising estimators whose properties hinge on whether T is large or not. The widely used conventional ones we have discussed here do not, but we have not exhausted the possibilities.

[23]In this connection, Chamberlain (1984) provided some innovative treatments of panel data that, in fact, take T as given in the model and that base consistency results solely on n increasing. Some additional results for dynamic models are given by Bhargava and Sargan (1983). Recent research on "bias reduction" in nonlinear panel models, such as Fernandez-Val (2010), do make use of large T approximations in explicitly small T settings.

11.5.6 HAUSMAN'S SPECIFICATION TEST FOR THE RANDOM EFFECTS MODEL

At various points, we have made the distinction between fixed and random effects models. An inevitable question is, which should be used? From a purely practical standpoint, the dummy variable approach is costly in terms of degrees of freedom lost. On the other hand, the fixed effects approach has one considerable virtue. There is little justification for treating the individual effects as uncorrelated with the other regressors, as is assumed in the random effects model. The random effects treatment, therefore, may suffer from the inconsistency due to this correlation between the included variables and the random effect.[24]

The **specification test** devised by Hausman (1978)[25] is used to test for orthogonality of the common effects and the regressors. The test is based on the idea that under the hypothesis of no correlation, both LSDV and FGLS estimators are consistent, but LSDV is inefficient,[26] whereas under the alternative, LSDV is consistent, but FGLS is not. Therefore, under the null hypothesis, the two estimates should not differ systematically, and a test can be based on the difference. The other essential ingredient for the test is the covariance matrix of the difference vector, $[\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE}]$,

$$\text{Var}[\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE}] = \text{Var}[\mathbf{b}_{FE}] + \text{Var}[\hat{\boldsymbol{\beta}}_{RE}] - \text{Cov}[\mathbf{b}_{FE}, \hat{\boldsymbol{\beta}}_{RE}] - \text{Cov}[\hat{\boldsymbol{\beta}}_{RE}, \mathbf{b}_{FE}]. \quad \textbf{(11-43)}$$

Hausman's essential result is that *the covariance of an efficient estimator with its difference from an inefficient estimator is zero*, which implies that

$$\text{Cov}[(\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE}), \hat{\boldsymbol{\beta}}_{RE}] = \text{Cov}[\mathbf{b}_{FE}, \hat{\boldsymbol{\beta}}_{RE}] - \text{Var}[\hat{\boldsymbol{\beta}}_{RE}] = \mathbf{0},$$

or that

$$\text{Cov}[\mathbf{b}_{FE}, \hat{\boldsymbol{\beta}}_{RE}] = \text{Var}[\hat{\boldsymbol{\beta}}_{RE}].$$

Inserting this result in (11-43) produces the required covariance matrix for the test,

$$\text{Var}[\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE}] = \text{Var}[\mathbf{b}_{FE}] - \text{Var}[\hat{\boldsymbol{\beta}}_{RE}] = \boldsymbol{\Psi}.$$

The chi-squared test is based on the Wald criterion,

$$W = \chi^2[K - 1] = [\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE}]'\hat{\boldsymbol{\Psi}}^{-1}[\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE}]. \quad \textbf{(11-44)}$$

For $\hat{\boldsymbol{\Psi}}$, we use the estimated covariance matrices of the slope estimator in the LSDV model and the estimated covariance matrix in the random effects model, excluding the constant term. Under the null hypothesis, W has a limiting chi-squared distribution with $K - 1$ degrees of freedom.

The *Hausman test* is a useful device for determining the preferred specification of the common effects model. As developed here, it has one practical shortcoming. The construction in (11-43) conforms to the theory of the test. However, it does not guarantee that the difference of the two covariance matrices will be positive definite in a finite sample. The implication is that nothing prevents the statistic from being negative when it is computed according to (11-44). One might, in that event, conclude that the random effects model is not rejected, because the similarity of the covariance matrices is what

[24] See Hausman and Taylor (1981) and Chamberlain (1978).

[25] Related results are given by Baltagi (1986).

[26] Referring to the FGLS matrix weighted average given earlier, we see that the efficient weight uses θ, whereas LSDV sets $\theta = 1$.

is causing the problem, and under the alternative (fixed effects) hypothesis, they should be significantly different. There are, however, several alternative methods of computing the statistic for the Hausman test, some asymptotically equivalent and others actually numerically identical. Baltagi (2005, pp. 65–73) provides an extensive analysis. One particularly convenient form of the test finesses the practical problem noted here. An asymptotically equivalent test statistic is given by

$$H' = (\mathbf{b}_{\text{FE}} - \mathbf{b}_{\text{MEANS}})'[\text{Asy.Var}[\mathbf{b}_{\text{FE}}] + \text{Asy.Var}[\mathbf{b}_{\text{MEANS}}]]^{-1} (\mathbf{b}_{\text{FE}} - \mathbf{b}_{\text{MEANS}}) \quad \textbf{(11-45)}$$

where $\mathbf{b}_{\text{MEANS}}$ is the group means estimator discussed in Section 11.3.4. As noted, this is one of several equivalent forms of the test. The advantage of this form is that the covariance matrix will always be nonnegative definite.

Imbens and Wooldridge (2007) have argued that in spite of the practical considerations about the Hausman test in (11-44) and (11-45), the test should be based on robust covariance matrices that do not depend on the assumption of the null hypothesis (the random effects model).[27] Their suggested approach amounts to the variable addition test described in the next section, with a robust covariance matrix.

11.5.7 EXTENDING THE UNOBSERVED EFFECTS MODEL: MUNDLAK'S APPROACH

Even with the Hausman test available, choosing between the fixed and random effects specifications presents a bit of a dilemma. Both specifications have unattractive shortcomings. The fixed effects approach is robust to correlation between the omitted heterogeneity and the regressors, but it proliferates parameters and cannot accommodate time-invariant regressors. The random effects model hinges on an unlikely assumption, that the omitted heterogeneity is uncorrelated with the regressors. Several authors have suggested modifications of the random effects model that would at least partly overcome its deficit. The failure of the random effects approach is that the mean independence assumption, $E[c_i|\mathbf{X}_i] = 0$, is untenable. **Mundlak's approach** (1978) suggests the specification

$$E[c_i|\mathbf{X}_i] = \bar{\mathbf{x}}_{i.}'\boldsymbol{\gamma}.^{28}$$

Substituting this in the random effects model, we obtain

$$\begin{aligned} y_{it} &= \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it} \\ &= \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + \bar{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + \varepsilon_{it} + (c_i - E[c_i|\mathbf{X}_i]) \\ &= \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + \bar{\mathbf{x}}_{i.}'\boldsymbol{\gamma} + \varepsilon_{it} + u_i. \end{aligned} \quad \textbf{(11-46)}$$

This preserves the specification of the random effects model, but (one hopes) deals directly with the problem of correlation of the effects and the regressors. Note that the

[27] That is, "It makes no sense to report a fully robust variance matrix for FE and RE but then to compute a Hausman test that maintains the full set of RE assumptions."

[28] Other analyses, for example, Chamberlain (1982) and Wooldridge (2010), interpret the linear function as the *projection* of c_i on the group means, rather than the conditional mean. The difference is that we need not make any particular assumptions about the conditional mean function while there always exists a linear projection. The conditional mean interpretation does impose an additional assumption on the model but brings considerable simplification. Several authors have analyzed the extension of the model to projection on the full set of individual observations rather than the means. The additional generality provides the bases of several other estimators including minimum distance [Chamberlain (1982)], GMM [Arellano and Bover (1995)], and constrained seemingly unrelated regressions and three-stage least squares [Wooldridge (2010)].

additional terms in $\bar{\mathbf{x}}_{i.}'\boldsymbol{\gamma}$ will only include the time-varying variables—the time-invariant variables are already group means.

Mundlak's approach is frequently used as a compromise between the fixed and random effects models. One side benefit of the specification is that it provides another convenient approach to the Hausman test. As the model is formulated above, the difference between the fixed effects model and the random effects model is the nonzero $\boldsymbol{\gamma}$. As such, a statistical test of the null hypothesis that $\boldsymbol{\gamma}$ equals zero should provide an alternative approach to the two methods suggested earlier. Estimation of (11-46) can be based on either pooled OLS (with a robust covariance matrix) or random effects FGLS. It turns out the coefficient vectors for the two estimators are identical, though the asymptotic covariance matrices will not be. The pooled OLS estimator is fully robust and seems preferable. The test of the null hypothesis that the common effects are uncorrelated with the regressors is then based on a Wald test.

Example 11.12 Hausman and Variable Addition Tests for Fixed versus Random Effects

Using the results in Examples 11.7 (fixed effects) and 11.11 (random effects), we retrieved the coefficient vector and estimated robust asymptotic covariance matrix, $\mathbf{b}_{FE}$ and $\mathbf{V}_{FE}$, from the fixed effects results and the nine elements of $\hat{\boldsymbol{\beta}}_{RE}$ and $\mathbf{V}_{RE}$ (excluding the constant term and the time-invariant variables) from the random effects results. The test statistic is

$$H = (\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE})'[\mathbf{V}_{FE} - \mathbf{V}_{RE}]^{-1}(\mathbf{b}_{FE} - \hat{\boldsymbol{\beta}}_{RE}),$$

The value of the test statistic is 739.374. The critical value from the chi-squared table is 16.919 so the null hypothesis of the random effects model is rejected. There is an additional subtle point to be checked. The difference of the covariance matrices, $\mathbf{V}_{FE} - \mathbf{V}_{RE}$, may not be positive definite. That might not prevent calculation of H if the analyst uses an ordinary inverse in the computation. In that case, a positive statistic might be obtained anyway. The statistic should not be used in this instance. However, that outcome should not lead one to conclude that the correct value for H is zero. The better response is to use the variable addition test we consider next. (For the example here, the smallest characteristic root of the difference matrix was, indeed positive.)

We conclude that the fixed effects model is the preferred specification for these data. This is an unfortunate turn of events, as the main object of the study is the impact of education, which is a time-invariant variable in this sample. We then used the variable addition test instead, based on the regression results in Table 11.12. We recovered the subvector of the estimates at the right in Table 11.12 corresponding to $\boldsymbol{\gamma}$, and the corresponding submatrix of the full covariance matrix. The test statistic is

$$H' = \hat{\boldsymbol{\gamma}}'[\text{Est.Asy.Var}(\hat{\boldsymbol{\gamma}})]^{-1}\,\hat{\boldsymbol{\gamma}}.$$

We obtained a value of 2267.32. This does not change the conclusion, so the null hypothesis of the random effects model is rejected. We conclude as before that the fixed effects estimator is the preferred specification for this model.

11.5.8 EXTENDING THE RANDOM AND FIXED EFFECTS MODELS: CHAMBERLAIN'S APPROACH

The linear unobserved effects model is

$$y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}. \tag{11-47}$$

The *random effects* model assumes that $E[c_i|\mathbf{X}_i] = \alpha$, where the T rows of $\mathbf{X}_i$ are $\mathbf{x}_{it}'$. As we saw in Section 11.5.1, this model can be estimated consistently by ordinary

TABLE 11.12 Wage Equation Estimated by OLS and LSDV

	Pooled OLS		*Augmented Regression*		*Group Means*	
Variable	*Least Squares Estimate*	*Clustered Std. Error*	*Least Squares Estimates*	*Robust Std. Error*	*Least Squares Estimates*	*Robust Std. Error*
R^2	*0.42861*		*0.57518*			
Constant	5.25112	0.12355	5.12143	0.20847	—	—
Exp	0.00401	0.00408	0.11321	0.00406	−0.08131	0.00614
ExpSq	−0.00067	0.00009	−0.00042	0.00008	−0.00015	0.00013
Wks	0.00422	0.00154	0.00084	0.00087	0.00835	0.00361
Occ	−0.14001	0.02724	−0.02148	0.01902	−0.14614	0.03821
Ind	0.04679	0.02366	0.01921	0.02271	0.03871	0.03509
South	−0.05564	0.02616	−0.00186	0.08943	−0.05519	0.09371
SMSA	0.15167	0.02410	−0.04247	0.02953	0.21824	0.03859
MS	0.04845	0.04094	−0.02973	0.02691	0.14451	0.05569
Union	0.09263	0.02367	0.03278	0.02510	0.07628	0.03828
Ed	0.05670	0.00556	0.05144	0.00588	—	—
Fem	−0.36779	0.04557	−0.31706	0.05122	—	—
Blk	−0.16694	0.04433	−0.15780	0.04367	—	—

least squares. Regardless of how ε_{it} is modeled, there is autocorrelation induced by the common, unobserved c_i, so the generalized regression model applies. The random effects formulation is based on the assumption $E[\mathbf{w}_i\mathbf{w}_i'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T + \sigma_u^2\mathbf{i}\mathbf{i}'$, where $w_{it} = (\varepsilon_{it} + u_i)$. We developed the GLS and FGLS estimators for this formulation as well as a strategy for robust estimation of the OLS and LSDV covariance matrices. Among the implications of the development of Section 11.5 is that this formulation of the disturbance covariance matrix is more restrictive than necessary, given the information contained in the data. The assumption that $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T$ assumes that the correlation across periods is equal for all pairs of observations, and arises solely through the persistent c_i. We found some contradictory empirical evidence in Example 11.11—the OLS covariances across periods in the Cornwell and Rupert model do not appear to conform to this specification. In Example 11.11, we estimated the equivalent model with an unrestricted covariance matrix, $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i] = \boldsymbol{\Sigma}$. The implication is that the random effects treatment includes two restrictive assumptions, mean independence, $E[c_i|\mathbf{X}_i] = \alpha$, and homoscedasticity, $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T$. [We do note that dropping the second assumption will cost us the identification of σ_u^2 as an estimable parameter. This makes sense—if the correlation across periods t and s can arise from either their common u_i or from correlation of $(\varepsilon_{it}, \varepsilon_{is})$ then there is no way for us separately to estimate a variance for u_i apart from the covariances of ε_{it} and ε_{is}.] It is useful to note, however, that the panel data model can be viewed and formulated as a seemingly unrelated regressions model with common coefficients in which each period constitutes an equation, Indeed, it is possible, albeit unnecessary, to impose the restriction $E[\mathbf{w}_i\mathbf{w}_i'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T + \sigma_u^2\mathbf{i}\mathbf{i}'$.

The mean independence assumption is the major shortcoming of the random effects model. The central feature of the fixed effects model in Section 11.4 is the possibility that

$E[c_i|\mathbf{X}_i]$ is a nonconstant $h(\mathbf{X}_i)$. As such, least squares regression of y_{it} on $\mathbf{x}_{it}$ produces an inconsistent estimator of $\boldsymbol{\beta}$. The dummy variable model considered in Section 11.4 is the natural alternative. The **fixed effects** approach has the advantage of dispensing with the unlikely assumption that c_i and $\mathbf{x}_{it}$ are uncorrelated. However, it has the shortcoming of requiring estimation of the n parameters, α_i.

Chamberlain (1982, 1984) and Mundlak (1978) suggested alternative approaches that lie between these two. Their modifications of the fixed effects model augment it with the **projections** of c_i on all the rows of $\mathbf{X}_i$ (Chamberlain) or the group means (Mundlak). (See Section 11.5.7.) Consider the first of these, and assume (as it requires) a balanced panel of T observations per group. For purposes of this development, we will assume $T = 3$. The generalization will be obvious at the conclusion. Then, the projection suggested by Chamberlain is

$$c_i = \alpha + \mathbf{x}_{i1}'\boldsymbol{\gamma}_1 + \mathbf{x}_{i2}'\boldsymbol{\gamma}_2 + \mathbf{x}_{i3}'\boldsymbol{\gamma}_3 + r_i, \tag{11-48}$$

where now, by construction, r_i is orthogonal to $\mathbf{x}_{it}$.[29] Insert (11-48) into (11-47) to obtain

$$y_{it} = \alpha + \mathbf{x}_{i1}'\boldsymbol{\gamma}_1 + \mathbf{x}_{i2}'\boldsymbol{\gamma}_2 + \mathbf{x}_{i3}'\boldsymbol{\gamma}_3 + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + r_i.$$

Estimation of the $1 + 3K + K$ parameters of this model presents a number of complications. [We do note that this approach has the potential to (wildly) proliferate parameters. For our quite small regional productivity model in Example 11.22. the original model with six main coefficients plus the treatment of the constants becomes a model with $1 + 6 + 17(6) = 109$ parameters to be estimated.]

If only the n observations for period 1 are used, then the parameter vector,

$$\boldsymbol{\theta}_1 = \alpha, (\boldsymbol{\beta} + \boldsymbol{\gamma}_1), \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3 = \alpha, \boldsymbol{\pi}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \tag{11-49}$$

can be estimated consistently, albeit inefficiently, by ordinary least squares. The model is

$$y_{i1} = \mathbf{z}_{i1}'\boldsymbol{\theta}_1 + w_{i1}, i = 1, \ldots, n.$$

Collecting the n observations, we have

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{w}_1.$$

If, instead, only the n observations from period 2 or period 3 are used, then OLS estimates, in turn,

$$\boldsymbol{\theta}_2 = \boldsymbol{\alpha}, \boldsymbol{\gamma}_1, (\boldsymbol{\beta} + \boldsymbol{\gamma}_2), \boldsymbol{\gamma}_3 = \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\pi}_2, \boldsymbol{\gamma}_3,$$

or

$$\boldsymbol{\theta}_3 = \boldsymbol{\alpha}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, (\boldsymbol{\beta} + \boldsymbol{\gamma}_3) = \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\pi}_3.$$

[29] There are some fine points here that can only be resolved theoretically. If the projection in (11-48) is not the conditional mean, then we have $E[r_i \times \mathbf{x}_{it}] = 0, t = 1, \ldots, T$ but not $E[r_i|\mathbf{X}_i] = 0$. This does not affect the asymptotic properties of the FGLS estimator to be developed here, although it does have implications, for example, for unbiasedness. Consistency will hold regardless. The assumptions behind (11-48) do not include that $\text{Var}[r_i|\mathbf{X}_i]$ is homoscedastic. It might not be. This *could* be investigated empirically. The implication here concerns efficiency, not consistency. The FGLS estimator to be developed here would remain consistent, but a GMM estimator would be more efficient—see Chapter 13. Moreover, without homoscedasticity, it is not certain that the FGLS estimator suggested here is more efficient than OLS (with a robust covariance matrix estimator). Our intent is to begin the investigation here. Further details can be found in Chamberlain (1984) and, for example, Im, Ahn, Schmidt, and Wooldridge (1999).

It remains to reconcile the multiple estimates of the same parameter vectors. In terms of the preceding layouts above, we have the following:

$$\begin{array}{lllll}
\text{OLS Estimates:} & a_1, \mathbf{p}_1, \mathbf{c}_{2,1}, \mathbf{c}_{3,1}, & a_2\, \mathbf{c}_{1,2}, \mathbf{p}_2, \mathbf{c}_{3,2}, & a_3, \mathbf{c}_{1,3}, \mathbf{c}_{2,3}, \mathbf{p}_3; \\
\text{Estimated Parameters:} & \alpha, (\boldsymbol{\beta} + \boldsymbol{\gamma}_1), \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, & \alpha, \boldsymbol{\gamma}_1, (\boldsymbol{\beta} + \boldsymbol{\gamma}_2), \boldsymbol{\gamma}_3, & \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, (\boldsymbol{\beta} + \boldsymbol{\gamma}_3); \\
\text{Structural Parameters:} & \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3. & &
\end{array} \tag{11-50}$$

Chamberlain suggested a minimum distance estimator (MDE). For this problem, the MDE is essentially a weighted average of the several estimators of each part of the parameter vector. We will examine the MDE for this application in more detail in Chapter 13. (For another simpler application of minimum distance estimation that shows the weighting procedure at work, see the reconciliation of four competing estimators of a single parameter at the end of Example 11.23.) There is an alternative way to formulate the estimator that is a bit more transparent. For the first period,

$$\mathbf{y}_1 = \begin{pmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n,1} \end{pmatrix} = \begin{bmatrix} 1 & \mathbf{x}_{1,1} & \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \mathbf{x}_{1,3} \\ 1 & \mathbf{x}_{2,1} & \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \mathbf{x}_{2,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{x}_{n,1} & \mathbf{x}_{n,1} & \mathbf{x}_{n,2} & \mathbf{x}_{n,3} \end{bmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \boldsymbol{\gamma}_3 \end{pmatrix} + \begin{pmatrix} r_{1,1} \\ r_{2,1} \\ \vdots \\ r_{n,1} \end{pmatrix} = \widetilde{\mathbf{X}}_1 \boldsymbol{\theta} + \mathbf{r}_1. \tag{11-51}$$

We treat this as the first equation in a T equation seemingly unrelated regressions model. The second equation, for period 2, is the same (same coefficients), with the data from the second period appearing in the blocks, then likewise for period 3 (and periods 4, . . ., T in the general case). Stacking the data for the T equations (periods), we have

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{X}}_1 \\ \widetilde{\mathbf{X}}_2 \\ \vdots \\ \widetilde{\mathbf{X}}_T \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_T \end{pmatrix} + \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_T \end{pmatrix} = \widetilde{\mathbf{X}} \boldsymbol{\theta} + \mathbf{r}, \tag{11-52}$$

where $E[\widetilde{\mathbf{X}}'\mathbf{r}] = \mathbf{0}$ and (by assumption), $E[\mathbf{r}\mathbf{r}' \,|\, \widetilde{\mathbf{X}}] = \boldsymbol{\Sigma} \otimes \mathbf{I}_n$. With the homoscedasticity assumption for $r_{i,t}$, this is precisely the application in Section 10.2.5. The parameters can be estimated by FGLS as shown in Section 10.2.5.

Example 11.13 Hospital Costs

Carey (1997) examined hospital costs for a sample of 1,733 hospitals observed in five years, 1987–1991. The model estimated is

$$\begin{aligned}
\ln (TC/P)_{it} = {} & \alpha_i + \beta_D\, DIS_{it} + \beta_O\, OPV_{it} + \beta_3\, ALS_{it} + \beta_4\, CM_{it} \\
& + \beta_5\, DIS_{it}^2 + \beta_6\, DIS_{it}^3 + \beta_7\, OPV_{it}^2 + \beta_8\, OPV_{it}^3 \\
& + \beta_9\, ALS_{it}^2 + \beta_{10}\, ALS_{it}^3 + \beta_{11} DIS_{it} \times OPV_{it} \\
& + \beta_{12} FA_{it} + \beta_{13} HI_{it} + \beta_{14} HT_i + \beta_{15} LT_i + \beta_{16}\, Large_i \\
& + \beta_{17}\, Small_i + \beta_{18}\, NonProfit_i + \beta_{19}\, Profit_i \\
& + \varepsilon_{it},
\end{aligned}$$

where

TC	= total cost,
P	= input price index,
DIS	= discharges
OPV	= outpatient visits,
ALS	= average length of stay,
CM	= case mix index,
FA	= fixed assets,
HI	= Hirfindahl index of market concentration at county level,
HT	= dummy variable for high teaching load hospital,
LT	= dummy variable for low teaching load hospital
Large	= dummy variable for large urban area
Small	= dummy variable for small urban area,
Nonprofit	= dummy variable for nonprofit hospital,
Profit	= dummy variable for for-profit hospital.

We have used subscripts "D" and "O" for the coefficients on DIS and OPV as these will be isolated in the following discussion. The model employed in the study is that in (11-47) and (11-48). Initial OLS estimates are obtained for the full cost function in each year. SUR estimates are then obtained using a restricted version of the Chamberlain system. This second step involved a hybrid model that modified (11-49) so that in each period the coefficient vector was

$$\boldsymbol{\theta}_t = [\alpha_t, \beta_{Dt}(\boldsymbol{\gamma}), \beta_{Ot}(\boldsymbol{\gamma}), \beta_{3t}(\boldsymbol{\gamma}), \beta_{4t}(\boldsymbol{\gamma}), \beta_{5t}, \ \ldots, \beta_{19t}],$$

where $\beta_{Dt}(\boldsymbol{\gamma})$ indicates that all five years of the variable (DIS_{it}) are included in the equation, and likewise, for $\beta_{Ot}(\boldsymbol{\gamma})(OPV)$, $\beta_{3t}(\boldsymbol{\gamma})(ALS)$, and $\beta_{4t}(\boldsymbol{\gamma})(CM)$. This is equivalent to using

$$c_i = \alpha + \boldsymbol{\Sigma}_{t=1987}^{1991} (DIS, OPV, ALS, CM)_{it}' \boldsymbol{\gamma}_t + r_i$$

in (11-48).

The unrestricted SUR system estimated at the second step provides multiple estimates of the various model parameters. For example, each of the five equations provides an estimate of ($\beta_5, \ \ldots, \beta_{19}$). The author added one more layer to the model in allowing the coefficients on DIS_{it} and OPV_{it} to vary over time. Therefore, the structural parameters of interest are ($\beta_{D1}, \ \ldots, \beta_{D5}$), ($\gamma_{D1} \ \ldots, \gamma_{D5}$) (the coefficients on DIS) and ($\beta_{O1}, \ \ldots, \beta_{O5}$), ($\gamma_{O1} \ \ldots, \gamma_{O5}$) (the coefficients on OPV). There are, altogether, 20 parameters of interest. The SUR estimates produce, in each year (equation), parameters on DIS for the five years and on OPV for the five years, so there is a total of 50 estimates. Reconciling all of them means imposing a total of 30 restrictions. Table 11.13 shows the relationships for the time-varying parameter on DIS_{it} in the five-equation model. The numerical values reported by the author are shown following the theoretical results. A similar table would apply for the coefficients on OPV, ALS, and CM. (In the latter two, the β coefficient was not assumed to be time varying.) It can be seen in the table, for example, that there are directly four different estimates of $\gamma_{D,87}$ in the second to fifth equations, and likewise for each of the other parameters. Combining the entries in Table 11.13 with the counterparts for the coefficients on OPV, we see 50 SUR/FGLS estimates to be used to estimate 20 underlying parameters. The author used a minimum distance approach to reconcile the different estimates. We will return to this example in Example 13.6, where we will develop the MDE in more detail.

TABLE 11.13 Coefficient Estimates in SUR Model for Hospital Costs

	Coefficient on Variable in the Equation				
Equation	***DIS87***	***DIS88***	***DIS89***	***DIS90***	***DIS91***
SUR87	$\beta_{D,87} + \gamma_{D,87}$ 1.76	$\gamma_{D,88}$ 0.116	$\gamma_{D,89}$ −0.0881	$\gamma_{D,90}$ 0.0570	$\gamma_{D,91}$ −0.0617
SUR88	$\gamma_{D,87}$ 0.254	$\beta_{D,88} + \gamma_{D,88}$ 1.61	$\gamma_{D,89}$ −0.0934	$\gamma_{D,90}$ 0.0610	$\gamma_{D,91}$ −0.0514
SUR89	$\gamma_{D,87}$ 0.217	$\gamma_{D,88}$ 0.0846	$\beta_{D,89} + \gamma_{D,89}$ 1.51	$\gamma_{D,90}$ 0.0454	$\gamma_{D,91}$ −0.0253
SUR90	$\gamma_{D,87}$ 0.179	$\gamma_{D,88}$ 0.0822[a]	$\gamma_{D,89}$ 0.0295	$\beta_{D,90} + \gamma_{D,90}$ 1.57	$\gamma_{D,91}$ 0.0244
SUR91	$\gamma_{D,87}$ 0.153	$\gamma_{D,88}$ 0.0363	$\gamma_{D,89}$ −0.0422	$\gamma_{D,90}$ 0.0813	$\beta_{D,91} + \gamma_{D,91}$ 1.70

[a]The value reported in the published paper is 8.22. The correct value is 0.0822. (Personal communication with the author.)

11.6 NONSPHERICAL DISTURBANCES AND ROBUST COVARIANCE MATRIX ESTIMATION

Because the models considered here are extensions of the classical regression model, we can treat heteroscedasticity in the same way that we did in Chapter 9. That is, we can compute the ordinary or feasible generalized least squares estimators and obtain an appropriate robust covariance matrix estimator, or we can impose some structure on the disturbance variances and use generalized least squares. In the panel data settings, there is greater flexibility for the second of these without making strong assumptions about the nature of the heteroscedasticity.

11.6.1 HETEROSCEDASTICITY IN THE RANDOM EFFECTS MODEL

Because the random effects model is a generalized regression model with a known structure, OLS with a robust estimator of the asymptotic covariance matrix is not the best use of the data. The GLS estimator is efficient whereas the OLS estimator is not. If a perfectly general covariance structure is assumed, then one might simply use Arellano's estimator, described in Section 11.4.3, with a single overall constant term rather than a set of fixed effects. But, within the setting of the random effects model, $\eta_{it} = \varepsilon_{it} + u_i$, allowing the disturbance variance to vary across groups would seem to be a useful extension.

The calculation in (11-33) has a type of heteroscedasticity due to the varying group sizes. The estimator there (and its feasible counterpart) would be the same if, instead of $\theta_i = 1 - \sigma_\varepsilon/(T_i\sigma_u^2 + \sigma_\varepsilon^2)^{1/2}$, the disturbances were specifically heteroscedastic with $E[\varepsilon_{it}{}^2|\mathbf{X}_i] = \sigma_{\varepsilon i}{}^2$ and

$$\theta_i = 1 - \frac{\sigma_{\varepsilon i}}{\sqrt{\sigma_{\varepsilon i}^2 + T_i\sigma_u^2}}.$$

Therefore, for computing the appropriate feasible generalized least squares estimator, once again we need only devise consistent estimators for the variance components and then apply the GLS transformation shown earlier. One possible way to proceed is as follows: Because pooled OLS is still consistent, OLS provides a usable set of residuals. Using the OLS residuals for the specific groups, we would have, for each group,

$$\sigma_{\varepsilon i}^2 + u_i^2 = \frac{\mathbf{e}_i'\mathbf{e}_i}{T}.$$

The residuals from the dummy variable model are purged of the individual specific effect, u_i, so $\sigma_{\varepsilon i}^2$ may be consistently (in T) estimated with

$$\sigma_{\varepsilon i}^2 = \frac{\mathbf{e}_i^{lsdv\prime}\mathbf{e}_i^{lsdv}}{T},$$

where $e_{it}^{lsdv} = y_{it} - \mathbf{x}_{it}'\mathbf{b}^{lsdv} - a_i$. Combining terms, then,

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\mathbf{e}_i^{ols\prime}\mathbf{e}_i^{ols}}{T}\right) - \left(\frac{\mathbf{e}_i^{lsdv\prime}\mathbf{e}_i^{lsdv}}{T}\right)\right] = \frac{1}{n}\sum_{i=1}^{n}(u_i^2).$$

We can now compute the FGLS estimator as before.

11.6.2 AUTOCORRELATION IN PANEL DATA MODELS

As we saw in Section 11.3.2 and Example 11.4, autocorrelation—that is, correlation across the observations in the groups in a panel—is likely to be a substantive feature of the model. Our treatment of the effect there, however, was meant to accommodate autocorrelation in its broadest sense, that is, nonzero covariances across observations in a group. The results there would apply equally to clustered observations, as observed in Section 11.3.3. An important element of that specification was that with clustered data, there might be no obvious structure to the autocorrelation. When the panel data set consists explicitly of groups of time series, and especially if the time series are relatively long as in Example 11.9, one might want to begin to invoke the more detailed, structured time-series models which are discussed in Chapter 20.

11.7 SPATIAL AUTOCORRELATION

The clustering effects suggested in Section 11.3.3 are motivated by an expectation that effects of neighboring locations would spill over into each other, creating a sort of correlation across space, rather than across time as we have focused on thus far. The effect should be common in cross-region studies, such as in agriculture, urban economics, and regional science. Studies of the phenomenon include Case's (1991) study of expenditure patterns, Bell and Bockstael's (2000) study of real estate prices, Baltagi and Li's (2001) analysis of R&D spillovers, Fowler, Cover and Kleit's (2014) study of fringe banking, Klier and McMillen's (2012) analysis of clustering of auto parts suppliers, and Flores-Lagunes and Schnier's (2012) model of cod fishing performance. Models of spatial regression and **spatial autocorrelation** are constructed to formalize this notion.[30]

[30]See Anselin (1988, 2001) for the canonical reference and Le Sage and Pace (2009) for a recent survey.

A model with spatial autocorrelation can be formulated as follows: The regression model takes the familiar panel structure,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + u_i, i = 1, \ldots, n; t = 1, \ldots, T.$$

The common u_i is the usual unit (e.g., country) effect. The correlation across space is implied by the spatial autocorrelation structure,

$$\varepsilon_{it} = \lambda \sum_{j=1}^{n} W_{ij}\varepsilon_{jt} + v_t.$$

The scalar λ is the **spatial autocorrelation coefficient**. The elements W_{ij} are spatial (or **contiguity**) weights that are assumed known. The elements that appear in the sum above are a row of the spatial weight or **contiguity matrix**, $\mathbf{W}$, so that for the n units, we have

$$\boldsymbol{\varepsilon}_t = \lambda \mathbf{W}\boldsymbol{\varepsilon}_t + \mathbf{v}_t, \mathbf{v}_t = v_t\mathbf{i}.$$

The structure of the model is embodied in the symmetric weight matrix, $\mathbf{W}$. Consider for an example counties or states arranged geographically on a grid or some linear scale such as a line from one coast of the country to another. Typically W_{ij} will equal one for i,j pairs that are neighbors and zero otherwise. Alternatively, W_{ij} may reflect distances across space, so that W_{ij} decreases with increases in $|i - j|$. In Flores-Lagunes and Schnier's (2012) study, the spatial weights were inversely proportional to the Euclidean distances between points in a grid. This would be similar to a temporal autocorrelation matrix. Assuming that $|\lambda|$ is less than one, and that the elements of $\mathbf{W}$ are such that $(\mathbf{I} - \lambda\mathbf{W})$ is nonsingular, we may write

$$\boldsymbol{\varepsilon}_t = (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\mathbf{v}_t,$$

so for the n observations at time t,

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\mathbf{v}_t + \mathbf{u}.$$

We further assume that u_i and v_i have zero means, variances σ_u^2 and σ_v^2, and are independent across countries and of each other. It follows that a generalized regression model applies to the n observations at time t,

$$E[\mathbf{y}_t|\mathbf{X}_t] = \mathbf{X}_t\boldsymbol{\beta},$$
$$\text{Var}[\mathbf{y}_t|\mathbf{X}_t] = (\mathbf{I}_n - \lambda\mathbf{W})^{-1}[\sigma_v^2\mathbf{i}\mathbf{i}'](\mathbf{I}_n - \lambda\mathbf{W})^{-1} + \sigma_u^2\mathbf{I}_n.$$

At this point, estimation could proceed along the lines of Chapter 9, save for the need to estimate λ. There is no natural residual-based estimator of λ. Recent treatments of this model have added a normality assumption and employed maximum likelihood methods.[31]

A natural first step in the analysis is a test for spatial effects. The standard procedure for a cross section is Moran's (1950) I statistic, which would be computed for each set of residuals, $\mathbf{e}_t$, using

$$I_t = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}(e_{it} - \bar{e}_t)(e_{jt} - \bar{e}_t)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{i,j}\right)\sum_{i=1}^{n}(e_{it} - \bar{e}_t)^2}. \quad \textbf{(11-53)}$$

[31]The log-likelihood function for this model and numerous references appear in Baltagi (2005, p. 196). Extensive analysis of the estimation problem is given in Bell and Bockstael (2000).

For a panel of T independent sets of observations, $\bar{I} = \frac{1}{T}\sum_{t=1}^{T} I_t$ would use the full set of information. A large sample approximation to the variance of the statistic under the null hypothesis of no spatial autocorrelation is

$$V^2 = \frac{1}{T}\frac{n^2\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}^2 + 3\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\right)^2 - n\sum_{i=1}^{n}\left(\sum_{j=1}^{n}W_{ij}\right)^2}{(n^2-1)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\right)^2}. \quad \textbf{(11-54)}$$

The statistic $\bar{I}/V$ will converge to standard normality under the null hypothesis and can form the basis of the test. (The assumption of independence across time is likely to be dubious at best, however.) Baltagi, Song, and Koh (2003) identify a variety of LM tests based on the assumption of normality. Two that apply to cross-section analysis are[32]

$$LM(1) = \frac{(\mathbf{e}'\mathbf{W}\mathbf{e}/s^2)^2}{tr(\mathbf{W}'\mathbf{W} + \mathbf{W}^2)}$$

for spatial autocorrelation and

$$LM(2) = \frac{(\mathbf{e}'\mathbf{W}\mathbf{y}/s^2)^2}{\mathbf{b}'\mathbf{X}'\mathbf{W}\mathbf{M}\mathbf{W}\mathbf{X}\mathbf{b}/s^2 + tr(\mathbf{W}'\mathbf{W} + \mathbf{W}^2)}$$

for spatially lagged dependent variables, where $\mathbf{e}$ is the vector of OLS residuals, $s^2 = \mathbf{e}'\mathbf{e}/n$, and $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. [33]

Anselin (1988) identifies several possible extensions of the spatial model to dynamic regressions. A "pure space-recursive model" specifies that the autocorrelation pertains to neighbors in the previous period,

$$y_{it} = \gamma[\mathbf{W}\mathbf{y}_{t-1}]_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

A "time-space recursive model" specifies dependence that is purely autoregressive with respect to neighbors in the previous period,

$$y_{it} = \rho y_{i,t-1} + \gamma[\mathbf{W}\mathbf{y}_{t-1}]_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

A "time-space simultaneous" model specifies that the spatial dependence is with respect to neighbors in the current period,

$$y_{it} = \rho y_{i,t-1} + \lambda[\mathbf{W}\mathbf{y}_{t}]_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

Finally, a "time-space dynamic model" specifies that autoregression depends on neighbors in both the current and last period,

$$y_{it} = \rho y_{i,t-1} + \lambda[\mathbf{W}\mathbf{y}_{t}]_i + \gamma[\mathbf{W}\mathbf{y}_{t-1}]_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

Example 11.14 Spatial Autocorrelation in Real Estate Sales

Bell and Bockstael analyzed the problem of modeling spatial autocorrelation in large samples. This is a common problem with GIS (geographic information system) data sets. The central problem is maximization of a likelihood function that involves a sparse matrix, $(\mathbf{I} - \lambda\mathbf{W})$. Direct approaches to the problem can encounter severe inaccuracies in evaluation of the inverse

[32] See Bell and Bockstael (2000, p. 78).

[33] See Anselin and Hudak (1992).

and determinant. Kelejian and Prucha (1999) have developed a moment-based estimator for λ that helps alleviate the problem. Once the estimate of λ is in hand, estimation of the spatial autocorrelation model is done by FGLS. The authors applied the method to analysis of a cross section of 1,000 residential sales in Anne Arundel County, Maryland, from 1993 to 1996. The parcels sold all involved houses built within one year prior to the sale. GIS software was used to measure attributes of interest.

The model is

$$
\begin{aligned}
&+ \beta_2 \ln \textit{Lot size} (\textit{LLT}) \\
&+ \beta_3 \ln \textit{Distance in km to Washington, DC} (\textit{LDC}) \\
&+ \beta_4 \ln \textit{Distance in km to Baltimore} (\textit{LBA}) \\
&+ \beta_5 \% \textit{ land surrounding parcel in publicly owned space} (\textit{POPN}) \\
&+ \beta_6 \% \textit{ land surrounding parcel in natural privately owned space} (\textit{PNAT}) \\
&+ \beta_7 \% \textit{ land surrounding parcel in intensively developed use} (\textit{PDEV}) \\
&+ \beta_8 \% \textit{ land surrounding parcel in low density residential use} (\textit{PLOW}) \\
&+ \beta_9 \textit{ Public sewer service} (1 \textit{ if existing or planned}, 0 \textit{ if not}) (\textit{PSEW}) \\
&+ \varepsilon.
\end{aligned}
$$

(Land surrounding the parcel is all parcels in the GIS data whose centroids are within 500 meters of the transacted parcel.) For the full model, the specification is

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\
\boldsymbol{\varepsilon} &= \lambda \mathbf{W}\boldsymbol{\varepsilon} + \mathbf{v}.
\end{aligned}
$$

The authors defined four contiguity matrices:

W1: $W_{ij} = 1/\text{distance between } i \text{ and } j \text{ if distance} < 600$ meters, 0 otherwise,
W2: $W_{ij} = 1$ if distance between i and $j < 200$ meters, 0 otherwise,
W3: $W_{ij} = 1$ if distance between i and $j < 400$ meters, 0 otherwise,
W4: $W_{ij} = 1$ if distance between i and $j < 600$ meters, 0 otherwise.

All contiguity matrices were row-standardized. That is, elements in each row are scaled so that the row sums to one. One of the objectives of the study was to examine the impact of row standardization on the estimation. It is done to improve the numerical stability of the optimization process. Because the estimates depend numerically on the normalization, it is not completely innocent.

Test statistics for spatial autocorrelation based on the OLS residuals are shown in Table 11.14. (These are taken from the authors' Table 3.) The Moran statistics are distributed as standard normal while the LM statistics are distributed as chi-squared with one degree of freedom. All but the LM(2) statistic for W3 are larger than the 99 percent critical value from the respective table, so we would conclude that there is evidence of spatial autocorrelation. Estimates from some of the regressions are shown in Table 11.15. In the remaining results in the study, the authors find that the outcomes are somewhat sensitive to the specification of the spatial weight matrix, but not particularly so to the method of estimating λ.

TABLE 11.14 Test Statistics for Spatial Autocorrelation

	W1	*W2*	*W3*	*W4*
Moran's *I*	7.89	9.67	13.66	6.88
LM(1)	49.95	84.93	156.48	36.46
LM(2)	7.40	17.22	2.33	7.42

TABLE 11.15 Estimated Spatial Regression Models

	OLS		*FGLS*[a]		*Spatial Based on W1 ML*		*Spatial Based on W1 Gen. Moments*	
Parameter	***Estimate***	***Std.Err.***	***Estimate***	***Std.Err.***	***Estimate***	***Std.Err.***	***Estimate***	***Std.Err.***
α	4.7332	0.2047	4.7380	0.2048	5.1277	0.2204	5.0648	0.2169
β_1	0.6926	0.0124	0.6924	0.0214	0.6537	0.0135	0.6638	0.0132
β_2	0.0079	0.0052	0.0078	0.0052	0.0002	0.0052	0.0020	0.0053
β_3	−0.1494	0.0195	−0.1501	0.0195	−0.1774	0.0245	−0.1691	0.0230
β_4	−0.0453	0.0114	−0.0455	0.0114	−0.0169	0.0156	−0.0278	0.0143
β_5	−0.0493	0.0408	−0.0484	0.0408	−0.0149	0.0414	−0.0269	0.0413
β_6	0.0799	0.0177	0.0800	0.0177	0.0586	0.0213	0.0644	0.0204
β_7	0.0677	0.0180	0.0680	0.0180	0.0253	0.0221	0.0394	0.0211
β_8	−0.0166	0.0194	−0.0168	0.0194	−0.0374	0.0224	−0.0313	0.0215
β_9	−0.1187	0.0173	−0.1192	0.0174	−0.0828	0.0180	−0.0939	0.0179
λ	—	—	—	—	0.4582	0.0454	0.3517	—

[a]The authors report using a heteroscedasticity model $\sigma_i^2 \times f(LIV_i, LIV_i^2)$. The function $f(.)$ is not identified.

Example 11.15 Spatial Lags in Health Expenditures

Moscone, Knapp, and Tosetti (2007) investigated the determinants of mental health expenditure over six years in 148 British local authorities using two forms of the spatial correlation model to incorporate possible interaction among authorities as well as unobserved spatial heterogeneity. The models estimated, in addition to pooled regression and a random effects model, were as follows. The first is a model with **spatial lags**,

$$\mathbf{y}_t = \gamma_t \mathbf{i} + \rho \mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{u}$ is a 148×1 vector of random effects and $\mathbf{i}$ is a 148×1 column of ones. For each local authority,

$$y_{it} = \gamma_t + \rho(\mathbf{w}_i'\mathbf{y}_t) + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it},$$

where $\mathbf{w}_i'$ is the ith row of the contiguity matrix, $\mathbf{W}$. Contiguities were defined in $\mathbf{W}$ as one if the locality shared a border or vertex and zero otherwise. (The authors also experimented with other contiguity matrices based on "sociodemographic" differences.) The second model estimated is of **spatial error correlation,**

$$\mathbf{y}_t = \gamma_t \mathbf{i} + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}_t,$$
$$\boldsymbol{\varepsilon}_t = \lambda \mathbf{W}\boldsymbol{\varepsilon}_t + \mathbf{v}_t.$$

For each local authority, this model implies

$$y_{it} = \gamma_t + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \lambda \Sigma_j w_{ij}\varepsilon_{jt} + v_{it}.$$

The authors use maximum likelihood to estimate the parameters of the model. To simplify the computations, they note that the maximization can be done using a two-step procedure. As we have seen in other applications, when $\boldsymbol{\Omega}$ in a generalized regression model is known, the appropriate estimator is GLS. For both of these models, with known spatial autocorrelation parameter, a GLS transformation of the data produces a classical regression model. [See (9-11).] The method used is to iterate back and forth between simple OLS estimation of γ_t, $\boldsymbol{\beta}$, and σ_ε^2 and maximization of the concentrated log-likelihood function which, given the other estimates, is a function of the spatial autocorrelation parameter, ρ or λ, and the variance of the heterogeneity, σ_u^2.

The dependent variable in the models is the log of per capita mental health expenditures. The covariates are the percentage of males and of people under 20 in the area, average mortgage rates, numbers of unemployment claims, employment, average house price, median weekly wage, percent of single parent households, dummy variables for Labour party or Liberal Democrat party authorities, and the density of population ("to control for supply-side factors"). The estimated spatial autocorrelation coefficients for the two models are 0.1579 and 0.1220, both more than twice as large as the estimated standard error. Based on the simple Wald tests, the hypothesis of no spatial correlation would be rejected. The log-likelihood values for the two spatial models were +206.3 and +202.8, compared to −211.1 for the model with no spatial effects or region effects, so the results seem to favor the spatial models based on a chi-squared test statistic (with one degree of freedom) of twice the difference. However, there is an ambiguity in this result as the improved "fit" could be due to the region effects rather than the spatial effects. A simple random effects model shows a log-likelihood value of +202.3, which bears this out. Measured against this value, the spatial lag model seems the preferred specification, whereas the spatial autocorrelation model does not add significantly to the log-likelihood function compared to the basic random effects model.

11.8 ENDOGENEITY

Recent **panel data** applications have relied heavily on the methods of instrumental variables. We will develop some of this methodology in detail in Chapter 13 where we consider generalized method of moments (GMM) estimation. At this point, we can examine three major building blocks in this set of methods, a panel data counterpart to two-stage least squares developed in Chapter 8, Hausman and Taylor's (1981) estimator for the random effects model and Bhargava and Sargan's (1983) proposals for estimating a dynamic panel data model. These tools play a significant role in the GMM estimators of dynamic panel models in Chapter 13.

11.8.1 INSTRUMENTAL VARIABLE ESTIMATION

The exogeneity assumption, $E[\mathbf{x}_{it}\varepsilon_{it}] = \mathbf{0}$, has been essential to the estimation strategies suggested thus far. For the generalized regression model (random effects), it was necessary to strengthen this to strict exogeneity, $E[\mathbf{x}_{it}\varepsilon_{is}] = \mathbf{0}$ for all t,s for given i. If these assumptions are not met, then $\mathbf{x}_{it}$ is endogenous in the model, and typically an instrumental variable approach to consistent estimation would be called for.

The fixed effects case is simpler, and can be based entirely on results we have already obtained. The model is $y_{it} = c_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$. We assume there is a set of $L \geq K$ instrumental variables, $\mathbf{z}_{it}$. The set of instrumental variables must be exogenous, that is, orthogonal to ε_{it}; the minimal assumption is $E[\mathbf{z}_{it}\varepsilon_{it}] = \mathbf{0}$. (It will turn out, at least initially, to be immaterial to estimation of $\boldsymbol{\beta}$ whether $E[\mathbf{z}_{it}c_i] = \mathbf{0}$, though one would expect it would be.) Then, the model in deviation form,

$$\begin{aligned} y_{it} - \bar{y}_{i.} &= (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_{i.}) \\ \ddot{y} &= \ddot{\mathbf{x}}_{it}'\boldsymbol{\beta} + \ddot{\varepsilon}_{it}, \end{aligned}$$

is amenable to 2SLS. The IV estimator can be written

$$\mathbf{b}_{IV,FE} = (\ddot{\mathbf{X}}'\ddot{\mathbf{Z}}(\ddot{\mathbf{Z}}'\ddot{\mathbf{Z}})^{-1}\ddot{\mathbf{Z}}'\ddot{\mathbf{X}})^{-1}(\ddot{\mathbf{X}}'\ddot{\mathbf{Z}}(\ddot{\mathbf{Z}}'\ddot{\mathbf{Z}})^{-1}\ddot{\mathbf{Z}}'\ddot{\mathbf{y}}).$$

We can see from this expression that this computation will break down if $\mathbf{Z}$ contains any time-invariant variables. Clearly if there are, then the corresponding columns in $\ddot{\mathbf{Z}}$ will be zero. But, even if $\mathbf{Z}$ is not transformed, columns of $\ddot{\mathbf{X}}'\mathbf{Z}$ will still turn to zeros because $\ddot{\mathbf{X}}'\mathbf{Z} = \mathbf{X}'\ddot{\mathbf{Z}}$ ($\mathbf{M}^0$ is idempotent). Assuming, then, that $\mathbf{Z}$ is also transformed to deviations from group means, the 2SLS estimator is

$$\begin{aligned}\mathbf{b}_{IV,FE} &= \left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{Z}}_i(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{Z}}_i)^{-1}\ddot{\mathbf{Z}}_i'\ddot{\mathbf{X}}_i\right]^{-1}\left[\sum_{i=1}^{n}\ddot{\mathbf{X}}_i'\ddot{\mathbf{Z}}_i(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{Z}}_i)^{-1}\ddot{\mathbf{Z}}_i'\ddot{\mathbf{y}}_i\right] \\ &= \left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\hat{\ddot{\mathbf{X}}}_i\right]^{-1}\left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\ddot{\mathbf{y}}_i\right].\end{aligned} \tag{11-55}$$

For computing the asymptotic covariance matrix, without correction, we would use

$$\text{Est.Asy.Var}[\mathbf{b}_{IV,FE}] = \hat{\sigma}_\varepsilon^2\left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\hat{\ddot{\mathbf{X}}}_i\right]^{-1}$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{t}(\ddot{y}_{it} - \ddot{\mathbf{X}}_{it}'\mathbf{b}_{IV,FE})^2}{n(T-1) - K}. \tag{11-56}$$

An asymptotic covariance matrix that is robust to heteroscedasticity and autocorrelation is

$$\text{Est.Asy.Var}[\mathbf{b}_{IV,FE}] = \left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\hat{\ddot{\mathbf{X}}}_i\right]^{-1}\left[\sum_{i=1}^{n}\left(\hat{\ddot{\mathbf{X}}}_i'\ddot{\mathbf{e}}_i\right)\left(\ddot{\mathbf{e}}_i'\hat{\ddot{\mathbf{X}}}_i\right)\right]\left[\sum_{i=1}^{n}\hat{\ddot{\mathbf{X}}}_i'\hat{\ddot{\mathbf{X}}}_i\right]^{-1}. \tag{11-57}$$

The procedure would be similar for the random effects model, but would (as before) require a first step to estimate the variances of ε and u. The steps follow the earlier prescription:

1. Use pooled 2SLS to compute $\hat{\boldsymbol{\beta}}_{IV,Pooled}$ and obtain residuals $\mathbf{w}$. The estimator of $\sigma_\varepsilon^2 + \sigma_u^2$ is $\mathbf{w}'\mathbf{w}/(nT\text{-}K)$. Use FE 2SLS as described above to obtain $\mathbf{b}_{\text{IV,FE}}$, then use (11-56) to estimate ${\sigma_\varepsilon}^2$. Use these two estimators to compute the estimator of ${\sigma_u}^2$, then $\boldsymbol{\Sigma}^{-1} = (1/\sigma_\varepsilon^2)[\mathbf{I}_T - (\theta(2-\theta)/T)\mathbf{i}\mathbf{i}']$. [The result for $\boldsymbol{\Sigma}^{-1/2}$ is given in (11-33).]
2. Use IV for the generalized regression model,

$$\hat{\boldsymbol{\beta}}_{IV,RE} = \left[\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i'\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right]. \tag{11-58}$$

3. The estimator for the asymptotic covariance matrix is the bracketed inverse. A robust covariance matrix is computed with

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}_{IV,RE}] =$$

$$\mathbf{A}^{-1}\left[\sum_{i=1}^{n}(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\varepsilon}}_i)(\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\varepsilon}}_i)'\right]\mathbf{A}^{-1}$$

$$\mathbf{A} = \left[\sum_{i=1}^{n}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_i(\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_1)^{-1})\mathbf{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_i\right]. \tag{11-59}$$

TABLE 11.16 Estimated Health Satisfaction Equations (Robust standard errors in parentheses)

Variable	*OLS*	*2SLS*	*FE*	*RE*	*FE/2SLS*	*RE/2SLS*
Constant	9.17989	10.7061	—	9.69595	—	12.1185
	(0.36704)	(0.36931)	—	(0.28573)	—	(0.75062)
ln *Income*	0.18045	1.16373	0.13957	0.13001	0.99046	1.24378
	(0.10931)	(0.20863)	(0.10246)	(0.06970)	(0.48337)	(0.33140)
Working	0.63475	0.34196	0.12963	0.29491	−0.05739	0.00243
	(0.12705)	(0.09007)	(0.11656)	(0.07392)	(0.15171)	(0.12932)
Public	−0.78176	−0.52551	−0.20282	−0.48854	−0.15991	−0.29334
	(0.15438)	(0.10963)	(0.17409)	(0.12775)	(0.16779)	(0.13964)
Add On	0.18664	−0.06131	−0.03252	0.04340	−0.01482	−0.02720
	(0.29279)	(0.24477)	(0.17287)	(0.21060)	(0.16327)	(0.15847)
Age	−0.04606	−0.05523	−0.07178	−0.05926	−0.10419	−0.08409
	(0.00583)	(0.00369)	(0.00900)	(0.00468)	(0.01992)	(0.00882)
σ_ε	2.17305	2.21080	1.57382	2.47692	1.59032	2.57864
σ_u	—	—	—	1.49841	—	1.53728

Example 11.16 Endogenous Income in a Health Production Model

In Example 10.8, we examined a health outcome, health satisfaction, in a two-equation model,

$$\begin{aligned} \textit{Health Satisfaction} &= \alpha_1 + \gamma_1 \ln \textit{Income} + \alpha_2 \textit{Female} + \alpha_3 \textit{Working} + \alpha_4 \textit{Public} + \alpha_5 \textit{AddOn} \\ &\quad + \alpha_6 \textit{Age} + \varepsilon_H, \\ \ln \textit{Income} &= \beta_1 + \gamma_2 \textit{Health Satisfaction} + \beta_2 \textit{Female} + \beta_3 \textit{Education} + \beta_4 \textit{Married} \\ &\quad + \beta_5 \textit{HHKids} + \beta_6 \textit{Age} + \varepsilon_I. \end{aligned}$$

The data are an unbalanced panel of 7,293 households. For simplicity, we will focus on the balanced panel of 887 households that were present for all 7 waves. The variable ln *Income* is endogenous in the health equation. There is also a time-invariant variable, *Female*, in the equation that will have to be dropped in this application as we are going to fit a fixed effects model. The instrumental variables are the constant, *Working*, *Public*, *AddOn*, *Age*, *Education*, *Married,* and *HHKids*. Table 11.16 presents the OLS, 2SLS, FE, RE, FE2SLS, and RE2SLS estimates for the health satisfaction equation. Robust standard errors are reported for each case. There is a clear pattern in the results; the instrumental variable estimates of the coefficient on ln Income are 7 to 10 times as large as the least squares estimates, and the estimated standard errors increase comparably.

11.8.2 HAUSMAN AND TAYLOR'S INSTRUMENTAL VARIABLES ESTIMATOR

Recall the original specification of the linear model for panel data in (11-1),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{it}. \tag{11-60}$$

The random effects model is based on the assumption that the unobserved person-specific effects, $\mathbf{z}_i$, are uncorrelated with the included variables, $\mathbf{x}_{it}$. This assumption is a major shortcoming of the model. However, the random effects treatment does allow the model to contain observed time-invariant characteristics, such as demographic characteristics, while the fixed effects model does not—if present, they are simply absorbed into the fixed effects. **Hausman and Taylor's** (1981) **estimator** for the random effects model suggests a way to overcome the first of these while accommodating the second.

Their model is of the form

$$y_{it} = \mathbf{x}_{1it}'\boldsymbol{\beta}_1 + \mathbf{x}_{2it}'\boldsymbol{\beta}_2 + \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 + z_{2i}'\boldsymbol{\alpha}_2 + \varepsilon_{it} + u_i,$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')'$. In this formulation, all individual effects denoted $\mathbf{z}_i$ are observed. As before, unobserved individual effects that are contained in $\mathbf{z}_i'\boldsymbol{\alpha}$ in (11-60) are contained in the person-specific random term, u_i. Hausman and Taylor define four sets of *observed* variables in the model:

$\mathbf{x}_{1it}$ is K_1 variables that are time varying and uncorrelated with u_i,
$\mathbf{z}_{1i}$ is L_1 variables that are time invariant and uncorrelated with u_i,
$\mathbf{x}_{2it}$ is K_2 variables that are time varying and are correlated with u_i,
$\mathbf{z}_{2i}$ is L_2 variables that are time invariant and are correlated with u_i.

The assumptions about the random terms in the model are

$$E[u_i|\mathbf{x}_{1it}, \mathbf{z}_{1i}] = 0 \text{ though } E[u_i|\mathbf{x}_{2it}, \mathbf{z}_{2i}] \neq 0,$$
$$\text{Var}[u_i|\mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \sigma_u^2,$$
$$\text{Cov}[\varepsilon_{it}, u_i|\mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = 0,$$
$$\text{Var}[\varepsilon_{it}+u_i|\mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \sigma^2 = \sigma_\varepsilon^2 + \sigma_u^2,$$
$$\text{Corr}[\varepsilon_{it} + u_i, \varepsilon_{is} + u_i|\mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] = \rho = \sigma_u^2/\sigma^2.$$

Note the crucial assumption that one can distinguish sets of variables $\mathbf{x}_1$ and $\mathbf{z}_1$ that are uncorrelated with u_i from $\mathbf{x}_2$ and $\mathbf{z}_2$ which are not. The likely presence of $\mathbf{x}_2$ and $\mathbf{z}_2$ is what complicates specification and estimation of the random effects model in the first place.

By construction, any OLS or GLS estimators of this model are inconsistent when the model contains variables that are correlated with the random effects. Hausman and Taylor have proposed an instrumental variables estimator that uses only the information within the model (i.e., as already stated). The strategy for estimation is based on the following logic: First, by taking deviations from group means, we find that

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i.})'\boldsymbol{\beta}_1 + (\mathbf{x}_{2it} - \bar{\mathbf{x}}_{2i.})'\boldsymbol{\beta}_2 + \varepsilon_{it} - \bar{\varepsilon}_{i.}, \quad \textbf{(11-61)}$$

which implies that both parts of $\boldsymbol{\beta}$ can be consistently estimated by least squares, in spite of the correlation between $\mathbf{x}_2$ and u. This is the familiar, fixed effects, least squares dummy variable estimator—the transformation to deviations from group means removes from the model the part of the disturbance that is correlated with $\mathbf{x}_{2it}$. In the original model, Hausman and Taylor show that the group mean deviations can be used as $(K_1 + K_2)$ instrumental variables for estimation of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. That is the implication of (11-61). Because $\mathbf{z}_1$ is uncorrelated with the disturbances, it can likewise serve as a set of L_1 instrumental variables. That leaves a necessity for L_2 instrumental variables. The authors show that the group means for $\mathbf{x}_1$ can serve as these remaining instruments, and the model will be identified so long as K_1 is greater than or equal to L_2. For identification purposes, then, K_1 must be at least as large as L_2. As usual, **feasible GLS** is better than OLS, and available. Likewise, FGLS is an improvement over simple instrumental variable estimation of the model, which is consistent but inefficient.

The authors propose the following set of steps for consistent and efficient estimation:

Step 1. Obtain the LSDV (fixed effects) estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ based on $\mathbf{x}_1$ and $\mathbf{x}_2$. The residual variance estimator from this step is a consistent estimator of σ_ε^2.

Step 2. Form the within-groups residuals, e_{it}, from the LSDV regression at step 1. Stack the group means of these residuals in a full-sample-length data vector. Thus, $e_{it}^* = \bar{e}_{i.} = \dfrac{1}{T}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\mathbf{b}_w)$, $t = 1, \ldots, T$, $i = 1, \ldots, n$. (The individual constant term, a_i, is not included in e_{it}^*.) (Note, from (11-16b), $e_{it}^* = \bar{e}_{i.}$ is a_i, the ith constant term.) These group means are used as the dependent variable in an instrumental variable regression on $\mathbf{z}_1$ and $\mathbf{z}_2$ with instrumental variables $\mathbf{z}_1$ and $\mathbf{x}_1$. (Note the identification requirement that K_1, the number of variables in $\mathbf{x}_1$, be at least as large as L_2, the number of variables in $\mathbf{z}_2$.) The time-invariant variables are each repeated T times in the data matrices in this regression. This provides a consistent estimator of $\boldsymbol{\alpha}$.

Step 3. The residual variance in the regression in step 2 is a consistent estimator of $\sigma_*^2 = \sigma_u^2 + \sigma_\varepsilon^2/T$. From this estimator and the estimator of σ_ε^2 in step 1, we deduce an estimator of $\sigma_u^2 = \sigma_*^2 - \sigma_\varepsilon^2/T$. We then form the weight for feasible GLS in this model by forming the estimate of

$$\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

Step 4. The final step is a weighted instrumental variable estimator. Let the full set of variables in the model be

$$\mathbf{w}_{it}' = (\mathbf{x}_{1it}', \mathbf{x}_{2it}', \mathbf{z}_{1i}', \mathbf{z}_{2i}').$$

Collect these nT observations in the rows of data matrix $\mathbf{W}$. The transformed variables for GLS are, as before when we first fit the random effects model,

$$\mathbf{w}_{it}^{*\prime} = \mathbf{w}_{it}' - \hat{\theta}\bar{\mathbf{w}}_i'. \quad \text{and} \quad y_{it}^* = y_{it} - \hat{\theta}\bar{y}_{i.},$$

where $\hat{\theta}$ denotes the sample estimate of θ. The transformed data are collected in the rows data matrix $\mathbf{W}^*$ and in column vector $\mathbf{y}^*$. Note in the case of the time-invariant variables in $\mathbf{w}_{it}$, the group mean is the original variable, and the transformation just multiplies the variable by $1 - \hat{\theta}$. The instrumental variables are

$$\mathbf{v}_{it}' = [(\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i.})', (\mathbf{x}_{2it} - \bar{\mathbf{x}}_{2i.})', \mathbf{z}_{1i}'\,\bar{\mathbf{x}}_{1i.}].$$

These are stacked in the rows of the $nT \times (K_1 + K_2 + L_1 + K_1)$ matrix $\mathbf{V}$. Note for the third and fourth sets of instruments, the time-invariant variables and group means are repeated for each member of the group. The instrumental variable estimator would be

$$(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}')_{\text{IV}}' = [(\mathbf{W}^{*\prime}\mathbf{V})(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{W}^*)]^{-1}[(\mathbf{W}^{*\prime}\mathbf{V})(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{y}^*)].^{34} \qquad \textbf{(11-62)}$$

The instrumental variable estimator is consistent if the data are not weighted, that is, if $\mathbf{W}$ rather than $\mathbf{W}^*$ is used in the computation. But this is inefficient, in the same way that OLS is consistent but inefficient in estimation of the simpler random effects model.

[34]Note that the FGLS random effects estimator would be $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}')_{RE}' = [\mathbf{W}^{*\prime}\mathbf{W}^*]^{-1}\mathbf{W}^{*\prime}\mathbf{y}^*$.

Example 11.17 The Returns to Schooling

The economic returns to schooling have been a frequent topic of study by econometricians. The PSID and NLS data sets have provided a rich source of panel data for this effort. In wage (or log wage) equations, it is clear that the economic benefits of schooling are correlated with latent, unmeasured characteristics of the individual such as innate ability, intelligence, drive, or perseverance. As such, there is little question that simple random effects models based on panel data will suffer from the effects noted earlier. The fixed effects model is the obvious alternative, but these rich data sets contain many useful variables, such as race, union membership, and marital status, which are generally time invariant. Worse yet, the variable most of interest, years of schooling, is also time invariant. Hausman and Taylor (1981) proposed the estimator described here as a solution to these problems. The authors studied the effect of schooling on (the log of) wages using a random sample from the PSID of 750 men aged 25 to 55, observed in two years, 1968 and 1972. The two years were chosen so as to minimize the effect of serial correlation apart from the persistent unmeasured individual effects. The variables used in their model were as follows:

Experience	= age − years of schooling − 5,
Years of schooling	= continuous variable
Bad Health	= a dummy variable indicating general health,
Race	= adummy variable indicating nonwhite (70 of 750 observations),
Union	= a dummy variable indicating union membership,
Unemployed	= a dummy variable indicating previous year's unemployment.

The model also included a constant term and a period indicator.[35]

The primary focus of the study is the coefficient on schooling in the log wage equation. Because Schooling and, probably, Experience and Unemployed, are correlated with the latent effect, there is likely to be serious bias in conventional estimates of this equation. Table 11.17 reports some of their reported results. The OLS and random effects GLS results in the first two columns provide the benchmark for the rest of the study. The schooling coefficient is estimated at 0.0669, a value which the authors suspected was far too small. As we saw earlier, even in the presence of correlation between measured and latent effects, in this model, the LSDV estimator provides a consistent estimator of the coefficients on the time-varying variables. Therefore, we can use it in the **Hausman specification test** for correlation between the included variables and the latent heterogeneity. The calculations are shown in Section 11.5.5, result (11-44). Because there are three variables remaining in the LSDV equation, the chi-squared statistic has three degrees of freedom. The reported value of 20.2 is far larger than the 95% critical value of 7.81, so the results suggest that the random effects model is misspecified.

Hausman and Taylor proceeded to reestimate the log wage equation using their proposed estimator. The fourth and fifth sets of results in Table 11.17 present the instrumental variable estimates. The specification test given with the fourth set of results suggests that the procedure has produced the expected result. The hypothesis of the modified random effects model is now not rejected; the chi-squared value of 2.24 is much smaller than the critical value. The schooling variable is treated as endogenous (correlated with u_i) in both cases. The difference between the two is the treatment of Unemployed and Experience. In the preferred equation, they are included in $\mathbf{x}_2$ rather than $\mathbf{x}_1$. The end result of the exercise is, again, the coefficient on schooling, which has risen from 0.0669 in the worst specification (OLS) to 0.2169 in the last one, an increase of over 200 %. As the authors note, at the same time, the measured effect of race nearly vanishes.

[35]The coding of the latter is not given, but any two distinct values, including 0 for 1968 and 1 for 1972, would produce identical results. (Why?)

TABLE 11.17 Estimated Log Wage Equations

	Variables	*OLS*	*GLS/RE*	*LSDV*	*HT/IV-GLS*	*HT/IV-GLS*
$\mathbf{x}_1$	*Experience*	0.0132	0.0133	0.0241	0.0217	
		(0.0011)[a]	(0.0017)	(0.0042)	(0.0031)	
	Bad health	−0.0843	−0.0300	−0.0388	−0.0278	−0.0388
		(0.0412)	(0.0363)	(0.0460)	(0.0307)	(0.0348)
	Unemployed	−0.0015	−0.0402	−0.0560	−0.0559	
	Last Year	(0.0267)	(0.0207)	(0.0295)	(0.0246)	
	Time	NR[b]	NR	NR	NR	NR
$\mathbf{x}_2$	*Experience*					0.0241
						(0.0045)
	Unemployed					−0.0560
						(0.0279)
$\mathbf{z}_1$	*Race*	−0.0853	−0.0878		−0.0278	−0.0175
		(0.0328)	(0.0518)		(0.0752)	(0.0764)
	Union	0.0450	0.0374		0.1227	0.2240
		(0.0191)	(0.0296)		(0.0473)	(0.2863)
	Schooling	**0.0669**	**0.0676**			
		(0.0033)	(0.0052)			
	Constant	NR	NR	NR	NR	NR
$\mathbf{z}_2$	*Schooling*				**0.1246**	**0.2169**
					(0.0434)	(0.0979)
	σ_ε	0.321	0.192	0.160	0.190	0.629
	$\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2)$		0.632		0.661	0.817
	Spec. Test [3]		20.2		2.24	0.00

[a]Estimated asymptotic standard errors are given in parentheses.

[b]NR indicates that the coefficient estimate was not reported in the study.

Example 11.18 The Returns to Schooling

In Example 11.17, Hausman and Taylor find that the estimated effect of education in a wage equation increases substantially (nearly doubles from 0.0676 to 0.1246) when it is treated as endogenous in a random effects model, then increases again by 75% to 0.2169 when experience and unemployment status are also treated as endogenous. In this exercise, we will examine whether these results reappear in Cornwell and Rupert's application. (We do not have the unemployment indicator.) Three sets of least squares results, ordinary, fixed effects, and feasible GLS random effects, appear at the left of Table 11.18. The education effect in the RE model is about 11%. (Time-invariant education falls out of the fixed effects model.) The effect increases by 29% to 13.8% when education is treated as endogenous, which is similar to Hausman and Taylor's 12.5%. When experience is treated as exogenous, instead, the education effect rises again by 72%. (The second such increase in the Hausman/Taylor results resulted from treating experience as endogenous, not exogenous.)

11.8.3 CONSISTENT ESTIMATION OF DYNAMIC PANEL DATA MODELS: ANDERSON AND HSIAO'S IV ESTIMATOR

Consider a heterogeneous dynamic panel data model,

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad \textbf{(11-63)}$$

TABLE 11.18 Hausman–Taylor Estimates of Wage Equation

	OLS	*LGLS/RE*	*FE*	*HT-RE/FGLS*		
				x_1 = *Exogenous Time Varying*		
OCC	−0.14001	−0.04322	−0.02148	−0.02004	−0.02070	−0.01445
South	−0.05564	−0.00825	−0.00186	0.00821	0.00746	0.01512
SMSA	0.15167	−0.02840	−0.04247	−0.04227	−0.04183	−0.05219
IND	0.04679	0.00378	0.01921	0.01392	0.01359	0.01971
Exp	0.04010	0.08748	0.11321			0.10919
Expsq	−0.00067	−0.00076	−0.00042			−0.00048
				x_2 = *Endogenous Time Varying*		
Exp				0.11313	0.11313	
ExpSq				−0.00042	−0.00042	
WKS	0.00422	0.00096	0.00084	0.00084	0.00084	0.00080
MS	0.04845	−0.07090	−0.02973	−0.02980	−0.02985	−0.03850
Union	0.09263	0.05835	0.03278	0.03293	0.03277	0.03773
				f_1 = *Exogenous Time Invariant*		
Constant	5.25112	4.04144		2.82907	2.91273	1.74978
FEM	−0.36779	−0.30938		−0.13209	−0.13093	−0.18008
Blk	−0.16694	−0.21950		−0.27726	−0.28575	−0.13633
Education	**0.05670**	**0.10707**		**0.14440**		
				f_2 = *Endogenous Time Invariant*		
Education					**0.13794**	**0.23726**
σ_ε	0.34936	0.15206	0.15206	0.15199	0.15199	0.15199
σ_u	—	0.31453		0.94179	0.94180	0.99443

where c_i is, as in the preceding sections of this chapter, individual unmeasured heterogeneity, that may or may not be correlated with $\mathbf{x}_{it}$. We consider methods of estimation for this model when T is fixed and relatively small, and n may be large and increasing.

Pooled OLS is obviously inconsistent. Rewrite (11-63) as

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + w_{it}.$$

The disturbance in this pooled regression may be correlated with $\mathbf{x}_{it}$, but either way, it is surely correlated with $y_{i,t-1}$. By substitution,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] = \sigma_c^2 + \gamma\, \text{Cov}[y_{i,t-2}, (c_i + \varepsilon_{it})],$$

and so on. By repeated substitution, it can be seen that for $|\gamma| < 1$ and moderately large T,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] \approx \sigma_c^2/(1 - \gamma). \qquad \textbf{(11-64)}$$

[It is useful to obtain this result from a different direction. If the stochastic process that is generating (y_{it}, c_i) is stationary, then $\text{Cov}[y_{i,t-1}, c_i] = \text{Cov}[y_{i,t-2}, c_i]$, from which we would obtain (11-64) directly. The assumption $|\gamma| < 1$ would be required for stationarity.]

Consequently, OLS and GLS are inconsistent. The fixed effects approach does not solve the problem either. Taking deviations from individual means, we have

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \gamma(y_{i,t-1} - \bar{y}_{i.}) + (\varepsilon_{it} - \bar{\varepsilon}_{i.}).$$

Anderson and Hsiao (1981, 1982) show that

$$\begin{aligned}\text{Cov}[(y_{it} - \bar{y}_{i.}), (\varepsilon_{it} - \bar{\varepsilon}_{i.})] &\approx \frac{-\sigma_\varepsilon^2}{T(1-\gamma)^2}\left[\frac{(T-1) - T\gamma + \gamma^T}{T}\right] \\ &= \frac{-\sigma_\varepsilon^2}{T(1-\gamma)^2}\left[(1-\gamma) - \frac{1-\gamma^T}{T}\right].\end{aligned}$$

This does converge to zero as T increases, but, again, we are considering cases in which T is small or moderate, say 5 to 15, in which case the bias in the OLS estimator could be 15% to 60%. The implication is that the "within" transformation does not produce a consistent estimator.

It is easy to see that taking first differences is likewise ineffective. The first differences of the observations are

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \gamma(y_{i,t-1} - y_{i,t-2}) + (\varepsilon_{it} - \varepsilon_{i,t-1}). \qquad \textbf{(11-65)}$$

As before, the correlation between the last regressor and the disturbance persists, so OLS or GLS based on first differences would also be inconsistent. There is another approach. Write the regression in differenced form as

$$\Delta y_{it} = \Delta \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma\, \Delta y_{i,t-1} + \Delta\varepsilon_{it},$$

or, defining $\mathbf{x}_{it}^* = [\Delta\mathbf{x}_{it}, \Delta y_{i,t-1}]$, $\varepsilon_{it}^* = \Delta\varepsilon_{it}$ and $\boldsymbol{\theta} = [\boldsymbol{\beta}', \gamma]'$,

$$y_{it}^* = \mathbf{x}_{it}^{*\prime}\boldsymbol{\theta} + \varepsilon_{it}^*.$$

For the pooled sample, beginning with $t = 3$, write this as

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon}^*.$$

The least squares estimator based on the first differenced data is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left[\frac{1}{n(T-3)}\mathbf{X}^{*\prime}\mathbf{X}^*\right]^{-1}\left(\frac{1}{n(T-3)}\mathbf{X}^{*\prime}\mathbf{y}^*\right) \\ &= \boldsymbol{\theta} + \left[\frac{1}{n(T-3)}\mathbf{X}^{*\prime}\mathbf{X}^*\right]^{-1}\left(\frac{1}{n(T-3)}\mathbf{X}^{*\prime}\boldsymbol{\varepsilon}^*\right).\end{aligned}$$

Assuming that the inverse matrix in brackets converges to a positive definite matrix—that remains to be shown—the inconsistency in this estimator arises because the vector in parentheses does not converge to zero. The last element is $\text{plim}_{n\to\infty}[1/(n(T-3))]\Sigma_{i=1}^n\Sigma_{t=3}^T(y_{i,t-1} - y_{i,t-2})(\varepsilon_{it} - \varepsilon_{i,t-1})$, which is not zero.

Suppose there were a variable $\mathbf{z}^*$ such that plim $[1/(n(T-3))]\mathbf{z}^{*\prime}\boldsymbol{\varepsilon}^* = 0$ (exogenous) and $\text{plim}[1/(n(T-3))]\mathbf{z}^{*\prime}\mathbf{X}^* \neq \mathbf{0}$ (relevant). Let $\mathbf{Z} = [\Delta\mathbf{X}, \mathbf{z}^*]$; z_{it}^* replaces $\Delta y_{i,t-1}$ in $\mathbf{x}_{it}^*$. By this construction, it appears we have a consistent estimator. Consider

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{IV} &= (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'\mathbf{y}^*. \\ &= (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'(\mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon}^*) \\ &= \boldsymbol{\theta} + (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}^*.\end{aligned}$$

Then, after multiplying throughout by $1/(n(T-3))$ as before, we find

$$\text{Plim } \hat{\boldsymbol{\theta}}_{\text{IV}} = \boldsymbol{\theta} + \text{plim}\{[1/(n(T-3))](\mathbf{Z}'\mathbf{X}^*)\}^{-1} \times \mathbf{0},$$

which seems to solve the problem of consistent estimation.

The variable z^* is an **instrumental variable**, and the estimator is an **instrumental variable estimator** (hence the subscript on the preceding estimator). Finding suitable, valid instruments, that is, variables that satisfy the necessary assumptions, for models in which the right-hand variables are correlated with omitted factors is often challenging. In this setting, there is a natural candidate—in fact, there are several. From (11-65), we have at period $t = 3$,

$$y_{i3} - y_{i2} = (\mathbf{x}_{i3} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \gamma(y_{i2} - y_{i1}) + (\varepsilon_{i3} - \varepsilon_{i2}).$$

We could use y_{i1} as the needed variable because it is not correlated $\varepsilon_{i3} - \varepsilon_{i2}$. Continuing in this fashion, we see that for $t = 3, 4, \ldots, T$, $y_{i,t-2}$ satisfies our requirements. Alternatively, beginning from period $t = 4$, we can see that $z_{it} = (y_{i,t-2} - y_{i,t-3})$ once again satisfies our requirements. This is Anderson and Hsiao's (1981) result for instrumental variable estimation of the dynamic panel data model. It now becomes a question of which approach, levels ($y_{i,t-2}, t = 3, \ldots, T$), or differences ($y_{i,t-2} - y_{i,t-3}, t = 4, \ldots, T$) is a preferable approach. Arellano (1989) and Kiviet (1995) obtain results that suggest that the estimator based on levels is more efficient.

11.8.4 EFFICIENT ESTIMATION OF DYNAMIC PANEL DATA MODELS: THE ARELLANO/BOND ESTIMATORS

A leading application of the methods of this chapter is the **dynamic panel data model**, which we now write as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + c_i + \varepsilon_{it}.$$

Several applications are described in Example 11.21. The basic assumptions of the model are

1. Strict exogeneity: $E[\varepsilon_{it}|\mathbf{X}_i, c_i] = 0$,

2. Homoscedasticity and Nonautocorrelation:

$$E[\varepsilon_{it}\varepsilon_{is}|\mathbf{X}_i, c_i] = \sigma_\varepsilon^2 \text{ if } i = j \text{ and } t = s \text{ and } = 0 \text{ if } i \neq j \text{ or } t \neq s,$$

3. Common effects: The rows of the $T \times K$ data matrix $\mathbf{X}_i$ are $\mathbf{x}_{it}'$. We will not assume mean independence. The "effects" may be fixed or random, so we allow

$$E[c_i|\mathbf{X}_i] = h(\mathbf{X}_i).$$

(See Section 11.2.1.) We will also assume a fixed number of periods, T, for convenience. The treatment here (and in the literature) can be modified to accommodate unbalanced panels, but it is a bit inconvenient. (It involves the placement of zeros at various places in the data matrices defined below and changing the terminal indexes in summations from 1 to T.)

The presence of the lagged dependent variable in this model presents a considerable obstacle to estimation. Consider, first, the straightforward application of Assumption A.I3 in Section 8.2. The compound disturbance in the model is

$(c_i + \varepsilon_{it})$. The correlation between $y_{i,t-1}$ and $(c_i + \varepsilon_{i,t})$ is obviously nonzero because $y_{i,t-1} = \mathbf{x}'_{i,t-1}\boldsymbol{\beta} + \delta y_{i,t-2} + c_i + \varepsilon_{i,t-1}$,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] = \sigma_c^2 + \delta\,\text{Cov}[y_{i,t-2}, (c_i + \varepsilon_{it})].$$

If T is large and $0 < \delta < 1$, then this covariance will be approximately $\sigma_c^2/(1 - \delta)$. The large T assumption is not going to be met in most cases. But because δ will generally be positive, we can expect that this covariance will be at least larger than σ_c^2. The implication is that both (pooled) OLS and GLS in this model will be inconsistent. Unlike the case for the static model ($\delta = 0$), the fixed effects treatment does not solve the problem. Taking group mean differences, we obtain

$$y_{i,t} - \bar{y}_{i.} = (\mathbf{x}_{i,t} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \delta(y_{i,t-1} - \bar{y}_{i.}) + (\varepsilon_{i,t} - \bar{\varepsilon}_{i.}).$$

As shown in Anderson and Hsiao (1981, 1982),

$$\text{Cov}[(y_{i,t-1} - \bar{y}_{i.}), (\varepsilon_{i,t} - \bar{\varepsilon}_{i.})] \approx \frac{-\sigma_\varepsilon^2}{T^2}\frac{(T-1) - T\delta + \delta^T}{(1-\delta)^2}.$$

This result is $O(1/T)$, which would generally be no problem if the asymptotics in the model were with respect to increasing T. But, in this panel data model, T is assumed to be fixed and relatively small. For conventional values of T, say 5 to 15, the proportional bias in estimation of δ could be on the order of, say, 15 to 60 percent.

Neither OLS nor GLS are useful as estimators. There are, however, instrumental variables available within the structure of the model. Anderson and Hsiao (1981, 1982) proposed an approach based on first differences rather than differences from group means,

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1}.$$

For the first full observation,

$$y_{i3} - y_{i2} = (\mathbf{x}_{i3} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \delta(y_{i2} - y_{i1}) + \varepsilon_{i3} - \varepsilon_{i2}, \qquad \textbf{(11-66)}$$

the variable y_{i1} (assuming initial point $t = 0$ is where our data-generating process begins) satisfies the requirements, because ε_{i1} is predetermined with respect to $(\varepsilon_{i3} - \varepsilon_{i2})$. [That is, if we used only the data from periods 1 to 3 constructed as in (11-66), then the instrumental variables for $(y_{i2} - y_{i1})$ would be $\mathbf{z}_{i(3)}$ where $\mathbf{z}_{i(3)} = (y_{1,1}, y_{2,1}, \ldots, y_{n,1})$ for the n observations.] For the next observation,

$$y_{i4} - y_{i3} = (\mathbf{x}_{i4} - \mathbf{x}_{i3})'\boldsymbol{\beta} + \delta(y_{i3} - y_{i2}) + \varepsilon_{i4} - \varepsilon_{i3},$$

variables y_{i2} and $(y_{i2} - y_{i1})$ are both available.

Based on the preceding paragraph, one might begin to suspect that there is, in fact, rather than a paucity of instruments, a large surplus. In this limited development, we have a choice between differences and levels. Indeed, we could use both and, moreover, in any period after the fourth, not only is y_{i2} available as an instrument, but so also is y_{i1}, and so on. This is the essential observation behind the Arellano, Bover, and Bond (1991, 1995) estimators, which are based on the very large number of candidates for instrumental variables in this panel data model. To begin, with the model in first differences form, for $y_{i3} - y_{i2}$, variable y_{i1} is available. For $y_{i4} - y_{i3}$, y_{i1} and y_{i2} are both available; for $y_{i5} - y_{i4}$,

we have y_{i1}, y_{i2}, and y_{i3}, and so on. Consider, as well, that we have not used the exogenous variables. With strictly exogenous regressors, not only are all lagged values of y_{is} for s previous to $t-1$, but all values of $\mathbf{x}_{it}$ are also available as instruments. For example, for $y_{i4} - y_{i3}$, the candidates are y_{i1}, y_{i2} and $(\mathbf{x}_{i1}', \mathbf{x}_{i2}', \ldots, \mathbf{x}_{iT}')$ for all T periods. The number of candidates for instruments is, in fact, potentially huge.[36] If the exogenous variables are only predetermined, rather than strictly exogenous, then only $E[\varepsilon_{it} | \mathbf{x}_{i,t}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}] = 0$, and only vectors $\mathbf{x}_{is}$ from 1 to $t-1$ will be valid instruments in the differenced equation that contains $\varepsilon_{it} - \varepsilon_{i,t-1}$.[37] This is hardly a limitation, given that in the end, for a moderate sized model, we may be considering potentially hundreds or thousands of instrumental variables for estimation of what is usually a small handful of parameters.

We now formulate the model in a more familiar form, so we can apply the instrumental variable estimator. In terms of the differenced data, the basic equation is

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1},$$

or

$$\Delta y_{it} = (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \delta(\Delta y_{i,t-1}) + \Delta\varepsilon_{it}, \tag{11-67}$$

where Δ is the first difference operator, $\Delta a_t = a_t - a_{t-1}$ for any time-series variable (or vector) a_t. (It should be noted that a constant term and any time-invariant variables in $\mathbf{x}_{it}$ will fall out of the first differences. We will recover these below after we develop the estimator for $\boldsymbol{\beta}$.) The parameters of the model to be estimated are $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)'$ and σ_ε^2. For convenience, write the model as

$$\widetilde{y}_{it} = \widetilde{\mathbf{x}}_{it}'\boldsymbol{\theta} + \widetilde{\varepsilon}_{it}.$$

We are going to define an instrumental variable estimator along the lines of (8-9) and (8-10). Because our data set is a panel, the counterpart to

$$\mathbf{Z}'\widetilde{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{z}_i \widetilde{\mathbf{x}}_i' \tag{11-68}$$

in the cross-section case would seem to be

$$\mathbf{Z}'\widetilde{\mathbf{X}} = \sum_{i=1}^{n}\sum_{t=3}^{T} \mathbf{z}_{it}\widetilde{\mathbf{x}}_{it}' = \sum_{i=1}^{n} \mathbf{Z}_i'\widetilde{\mathbf{X}}_i', \tag{11-69}$$

$$\widetilde{\mathbf{y}}_i = \begin{bmatrix} \Delta y_{i3} \\ \Delta y_{i4} \\ \vdots \\ \Delta y_{iT_i} \end{bmatrix}, \widetilde{\mathbf{X}}_i = \begin{bmatrix} \Delta \mathbf{x}_{i3}' & \Delta y_{i2} \\ \Delta \mathbf{x}_{i4}' & \Delta y_{i3} \\ \cdots & \\ \Delta \mathbf{x}_{iT}' & \Delta y_{i,T-1} \end{bmatrix},$$

where there are $(T-2)$ observations (rows) and $K+1$ columns in $\widetilde{\mathbf{X}}_i$. There is a complication, however, in that the number of instruments we have defined may vary by period, so the matrix computation in (11-69) appears to sum matrices of different sizes.

[36] See Ahn and Schmidt (1995) for a very detailed analysis.

[37] See Baltagi and Levin (1986) for an application.

Consider an alternative approach. If we used only the first full observations defined in (11-67), then the cross-section version would apply, and the set of instruments $\mathbf{Z}$ in (11-68) with strictly exogenous variables would be the $n \times (1 + KT)$ matrix,

$$\mathbf{Z}_{(3)} = \begin{bmatrix} y_{1,1}, \mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \cdots \mathbf{x}'_{1,T} \\ y_{2,1}, \mathbf{x}'_{2,1}, \mathbf{x}'_{2,2}, \cdots \mathbf{x}'_{2,T} \\ \vdots \\ y_{n,1}, \mathbf{x}'_{n,1}, \mathbf{x}'_{n,2}, \cdots \mathbf{x}'_{n,T} \end{bmatrix},$$

and the instrumental variable estimator of (8-9) would be based on

$$\widetilde{\mathbf{X}}_{(3)} = \begin{bmatrix} \mathbf{x}'_{1,3} - \mathbf{x}'_{1,2} & y_{1,4} - y_{1,3} \\ \mathbf{x}'_{2,3} - \mathbf{x}'_{2,2} & y_{2,4} - y_{2,3} \\ \vdots & \vdots \\ \mathbf{x}'_{n,3} - \mathbf{x}'_{n,2} & y_{n,4} - y_{n,3} \end{bmatrix} \text{ and } \widetilde{\mathbf{y}}_{(3)} = \begin{bmatrix} y_{1,3} - y_{1,2} \\ y_{2,3} - y_{2,2} \\ \vdots \\ y_{n,3} - y_{n,2} \end{bmatrix}.$$

The subscript "(3)" indicates the first observation used for the left-hand side of the equation. Neglecting the other observations, then, we could use these data to form the IV estimator in (8-9), which we label for the moment $\hat{\boldsymbol{\theta}}_{\mathrm{IV}(3)}$. Now, repeat the construction using the next (fourth) observation as the first, and, again, using only a single year of the panel. The data matrices are now

$$\widetilde{\mathbf{X}}_{(4)} = \begin{bmatrix} \mathbf{x}'_{1,4} - \mathbf{x}'_{1,3} & y_{1,3} - y_{1,2} \\ \mathbf{x}'_{2,4} - \mathbf{x}'_{2,3} & y_{2,3} - y_{2,2} \\ \vdots & \vdots \\ \mathbf{x}'_{n,4} - \mathbf{x}'_{n,3} & y_{n,3} - y_{n,2} \end{bmatrix}, \widetilde{\mathbf{y}}_{(4)} = \begin{bmatrix} y_{1,4} - y_{1,3} \\ y_{2,4} - y_{2,3} \\ \vdots \\ y_{n,4} - y_{n,3} \end{bmatrix}, \text{ and}$$

$$\mathbf{Z}_{(4)} = \begin{bmatrix} y_{1,1}, y_{1,2}, \mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \cdots \mathbf{x}'_{1,T} \\ y_{2,1}, y_{2,2}, \mathbf{x}'_{2,1}, \mathbf{x}'_{2,2}, \cdots \mathbf{x}'_{2,T} \\ \vdots \\ y_{n,1}, y_{n,2}, \mathbf{x}'_{n,1}, \mathbf{x}'_{n,2}, \cdots \mathbf{x}_{n,T} \end{bmatrix}, \tag{11-70}$$

and we have a second IV estimator, $\hat{\boldsymbol{\theta}}_{\mathrm{IV}(4)}$, also based on n observations, but, now, $2 + KT$ instruments. And so on.

We now need to reconcile the $T - 2$ estimators of $\boldsymbol{\theta}$ that we have constructed, $\hat{\boldsymbol{\theta}}_{\mathrm{IV}(3)}, \hat{\boldsymbol{\theta}}_{\mathrm{IV}(4)}, \ldots, \hat{\boldsymbol{\theta}}_{\mathrm{IV}(T)}$. We faced this problem in Section 11.5.8 where we examined Chamberlain's formulation of the fixed effects model. The minimum distance estimator suggested there and used in Carey's (1997) study of hospital costs in Example 11.13 provides a means of efficiently "averaging" the multiple estimators of the parameter vector. We will return to the MDE in Chapter 13. For the present, we consider, instead, **Arellano and Bond's approach** (1991)[38] to this problem. We will collect the full set of estimators in a counterpart to (11-56) and (11-57). First, combine the sets of instruments in a single matrix, $\mathbf{Z}$, where for each individual, we obtain the $(T - 2) \times L$ matrix $\mathbf{Z}_i$. The definition of the rows of $\mathbf{Z}_i$ depend on whether the regressors are assumed to be strictly exogenous or predetermined. For strictly exogenous variables,

[38] And Arellano and Bover's (1995).

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} & 0 & \ldots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & y_{i,1}, y_{i,2}, \ldots, y_{i,T-2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} \end{bmatrix}, \tag{11.71a}$$

and $L = \sum_{i=1}^{T-2}(i + TK) = (T-2)(T-1)/2 + (T-2)TK$. For only predetermined variables, the matrix of instrumental variables is

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2} & 0 & \ldots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \mathbf{x}'_{i,3} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & y_{i,1}, y_{i,2}, \ldots, y_{i,T-2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T-1} \end{bmatrix}, \tag{11.71b}$$

and $L = \Sigma_{i=1}^{T-2}(i(K+1) + K) = [(T-2)(T-1)/2](1+K) + (T-2)K$. This construction does proliferate instruments (moment conditions, as we will see in Chapter 13). In the application in Example 11.18, we have a small panel with only $T = 7$ periods, and we fit a model with only $K = 4$ regressors in $\mathbf{x}_{it}$, plus the lagged dependent variable. The strict exogeneity assumption produces a $\mathbf{Z}_i$ matrix that is (5×135) for this case. With only the assumption of predetermined $\mathbf{x}_{it}$, $\mathbf{Z}_i$ collapses slightly to (5×95). For purposes of the illustration, we have used only the two previous observations on $\mathbf{x}_{it}$. This further reduces the matrix to

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2} & 0 & \ldots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}_{i,2}, \mathbf{x}'_{i,3} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & y_{i,1}, y_{i,2}, \ldots, y_{i,T-2}, \mathbf{x}'_{i,T-2}, \mathbf{x}'_{i,T-1} \end{bmatrix}, \tag{11.71c}$$

which, with $T = 7$ and $K = 4$, will be (5×55).[39]

Now, we can compute the two-stage least squares estimator in (11-55) using our definitions of the data matrices $\mathbf{Z}_i$, $\widetilde{\mathbf{X}}_i$, and $\widetilde{\mathbf{y}}_i$ and (11-69). This will be

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\mathrm{IV}} = & \left[\left(\sum_{i=1}^{n} \widetilde{\mathbf{X}}'_i \mathbf{Z}_i \right) \left(\sum_{i=1}^{n} \mathbf{Z}'_i \mathbf{Z}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{Z}'_i \widetilde{\mathbf{X}}_i \right) \right]^{-1} \\ & \times \left[\left(\sum_{i=1}^{n} \widetilde{\mathbf{X}}'_i \mathbf{Z}_i \right) \left(\sum_{i=1}^{n} \mathbf{Z}'_i \mathbf{Z}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{Z}'_i \widetilde{\mathbf{y}}_i \right) \right]. \end{aligned} \tag{11-72}$$

The natural estimator of the asymptotic covariance matrix for the estimator would be

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}_{\mathrm{IV}}] = \hat{\sigma}^2_{\Delta\varepsilon} \left[\left(\sum_{i=1}^{n} \widetilde{\mathbf{X}}'_i \mathbf{Z}_i \right) \left(\sum_{i=1}^{n} \mathbf{Z}'_i \mathbf{Z}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{Z}'_i \widetilde{\mathbf{X}}_i \right) \right]^{-1}, \tag{11-73}$$

[39] Baltagi (2005, Chapter 8) presents some alternative configurations of $\mathbf{Z}_i$ that allow for mixtures of strictly exogenous and predetermined variables.

where

$$\hat{\sigma}^2_{\Delta\varepsilon} = \frac{\sum_{i=1}^{n}\sum_{t=3}^{T}[(y_{it} - y_{i,t-1}) - (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\hat{\boldsymbol{\beta}} - \hat{\delta}(y_{i,t-1} - y_{i,t-2})]^2}{n(T-2)}. \tag{11-74}$$

However, this variance estimator is likely to understate the true asymptotic variance because the observations are autocorrelated for one period. Because $(y_{it} - y_{i,t-1}) = \widetilde{\mathbf{x}}'_{it}\boldsymbol{\theta} + (\varepsilon_{it} - \varepsilon_{i,t-1}) = \widetilde{\mathbf{x}}'_{it}\boldsymbol{\theta} + v_{it}$, $\text{Cov}[v_{it}, v_{i,t-1}] = \text{Cov}[v_{it}, v_{i,t+1}] = -\sigma^2_\varepsilon$. Covariances at longer lags or leads are zero. In the differenced model, though the disturbance covariance matrix is not $\sigma^2_v\mathbf{I}$, it does take a particularly simple form,

$$\text{Cov}\begin{pmatrix} \varepsilon_{i,3} - \varepsilon_{i,2} \\ \varepsilon_{i,4} - \varepsilon_{i,3} \\ \varepsilon_{i,5} - \varepsilon_{i,4} \\ \cdots \\ \varepsilon_{i,T} - \varepsilon_{i,T-1} \end{pmatrix} = \sigma^2_\varepsilon \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ 0 & -1 & 2 & \ldots & 0 \\ \ldots & \ldots & -1 & \ldots & -1 \\ 0 & 0 & \ldots & -1 & 2 \end{bmatrix} = \sigma^2_\varepsilon \boldsymbol{\Omega}_i. \tag{11-75}$$

The implication is that the estimator in (11-74) estimates not σ^2_ε but $2\sigma^2_\varepsilon$. However, simply dividing the estimator by two does not produce the correct asymptotic covariance matrix because the observations themselves are autocorrelated. As such, the matrix in (11-73) is inappropriate. A robust correction can be based on the counterpart to the White estimator that we developed in (11-3). For simplicity, let

$$\hat{\mathbf{A}} = \left[\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}'_i\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}'_i\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}'_i\widetilde{\mathbf{X}}_i\right)\right]^{-1}.$$

Then, a robust covariance matrix that accounts for the autocorrelation would be

$$\hat{\mathbf{A}}\left[\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}'_i\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}'_i\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}'_i\hat{\mathbf{v}}_i\hat{\mathbf{v}}'_i\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}'_i\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}'_i\widetilde{\mathbf{X}}_i\right)\right]\hat{\mathbf{A}}. \tag{11-76}$$

[One could also replace the $\hat{\mathbf{v}}_i\hat{\mathbf{v}}'_i$ in (11-76) with $\hat{\sigma}^2_\varepsilon\boldsymbol{\Omega}_i$ in (11-75) because this is the known expectation.]

It will be useful to digress briefly and examine the estimator in (11-72). The computations are less formidable than it might appear. Note that the rows of $\mathbf{Z}_i$ in (11-71a,b,c) are orthogonal. It follows that the matrix $\mathbf{F} = \sum_{i=1}^{n}\mathbf{Z}'_i\mathbf{Z}_i$ in (11-72) is block-diagonal with $T-2$ blocks. The specific blocks in $\mathbf{F}$ are $\mathbf{F}_t = \sum_{i=1}^{n}\mathbf{z}_{it}\mathbf{z}'_{it} = \mathbf{Z}'_{(t)}\mathbf{Z}_{(t)}$, for $t = 3, \ldots, T$. Because the number of instruments is different in each period—see (11-71)—these blocks are of different sizes, say, $(L_t \times L_t)$. The same construction shows that the matrix $\sum_{i=1}^{n}\widetilde{\mathbf{X}}'_i\mathbf{Z}_i$ is actually a partitioned matrix of the form

$$\sum_{i=1}^{n}\widetilde{\mathbf{X}}'_i\mathbf{Z}_i = \begin{bmatrix} \widetilde{\mathbf{X}}'_{(3)}\mathbf{Z}_{(3)} & \widetilde{\mathbf{X}}'_{(4)}\mathbf{Z}_{(4)} & \ldots & \widetilde{\mathbf{X}}'_{(T)}\mathbf{Z}_{(T)} \end{bmatrix},$$

where, again, the matrices are of different sizes; there are $T-2$ rows in each but the number of columns differs. It follows that the inverse matrix, $\left(\sum_{i=1}^{n}\mathbf{Z}'_i\mathbf{Z}_i\right)^{-1}$, is also block-diagonal, and that the matrix quadratic form in (11-72) can be written

$$\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{X}}_i\right) = \sum_{t=3}^{T}(\widetilde{\mathbf{X}}_{(t)}'\mathbf{Z}_{(t)})(\mathbf{Z}_{(t)}'\mathbf{Z}_{(t)})^{-1}(\mathbf{Z}_{(t)}'\widetilde{\mathbf{X}}_{(t)})$$
$$= \sum_{t=3}^{T}\left(\hat{\widetilde{\mathbf{X}}}_{(t)}'\hat{\widetilde{\mathbf{X}}}_{(t)}\right)$$
$$= \sum_{t=3}^{T}\mathbf{W}_{(t)},$$

[see (8-9) and the preceding result]. Continuing in this fashion, we find

$$\left(\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\widetilde{\mathbf{y}}_i\right) = \sum_{t=3}^{T}\hat{\widetilde{\mathbf{X}}}_{(t)}'\mathbf{y}_{(t)}.$$

From (8-10), we can see that

$$\hat{\widetilde{\mathbf{X}}}_{(t)}'\mathbf{y}_{(t)} = \left(\hat{\widetilde{\mathbf{X}}}_{(t)}'\hat{\widetilde{\mathbf{X}}}_{(t)}\right)\hat{\boldsymbol{\theta}}_{\mathrm{IV}}(t)$$
$$= \mathbf{W}_{(t)}\hat{\boldsymbol{\theta}}_{\mathrm{IV}}(t).$$

Combining the terms constructed thus far, we find that the estimator in (11-72) can be written in the form

$$\hat{\boldsymbol{\theta}}_{\mathrm{IV}} = \left(\sum_{t=3}^{T}\mathbf{W}_{(t)}\right)^{-1}\left(\sum_{t=3}^{T}\mathbf{W}_{(t)}\hat{\boldsymbol{\theta}}_{\mathrm{IV}}(t)\right)$$
$$= \sum_{t=3}^{T}\mathbf{R}_{(t)}\hat{\boldsymbol{\theta}}_{\mathrm{IV}}(t),$$

where

$$\mathbf{R}_{(t)} = \left(\sum_{t=3}^{T}\mathbf{W}_{(t)}\right)^{-1}\mathbf{W}_{(t)} \text{ and } \sum_{t=3}^{T}\mathbf{R}_{(t)} = \mathbf{I}.$$

In words, we find that, as might be expected, the Arellano and Bond estimator of the parameter vector is a matrix weighted average of the $T - 2$ period-specific two-stage least squares estimators, where the instruments used in each period may differ. Because the estimator is an average of estimators, a question arises, is it an efficient average—are the weights chosen to produce an efficient estimator? Perhaps not surprisingly, the answer for this $\hat{\boldsymbol{\theta}}$ is no; there is a more efficient set of weights that can be constructed for this model. We will assemble them when we examine the generalized method of moments estimator in Chapter 13.

There remains a loose end in the preceding. After (11-67), it was noted that this treatment discards a constant term and any time-invariant variables that appear in the model. The Hausman and Taylor (1981) approach developed in the preceding section suggests a means by which the model could be completed to accommodate this possibility. Expand the basic formulation to include the time-invariant effects, as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + \boldsymbol{\alpha} + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \varepsilon_{it},$$

where $\mathbf{f}_i$ is the set of time-invariant variables and $\boldsymbol{\gamma}$ is the parameter vector yet to be estimated. This model is consistent with the entire preceding development, as the component $\alpha + \mathbf{f}_i'\boldsymbol{\gamma}$ would have fallen out of the differenced equation along with c_i at

the first step at (11-63). Having developed a consistent estimator for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)'$, we now turn to estimation of $(\alpha, \boldsymbol{\gamma}')'$. The residuals from the IV regression (11-72),

$$w_{it} = \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{IV} - \hat{\delta}_{IV} y_{i,t-1},$$

are pointwise consistent estimators of

$$\omega_{it} = \alpha + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \varepsilon_{it}.$$

Thus, the group means of the residuals can form the basis of a second-step regression,

$$\overline{w}_i = \alpha + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \overline{\varepsilon}_i + \eta_i, \qquad \textbf{(11-77)}$$

where $\eta_i = (\overline{w}_i. - \overline{\omega}_i.)$ is the estimation error that converges to zero as $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$. The implication would seem to be that we can now linearly regress these group mean residuals on a constant and the time-invariant variables $\mathbf{f}_i$ to estimate α and $\boldsymbol{\gamma}$. The flaw in the strategy, however, is that the initial assumptions of the model do not state that c_i is uncorrelated with the other variables in the model, including the implicit time-invariant terms, $\mathbf{f}_i$. Therefore, least squares is not a usable estimator here unless the random effects model is assumed, which we specifically sought to avoid at the outset. As in Hausman and Taylor's treatment, there is a workable strategy if it can be assumed that there are some variables in the model, including possibly some among the $\mathbf{f}_i$ as well as others among $\mathbf{x}_{it}$ that are uncorrelated with c_i and ε_{it}. These are the $\mathbf{z}_1$ and $\mathbf{x}_1$ in the Hausman and Taylor estimator (see step 2 in the development of the preceding section). Assuming that these variables are available—this is an identification assumption that must be added to the model—then we do have a usable instrumental variable estimator, using as instruments the constant term (1), any variables in $\mathbf{f}_i$ that are uncorrelated with the latent effects or the disturbances (call this $\mathbf{f}_{i1}$), and the group means of any variables in $\mathbf{x}_{it}$ that are also exogenous. There must be enough of these to provide a sufficiently large set of instruments to fit all the parameters in (11-77). This is, once again, the same identification we saw in step 2 of the Hausman and Taylor estimator, K_1, the number of exogenous variables in $\mathbf{x}_{it}$ must be at least as large as L_2, which is the number of endogenous variables in $\mathbf{f}_i$. With all this in place, we then have the instrumental variable estimator in which the dependent variable is $\overline{w}_{i.}$, the right-hand-side variables are $(1, \mathbf{f}_i)$, and the instrumental variables are $(1, \mathbf{f}_{i1}, \overline{\mathbf{x}}_{i1}.)$.

There is yet another direction that we might extend this estimation method. In (11-76), we have implicitly allowed a more general covariance matrix to govern the generation of the disturbances ε_{it} and computed a robust covariance matrix for the simple IV estimator. We could take this a step further and look for a more efficient estimator. As a library of recent studies has shown, panel data sets are rich in information that allows the analyst to specify highly general models and to exploit the implied relationships among the variables to construct much more efficient generalized method of moments (GMM) estimators.[40] We will return to this development in Chapter 13.

Example 11.19 Dynamic Labor Supply Equation

In Example 8.5, we used instrumental variables to fit a labor supply equation,

$$Wks_{it} = \gamma_1 + \gamma_2 \ln Wage_{it} + \gamma_3 Ed_i + \gamma_4 Union_{it} + \gamma_5 Fem_i + u_{it}.$$

[40] See, in particular, Arellano and Bover (1995) and Blundell and Bond (1998).

To illustrate the computations of this section, we will extend this model as follows,

$$Wks_{it} = \beta_1 \ln Wage_{it} + \beta_2 Union_{it} + \beta_3 Occ_{it} + \beta_4 Exp_{it} + \delta Wks_{i,t-1} + \alpha + \gamma_1 Ed_i + \gamma_2 Fem_i + c_i + \varepsilon_{it}.$$

(We have rearranged the variables and parameter names to conform to the notation in this section.) We note, in theoretical terms, as suggested in the earlier example, it may not be appropriate to treat ln $Wage_{it}$ as uncorrelated with ε_{it} or c_i. However, we will be analyzing the model in first differences. It may well be appropriate to treat changes in wages as exogenous. That would depend on the theoretical underpinnings of the model. We will treat the variable as predetermined here, and proceed. There are two time-invariant variables in the model, Fem_i, which is clearly exogenous, and Ed_i, which might be endogenous. The identification requirement for estimation of $(\alpha, \gamma_1, \gamma_2)$ is met by the presence of three exogenous variables, $Union_{it}$, Occ_{it}, and Exp_{it} ($K_1 = 3$ and $L_2 = 1$).

The differenced equation analyzed at the first step is

$$\Delta Wks_{it} = \beta_1 \Delta \ln Wage_{it} + \beta_2 \Delta Union_{it} + \beta_3 \Delta Occ_{it} + \beta_4 \Delta Exp_{it} + \delta \Delta Wks_{i,t-1} + \Delta \varepsilon_{it}.$$

We estimated the parameters and the asymptotic covariance matrix according to (11-73) and (11-76). For specification of the instrumental variables, we used the one previous observation on $\mathbf{x}_{it}$, as shown in the text. Table 11.19 presents the computations with several other inconsistent estimators.

The various estimates are quite far apart. In the absence of the common effects (and autocorrelation of the disturbances), all five estimators shown would be consistent. Given the very wide disparities, one might suspect that common effects are an important feature

TABLE 11.19 Estimated Dynamic Panel Data Model Using Arellano and Bond Estimator

(Estimated standard errors in parentheses)

Variable	*OLS Full Equation*	*OLS Differenced*	*IV Differenced*	*Random Effects*	*Fixed Effects*
ln *Wage*	0.2966	−0.1100	−1.1402	0.2281	0.5886
	(0.2052)	(0.4565)	(0.2639) [0.8768]	(0.2405)	(0.4790)
Union	−1.2945	1.1640	2.7089	−1.4104	0.1444
	(0.1713)	(0.4222)	(0.3684) [0.8676]	(0.2199)	(0.4369)
Occ	0.4163	0.8142	2.2808	0.5191	1.0064
	(0.2005)	(0.3924)	(1.3105) [0.7220]	(2.2484)	(0.4030)
Exp	−0.0295	−0.0742	−0.0208	−0.0353	−0.1683
	(0.0073)	(0.0975)	(0.1126) [0.1104]	(0.0102)	(0.0595)
Wks_{t-1}	0.3804	−0.3527	0.1304	0.2100	0.0148
	(0.0148)	(0.0161)	(0.0476) [0.0213]	(0.0151)	(0.0171)
Constant	28.918	—	−0.4110	37.4610	—
	(1.4490)	—	(0.3364)	(1.6778)	—
Ed	−0.0690	—	0.0321	−0.0657	—
	(0.0370)	—	(0.0259)	(0.0499)	—
Fem	−0.8607	—	−0.0122	−1.1463	—
	(0.2544)	—	(0.1554)	(0.3513)	—
Sample	$t = 2$ to 7	$t = 3$ to 7	$t = 3$ to 7	$t = 2$ to 7	$t = 2$ to 7
Observations	595	595	595, Means used $t = 7$	595	595

of the data. The second standard errors given in brackets with the IV estimates are based on the uncorrected matrix in (11-73) with $\hat{\sigma}^2_{\Delta\varepsilon}$ in (11-74) divided by two. We found the estimator to be quite volatile, as can be seen in the table. The estimator is also very sensitive to the choice of instruments that comprise $\mathbf{Z}_i$. Using (11-71a) instead of (11-71b) produces wild swings in the estimates and, in fact, produces implausible results. One possible explanation in this particular example is that the instrumental variables we are using are dummy variables that have relatively little variation over time.

11.8.5 NONSTATIONARY DATA AND PANEL DATA MODELS

Some of the discussion thus far (and to follow) focuses on "small T" statistical results. Panels are taken to contain a fixed and small T observations on a large n individual units. Recent research using cross-country data sets such as the Penn World Tables (http://cid.econ.ucdavis.edu/pwt.html), which now include data on over 150 countries for well over 50 years, have begun to analyze panels with T sufficiently large that the time-series properties of the data become an important consideration. In particular, the recognition and accommodation of nonstationarity that is now a standard part of single time-series analyses (as in Chapter 21) are now seen to be appropriate for large-scale cross-country studies, such as income growth studies based on the Penn World Tables, cross-country studies of health care expenditure, and analyses of purchasing power parity.

The analysis of long panels, such as in the growth and convergence literature, typically involves dynamic models, such as

$$y_{it} = \alpha_i + \gamma_i y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}. \tag{11-78}$$

In single time-series analysis involving low-frequency macroeconomic flow data such as income, consumption, investment, the current account deficit, and so on, it has long been recognized that estimated regression relations can be distorted by nonstationarity in the data. What appear to be persistent and strong regression relationships can be entirely spurious and due to underlying characteristics of the time-series processes rather than actual connections among the variables. Hypothesis tests about long-run effects will be considerably distorted by unit roots in the data. It has become evident that the same influences, with the same deletarious effects, will be found in long panel data sets. The panel data application is further complicated by the possible heterogeneity of the parameters. The coefficients of interest in many cross-country studies are the lagged effects, such as γ_i in (11-78), and it is precisely here that the received results on nonstationary data have revealed the problems of estimation and inference. Valid tests for unit roots in panel data have been proposed in many studies. Three that are frequently cited are Levin and Lin (1992), Im, Pesaran, and Shin (2003), and Maddala and Wu (1999).

There have been numerous empirical applications of time-series methods for nonstationary data in panel data settings, including Frankel and Rose's (1996) and Pedroni's (2001) studies of purchasing power parity, Fleissig and Strauss (1997) on real wage stationarity, Culver and Papell (1997) on inflation, Wu (2000) on the current account balance, McCoskey and Selden (1998) on health care expenditure, Sala-i-Martin (1996) on growth and convergence, McCoskey and Kao (1999) on urbanization and production, and Coakely et al. (1996) on savings and investment. An extensive enumeration appears in Baltagi (2005, Chapter 12).

A subtle problem arises in obtaining results useful for characterizing the properties of estimators of the model in (11-78). The asymptotic results based on large n and large

T are not necessarily obtainable simultaneously, and great care is needed in deriving the asymptotic behavior of useful statistics. Phillips and Moon (1999, 2000) are standard references on the subject.

We will return to the topic of nonstationary data in Chapter 21. This is an emerging literature, most of which is beyond the level of this text. We will rely on the several detailed received surveys, such as Bannerjee (1999), Smith (2000), and Baltagi and Kao (2000), to fill in the details.

11.9 NONLINEAR REGRESSION WITH PANEL DATA

The extension of the panel data models to the nonlinear regression case is, perhaps surprisingly, not at all straightforward. Thus far, to accommodate the nonlinear model, we have generally applied familiar results to the linearized regression. This approach will carry forward to the case of clustered data. (See Section 11.3.3.) Unfortunately, this will not work with the standard panel data methods. The nonlinear regression will be the first of numerous panel data applications that we will consider in which the wisdom of the linear regression model cannot be extended to the more general framework.

11.9.1 A ROBUST COVARIANCE MATRIX FOR NONLINEAR LEAST SQUARES

The counterpart to (11-3) or (11-4) would simply replace $\mathbf{X}_i$ with $\hat{\mathbf{X}}_i^0$ where the rows are the pseudo regressors for cluster i as defined in (7-12) and "$\hat{}$" indicates that it is computed using the nonlinear least squares estimates of the parameters.

Example 11.20 Health Care Utilization

The recent literature in health economics includes many studies of health care utilization. A common measure of the dependent variable of interest is a count of the number of encounters with the health care system, either through visits to a physician or to a hospital. These counts of occurrences are usually studied with the Poisson regression model described in Section 18.4. The nonlinear regression model is

$$E[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}).$$

A recent study in this genre is "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation" by Riphahn, Wambach, and Million (2003). The authors were interested in counts of physician visits and hospital visits. In this application, they were particularly interested in the impact of the presence of private insurance on the utilization counts of interest, that is, whether the data contain evidence of moral hazard.

The raw data are published on the *Journal of Applied Econometrics* data archive Web site, The URL for the data file is http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/. The variables in the data file are listed in Appendix Table F7.1. The sample is an unbalanced panel of 7,293 households, the German Socioeconomic Panel data set. The number of observations varies from one to seven (1,525; 1,079; 825; 926; 1,311; 1,000; 887), with a total number of observations of 27,326. We will use these data in several examples here and later in the book.

The following model uses a simple specification for the count of number of visits to the physican in the observation year,

$$\mathbf{x}_{it} = (1, age_{it}, educ_{it}, income_{it}, kids_{it}).$$

Table 11.20 details the nonlinear least squares iterations and the results. The convergence criterion for the iterations is $\mathbf{e}^{0\prime}\mathbf{X}^0(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\mathbf{X}^{0\prime}\mathbf{e}^0 < 10^{-10}$. Although this requires 11 iterations,

TABLE 11.20 Nonlinear Least Squares Estimates of a Health Care Utilization Equation

Begin NLSQ iterations. Linearized regression.

Iteration = 1; Sum of squares = 1014865.00; Gradient = 156281.794
Iteration = 2; Sum of squares = 8995221.17; Gradient = 8131951.67
Iteration = 3; Sum of squares = 1757006.18; Gradient = 897066.012
Iteration = 4; Sum of squares = 930876.806; Gradient = 73036.2457
Iteration = 5; Sum of squares = 860068.332; Gradient = 2430.80472
Iteration = 6; Sum of squares = 857614.333; Gradient = 12.8270683
Iteration = 7; Sum of squares = 857600.927; Gradient = 0.411851239E-01
Iteration = 8; Sum of squares = 857600.883; Gradient = 0.190628165E-03
Iteration = 9; Sum of squares = 857600.883; Gradient = 0.904650588E-06
Iteration = 10; Sum of squares = 857600.883; Gradient = 0.430441193E-08
Iteration = 11; Sum of squares = 857600.883; Gradient = 0.204875467E-10
Convergence achieved

Variable	*Estimate*	*Std. Error*	*Robust Std. Error*
Constant	0.9801	0.08927	0.12522
Age	0.0187	0.00105	0.00142
Education	−0.0361	0.00573	0.00780
Income	−0.5911	0.07173	0.09702
Kids	−0.1692	0.02642	0.03330

the function actually reaches the minimum in 7. The estimates of the asymptotic standard errors are computed using the conventional method, $s^2(\hat{\mathbf{X}}^{0\prime}\hat{\mathbf{X}}^0)^{-1}$, and then by the cluster correction in (11-4). The corrected standard errors are considerably larger, as might be expected given that these are a panel data set.

11.9.2 FIXED EFFECTS IN NONLINEAR REGRESSION MODELS

The nonlinear panel data regression model would appear as

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}) + \varepsilon_{it}, t = 1, \ldots, T_i, i = 1, \ldots, n.$$

Consider a model with latent heterogeneity, c_i. An ambiguity immediately emerges; how should heterogeneity enter the model? Building on the linear model, an additive term might seem natural, as in

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}) + c_i + \varepsilon_{it}, t = 1, \ldots, T_i, i = 1, \ldots, n. \quad \textbf{(11-79)}$$

But we can see in the previous application that this is likely to be inappropriate. The loglinear model of the previous section is constrained to ensure that $E[y_{it}|\mathbf{x}_{it}]$ is positive. But an additive random term c_i as in (11-79) could subvert this; unless the range of c_i is restricted, the conditional mean could be negative. The most common application of nonlinear models is the **index function model**,

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + c_i) + \varepsilon_{it}.$$

This is the natural extension of the linear model, but only in the appearance of the conditional mean. Neither the fixed effects nor the random effects model can be estimated as they were in the linear case.

Consider the fixed effects model first. We would write this as

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) + \varepsilon_{it}, \tag{11-80}$$

where the parameters to be estimated are $\boldsymbol{\beta}$ and α_i, $i = 1, \ldots, n$. Transforming the data to deviations from group means does not remove the fixed effects from the model. For example,

$$y_{it} - \bar{y}_{i.} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) - \frac{1}{T_i}\sum_{s=1}^{T_i} h(\mathbf{x}_{is}'\boldsymbol{\beta} + \alpha_i),$$

which does not simplify things at all. Transforming the regressors to deviations is likewise pointless. To estimate the parameters, it is necessary to minimize the sum of squares with respect to all $n + K$ parameters simultaneously. Because the number of dummy variable coefficients can be huge—the preceding example is based on a data set with 7,293 groups—this can be a difficult or impractical computation. A method of maximizing a function (such as the negative of the sum of squares) that contains an unlimited number of dummy variable coefficients is shown in Chapter 17. As we will examine later in the book, the difficulty with nonlinear models that contain large numbers of dummy variable coefficients is not necessarily the practical one of computing the estimates. That is generally a solvable problem. The difficulty with such models is an intriguing phenomenon known as the **incidental parameters problem**. (See footnote 12.) In most (not all, as we shall find) nonlinear panel data models that contain n dummy variable coefficients, such as the one in (11-80), as a consequence of the fact that the number of parameters increases with the number of individuals in the sample, the estimator of $\boldsymbol{\beta}$ is biased and inconsistent, to a degree that is $O(1/T)$. Because T is only 7 or less in our application, this would seem to be a case in point.

Example 11.21 Exponential Model with Fixed Effects

The exponential model of the preceding example is actually one of a small handful of known special cases in which it is possible to "condition" out the dummy variables. Consider the sum of squared residuals,

$$S_n = \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}[y_{it} - \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i)]^2.$$

The first-order condition for minimizing S_n with respect to α_i is

$$\frac{\partial S_n}{\partial \alpha_i} = \sum_{t=1}^{T_i} -[y_{it} - \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i)]\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) = 0. \tag{11-81}$$

Let $\gamma_i = \exp(\alpha_i)$. Then, an equivalent necessary condition would be

$$\frac{\partial S_n}{\partial \gamma_i} = \sum_{t=1}^{T_i} -[y_{it} - \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})][\gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})] = 0,$$

or

$$\gamma_i \sum_{t=1}^{T_i}[y_{it}\exp(\mathbf{x}_{it}'\boldsymbol{\beta})] = \gamma_i^2 \sum_{t=1}^{T_i}[\exp(\mathbf{x}_{it}'\boldsymbol{\beta})]^2.$$

Obviously, if we can solve the equation for γ_i, we can obtain $\alpha_i = \ln \gamma_i$. The preceding equation can, indeed, be solved for γ_i, at least conditionally. At the minimum of the sum of squares, it will be true that

$$\hat{\gamma}_i = \frac{\sum_{t=1}^{T_i} y_{it}\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})}{\sum_{t=1}^{T_i}[\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})]^2}. \tag{11-82}$$

We can now insert (11-82) into (11-81) to eliminate α_i. (This is a counterpart to taking deviations from means in the linear case. As noted, this is possible only for a very few special models—this happens to be one of them. The process is also known as "concentrating out" the parameters γ_i. Note that at the solution, $\hat{\gamma}_i$ is obtained as the slope in a regression without a constant term of y_{it} on $\hat{\mathbf{z}}_{it} = \exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})$ using T_i observations.) The result in (11-82) must hold at the solution. Thus, (11-82) inserted in (11-81) restricts the search for $\boldsymbol{\beta}$ to those values that satisfy the restrictions in (11-82). The resulting sum of squares function is now a function only of the data and $\boldsymbol{\beta}$, and can be minimized with respect to this vector of K parameters. With the estimate of $\boldsymbol{\beta}$ in hand, α_i can be estimated using the log of the result in (11-82) (which is positive by construction).

The preceding example presents a mixed picture for the fixed effects model. In nonlinear cases, two problems emerge that were not present earlier, the practical one of actually computing the dummy variable parameters and the theoretical incidental parameters problem that we have yet to investigate, but which promises to be a significant shortcoming of the fixed effects model. We also note we have focused on a particular form of the model, the single index function, in which the conditional mean is a nonlinear function of a linear function. In more general cases, it may be unclear how the unobserved heterogeneity should enter the regression function.

11.9.3 RANDOM EFFECTS

The random effects nonlinear model also presents complications both for specification and for estimation. We might begin with a general model,

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}, u_i) + \varepsilon_{it}.$$

The "random effects" assumption would be, as usual, mean independence,

$$E[u_i | \mathbf{X}_i] = 0.$$

Unlike the linear model, the nonlinear regression cannot be consistently estimated by (nonlinear) least squares. In practical terms, we can see why in (7-28) through (7-30). In the linearized regression, the conditional mean at the expansion point $\boldsymbol{\beta}^0$ [see (7-28)] as well as the pseudoregressors are both functions of the unobserved u_i. This is true in the general case as well as the simpler case of a single index model,

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) + \varepsilon_{it}. \tag{11-83}$$

Thus, it is not possible to compute the iterations for nonlinear least squares. As in the fixed effects case, neither deviations from group means nor first differences solves the problem. Ignoring the problem—that is, simply computing the nonlinear least squares estimator without accounting for heterogeneity—does not produce a consistent estimator, for the same reasons. In general, the benign effect of latent heterogeneity (random effects) that we observe in the linear model only carries over to a very few nonlinear models and, unfortunately, this is not one of them.

The problem of computing partial effects in a random effects model such as (11-83) is that when $E[y_{it} | \mathbf{x}_{it}, u_i]$ is given by (11-83), then

$$\frac{\partial E[y_{it} | \mathbf{x}_{it}'\boldsymbol{\beta} + u_i]}{\partial \mathbf{x}_{it}} = [h'(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)]\boldsymbol{\beta}$$

is a function of the unobservable u_i. Two ways to proceed from here are the fixed effects approach of the previous section and a random effects approach. The fixed

effects approach is feasible but may be hindered by the incidental parameters problem noted earlier. A random effects approach might be preferable, but comes at the price of assuming that $\mathbf{x}_{it}$ and u_i are uncorrelated, which may be unreasonable. Papke and Wooldridge (2008) examined several cases and proposed the Mundlak approach of projecting u_i on the group means of $\mathbf{x}_{it}$. The working specification of the model is then

$$E^*[y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i, v_i] = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\theta} + v_i).$$

This leaves the practical problem of how to compute the estimates of the parameters and how to compute the partial effects. Papke and Wooldridge (2008) suggest a useful result if it can be assumed that v_i is normally distributed with mean zero and variance σ_v^2. In that case,

$$E[y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}] = E_{v_i}E[y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}, v_i] = h\left(\frac{\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\theta}}{\sqrt{1 + \sigma_v^2}}\right) = h(\mathbf{x}_{it}'\boldsymbol{\beta}_v + \alpha_v + \mathbf{x}_i'\boldsymbol{\theta}_v).$$

The implication is that nonlinear least squares regression will estimate the scaled coefficients, after which the average partial effect can be estimated for a particular value of the covariates, $\mathbf{x}_0$, with

$$\hat{\Delta}(\mathbf{x}_0) = \frac{1}{n}\sum_{i=1}^{n} h'(\mathbf{x}_0'\hat{\boldsymbol{\beta}}_v + \hat{\alpha}_v + \bar{\mathbf{x}}_i'\hat{\boldsymbol{\theta}}_v)\hat{\boldsymbol{\beta}}_v.$$

They applied the technique to a case of test pass rates, which are a fraction bounded by zero and one. Loudermilk (2007) is another application with an extension to a dynamic model.

11.10 PARAMETER HETEROGENEITY

The treatment so far has assumed that the slope parameters of the model are fixed constants, and the intercept varies randomly from group to group. An equivalent formulation of the pooled, fixed, and random effects models is

$$y_{it} = (\alpha + u_i) + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$

where u_i is a person-specific random variable with conditional variance zero in the pooled model, positive in the others, and conditional mean dependent on $\mathbf{X}_i$ in the fixed effects model and constant in the random effects model. By any of these, the heterogeneity in the model shows up as variation in the constant terms in the regression model. There is ample evidence in many studies—we will examine two later—that suggests that the other parameters in the model also vary across individuals. In the dynamic model we consider in Section 11.10.3, cross-country variation in the slope parameter in a production function is the central focus of the analysis. This section will consider several approaches to analyzing parameter heterogeneity in panel data models.

11.10.1 A RANDOM COEFFICIENTS MODEL

Parameter heterogeneity across individuals or groups can be modeled as stochastic variation.[41] Suppose that we write

[41]The most widely cited studies are Hildreth and Houck (1968), Swamy (1970, 1971, 1974), Hsiao (1975), and Chow (1984). See also Breusch and Pagan (1979). Some recent discussions are Swamy and Tavlas (1995, 2001) and Hsiao (2003). The model bears some resemblance to the Bayesian approach of Chapter 16. But the similarity is only superficial. We are maintaining the classical approach to estimation throughout.

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \\ E[\boldsymbol{\varepsilon}_i|\mathbf{X}_i] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{X}_i] &= \sigma_\varepsilon^2\mathbf{I}_T, \end{aligned} \tag{11-84}$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i \tag{11-85}$$

and

$$\begin{aligned} E[\mathbf{u}_i|\mathbf{X}_i] &= \mathbf{0}, \\ E[\mathbf{u}_i\mathbf{u}_i', |\mathbf{X}_i] &= \boldsymbol{\Gamma}. \end{aligned} \tag{11-86}$$

(Note that if only the constant term in $\boldsymbol{\beta}$ is random in this fashion and the other parameters are fixed as before, then this reproduces the random effects model we studied in Section 11.5.) Assume for now that there is no autocorrelation or cross-section correlation in $\boldsymbol{\varepsilon}_i$. We also assume for now that $T > K$, so that, when desired, it is possible to compute the linear regression of $\mathbf{y}_i$ on $\mathbf{X}_i$ for each group. Thus, the $\boldsymbol{\beta}_i$ that applies to a particular cross-sectional unit is the outcome of a random process with mean vector $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Gamma}$.[42] By inserting (11-85) into (11-84) and expanding the result, we obtain a generalized regression model for each block of observations,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + (\boldsymbol{\varepsilon}_i + \mathbf{X}_i\mathbf{u}_i),$$

so

$$\boldsymbol{\Omega}_{ii} = E[(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T + \mathbf{X}_i\boldsymbol{\Gamma}\mathbf{X}_i'.$$

For the system as a whole, the disturbance covariance matrix is block diagonal, with $T \times T$ diagonal block $\boldsymbol{\Omega}_{ii}$. We can write the GLS estimator as a matrix weighted average of the group-specific OLS estimators,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} = \sum_{i=1}^{n}\mathbf{W}_i\mathbf{b}_i, \tag{11-87}$$

where

$$\mathbf{W}_i = \left[\sum_{i=1}^{n}\left(\boldsymbol{\Gamma} + \sigma_\varepsilon^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right)^{-1}\right]^{-1}\left(\boldsymbol{\Gamma} + \sigma_\varepsilon^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right)^{-1}.$$

Empirical implementation of this model requires an estimator of $\boldsymbol{\Gamma}$. One approach[43] is to use the empirical variance of the set of n least squares estimates, $\mathbf{b}_i$ minus the average value of $s_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$,

$$\mathbf{G} = [1/(n-1)][\Sigma_i\mathbf{b}_i\mathbf{b}_i' - n\bar{\mathbf{b}}\bar{\mathbf{b}}'] - (1/N)\Sigma_i\mathbf{V}_i, \tag{11-88}$$

where

$$\bar{\mathbf{b}} = (1/n)\Sigma_i\mathbf{b}_i$$

and

$$\mathbf{V}_i = s_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}.$$

[42]Swamy and Tavlas (2001) label this the "first-generation random coefficients model" (RCM). We will examine the "second generation" (the current generation) of random coefficients models in the next section.

[43]See, for example, Swamy (1971).

This matrix may not be positive definite, however, in which case [as Baltagi (2005) suggests], one might drop the second term.

A chi-squared test of the random coefficients model against the alternative of the classical regression[44] (no randomness of the coefficients) can be based on

$$C = \Sigma_i(\mathbf{b}_i - \mathbf{b}_*)'\mathbf{V}_i^{-1}(\mathbf{b}_i - \mathbf{b}_*),$$

where

$$\mathbf{b}_* = [\Sigma_i \mathbf{V}_i^{-1}]^{-1}\Sigma_i \mathbf{V}_i^{-1}\mathbf{b}_i.$$

Under the null hypothesis of homogeneity, C has a limiting chi-squared distribution with $(n - 1)K$ degrees of freedom. The best linear unbiased individual predictors of the group-specific coefficient vectors are matrix weighted averages of the GLS estimator, $\hat{\boldsymbol{\beta}}$, and the group-specific OLS estimates, $\mathbf{b}_i$,[45]

$$\hat{\boldsymbol{\beta}}_i = \mathbf{Q}_i\hat{\boldsymbol{\beta}} + [\mathbf{I} - \mathbf{Q}_i]\mathbf{b}_i, \tag{11-89}$$

where

$$\mathbf{Q}_i = [(1/s_i^2)\mathbf{X}_i'\mathbf{X}_i + \mathbf{G}^{-1}]^{-1}\mathbf{G}^{-1}.$$

Example 11.22 Random Coefficients Model

In Examples 10.1 and 11.9, we examined Munell's production model for gross state product,

$$\ln gsp_{it} = \beta_1 + \beta_2 \ln pc_{it} + \beta_3 \ln hwy_{it} + \beta_4 \ln \text{water}_{it} + \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7 unemp_{it} + \varepsilon_{it}, \quad i = 1, \ldots, 48; t = 1, \ldots, 17.$$

The panel consists of state-level data for 17 years. The model in Example 10.1 (and Munnell's) provides no means for parameter heterogeneity save for the constant term. We have reestimated the model using the Hildreth and Houck approach. The OLS and Feasible GLS estimates are given in Table 11.21. The chi-squared statistic for testing the null hypothesis of parameter homogeneity is 25,556.26, with 7(47) = 329 degrees of freedom. The critical value from the table is 372.299, so the hypothesis would be rejected.

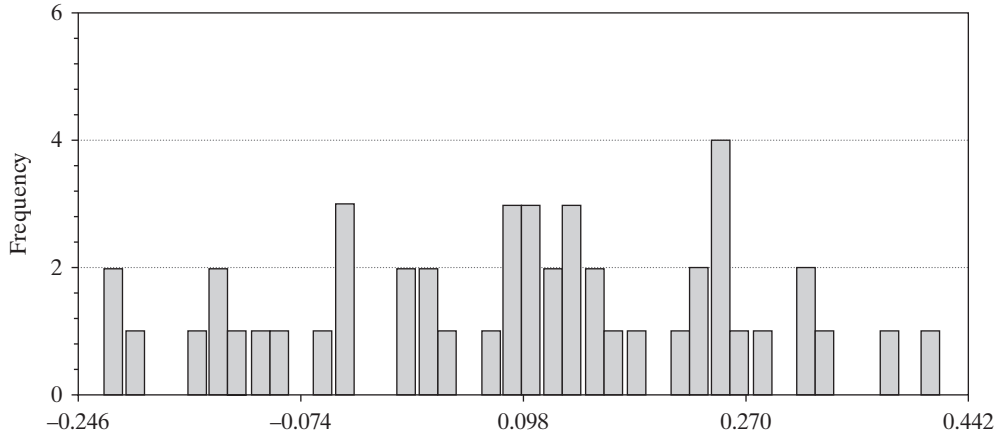
TABLE 11.21 Estimated Random Coefficients Models

	Least Squares		*Feasible GLS*		
Variable	***Estimate***	***Standard Error***	***Estimate***	***Std. Error***	***Popn. Std. Deviation***
Constant	1.9260	0.05250	1.6533	1.08331	7.0782
ln *pc*	0.3120	0.01109	0.09409	0.05152	0.3036
ln *hwy*	0.05888	0.01541	0.1050	0.1736	1.1112
ln *water*	0.1186	0.01236	0.07672	0.06743	0.4340
ln *util*	0.00856	0.01235	−0.01489	0.09886	0.6322
ln *emp*	0.5497	0.01554	0.9190	0.1044	0.6595
unemp	−0.00727	0.00138	−0.00471	0.00207	0.01266
σ_ε	0.08542			0.2129	
ln *L*	853.13720				

[44]See Swamy (1971).

[45]See Hsiao (2003, pp. 144–149).

FIGURE 11.1 Estimates of Coefficient on Private Capital.



Unlike the other cases we have examined in this chapter, the FGLS estimates are very different from OLS in these estimates, in spite of the fact that both estimators are consistent and the sample is fairly large. The underlying standard deviations are computed using **G** as the covariance matrix. [For these data, subtracting the second matrix rendered **G** not positive definite, so in the table, the standard deviations are based on the estimates using only the first term in (11-88).] The increase in the standard errors is striking. This suggests that there is considerable variation in the parameters across states. We have used (11-89) to compute the estimates of the state-specific coefficients. Figure 11.1 shows a histogram for the coefficient on private capital. As suggested, there is a wide variation in the estimates.

11.10.2 A HIERARCHICAL LINEAR MODEL

Many researchers have employed a two-step approach to estimate two-level models. In a common form of the application, a panel data set is employed to estimate the model,

$$\mathbf{y}_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}, i = 1, \ldots n, t = 1, \ldots, T,$$
$$\beta_{i,k} = \mathbf{z}_i'\boldsymbol{\alpha}_k + u_{i,k}, i = 1, \ldots, n.$$

Assuming the panel is long enough, the first equation is estimated n times, once for each individual i, and then the estimated coefficient on x_{itk} in each regression forms an observation for the second-step regression.[46] [This is the approach we took in (11-16) in Section 11.4; each a_i is computed by a linear regression of $\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{LSDV}$ on a column of ones.]

Example 11.23 Fannie Mae's Pass Through

Fannie Mae is the popular name for the Federal National Mortgage Corporation. Fannie Mae is the secondary provider for mortgage money for nearly all the small- and moderate-sized home mortgages in the United States. Loans in the study described here are termed "small" if they are for less than $100,000. A loan is termed as *conforming* in the language

[46] An extension of the model in which "u_i" is heteroscedastic is developed at length in Saxonhouse (1976) and revisited by Achen (2005).

of the literature on this market if (as of 2016), it is for no more than $417,000. A larger than conforming loan is called a *jumbo* mortgage. Fannie Mae provides the capital for nearly all conforming loans and no nonconforming loans. (See Exercise 6.14 for another study of Fannie Mae and Freddie Mac.) The question pursued in the study described here was whether the clearly observable spread between the rates on jumbo loans and conforming loans reflects the cost of raising the capital in the market. Fannie Mae is a government sponsored enterprice (GSE). It was created by the U.S. Congress, but it is not an arm of the government; it is a private corporation. In spite of, or perhaps because of, this ambiguous relationship to the government, apparently, capital markets believe that there is some benefit to Fannie Mae in raising capital. Purchasers of the GSE's debt securities seem to believe that the debt is implicitly backed by the government—this in spite of the fact that Fannie Mae explicitly states otherwise in its publications. This emerges as a funding advantage (GFA) estimated by the authors of the study of about 16 basis points (hundredths of one percent). In a study of the residential mortgage market, Passmore (2005) and Passmore, Sherlund, and Burgess (2005) sought to determine whether this implicit subsidy to the GSE was passed on to the mortgagees or was, instead, passed on to the stockholders. Their approach utilitized a very large data set and a two-level, two-step estimation procedure. The first step equation estimated was a mortgage rate equation using a sample of roughly 1 million closed mortgages. All were conventional 30-year, fixed-rate loans closed between April 1997 and May 2003. The dependent variable of interest is the rate on the mortgage, RM_{it}. The first-level equation is

$$RM_{it} = \beta_{1i} + \beta_{2,i} J_{it} + \text{terms for ``loan to value ratio,'' ``new home dummy variable,'' ``small mortgage''} + \text{terms for ``fees charged'' and whether the mortgage was originated by a mortgage company} + \varepsilon_{it}.$$

The main variable of interest in this model is J_{it}, which is a dummy variable for whether the loan is a jumbo mortgage. The "i" in this setting is a (state, time) pair for California, New Jersey, Maryland, Virginia, and all other states, and months from April 1997 to May 2003. There were 370 groups in total. The regression model was estimated for each group. At the second step, the coefficient of interest is $\beta_{2,i}$. On overall average, the spread between jumbo and conforming loans at the time was roughly 16 basis points. The second-level equation is

$$\begin{aligned}\beta_{2,i} = {} & \alpha_1 + \alpha_2 \,\text{GFA}_i \\ & + \alpha_3 \text{ one-year treasury rate} \\ & + \alpha_4 \text{ 10-year treasury rate} \\ & + \alpha_5 \text{ credit risk} \\ & + \alpha_6 \text{ prepayment risk} \\ & + \text{measures of maturity mismatch risk} \\ & + \text{quarter and state fixed effects} \\ & + \text{mortgage market capacity} \\ & + \text{mortgage market development} \\ & + u_i.\end{aligned}$$

The result ultimately of interest is the coefficient on GFA, α_2, which is interpreted as the fraction of the GSE funding advantage that is passed through to the mortgage holders. Four different estimates of α_2 were obtained, based on four different measures of corporate debt liquidity; the estimated values were $(\hat{\alpha}_2^1, \hat{\alpha}_2^2, \hat{\alpha}_2^3, \hat{\alpha}_2^4) = (0.07, 0.31, 0.17, 0.10)$. The four

estimates were averaged using a minimum distance estimator (MDE). Let $\hat{\mathbf{\Omega}}$ denote the estimated 4×4 asymptotic covariance matrix for the estimators. Denote the distance vector

$$\mathbf{d} = (\hat{\alpha}_2^1 - \alpha_2, \hat{\alpha}_2^2 - \alpha_2, \hat{\alpha}_2^3 - \alpha_2, \hat{\alpha}_2^4 - \alpha_2)'.$$

The minimum distance estimator is the value for α_2 that minimizes $\mathbf{d}'\hat{\mathbf{\Omega}}^{-1}\mathbf{d}$. For this study, $\hat{\mathbf{\Omega}}$ is a diagonal matrix. It is straightforward to show that in this case, the MDE is

$$\hat{\alpha}_2 = \sum_{j=1}^{4} \hat{\alpha}_2^j \left(\frac{1/\hat{\omega}_j}{\sum_{m=1}^{4} 1/\hat{\omega}_m} \right).$$

The final answer is roughly 16%. By implication, then, the authors estimated that $100 - 16 = 84$ percent of the GSE funding advantage was kept within the company or passed through to stockholders.

11.10.3 PARAMETER HETEROGENEITY AND DYNAMIC PANEL DATA MODELS

The analysis in this section has involved static models and relatively straightforward estimation problems. We have seen as this section has progressed that parameter heterogeneity introduces a fair degree of complexity to the treatment. Dynamic effects in the model, with or without heterogeneity, also raise complex new issues in estimation and inference. There are numerous cases in which dynamic effects and parameter heterogeneity coincide in panel data models. This section will explore a few of the specifications and some applications. The familiar estimation techniques (OLS, FGLS, etc.) are not effective in these cases. The proposed solutions are developed in Chapter 8 where we present the technique of instrumental variables and in Chapter 13 where we present the GMM estimator and its application to dynamic panel data models.

Example 11.24 Dynamic Panel Data Models

The antecedent of much of the current research on panel data is Balestra and Nerlove's (1966) study of the natural gas market.[47] The model is a stock-flow description of the derived demand for fuel for gas using appliances. The central equation is a model for total demand,

$$G_{it} = G_{it}^{*} + (1 - r)G_{i,t-1},$$

where G_{it} is current total demand. Current demand consists of new demand, G_{it}^{*}, that is created by additions to the stock of appliances plus old demand, which is a proportion of the previous period's demand, r being the depreciation rate for gas using appliances. New demand is due to net increases in the stock of gas using appliances, which is modeled as

$$G_{it}^{*} = \beta_0 + \beta_1 Price_{it} + \beta_2 \Delta Pop_{it} + \beta_3 Pop_{it} + \beta_4 \Delta Income_{it} + \beta_5 Income_{it} + \varepsilon_{it},$$

where Δ is the first difference (change) operator, $\Delta X_t = X_t - X_{t-1}$. The reduced form of the model is a dynamic equation,

$$G_{it} = \beta_0 + \beta_1 Price_{it} + \beta_2 \Delta Pop_{it} + \beta_3 Pop_{it} + \beta_4 \Delta Income_{it} + \beta_5 Income_{it} + \gamma G_{i,t-1} + \varepsilon_{it}.$$

The authors analyzed a panel of 36 states over a six-year period (1957–1962). Both fixed effects and random effects approaches were considered.

An equilibrium model for steady-state growth has been used by numerous authors [e.g., Robertson and Symons (1992), Pesaran and Smith (1995), Lee, Pesaran, and Smith (1997),

[47] See, also, Nerlove (2002, Chapter 2).

Pesaran, Shin, and Smith (1999), Nerlove (2002) and Hsiao, Pesaran, and Tahmiscioglu (2002)] for cross-industry or -country comparisons. Robertson and Symons modeled real wages in 13 OECD countries over the period 1958–1986 with a wage equation

$$W_{it} = \alpha_i + \beta_{1i} k_{it} + \beta_{2i} \Delta\, wedge_{it} + \gamma_i W_{i,t-1} + \varepsilon_{it},$$

where W_{it} is the real product wage for country i in year t, k_{it} is the capital-labor ratio, and *wedge* is the "tax and import price wedge."

Lee, Pesaran, and Smith (1997) compared income growth across countries with a steady-state income growth model of the form

$$\ln y_{it} = \alpha_i + \theta_i t + \lambda_i \ln y_{i,t-1} + \varepsilon_{it},$$

where $\theta_i = (1 - \lambda_i)\delta_i$, δ_i is the technological growth rate for country i, and λ_i is the convergence parameter. The rate of convergence to a steady state is $1 - \lambda_i$.

Pesaran and Smith (1995) analyzed employment in a panel of 38 UK industries observed over 29 years, 1956–1984. The main estimating equation was

$$\ln e_{it} = \alpha_i + \beta_{1i} t + \beta_{2i} \ln y_{it} + \beta_{3i} \ln y_{i,t-1} + \beta_{4i} \ln \bar{y}_t + \beta_{5i} \ln \bar{y}_{t-1}$$
$$+ \beta_{6i} \ln w_{it} + \beta_{7i} \ln w_{i,t-1} + \gamma_{1i} \ln e_{i,t-1} + \gamma_{2i} \ln e_{i,t-2} + \varepsilon_{it},$$

where y_{it} is industry output, $\bar{y}_t$ is total (not average) output, and w_{it} is real wages.

In the growth models, a quantity of interest is the **long-run multiplier** or **long-run elasticity**. Long-run effects are derived through the following conceptual experiment. The essential feature of the models above is a dynamic equation of the form

$$y_t = \alpha + \beta x_t + \gamma y_{t-1}.$$

Suppose at time t, x_t is fixed from that point forward at $\bar{x}$. The value of y_t at that time will then be $\alpha + \beta\bar{x} + \gamma y_{t-1}$, given the previous value. If this process continues, and if $|\gamma| < 1$, then eventually y_s will reach an equilibrium at a value such that $y_s = y_{s-1} = \bar{y}$. If so, then $\bar{y} = \alpha + \beta\bar{x} + \gamma\bar{y}$, from which we can deduce that $\bar{y} = (\alpha + \bar{x})/(1 - \gamma)$. The path to this equilibrium from time t into the future is governed by the **adjustment equation**

$$y_s - \bar{y} = (y_t - \bar{y})\gamma^{s-t}, s \geq t.$$

The experiment, then, is to ask: What is the impact on the equilibrium of a change in the input, $\bar{x}$? The result is $\partial\bar{y}/\partial\bar{x} = \beta/(1 - \gamma)$. This is the long-run multiplier, or **equilibrium multiplier**, in the model. In the preceding Pesaran and Smith model, the inputs are in logarithms, so the multipliers are long-run elasticities. For example, with two lags of $\ln e_{it}$ in Pesaran and Smith's model, the long-run effects for wages are

$$\phi_i = (\beta_{6i} + \beta_{7i})/(1 - \gamma_{1i} - \gamma_{2i}).$$

In this setting, in contrast to the preceding treatments, the number of units, n, is generally taken to be fixed, though often it will be fairly large. The Penn World Tables (http://cid.econ.ucdavis.edu/pwt.html) that provide the database for many of these analyses now contain information on more than 150 countries for well more than 50 years. Asymptotic results for the estimators are with respect to increasing T, though we will consider, in general, cases in which T is small. Surprisingly, increasing T and n at the same time need not simplify the derivations.

The parameter of interest in many studies is the average long-run effect, say $\bar{\phi} = (1/n)\Sigma_i \phi_i$, in the Pesaran and Smith example. Because n is taken to be fixed, the "parameter" $\bar{\phi}$ is a definable object of estimation—that is, with n fixed, we can speak

of $\bar{\phi}$ as a parameter rather than as an estimator of a parameter. There are numerous approaches one might take. For estimation purposes, pooling, fixed effects, random effects, group means, or separate regressions are all possibilities. (Unfortunately, nearly all are inconsistent.) In addition, there is a choice to be made whether to compute the average of long-run effects or to compute the long-run effect from averages of the parameters. The choice of the average of functions, $\bar{\phi}$ versus the function of averages,

$$\bar{\phi}^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{6i} + \hat{\beta}_{7i})}{1 - \frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_{1i} + \hat{\gamma}_{2i})},$$

turns out to be of substance. For their UK industry study, Pesaran and Smith report estimates of -0.33 for $\bar{\phi}$ and -0.45 for $\bar{\phi}^*$. (The authors do not express a preference for one over the other.)

The development to this point is implicitly based on estimation of separate models for each unit (country, industry, etc.). There are also a variety of other estimation strategies one might consider. We will assume for the moment that the data series are stationary in the dimension of T. (See Chapter 21.) This is a transparently false assumption, as revealed by a simple look at the trends in macroeconomic data, but maintaining it for the moment allows us to proceed. We will reconsider it later.

We consider the generic, dynamic panel data model,

$$y_{it} = \alpha_i + \beta_i x_{it} + \gamma_i y_{i,t-1} + \varepsilon_{it}. \quad \textbf{(11-90)}$$

Assume that T is large enough that the individual regressions can be computed. In the absence of autocorrelation in ε_{it}, it has been shown[48] that the OLS estimator of γ_i is biased downward, but consistent in T. Thus, $E[\hat{\gamma}_i - \gamma_i] = \theta_i/T$ for some θ_i. The implication for the individual estimator of the long-run multiplier, $\phi_i = \beta_i/(1 - \gamma_i)$, is unclear in this case, however. The denominator is overestimated. But it is not clear whether the estimator of β_i is overestimated or underestimated. It is true that whatever bias there is is $O(1/T)$. For this application, T is fixed and possibly quite small. The end result is that it is unlikely that the individual estimator of ϕ_i is unbiased, and by construction, it is inconsistent, because T cannot be assumed to be increasing. If that is the case, then $\hat{\bar{\phi}}$ is likewise inconsistent for $\bar{\phi}$. We are averaging n estimators, each of which has bias and variance that are $O(1/T)$. The variance of the mean is, therefore, $O(1/nT)$ which goes to zero, but the bias remains $O(1/T)$. It follows that the average of the n means is not converging to $\bar{\phi}$; it is converging to the average of whatever these biased estimators are estimating. The problem vanishes with large T, but that is not relevant to the current context. However, in the Pesaran and Smith study, T was 29, which is large enough that these effects are probably moderate. For macroeconomic cross-country studies such as those based on the Penn World Tables, the data series may be even longer than this.

One might consider aggregating the data to improve the results. Pesaran and Smith (1995) suggest an average based on country means. Averaging the observations over T in (11-90) produces

$$\bar{y}_{i.} = \alpha_i + \beta_i \bar{x}_{i.} + \gamma_i \bar{y}_{-1,i} + \bar{\varepsilon}_{i.}. \quad \textbf{(11-91)}$$

A linear regression using the n observations would be inconsistent for two reasons: First, $\bar{\varepsilon}_{i.}$ and $\bar{y}_{-1,i}$ must be correlated. Second, because of the parameter heterogeneity, it is not clear

[48] For example, Griliches (1961) and Maddala and Rao (1973).

without further assumptions what the OLS slopes estimate under the false assumption that all coefficients are equal. But $\bar{y}_{i.}$ and $\bar{y}_{-1,i}$ differ by only the first and last observations; $\bar{y}_{-1,i} = \bar{y}_{i.} - (y_{iT} - y_{i0})/T = \bar{y}_{i.} - [\Delta_T(y)/T]$. Inserting this in (11-91) produces

$$\begin{aligned}\bar{y}_{i.} &= \alpha_i + \beta_i \bar{x}_{i.} + \gamma_i \bar{y}_{i.} - \gamma_i[\Delta_T(y)/T] + \bar{\varepsilon}_{i.} \\ &= \frac{\alpha_i}{1-\gamma_i} + \frac{\beta_i}{1-\gamma_i}\bar{x}_{i.} - \frac{\gamma_i}{1-\gamma_i}[\Delta_T(y)/T] + \bar{\varepsilon}_{i.} \\ &= \delta_i + \phi_i \bar{x}_{i.} + \tau_i[\Delta_T(y)/T] + \bar{\varepsilon}_{i.}.\end{aligned} \tag{11-92}$$

We still seek to estimate $\overline{\phi}$. The form in (11-92) does not solve the estimation problem, because the regression suggested using the group means is still heterogeneous. If it could be assumed that the individual long-run coefficients differ randomly from the averages in the fashion of the random parameters model of Section 11.10.1, so $\delta_i = \overline{\delta} + u_{\delta,i}$ and likewise for the other parameters, then the model could be written

$$\begin{aligned}\bar{y}_{i.} &= \overline{\delta} + \overline{\phi}\bar{x}_{i.} + \overline{\tau}[\Delta_T(y)/T]_i + \bar{\varepsilon}_{i.} + \{u_{\delta,i} + u_{\phi,i}\bar{x}_i + u_{\tau,i}[\Delta_T(y)/T]_i\} \\ &= \overline{\delta} + \overline{\phi}\bar{x}_{i.} + \overline{\tau}[\Delta_T(y)/T]_i + \bar{\varepsilon}_i + w_i.\end{aligned}$$

At this point, the equation appears to be a heteroscedastic regression amenable to least squares estimation, but for one loose end. Consistency follows if the terms $[\Delta_T(y)/T]_i$ and $\bar{\varepsilon}_i$ are uncorrelated. Because the first is a rate of change and the second is in levels, this should generally be the case. Another interpretation that serves the same purpose is that the rates of change in $[\Delta_T(y)/T]_i$ should be uncorrelated with the levels in $\bar{x}_{i.}$, in which case, the regression can be partitioned, and simple linear regression of the country means of y_{it} on the country means of x_{it} and a constant produces consistent estimates of $\overline{\phi}$ and $\overline{\delta}$.

Alternatively, consider a time-series approach. We average the observation in (11-90) across countries at each time period rather than across time within countries. In this case, we have

$$\bar{y}_{.t} = \overline{\alpha} + \frac{1}{n}\sum_{i=1}^{n}\beta_i x_{it} + \frac{1}{n}\sum_{i=1}^{n}\gamma_i y_{i,t-1} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{it}.$$

Let $\overline{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i$ so that $\gamma_i = \overline{\gamma} + (\gamma_i - \overline{\gamma})$ and $\beta_i = \overline{\beta} + (\beta_i - \overline{\beta})$. Then,

$$\begin{aligned}\bar{y}_{.t} &= \overline{\alpha} + \overline{\beta}\bar{x}_{.t} + \overline{\gamma}\bar{y}_{-1,t} + [\bar{\varepsilon}_{.t} + (\beta_i - \overline{\beta})\bar{x}_{.t} + (\gamma_i - \overline{\gamma})\bar{y}_{-1,t}] \\ &= \overline{\alpha} + \overline{\beta}\bar{x}_{.t} + \overline{\gamma}\bar{y}_{-1,t} + \bar{\varepsilon}_{.t} + w_{.t}.\end{aligned}$$

Unfortunately, the regressor, $\overline{\gamma}\bar{y}_{-1,t}$ is surely correlated with $w_{.t}$, so neither OLS or GLS will provide a consistent estimator for this model. (One might consider an instrumental variable estimator; however, there is no natural instrument available in the model as constructed.) Another possibility is to pool the entire data set, possibly with random or fixed effects for the constant terms. Because pooling, even with country-specific constant terms, imposes homogeneity on the other parameters, the same problems we have just observed persist.

Finally, returning to (11-90), one might treat it as a formal random parameters model,

$$\begin{aligned}y_{it} &= \alpha_i + \beta_i x_{it} + \gamma_i y_{i,t-1} + \varepsilon_{it}, \\ \alpha_i &= \alpha + u_{\alpha,i}, \\ \beta_i &= \beta + u_{\beta,i}, \\ \gamma_i &= \gamma + u_{\gamma,i}.\end{aligned} \tag{11-93}$$

The assumptions needed to formulate the model in this fashion are those of the previous section. As Pesaran and Smith (1995) observe, this model can be estimated using the Swamy (1971) estimator, which is the matrix weighted average of the least squares estimators discussed in Section 11.11.1. The estimator requires that T be large enough to fit each country regression by least squares. That has been the case for the received applications. Indeed, for the applications we have examined, both n and T are relatively large. If not, then one could still use the mixed models approach developed in Chapter 15. A compromise that appears to work well for panels with moderate sized n and T is the "mixed-fixed" model suggested in Hsiao (1986, 2003) and Weinhold (1999). The dynamic model in (11-92) is formulated as a partial fixed effects model,

$$y_{it} = \alpha_i d_{it} + \beta_i x_{it} + \gamma_i d_{it} y_{i,t-1} + \varepsilon_{it},$$
$$\beta_i = \beta + u_{\beta,i},$$

where d_{it} is a dummy variable that equals one for country i in every period and zero otherwise (i.e., the usual fixed effects approach). Note that d_{it} also appears with $y_{i,t-1}$. As stated, the model has "fixed effects," one random coefficient, and a total of $2n + 1$ coefficients to estimate, in addition to the two variance components, σ_ε^2 and σ_u^2. The model could be estimated inefficiently by using ordinary least squares—the random coefficient induces heteroscedasticity (see Section 11.10.1)—by using the Hildreth–Houck–Swamy approach, or with the mixed linear model approach developed in Chapter 15.

Example 11.25 A Mixed Fixed Growth Model for Developing Countries

Weinhold (1996) and Nair–Reichert and Weinhold (2001) analyzed growth and development in a panel of 24 developing countries observed for 25 years, 1971–1995. The model they employed was a variant of the mixed-fixed model proposed by Hsiao (1986, 2003). In their specification,

$$GGDP_{i,t} = \alpha_i d_{it} + \gamma_i d_{it} GGDP_{i,t-1} + \beta_{1i} GGDI_{i,t-1} + \beta_{2i} GFDI_{i,t-1} + \beta_{3i} GEXP_{i,t-1} + \beta_4 INFL_{i,t-1} + \varepsilon_{it},$$

where

GGDP = Growth rate of gross domestic product,
GGDI = Growth rate of gross domestic investment,
GFDI = Growthrate of foreign direct investment (inflows),
GEXP = Growth rate of exports of goods and services,
INFL = Inflation rate.

11.11 SUMMARY AND CONCLUSIONS

This chapter has shown a few of the extensions of the classical model that can be obtained when panel data are available. In principle, any of the models we have examined before this chapter and all those we will consider later, including the multiple equation models, can be extended in the same way. The main advantage, as we noted at the outset, is that with panel data, one can formally model dynamic effects and the heterogeneity across groups that are typical in microeconomic data.

Key Terms and Concepts

- Adjustment equation
- Arellano and Bond's estimator
- Balanced panel
- Between groups
- Contiguity
- Contiguity matrix
- Contrasts
- Dynamic panel data model
- Equilibrium multiplier
- Error components model
- Estimator
- Feasible GLS
- First difference
- Fixed effects
- Fixed panel
- Group means
- Group means estimator
- Hausman specification test
- Heterogeneity
- Hierarchical model
- Incidental parameters problem
- Index function model
- Individual effect
- Instrumental variable
- Instrumental variable estimator
- Lagrange multiplier test
- Least squares dummy variable model (LSDV)
- Long run elasticity
- Long run multiplier
- Longitudinal data set
- Matrix weighted average
- Mundlak's approach
- Panel data
- Partial effects
- Pooled model
- Projections
- Rotating panel
- Spatial autocorrelation
- Spatial autoregression coefficient
- Spatial error correlation
- Spatial lags
- Specification test
- Strict exogeneity
- Time invariant
- Unbalanced panel
- Within groups

Exercises

1. The following is a panel of data on investment (y) and profit (x) for $n = 3$ firms over $T = 10$ periods.

	$i = 1$		$i = 2$		$i = 3$	
t	y	x	y	x	y	x
1	13.32	12.85	20.30	22.93	8.85	8.65
2	26.30	25.69	17.47	17.96	19.60	16.55
3	2.62	5.48	9.31	9.16	3.87	1.47
4	14.94	13.79	18.01	18.73	24.19	24.91
5	15.80	15.41	7.63	11.31	3.99	5.01
6	12.20	12.59	19.84	21.15	5.73	8.34
7	14.93	16.64	13.76	16.13	26.68	22.70
8	29.82	26.45	10.00	11.61	11.49	8.36
9	20.32	19.64	19.51	19.55	18.49	15.44
10	4.77	5.43	18.32	17.06	20.84	17.87

   a. Pool the data and compute the least squares regression coefficients of the model

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}.$$

   b. Estimate the fixed effects model of (11-11), and then test the hypothesis that the constant term is the same for all three firms.
   c. Estimate the random effects model of (11-28), and then carry out the Lagrange multiplier test of the hypothesis that the classical model without the common effect applies.
   d. Carry out Hausman's specification test for the random versus the fixed effect model.

2. Suppose that the fixed effects model is formulated with an overall constant term and $n - 1$ dummy variables (dropping, say, the last one). Investigate the effect that this supposition has on the set of dummy variable coefficients and on the least squares estimates of the slopes, compared to (11-13).
3. *Unbalanced design for random effects.* Suppose that the random effects model of Section 11.5 is to be estimated with a panel in which the groups have different numbers of observations. Let T_i be the number of observations in group i.
   a. Show that the pooled least squares estimator is unbiased and consistent despite this complication.
   b. Show that the estimator in (11-40) based on the pooled least squares estimator of $\boldsymbol{\beta}$ (or, for that matter, *any* consistent estimator of $\boldsymbol{\beta}$) is a consistent estimator of σ_ε^2.
4. What are the probability limits of $(1/n)$ LM, where LM is defined in (11-42) under the null hypothesis that $\sigma_u^2 = 0$ and under the alternative that $\sigma_u^2 \neq 0$?
5. *A two-way fixed effects model.* Suppose that the fixed effects model is modified to include a time-specific dummy variable as well as an individual-specific variable. Then $y_{it} = \alpha_i + \gamma_t + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$. At every observation, the individual- and time-specific dummy variables sum to 1, so there are some redundant coefficients. The discussion in Section 11.4.4 shows that one way to remove the redundancy is to include an overall constant and drop one of the time-specific *and* one of the time dummy variables. The model is, thus,

$$y_{it} = \mu + (\alpha_i - \alpha_1) + (\gamma_t - \gamma_1) + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

(Note that the respective time- or individual-specific variable is zero when t or i equals one.) Ordinary least squares estimates of $\boldsymbol{\beta}$ are then obtained by regression of $y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{\bar{y}}$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{.t} + \bar{\bar{\mathbf{x}}}$. Then $(\alpha_i - \alpha_1)$ and $(\gamma_t - \gamma_1)$ are estimated using the expressions in (11-25). Using the following data, estimate the full set of coefficients for the least squares dummy variable model:

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$
					$i = 1$					
y	21.7	10.9	33.5	22.0	17.6	16.1	19.0	18.1	14.9	23.2
x_1	26.4	17.3	23.8	17.6	26.2	21.1	17.5	22.9	22.9	14.9
x_2	5.79	2.60	8.36	5.50	5.26	1.03	3.11	4.87	3.79	7.24
					$i = 2$					
y	21.8	21.0	33.8	18.0	12.2	30.0	21.7	24.9	21.9	23.6
x_1	19.6	22.8	27.8	14.0	11.4	16.0	28.8	16.8	11.8	18.6
x_2	3.36	1.59	6.19	3.75	1.59	9.87	1.31	5.42	6.32	5.35
					$i = 3$					
y	25.2	41.9	31.3	27.8	13.2	27.9	33.3	20.5	16.7	20.7
x_1	13.4	29.7	21.6	25.1	14.1	24.1	10.5	22.1	17.0	20.5
x_2	9.57	9.62	6.61	7.24	1.64	5.99	9.00	1.75	1.74	1.82
					$i = 4$					
y	15.3	25.9	21.9	15.5	16.7	26.1	34.8	22.6	29.0	37.1
x_1	14.2	18.0	29.9	14.1	18.4	20.1	27.6	27.4	28.5	28.6
x_2	4.09	9.56	2.18	5.43	6.33	8.27	9.16	5.24	7.92	9.63

Test the hypotheses that (1) the *period* effects are all zero, (2) the *group* effects are all zero, and (3) both period and group effects are zero. Use an F test in each case.

6. *Two-way random effects model.* We modify the random effects model by the addition of a time-specific disturbance. Thus,

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + u_i + v_t,$$

where

$$\begin{aligned}
E[\varepsilon_{it}|\mathbf{X}] &= E[u_i|\mathbf{X}] = E[v_t|\mathbf{X}] = 0,\\
E[\varepsilon_{it}u_j|\mathbf{X}] &= E[\varepsilon_{it}v_s|\mathbf{X}] = E[u_iv_t|\mathbf{X}] = 0 \quad \text{for all } i, j, t, s,\\
\text{Var}[\varepsilon_{it}|\mathbf{X}] &= \sigma_\varepsilon^2, \quad \text{Cov}[\varepsilon_{it}, \varepsilon_{js}|\mathbf{X}] = 0 \quad \text{for all } i, j, t, s,\\
\text{Var}[u_i|\mathbf{X}] &= \sigma_u^2, \quad \text{Cov}[u_i, u_j|\mathbf{X}] = 0 \quad \text{for all } i, j,\\
\text{Var}[v_t|\mathbf{X}] &= \sigma_v^2, \quad \text{Cov}[v_t, v_s|\mathbf{X}] = 0 \quad \text{for all } t, s.
\end{aligned}$$

Write out the full disturbance covariance matrix for a data set with $n = 2$ and $T = 2$.

7. In Section 11.4.5, we found that the group means of the time-varying variables would work as a control function in estimation of the fixed effects model. That is, although regression of $\mathbf{y}$ on $\mathbf{X}$ is inconsistent for $\boldsymbol{\beta}$, the Mundlak estimator, regression of $\mathbf{y}$ on $\mathbf{X}$ and $\overline{\mathbf{X}} = \mathbf{P_D X} = (\mathbf{I} - \mathbf{M_D})\mathbf{X}$ is a consistent estimator. Would the deviations from group means, $\ddot{\mathbf{X}} = \mathbf{M_D X} = (\mathbf{X} - \overline{\mathbf{X}})$, also be useable as a control function estimator. That is, does regression of $\mathbf{y}$ on $(\mathbf{X}, \ddot{\mathbf{X}})$ produce a consistent estimator of $\boldsymbol{\beta}$?
8. Prove plim $(1/nT)\mathbf{X}'\mathbf{M_D}\boldsymbol{\varepsilon} = \mathbf{0}$.
9. If the panel has $T = 2$ periods, the LSDV (within groups) estimator gives the same results as first differences. Prove this claim.

Applications

The following applications require econometric software.

1. Several applications in this and previous chapters have examined the returns to education in panel data sets. Specifically, we applied Hausman and Taylor's approach in Examples 11.17 and 11.18. Example 11.18 used Cornwell and Rupert's data for the analysis. Koop and Tobias's (2004) study that we used in Chapters 3 and 5 provides yet another application that we can use to continue this analysis. The data may be downloaded from the *Journal of Applied Econometrics* data archive at http://qed.econ.queensu.ca/jae/2004-vl9.7/koop-tobias/. The data file is in two parts. The first file contains the full panel of 17,919 observations on variables:

Column 1; *Person id* (ranging from 1 to 2,178),
Column 2; *Education*,
Column 3; *Log of hourly wage*,
Column 4; *Potential experience*,
Column 5; *Time trend*.

Columns 2 through 5 contain time-varying variables. The second part of the data set contains time-invariant variables for the 2,178 households. These are:

Column 1; *Ability*,
Column 2; *Mother's education*,
Column 3; *Father's education*,
Column 4; *Dummy variable for residence in a broken home*,
Column 5; *Number of siblings*.

To create the data set for this exercise, it is necessary to merge these two data files. The ith observation in the second file will be replicated T_i times for the set of T_i observations in the first file. The *person id* variable indicates which rows must contain the data from the second file. (How this preparation is carried out will vary from one computer package to another.) The panel is quite unbalanced; the number of observations by group size is:

Value of T_i

1:83,	2:104,	3:102,	4:116
5:148,	6:165,	7:201,	8:202
9:200,	10:202,	11:182,	12:148
13:136,	14:96,	15:93	

a. Using these data, fit fixed and random effects models for log wage and examine the result for the return to education.
b. For a Hausman–Taylor specification, consider the following:

$\mathbf{x}_1$ = potential experience, ability
$\mathbf{x}_2$ = education
$\mathbf{f}_1$ = constant, number of siblings, broken home
$\mathbf{f}_2$ = mother's education, father's education

Based on this specification, what is the estimated return to education? (*Note:* you may need the average value of $1/T_i$ for your calculations. This is 0.1854.)
c. It might seem natural to include ability with education in $\mathbf{x}_2$. What becomes of the Hausman and Taylor estimator if you do so?
d. Using a different specification, compute an estimate of the return to education using the instrumental variables method.
e. Compare your results in parts b and d to the results in Examples 11.17 and 11.18. The estimated return to education is surprisingly stable.

2. The data in Appendix Table F10.4 were used by Grunfeld (1958) and dozens of researchers since, including Zellner (1962, 1963) and Zellner and Huang (1962), to study different estimators for panel data and linear regression systems. [See Kleiber and Zeileis (2010).] The model is an investment equation,

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it}, t = 1, \ldots, 20, i = 1, \ldots, 10,$$

where

I_{it} = real gross investment for firm i in year t,
F_{it} = real value of the firm—shares outstanding,
C_{it} = real value of the capital stock.

For present purposes, this is a balanced panel data set.

a. Fit the pooled regression model.
b. Referring to the results in part a, is there evidence of within-groups correlation? Compute the robust standard errors for your pooled OLS estimator and compare them to the conventional ones.
c. Compute the fixed effects estimator for these data. Then, using an F test, test the hypothesis that the constants for the 10 firms are all the same.
d. Use a Lagrange multiplier statistic to test for the presence of common effects in the data.
e. Compute the one-way random effects estimator and report all estimation results. Explain the difference between this specification and the one in part c.
f. Use a Hausman test to determine whether a fixed or random effects specification is preferred for these data.

3. The data in Appendix Table F6.1 are an unbalanced panel on 25 U.S. airlines in the pre-deregulation days of the 1970s and 1980s. The group sizes range from 2 to 15. Data in the file are the following variables. (Variable names contained in the data file are constructed to indicate the variable contents.)

   Total cost,
   Expenditures on Capital, Labor, Fuel, Materials, Property, and Equipment,
   Price measures for the six inputs,
   Quantity measures for the six inputs,
   Output measured in revenue passenger miles, converted to an index number for the airline,
   Load factor = the average percentage capacity utilization of the airline's fleet,
   Stage = the average flight (stage) length in miles,
   Points = the number of points served by the airline,
   Year = the calendar year,
   T Year = 1969,
   TI = the number of observations for the airline, repeated for each year.

   Use these data to build a cost model for airline service. Allow for cross-airline heterogeneity in the constants in the model. Use both random and fixed effects specifications, and use available statistical tests to determine which is the preferred model. An appropriate cost model to begin the analysis with would be

$$\ln cost_{it} = \alpha_i + \sum_{k=1}^{6} \beta_k \ln Price_{k,it} + \gamma \ln Output_{it} + \varepsilon_{it}.$$

   It is necessary to impose linear homogeneity in the input prices on the cost function, which you would do by dividing five of the six prices and the total cost by the sixth price (choose any one), then using $\ln(cost/P_6)$ and $\ln(P_k/P_6)$ in the regression. You might also generalize the cost function by including a quadratic term in the log of output in the function. A translog model would include the unique squares and cross products of the input prices and products of log output with the logs of the prices. The data include three additional factors that may influence costs, stage length, load factor, and number of points served. Include them in your model, and use the appropriate test statistic to test whether they are, indeed, relevant to the determination of (log) total cost.

12

ESTIMATION FRAMEWORKS IN ECONOMETRICS

12.1 INTRODUCTION

This chapter begins our treatment of methods of estimation. Contemporary econometrics offers the practitioner a remarkable variety of estimation methods, ranging from tightly parameterized likelihood-based techniques at one end to thinly stated nonparametric methods that assume little more than mere association between variables at the other, and a rich variety in between. Even the experienced researcher could be forgiven for wondering how to choose from this long menu. It is certainly beyond our scope to answer this question here, but a few principles will be suggested. Recent research has leaned, when possible, toward methods that require few (or fewer) possibly unwarranted or improper assumptions. This explains the ascendance of the GMM estimator in situations where strong likelihood-based parameterizations can be avoided and robust estimation can be done in the presence of heteroscedasticity and serial correlation. (It is intriguing to observe that this is occurring at a time when advances in computation have helped bring about *increased* acceptance of very heavily parameterized Bayesian methods.)

As a general proposition, the progression from full to semiparametric to nonparametric estimation relaxes strong assumptions, but at the cost of weakening the conclusions that can be drawn from the data. As much as anywhere else, this is clear in the analysis of discrete choice models, which provide one of the most active literatures in the field. (A sampler appears in Chapter 17.) A formal probit or logit model allows estimation of probabilities, partial effects, and a host of ancillary results, but at the cost of imposing the normal or logistic distribution on the data. **Semiparametric estimators** and **nonparametric estimators** allow one to relax the restriction but often provide, in return, only ranges of probabilities, if that, and in many cases, preclude estimation of probabilities or useful partial effects. The conclusions drawn based on the nonparametric and semiparametric estimators, such as they are, are robust.[1]

Estimation properties is another arena in which the different approaches can be compared. Within a class of estimators, one can define the best (most efficient) means of using the data. (See Example 12.2 for an application.) Sometimes comparisons can be made across classes as well. For example, when they are estimating the same parameters—this remains to be established—the best parametric estimator will generally outperform the best semiparametric estimator. That is the value of the additional information used by the parametric estimator, of course. The other side of the comparison, however, is that the semiparametric estimator will carry the day if the parametric model is misspecified in a fashion to which the semiparametric estimator is robust (and the parametric model is not).

[1]See, for example, the symposium in Angrist and Pischke (2010) for a spirited discussion on these points.

Schools of thought have punctuated this conversation. Proponents of **Bayesian estimation** often took an almost theological viewpoint in their criticism of their classical colleagues.[2] Contemporary practitioners are usually more pragmatic than this. Bayesian estimation has gained currency as a set of techniques that can, in very many cases, provide both elegant and tractable solutions to problems that have heretofore been out of reach.[3] Thus, for example, the **simulation-based estimation** advocated in the many papers of Chib and Greenberg (for example, 1996) have provided solutions to a variety of computationally challenging problems. Arguments as to the methodological virtue of one approach or the other have received much less attention than before.

Chapters 2 through 7 of this book have focused on the classical regression model and a particular estimator, least squares (linear and nonlinear). In this and the next four chapters, we will examine several general estimation strategies that are used in a wide variety of situations. This chapter will survey a few methods in the three broad areas we have listed. Chapter 13 discusses the **generalized method of moments**, which has emerged as the centerpiece of semiparametric estimation. Chapter 14 presents the method of **maximum likelihood**, the broad platform for parametric, classical estimation in econometrics. Chapter 15 discusses simulation-based estimation and bootstrapping. This is a body of techniques that have been made feasible by advances in estimation technology and which have made quite straightforward many estimators that were previously only scarcely used because of the sheer difficulty of the computations. Finally, Chapter 16 introduces the methods of Bayesian econometrics.

The list of techniques presented here is far from complete. We have chosen a set that constitutes the mainstream of econometrics. Certainly there are others that might be considered.[4] Virtually all of them are the subjects of excellent monographs on the subject. In this chapter we will present several applications, some from the literature, some home grown, to demonstrate the range of techniques that are current in econometric practice. We begin in Section 12.2 with parametric approaches, primarily maximum likelihood. Because this is the subject of much of the remainder of this book, this section is brief. Section 12.2 also introduces Bayesian estimation, which in its traditional form is as heavily parameterized as maximum likelihood estimation. Section 12.3 is on semiparametric estimation. GMM estimation is the subject of all of Chapter 13, so it is only introduced here. The technique of least absolute deviations is presented here as well. A range of applications from the recent literature is also surveyed. Section 12.4 describes nonparametric estimation. The fundamental tool, the kernel density estimator, is developed, then applied to a problem in regression analysis. Two applications are presented here as well. Being focused on application, this chapter will say very little about the statistical theory for these techniques—such as their asymptotic properties. (The results are developed at length in the literature, of course.) We will turn to the subject of the properties of estimators briefly at the end of the chapter, in Section 12.5, then in greater detail in Chapters 13 through 16.

[2]See, for example, Poirier (1995).

[3]The penetration of Bayesian methods in econometrics could be overstated. It is quite well represented in current journals such as the *Journal of Econometrics, Journal of Applied Econometrics, Journal of Business and Economic Statistics*, and so on. On the other hand, of the six major general treatments of econometrics published in 2000, four (Hayashi, Ruud, Patterson, Davidson) do not mention Bayesian methods at all. A buffet of 32 essays (Baltagi) devotes only one to the subject. Likewise, Wooldridge's (2010) widely cited treatise contains no mention of Bayesian econometrics. The one that displays any preference [for example, Mittelhammer et al. (2000)] devotes nearly 10% (70) of its pages to Bayesian estimation, but all to the broad metatheory of the linear regression model and none to the more elaborate applications that form the received applications in the many journals in the field.

[4]See, for example, Mittelhammer, Judge, and Miller (2000) for a lengthy catalog.

12.2 PARAMETRIC ESTIMATION AND INFERENCE

Parametric estimation departs from a full statement of the **density** or probability model that provides the **data-generating mechanism** for a random variable of interest. For the sorts of applications we have considered thus far, we might say that the joint density of a scalar random variable, y, and a random vector, $\mathbf{x}$, of interest can be specified by

$$f(y, \mathbf{x}) = g(y|\mathbf{x}, \boldsymbol{\beta}) \times h(\mathbf{x}|\boldsymbol{\theta}), \tag{12-1}$$

with unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. To continue the application that has occupied us since Chapter 2, consider the linear regression model with normally distributed disturbances. The assumption produces a full statement of the **conditional density** that is the population from which an observation is drawn,

$$y_i|\mathbf{x}_i \sim N[\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2].$$

All that remains for a full definition of the population is knowledge of the specific values taken by the *unknown*, but *fixed*, parameters. With those in hand, the conditional probability distribution for y_i is completely defined—mean, variance, probabilities of certain events, and so on. (The marginal density for the conditioning variables is usually not of particular interest.) Thus, the signature features of this modeling platform are specifications of both the density and the features (parameters) of that density.

The **parameter space** for the parametric model is the set of allowable values of the parameters that satisfy some prior specification of the model. For example, in the regression model specified previously, the K regression slopes may take any real value, but the variance must be a positive number. Therefore, the parameter space for that model is $[\beta, \sigma^2] \in \mathbb{R}^K \times \mathbb{R}+$. *Estimation* in this context consists of specifying a criterion for ranking the points in the parameter space, then choosing that point (a point estimate) or a set of points (an interval estimate) that optimizes that criterion, that is, has the best ranking. Thus, for example, we chose linear least squares as one estimation criterion for the linear model. *Inference* in this setting is a process by which some regions of the (already specified) parameter space are deemed not to contain the unknown parameters, though, in more practical terms, we typically define a criterion and then state that, by that criterion, certain regions are *unlikely* to contain the true parameters.

12.2.1 CLASSICAL LIKELIHOOD-BASED ESTIMATION

The most common (by far) class of parametric estimators used in econometrics is the maximum likelihood estimators. The underlying philosophy of this class of estimators is the idea of sample information. When the density of a sample of observations is completely specified, apart from the unknown parameters, then the joint density of those observations (assuming they are independent), is the **likelihood function**

$$f(y_1, y_2, \ldots, \mathbf{x}_1, \mathbf{x}_2, \ldots) = \prod_{i=1}^{n} f(y_i, \mathbf{x}_i|\boldsymbol{\beta}, \boldsymbol{\theta}). \tag{12-2}$$

This function contains all the information available in the sample about the population from which those observations were drawn. The strategy by which that information is used in estimation constitutes the estimator.

The **maximum likelihood estimator** [Fisher (1925)] is the function of the data that (as its name implies) maximizes the likelihood function (or, because it is usually more

convenient, the log of the likelihood function). The motivation for this approach is most easily visualized in the setting of a discrete random variable. In this case, the likelihood function gives the joint probability for the sample data, and the maximum likelihood estimator is the function of the sample information that makes the observed data most probable (at least by that criterion). Though the analogy is most intuitively appealing for a discrete variable, it carries over to continuous variables as well. Because this estimator is the subject of Chapter 14, which is quite lengthy, we will defer any formal discussion until then and consider instead two applications to illustrate the techniques and underpinnings.

Example 12.1 The Linear Regression Model

Least squares weighs negative and positive deviations equally and gives disproportionate weight to large deviations in the calculation. This property can be an advantage or a disadvantage, depending on the data-generating process. For normally distributed disturbances, this method is precisely the one needed to use the data most efficiently. If the data are generated by a normal distribution, then the log of the likelihood function is

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

You can easily show that least squares is the estimator of choice for this model. Maximizing the function means minimizing the exponent, which is done by least squares for $\boldsymbol{\beta}$, then $\mathbf{e}'\mathbf{e}/n$ follows as the estimator for σ^2.

If the appropriate distribution is deemed to be something other than normal—perhaps on the basis of an observation that the tails of the disturbance distribution are too thick (see Example 14.8 and Section 14.9.2) then there are three ways one might proceed. First, as we have observed, the consistency of least squares is robust to this failure of the specification so long as the conditional mean of the disturbances is still zero. Some correction to the standard errors is necessary for proper inferences. Second, one might want to proceed to an estimator with better finite sample properties. The least absolute deviations estimator discussed in Section 12.3.3 is a candidate. Finally, one might consider some other distribution which accommodates the observed discrepancy. For example, Ruud (2000) examines in some detail a linear regression model with disturbances distributed according to the t distribution with v degrees of freedom. As long as v is finite, this random variable will have a larger variance than the normal. Which way should one proceed? The third approach is the least appealing. Surely if the normal distribution is inappropriate, then it would be difficult to come up with a plausible mechanism whereby the t distribution would be. The LAD estimator might well be preferable if the sample were small. If not, then least squares would probably remain the estimator of choice, with some allowance for the fact that standard inference tools would probably be misleading. Current practice is generally to adopt the first strategy.

Example 12.2 The Stochastic Frontier Model

The **stochastic frontier model**, discussed in detail in Chapter 19, is a regression-like model with a disturbance distribution that is asymmetric and distinctly nonnormal. The conditional density for the dependent variable in this skew normal model is

$$f(y|\mathbf{x}, \boldsymbol{\beta}, \sigma, \lambda) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp\left[\frac{-(y - \alpha - \mathbf{x}'\boldsymbol{\beta})^2}{2\sigma^2}\right]\Phi\left(\frac{-\lambda(y - \alpha - \mathbf{x}'\boldsymbol{\beta})}{\sigma}\right).$$

This produces a log-likelihood function for the model,

$$\ln L = -n\ln\sigma - \frac{n}{2}\ln\frac{2}{\pi} - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sigma}\right)^2 + \sum_{i=1}^{n}\ln\Phi\left(\frac{-\lambda\varepsilon_i}{\sigma}\right).$$

There are at least two fully parametric estimators for this model. The maximum likelihood estimator is discussed in Section 19.2.4. Greene (2007a) presents the following **method of moments** estimator: For the regression slopes, excluding the constant term, use least squares. For the parameters α, σ, and λ, based on the second and third moments of the least squares residuals and the least squares constant, solve

$$m_2 = \sigma_v^2 + [1 - 2/\pi]\sigma_u^2,$$
$$m_3 = (2/\pi)^{1/2}[1 - 4/\pi]\sigma_u^3,$$
$$a = \alpha + (2/\pi)^2\sigma_u,$$

where $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_u^2/\sigma_v^2$.

Both estimators are fully parametric. The maximum likelihood estimator is for the reasons discussed earlier. The method of moments estimators (see Section 13.2) are appropriate only for this distribution. Which is preferable? As we will see in Chapter 19, both estimators are consistent and asymptotically normally distributed. By virtue of the Cramér–Rao theorem, the maximum likelihood estimator has a smaller asymptotic variance. Neither has any small sample optimality properties. Thus, the only virtue of the method of moments estimator is that one can compute it with any standard regression/statistics computer package and a hand calculator whereas the maximum likelihood estimator requires specialized software (only somewhat—it is reasonably common).

12.2.2 MODELING JOINT DISTRIBUTIONS WITH COPULA FUNCTIONS

Specifying the likelihood function commits the analyst to a possibly strong assumption about the distribution of the random variable of interest. The payoff, of course, is the stronger inferences that this permits. However, when there is more than one random variable of interest, such as in a joint household decision on health care usage in the example to follow, formulating the full likelihood involves specifying the marginal distributions, which might be comfortable, and a full specification of the joint distribution, which is likely to be less so. In the typical situation, the model might involve two similar random variables and an ill-formed specification of correlation between them. Implicitly, this case involves specification of the marginal distributions. The joint distribution is an empirical necessity to allow the correlation to be nonzero. The **copula function** approach provides a mechanism that the researcher can use to steer around this situation.

Trivedi and Zimmer (2007) suggest a variety of applications that fit this description:

- Financial institutions are often concerned with the prices of different, related (dependent) assets. The typical multivariate normality assumption is problematic because of GARCH effects (see Section 20.13) and thick tails in the distributions. While specifying appropriate marginal distributions may be reasonably straightforward, specifying the joint distribution is anything but that. Klugman and Parsa (2000) is an application.
- There are many microeconometric applications in which straightforward marginal distributions cannot be readily combined into a natural joint distribution. The bivariate event count model analyzed in Munkin and Trivedi (1999) and in the next example is an application.
- In the linear self-selection model of Chapter 19, the necessary joint distribution is part of a larger model. The likelihood function for the observed outcome involves the joint distribution of a variable of interest, hours, wages, income, and so on, and

the probability of observation. The typical application is based on a joint normal distribution. Smith (2003, 2005) suggests some applications in which a flexible copula representation is more appropriate. [In an intriguing early application of copula modeling that was not labeled as such, since it greatly predates the econometric literature, Lee (1983) modeled the outcome variable in a selectivity model as normal, the observation probability as logistic, and the connection between them using what amounted to the "Gaussian" copula function shown next.]

- Cheng and Trivedi (2015) used a copula function as an alternative to the bivariate normal distribution in analyzing attrition in a panel data set. (This application is examined in Example 11.3.)

Although the antecedents in the statistics literature date to Sklar's (1973) derivations, the applications in econometrics and finance are quite recent, with most applications appearing since 2000.[5]

Consider a modeling problem in which the marginal cdfs of two random variables can be fully specified as $F_1(y_1|\bullet)$ and $F_2(y_2|\bullet)$, where we condition on sample information (data) and parameters denoted "$\bullet$." For the moment, assume these are continuous random variables that obey all the axioms of probability. The bivariate cdf is $F_{12}(y_1, y_2|\bullet)$. A (bivariate) copula function (the results also extend to multivariate functions) is a function $C(u_1, u_2)$ defined over the unit square $[(0 \leq u_1 \leq 1) \times (0 \leq u_2 \leq 1)]$ that satisfies

$$\begin{aligned}
&(1)\quad C(1, u_2) = u_2 \text{ and } C(u_1, 1) = u_1,\\
&(2)\quad C(0, u_2) = C(u_1, 0) = 0,\\
&(3)\quad \partial C(u_1, u_2)/\partial u_1 \geq 0 \text{ and } \partial C(u_1, u_2)/\partial u_2 \geq 0.
\end{aligned}$$

These are properties of bivariate cdfs for random variables u_1 and u_2 that are bounded in the unit square. It follows that the copula function is a two-dimensional cdf defined over the unit square that has one-dimensional marginal distributions that are standard uniform in the unit interval [that is, property (1)]. To make profitable use of this relationship, we note that the cdf of a random variable, $F_1(y_1|\bullet)$, is, itself, a uniformly distributed random variable. This is the **fundamental probability transform** that we use for generating random numbers. (See Section 15.2.) In **Sklar's theorem**(1973), the marginal cdfs play the roles of u_1 and u_2. The theorem states that there exists a copula function, $C(...,...)$ such that

$$F_{12}(y_1, y_2|\bullet) = C[F_1(y_1|\bullet), F_2(y_2|\bullet)].$$

If $F_{12}(y_1, y_2|\bullet) = C[F_1(y_1|\bullet), F_2(y_2|\bullet)]$ is continuous and if the marginal cdfs have quantile (inverse) functions $F_j^{-1}(u_j)$ where $0 \leq u_j \leq 1$, then the copula function can be expressed as

$$\begin{aligned}
F_{12}(y_1, y_2|\bullet) &= F_{12}[F_1^{-1}(u_1|\bullet), F_2^{-1}(u_2|\bullet)]\\
&= \text{Prob}[U_1 \leq u_1, U_2 \leq u_2]\\
&= C(u_1, u_2).
\end{aligned}$$

[5]See the excellent survey by Trivedi and Zimmer (2007) for an extensive description.

In words, the theorem implies that the joint density can be written as the copula function evaluated at the two cumulative probability functions.

Copula functions allow the analyst to assemble joint distributions when only the marginal distributions can be specified. To fill in the desired element of correlation between the random variables, the copula function is written

$$F_{12}(y_1, y_2|\bullet) = C[F_1(y_1|\bullet), F_2(y_2|\bullet), \theta],$$

where θ is a dependence parameter. For continuous random variables, the joint pdf is then the mixed partial derivative,

$$\begin{aligned} f_{12}(y_1, y_2|\bullet) &= c_{12}[F_1(y_1|\bullet), F_2(y_2|\bullet), \theta] \\ &= \partial^2 C[F_1(y_1|\bullet), F_2(y_2|\bullet), \theta]/\partial y_1 \partial y_2 \\ &= [\partial^2 C(., ., \theta)/\partial F_1 \partial F_2] f_1(y_1|\bullet) f_2(y_2|\bullet). \end{aligned} \tag{12-3}$$

A log-likelihood function can now be constructed using the logs of the right-hand sides of (12-3). Taking logs of (12-3) reveals the utility of the copula approach. The contribution of the joint observation to the log likelihood is

$$\ln f_{12}(y_1, y_2|\bullet) = \ln[\partial^2 C(., ., \theta)/\partial F_1 \partial F_2] + \ln f_1(y_1|\bullet) + \ln f_2(y_2|\bullet).$$

Some of the common copula functions that have been used in applications are as follows:

Product: $C[u_1, u_2, \theta] = u_1 u_2,$

FGM: $C[u_1, u_2, \theta] = u_1 u_2[1 + \theta(1 - u_1)(1 - u_2)],$

Gaussian: $C[u_1, u_2, \theta] = \Phi_2[\Phi^{-1}(u_1), \Phi^{-1}(u_2), \theta],$

Clayton: $C[u_1, u_2, \theta] = [u_1^{-\theta} + u_2^{-\theta} - 1]^{-1/\theta},$

Frank: $C[u_1, u_2, \theta] = \dfrac{1}{\theta} \ln\left[1 + \dfrac{\exp(\theta u_1 - 1)\exp(\theta u_2 - 1)}{\exp(\theta) - 1}\right],$

Plackett: $C[u_1, u_2, \theta] = \dfrac{1 + (\theta - 1)(u_1 + u_2) - \sqrt{[1 + (\theta - 1)(u_1 + u_2)]^2 - 4\theta(\theta - 1)(u_1 u_2)}}{2(\theta - 1)}.$

The product copula implies that the random variables are independent because it implies that the joint cdf is the product of the marginals. In the FGM (Fairlie, Gumbel, Morgenstern) copula, it can be seen that $\theta = 0$ implies the product copula, or independence. The same result can be shown for the Clayton copula. Independence in the Plackett copula follows if $\theta = 1$. In the Gaussian function, the copula is the bivariate normal cdf if the marginals happen to be normal to begin with. The essential point is that the marginals need not be normal to construct the copula function, so long as the marginal cdfs can be specified. (The dependence parameter is not the correlation between the variables. Trivedi and Zimmer provide transformations of θ that are closely related to correlations for each copula function listed.)

The essence of the copula technique is that the researcher can specify and analyze the marginals and the copula functions separately. The likelihood function is obtained by formulating the cdfs [or the densities because the differentiation in (12-3) will reduce the joint density to a convenient function of the marginal densities] and the copula.

Example 12.3 Joint Modeling of a Pair of Event Counts

The standard regression modeling approach for a random variable, y, that is a count of events is the Poisson regression model,

$$\text{Prob}[Y = y|\mathbf{x}] = \exp(-\lambda)\lambda^y/y!, \text{ where } \lambda = \exp(\mathbf{x}'\boldsymbol{\beta}), y = 0, 1, \ldots.$$

More intricate specifications use the negative binomial model (version 2, NB2),

$$\text{Prob}[Y = y|\mathbf{x}] = \frac{\Gamma(y + \alpha)}{\Gamma(\alpha)\Gamma(y + 1)}\left(\frac{\alpha}{\lambda + \alpha}\right)^{\alpha}\left(\frac{\lambda}{\lambda + \alpha}\right)^{y}, y = 0, 1, \ldots,$$

where α is an overdispersion parameter. (See Section 18.4.) A satisfactory, appropriate specification for bivariate outcomes has been an ongoing topic of research. Early suggestions were based on a latent mixture model,

$$y_1 = z + w_1,$$
$$y_2 = z + w_2,$$

where w_1 and w_2 have the Poisson or NB2 distributions specified earlier with conditional means λ_1 and λ_2 and z is taken to be an unobserved Poisson or NB variable. This formulation induces correlation between the variables but is unsatisfactory because that correlation must be positive. In a natural application, y_1 is doctor visits and y_2 is hospital visits. These could be negatively correlated. Munkin and Trivedi (1999) specified the jointness in the conditional mean functions, in the form of latent, common heterogeneity,

$$\lambda_j = \exp(\mathbf{x}_j'\boldsymbol{\beta}_j + \varepsilon),$$

where ε is common to the two functions. Cameron et al. (2004) used a bivariate copula approach to analyze Australian data on self-reported and actual physician visits (the latter maintained by the Health Insurance Commission). They made two adjustments to the preceding model we developed above. First, they adapted the basic copula formulation to these discrete random variables. Second, the variable of interest to them was not the actual or self-reported count but the difference. Both of these are straightforward modifications of the basic copula model.

Example 12.4 The Formula That Killed Wall Street[6]

David Li (2000) designed a bivariate normal (Gaussian) copula model for the pricing of collateralized debt obligations (CDOs) such as mortgage-backed securities. The methodology he proposed became a widely used tool in the mortgage-backed securities market. The model appeared to work well when markets were stable, but failed spectacularly in the turbulent period around the financial crisis of 2008–2009. Li has been (surely unfairly) deemed partly to blame for the financial crash of 2008.[7]

12.3 SEMIPARAMETRIC ESTIMATION

Semiparametric estimation is based on fewer assumptions than parametric estimation. In general, the distributional assumption is removed, and an estimator is devised from certain more general characteristics of the population. Intuition suggests two (correct) conclusions. First, the semiparametric estimator will be more robust than the parametric estimator—it will retain its properties, notably consistency across a greater range of specifications.

[6]Salmon (2000) and Li (1999, 2000).

[7]For example, Lee (2009), Hombrook (2009), Jones (2009), many others. From the CBC article: "… David Li is a Canadian math whiz who, some now say, developed the risk formula that destroyed Wall Street."

Consider our most familiar example. The least squares slope estimator is consistent whenever the data are well behaved and the disturbances and the regressors are uncorrelated. This is even true for the frontier function in Example 12.2, which has an asymmetric, nonnormal disturbance. But, second, this robustness comes at a cost. The distributional assumption usually makes the preferred estimator more efficient than a robust one. The best robust estimator in its class will usually be inferior to the parametric estimator when the assumption of the distribution is correct. Once again, in the frontier function setting, least squares may be robust for the slopes, and it is the most efficient estimator that uses only the orthogonality of the disturbances and the regressors, but it will be inferior to the maximum likelihood estimator when the two-part normal distribution is the correct assumption.

12.3.1 GMM ESTIMATION IN ECONOMETRICS

Recent applications in economics include many that base estimation on the **method of moments**. The generalized method of moments departs from a set of model-based moment equations, $E[\mathbf{m}(y_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0}$, where the set of equations specifies a relationship known to hold in the population. We used one of these in the preceding paragraph. The least squares estimator can be motivated by noting that the essential assumption is that $E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$. The estimator is obtained by seeking a parameter estimator $\mathbf{b}$ which mimics the population result, $(1/n)\Sigma_i[\mathbf{x}_i(y_i - \mathbf{x}_i'\mathbf{b})] = \mathbf{0}$. These are, of course, the normal equations for least squares. Note that the estimator is specified without benefit of any distributional assumption. Method of moments estimation is the subject of Chapter 13, so we will defer further analysis until then.

12.3.2 MAXIMUM EMPIRICAL LIKELIHOOD ESTIMATION

Empirical likelihood methods are suggested as a semiparametric alternative to maximum likelihood. As we shall see shortly, the estimator is closely related to the GMM estimator. Let π_i denote generically the probability that $y_i|\mathbf{x}_i$ takes the realized value in the sample. Intuition suggests (correctly) that with no further information, π_i will equal $1/n$. The **empirical likelihood function** is

$$EL = \prod_{i=1}^{n} \pi_i^{1/n}.$$

The **maximum empirical likelihood estimator** maximizes EL. Equivalently, we maximize the log of the empirical likelihood,

$$ELL = \frac{1}{n}\sum_{i=1}^{n} \ln \pi_i.$$

As a maximization problem, this program lacks sufficient structure to admit a solution—the solutions for π_i are unbounded. If we impose the restrictions that π_i are probabilities that sum to one, we can use a Lagrangean formulation to solve the optimization problem,

$$ELL = \left[\frac{1}{n}\sum_{i=1}^{n} \ln \pi_i\right] + \lambda\left[1 - \sum_{i=1}^{n} \pi_i\right].$$

This slightly restricts the problem since with $0 < \pi_i < 1$ and $\Sigma_i \pi_i = 1$, the solution suggested earlier becomes obvious. (There is nothing in the problem that differentiates the π_i's so they must all be equal to each other.) Inserting this result in the derivative with respect to any specific π_i produces the remaining result, $\lambda = 1$.

The maximization problem becomes meaningful when we impose a structure on the data. To develop an example, we'll recall Example 7.6, a nonlinear regression equation for *Income* for the German Socioeconomic Panel data, where we specified

$$E[Income|Age, Sex, Education] = \exp(\mathbf{x}'\boldsymbol{\beta}) = h(\mathbf{x}, \boldsymbol{\beta}).$$

For an example, assume that *Education* may be endogenous in this equation, but we have available a set of instruments, $\mathbf{z}$, say (*Age*, *Health*, *Sex*, *Market Condition*). We have assumed that there are more instruments (4) than included variables (3), so that the parameters will be overidentified (and the example will be complicated enough to be interesting). (See Sections 8.3.4 and 8.9.) The orthogonality conditions for nonlinear instrumental variable estimation are that the disturbances be uncorrelated with the instrumental variables, so

$$E\{\mathbf{z}_i[Income_i - h(\mathbf{x}_i, \boldsymbol{\beta})]\} = E[\mathbf{m}_i(\boldsymbol{\beta})] = \mathbf{0}.$$

The nonlinear least squares solution to this problem was developed in Section 8.9. A GMM estimator will minimize with respect to $\boldsymbol{\beta}$ the criterion function

$$q = \overline{\mathbf{m}}'(\boldsymbol{\beta})\mathbf{A}\overline{m}(\boldsymbol{\beta}),$$

where $\mathbf{A}$ is the chosen weighting matrix. Note that for our example, including the constant term, there are four elements in $\boldsymbol{\beta}$ and five moment equations, so the parameters are overidentified.

If, instead, we impose the restrictions implied by our moment equations on the empirical likelihood function, we obtain the population moment condition,

$$\left[\sum_{i=1}^{n} \pi_i \mathbf{z}_i \times (Income_i - h(\mathbf{x}_i, \boldsymbol{\beta}))\right] = \mathbf{0}.$$

(The probabilities are population quantities, so this is the expected value.) This produces the constrained empirical log likelihood,

$$ELL = \left[\frac{1}{n}\sum_{i=1}^{n} \ln \pi_i\right] + \lambda\left[1 - \sum_{i=1}^{n} \pi_i\right] + \boldsymbol{\gamma}'\left[\sum_{i=1}^{n} \pi_i \mathbf{z}_i\,(Income_i - h(\mathbf{x}_i, \boldsymbol{\beta}))\right].$$

The function is now maximized with respect to π_i, λ, $\boldsymbol{\beta}$ (K elements), and $\boldsymbol{\gamma}$ (L elements, the number of instrumental variables). At the solution, the values of π_i provide, essentially, a set of weights. Cameron and Trivedi (2005, p. 205) provide a solution for $\hat{\pi}_i$ in terms of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and show, once again, that $\lambda = 1$. The concentrated ELL function with these inserted provides a function of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ that remains to be maximized.

The empirical likelihood estimator has the same asymptotic properties as the GMM estimator. (This makes sense, given the resemblance of the estimation criteria—ultimately, both are focused on the moment equations.) There is evidence that, at least in some cases, the finite sample properties of the empirical likelihood estimator might be better than GMM. A survey appears in Imbens (2002). One suggested modification of the procedure is to replace the core function in $(1/n)\Sigma_i \ln \pi_i$ with the **entropy** measure,

$$Entropy = -(1/n)\Sigma_i \pi_i \ln \pi_i.$$

The **maximum entropy** estimator is developed in Golan, Judge, and Miller (1996) and Golan (2009).

12.3.3 LEAST ABSOLUTE DEVIATIONS ESTIMATION AND QUANTILE REGRESSION

Least squares can be severely distorted by outlying observations in a small sample. Recent applications in microeconomics and financial economics involving thick-tailed disturbance distributions, for example, are particularly likely to be affected by precisely these sorts of observations. (Of course, in those applications in finance involving hundreds of thousands of observations, which are becoming commonplace, this discussion is moot.) These applications have led to the proposal of robust estimators that are unaffected by outlying observations. One of these, the least absolute deviations, or LAD estimator discussed in Section 7.3.1, is also useful in its own right as an estimator of the conditional median function in the modified model

$$\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}_{.50}.$$

That is, rather than providing a robust alternative to least squares as an estimator of the slopes of $E[y|\mathbf{x}]$, LAD is an estimator of a different feature of the population. This is essentially a semiparametric specification in that it specifies only a particular feature of the distribution, its median, but not the distribution itself. It also specifies that the conditional median be a linear function of $\mathbf{x}$.

The median, in turn, is only one possible quantile of interest. If the model is extended to other quantiles of the conditional distribution, we obtain

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that } \text{Prob}[y \leq \mathbf{x}'\boldsymbol{\beta}_q|\mathbf{x}] = q, 0 < q < 1.$$

This is essentially a semiparametric specification. No assumption is made about the distribution of $y|\mathbf{x}$ or about its conditional variance. The fact that q can vary continuously (strictly) between zero and one means that there is an infinite number of possible parameter vectors. It seems reasonable to view the coefficients, which we might write $\boldsymbol{\beta}(q)$ less as fixed parameters, as we do in the linear regression model, than loosely as features of the distribution of $y|\mathbf{x}$. For example, it is not likely to be meaningful to view $\boldsymbol{\beta}(.49)$ to be discretely different from $\boldsymbol{\beta}(.50)$ or to compute precisely a particular difference such as $\boldsymbol{\beta}(.5) - \boldsymbol{\beta}(.3)$. On the other hand, the qualitative difference, or possibly the lack of a difference, between $\boldsymbol{\beta}(.3)$ and $\boldsymbol{\beta}(.5)$ may well be an interesting characteristic of the population. The quantile regression model is examined in Section 7.3.2.

12.3.4 KERNEL DENSITY METHODS

The kernel density estimator is an inherently nonparametric tool, so it fits more appropriately into the next section. But some models that use kernel methods are not completely nonparametric. The partially linear model in Section 7.4 is a case in point. Many models retain an index function formulation, that is, build the specification around a linear function $\mathbf{x}'\boldsymbol{\beta}$, which makes them at least semiparametric, but nonetheless still avoid distributional assumptions by using kernel methods. Lewbel's (2000) estimator for the binary choice model is another example.

Example 12.5 Semiparametric Estimator for Binary Choice Models

The core binary choice model analyzed in Section 17.3, the probit model, is a fully parametric specification. Under the assumptions of the model, maximum likelihood is the efficient (and appropriate) estimator. However, as documented in a voluminous literature, the estimator of $\boldsymbol{\beta}$ is fragile with respect to failures of the distributional assumption. We will examine

a few semiparametric and nonparametric estimators in Section 17.4.7. To illustrate the nature of the modeling process, we consider an estimator suggested by Lewbel (2000). The probit model is based on the normal distribution, with $\text{Prob}[y_i = 1|\mathbf{x}_i] = \text{Prob}[\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0]$ where $\varepsilon_i \sim \mathbf{N}[0, 1]$. The estimator of $\boldsymbol{\beta}$ under this specification may be inconsistent if the distribution is not normal or if ε_i is heteroscedastic. Lewbel suggests the following: If (a) it can be assumed that $\mathbf{x}_i$ contains a "special" variable v_i whose coefficient has a known sign, a method is developed for determining the sign, and (b) the density of ε_i is independent of this variable, then a consistent estimator of $\boldsymbol{\beta}$ can be obtained by *regression* of $[y_i - s(v_i)]/f(v_i|\mathbf{x}_i)$ on $\mathbf{x}_i$ where $s(v_i) = 1$ if $v_i > 0$ and 0 otherwise and $f(v_i|\mathbf{x}_i)$ is a kernel density estimator of the density of $v_i|\mathbf{x}_i$. Lewbel's estimator is robust to heteroscedasticity and distribution. A method is also suggested for estimating the distribution of ε_i. Note that Lewbel's estimator is semiparametric. His underlying model is a function of the parameters $\boldsymbol{\beta}$ but the distribution is unspecified.

12.3.5 COMPARING PARAMETRIC AND SEMIPARAMETRIC ANALYSES

It is often of interest to compare the outcomes of parametric and semiparametric models. As we have noted earlier, the strong assumptions of the fully parametric model come at a cost; the inferences from the model are only as robust as the underlying assumptions. Of course, the other side of that argument is that when the assumptions are met, parametric models represent efficient strategies for analyzing the data. The alternative, semiparametric approaches, relax assumptions such as normality and homoscedasticity. It is important to note that the model extensions to which semiparametric estimators are typically robust render the more heavily parameterized estimators inconsistent. The comparison is not just one of efficiency. As a consequence, comparison of parameter estimates can be misleading—the parametric and semiparametric estimators are often estimating very different quantities.

Example 12.6 A Model of Vacation Expenditures

Melenberg and van Soest (1996) analyzed the 1981 vacation expenditures of a sample of 1,143 Dutch families. The important feature of the data that complicated the analysis was that 37% (423) of the families reported zero expenditures. A linear regression that ignores this feature of the data would be heavily skewed toward underestimating the response of expenditures to the covariates such as total family expenditures (budget), family size, age, or education. (See Section 19.3.) The standard parametric approach to analyzing data of this sort is the Tobit, or censored, regression model,

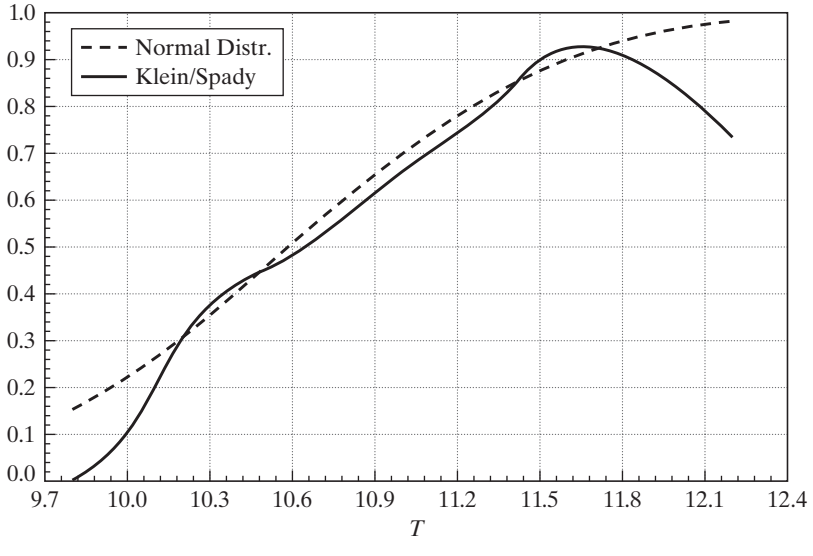
$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i \sim \text{N}[0, \sigma^2]$$
$$y_i = \max(0, y_i^*),$$

or a two-part model that models the participation (zero or positive expenditure) and intensity (expenditure given positive expenditure) as separate decisions. (Maximum likelihood estimation of this model is examined in detail in Section 19.3.) The model rests on two strong assumptions, normality and homoscedasticity. Both assumptions can be relaxed in a more elaborate parametric framework, but the authors found that test statistics persistently rejected one or both of the assumptions even with the extended specifications. An alternative approach that is robust to both is Powell's (1984, 1986a,b) censored least absolute deviations estimator, which is a more technically demanding computation based on the LAD estimator in Section 7.3.1. Not surprisingly, the parameter estimates produced by the two approaches vary widely. The authors computed a variety of estimators of $\boldsymbol{\beta}$. A useful exercise that they

did not undertake would be to compare the partial effects from the different models. This is a benchmark on which the differences between the different estimators can sometimes be reconciled. In the Tobit model, $\partial E[y_i|\mathbf{x}_i]/\partial \mathbf{x}_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)\boldsymbol{\beta}$ (see Section 19.3). It is unclear how to compute the counterpart in the semiparametric model, since the underlying specification holds only that $\text{Med}[\varepsilon_i|\mathbf{x}_i] = 0$. (The authors report on the *Journal of Applied Econometrics* data archive site that these data are proprietary. As such, we were unable to extend the analysis to obtain estimates of partial effects.) This highlights a significant difficulty with the semiparametric approach to estimation. In a nonlinear model such as this one, it is often the partial effects that are of interest, not the coefficients. But one of the byproducts of the more robust specification is that the partial effects are not defined.

In a second stage of the analysis, the authors decomposed their expenditure equation into a participation equation that modeled probabilities for the binary outcome "expenditure $= 0$ or > 0" and a conditional expenditure equation for those with positive expenditure.[8] For this step, the authors once again used a parametric model based on the normal distribution (the probit model—see Section 17.3) and a semiparametric model that is robust to distribution and heteroscedasticity developed by Klein and Spady (1993). As before, the coefficient estimates differ substantially. However, in this instance, the specification tests are considerably more sympathetic to the parametric model. Figure 12.1, which reproduces their Figure 2, compares the predicted probabilities from the two models. The dashed curve is the probit model. Within the range of most of the data, the models give quite similar predictions. Once again, however, it is not possible to compare partial effects. The interesting outcome from this part of the analysis seems to be that the failure of the parametric specification resides more in the modeling of the continuous expenditure variable than with the model that separates the two subsamples based on zero or positive expenditures.

FIGURE 12.1 Predicted Probabilities of Positive Expenditure.



[8]In Section 18.4.8, we will label this a "hurdle" model. See Mullahy (1986).

12.4 NONPARAMETRIC ESTIMATION

Researchers have long held reservations about the strong assumptions made in parametric models fit by maximum likelihood. The linear regression model with normal disturbances is a leading example. Splines, translog models, and polynomials all represent attempts to generalize the functional form. Nonetheless, questions remain about how much generality can be obtained with such approximations. The techniques of nonparametric estimation discard essentially all fixed assumptions about functional form and distribution. Given their very limited structure, it follows that nonparametric specifications rarely provide very precise inferences. The benefit is that what information is provided is extremely robust. The centerpiece of this set of techniques is the kernel density estimator that we have used in the preceding examples. We will examine some examples, then examine an application to a bivariate regression.[9]

12.4.1 KERNEL DENSITY ESTIMATION

Sample statistics such as mean, variance, and range give summary information about the values that a random variable may take. But they do not suffice to show the distribution of values that the random variable takes, and these may be of interest as well. The density of the variable is used for this purpose. A fully parametric approach to density estimation begins with an assumption about the form of a distribution. Estimation of the density is accomplished by estimation of the parameters of the distribution. To take the canonical example, if we decide that a variable is generated by a normal distribution with mean μ and variance σ^2, then the density is fully characterized by these parameters. It follows that

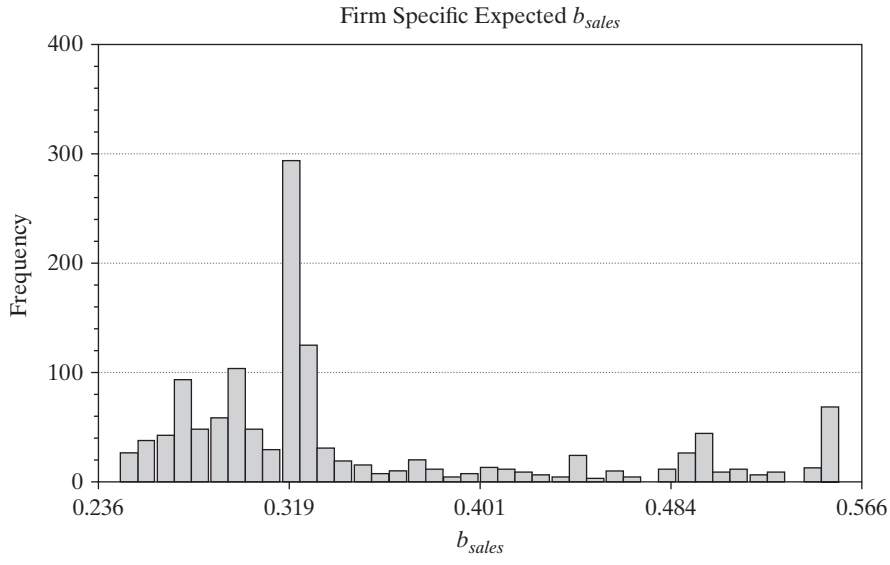
$$\hat{f}(x) = f(x|\hat{\mu}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\mu}}{\hat{\sigma}} \right)^2 \right].$$

One may be unwilling to make a narrow distributional assumption about the density. The usual approach in this case is to begin with a **histogram** as a descriptive device. Consider an example. In Example 15.17 and in Greene (2004c), we estimate a model that produces a conditional estimator of a slope vector for each of the 1,270 firms in the sample. We might be interested in the distribution of these estimators across firms. In particular, the conditional estimates of the estimated slope on ln *sales* for the 1,270 firms have a sample mean of 0.3428, a standard deviation of 0.08919, a minimum of 0.2361, and a maximum of 0.5664. This tells us little about the distribution of values, though the fact that the mean is well below the midrange of 0.4013 might suggest some skewness. The histogram in Figure 12.2 is much more revealing. Based on what we see thus far, an assumption of normality might not be appropriate. The distribution seems to be bimodal, but certainly no particular functional form seems natural.

The histogram is a crude density estimator. The rectangles in the figure are called bins. By construction, they are of equal width. (The parameters of the histogram are the number of bins, the bin width, and the leftmost starting point. Each is important in the shape of the end result.) Because the frequency count in the bins sums to the sample size, by dividing each by n, we have a density estimator that satisfies an obvious

[9]The set of literature in this area of econometrics is large and rapidly growing. Major references which provide an applied and theoretical foundation are Härdle (1990), Pagan and Ullah (1999), and Li and Racine (2007).

FIGURE 12.2 Histogram for Estimated b_{sales} Coefficients.



requirement for a density; it sums (integrates) to one. We can formalize this by laying out the method by which the frequencies are obtained. Let x_k be the midpoint of the kth bin and let h be the width of the bin—we will shortly rename h to be the bandwidth for the density estimator. The distances to the left and right boundaries of the bins are $h/2$. The frequency count in each bin is the number of observations in the sample which fall in the range $x_k \pm h/2$. Collecting terms, we have our estimator

$$\hat{f}(x) = \frac{1}{n}\frac{\text{frequency in bin}_x}{\text{width of bin}_x} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathbf{1}\left(x - \frac{h}{2} < x_i < x + \frac{h}{2}\right),$$

where **1** (*statement*) denotes an indicator function that equals 1 if the statement is true and 0 if it is false and bin_x denotes the bin which has x as its midpoint. We see, then, that the histogram is an estimator, at least in some respects, like other estimators we have encountered. The event in the indicator can be rearranged to produce an equivalent form,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathbf{1}\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right).$$

This form of the estimator simply counts the number of points that are within one half-bin width of x_k.

Albeit rather crude, this "naïve" (its formal name in the literature) estimator is in the form of **kernel density estimators** that we have met at various points,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left[\frac{x_i - x}{h}\right], \quad \text{where } K[z] = \mathbf{1}[-1/2 < z < 1/2].$$

The naïve estimator has several shortcomings. It is neither smooth nor continuous. Its shape is partly determined by where the leftmost and rightmost terminals of the

histogram are set. (In constructing a histogram, one often chooses the bin width to be a specified fraction of the sample range. If so, then the terminals of the lowest and highest bins will equal the minimum and maximum values in the sample, and this will partly determine the shape of the histogram. If, instead, the bin width is set irrespective of the sample values, then this problem is resolved.) More importantly, the shape of the histogram will be crucially dependent on the bandwidth itself. (Unfortunately, this problem remains even with more sophisticated specifications.)

The crudeness of the weighting function in the estimator is easy to remedy. Rosenblatt's (1956) suggestion was to substitute for the naïve estimator some other weighting function which is continuous and which also integrates to one. A number of candidates have been suggested, including the (long) list in Table 12.1. Each of these is smooth, continuous, symmetric, and equally attractive. The logit and normal kernels are defined so that the weight only asymptotically falls to zero whereas the others fall to zero at specific points. It has been observed that in constructing a density estimator, the choice of kernel function is rarely crucial, and is usually minor in importance compared to the more difficult problem of choosing the bandwidth. (The logit, normal and Epanechnikov kernels appear to be the default choices in many applications.)

The kernel density function is an estimator. For any specific x, $\hat{f}(x)$ is a sample statistic,

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} g(x_i|\, z, h).$$

Because $g(x_i|\, z, h)$ is nonlinear, we should expect a bias in a finite sample. It is tempting to apply our usual results for sample moments, but the analysis is more complicated because the bandwidth is a function of n. Pagan and Ullah (1999) have examined the properties of kernel estimators in detail and found that under certain assumptions, the estimator is consistent and asymptotically normally distributed but biased in finite samples.[10] The bias is a function of the bandwidth, but for an appropriate choice of h, the bias does vanish asymptotically. As intuition might suggest, the larger the bandwidth, the greater the bias, but at the same time, the smaller the variance. This might suggest a search for an optimal bandwidth. After a lengthy analysis of the subject, however, the authors' conclusion provides little guidance for finding one. One consideration does seem useful. For the proportion of observations captured in the bin to converge to the

TABLE 12.1 Kernel Functions for Density Estimation

Kernel	***Formula K[z]***
Epanechnikov	$0.75(1 - 0.2z^2)/\sqrt{5}$ if $\lvert z\rvert \leq \sqrt{5}$, 0 else
Normal	$\phi(z)$ (normal density)
Logit	$\Lambda(z)[1 - \Lambda(z)]$ (logistic density)
Uniform	0.5 if $\lvert z\rvert \leq 1$, 0 else
Beta	$0.75(1 - z)(1 + z)$ if $\lvert z\rvert \leq 1$, 0 else
Cosine	$1 + \cos(2\pi z)$ if $\lvert z\rvert \leq 0.5$, 0 else
Triangle	$1 - \lvert z\rvert$, if $\lvert z\rvert \leq 1$, 0 else
Parzen	$4/3 - 8z^2 + 8\lvert z\rvert^3$ if $\lvert z\rvert \leq 0.5$, $8(1 - \lvert z\rvert)^3/3$ if $0.5 < \lvert z\rvert \leq 1$, 0 else

[10]See also Li and Racine (2007) and Henderson and Parmeter (2015).

corresponding area under the density, the width itself must shrink more slowly than $1/n$. Common applications typically use a bandwidth equal to some multiple of $n^{-1/5}$ for this reason. Thus, the one we used earlier is Silverman's (1986) bandwidth, $h = 0.9 \times s/n^{1/5}$. To conclude the illustration begun earlier, Figure 12.3 is a logit-based kernel density estimator for the distribution of slope estimates for the model estimated earlier. The resemblance to the histogram in Figure 12.2 is to be expected.

12.5 PROPERTIES OF ESTIMATORS

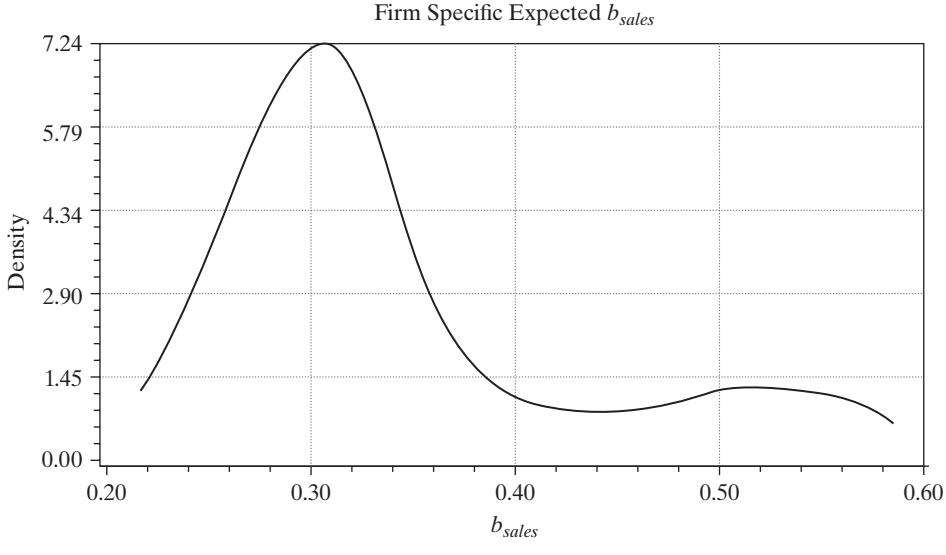
The preceding has been concerned with methods of estimation. We have surveyed a variety of techniques that have appeared in the applied literature. We have not yet examined the statistical properties of these estimators. Although, as noted earlier, we will leave extensive analysis of the asymptotic theory for more advanced treatments, it is appropriate to spend at least some time on the fundamental theoretical platform that underlies these techniques.

12.5.1 STATISTICAL PROPERTIES OF ESTIMATORS

Properties that we have considered are as follows:

- **Unbiasedness:** This is a finite sample property that can be established in only a very small number of cases. Strict unbiasedness is rarely of central importance outside the linear regression model. However, asymptotic unbiasedness (whereby the expectation of an estimator converges to the true parameter as the sample size grows), might be of interest.[11] In most cases, however, discussions of asymptotic

FIGURE 12.3 Kernel Density for b_{sales}.



[11]See, for example, Pagan and Ullah (1999, Section 2.5.1) and Henderson and Parmeter (2015, Section 2.2) on the subject of the kernel density estimator.

unbiasedness are actually directed toward consistency, which is a more desirable property.

- **Consistency:** This is a much more important property. Econometricians are rarely willing to place much credence in an estimator for which consistency cannot be established. In some instances, the inconsistency can be more precisely quantified. For example, the "incidental parameters problem" (see Section 17.7.3) relates to estimation of fixed effects models in panel data settings in which an estimator is inconsistent for fixed T but is consistent in T (and tolerably biased for moderate sized T).
- **Asymptotic normality:** This property forms the platform for most of the statistical inference that is done with common estimators. When asymptotic normality cannot be established, it sometimes becomes difficult to find a method of progressing beyond simple presentation of the numerical values of estimates (with caveats). However, most of the contemporary literature in macroeconomics and time-series analysis is strongly focused on estimators that are decidedly not asymptotically normally distributed. The implication is that this property takes its importance only in context, not as an absolute virtue.
- **Asymptotic efficiency:** Efficiency can rarely be established in absolute terms. Efficiency within a class often can, however. Thus, for example, a great deal can be said about the relative efficiency of maximum likelihood and GMM estimators in the class of consistent and asymptotically normally distributed (CAN) estimators. There are two important practical considerations in this setting. First, the researcher will want to know that he or she has not made demonstrably suboptimal use of the data. (The literature contains discussions of GMM estimation of fully specified parametric probit models—GMM estimation in this context is unambiguously inferior to maximum likelihood.) Thus, when possible, one would want to avoid obviously inefficient estimators. On the other hand, it will usually be the case that the researcher is not choosing from a list of available estimators; he or she has one at hand, and questions of relative efficiency are moot.

12.5.2 EXTREMUM ESTIMATORS

An **extremum estimator** is one that is obtained as the optimizer of a **criterion function** $q(\boldsymbol{\theta}|\mathbf{data})$. Three that have occupied much of our effort thus far are:

- Least squares: $\hat{\boldsymbol{\theta}}_{LS} = \text{Argmax}[-(1/n)\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i, \boldsymbol{\theta}_{LS}))^2]$,
- Maximum likelihood: $\hat{\boldsymbol{\theta}}_{ML} = \text{Argmax}[(1/n)\sum_{i=1}^{n}\ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}_{ML})]$, and
- GMM: $\hat{\boldsymbol{\theta}}_{GMM} = \text{Argmax}[-\overline{\mathbf{m}}(\mathbf{data}, \boldsymbol{\theta}_{GMM})'\mathbf{W}\overline{\mathbf{m}}(\mathbf{data}, \boldsymbol{\theta}_{GMM})]$.

(We have changed the signs of the first and third only for convenience so that all three may be cast as the same type of optimization problem.) The least squares and maximum likelihood estimators are examples of **M estimators**, which are defined by optimizing over a sum of terms. Most of the familiar theoretical results developed here and in other treatises concern the behavior of extremum estimators. Several of the estimators considered in this chapter are extremum estimators, but a few—including the Bayesian estimators, some of the semiparametric estimators, and all of the nonparametric estimators—are not. Nonetheless, we are interested in establishing the properties of estimators in all these cases, whenever possible. The end result for the practitioner

will be the set of statistical properties that will allow him or her to draw with confidence conclusions about the data-generating process(es) that have motivated the analysis in the first place.

Derivations of the behavior of extremum estimators are pursued at various levels in the literature. (See, for example, any of the sources mentioned in Footnote 1 of this chapter.) Amemiya (1985) and Davidson and MacKinnon (2004) are very accessible treatments. Newey and McFadden (1994) is a rigorous analysis that provides a current, standard source. Our discussion at this point will only suggest the elements of the analysis. The reader is referred to one of these sources for detailed proofs and derivations.

12.5.3 ASSUMPTIONS FOR ASYMPTOTIC PROPERTIES OF EXTREMUM ESTIMATORS

Some broad results are needed in order to establish the asymptotic properties of the classical (not Bayesian) conventional extremum estimators noted above.

1. **The parameter space** (see Section 12.2) must be convex and the parameter vector that is the object of estimation must be a point in its interior. The first requirement rules out ill-defined estimation problems such as estimating a parameter which can only take one of a finite discrete set of values. Thus, searching for the date of a structural break in a time-series model as if it were a conventional parameter leads to a nonconvexity. Some proofs in this context are simplified by assuming that the parameter space is compact. (A compact set is closed and bounded.) However, assuming compactness is usually restrictive, so we will opt for the weaker requirement.
2. **The criterion function** must be concave in the parameters. (See Section A.8.2.) This assumption implies that with a given data set, the objective function has an interior optimum and that we can locate it. Criterion functions need not be globally concave; they may have multiple optima. But, if they are not at least locally concave, then we cannot speak meaningfully about optimization. One would normally only encounter this problem in a badly structured model, but it is possible to formulate a model in which the estimation criterion is monotonically increasing or decreasing in a parameter. Such a model would produce a nonconcave criterion function.[12] The distinction between compactness and concavity in the preceding condition is relevant at this point. If the criterion function is strictly continuous in a compact parameter space, then it has a maximum in that set and assuming concavity is not necessary. The problem for estimation, however, is that this does not rule out having that maximum occur on the (assumed) boundary of the parameter space. This case interferes with proofs of consistency and asymptotic normality. The overall problem is solved by assuming that the criterion function is concave in the neighborhood of the true parameter vector.
3. **Identifiability of the parameters.** Any statement that begins with "the true parameters of the model, $\boldsymbol{\theta}_0$ are identified if …" is problematic because if the parameters are "not identified," then arguably, they are not *the* parameters of the (any) model. (For example, there is no "true" parameter vector in the unidentified

[12]In their Exercise 23.6, Griffiths, Hill, and Judge (1993), based (alas) on the first edition of this text, suggest a probit model for statewide voting outcomes that includes dummy variables for region: Northeast, Southeast, West, and Mountain. One would normally include three of the four dummy variables in the model, but Griffiths et al. carefully dropped two of them because, in addition to the dummy variable trap, the Southeast variable is always zero when the dependent variable is zero. Inclusion of this variable produces a nonconcave likelihood function—the parameter on this variable diverges. Analysis of a closely related case appears as a caveat in Amemiya (1985, p. 272).

model of Example 2.5.) A useful way to approach this question that avoids the ambiguity of trying to define *the* true parameter vector first and then asking if it is identified (estimable) is as follows, where we borrow from Davidson and MacKinnon (1993, p. 591): Consider the parameterized model, M, and the set of allowable data generating processes for the model, μ. Under a particular parameterization μ, let there be an assumed "true" parameter vector, $\boldsymbol{\theta}(\mu)$. Consider any parameter vector $\boldsymbol{\theta}$ in the parameter space, Θ. Define

$$q_\mu(\mu, \boldsymbol{\theta}) = \text{plim}_\mu q_n(\boldsymbol{\theta} \,|\, \textbf{data}).$$

This function is the probability limit of the objective function under the assumed parameterization μ. If this probability limit exists (is a finite constant) and moreover, if

$$q_\mu[\mu, \boldsymbol{\theta}(\mu)] > q_\mu(\mu, \boldsymbol{\theta}) \quad \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}(\mu),$$

then, if the parameter space is compact, the parameter vector is identified by the criterion function. We have not assumed compactness. For a convex parameter space, we would require the additional condition that there exist no sequences without limit points $\boldsymbol{\theta}^m$ such that $q(\mu, \boldsymbol{\theta}^m)$ converges to $q[\mu, \boldsymbol{\theta}(\mu)]$.

The approach taken here is to assume first that the model has *some* set of parameters. The identifiability criterion states that assuming this is the case, the probability limit of the criterion is maximized at these parameters. This result rests on convergence of the criterion function to a finite value at any point in the interior of the parameter space. Because the criterion function is a function of the data, this convergence requires a statement of the properties of the data, for example, well behaved in some sense. Leaving that aside for the moment, interestingly, the results to this point already establish the consistency of the M estimator. In what might seem to be an extremely terse fashion, Amemiya (1985) defined identifiability simply as "existence of a consistent estimator." We see that identification and the conditions for consistency of the M estimator are substantively the same.

This form of identification is necessary, in theory, to establish the consistency arguments. In any but the simplest cases, however, it will be extremely difficult to verify in practice. Fortunately, there are simpler ways to secure identification that will appeal more to the intuition:

- For the least squares estimator, a sufficient condition for identification is that any two different parameter vectors, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, must be able to produce different values of the conditional mean function. This means that for any two different parameter vectors, there must be an $\mathbf{x}_i$ that produces different values of the conditional mean function. You should verify that for the linear model, this is the full rank assumption A.2. For the model in Example 2.5, we have a regression in which $x_2 = x_3 + x_4$. In this case, any parameter vector of the form $(\beta_1, \beta_2 - a, \beta_3 + a, \beta_4 + a)$ produces the same conditional mean as $(\beta_1, \beta_2, \beta_3, \beta_4)$ regardless of $\mathbf{x}_i$, so this model is not identified. The full rank assumption is needed to preclude this problem. For nonlinear regressions, the problem is much more complicated, and there is no simple generality. Example 7.2 shows a nonlinear regression model that is not identified and how the lack of identification is remedied.
- For the maximum likelihood estimator, a condition similar to that for the regression model is needed. For any two parameter vectors, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, it must be possible to

produce different values of the density $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ for some data vector $(y_i, \mathbf{x}_i)$. Many econometric models that are fit by maximum likelihood are "index function" models that involve densities of the form $f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = f(y_i|\mathbf{x}_i'\boldsymbol{\theta})$. When this is the case, the same full rank assumption that applies to the regression model may be sufficient. (If there are no other parameters in the model, then it will be sufficient.)

- For the GMM estimator, not much simplicity can be gained. A sufficient condition for identification is that $\mathrm{E}[\overline{\mathbf{m}}(\mathbf{data}, \boldsymbol{\theta})] \neq \mathbf{0}$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

4. **Behavior of the data** has been discussed at various points in the preceding text. The estimators are based on means of functions of observations. (You can see this in all three of the preceding definitions. Derivatives of these criterion functions will likewise be means of functions of observations.) Analysis of their large sample behaviors will turn on determining conditions under which certain sample means of functions of observations will be subject to laws of large numbers such as the Khinchine (D.5) or Chebychev (D.6) theorems, and what must be assumed in order to assert that "root-n" times sample means of functions will obey central limit theorems such as the Lindeberg–Feller (D.19) or Lyapounov (D.20) theorems for cross sections or the Martingale Difference Central Limit theorem for dependent observations (Theorem 20.3). Ultimately, this is the issue in establishing the statistical properties. The convergence property claimed above must occur in the context of the data. These conditions have been discussed in Sections 4.4.1 and 4.4.2 under the heading of "well-behaved data." At this point, we will assume that the data are well behaved.

12.5.4 ASYMPTOTIC PROPERTIES OF ESTIMATORS

With all this apparatus in place, the following are the standard results on asymptotic properties of M estimators:

THEOREM 12.1 Consistency of M Estimators
If (a) the parameter space is convex and the true parameter vector is a point in its interior, (b) the criterion function is concave, (c) the parameters are identified by the criterion function, and (d) the data are well behaved, then the M estimator converges in probability to the true parameter vector.

Proofs of consistency of M estimators rely on a fundamental convergence result that, itself, rests on assumptions (a) through (d) in Theorem 12.1. We have assumed identification. The fundamental device is the following: Because of its dependence on the data, $q(\boldsymbol{\theta}|\mathbf{data})$ is a random variable. We assumed in (c) that plim $q(\boldsymbol{\theta}|\mathbf{data}) = q_0(\boldsymbol{\theta})$ for any point in the parameter space. Assumption (c) states that the maximum of $q_0(\boldsymbol{\theta})$ occurs at $q_0(\boldsymbol{\theta}_0)$, so $\boldsymbol{\theta}_0$ is the maximizer of the probability limit. By its definition, the estimator, $\hat{\boldsymbol{\theta}}$, is the maximizer of $q(\boldsymbol{\theta}|\mathbf{data})$. Therefore, consistency requires the limit of the maximizer, $\hat{\boldsymbol{\theta}}$, be equal to the maximizer of the limit, $\boldsymbol{\theta}_0$. Our identification condition establishes this. We will use this approach in somewhat greater detail in Section 14.4.5.a where we establish consistency of the maximum likelihood estimator.

THEOREM 12.2 Asymptotic Normality of M Estimators
If:

(i) $\hat{\boldsymbol{\theta}}$ *is a consistent estimator of* $\boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ *is a point in the interior of the parameter space*;

(ii) $q(\boldsymbol{\theta}|\mathbf{data})$ *is concave and twice continuously differentiable in* $\boldsymbol{\theta}$ *in a neighborhood of* $\boldsymbol{\theta}_0$;

(iii) $\sqrt{n}[\partial q(\boldsymbol{\theta}_0|\mathbf{data})/\partial\boldsymbol{\theta}_0] \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}]$;

(iv) *for any* $\boldsymbol{\theta}$ *in* $\boldsymbol{\Theta}$, $\lim_{n\to\infty}\Pr[|(\partial^2 q(\boldsymbol{\theta}|\mathbf{data})/\partial\theta_k\partial\theta_m) - h_{km}(\theta)| > \varepsilon] = 0\ \forall\ \varepsilon > 0$ *where* $h_{km}(\boldsymbol{\theta})$ *is a continuous finite valued function of* $\boldsymbol{\theta}$;

(v) *the matrix of elements* $\mathbf{H}(\boldsymbol{\theta})$ *is nonsingular at* $\boldsymbol{\theta}_0$, *then* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, [\mathbf{H}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Phi}\mathbf{H}^{-1}(\boldsymbol{\theta}_0)]\}$.

The proof of asymptotic normality is based on the mean value theorem from calculus and a Taylor series expansion of the derivatives of the maximized criterion function around the true parameter vector,

$$\sqrt{n}\frac{\partial q(\hat{\boldsymbol{\theta}}|\mathbf{data})}{\partial\hat{\boldsymbol{\theta}}} = \mathbf{0} = \sqrt{n}\frac{\partial q(\boldsymbol{\theta}_0|\mathbf{data})}{\partial\boldsymbol{\theta}_0} + \frac{\partial^2 q(\overline{\boldsymbol{\theta}}|\mathbf{data})}{\partial\overline{\boldsymbol{\theta}}\partial\overline{\boldsymbol{\theta}}'}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Each derivative is evaluated at a point $\overline{\boldsymbol{\theta}}$ that is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, that is, $\overline{\boldsymbol{\theta}} = w\hat{\boldsymbol{\theta}} + (1 - w)\boldsymbol{\theta}_0$ for some $0 < w < 1$. Because we have assumed plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, we see that the matrix in the second term on the right must be converging to $\mathbf{H}(\boldsymbol{\theta}_0)$. The assumptions in the theorem can be combined to produce the claimed normal distribution. Formal proof of this set of results appears in Newey and McFadden (1994). A somewhat more detailed analysis based on this theorem appears in Section 14.4.5.b, where we establish the asymptotic normality of the maximum likelihood estimator.

The preceding was restricted to M estimators, so it remains to establish counterparts for the important GMM estimator. Consistency follows along the same lines used earlier, but asymptotic normality is a bit more difficult to establish. We will return to this issue in Chapter 13, where, once again, we will sketch the formal results and refer the reader to a source such as Newey and McFadden (1994) for rigorous derivation.

The preceding results are not straightforward in all estimation problems. For example, the least absolute deviations (LAD) is not among the estimators noted earlier, but it is an M estimator and it shares the results given here. The analysis is complicated because the criterion function is not continuously differentiable. Nonetheless, consistency and asymptotic normality have been established.[13] Some of the semiparametric and all of the nonparametric estimators noted require somewhat more intricate treatments. For example, Pagan and Ullah (1999, Sections 2.5–2.6) and Li and Racine (2007, Sections 1.9–1.12) are able to establish the familiar desirable properties for the kernel density estimator $\hat{f}(x^*)$, but it requires a somewhat more involved analysis of the function and the data than is necessary, say, for the linear regression or binomial logit model. The interested reader can find many lengthy and detailed analyses of asymptotic properties of estimators in, for example, Amemiya (1985), Newey and McFadden (1994), Davidson

[13]See Koenker and Bassett (1982) and Amemiya (1985, pp. 152–154).

and MacKinnon (2004), and Hayashi (2000). In practical terms, it is rarely possible to verify the conditions for an estimation problem at hand, and they are usually simply assumed. However, finding violations of the conditions is sometimes more straightforward, and this is worth pursuing. For example, lack of parametric identification can often be detected by analyzing the model itself.

12.5.5 TESTING HYPOTHESES

The preceding describes a set of results that (more or less) unifies the theoretical underpinnings of three of the major classes of estimators in econometrics, least squares, maximum likelihood, and GMM. A similar body of theory has been produced for the familiar test statistics, Wald, Likelihood Ratio (LR), and Lagrange multiplier (LM).[14] All of these have been laid out in practical terms elsewhere in this text, so in the interest of brevity, we will refer the interested reader to the background sources listed for the technical details.

12.6 SUMMARY AND CONCLUSIONS

This chapter has presented a short overview of estimation in econometrics. There are various ways to approach such a survey. The current literature can be broadly grouped by three major types of estimators—parametric, semiparametric, and nonparametric. It has been suggested that the overall drift in the literature is from the first toward the third of these, but on a closer look, we see that this is probably not the case. Maximum likelihood is still the estimator of choice in many settings. New applications have been found for the GMM estimator, but at the same time, new Bayesian and simulation estimators, all fully parametric, are emerging at a rapid pace. Certainly, the range of tools that can be applied in any setting is growing steadily.

Key Terms and Concepts

- Bayesian estimation
- Conditional density
- Copula function
- Criterion function
- Data-generating mechanism
- Density
- Empirical likelihood function
- Entropy
- Estimation criterion
- Extremum estimator
- Fundamental probability transform
- Generalized method of moments
- Histogram
- Kernel density estimator
- M estimator
- Maximum empirical likelihood estimator
- Maximum entropy
- Maximum likelihood estimator
- Method of moments
- Nonparametric estimators
- Semiparametric estimation
- Simulation-based estimation
- Sklar's theorem
- Stochastic frontier model

Exercise and Question

1. Compare the fully parametric and semiparametric approaches to estimation of a discrete choice model such as the multinomial logit model discussed in Chapter 17. What are the benefits and costs of the semiparametric approach?

[14] See Newey and McFadden (1994).

13

MINIMUM DISTANCE ESTIMATION AND THE GENERALIZED METHOD OF MOMENTS

13.1 INTRODUCTION

The **maximum likelihood estimator** presented in Chapter 14 is fully efficient among consistent and asymptotically normally distributed estimators in the context of the specified parametric model. The possible shortcoming in this result is that to attain that efficiency, it is necessary to make possibly strong, restrictive assumptions about the distribution, or data-generating process. The generalized method of moments (GMM) estimators discussed in this chapter move away from parametric assumptions, toward estimators that are robust to some variations in the underlying data-generating process.

This chapter will present a number of fairly general results on parameter estimation. We begin with perhaps the oldest formalized theory of estimation, the classical theory of the method of moments. This body of results dates to the pioneering work of Fisher (1925). The use of sample moments as the building blocks of estimating equations is fundamental in econometrics. GMM is an extension of this technique that, as will be clear shortly, encompasses nearly all the familiar estimators discussed in this book. Section 13.2 will introduce the estimation framework with the method of moments. The technique of minimum distance estimation is developed in Section 13.3. Formalities of the GMM estimator are presented in Section 13.4. Section 13.5 discusses hypothesis testing based on moment equations. Major applications, including dynamic panel data models, are described in Section 13.6.

Example 13.1 Euler Equations and Life Cycle Consumption

One of the most often cited applications of the GMM principle for estimating econometric models is Hall's (1978) permanent income model of consumption. The original form of the model (with some small changes in notation) posits a hypothesis about the optimizing behavior of a consumer over the life cycle. Consumers are hypothesized to act according to the model,

$$\text{Maximize } E_t\left[\sum_{\tau=0}^{T-t}\left(\frac{1}{1+\delta}\right)^{\tau} U(c_{t+\tau})\,|\,\boldsymbol{\Omega}_t\right] \text{ subject to } \sum_{\tau=0}^{T-t}\left(\frac{1}{1+r}\right)^{\tau}(c_{t+\tau}-w_{t+\tau}) = A_t.$$

The information available at time t is denoted $\boldsymbol{\Omega}_t$ so that E_t denotes the expectation formed at time t based on the information set $\boldsymbol{\Omega}_t$. The maximand is the expected discounted stream of future utility from consumption from time t until the end of life at time T. The individual's subjective rate of time preference is $\beta = 1/(1+\delta)$. The real rate of interest, $r \geq \delta$, is assumed to be constant. The utility function $U(c_t)$ is assumed to be strictly concave and time separable (as shown in the model). One period's consumption is c_t. The intertemporal budget constraint states that the present discounted excess of c_t over earnings, w_t, over the lifetime equals

total assets A_t not including human capital. In this model, it is claimed that the only source of uncertainty is w_t. No assumption is made about the stochastic properties of w_t except that there exists an expected future earnings, $E_t[w_{t+\tau} | \boldsymbol{\Omega}_t]$. Successive values are not assumed to be independent and w_t is not assumed to be stationary.

Hall's major theorem in the paper is the solution to the optimization problem, which states

$$E_t[U'(c_{t+1}) | \boldsymbol{\Omega}_t] = \frac{1 + \delta}{1 + r} U'(c_t).$$

For our purposes, the major conclusion of the paper is "Corollary 1," which states, "No information available in time t apart from the level of consumption, c_t, helps predict future consumption, c_{t+1}, in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods t or earlier are irrelevant once c_t is known." We can use this as the basis of a model that can be placed in the GMM framework. To proceed, it is necessary to assume a form of the utility function. A common (convenient) form of the utility function is $U(c_t) = c_t^{1-\alpha}/(1 - \alpha)$, which is monotonic, $U' = c_t^{-\alpha} > 0$ and concave, $U''/U' = -\alpha/c_t < 0$. Inserting this form into the solution, rearranging the terms, and reparameterizing it for convenience, we have

$$E_t\left[(1 + r)\left(\frac{1}{1 + \delta}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\alpha} - 1 \,|\, \boldsymbol{\Omega}_t\right] = E_t[\beta(1 + r)R_{t+1}^{\lambda} - 1 \,|\, \boldsymbol{\Omega}_t] = 0,$$

where $R_{t+1} = c_{t+1}/c_t$ and $\lambda = -\alpha$.

Hall assumed that r was constant over time. Other applications of this modeling framework modified the framework so as to involve a forecasted interest rate, r_{t+1}.[1] How one proceeds from here depends on what is in the information set. The unconditional mean does not identify the two parameters. The corollary states that the only relevant information in the information set is c_t. Given the form of the model, the more natural instrument might be R_t. This assumption exactly identifies the two parameters in the model,

$$E_t\left[(\beta(1 + r_{t+1})R_{t+1}^{\lambda} - 1)\begin{pmatrix}1\\R_t\end{pmatrix}\right] = \begin{bmatrix}0\\0\end{bmatrix}.$$

As stated, the model has no testable implications. These two moment equations would exactly identify the two unknown parameters. Hall hypothesized several models involving income and consumption, which would overidentify and thus place restrictions on the model.

13.2 CONSISTENT ESTIMATION: THE METHOD OF MOMENTS

Sample statistics, such as the mean and variance, can be treated as simple descriptive measures. In our discussion of estimation in Appendix C, however, we argue that, in general, sample statistics each have a counterpart in the population, for example, the correspondence between the sample mean and the population expected value. The natural (perhaps obvious) next step in the analysis is to use this analogy to justify using the sample moments as estimators of these population parameters. What remains to establish is whether this approach is the best, or even a good way, to use the sample data to infer the characteristics of the population.

The basis of the **method of moments** is as follows: In random sampling, under generally benign assumptions, a sample statistic will converge in probability to some constant. For example, with i.i.d. random sampling, $\overline{m}_2' = (1/n)\sum_{i=1}^{n} y_i^2$ will converge in mean square to the variance plus the square of the mean of the random variable, y. This

[1]For example, Hansen and Singleton (1982).

constant will, in turn, be a function of the unknown parameters of the distribution. To estimate K parameters, $\theta_1, \ldots, \theta_K$, we can compute K such statistics, $\overline{m}_1, \ldots, \overline{m}_K$, whose **probability limits** are known functions of the parameters. These K moments are equated to the K functions, and the functions are inverted to express the parameters as functions of the moments. The moments will be consistent by virtue of a law of large numbers (Theorems D.4–D.9). They will be asymptotically normally distributed by virtue of the Lindeberg–Levy **central limit theorem** (D.18). The derived parameter estimators will inherit consistency by virtue of the Slutsky theorem (D.12) and asymptotic normality by virtue of the delta method (Theorem D.21, sometimes called the *law of propagated error*).

This section will develop this technique in some detail, partly to present it in its own right and partly as a prelude to the discussion of the generalized method of moments, or GMM, estimation technique, which is treated in Section 13.4.

13.2.1 RANDOM SAMPLING AND ESTIMATING THE PARAMETERS OF DISTRIBUTIONS

Consider independent, identically distributed random sampling from a distribution $f(y|\theta_1, \ldots \theta_K)$ with finite moments up to $E[y^{2K}]$. The **random sample** consists of n observations, $y_1, \ldots, y_n$. The kth "raw" or **uncentered moment** is

$$\overline{m}'_k = \frac{1}{n}\sum_{i=1}^{n} y_i^k.$$

By Theorem D.4,

$$E[\overline{m}'_k] = \mu'_k = E[y_i^k],$$

and

$$\text{Var}[\overline{m}'_k] = \frac{1}{n}\text{Var}[y_i^k] = \frac{1}{n}(\mu'_{2k} - \mu'^2_k).$$

By convention, $\mu'_1 = E[y_i] = \mu$. By the Khinchine theorem, D.5,

$$\text{plim}\, \overline{m}'_k = \mu'_k = E[y_i^k].$$

Finally, by the Lindeberg–Levy central limit theorem,

$$\sqrt{n}(\overline{m}'_k - \mu'_k) \xrightarrow{d} N[0, \mu'_{2k} - \mu'^2_k].$$

In general, μ'_k will be a function of the underlying parameters. By computing K raw moments and equating them to these functions, we obtain K equations that can (in principle) be solved to provide estimates of the K unknown parameters.

Example 13.2 Method of Moments Estimator for N[μ, σ²]

In random sampling from $N[\mu, \sigma^2]$,

$$\text{plim}\, \frac{1}{n}\sum_{i=1}^{n} y_i = \text{plim}\, \overline{m}'_1 = E[y] = \mu,$$

and

$$\text{plim}\, \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \text{plim}\, \overline{m}'_2 = Var[y] + \mu^2 = \sigma^2 + \mu^2.$$

Equating the right- and left-hand sides of the probability limits gives moment estimators

$$\hat{\mu} = \overline{m}'_1 = \bar{y},$$

and

$$\hat{\sigma}^2 = \overline{m}_2' - \overline{m}_1'^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Note that $\hat{\sigma}^2$ is biased, although both estimators are consistent.

Although the moments based on powers of y provide a natural source of information about the parameters, other functions of the data may also be useful. Let $m_k(\cdot)$ be a continuous and differentiable function not involving the sample size n, and let

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{n} m_k(y_i), \quad k = 1, 2, \ldots, K.$$

These are also moments of the data. It follows from Theorem D.4 and the corollary, (D-5), that

$$\text{plim } \overline{m}_k = E[m_k(y_i)] = \mu_k(\theta_1, \ldots, \theta_K).$$

We assume that $\mu_k(\cdot)$ involves some or all of the parameters of the distribution. With K parameters to be estimated, the K **moment equations,**

$$\begin{aligned}
&\overline{m}_1 - \mu_1(\theta_1, \ldots, \theta_K) = 0,\\
&\overline{m}_2 - \mu_2(\theta_1, \ldots, \theta_K) = 0,\\
&\cdots\\
&\overline{m}_K - \mu_K(\theta_1, \ldots, \theta_K) = 0,
\end{aligned}$$

provide K equations in K unknowns, $\theta_1, \ldots, \theta_K$. If the equations are continuous and functionally independent, then **method of moments estimators** can be obtained by solving the system of equations for

$$\hat{\theta}_k = \hat{\theta}_k[\overline{m}_1, \ldots, \overline{m}_K].$$

As suggested, there may be more than one set of moments that one can use for estimating the parameters, or there may be more moment equations available than are necessary.

Example 13.3 Inverse Gaussian (Wald) Distribution

The inverse Gaussian distribution is used to model survival times, or elapsed times, from some beginning time until some kind of transition takes place. The standard form of the density for this random variable is

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right], \quad y > 0, \lambda > 0, \mu > 0.$$

The mean is μ while the variance is μ^3/λ. The efficient maximum likelihood estimators of the two parameters are based on $(1/n)\sum_{i=1}^{n} y_i$ and $(1/n)\sum_{i=1}^{n}(1/y_i)$. Because the mean and variance are simple functions of the underlying parameters, we can also use the sample mean and sample variance as moment estimators of these functions. Thus, an alternative pair of method of moments estimators for the parameters of the Wald distribution can be based on $(1/n)\sum_{i=1}^{n} y_i$ and $(1/n)\sum_{i=1}^{n} y_i^2$. The precise formulas for this pair of moment estimators are left as an exercise.

Example 13.4 Mixture of Normal Distributions

Quandt and Ramsey (1978) analyzed the problem of estimating the parameters of a mixture of two normal distributions. Suppose that each observation in a random sample is drawn from one of two different normal distributions. The probability that the observation is drawn

from the first distribution, $N[\mu_1, \sigma_1^2]$, is λ and the probability that it is drawn from the second is $(1 - \lambda)$. The density for the observed y is $f(y) = \lambda\, N[\mu_1, {\sigma_1}^2] + (1 - \lambda)N[\mu_2, {\sigma_2}^2]$, $0 < \lambda < 1$. Inserting the definitions gives

$$f(y) = \frac{\lambda}{(2\pi\sigma_1^2)^{1/2}} e^{-1/2[(y-\mu_1)/\sigma_1]^2} + \frac{1-\lambda}{(2\pi\sigma_2^2)^{1/2}} e^{-1/2[(y-\mu_2)/\sigma_2]^2}.$$

Before proceeding, we note that this density is precisely the same as the finite mixture model described in Section 14.15.1. Maximum likelihood estimation of the model using the method described there would be simpler than the method of moment generating functions developed here.

The sample mean and second through fifth central moments,

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^k, \quad k = 2, 3, 4, 5,$$

provide five equations in five unknowns that can be solved (via a ninth-order polynomial) for consistent estimators of the five parameters. Because $\overline{y}$ converges in probability to $E[y_i] = \mu$, the theorems given earlier for $\overline{m}_k'$ as an estimator of μ_k' apply as well to $\overline{m}_k$ as an estimator of

$$\mu_k = E[(y_i - \mu)^k].$$

For the mixed normal distribution, the mean and variance are

$$\mu = E[y] = \lambda\mu_1 + (1 - \lambda)\mu_2$$

and

$$\sigma^2 = \text{Var}[y] = \lambda\sigma_1^2 + (1 - \lambda)\sigma_2^2 + 2\lambda(1 - \lambda)(\mu_1 - \mu_2)^2,$$

which suggests how complicated the familiar method of moments is likely to become. An alternative method of estimation proposed by the authors is based on

$$E[e^{ty}] = \lambda e^{t\mu_1 + t^2{\sigma_1}^2/2} + (1 - \lambda)e^{t\mu_2 + t^2\sigma^{2/2}_2} = \Lambda_t,$$

where t is any value not necessarily an integer. Quandt and Ramsey (1978) suggest choosing five values of t that are not too close together and using the statistics

$$\overline{M}_t = \frac{1}{n}\sum_{i=1}^{n} e^{ty_i}$$

to estimate the parameters. The moment equations are $\overline{M}_t - \Lambda_t(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = 0$. They label this procedure the **method of moment generating functions**. (See Section B.6 for the definition of the moment generating function.)

In most cases, method of moments estimators are not efficient. The exception is in random sampling from **exponential families** of distributions.

DEFINITION 13.1 Exponential Family

An exponential (parametric) family of distributions is one whose log-likelihood is of the form

$$\ln L(\boldsymbol{\theta}|\,\textbf{data}) = a(\textbf{data}) + b(\boldsymbol{\theta}) + \sum_{k=1}^{K} c_k(\textbf{data})s_k(\boldsymbol{\theta}),$$

where $a(\cdot)$, $b(\cdot)$, $c_k(\cdot)$, and $s_k(\cdot)$ are functions. The members of the "family" are distinguished by the different parameter values. The normal distribution and the Wald distribution in Example 13.3 are examples.

If the log-likelihood function is of this form, then the functions $c_k(\cdot)$ are called **sufficient statistics**.[2] When sufficient statistics exist, method of moments estimator(s) can be functions of them. In this case, the method of moments estimators will also be the maximum likelihood estimators, so, of course, they will be efficient, at least asymptotically. We emphasize, in this case, the probability distribution is fully specified. Because the normal distribution is an exponential family with sufficient statistics $\overline{m}'_1$ and $\overline{m}'_2$, the estimators described in Example 13.2 are fully efficient. (They are the maximum likelihood estimators.) The mixed normal distribution is not an exponential family. We leave it as an exercise to show that the Wald distribution in Example 13.3 is an exponential family. You should be able to show that the sufficient statistics are the ones that are suggested in Example 13.3 as the bases for the MLEs of μ and λ.

Example 13.5 Gamma Distribution

The gamma distribution (see Section B.4.5) is

$$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda y} y^{P-1}, \quad y \geq 0, P > 0, \lambda > 0.$$

The log-likelihood function for this distribution is

$$\frac{1}{n} \ln L = [P \ln \lambda - \ln \Gamma(P)] - \lambda \frac{1}{n} \sum_{i=1}^{n} y_i + (P - 1) \frac{1}{n} \sum_{i=1}^{n} \ln y_i.$$

This function is an exponential family with $a(\mathbf{data}) = 0$, $b(\boldsymbol{\theta}) = n[P \ln \lambda - \ln \Gamma(P)]$ and two sufficient statistics, $\frac{1}{n}\sum_{i=1}^{n} y_i$ and $\frac{1}{n}\sum_{i=1}^{n} \ln y_i$. The method of moments estimators based on $\frac{1}{n}\sum_{i=1}^{n} y_i$ and $\frac{1}{n}\sum_{i=1}^{n} \ln y_i$ would be the maximum likelihood estimators. But we also have

$$\text{plim} \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} y_i \\ y_i^2 \\ \ln y_i \\ 1/y_i \end{bmatrix} = \begin{bmatrix} P/\lambda \\ P(P+1)/\lambda^2 \\ \Psi(P) - \ln \lambda \\ \lambda/(P-1) \end{bmatrix}.$$

(The functions $\Gamma(P)$ and $\Psi(P) = \text{d} \ln \Gamma(P)/\text{d}P$ are discussed in Section E.2.3.) Any two of these can be used to estimate λ and P.

For the income data in Example C.1, the four moments listed earlier are

$$(\overline{m}'_1, \overline{m}'_2, \overline{m}'_*, \overline{m}'_{-1}) = \frac{1}{n} \sum_{i=1}^{n} (y_i, y_i^2, \ln y_i, 1/y_i) = (31.278, 1453.96, 3.22139, 0.050014).$$

The method of moments estimators of $\boldsymbol{\theta} = (P, \lambda)$ based on the six possible pairs of these moments are as follows:

$$(\hat{P}, \hat{\lambda}) = \begin{bmatrix} & \overline{m}'_1 & \overline{m}'_2 & \overline{m}'_{-1} \\ \overline{m}'_2 & 2.05682, 0.065759 & & \\ \overline{m}'_{-1} & 2.77198, 0.0886239 & 2.60905, 0.080475 & \\ \overline{m}'_* & 2.4106, \quad 0.0770702 & 2.26450, 0.071304 & 3.03580, 0.1018202 \end{bmatrix}.$$

The maximum likelihood estimates are $\hat{\boldsymbol{\theta}}(\overline{m}'_1, \overline{m}'_*) = (2.4106, 0.0770702)$.

13.2.2 ASYMPTOTIC PROPERTIES OF THE METHOD OF MOMENTS ESTIMATOR

In a few cases, we can obtain the exact distribution of the method of moments estimator. For example, in sampling from the normal distribution, $\hat{\mu}$ has mean μ and variance

[2] Stuart and Ord (1989, pp. 1–29) give a discussion of sufficient statistics and exponential families of distributions. A result that we will use in Chapter 17 is that if the statistics, $c_k(\mathbf{data})$, are sufficient statistics, then the conditional density, $f[y_1, \ldots, y_n | c_k(\mathbf{data}), k = 1, \ldots, K]$, is not a function of the parameters.

σ^2/n and is normally distributed, while $\hat{\sigma}^2$ has mean $[(n-1)/n]\sigma^2$ and variance $[(n-1)/n]^2 2\sigma^4/(n-1)$ and is exactly distributed as a multiple of a chi-squared variate with $(n-1)$ degrees of freedom. If sampling is not from the normal distribution, the exact variance of the sample mean will still be $Var[y]/n$, whereas an asymptotic variance for the moment estimator of the population variance could be based on the leading term in (D-27), in Example D.10, but the precise distribution may be intractable.

There are cases in which no explicit expression is available for the variance of the underlying sample moment. For instance, in Example 13.4, the underlying sample statistic is

$$\overline{M}_t = \frac{1}{n}\sum_{i=1}^{n} e^{ty_i} = \frac{1}{n}\sum_{i=1}^{n} M_{it}.$$

The exact variance of $\overline{M}_t$ is known only if t is an integer. But if sampling is random, and if $\overline{M}_t$ is a sample mean, we can estimate its variance with $1/n$ times the sample variance of the observations on M_{it}. We can also construct an estimator of the covariance of $\overline{M}_t$ and $\overline{M}_s$ with

$$\text{Est.Asy.Cov}[\overline{M}_t, \overline{M}_s] = \frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}[(e^{ty_i} - \overline{M}_t)(e^{sy_i} - \overline{M}_s)]\right\}.$$

In general, when the moments are computed as

$$\overline{m}_{n,k} = \frac{1}{n}\sum_{i=1}^{n} m_k(\mathbf{y}_i), \quad k = 1, \ldots, K,$$

where $\mathbf{y}_i$ is an observation on a vector of variables, an appropriate estimator of the asymptotic covariance matrix of $\overline{\mathbf{m}}_n = [\overline{m}_{n,1}, \ldots, \overline{m}_{n,k}]$ can be computed using

$$\frac{1}{n}\mathbf{F}_{jk} = \frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}[(m_j(\mathbf{y}_i) - \overline{m}_j)(m_k(\mathbf{y}_i) - \overline{m}_k)]\right\}, \quad j, k = 1, \ldots, K.$$

(One might divide the inner sum by $n-1$ rather than n. Asymptotically it is the same.) This estimator provides the asymptotic covariance matrix for the moments used in computing the estimated parameters. Under the assumption of i.i.d. random sampling from a distribution with finite moments, $\mathbf{F}$ will converge in probability to the appropriate covariance matrix of the normalized vector of moments, $\boldsymbol{\Phi} = \text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\theta})]$. Finally, under our assumptions of random sampling, although the precise distribution is likely to be unknown, we can appeal to the Lindeberg–Levy central limit theorem (D.18) to obtain an asymptotic approximation.

To formalize the remainder of this derivation, refer back to the moment equations, which we will now write as

$$\overline{m}_{n,k}(\theta_1, \theta_2, \ldots, \theta_K) = 0, \quad k = 1, \ldots, K.$$

The subscript n indicates the dependence on a data set of n observations. We have also combined the sample statistic (sum) and function of parameters, $\mu(\theta_1, \ldots, \theta_K)$ in this general form of the moment equation. Let $\overline{\mathbf{G}}_n(\boldsymbol{\theta})$ be the $K \times K$ matrix whose kth row is the vector of partial derivatives,

$$\overline{\mathbf{G}}'_{n,k} = \frac{\partial \overline{m}_{n,k}}{\partial \boldsymbol{\theta}'}.$$

Now, expand the set of solved moment equations around the true values of the parameters $\boldsymbol{\theta}_0$ in a linear **Taylor series**. The linear approximation is

$$\mathbf{0} \approx [\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)] + \overline{\mathbf{G}}_n'(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx -[\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}[\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)]. \tag{13-1}$$

(We have treated this as an approximation because we are not dealing formally with the higher-order term in the Taylor series. We will make this explicit in the treatment of the GMM estimator in Section 13.4.) The argument needed to characterize the large sample behavior of the estimator, $\hat{\boldsymbol{\theta}}$, is discussed in Appendix D. We have from Theorem D.18 (the central limit theorem) that $\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)$ has a limiting normal distribution with mean vector $\mathbf{0}$ and covariance matrix equal to $\boldsymbol{\Phi}$. Assuming that the functions in the moment equation are continuous and functionally independent, we can expect $\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0)$ to converge to a nonsingular matrix of constants, $\Gamma(\boldsymbol{\theta}_0)$. Under general conditions, the limiting distribution of the right-hand side of (13-1) will be that of a linear function of a normally distributed vector. Jumping to the conclusion, we expect the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ to be normal with mean vector $\boldsymbol{\theta}_0$ and covariance matrix $(1/n) \times \{-[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1}\}\boldsymbol{\Phi}\{-[\boldsymbol{\Gamma}'(\boldsymbol{\theta}_0)]^{-1}\}$. Thus, the asymptotic covariance matrix for the method of moments estimator may be estimated with

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}] = \frac{1}{n}[\overline{\mathbf{G}}_n'(\hat{\boldsymbol{\theta}})\mathbf{F}^{-1}\overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}})]^{-1}.$$

Example 13.5 (Continued)

Using the estimates $\hat{\boldsymbol{\theta}}(m_1', m_*') = (2.4106, 0.0770702)$,

$$\hat{\overline{\mathbf{G}}} = \begin{bmatrix} -1/\hat{\lambda} & \hat{P}/\hat{\lambda}^2 \\ -\hat{\Psi}' & 1/\hat{\lambda} \end{bmatrix} = \begin{bmatrix} -12.97515 & 405.8353 \\ -0.51241 & 12.97515 \end{bmatrix}.$$

[The function $\Psi'(P)$ is $d^2 \ln \Gamma(P)/dP^2 = (\Gamma\Gamma'' - \Gamma')/\Gamma^2$. With $\hat{P} = 2.4106$, $\hat{\Gamma} = 1.250832$, $\hat{\Psi} = 0.658347$, and $\hat{\Psi}' = 0.512408$].[3] The matrix $\mathbf{F}$ is the sample covariance matrix of y and $\ln y$ (using 19 as the divisor),

$$\mathbf{F} = \begin{bmatrix} 500.68 & 14.31 \\ 14.31 & 0.47746 \end{bmatrix}.$$

The product is

$$\frac{1}{n}[\hat{\mathbf{G}}'\mathbf{F}^{-1}\hat{\mathbf{G}}]^{-1} = \begin{bmatrix} 0.38978 & 0.014605 \\ 0.014605 & 0.00068747 \end{bmatrix}.$$

For the maximum likelihood estimator, the estimate of the asymptotic covariance matrix based on the expected (and actual) Hessian is

$$[-\mathbf{H}]^{-1} = \frac{1}{n}\begin{bmatrix} \Psi' & -1/\lambda \\ -1/\lambda & P/\lambda^2 \end{bmatrix}^{-1} = \begin{bmatrix} 0.51243 & 0.01638 \\ 0.01638 & 0.00064654 \end{bmatrix}.$$

The Hessian has the same elements as $\mathbf{G}$ because we chose to use the sufficient statistics for the moment estimators, so the moment equations that we differentiated are, apart from

[3] Ψ' is the trigamma function. Values for $\Gamma(P)$, $\Psi(P)$, and $\Psi'(P)$ are tabulated in Abramovitz and Stegun (1971). The values given were obtained using the IMSL computer program library.

a sign change, also the derivatives of the log-likelihood. The estimates of the two variances are 0.51203 and 0.00064654, respectively, which agrees reasonably well with the method of moments estimates. The difference would be due to sampling variability in a finite sample and the presence of **F** in the first variance estimator.

13.2.3 SUMMARY—THE METHOD OF MOMENTS

In the simplest cases, the method of moments is robust to differences in the specification of the data-generating process (DGP). A sample mean or variance estimates its population counterpart (assuming it exists), regardless of the underlying process. It is this freedom from unnecessary distributional assumptions that has made this method so popular in recent years. However, this comes at a cost. If more is known about the DGP, its specific distribution for example, then the method of moments may not make use of all of the available information. Thus, in Example 13.3, the natural estimators of the parameters of the distribution based on the sample mean and variance turn out to be inefficient. The method of maximum likelihood, which remains the foundation of much work in econometrics, is an alternative approach which utilizes this out of sample information and is, therefore, more efficient.

13.3 MINIMUM DISTANCE ESTIMATION

The preceding analysis has considered **exactly identified cases**. In each example, there were K parameters to estimate and we used K moments to estimate them. In Example 13.5, we examined the gamma distribution, a two-parameter family, and considered different pairs of moments that could be used to estimate the two parameters. The most efficient estimator for the parameters of this distribution will be based on $(1/n)\Sigma_i y_i$ and $(1/n)\Sigma_i \ln y_i$. This does raise a general question: How should we proceed if we have more moments than we need? It would seem counterproductive to simply discard the additional information. In this case, logically, the sample information provides more than one estimate of the model parameters, and it is now necessary to reconcile those competing estimators.

We have encountered this situation in several earlier examples: In Example 11.23, in Passmore et al.'s (2005) study of Fannie Mae, we have four independent estimators of a single parameter, $\hat{\alpha}_j$, each with estimated asymptotic variance $\hat{V}_j$, $j = 1, \ldots, 4$. The estimators were combined using a **criterion function**,

$$\text{minimize with respect to } \alpha : q = \sum_{j=i}^{4} \frac{(\hat{\alpha}_j - \alpha)^2}{\hat{V}_j}.$$

The solution to this minimization problem is a minimum distance estimator,

$$\hat{\alpha}_{\text{MDE}} = \sum_{j=1}^{4} w_j \hat{\alpha}_j,\ w_j = \frac{1/\hat{V}_j}{\sum_{s=1}^{4}(1/\hat{V}_s)},\ j = 1, \ldots, 4 \text{ and } \sum_{j=1}^{4} w_j = 1.$$

In forming the two-stage least squares estimator of the parameters in a dynamic panel data model in Section 11.10.3, we obtained $T - 2$ instrumental variable estimators of the parameter vector $\boldsymbol{\theta}$ by forming different instruments for each period for which we had sufficient data. The $T - 2$ estimators of the same parameter vector are $\hat{\boldsymbol{\theta}}_{\text{IV}(t)}$. The Arellano–Bond estimator of the single parameter vector in this setting is

$$\hat{\boldsymbol{\theta}}_{\text{IV}} = \left(\sum_{t=3}^{T} \mathbf{W}_{(t)} \right)^{-1} \left(\sum_{t=3}^{T} \mathbf{W}_{(t)} \hat{\boldsymbol{\theta}}_{\text{IV}(t)} \right) = \sum_{t=3}^{T} \mathbf{R}_{(t)} \hat{\boldsymbol{\theta}}_{\text{IV}(t)},$$

where

$$\mathbf{W}_{(t)} = \left(\hat{\bar{\mathbf{X}}}'_{(t)}\hat{\bar{\mathbf{X}}}_{(t)}\right), \mathbf{R}_{(t)} = \left(\sum_{t=3}^{T}\mathbf{W}_{(t)}\right)^{-1}\mathbf{W}_{(t)} \text{ and } \sum_{t=3}^{T}\mathbf{R}_{(t)} = \mathbf{I}.$$

Finally, Carey's (1997) analysis of hospital costs that we examined in Example 11.13 involved a seemingly unrelated regressions model that produced multiple estimates of several of the model parameters. We will revisit this application in Example 13.6.

A **minimum distance estimator (MDE)** is defined as follows: Let $\overline{m}_{n,l}$ denote a sample statistic based on n observations such that

$$\text{plim } \overline{m}_{n,l} = g_l(\boldsymbol{\theta}_0), l = 1, \ldots, L,$$

where $\boldsymbol{\theta}_0$ is a vector of $K \leq L$ parameters to be estimated. Arrange these moments and functions in $L \times 1$ vectors $\overline{\mathbf{m}}_n$ and $\mathbf{g}(\boldsymbol{\theta}_0)$ and further assume that the statistics are jointly asymptotically normally distributed with plim $\overline{\mathbf{m}}_n = \mathbf{g}(\boldsymbol{\theta})$ and Asy.Var$[\overline{\mathbf{m}}_n] = (1/n)\boldsymbol{\Phi}$. Define the criterion function

$$q = [\overline{\mathbf{m}}_n - \mathbf{g}(\boldsymbol{\theta})]'\mathbf{W}[\overline{\mathbf{m}}_n - \mathbf{g}(\boldsymbol{\theta})]$$

for a positive definite **weighting matrix**, $\mathbf{W}$. The minimum distance estimator is the $\hat{\boldsymbol{\theta}}_{\text{MDE}}$ that minimizes q. Different choices of $\mathbf{W}$ will produce different estimators, but the estimator has the following properties for any $\mathbf{W}$:

THEOREM 13.1 Asymptotic Distribution of the Minimum Distance Estimator

Under the assumption that $\sqrt{n}[\overline{\mathbf{m}}_n - \mathbf{g}(\boldsymbol{\theta}_0)] \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}]$, *the asymptotic properties of the minimum distance estimator are as follows:*

$$\text{plim } \hat{\boldsymbol{\theta}}_{\text{MDE}} = \boldsymbol{\theta}_0,$$

$$\text{Asy.Var}[\boldsymbol{\theta}_{\text{MDE}}] = \frac{1}{n}[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1}[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)][\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1} = \frac{1}{n}\mathbf{V},$$

where

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}_0) = \text{plim } \mathbf{G}(\hat{\boldsymbol{\theta}}_{\text{MDE}}) = \text{plim } \frac{\partial \mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MDE}})}{\partial \hat{\boldsymbol{\theta}}'_{\text{MDE}}},$$

and

$$\hat{\boldsymbol{\theta}}_{\text{MDE}} \xrightarrow{a} N\left[\boldsymbol{\theta}_0, \frac{1}{n}\mathbf{V}\right].$$

Proofs may be found in Malinvaud (1970) and Amemiya (1985). For our purposes, we note that the MDE is an extension of the method of moments presented in the preceding section. One implication is that the estimator is consistent for any $\mathbf{W}$, but the asymptotic covariance matrix is a function of $\mathbf{W}$. This suggests that the choice of $\mathbf{W}$ might be made with an eye toward the size of the covariance matrix and that there might be an optimal choice. That does, indeed, turn out to be the case. For minimum distance estimation, the weighting matrix that produces the smallest variance is

$$\textbf{optimal weighting matrix: } \mathbf{W}^* = [\text{Asy.Var.}\sqrt{n}\{\overline{\mathbf{m}}_n - \mathbf{g}(\boldsymbol{\theta})\}]^{-1} = \boldsymbol{\Phi}^{-1}.$$

[See Hansen (1982) for discussion.] With this choice of $\mathbf{W}$,

$$\text{Asy.Var}[\hat{\boldsymbol{\theta}}_{\text{MDE}}] = \frac{1}{n}[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1},$$

which is the result we had earlier for the method of moments estimator.

The solution to the MDE estimation problem is found by locating the $\hat{\boldsymbol{\theta}}_{\text{MDE}}$ such that

$$\frac{\partial q}{\partial \hat{\boldsymbol{\theta}}_{\text{MDE}}} = -\mathbf{G}(\hat{\boldsymbol{\theta}}_{\text{MDE}})'\mathbf{W}[\overline{\mathbf{m}}_n - \mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MDE}})] = \mathbf{0}.$$

An important aspect of the MDE arises in the exactly identified case. If K equals L, and if the functions $g_l(\boldsymbol{\theta})$ are functionally independent, that is, $\mathbf{G}(\boldsymbol{\theta})$ has full row rank, K, then it is possible to solve the moment equations exactly. That is, the minimization problem becomes one of simply solving the K moment equations, $\overline{m}_{n,l} = g_l(\boldsymbol{\theta}_0)$ in the K unknowns, $\hat{\boldsymbol{\theta}}_{\text{MDE}}$. This is the method of moments estimator examined in the preceding section. In this instance, the weighting matrix, $\mathbf{W}$, is irrelevant to the solution, because the MDE will now satisfy the moment equations

$$[\overline{\mathbf{m}}_n - \mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MDE}})] = \mathbf{0}.$$

For the examples listed earlier, which are all for **overidentified cases**, the minimum distance estimators are defined by

$$q = ((\hat{\alpha}_1 - \alpha)(\hat{\alpha}_2 - \alpha)(\hat{\alpha}_3 - \alpha)(\hat{\alpha}_4 - \alpha))\begin{bmatrix} \hat{V}_1 & 0 & 0 & 0 \\ 0 & \hat{V}_2 & 0 & 0 \\ 0 & 0 & \hat{V}_3 & 0 \\ 0 & 0 & 0 & \hat{V}_4 \end{bmatrix}^{-1}\begin{pmatrix} (\hat{\alpha}_1 - \alpha) \\ (\hat{\alpha}_2 - \alpha) \\ (\hat{\alpha}_3 - \alpha) \\ (\hat{\alpha}_4 - \alpha) \end{pmatrix}$$

for Passmore's analysis of Fannie Mae, and

$$q = ((\mathbf{b}_{\text{IV}(3)} - \boldsymbol{\theta}) \ldots (\mathbf{b}_{\text{IV}(T)} - \boldsymbol{\theta}))'\begin{bmatrix} (\hat{\widetilde{\mathbf{X}}}'_{(3)}\hat{\widetilde{\mathbf{X}}}'_{(3)}) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & (\hat{\widetilde{\mathbf{X}}}'_{(T)}\hat{\widetilde{\mathbf{X}}}_{(T)}) \end{bmatrix}^{-1}\begin{pmatrix} (\mathbf{b}_{\text{IV}(3)} - \boldsymbol{\theta}) \\ \vdots \\ (\mathbf{b}_{\text{IV}(T)} - \boldsymbol{\theta}) \end{pmatrix}$$

for the Arellano–Bond estimator of the dynamic panel data model.

Example 13.6 Minimum Distance Estimation of a Hospital Cost Function

In Carey's (1997) study of hospital costs in Example 11.13, Chamberlain's (1984) seemingly unrelated regressions (SUR) approach to a panel data model produces five period-specific estimates of a parameter vector, $\boldsymbol{\theta}_t$. Some of the parameters are specific to the year while others (it is hypothesized) are common to all five years. There are two specific parameters of interest, β_D and β_O, that are allowed to vary by year, but are each estimated multiple times by the SUR model. We focus on just these parameters. The model states

$$y_{it} = \alpha_i + A_{it} + \beta_{D,t}\, DIS_{it} + \beta_{O,t}\, OUT_{it} + \varepsilon_{it},$$

where

$$\alpha_i = B_i + \Sigma_t \gamma_{D,t}\, DIS_{it} + \Sigma_t \gamma_{O,t}\, OUT_{it} + u_i,\ t = 1987, \ldots, 1991.$$

DIS_{it} is patient discharges, and OUT_{it} is outpatient visits. (We are changing Carey's notation slightly and suppressing parts of the model that are extraneous to the development here. The terms A_{it} and B_i contain those additional components.) The preceding model is

estimated by inserting the expression for α_i in the main equation, then fitting an unrestricted seemingly unrelated regressions model by FGLS. There are five years of data, hence five sets of estimates. Note, however, with respect to the discharge variable, *DIS*, although each equation provides separate estimates of ($\gamma_{D,1}$, . . . , ($\beta_{D,t} + \gamma_{D,t}$), . . . , $\gamma_{D,5}$), a total of five parameter estimates in each equation (year), there are only 10, not 25 parameters to be estimated in total. The parameters on OUT_{it} are likewise overidentified. Table 13.1 reproduces the estimates in Table 11.7 for the discharge coefficients and adds the estimates for the outpatient variable.

TABLE 13.1a Coefficient Estimates for DIS in SUR Model for Hospital Costs

Coefficient on Variable in the Equation

Equation	*DIS87*	*DIS88*	*DIS89*	*DIS90*	*DIS91*
SUR87	$\beta_{D,87} + \gamma_{D,87}$	$\gamma_{D,88}$	$\gamma_{D,89}$	$\gamma_{D,90}$	$\gamma_{D,91}$
	1.76	0.116	−0.0881	0.0570	−0.0617
SUR88	$\gamma_{D,87}$	$\beta_{D,88} + \gamma_{D,88}$	$\gamma_{D,89}$	$\gamma_{D,90}$	$\gamma_{D,91}$
	0.254	1.61	−0.0934	0.0610	−0.0514
SUR89	$\gamma_{D,87}$	$\gamma_{D,88}$	$\beta_{D,89} + \gamma_{D,89}$	$\gamma_{D,90}$	$\gamma_{D,91}$
	0.217	0.0846	1.51	0.0454	−0.0253
SUR90	$\gamma_{D,87}$	$\gamma_{D,88}$	$\gamma_{D,89}$	$\beta_{D,90} + \gamma_{D,90}$	$\gamma_{D,91}$
	0.179	0.0822	0.0295	1.57	0.0244
SUR91	$\gamma_{D,87}$	$\gamma_{D,88}$	$\gamma_{D,89}$	$\gamma_{D,90}$	$\beta_{D,91} + \gamma_{D,91}$
	0.153	0.0363	−0.0422	0.0813	1.70
MDE	$\beta = 1.50$	$\beta = 1.58$	$\beta = 1.54$	$\beta = 1.57$	$\beta = 1.63$
	$\gamma = 0.219$	$\gamma = 0.0666$	$\gamma = -0.0539$	$\gamma = 0.0690$	$\gamma = -0.0213$

TABLE 13.1b Coefficient Estimates for OUT in SUR Model for Hospital Costs

Coefficient on Variable in the Equation

Equation	*OUT87*	*OUT88*	*OUT89*	*OUT90*	*OUT91*
SUR87	$\beta_{O,87} + \gamma_{D,87}$	$\gamma_{O,88}$	$\gamma_{O,89}$	$\gamma_{O,90}$	$\gamma_{O,91}$
	0.0139	0.00292	0.00157	0.000951	0.000678
SUR88	$\gamma_{O,87}$	$\beta_{O,88} + \gamma_{O,88}$	$\gamma_{O,89}$	$\gamma_{O,90}$	$\gamma_{O,91}$
	0.00347	0.0125	0.00501	0.00550	0.00503
SUR89	$\gamma_{O,87}$	$\gamma_{O,88}$	$\beta_{O,89} + \gamma_{O,89}$	$\gamma_{O,90}$	$\gamma_{O,91}$
	0.00118	0.00159	0.00832	−0.00220	−0.00156
SUR90	$\gamma_{O,87}$	$\gamma_{O,88}$	$\gamma_{O,89}$	$\beta_{O,90} + \gamma_{O,90}$	$\gamma_{O,91}$
	−0.00226	−0.00155	0.000401	0.00897	0.000450
SUR91	$\gamma_{O,87}$	$\gamma_{O,88}$	$\gamma_{O,89}$	$\gamma_{O,90}$	$\beta_{O,91} + \gamma_{O,91}$
	0.00278	0.00255	0.00233	0.00305	0.0105
MDE	$\beta = 0.0112$	$\beta = 0.00999$	$\beta = 0.0100$	$\beta = 0.00915$	$\beta = 0.00793$
	$\gamma = 0.00177$	$\gamma = 0.00408$	$\gamma = -0.00011$	$\gamma = -0.00073$	$\gamma = 0.00267$

Looking at the tables we see that the SUR model provides four direct estimates of $\gamma_{D,87}$, based on the 1988–1991 equations. It also implicitly provides four estimates of $\beta_{D,87}$ because any of the four estimates of $\gamma_{D,87}$ from the last four equations can be subtracted from the coefficient on DIS in the 1987 equation to estimate $\beta_{D,87}$. There are 50 parameter estimates of different functions of the 20 underlying parameters,

$$\theta = (\beta_{D,87}, \ldots, \beta_{D,91}), (\gamma_{D,87}, \ldots, \gamma_{D,91}), (\beta_{O,87}, \ldots, \beta_{O,91}), (\gamma_{O,87}, \ldots, \gamma_{O,91}),$$

and, therefore, 30 constraints to impose in finding a common, restricted estimator. An MDE was used to reconcile the competing estimators.

Let $\hat{\boldsymbol{\beta}}_t$ denote the 10×1 period-specific estimator of the model parameters. Unlike the other cases we have examined, the individual estimates here are not uncorrelated. In the SUR model, the estimated asymptotic covariance matrix is the partitioned matrix given in (10-7). For the estimators of two equations,

$$\text{Est.Asy.Cov}[\hat{\boldsymbol{\beta}}_t, \hat{\boldsymbol{\beta}}_s] = \text{the } t, s \text{ block of } \begin{bmatrix} \hat{\sigma}^{11}\mathbf{X}_1'\mathbf{X}_1 & \hat{\sigma}^{12}\mathbf{X}_1'\mathbf{X}_2 & \cdots & \hat{\sigma}^{15}\mathbf{X}_1'\mathbf{X}_5 \\ \hat{\sigma}^{21}\mathbf{X}_2'\mathbf{X}_1 & \hat{\sigma}^{22}\mathbf{X}_2'\mathbf{X}_2 & \cdots & \hat{\sigma}^{25}\mathbf{X}_2'\mathbf{X}_5 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}^{51}\mathbf{X}_5'\mathbf{X}_1 & \hat{\sigma}^{52}\mathbf{X}_5'\mathbf{X}_2 & \cdots & \hat{\sigma}^{55}\mathbf{X}_5'\mathbf{X}_5 \end{bmatrix}^{-1} = \hat{\mathbf{V}}_{ts},$$

where $\hat{\sigma}^{ts}$ is the t,s element of $\hat{\boldsymbol{\Sigma}}^{-1}$. (We are extracting a submatrix of the relevant matrices here because Carey's SUR model contained 26 other variables in each equation in addition to the five periods of DIS and OUT). The 50×50 weighting matrix for the MDE is

$$\mathbf{W} = \begin{bmatrix} \hat{\mathbf{V}}_{87,87} & \hat{\mathbf{V}}_{87,88} & \hat{\mathbf{V}}_{87,89} & \hat{\mathbf{V}}_{87,90} & \hat{\mathbf{V}}_{87,91} \\ \hat{\mathbf{V}}_{88,87} & \hat{\mathbf{V}}_{88,88} & \hat{\mathbf{V}}_{88,89} & \hat{\mathbf{V}}_{88,90} & \hat{\mathbf{V}}_{88,91} \\ \hat{\mathbf{V}}_{89,87} & \hat{\mathbf{V}}_{89,88} & \hat{\mathbf{V}}_{89,89} & \hat{\mathbf{V}}_{89,90} & \hat{\mathbf{V}}_{89,91} \\ \hat{\mathbf{V}}_{90,87} & \hat{\mathbf{V}}_{90,88} & \hat{\mathbf{V}}_{90,89} & \hat{\mathbf{V}}_{90,90} & \hat{\mathbf{V}}_{90,91} \\ \hat{\mathbf{V}}_{91,87} & \hat{\mathbf{V}}_{91,88} & \hat{\mathbf{V}}_{91,89} & \hat{\mathbf{V}}_{91,90} & \hat{\mathbf{V}}_{91,91} \end{bmatrix}^{-1} = [\hat{\mathbf{V}}^{ts}].$$

The vector of the quadratic form is a stack of five 10×1 vectors; the first is

$$\overline{\mathbf{m}}_{n,87} - \mathbf{g}_{87}(\boldsymbol{\theta})$$

$$= \begin{bmatrix} \{\hat{\beta}^{87}_{D,87} - (\beta_{D,87} + \gamma_{D,87})\}, \{\hat{\beta}^{87}_{D,88} - \gamma_{D,88}\}, \{\hat{\beta}^{87}_{D,89} - \gamma_{D,89}\}, \{\hat{\beta}^{87}_{D,90} - \gamma_{D,90}\}, \{\hat{\beta}^{87}_{D,91} - \gamma_{D,90}\}, \\ \{\hat{\beta}^{87}_{O,87} - (\beta_{O,87} + \gamma_{O,87})\}, \{\hat{\beta}^{87}_{O,88} - \gamma_{O,88}\}, \{\hat{\beta}^{87}_{O,89} - \gamma_{O,89}\}, \{\hat{\beta}^{87}_{O,90} - \gamma_{O,90}\}, \{\hat{\beta}^{87}_{O,91} - \gamma_{O,90}\} \end{bmatrix}'$$

for the 1987 equation and likewise for the other four equations. The MDE criterion function for this model is

$$q = \sum_{t=1987}^{1991} \sum_{s=1987}^{1991} [\overline{\mathbf{m}}_t - \mathbf{g}_t(\boldsymbol{\theta})]' \hat{V}^{ts} [\overline{\mathbf{m}}_s - \mathbf{g}_s(\boldsymbol{\theta})].$$

Note there are 50 estimated parameters from the SUR equations (those are listed in Table 13.1) and 20 unknown parameters to be calibrated in the criterion function. The reported minimum distance estimates are shown in the last row of each table.

13.4 THE GENERALIZED METHOD OF MOMENTS (GMM) ESTIMATOR

A large proportion of the recent empirical work in econometrics, particularly in macroeconomics and finance, has employed GMM estimators. As we shall see, this broad class of estimators, in fact, includes most of the estimators discussed elsewhere in this book.

The GMM estimation technique is an extension of the minimum distance estimator described in Section 13.3.[4] In the following, we will extend the generalized method of moments to other models beyond the generalized linear regression, and we will fill in some gaps in the derivation in Section 13.2.

13.4.1 ESTIMATION BASED ON ORTHOGONALITY CONDITIONS

Consider the least squares estimator of the parameters in the classical linear regression model. An important assumption of the model is

$$E[\mathbf{x}_i\varepsilon_i] = E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

The sample analog is

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

The estimator of $\boldsymbol{\beta}$ is the one that satisfies these moment equations, which are just the normal equations for the least squares estimator. So we see that the OLS estimator is a method of moments estimator.

For the instrumental variables estimator of Chapter 8, we relied on a large sample analog to the moment condition,

$$\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right) = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right) = \mathbf{0}.$$

We resolved the problem of having more instruments than parameters by solving the equations

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}'\hat{\boldsymbol{\varepsilon}}\right) = \frac{1}{n}\hat{\mathbf{X}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i\hat{\varepsilon}_i = \mathbf{0},$$

where the columns of $\hat{\mathbf{X}}$ are the fitted values in regressions on all the columns of $\mathbf{Z}$ (that is, the projections of these columns of $\mathbf{X}$ into the column space of $\mathbf{Z}$). (See Section 8.3.4 for further details.)

The nonlinear least squares estimator was defined similarly, although in this case the normal equations are more complicated because the estimator is only implicit. The population **orthogonality condition** for the nonlinear regression model is $E[\mathbf{x}_i^0\varepsilon_i] = \mathbf{0}$. The **empirical moment equation** is

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial E[y_i|\mathbf{x}_i, \boldsymbol{\beta}]}{\partial\boldsymbol{\beta}}\right)(y_i - E[y_i|\mathbf{x}_i, \boldsymbol{\beta}]) = \mathbf{0}.$$

Maximum likelihood estimators are obtained by equating the derivatives of a log-likelihood to zero. The scaled log-likelihood function is

$$\frac{1}{n}\ln L = \frac{1}{n}\sum_{i=1}^{n}\ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}),$$

[4]Formal presentation of the results required for this analysis are given by Hansen (1982); Hansen and Singleton (1988); Chamberlain (1987); Cumby, Huizinga, and Obstfeld (1983); Newey (1984, 1985a,b); Davidson and MacKinnon (1993); and Newey and McFadden (1994). Useful summaries of GMM estimation are provided by Pagan and Wickens (1989) and Matyas (1999). An application of some of these techniques that contains useful summaries is Pagan and Vella (1989). Some further discussion can be found in Davidson and MacKinnon (2004). Ruud (2000) provides many of the theoretical details. Hayashi (2000) is another extensive treatment of estimation centered on GMM estimators.

where $f(\cdot)$ is the density function and $\boldsymbol{\theta}$ is the parameter vector. For densities that satisfy the regularity conditions [see Section 14.4.1],

$$E\left[\frac{\partial \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}.$$

The maximum likelihood estimator is obtained by equating the sample analog to zero:

$$\frac{1}{n}\frac{\partial \ln L}{\partial \hat{\boldsymbol{\theta}}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} = \mathbf{0}.$$

(Dividing by n to make this result comparable to our earlier ones does not change the solution.) The upshot is that nearly all the estimators we have discussed and will encounter later can be construed as method of moments estimators. [Manski's (1992) treatment of **analog estimation** provides some interesting extensions and methodological discourse.]

As we extend this line of reasoning, it will emerge that most of the estimators defined in this book can be viewed as generalized method of moments estimators.

13.4.2 GENERALIZING THE METHOD OF MOMENTS

The preceding examples all have a common aspect. In each case listed, save for the general case of the instrumental variable estimator, there are exactly as many moment equations as there are parameters to be estimated. Thus, each of these are **exactly identified** cases. There will be a single solution to the moment equations, and at that solution, the equations will be exactly satisfied.[5] But there are cases in which there are more moment equations than parameters, so the system is overdetermined.

In Example 13.5, we defined four sample moments,

$$\bar{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i, y_i^2, \frac{1}{y_i}, \ln y_i\right],$$

with probability limits P/λ, $P(P+1)/\lambda^2$, $\lambda/(P-1)$, and $\psi(P) - \ln\lambda$, respectively. Any pair could be used to estimate the two parameters, but as shown in the earlier example, the six pairs produce six somewhat different estimates of $\boldsymbol{\theta} = (P, \lambda)$.

In such a case, to use all the information in the sample it is necessary to devise a way to reconcile the conflicting estimates that may emerge from the overdetermined system. More generally, suppose that the model involves K parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)'$, and that the theory provides a set of $L > K$ moment conditions,

$$E[m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta})] = E[m_{il}(\boldsymbol{\theta})] = 0,$$

where y_i, $\mathbf{x}_i$, and $\mathbf{z}_i$ are variables that appear in the model and the subscript i on $m_{il}(\boldsymbol{\theta})$ indicates the dependence on $(y_i, \mathbf{x}_i, \mathbf{z}_i)$. Denote the corresponding sample means as

$$\overline{m}_l(\mathbf{y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_{il}(\boldsymbol{\theta}).$$

Unless the equations are functionally dependent, the system of L equations in K unknown parameters,

$$\overline{m}_l(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = 0, \quad l = 1, \ldots, L,$$

[5] That is, of course if there is *any* solution. In the regression model with multicollinearity, there are K parameters but fewer than K independent moment equations.

will not have a unique solution.[6] For convenience, the moment equations are defined implicitly here as opposed to equalities of moments to functions as in Section 13.3. It will be necessary to reconcile the $\binom{L}{K}$ different sets of estimates that can be produced. One possibility is to minimize a criterion function, such as the sum of squares,[7]

$$q = \sum_{l=1}^{L} \bar{m}_l^2 = \bar{\mathbf{m}}(\boldsymbol{\theta})' \bar{\mathbf{m}}(\boldsymbol{\theta}). \tag{13-2}$$

It can be shown that under the assumptions we have made so far, specifically that plim $\bar{\mathbf{m}}(\boldsymbol{\theta}) = E[\bar{\mathbf{m}}(\boldsymbol{\theta})] = \mathbf{0}$, the minimizer of q in (13-2) produces a consistent, though possibly inefficient, estimator of $\boldsymbol{\theta}$.[8] We can use as the criterion a weighted sum of squares,

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W}_n \bar{\mathbf{m}}(\boldsymbol{\theta}),$$

where $\mathbf{W}_n$ is *any* positive definite matrix that may depend on the data but is not a function of $\boldsymbol{\theta}$, such as $\mathbf{I}$ in (13-2), to produce a consistent estimator of $\boldsymbol{\theta}$.[9] For example, we might use a diagonal matrix of weights if some information were available about the importance (by some measure) of the different moments. We do make the additional assumption that plim $\mathbf{W}_n =$ a positive definite matrix, $\mathbf{W}$.

By the same logic that makes generalized least squares preferable to ordinary least squares, it should be beneficial to use a weighted criterion in which the weights are inversely proportional to the variances of the moments. Let $\mathbf{W}$ be a diagonal matrix whose diagonal elements are the reciprocals of the variances of the individual moments,

$$w_{ll} = \frac{1}{\text{Asy.Var}[\sqrt{n}\bar{m}_l]} = \frac{1}{\phi_{ll}}.$$

(We have written it in this form to emphasize that the right-hand side involves the variance of a sample mean which is of order $(1/n)$.) Then, a **weighted least squares** estimator would minimize

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \boldsymbol{\Phi}^{-1} \bar{\mathbf{m}}(\boldsymbol{\theta}). \tag{13-3}$$

In general, the L elements of $\bar{\mathbf{m}}$ are freely correlated. In (13-3), we have used a diagonal $\mathbf{W}$ that ignores this correlation. To use generalized least squares, we would define the full matrix,

$$\mathbf{W} = \{\text{Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}]\}^{-1} = \boldsymbol{\Phi}^{-1}. \tag{13-4}$$

The estimators defined by choosing $\boldsymbol{\theta}$ to minimize

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W}_n \bar{\mathbf{m}}(\boldsymbol{\theta})$$

[6] It may if L is greater than the sample size, n. We assume that L is strictly less than n.

[7] This approach is one that Quandt and Ramsey (1978) suggested for the problem in Example 13.4.

[8] See, for example, Hansen (1982).

[9] In principle, the weighting matrix can be a function of the parameters as well. See Hansen, Heaton, and Yaron (1996) for discussion. Whether this provides any benefit in terms of the asymptotic properties of the estimator seems unlikely. The one payoff the authors do note is that certain estimators become invariant to the sort of normalization that is discussed in Example 14.1. In practical terms, this is likely to be a consideration only in a fairly small class of cases.

are minimum distance estimators as defined in Section 13.3. The general result is that if $\mathbf{W}_n$ is a positive definite matrix and if

$$\text{plim } \bar{\mathbf{m}}(\boldsymbol{\theta}) = \mathbf{0},$$

then the minimum distance (GMM) estimator of $\boldsymbol{\theta}$ is consistent.[10] Because the OLS criterion in (13-2) uses $\mathbf{I}$, this method produces a consistent estimator, as does the weighted least squares estimator and the full GLS estimator. What remains to be decided is the best $\mathbf{W}$ to use. Intuition might suggest (correctly) that the one defined in (13-4) would be optimal, once again based on the logic that motivates generalized least squares. This result is the now-celebrated one of Hansen (1982).

The asymptotic covariance matrix of this **generalized method of moments (GMM) estimator** is

$$\mathbf{V}_{GMM} = \frac{1}{n}[\boldsymbol{\Gamma}'\mathbf{W}\boldsymbol{\Gamma}]^{-1} = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}]^{-1}, \tag{13-5}$$

where $\boldsymbol{\Gamma}$ is the matrix of derivatives with jth row equal to

$$\boldsymbol{\Gamma}^j = \text{plim } \frac{\partial \bar{m}_j(\boldsymbol{\theta})}{\partial \bar{\boldsymbol{\theta}}},$$

and $\boldsymbol{\Phi} = \text{Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}]$. Finally, by virtue of the central limit theorem applied to the sample moments and the **Slutsky theorem** applied to this manipulation, we can expect the estimator to be asymptotically normally distributed. We will revisit the asymptotic properties of the estimator in Section 13.4.3.

Example 13.7 GMM Estimation of a Nonlinear Regression Model

In Example 7.6, we examined a nonlinear regression model for income using the German Socioeconomic Panel Data set. The regression model was

$$\textit{Income} = h(1, \textit{Age}, \textit{Education}, \textit{Female}, \boldsymbol{\gamma}) + \varepsilon,$$

where $h(.)$ is an exponential function of the variables. In the example, we used several interaction terms. In this application, we will simplify the conditional mean function somewhat, and use

$$\textit{Income} = \exp(\gamma_1 + \gamma_2 \textit{Age} + \gamma_3 \textit{Education} + \gamma_4 \textit{Female}) + \varepsilon,$$

which, for convenience, we will write $y_i = \exp(\mathbf{x}_i'\boldsymbol{\gamma}) + \varepsilon_i = \mu_i + \varepsilon_i$.[11] The sample consists of the 1988 wave of the panel, less two observations for which *Income* equals zero. The resulting sample contains 4,481 observations. Descriptive statistics for the sample data are given in Table 7.2. We will first consider nonlinear least squares estimation of the parameters. The normal equations for nonlinear least squares will be

$$(1/n)\Sigma_i[(y_i - \mu_i)\mu_i\mathbf{x}_i] = (1/n)\Sigma_i[\varepsilon_i\mu_i\mathbf{x}_i] = \mathbf{0}.$$

Note that the orthogonality condition involves the pseudoregressors, $\partial\mu_i/\partial\gamma = \mathbf{x}_i^0 = \mu_i\mathbf{x}_i$. The implied population moment equation is $E[\varepsilon_i(\mu_i\mathbf{x}_i)] = \mathbf{0}$. Computation of the nonlinear least squares estimator is discussed in Section 7.2.8. The estimator of the asymptotic covariance matrix is

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}_{\text{NLSQ}}] = \frac{\Sigma_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{(4{,}481 - 4)}\left[\sum_{i=1}^{4{,}481}(\hat{\mu}_i\mathbf{x}_i)(\hat{\mu}_i\mathbf{x}_i)'\right]^{-1}, \quad \text{where } \hat{\mu}_i = \exp(\mathbf{x}_i'\hat{\boldsymbol{\gamma}}).$$

[10] In the most general cases, a number of other subtle conditions must be met so as to assert consistency and the other properties we discuss. For our purposes, the conditions given will suffice. Minimum distance estimators are discussed in Malinvaud (1970), Hansen (1982), and Amemiya (1985).

[11] We note that in this model, it is likely that *Education* is endogenous. It would be straightforward to accommodate that in the GMM estimator. However, for purposes of a straightforward numerical example, we will proceed assuming that *Education* is exogenous.

A simple method of moments estimator might be constructed from the hypothesis that $\mathbf{x}_i$ (not $\mathbf{x}_i^0$) is orthogonal to ε_i. Then,

$$E[\varepsilon_i \mathbf{x}_i] = E\left[\varepsilon_i \begin{pmatrix} 1 \\ Age_i \\ Education_i \\ Female_i \end{pmatrix}\right] = \mathbf{0}$$

implies four moment equations. The sample counterparts will be

$$\bar{m}_k(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i)\, x_{ik} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_{ik}.$$

In order to compute the method of moments estimator, we will minimize the sum of squares,

$$\bar{\mathbf{m}}'(\boldsymbol{\gamma})\,\bar{\mathbf{m}}(\boldsymbol{\gamma}) = \sum_{k=1}^{4}\bar{m}_k^2(\boldsymbol{\gamma}).$$

This is a nonlinear optimization problem that must be solved iteratively using the methods described in Section E.3.

With the first-step estimated parameters, $\hat{\boldsymbol{\gamma}}^0$. in hand, the covariance matrix is estimated using (13-5).

$$\hat{\boldsymbol{\Phi}} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{4{,}481}\mathbf{m}_i(\hat{\boldsymbol{\gamma}}^0)\mathbf{m}_i'(\hat{\boldsymbol{\gamma}}^0)\right\} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{4{,}481}(\hat{\varepsilon}_i^0\mathbf{x}_i)(\hat{\varepsilon}_i^0\mathbf{x}_i)'\right\}$$

$$\bar{\mathbf{G}} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{n}\mathbf{x}_i(-\mu_i^0\mathbf{x}_i)'\right\}.$$

The asymptotic covariance matrix for the MOM estimator is computed using (13-5),

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}_{\text{MOM}}] = \frac{1}{n}[\bar{\mathbf{G}}\hat{\boldsymbol{\Phi}}^{-1}\bar{\mathbf{G}}']^{-1}.$$

Suppose we have in hand additional variables, *Health Satisfaction* and *Marital Status*, such that although the conditional mean function remains as given previously, we will use them to form a GMM estimator. This provides two additional moment equations,

$$E\left[\varepsilon_i \begin{pmatrix} Health\ Satisfaction_i \\ Marital\ Status_i \end{pmatrix}\right],$$

for a total of six moment equations for estimating the four parameters. We constuct the generalized method of moments estimator as follows: The initial step is the same as before, except the sum of squared moments, $\bar{\mathbf{m}}'(\boldsymbol{\gamma})\bar{\mathbf{m}}(\boldsymbol{\gamma})$, is summed over six rather than four terms. We then construct

$$\boldsymbol{\Phi} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{4{,}481}\bar{\mathbf{m}}_i(\hat{\boldsymbol{\gamma}})\mathbf{m}_i'(\hat{\boldsymbol{\gamma}})\right\} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{4{,}481}(\hat{\varepsilon}_i\mathbf{z}_i)(\hat{\varepsilon}_i\mathbf{z}_i)'\right\},$$

where now $\mathbf{z}_i$ in the second term is the six exogenous variables, rather than the original four (including the constant term). Thus, $\hat{\boldsymbol{\Phi}}$ is now a 6×6 moment matrix. The optimal weighting matrix for estimation (developed in the next section) is $\hat{\boldsymbol{\Phi}}^{-1}$. The GMM estimator is computed by minimizing with respect to $\boldsymbol{\gamma}$

$$q = \bar{\mathbf{m}}'(\boldsymbol{\gamma})\hat{\boldsymbol{\Phi}}^{-1}\bar{\mathbf{m}}(\boldsymbol{\gamma}).$$

The asymptotic covariance matrix is computed using (13-5) as it was for the simple method of moments estimator.

TABLE 13.2 Nonlinear Regression Estimates (Standard errors in parentheses)

Estimate	*Nonlinear Least Squares*	*Method of Moments*	*First Step GMM*	*GMM*
Constant	−1.69331	−1.62969	−1.45551	−1.61192
	(0.04408)	(0.04214)	(0.10102)	(0.04163)
Age	0.00207	0.00178	−0.00028	0.00092
	(0.00061)	(0.00057)	(0.00100)	(0.00056)
Education	0.04792	0.04861	0.03731	0.04647
	(0.00247)	(0.00262)	(0.00518)	(0.00262)
Female	−0.00658	0.00070	−0.02205	−0.01517
	(0.01373)	(0.01384)	(0.01445)	(0.01357)

Table 13.2 presents four sets of estimates, nonlinear least squares, method of moments, first-step GMM, and GMM using the optimal weighting matrix. Two comparisons are noted. The method of moments produces slightly different results from the nonlinear least squares estimator. This is to be expected because they are different criteria. Judging by the standard errors, the GMM estimator seems to provide a very slight improvement over the nonlinear least squares and method of moments estimators. The conclusion, though, would seem to be that the two additional moments (variables) do not provide very much additional information for estimation of the parameters.

13.4.3 PROPERTIES OF THE GMM ESTIMATOR

We will now examine the properties of the GMM estimator in some detail. Because the GMM estimator includes other familiar estimators that we have already encountered, including least squares (linear and nonlinear) and instrumental variables, these results will extend to those cases. The discussion given here will only sketch the elements of the formal proofs. The assumptions we make here are somewhat narrower than a fully general treatment might allow, but they are broad enough to include the situations likely to arise in practice. More detailed and rigorous treatments may be found in, for example, Newey and McFadden (1994), White (2001), Hayashi (2000), Mittelhammer et al. (2000), or Davidson (2000).

The GMM estimator is based on the set of population orthogonality conditions,

$$E[\mathbf{m}_i(\boldsymbol{\theta}_0)] = \mathbf{0},$$

where we denote the true parameter vector by $\boldsymbol{\theta}_0$. The subscript i on the term on the left-hand side indicates dependence on the observed data, $(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i)$. Averaging this over the sample observations produces the sample moment equation

$$E[\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)] = \mathbf{0},$$

where

$$\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\boldsymbol{\theta}_0).$$

This moment is a set of L equations involving the K parameters. We will assume that this expectation exists and that the sample counterpart converges to it. The definitions are cast in terms of the population parameters and are indexed by the sample size. To fix the ideas, consider, once again, the empirical moment equations that define the instrumental variable estimator for a linear or nonlinear regression model.

Example 13.8 Empirical Moment Equation for Instrumental Variables

For the IV estimator in the linear or nonlinear regression model, we assume

$$E\left[\overline{\mathbf{m}}_n(\boldsymbol{\beta})\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i[y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]\right] = \mathbf{0}.$$

There are L instrumental variables in $\mathbf{z}_i$ and K parameters in $\boldsymbol{\beta}$. This statement defines L moment equations, one for each instrumental variable.

We make the following assumptions about the model and these empirical moments:

ASSUMPTION 13.1 Convergence of the Empirical Moments
The data-generating process is assumed to meet the conditions for a law of large numbers to apply, so that we may assume that the empirical moments converge in probability to their expectation. Appendix D lists several different laws of large numbers that increase in generality. What is required for this assumption is that

$$\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\boldsymbol{\theta}_0) \xrightarrow{\ p\ } \mathbf{0}.$$

The laws of large numbers that we examined in Appendix D accommodate cases of independent observations. Cases of dependent or correlated observations can be gathered under the **Ergodic theorem** (20.1). For this more general case, then, we would assume that the sequence of observations $\mathbf{m}(\boldsymbol{\theta})$ constitutes a jointly $(L \times 1)$ stationary and ergodic process.

The empirical moments are assumed to be continuous and continuously differentiable functions of the parameters. For our earlier example, this would mean that the conditional mean function, $h(\mathbf{x}_i, \boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ (although not necessarily of $\mathbf{x}_i$). With continuity and differentiability, we will also be able to assume that the derivatives of the moments,

$$\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \frac{\partial \overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{m}_{i,n}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'},$$

converge to a probability limit, say, plim $\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \overline{\mathbf{G}}(\boldsymbol{\theta}_0)$. [See (13-1), (13-5), and Theorem 13.1.] For sets of *independent* observations, the continuity of the functions and the derivatives will allow us to invoke the Slutsky theorem to obtain this result. For the more general case of sequences of *dependent* observations, Theorem 20.2, Ergodicity of Functions, will provide a counterpart to the Slutsky theorem for time-series data. In sum, if the moments themselves obey a law of large numbers, then it is reasonable to assume that the derivatives do as well.

ASSUMPTION 13.2 Identification
For any $n \geq K$, if $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two different parameter vectors, then there exist data sets such that $\overline{\mathbf{m}}_n(\boldsymbol{\theta}_1) \neq \overline{\mathbf{m}}_n(\boldsymbol{\theta}_2)$. Formally, in Section 12.5.3, identification is defined to imply that the probability limit of the GMM criterion function is uniquely minimized at the true parameters, θ_0.

Assumption 13.2 is a practical prescription for identification. More formal conditions are discussed in Section 12.5.3. We have examined two violations of this crucial assumption. In the linear regression model, one of the assumptions is full rank of the matrix of exogenous variables—the absence of multicollinearity in $\mathbf{X}$. In our discussion of the maximum likelihood estimator, we will encounter a case (Example 14.1) in which a normalization is needed to identify the vector of parameters.[12] Both of these cases are included in this assumption. The identification condition has three important implications:

1. **Order condition.** The number of moment conditions is at least as large as the number of parameters, $L \geq K$. This is necessary, but not sufficient, for identification.
2. **Rank condition.** The $L \times K$ matrix of derivatives, $\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0)$, will have row rank equal to K. (Again, note that the number of rows must equal or exceed the number of columns.)
3. **Uniqueness.** With the continuity assumption, the identification assumption implies that the parameter vector that satisfies the population moment condition is unique. We know that at the true parameter vector, plim $\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \mathbf{0}$. If $\boldsymbol{\theta}_1$ is any parameter vector that satisfies this condition, then $\boldsymbol{\theta}_1$ must equal $\boldsymbol{\theta}_0$.

Assumptions 13.1 and 13.2 characterize the parameterization of the model. Together they establish that the parameter vector will be estimable. We now make the statistical assumption that will allow us to establish the properties of the GMM estimator.

ASSUMPTION 13.3 Asymptotic Distribution of Empirical Moments
We assume that the empirical moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix, $(1/n)\boldsymbol{\Phi}$, so that $\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}]$.

The underlying requirements on the data for this assumption to hold will vary and will be complicated if the observations comprising the empirical moment are not independent. For samples of independent observations, we assume the conditions underlying the Lindeberg–Feller (D.19) or Liapounov central limit theorem (D.20) will suffice. For the more general case, it is once again necessary to make some assumptions about the data. We have assumed that $E[\mathbf{m}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$. If we can go a step further and assume that the functions $\mathbf{m}_i(\boldsymbol{\theta}_0)$ are an ergodic, stationary **martingale difference sequence**, $E[\mathbf{m}_i(\boldsymbol{\theta}_0) \,|\, \mathbf{m}_{i-1}(\boldsymbol{\theta}_0), \mathbf{m}_{i-2}(\boldsymbol{\theta}_0) \ldots] = \mathbf{0}$, then we can invoke Theorem 20.3, the central limit theorem for the martingale difference series. It will generally be fairly complicated to verify this assumption for nonlinear models, so it will usually be assumed outright. On the other hand, the assumptions are likely to be fairly benign in a typical application. For regression models, the assumption takes the form

$$E[\mathbf{z}_i\varepsilon_i \,|\, \mathbf{z}_{i-1}\varepsilon_{i-1}, \ \ldots] = \mathbf{0},$$

which will often be part of the central structure of the model.

[12] See Hansen et al. (1996) for discussion of this case.

With the assumptions in place, we have

THEOREM 13.2 Asymptotic Distribution of the GMM Estimator
Under the preceding assumptions,

$$\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{p} \boldsymbol{\theta}_0,$$

$$\hat{\boldsymbol{\theta}}_{GMM} \overset{a}{} N[\boldsymbol{\theta}_0, \mathbf{V}_{GMM}], \tag{13-6}$$

where $\mathbf{V}_{GMM}$ *is defined in* (*13-5*).

We will now sketch a proof of Theorem 13.2. The GMM estimator is obtained by minimizing the criterion function,

$$q_n(\boldsymbol{\theta}) = \overline{\mathbf{m}}_n(\boldsymbol{\theta})'\mathbf{W}_n\overline{\mathbf{m}}_n(\boldsymbol{\theta}),$$

where $\mathbf{W}_n$ is the weighting matrix used. Consistency of the estimator that minimizes this criterion can be established by the same logic that will be used for the maximum likelihood estimator. It must first be established that $q_n(\boldsymbol{\theta})$ converges to a value $q_0(\boldsymbol{\theta})$. By our assumptions of strict continuity and Assumption 13.1, $q_n(\boldsymbol{\theta}_0)$ converges to 0. (We could apply the Slutsky theorem to obtain this result.) We will assume that $q_n(\boldsymbol{\theta})$ converges to $q_0(\boldsymbol{\theta})$ for other points in the parameter space as well. Because $\mathbf{W}_n$ is positive definite, for any finite n, we know that

$$0 \leq q_n(\hat{\boldsymbol{\theta}}_{GMM}) \leq q_n(\boldsymbol{\theta}_0). \tag{13-7}$$

That is, in the finite sample, $\hat{\boldsymbol{\theta}}_{GMM}$ actually minimizes the function, so the sample value of the criterion is not larger at $\hat{\boldsymbol{\theta}}_{GMM}$ than at any other value, including the true parameters. But, at the true parameter values, $q_n(\boldsymbol{\theta}_0) \xrightarrow{p} 0$. So, if (13-7) is true, then it must follow that $q_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{p} 0$ as well because of the identification assumption, 13.2. As $n \to \infty$, $q_n(\hat{\boldsymbol{\theta}}_{GMM})$ and $q_n(\boldsymbol{\theta})$ converge to the same limit. It must be the case, then, that as $n \to \infty$, $\overline{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) \to \overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)$, because the function is quadratic and $\mathbf{W}$ is positive definite. The identification condition that we assumed earlier now assures that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_{GMM}$ must equal $\boldsymbol{\theta}_0$. This establishes consistency of the estimator.

We will now sketch a proof of the asymptotic normality of the estimator. The first-order conditions for the GMM estimator are

$$\frac{\partial q_n(\hat{\boldsymbol{\theta}}_{GMM})}{\partial \hat{\boldsymbol{\theta}}_{GMM}} = 2\overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\overline{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \mathbf{0}. \tag{13-8}$$

(The leading 2 is irrelevant to the solution, so it will be dropped at this point.) The orthogonality equations are assumed to be continuous and continuously differentiable. This allows us to employ the **mean value theorem** as we expand the empirical moments in a linear Taylor series around the true value, $\boldsymbol{\theta}_0$,

$$\overline{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) + \overline{\mathbf{G}}_n(\overline{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0), \tag{13-9}$$

where $\overline{\boldsymbol{\theta}}$ is a point between $\hat{\boldsymbol{\theta}}_{GMM}$ and the true parameters, $\boldsymbol{\theta}_0$. Thus, for each element $\overline{\theta}_k = w_k\hat{\theta}_{k,GMM} + (1 - w_k)\theta_{0,k}$ for some w_k such that $0 < w_k < 1$. Insert (13-9) in (13-8) to obtain

$$\overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) + \overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\overline{\mathbf{G}}_n(\overline{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) = \mathbf{0}.$$

Solve this equation for the estimation error and multiply by $\sqrt{n}$. This produces

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) = -[\overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\overline{\mathbf{G}}_n(\overline{\boldsymbol{\theta}})]^{-1}\overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0).$$

Assuming that they have them, the quantities on the left- and right-hand sides have the same limiting distributions. By the consistency of $\hat{\boldsymbol{\theta}}_{GMM}$, we know that $\hat{\boldsymbol{\theta}}_{GMM}$ and $\overline{\boldsymbol{\theta}}$ both converge to $\boldsymbol{\theta}_0$. By the strict continuity assumed, it must also be the case that

$$\overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}) \xrightarrow{\ p\ } \overline{\mathbf{G}}(\boldsymbol{\theta}_0) \text{ and } \overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{\ p\ } \overline{\mathbf{G}}(\boldsymbol{\theta}_0).$$

We have also assumed that the weighting matrix, $\mathbf{W}_n$, converges to a matrix of constants, $\mathbf{W}$. Collecting terms, we find that the limiting distribution of the vector on the left-hand side must be the same as that on the right-hand side in (13-10),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \xrightarrow{\ d\ } \{-[\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\}\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0). \tag{13-10}$$

We now invoke Assumption 13.3. The matrix in curled brackets is a set of constants. The last term has the normal limiting distribution given in Assumption 13.3. The mean and variance of this limiting distribution are zero and $\boldsymbol{\Phi}$, respectively. Collecting terms, we have the result in Theorem 13.2, where

$$\mathbf{V}_{GMM} = \frac{1}{n}[\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)[\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}. \tag{13-11}$$

The final result is a function of the choice of weighting matrix, $\mathbf{W}$. If the optimal weighting matrix, $\mathbf{W} = \boldsymbol{\Phi}^{-1}$, is used, then the expression collapses to

$$\mathbf{V}_{GMM,optimal} = \frac{1}{n}[\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\boldsymbol{\Phi}^{-1}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}. \tag{13-12}$$

Returning to (13-11), there is a special case of interest. If we use least squares or instrumental variables with $\mathbf{W} = \mathbf{I}$, then

$$\mathbf{V}_{GMM} = \frac{1}{n}(\overline{\mathbf{G}}'\mathbf{G})^{-1}\overline{\mathbf{G}}'\boldsymbol{\Phi}\overline{\mathbf{G}}(\overline{\mathbf{G}}'\overline{\mathbf{G}})^{-1}.$$

This equation prescribes essentially the White or **Newey–West estimator**, which returns us to our departure point and provides a neat symmetry to the GMM principle. We will formalize this in Section 13.6.1.

13.5 TESTING HYPOTHESES IN THE GMM FRAMEWORK

The estimation framework developed in the previous section provides the basis for a convenient set of statistics for testing hypotheses. We will consider three groups of tests. The first is a pair of statistics that is used for testing the validity of the restrictions that produce the moment equations. The second is a trio of tests that correspond to the familiar Wald, LM, and LR tests. The third is a class of tests based on the theoretical underpinnings of the conditional moments that we used earlier to devise the GMM estimator.

13.5.1 TESTING THE VALIDITY OF THE MOMENT RESTRICTIONS

In the exactly identified cases we examined earlier (least squares, instrumental variables, maximum likelihood), the criterion for GMM estimation,

$$q = \overline{\mathbf{m}}(\boldsymbol{\theta})'\mathbf{W}\overline{\mathbf{m}}(\boldsymbol{\theta}),$$

would be exactly zero because we can find a set of estimates for which $\bar{\mathbf{m}}(\boldsymbol{\theta})$ is exactly zero. Thus, in the exactly identified case when there are the same number of moment equations as there are parameters to estimate, the weighting matrix $\mathbf{W}$ is irrelevant to the solution. But if the parameters are overidentified by the moment equations, then these equations imply substantive restrictions. As such, if the hypothesis of the model that led to the moment equations in the first place is incorrect, at least some of the sample moment restrictions will be systematically violated. This conclusion provides the basis for a test of the **overidentifying restrictions**. By construction, when the optimal weighting matrix is used,

$$nq = [\sqrt{n}\,\bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})']\{\text{Est.Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})]\}^{-1}[\sqrt{n}\,\bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})],$$

so nq is a Wald statistic. Therefore, under the hypothesis of the model,

$$nq \xrightarrow{d} \chi^2[L - K].$$

(For the exactly identified case, there are zero degrees of freedom and $q = 0$.)

Example 13.9 Overidentifying Restrictions

In Hall's consumption model, two orthogonality conditions noted in Example 13.1 exactly identify the two parameters. But his analysis of the model suggests a way to test the specification. The conclusion, "No information available in time t apart from the level of consumption, c_t, helps predict future consumption, c_{t+1}, in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods t or earlier are irrelevant once c_t is known," suggests how one might test the model. If lagged values of income (Y_t might equal the ratio of current income to the previous period's income) are added to the set of instruments, then the model is now overidentified by the orthogonality conditions,

$$E_t\left[(\beta(1 + r_{t+1})R_{t+1}^{\lambda} - 1) \times \begin{pmatrix} 1 \\ R_t \\ Y_{t-1} \\ Y_{t-2} \end{pmatrix}\right] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

A simple test of the overidentifying restrictions would be suggestive of the validity of the corollary. Rejecting the restrictions casts doubt on the original model. Hall's proposed tests to distinguish the life cycle–permanent income model from other theories of consumption involved adding two lags of income to the information set. Hansen and Singleton (1982) operated directly on this form of the model. Other studies, for example, Campbell and Mankiw's (1989) as well as Hall's, used the model's implications to formulate more conventional instrumental variable regression models.

The preceding is a **specification test**, not a test of parametric restrictions. However, there is a symmetry between the moment restrictions and restrictions on the parameter vector. Suppose $\boldsymbol{\theta}$ is subjected to J restrictions (linear or nonlinear) that restrict the number of free parameters from K to $K - J$. (That is, reduce the dimensionality of the parameter space from K to $K - J$.) The nature of the GMM estimation problem we have posed is not changed at all by the restrictions. The constrained problem may be stated in terms of

$$q_R = \bar{\mathbf{m}}(\boldsymbol{\theta}_R)'\mathbf{W}\bar{\mathbf{m}}(\boldsymbol{\theta}_R).$$

Note that the weighting matrix, $\mathbf{W}$, is unchanged. The precise nature of the solution method may be changed—the restrictions mandate a constrained optimization. However, the criterion is essentially unchanged. It follows then that

$$nq_R \xrightarrow{d} \chi^2[L - (K - J)].$$

This result suggests a method of testing the restrictions, although the distribution theory is not obvious. The weighted sum of squares with the restrictions imposed, nq_R, must be larger than the weighted sum of squares obtained without the restrictions, nq. The difference is

$$(nq_R - nq) \xrightarrow{d} \chi^2[J]. \tag{13-13}$$

The test is attributed to Newey and West (1987b). This provides one method of testing a set of restrictions. (The small-sample properties of this test will be the central focus of the application discussed in Section 13.6.5.) We now consider several alternatives.

13.5.2 GMM WALD COUNTERPARTS TO THE WALD, LM, AND LR TESTS

Section 14.6 describes a trio of testing procedures that can be applied to a hypothesis in the context of maximum likelihood estimation. To reiterate, let the hypothesis to be tested be a set of J possibly nonlinear restrictions on K parameters $\boldsymbol{\theta}$ in the form H_0: $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$. Let $\mathbf{c}_1$ be the maximum likelihood estimates of $\boldsymbol{\theta}$ estimated without the restrictions, and let $\mathbf{c}_0$ denote the restricted maximum likelihood estimates, that is, the estimates obtained while imposing the null hypothesis. The three statistics, which are asymptotically equivalent, are obtained as follows:

$$\text{LR} = \text{likelihood ratio} = -2(\ln L_0 - \ln L_1),$$

where

$$\ln L_j = \text{log-likelihood function evaluated at } \mathbf{c}_j, \quad j = 0, 1.$$

The **likelihood ratio statistic** requires that both estimates be computed. The Wald statistic is

$$W = \text{Wald} = [\mathbf{r}(\mathbf{c}_1)]'\{\text{Est.Asy.Var}[\mathbf{r}(\mathbf{c}_1)]\}^{-1}[\mathbf{r}(\mathbf{c}_1)]. \tag{13-14}$$

The **Wald statistic** is the distance measure for the degree to which the unrestricted estimator fails to satisfy the restrictions. The usual estimator for the asymptotic covariance matrix would be

$$\text{Est.Asy.Var}[\mathbf{r}(\mathbf{c}_1)] = \mathbf{R}_1\{\text{Est.Asy.Var}[\mathbf{c}_1]\}\mathbf{R}_1', \tag{13-15}$$

where

$$\mathbf{R}_1 = \partial \mathbf{r}(\mathbf{c}_1)/\partial \mathbf{c}_1' \quad (\mathbf{R}_1 \text{ is a } J \times K \text{ matrix}).$$

The Wald statistic can be computed using only the unrestricted estimate. The LM statistic is

$$\text{LM} = \text{Lagrange multiplier} = \mathbf{g}_1'(\mathbf{c}_0)\{\text{Est.Asy.Var}[\mathbf{g}_1(\mathbf{c}_0)]\}^{-1}\mathbf{g}_1(\mathbf{c}_0), \tag{13-16}$$

where

$$\mathbf{g}_1(\mathbf{c}_0) = \partial \ln L_1(\mathbf{c}_0)/\partial \mathbf{c}_0,$$

that is, the first derivatives of the *unconstrained* log-likelihood computed at the *restricted* estimates. The term $\text{Est.Asy.Var}[\mathbf{g}_1(\mathbf{c}_0)]$ is the inverse of any of the usual estimators of the asymptotic covariance matrix of the maximum likelihood estimators of the parameters, computed using the restricted estimates. The most convenient choice is usually the BHHH estimator. The LM statistic is based on the restricted estimates.

Newey and West (1987b) have devised counterparts to these test statistics for the GMM estimator. The Wald statistic is computed identically, using the results of GMM estimation rather than maximum likelihood.[13] That is, in (13-14), we would use the unrestricted GMM estimator of $\boldsymbol{\theta}$. The appropriate asymptotic covariance matrix is (13-12). The computation is exactly the same. The counterpart to the LR statistic is the difference in the values of nq in (13-13). It is necessary to use the same weighting matrix, $\mathbf{W}$, in both restricted and unrestricted estimators. Because the unrestricted estimator is consistent under both H_0 and H_1, a consistent, unrestricted estimator of $\boldsymbol{\theta}$ is used to compute $\mathbf{W}$. Label this $\boldsymbol{\Phi}_1^{-1} = \{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}_1(\mathbf{c}_1)]\}^{-1}$. In each occurrence, the subscript 1 indicates reference to the unrestricted estimator. Then q is minimized without restrictions to obtain q_1 and then subject to the restrictions to obtain q_0. The statistic is then $(nq_0 - nq_1)$.[14] Because we are using the same $\mathbf{W}$ in both cases, this statistic is necessarily nonnegative. (This is the statistic discussed in Section 13.5.1.)

Finally, the counterpart to the LM statistic would be

$$\text{LM}_{GMM} = n[\overline{\mathbf{m}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\hat{\mathbf{G}}_1(\mathbf{c}_0)][\mathbf{G}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0)]^{-1}[\overline{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{m}}_1(\mathbf{c}_0)].$$

The logic for this LM statistic is the same as that for the MLE. The derivatives of the minimized criterion q in (13-3) evaluated at the restricted estimator are

$$\mathbf{g}_1(\mathbf{c}_0) = \frac{\partial q}{\partial \mathbf{c}_0} = 2\overline{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{m}}(\mathbf{c}_0).$$

The **LM statistic**, LM_{GMM}, is a Wald statistic for testing the hypothesis that this vector equals zero under the restrictions of the null hypothesis. From our earlier results, we would have

$$\text{Est.Asy.Var}[\mathbf{g}_1(\mathbf{c}_0)] = \frac{4}{n}\overline{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\{\text{Est.Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\mathbf{c}_0)]\}\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0).$$

The estimated asymptotic variance of $\sqrt{n}\,\overline{\mathbf{m}}(\mathbf{c}_0)$ is $\hat{\boldsymbol{\Phi}}_1$, so

$$\text{Est.Asy.Var}[\mathbf{g}_1(\mathbf{c}_0)] = \frac{4}{n}\overline{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0).$$

The Wald statistic would be

$$\begin{aligned}\text{Wald} &= \mathbf{g}_1(\mathbf{c}_0)'\{\text{Est.Asy.Var}[\mathbf{g}_1(\mathbf{c}_0)]\}^{-1}\mathbf{g}_1(\mathbf{c}_0)\\ &= n\overline{\mathbf{m}}_1'(\mathbf{c}_0)\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0)\{\overline{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0)\}^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{m}}_1(\mathbf{c}_0). \quad \textbf{(13-17)}\end{aligned}$$

13.6 GMM ESTIMATION OF ECONOMETRIC MODELS

The preceding has suggested that the GMM approach to estimation broadly encompasses most of the estimators we will encounter in this book. We have implicitly examined least squares and the general method of instrumental variables in the process. In this section,

[13] See Burnside and Eichenbaum (1996) for some small-sample results on this procedure. Newey and McFadden (1994) have shown the asymptotic equivalence of the three procedures.

[14] Newey and West label this test the *D* test.

we will formalize more specifically the GMM estimators for several of the estimators that appear in the earlier chapters. Section 13.6.1 examines the generalized regression model of Chapter 9. Section 13.6.2 describes a relatively minor extension of the GMM/IV estimator to nonlinear regressions. Section 13.6.3 describes the GMM estimators for our models of systems of seemingly unrelated regression (SUR) model. Finally, in Section 13.6.4, we develop one of the major applications of GMM estimation, the Arellano–Bond–Bover estimator for dynamic panel data models.

13.6.1 SINGLE-EQUATION LINEAR MODELS

It is useful to confine attention to the instrumental variables case, as it is fairly general and we can easily specialize it to the simpler regression models if that is appropriate. Thus, we depart from the usual linear model (8-1), but we no longer require that $E[\varepsilon_i|\mathbf{x}_i] = 0$. Instead, we adopt the instrumental variables formulation in Chapter 8. That is, the model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$
$$E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$$

for K variables in $\mathbf{x}_i$ and for some set of L instrumental variables, $\mathbf{z}_i$, where $L \geq K$. The earlier case of the generalized regression model arises if $\mathbf{z}_i = \mathbf{x}_i$, and the classical regression results if we add $\boldsymbol{\Omega} = \mathbf{I}$ as well, so this is a convenient encompassing model framework.

In Chapter 9 on generalized least squares estimation, we considered two cases, first one with a known $\boldsymbol{\Omega}$, then one with an unknown $\boldsymbol{\Omega}$ that must be estimated. In estimation by the generalized method of moments, neither of these approaches is relevant because we begin with much less (assumed) knowledge about the data-generating process. We will consider three cases:

- Classical regression: $\text{Var}[\varepsilon_i|\mathbf{X}, \mathbf{Z}] = \sigma^2$,
- Heteroscedasticity: $\text{Var}[\varepsilon_i|\mathbf{X}, \mathbf{Z}] = \sigma_i^2$,
- Generalized model: $\text{Cov}[\varepsilon_t, \varepsilon_s|\mathbf{X}, \mathbf{Z}] = \sigma^2\omega_{ts}$,

where $\mathbf{Z}$ and $\mathbf{X}$ are the $n \times L$ and $n \times K$ observed data matrices, respectively. (We assume, as will often be true, that the fully general case will apply in a time-series setting. Hence the change in the subscripts.) No specific distribution is assumed for the disturbances, conditional or unconditional.

The assumption $E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$ implies the following orthogonality condition,

$$\text{Cov}[\mathbf{z}_i, \varepsilon_i] = \mathbf{0}, \quad \text{or} \quad E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

By summing the terms, we find that this further implies the **population moment equation**,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right] = E[\overline{\mathbf{m}}(\boldsymbol{\beta})] = \mathbf{0}. \tag{13-18}$$

This relationship suggests how we might now proceed to estimate $\boldsymbol{\beta}$. Note, in fact, that if $\mathbf{z}_i = \mathbf{x}_i$, then this is just the population counterpart to the least squares normal equations. So, as a guide to estimation, this would return us to least squares. Suppose we now translate this population expectation into a sample analog and use that as our guide for estimation. That is, if the population relationship holds for the true parameter vector, $\boldsymbol{\beta}$,

suppose we attempt to mimic this result with a sample counterpart, or empirical moment equation,

$$\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\right] = \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\hat{\boldsymbol{\beta}})\right] = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{13-19}$$

In the absence of other information about the data-generating process, we can use the empirical moment equation as the basis of our estimation strategy.

The empirical moment condition is L equations (the number of variables in $\mathbf{Z}$) in K unknowns (the number of parameters we seek to estimate). There are three possibilities to consider:

1. **Underidentified.** $L < K$. If there are fewer moment equations than there are parameters, then it will not be possible to find a solution to the equation system in (13-19). With no other information, such as restrictions that would reduce the number of free parameters, there is no need to proceed any further with this case.

   For the identified cases, it is convenient to write (13-19) as

   $$\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n}\mathbf{Z}'\mathbf{y}\right) - \left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}. \tag{13-20}$$

2. **Exactly identified.** If $L = K$, then you can easily show (we leave it as an exercise) that the single solution to our equation system is the familiar instrumental variables estimator from Section 8.3.2,

   $$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{13-21}$$

3. **Overidentified.** If $L > K$, then there is no unique solution to the equation system $\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. In this instance, we need to formulate some strategy to choose an estimator. One intuitively appealing possibility which has served well thus far is least squares. In this instance, that would mean choosing the estimator based on the criterion function,

   $$\text{Min}_{\boldsymbol{\beta}}\, q = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}})'\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}).$$

   We do keep in mind that we will only be able to minimize this at some positive value; there is no exact solution to (13-19) in the overidentified case. Also, you can verify that if we treat the exactly identified case as if it were overidentified, that is, use least squares anyway, we will still obtain the IV estimator shown in (13-21) for the solution to case (2). For the overidentified case, the first-order conditions are

   $$\begin{aligned}\frac{\partial q}{\partial \hat{\boldsymbol{\beta}}} &= 2\left(\frac{\partial \overline{\mathbf{m}}'(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = 2\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) \\ &= 2\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{0}.\end{aligned} \tag{13-22}$$

   We leave as exercise to show that the solution in both cases (2) and (3) is now

   $$\hat{\boldsymbol{\beta}} = [(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{y}). \tag{13-23}$$

   The estimator in (13-23) is a hybrid that we have not encountered before, though if $L = K$, then it does reduce to the earlier one in (13-21). (In the overidentified case, (13-23) is not an IV estimator, it is, as we have sought, a **method of moments estimator**.)

It remains to establish consistency and to obtain the asymptotic distribution and an asymptotic covariance matrix for the estimator. The intermediate results we need are Assumptions 13.1, 13.2, and 13.3 in Section 13.4.3:

- **Convergence of the moments.** The sample moment converges in probability to its population counterpart. That is, $\overline{\mathbf{m}}(\boldsymbol{\beta}) \rightarrow \mathbf{0}$. Different circumstances will produce different kinds of convergence, but we will require it in some form. For the simplest cases, such as a model of heteroscedasticity, this will be convergence in mean square. Certain time-series models that involve correlated observations will necessitate some other form of convergence. But, in any of the cases we consider, we will require the general result: plim $\overline{\mathbf{m}}(\boldsymbol{\beta}) = \mathbf{0}$.
- **Identification.** The parameters are identified in terms of the moment equations. Identification means, essentially, that a large enough sample will contain sufficient information for us actually to estimate $\boldsymbol{\beta}$ consistently using the sample moments. There are two conditions which must be met—an **order condition**, which we have already assumed ($L \geq K$), and a **rank condition**, which states that the moment equations are not redundant. The rank condition implies the order condition, so we need only formalize it:
- **Identification condition for GMM estimation.** The $L \times K$ matrix,

$$\boldsymbol{\Gamma}(\boldsymbol{\beta}) = E[\overline{\mathbf{G}}(\boldsymbol{\beta})] = \text{plim } \overline{\mathbf{G}}(\boldsymbol{\beta}) = \text{plim } \frac{\partial \overline{\mathbf{m}}}{\partial \boldsymbol{\beta}'} = \text{plim } \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{m}_i}{\partial \boldsymbol{\beta}'},$$

  must have row rank equal to K.[15] Because this requires $L \geq K$, this implies the order condition. This assumption means that this derivative matrix converges in probability to its expectation. Note that we have assumed, in addition, that the derivatives, like the moments themselves, obey a law of large numbers—they converge in probability to their expectations.
- **Limiting Normal Distribution for the Sample Moments.** The population moment obeys a central limit theorem. Because we are studying a generalized regression model, Lindeberg–Levy (D.18) will be too narrow—the observations will have different variances. Lindeberg–Feller (D.19.A) suffices in the heteroscedasticity case, but in the general case, we will ultimately require something more general. See Section 13.4.3.

It will follow from Assumptions 13.1–13.3 (again, at this point we do this without proof) that the GMM estimators that we obtain are, in fact, consistent. By virtue of the Slutsky theorem, we can transfer our limiting results to the empirical moment equations.

To obtain the asymptotic covariance matrix we will simply invoke the general result for GMM estimators in Section 13.4.3. That is,

$$\text{Asy.Var}[\hat{\boldsymbol{\beta}}] = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1}\boldsymbol{\Gamma}'\{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})]\}\boldsymbol{\Gamma}[\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1}.$$

For the particular model we are studying here,

$$\begin{aligned}
\overline{\mathbf{m}}(\boldsymbol{\beta}) &= (1/n)(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\boldsymbol{\beta}),\\
\overline{\mathbf{G}}(\boldsymbol{\beta}) &= (1/n)\mathbf{Z}'\mathbf{X},\\
\boldsymbol{\Gamma}(\boldsymbol{\beta}) &= \mathbf{Q}_{\mathbf{ZX}} \text{ (see Section 8.3.2)}
\end{aligned}$$

[15] We require that the row rank be at least as large as K. There could be redundant, that is, functionally dependent, moments, so long as there are at least K that are functionally independent.

(You should check in the preceding expression that the dimensions of the particular matrices and the dimensions of the various products produce the correctly configured matrix that we seek.) The remaining detail, which is the crucial one for the model we are examining, is for us to determine,

$$\mathbf{V} = \text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})].$$

Given the form of $\overline{\mathbf{m}}(\boldsymbol{\beta})$,

$$\mathbf{V} = \frac{1}{n}\text{Var}\left[\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma^2\omega_{ij}\mathbf{z}_i\mathbf{z}_j' = \sigma^2\frac{\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}}{n}$$

for the most general case. Note that this is precisely the expression that appears in (9-9), so the question that arose there arises here once again. That is, under what conditions will this converge to a constant matrix? We take the discussion there as given. The only remaining detail is how to estimate this matrix. The answer appears in Section 9.2, where we pursued this same question in connection with robust estimation of the asymptotic covariance matrix of the least squares estimator. To review then, what we have achieved to this point is to provide a theoretical foundation for the instrumental variables estimator. As noted earlier, this specializes to the least squares estimator. The estimators of $\mathbf{V}$ for our three cases will be

- Classical regression:

$$\hat{\mathbf{V}} = \frac{(\mathbf{e}'\mathbf{e}/n)}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i' = \frac{(\mathbf{e}'\mathbf{e}/n)}{n}\mathbf{Z}'\mathbf{Z}.$$

- Heteroscedastic regression:

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}e_i^2\mathbf{z}_i\mathbf{z}_i'. \tag{13-24}$$

- Generalized regression:

$$\hat{\mathbf{V}} = \frac{1}{n}\left[\sum_{t=1}^{n}e_t^2\mathbf{z}_t\mathbf{z}_t' + \sum_{\ell=1}^{p}\left(1 - \frac{\ell}{(p+1)}\right)\sum_{t=\ell+1}^{n}e_te_{t-\ell}(\mathbf{z}_t\mathbf{z}_{t-\ell}' + \mathbf{z}_{t-\ell}\mathbf{z}_t')\right].$$

We should observe that in each of these cases, we have actually used some information about the structure of $\boldsymbol{\Omega}$. If it is known only that the terms in $\overline{\mathbf{m}}(\boldsymbol{\beta})$ are uncorrelated, then there is a convenient estimator available, $\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\hat{\boldsymbol{\beta}})\mathbf{m}_i(\hat{\boldsymbol{\beta}})'$, that is, the natural, empirical variance estimator. Note that this is what is being used in the heteroscedasticity case in (13-24).

Collecting all the terms so far, then, we have

$$\begin{aligned}\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] &= \frac{1}{n}[\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})]^{-1}\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})[\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})]^{-1}\\ &= n[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}(\mathbf{Z}'\mathbf{X})[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}.\end{aligned} \tag{13-25}$$

The preceding might seem to endow the least squares or method of moments estimators with some degree of optimality, but that is not the case. We have only provided them with a different statistical motivation (and established consistency). We now consider the question of whether, because this is the generalized regression model, there is some better (more efficient) means of using the data.

The class of minimum distance estimators for this model is defined by the solutions to the criterion function, $\text{Min}_{\boldsymbol{\beta}}\, q = \overline{\mathbf{m}}(\boldsymbol{\beta})'\mathbf{W}\overline{\mathbf{m}}(\boldsymbol{\beta})$, where $\mathbf{W}$ is *any* positive definite **weighting matrix**. Based on the assumptions just made, we can invoke Theorem 13.1 to obtain

$$\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{MD}] = \frac{1}{n}[\overline{\mathbf{G}}'\mathbf{W}\overline{\mathbf{G}}]^{-1}\overline{\mathbf{G}}'\mathbf{W}\mathbf{V}\mathbf{W}\overline{\mathbf{G}}[\overline{\mathbf{G}}'\mathbf{W}\overline{\mathbf{G}}]^{-1}.$$

Note that our entire preceding analysis was of the simplest minimum distance estimator, which has $\mathbf{W} = \mathbf{I}$. The obvious question now arises, if any $\mathbf{W}$ produces a consistent estimator, is any $\mathbf{W}$ better than any other one, or is it simply arbitrary? There is a firm answer, for which we have to consider two cases separately:

- **Exactly identified case.** If $L = K$; that is, if the number of moment conditions is the same as the number of parameters being estimated, then $\mathbf{W}$ is irrelevant to the solution, so on the basis of simplicity alone, the optimal $\mathbf{W}$ is $\mathbf{I}$.
- **Overidentified case.** In this case, the "optimal" weighting matrix, that is, the $\mathbf{W}$ that produces the most efficient estimator, is $\mathbf{W} = \mathbf{V}^{-1}$. The best weighting matrix is the inverse of the asymptotic covariance of the moment vector. In this case, the MDE will be the GMM estimator with

$$\hat{\boldsymbol{\beta}}_{GMM} = [(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{y}),$$

and

$$\begin{aligned}\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{GMM}] &= \frac{1}{n}[\overline{\mathbf{G}}'\mathbf{V}^{-1}\overline{\mathbf{G}}]^{-1} \\ &= n[(\mathbf{X}'\mathbf{Z})\mathbf{V}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}.\end{aligned}$$

We conclude this discussion by tying together what should seem to be a loose end. The GMM estimator is computed as the solution to

$$\text{Min}_{\boldsymbol{\beta}}\, q = \overline{\mathbf{m}}(\boldsymbol{\beta})'\{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})\}^{-1}\overline{\mathbf{m}}(\boldsymbol{\beta}),$$

which might suggest that the weighting matrix is a function of the thing we are trying to estimate. The process of GMM estimation will have to proceed in two steps: Step 1 is to obtain an estimate of $\mathbf{V}$; Step 2 will consist of using the inverse of this $\mathbf{V}$ as the weighting matrix in computing the GMM estimator. The following is a common two-step strategy:

Step 1. Use $\mathbf{W} = \mathbf{I}$ to obtain a consistent estimator of $\boldsymbol{\beta}$. Then, in the heteroscedasticity case (i.e., the White estimator), estimate $\mathbf{V}$ with $\hat{\mathbf{V}} = (1/n)\sum_{i=1}^{n} e_i^2\mathbf{z}_i\mathbf{z}_i'$. For the more general case, use the Newey–West estimator.

Step 2. Use $\mathbf{W} = \hat{\mathbf{V}}^{-1}$ to compute the GMM estimator.

By this point, the observant reader should have noticed that in all of the preceding, we have never actually encountered the two-stage least squares estimator that we introduced in Section 8.4.1. To obtain this estimator, we must revert back to the classical, that is, homoscedastic and nonautocorrelated disturbances case. In that instance, the

weighting matrix in Theorem 13.2 will be $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ and we will obtain the apparently missing result.

The **GMM estimator** in the heteroscedastic regression model is produced by the empirical moment equations

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GMM}) = \frac{1}{n}\mathbf{X}'\hat{\boldsymbol{\varepsilon}}(\hat{\boldsymbol{\beta}}_{GMM}) = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}) = \mathbf{0}. \tag{13-26}$$

The estimator is obtained by minimizing

$$q = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM})\mathbf{W}\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}),$$

where $\mathbf{W}$ is a positive definite weighting matrix. The optimal weighting matrix would be

$$\mathbf{W} = \{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})]\}^{-1},$$

which is the inverse of

$$\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})] = \text{Asy.Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\boldsymbol{\varepsilon}_i\right] = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sigma^2\omega_i\mathbf{x}_i\mathbf{x}_i' = \sigma^2\mathbf{Q}^*.$$

(See Section 9.4.1.) The optimal weighting matrix would be $[\sigma^2\mathbf{Q}^*]^{-1}$. But recall that this minimization problem is an exactly identified case, so the weighting matrix is irrelevant to the solution. You can see the result in the moment equation—that equation is simply the normal equation for ordinary least squares. We can solve the moment equations exactly, so there is no need for the weighting matrix. Regardless of the covariance matrix of the moments, the GMM estimator for the heteroscedastic regression model is ordinary least squares. We can use the results we have already obtained to find its asymptotic covariance matrix. The implied estimator is the White estimator in (9-5). (Once again, see Theorem 13.2.) The conclusion to be drawn at this point is that until we make some specific assumptions about the variances, we do not have a more efficient estimator than least squares, but we do have to modify the estimated asymptotic covariance matrix.

13.6.2 SINGLE-EQUATION NONLINEAR MODELS

Suppose that the theory specifies a relationship, $y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i$, where $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector that we wish to estimate. This may not be a regression relationship, because it is possible that $\text{Cov}[\varepsilon_i, h(\mathbf{x}_i, \boldsymbol{\beta})] \neq 0$, or even $\text{Cov}[\varepsilon_i, \mathbf{x}_j] \neq 0$ for all i and j. Consider, for example, a model that contains lagged dependent variables and autocorrelated disturbances. (See Section 20.9.3.) For the present, we assume that $E[\boldsymbol{\varepsilon}|\mathbf{X}] \neq \mathbf{0}$, and $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\boldsymbol{\Omega} = \boldsymbol{\Sigma}$ where Σ is symmetric and positive definite but otherwise unrestricted. The disturbances may be heteroscedastic and/or autocorrelated. But for the possibility of correlation between regressors and disturbances, this model would be a generalized, possibly nonlinear, regression model. Suppose that at each observation i we observe a vector of L variables, $\mathbf{z}_i$, such that $\mathbf{z}_i$ is uncorrelated with $\boldsymbol{\varepsilon}_i$. You will recognize $\mathbf{z}_i$ as a set of **instrumental variables**. The assumptions thus far have implied a set of orthogonality conditions, $E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$, which may be sufficient to identify (if $L = K$) or even overidentify (if $L > K$) the parameters of the model. (See Section 8.3.4.)

For convenience, define

$$\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}}) = y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}), \quad i = 1, \ldots, n,$$

and

$$\mathbf{Z} = n \times L \text{ matrix whose } i\text{th row is } \mathbf{z}_i'.$$

By a straightforward extension of our earlier results, we can produce a GMM estimator of $\boldsymbol{\beta}$. The sample moments will be

$$\overline{\mathbf{m}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{n}\mathbf{Z}'\mathbf{e}(\mathbf{X}, \boldsymbol{\beta}).$$

The minimum distance estimator will be the $\hat{\boldsymbol{\beta}}$ that minimizes

$$q = \overline{\mathbf{m}}_n(\hat{\boldsymbol{\beta}})'\mathbf{W}\overline{\mathbf{m}}_n(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n}[\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})'\mathbf{Z}]\right)\mathbf{W}\left(\frac{1}{n}[\mathbf{Z}'\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})]\right) \tag{13-27}$$

for some choice of $\mathbf{W}$ that we have yet to determine. The criterion given earlier produces the **nonlinear instrumental variable estimator**. If we use $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$, then we have exactly the estimation criterion we used in Section 8.9, where we defined the nonlinear instrumental variables estimator. Apparently (13-27) is more general, because we are not limited to this choice of $\mathbf{W}$. For any given choice of $\mathbf{W}$, as long as there are enough orthogonality conditions to identify the parameters, estimation by minimizing q is, at least in principle, a straightforward problem in nonlinear optimization. The optimal choice of $\mathbf{W}$ for this estimator is

$$\begin{aligned}\mathbf{W}_{\text{GMM}} &= \{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\beta})]\}^{-1} \\ &= \left\{\text{Asy.Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right]\right\}^{-1} = \left\{\text{Asy.Var}\left[\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{e}(\mathbf{X}, \boldsymbol{\beta})\right]\right\}^{-1}.\end{aligned} \tag{13-28}$$

For our model, this is

$$\mathbf{W} = \left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}[\mathbf{z}_i\varepsilon_i, \mathbf{z}_j\varepsilon_j]\right]^{-1} = \left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{z}_i\mathbf{z}_j'\right]^{-1} = \left[\frac{\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}}{n}\right]^{-1}.$$

If we insert this result in (13-27), we obtain the criterion for the GMM estimator,

$$q = \left[\left(\frac{1}{n}\right)\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})'\mathbf{Z}\right]\left(\frac{\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}}{n}\right)^{-1}\left[\left(\frac{1}{n}\right)\mathbf{Z}'\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})\right].$$

There is a possibly difficult detail to be considered. The GMM estimator involves

$$\frac{1}{n}\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{z}_i\mathbf{z}_j'\,\text{Cov}[\varepsilon_i, \varepsilon_j] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{z}_i\mathbf{z}_j'\,\text{Cov}[(y_i - h(\mathbf{x}_i, \boldsymbol{\beta})),(y_j - h(\mathbf{x}_j, \boldsymbol{\beta}))].$$

The conditions under which such a double sum might converge to a positive definite matrix are sketched in Section 9.3.2. Assuming that they do hold, estimation appears to require that an estimate of $\boldsymbol{\beta}$ be in hand already, even though it is the object of estimation. It may be that a consistent but inefficient estimator of $\boldsymbol{\beta}$ is available. Suppose for the present that one is. If observations are uncorrelated, then the cross-observation terms may be omitted, and what is required is

$$\frac{1}{n}\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i' \operatorname{Var}[(y_i - h(\mathbf{x}_i, \boldsymbol{\beta}))].$$

We can use a counterpart to the White (1980) estimator discussed in Section 9.2 for this case,

$$\mathbf{S}_0 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'(y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))^2. \tag{13-29}$$

If the disturbances are autocorrelated but the process is stationary, then Newey and West's (1987a) estimator is available (assuming that the autocorrelations are sufficiently small at a reasonable lag, p),

$$\mathbf{S} = \left[\mathbf{S}_0 + \frac{1}{n}\sum_{\ell=1}^{p} w(\ell)\sum_{i=\ell+1}^{n}(e_i e_{i-\ell})(\mathbf{z}_i\mathbf{z}_{i-\ell}' + \mathbf{z}_{i-\ell}\mathbf{z}_i')\right] = \sum_{\ell=0}^{p} w(\ell)\mathbf{S}_\ell, \tag{13-30}$$

where $\mathrm{w}(\ell) = 1 - \ell/(p+1)$. (This is the *Bartlett weight*.) The maximum lag length p must be determined in advance. We will require that observations that are far apart in time—that is, for which $|i - \ell|$ is large—must have increasingly smaller covariances for us to establish the convergence results that justify OLS, GLS, and now GMM estimation. The choice of p is a reflection of how far back in time one must go to consider the autocorrelation negligible for purposes of estimating $(1/n)\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}$. Current practice suggests using the smallest integer greater than or equal to $n^{1/4}$.

Still left open is the question of where the initial consistent estimator should be obtained. One possibility is to obtain an inefficient but consistent GMM estimator by using $\mathbf{W} = \mathbf{I}$ in (13-27). That is, use a nonlinear (or linear, if the equation is linear) instrumental variables estimator. This first-step estimator can then be used to construct $\mathbf{W}$, which, in turn, can then be used in the GMM estimator. Another possibility is that $\boldsymbol{\beta}$ may be consistently estimable by some straightforward procedure other than GMM.

Once the GMM estimator has been computed, its asymptotic covariance matrix and asymptotic distribution can be estimated based on Theorem 13.2. Recall that

$$\overline{\mathbf{m}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i,$$

which is a sum of $L \times 1$ vectors. The derivative, $\partial\overline{\mathbf{m}}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$, is a sum of $L \times K$ matrices, so

$$\overline{\mathbf{G}}(\boldsymbol{\beta}) = \partial\overline{\mathbf{m}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{G}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\left[\frac{\partial\varepsilon_i}{\partial\boldsymbol{\beta}'}\right]. \tag{13-31}$$

In the model we are considering here, $\dfrac{\partial\varepsilon_i}{\partial\boldsymbol{\beta}'} = \dfrac{-\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}$. The derivatives are the pseudoregressors in the linearized regression model that we examined in Section 7.2.3. Using the notation defined there, $\dfrac{\partial\varepsilon_i}{\partial\boldsymbol{\beta}} = -\mathbf{x}_i^0$, so

$$\overline{\mathbf{G}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{G}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}-\mathbf{z}_i\mathbf{x}_i^{0\prime} = -\frac{1}{n}\mathbf{Z}'\mathbf{X}^0. \tag{13-32}$$

With this matrix in hand, the estimated asymptotic covariance matrix for the GMM estimator is

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \frac{1}{n}\left[\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\left(\frac{1}{n}\mathbf{Z}'\hat{\boldsymbol{\Sigma}}\mathbf{Z}\right)^{-1}\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})\right]^{-1} = [(\mathbf{X}^{0\prime}\mathbf{Z})(\mathbf{Z}'\hat{\boldsymbol{\Sigma}}\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X}^0)]^{-1}. \tag{13-33}$$

(The two minus signs, a $1/n^2$, and an n^2 all fall out of the result.)

If the $\boldsymbol{\Sigma}$ that appears in (13-33) were $\sigma^2\mathbf{I}$, then (13-33) would be precisely the asymptotic covariance matrix that appears in Theorem 8.1 for linear models and Theorem 8.2 for nonlinear models. But there is an interesting distinction between this estimator and the IV estimators discussed earlier. In the earlier cases, when there were more instrumental variables than parameters, we resolved the overidentification by specifically choosing a set of K instruments, the K projections of the columns of $\mathbf{X}$ or $\mathbf{X}^0$ into the column space of $\mathbf{Z}$. Here, in contrast, we do not attempt to resolve the overidentification; we simply use all the instruments and minimize the GMM criterion. You should be able to show that when $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ *and we use this information,* the same parameter estimates will be obtained when all is said and done. But, if we use a weighting matrix that differs from $\mathbf{W} = (\mathbf{Z}'\mathbf{Z}/n)^{-1}$, then they are not.

13.6.3 SEEMINGLY UNRELATED REGRESSION EQUATIONS

In Section 10.2.3, we considered FGLS estimation of the equation system

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{h}_1(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_1,\\ \mathbf{y}_2 &= \mathbf{h}_2(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_2,\\ &\vdots\\ \mathbf{y}_M &= \mathbf{h}_M(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_M.\end{aligned}$$

The development there extends backward to the linear system as well. However, none of the estimators considered is consistent if the pseudoregressors, $\mathbf{x}^0_{tm}$, or the actual regressors, $\mathbf{x}_{tm}$, for the linear model, are correlated with the disturbances, ε_{tm}. Suppose we allow for this correlation both within and across equations. (If it is, in fact, absent, then the GMM estimator developed here will remain consistent.) For simplicity in this section, we will denote observations with subscript t and equations with subscripts i and j. Suppose, as well, that there are a set of instrumental variables, $\mathbf{z}_t$, such that

$$E[\mathbf{z}_t\varepsilon_{tm}] = \mathbf{0}, t = 1, \ldots, T \text{ and } m = 1, \ldots, M. \tag{13-34}$$

(We could allow a separate set of instrumental variables for each equation, but it would needlessly complicate the presentation.)

Under these assumptions, the nonlinear FGLS and ML estimators given earlier will be inconsistent. But a relatively minor extension of the instrumental variables technique developed for the single-equation case in Section 8.4 can be used instead. The sample analog to (13-34) is

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t[y_{ti} - h_i(\mathbf{X}_t, \boldsymbol{\beta})] = \mathbf{0}, \quad i = 1, \ldots, M.$$

If we use this result for each equation in the system, one at a time, then we obtain exactly the GMM estimator discussed in Section 13.6.2. But, in addition to the efficiency loss

that results from not imposing the cross-equation constraints in $\boldsymbol{\beta}$, we would also neglect the correlation between the disturbances. Let

$$\frac{1}{T}\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z} = E\left[\frac{\mathbf{Z}'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j'\mathbf{Z}}{T}\right]. \tag{13-35}$$

The GMM criterion for estimation in this setting is

$$\begin{aligned} q &= \sum_{i=1}^{M}\sum_{j=1}^{M}[(\mathbf{y}_i - \mathbf{h}_i(\mathbf{X},\boldsymbol{\beta}))'\mathbf{Z}/T][\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'(\mathbf{y}_j - \mathbf{h}_j(\mathbf{X},\boldsymbol{\beta}))/T] \\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}[\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})'\mathbf{Z}/T][\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'\boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T], \end{aligned} \tag{13-36}$$

where $[\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}$ denotes the ijth block of the inverse of the matrix with the ijth block equal to $\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T$.

GMM estimation would proceed in several passes. To compute any of the variance parameters, we will require an initial consistent estimator of $\boldsymbol{\beta}$. This step can be done with equation-by-equation nonlinear instrumental variables—see Section 8.9—although if equations have parameters in common, then a choice must be made as to which to use. At the next step, the familiar White or Newey–West technique is used to compute, block by block, the matrix in (13-35). Because it is based on a consistent estimator of $\boldsymbol{\beta}$ (we assume), this matrix need not be recomputed. Now, with this result in hand, an iterative solution to the maximization problem in (13-36) can be sought, for example, using the methods of Appendix E. The first-order conditions are

$$\frac{\partial q}{\partial \boldsymbol{\beta}} = -2\sum_{i=1}^{M}\sum_{j=1}^{M}[\mathbf{X}_i^0(\boldsymbol{\beta})'\mathbf{Z}/T][\mathbf{Z}'\mathbf{W}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'\boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T] = \mathbf{0}. \tag{13-37}$$

Note again that the blocks of the inverse matrix in the center are extracted from the larger constructed matrix *after inversion*.[16] At completion, the asymptotic covariance matrix for the GMM estimator is estimated with

$$\mathbf{V}_{\text{GMM}} = \frac{1}{T}\left[\sum_{i=1}^{M}\sum_{j=1}^{M}[\mathbf{X}_i^0(\boldsymbol{\beta})'\mathbf{Z}/T][\mathbf{Z}'\mathbf{W}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'\mathbf{X}_j^0(\boldsymbol{\beta})/T]\right]^{-1}.$$

13.6.4 GMM ESTIMATION OF DYNAMIC PANEL DATA MODELS

Panel data are well suited for examining dynamic effects, as in the first-order model,

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + c_i + \varepsilon_{it} \\ &= \mathbf{w}_{it}'\boldsymbol{\theta} + \alpha_i + \varepsilon_{it}, \end{aligned}$$

where the set of right-hand-side variables, $\mathbf{w}_{it}$, now includes the lagged dependent variable, $y_{i,t-1}$. Adding dynamics to a model in this fashion creates a major change in the interpretation of the equation. Without the lagged variable, the independent variables represent the full set of information that produce observed outcome y_{it}. With the lagged variable, we now have in the equation the entire history of the right-hand-side variables, so that any measured influence is conditioned on this history; in this case, any impact of $\mathbf{x}_{it}$

[16]This brief discussion might understate the complexity of the optimization problem in (13-36), but that is inherent in the procedure.

represents the effect of *new* information. Substantial complications arise in estimation of such a model. In both the fixed and random effects settings, the difficulty is that the lagged dependent variable is correlated with the disturbance, even if it is assumed that ε_{it} is not itself autocorrelated. For the moment, consider the fixed effects model as an ordinary regression with a lagged dependent variable that is dependent across observations. In that dynamic regression model, the estimator based on T observations is biased in finite samples, but it is consistent in T. The finite sample bias is of order $1/T$. The same result applies here, but the difference is that whereas before we obtained our large sample results by allowing T to grow large, in this setting, T is assumed to be small and fixed, and large-sample results are obtained with respect to n growing large, not T. The fixed effects estimator of $\boldsymbol{\theta} = [\boldsymbol{\beta}, \delta]$ can be viewed as an average of n such estimators. Assume for now that $T \geq K + 1$ where K is the number of variables in $\mathbf{x}_{it}$. Then, from (11-14),

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\right]^{-1}\left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{y}_i\right] \\ &= \left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\right]^{-1}\left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\mathbf{d}_i\right] \\ &= \sum_{i=1}^{n} \mathbf{F}_i\mathbf{d}_i,\end{aligned}$$

where the rows of the $T \times (K + 1)$ matrix $\mathbf{W}_i$ are $\mathbf{w}_{it}'$ and $\mathbf{M}^0$ is the $T \times T$ matrix that creates deviations from group means [see (11-14)]. Each group-specific estimator, $\mathbf{d}_i$, is inconsistent, as it is biased in finite samples and its variance does not go to zero as n increases. This matrix weighted average of n inconsistent estimators will also be inconsistent. (This analysis is only heuristic. If $T < K + 1$, then the individual coefficient vectors cannot be computed.[17])

The problem is more transparent in the random effects model. In the model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + u_i + \varepsilon_{it},$$

the lagged dependent variable is correlated with the compound disturbance in the model because the same u_i enters the equation for every observation in group i.

Neither of these results renders the model inestimable, but they do make necessary some technique other than our familiar LSDV or FGLS estimators. The general approach, which has been developed in several stages in the literature,[18] relies on instrumental variables estimators and, most recently, on a GMM estimator. For example, in either the fixed or random effects cases, the heterogeneity can be swept from the model by taking first differences, which produces

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + (\varepsilon_{it} - \varepsilon_{i,t-1}).$$

This model is still complicated by correlation between the lagged dependent variable and the disturbance (and by its first-order moving average disturbance). But without the

[17]Further discussion is given by Nickell (1981), Ridder and Wansbeek (1990), and Kiviet (1995).

[18]The model was first proposed in this form by Balestra and Nerlove (1966). See, for example, Anderson and Hsiao (1981, 1982), Bhargava and Sargan (1983), Arellano (1989), Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), and Nerlove (1971a,b).

group effects, there is a simple instrumental variables estimator available. Assuming that the time series is long enough, one could use the lagged differences, $(y_{i,t-2} - y_{i,t-3})$, or the lagged levels, $y_{i,t-2}$ and $y_{i,t-3}$, as one or two instrumental variables for $(y_{i,t-1} - y_{i,t-2})$. (The other variables can serve as their own instruments.) This is the Anderson and Hsiao estimator developed for this model in Section 11.8.3. By this construction, then, the treatment of this model is a standard application of the instrumental variables technique that we developed in Section 11.8.[19] This illustrates the flavor of an instrumental variables approach to estimation. But, as Arellano et al. and Ahn and Schmidt (1995) have shown, there is still more information in the sample that can be brought to bear on estimation, in the context of a GMM estimator, which we now consider.

We can extend the Hausman and Taylor (HT) formulation of the random effects model in Section 11.8.2 to include the lagged dependent variable,

$$\begin{aligned} y_{it} &= \delta y_{i,t-1} + \mathbf{x}_{1it}'\boldsymbol{\beta}_1 + \mathbf{x}_{2it}'\boldsymbol{\beta}_2 + \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 + \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 + \varepsilon_{it} + u_i \\ &= \boldsymbol{\theta}'\mathbf{w}_{it} + \varepsilon_{it} + u_i \\ &= \boldsymbol{\theta}'\mathbf{w}_{it} + \eta_{it}, \end{aligned}$$

where

$$\mathbf{w}_{it} = [y_{i,t-1}, \mathbf{x}_{1it}', \mathbf{x}_{2it}', \mathbf{z}_{1i}', \mathbf{z}_{2i}']'$$

is now a $(1 + K_1 + K_2 + L_1 + L_2) \times 1$ vector. The terms in the equation are the same as in the Hausman and Taylor model. Instrumental variables estimation of the model without the lagged dependent variable is discussed in Section 11.8.1 on the HT estimator. Moreover, by just including $y_{i,t-1}$ in $\mathbf{x}_{2it}$, we see that the HT approach extends to this setting as well, essentially without modification. Arellano et al. suggest a GMM estimator and show that efficiency gains are available by using a larger set of moment conditions. In the previous treatment, we used a GMM estimator constructed as follows: the set of moment conditions we used to formulate the instrumental variables were

$$E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{it} - \bar{\eta}_i)\right] = E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\varepsilon_{it} - \bar{\varepsilon}_i)\right] = \mathbf{0}.$$

This moment condition is used to produce the instrumental variable estimator. We could ignore the nonscalar variance of η_{it} and use simple instrumental variables at this point. However, by accounting for the random effects formulation and using the counterpart to feasible GLS, we obtain the more efficient estimator in Section 11.8.4. As usual, this can be done in two steps. The inefficient estimator is computed to obtain the residuals needed to estimate the variance components. This is Hausman and Taylor's steps 1 and 2. Steps 3 and 4 are the GMM estimator based on these estimated variance components.

[19]There is a question as to whether one should use differences or levels as instruments. Arellano (1989) and Kiviet (1995) give evidence that the latter is preferable.

Arellano et al. suggest that the preceding does not exploit all the information in the sample. In simple terms, within the T observations in group i, we have not used the fact that

$$E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{is}-\bar{\eta}_i)\right]=\mathbf{0}\quad\text{for some } s\neq t.$$

Thus, for example, not only are disturbances at time t uncorrelated with these variables at time t, arguably, they are uncorrelated with the same variables at time $t-1, t-2$, possibly $t+1$, and so on. In principle, the number of valid instruments is potentially enormous. Suppose, for example, that the set of instruments listed above is strictly exogenous with respect to η_{it} in every period including current, lagged, and future. Then, there are a total of $[T(K_1+K_2)+L_1+K_1)]$ moment conditions for every observation. Consider, for example, a panel with two periods. We would have for the two periods,

$$E\left[\begin{pmatrix}\mathbf{x}_{1i1}\\ \mathbf{x}_{2i1}\\ \mathbf{x}_{1i2}\\ \mathbf{x}_{2i2}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{i1}-\bar{\eta}_i)\right]=\mathbf{0}\qquad\text{and}\qquad E\left[\begin{pmatrix}\mathbf{x}_{1i1}\\ \mathbf{x}_{2i1}\\ \mathbf{x}_{1i2}\\ \mathbf{x}_{2i2}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{i2}-\bar{\eta}_i)\right]=\mathbf{0}. \tag{13-38}$$

How much useful information is brought to bear on estimation of the parameters is uncertain, as it depends on the correlation of the instruments with the included exogenous variables in the equation. The farther apart in time these sets of variables become, the less information is likely to be present. (The literature on this subject contains reference to *strong* versus *weak* instrumental variables.[20]) To proceed, as noted, we can include the lagged dependent variable in $\mathbf{x}_{2i}$. This set of instrumental variables can be used to construct the estimator, actually whether the lagged variable is present or not. We note, at this point, that on this basis, Hausman and Taylor's estimator did not actually use all the information available in the sample. We now have the elements of the Arellano et al. estimator in hand; what remains is essentially the (unfortunately, fairly involved) algebra, which we now develop.

Let

$$\mathbf{W}_i=\begin{bmatrix}\mathbf{w}_{i1}'\\ \mathbf{w}_{i2}'\\ \vdots\\ \mathbf{w}_{iT}'\end{bmatrix}=\text{the full set of rhs data for group } i,\quad\text{and}\quad \mathbf{y}_i=\begin{bmatrix}y_{i1}\\ y_{i2}\\ \vdots\\ y_{iT}\end{bmatrix}.$$

Note that $\mathbf{W}_i$ is assumed to be a $T\times(1+K_1+K_2+L_1+L_2)$ matrix. Because there is a lagged dependent variable in the model, it must be assumed that there are actually $T+1$ observations available on y_{it}. To avoid cumbersome, cluttered notation, we will leave this distinction embedded in the notation for the moment. Later, when necessary,

[20]See West (2001).

we will make it explicit. It will reappear in the formulation of the instrumental variables. A total of T observations will be available for constructing the IV estimators. We now form a matrix of instrumental variables.[21] We will form a matrix $\mathbf{V_i}$ consisting of $T_i - 1$ rows constructed the same way for $T_i - 1$ observations and a final row that will be different, as discussed later.[22] The matrix will be of the form

$$\mathbf{V}_i = \begin{bmatrix} \mathbf{v}_{i1}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{v}_{i2}' & \cdots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{a}_i' \end{bmatrix}. \tag{13-39}$$

The instrumental variable sets contained in $\mathbf{v}_{it}'$ which have been suggested might include the following from within the model:

$\mathbf{x}_{it}$ and $\mathbf{x}_{i,t-1}$ (i.e., current and one lag of all the time-varying variables),
$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$ (i.e., all current, past, and future values of all the time-varying variables),
$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}$ (i.e., all current and past values of all the time-varying variables).

The time-invariant variables that are uncorrelated with u_i, that is, $\mathbf{z}_{1i}$, are appended at the end of the nonzero part of each of the first $T - 1$ rows. It may seem that including $\mathbf{x}_2$ in the instruments would be invalid. However, we will be converting the disturbances to deviations from group means which are free of the latent effects—that is, this set of moment conditions will ultimately be converted to what appears in (13-38). While the variables are correlated with u_i by construction, they are not correlated with $\varepsilon_{it} - \bar{\varepsilon}_i$. The final row of $\mathbf{V}_i$ is important to the construction. Two possibilities have been suggested:

$\mathbf{a}_i' = [\mathbf{z}_{1i}' \quad \bar{\mathbf{x}}_{i1}]$ (produces the Hausman and Taylor estimator),
$\mathbf{a}_i' = [\mathbf{z}_{1i}' \quad \mathbf{x}_{1i1}', \mathbf{x}_{1i2}', \ldots, x_{1iT}]$ (produces Amemiya and MaCurdy's estimator).

Note that the $\mathbf{a}$ variables are exogenous time-invariant variables, $\mathbf{z}_{1i}$ and the exogenous time-varying variables, either condensed into the single group mean or in the raw form, with the full set of T observations.

To construct the estimator, we will require a transformation matrix, $\mathbf{H}$, constructed as follows. Let $\mathbf{M}^{01}$ denote the first $T - 1$ rows of $\mathbf{M}^0$, the matrix that creates deviations from group means. Then,

$$\mathbf{H} = \begin{bmatrix} \mathbf{M}^{01} \\ \frac{1}{T}\mathbf{i}_T' \end{bmatrix}.$$

Thus, $\mathbf{H}$ replaces the last row of $\mathbf{M}^0$ with a row of $1/T$. The effect is as follows: if $\mathbf{q}$ is T observations on a variable, then $\mathbf{Hq}$ produces $\mathbf{q}^*$ in which the first $T - 1$ observations are converted to deviations from group means and the last observation is the group mean. In particular, let the $T \times 1$ column vector of disturbances,

$$\eta_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}] = [(\varepsilon_{i1} + u_i), (\varepsilon_{i2} + u_i), \ldots, (\varepsilon_{iT} + u_i)]',$$

[21] Different approaches to this have been considered by Hausman and Taylor (1981), Arellano et al. (1991, 1995, 1999), Ahn and Schmidt (1995), and Amemiya and MaCurdy (1986), among others.

[22] This is to exploit a useful algebraic result discussed by Arellano and Bover (1995).

then

$$\mathbf{H}\boldsymbol{\eta} = \begin{bmatrix} \eta_{i1} - \overline{\eta}_i \\ \vdots \\ \eta_{i,T-1} - \overline{\eta}_i \\ \overline{\eta}_i \end{bmatrix}.$$

We can now construct the moment conditions. With all this machinery in place, we have the result that appears in (13-40), that is,

$$E[\mathbf{V}_i'\mathbf{H}\boldsymbol{\eta}_i] = E[\mathbf{g}_i] = \mathbf{0}.$$

It is useful to expand this for a particular case. Suppose $T = 3$ and we use as instruments the current values in period 1, and the current and previous values in period 2 and the Hausman and Taylor form for the invariant variables. Then the preceding is

$$E\left[\begin{pmatrix} \mathbf{x}_{1i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{x}_{2i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{z}_{1i} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{1i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{z}_{1i} \\ \mathbf{0} & \mathbf{0} & \overline{\mathbf{x}}_{1i} \end{pmatrix} \begin{pmatrix} \eta_{i1} - \overline{\eta}_i \\ \eta_{i2} - \overline{\eta}_i \\ \overline{\eta}_i \end{pmatrix}\right] = \mathbf{0}. \tag{13-40}$$

This is the same as (13-38).[23] The empirical moment condition that follows from this is

$$\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} \mathbf{V}_i'\mathbf{H}\boldsymbol{\eta}_i$$
$$= \operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} \mathbf{V}_i'\mathbf{H} \begin{pmatrix} y_{i1} - \delta y_{i0} - \mathbf{x}_{1i1}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i1}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \\ y_{i2} - \delta y_{i1} - \mathbf{x}_{1i2}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i2}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \\ \vdots \\ y_{iT} - \delta y_{i,T-1} - \mathbf{x}_{1iT}'\boldsymbol{\beta}_1 - \mathbf{x}_{2iT}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \end{pmatrix} = \mathbf{0}.$$

Write this as

$$\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}_i = \operatorname{plim} \overline{\mathbf{m}} = \mathbf{0}.$$

The GMM estimator $\hat{\boldsymbol{\theta}}$ is then obtained by minimizing $q = \overline{\mathbf{m}}'\mathbf{A}\overline{\mathbf{m}}$ with an appropriate choice of the weighting matrix, $\mathbf{A}$. The optimal weighting matrix will be the inverse of

[23] In some treatments—for example, Blundell and Bond (1998)—an additional condition is assumed for the initial value, y_{i0}, namely $E[y_{i0}|\textit{exogenous data}] = \mu_0$. This would add a row at the top of the matrix in (13-40) containing $[(y_{i0} - \mu_0), 0, 0]$.

the asymptotic covariance matrix of $\sqrt{n}\,\bar{\mathbf{m}}$. With a consistent estimator of $\boldsymbol{\theta}$ in hand, this can be estimated empirically using

$$\text{Est.Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{m}}_i\hat{\mathbf{m}}_i' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\hat{\eta}_i\hat{\eta}_i'\mathbf{H}'\mathbf{V}_i.$$

This is a robust estimator that allows an unrestricted $T \times T$ covariance matrix for the T disturbances, $\varepsilon_{it} + u_i$. But we have assumed that this covariance matrix is the $\boldsymbol{\Sigma}$ defined in (11-31) for the random effects model. To use this information we would, instead, use the residuals in

$$\hat{\boldsymbol{\eta}}_i = \mathbf{y}_i - \mathbf{W}_i\hat{\boldsymbol{\theta}}$$

to estimate σ_u^2 and σ_ε^2 and then $\boldsymbol{\Sigma}$, which produces

$$\text{Est.Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}'\mathbf{V}_i.$$

We now have the full set of results needed to compute the GMM estimator. The solution to the optimization problem of minimizing q with respect to the parameter vector $\boldsymbol{\theta}$ is

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{GMM} = &\left[\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}\mathbf{V}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{W}_i\right)\right]^{-1}\\ &\times\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}\mathbf{V}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{y}_i\right).\end{aligned} \qquad \textbf{(13-41)}$$

The estimator of the asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_{GMM}$ is the inverse matrix in brackets.

The remaining loose end is how to obtain the consistent estimator of $\hat{\boldsymbol{\theta}}$ to compute $\hat{\boldsymbol{\Sigma}}$. Recall that the GMM estimator is consistent with any positive definite weighting matrix, $\mathbf{A}$, in our preceding expression. Therefore, for an initial estimator, we can set $\mathbf{A} = \mathbf{I}$ and use the simple instrumental variables estimator,

$$\hat{\boldsymbol{\theta}}_{\text{IV}} = \left[\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{W}_i\right)\right]^{-1}\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{y}_i\right).$$

It is more common to proceed directly to the 2SLS estimator (see Sections 8.3.4 and 11.8.2), which uses

$$\mathbf{A} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{H}\mathbf{V}_i\right)^{-1}.$$

The estimator is, then, the one given earlier in (13-41) with $\hat{\boldsymbol{\Sigma}}$ replaced by $\mathbf{I}_T$. Either estimator is a function of the sample data only and provides the initial estimator we need.

Ahn and Schmidt (among others) observed that the IV estimator proposed here, as extensive as it is, still neglects quite a lot of information and is therefore (relatively) inefficient. For example, in the first differenced model,

$$E[y_{is}(\varepsilon_{it} - \varepsilon_{i,t-1})] = 0, \quad s = 0, \ldots, t-2, \quad t = 2, \ldots, T.$$

That is, the *level* of y_{is} is uncorrelated with the differences of disturbances that are at least two periods subsequent.[24] (The differencing transformation, as the transformation to deviations from group means, removes the individual effect.) The corresponding moment equations that can enter the construction of a GMM estimator are

$$\frac{1}{n}\sum_{i=1}^{n} y_{is}[(y_{it} - y_{i,t-1}) - \delta(y_{i,t-1} - y_{i,t-2}) - (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta}] = 0$$
$$s = 0, \ldots, t-2, \quad t = 2, \ldots, T.$$

Altogether, Ahn and Schmidt identify $T(T-1)/2 + T - 2$ such equations that involve mixtures of the levels and differences of the variables. The main conclusion that they demonstrate is that in the dynamic model, there is a large amount of information to be gleaned not only from the familiar relationships among the levels of the variables, but also from the implied relationships between the levels and the first differences. The issue of correlation between the transformed y_{it} and the deviations of ε_{it} is discussed in the papers cited.[25]

The number of orthogonality conditions (instrumental variables) used to estimate the parameters of the model is determined by the number of variables in $\mathbf{v}_{it}$ and $\mathbf{a}_i$ in (13-39). In most cases, the model is vastly overidentified—there are far more orthogonality conditions than parameters. As usual in GMM estimation, a test of the overidentifying restrictions can be based on q, the estimation criterion. At its minimum, the limiting distribution of nq is chi squared with degrees of freedom equal to the number of instrumental variables in total minus

$$(1 + K_1 + K_2 + L_1 + L_2).$$[26]

Example 13.10 GMM Estimation of a Dynamic Panel Data Model of Local Government Expenditures

Dahlberg and Johansson (2000) estimated a model for the local government expenditure of several hundred municipalities in Sweden observed over the nine-year period $t = 1979$ to 1987. The equation of interest is

$$S_{i,t} = \alpha_t + \sum_{j=1}^{m}\beta_j S_{i,t-j} + \sum_{j=1}^{m}\gamma_j R_{i,t-j} + \sum_{j=1}^{m}\delta_j G_{i,t-j} + f_i + \varepsilon_{it},$$

for $i = 1, \ldots, n = 265$, and $t = m + 1, \ldots, 9$. (We have changed their notation slightly to make it more convenient.) $S_{i,t}$, $R_{i,t}$, and $G_{i,t}$ are municipal spending, receipts (taxes and fees), and central government grants, respectively. Analogous equations are specified for the current values of $R_{i,t}$ and $G_{i,t}$. The appropriate lag length, m, is one of the features of interest to be determined by the empirical study. The model contains a municipality specific effect, f_i,

[24] This is the approach suggested by Holtz-Eakin (1988) and Holtz-Eakin, Newey, and Rosen (1988).

[25] As Ahn and Schmidt show, there are potentially huge numbers of additional orthogonality conditions in this model owing to the relationship between first differences and second moments. We do not consider those. The matrix $\mathbf{V}_i$ could be huge. Consider a model with 10 time-varying, right-hand-side variables and suppose T_i is 15. Then, there are 15 rows and roughly $15 \times (10 \times 15)$ or 2,250 columns. The Ahn and Schmidt estimator, which involves potentially thousands of instruments in a model containing only a handful of parameters may become a bit impractical at this point. The common approach is to use only a small subset of the available instrumental variables. The order of the computation grows as the number of parameters times the square of T.

[26] This is true generally in GMM estimation. It was proposed for the dynamic panel data model by Bhargava and Sargan (1983).

which is not specified as being either *fixed* or *random*. To eliminate the individual effect, the model is converted to first differences. The resulting equation is

$$\Delta S_{i,t} = \lambda_t + \sum_{j=1}^{m} \beta_j \Delta S_{i,t-j} + \sum_{j=1}^{m} \gamma_j \Delta R_{i,t-j} + \sum_{j=1}^{m} \delta_j \Delta G_{i,t-j} + u_{it},$$

or

$$y_{i,t} = \mathbf{x}_{i,t}'\boldsymbol{\theta} + u_{i,t},$$

where $\Delta S_{i,t} = S_{i,t} - S_{i,t-1}$ and so on and $u_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$. This removes the group effect and leaves the time effect. Because the time effect was unrestricted to begin with, $\Delta\alpha_t = \lambda_t$ remains an unrestricted time effect, which is treated as fixed and modeled with a time-specific dummy variable. The maximum lag length is set at $m = 3$. With nine years of data, this leaves usable observations from 1983 to 1987 for estimation, that is, $t = m + 2, \ldots, 9$. Similar equations were fit for $R_{i,t}$ and $G_{i,t}$.

The orthogonality conditions claimed by the authors are

$$E[S_{i,s}u_{i,t}] = E[R_{i,s}u_{i,t}] = E[G_{i,s}u_{i,t}] = 0, \quad s = 1, \ldots, t-2.$$

The orthogonality conditions are stated in terms of the levels of the financial variables and the differences of the disturbances. The issue of this formulation as opposed to, for example, $E[\Delta S_{i,s}\Delta\varepsilon_{i,t}] = 0$ (which is implied) is discussed by Ahn and Schmidt (1995). As we shall see, this set of orthogonality conditions implies a total of 80 instrumental variables. The authors use only the first of the three sets listed, which produces a total of 30. For the five observations, using the formulation developed in Section 13.6.5, we have the following matrix of instrumental variables for the orthogonality conditions,

$$\mathbf{Z}_i = \begin{bmatrix} \boldsymbol{S}_{81-79} & d_{83} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \boldsymbol{S}_{82-79} & d_{84} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \boldsymbol{S}_{83-79} & d_{85} & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \boldsymbol{S}_{84-79} & d_{86} & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \boldsymbol{S}_{85-79} & d_{87} \end{bmatrix} \begin{matrix} 1983 \\ 1984 \\ 1985, \\ 1986 \\ 1987 \end{matrix}$$

where the notation $\boldsymbol{S}_{t1-t0}$ indicates the range of years for that variable. For example, $\boldsymbol{S}_{83-79}$ denotes $[S_{i,1983}, S_{i,1982}, S_{i,1981}, S_{i,1980}, S_{i,1979}]$ and d_{year} denotes the year-specific dummy variable. Counting columns in $\mathbf{Z}_i$ we see that using only the lagged values of the dependent variable and the time dummy variables, we have $(3 + 1) + (4 + 1) + (5 + 1) + (6 + 1) + (7 + 1) = 30$ instrumental variables. Using the lagged values of the other two variables in each equation would add 50 more, for a total of 80 if all the orthogonality conditions suggested earlier were employed. Given the preceding construction, the orthogonality conditions are now $E[\mathbf{Z}_i'\mathbf{u}_i] = \mathbf{0}$, where $\mathbf{u}_i = [u_{i,1983}, u_{i,1984}, u_{i,1985}, u_{i,1986}, u_{i,1987}]'$. The empirical moment equation is

$$\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{u}_i\right] = \text{plim}\ \overline{\mathbf{m}}(\boldsymbol{\theta}) = \mathbf{0}.$$

The parameters are vastly overidentified. Using only the lagged values of the dependent variable in each of the three equations estimated, there are 30 moment conditions and 14 parameters being estimated when $m = 3$, 11 when $m = 2$, 8 when $m = 1$, and 5 when $m = 0$. (As we do our estimation of each of these, we will retain the same matrix of instrumental variables in each case.) GMM estimation proceeds in two steps. In the first step, basic, unweighted instrumental variables is computed using

$$\hat{\boldsymbol{\theta}}_{IV}' = \left[\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{y}_i\right),$$

where

$$\mathbf{y}_i' = (\Delta S_{83} \quad \Delta S_{84} \quad \Delta S_{85} \quad \Delta S_{86} \quad \Delta S_{87}),$$

and

$$\mathbf{X}_i = \begin{bmatrix} \Delta S_{82} & \Delta S_{81} & \Delta S_{80} & \Delta R_{82} & \Delta R_{81} & \Delta R_{80} & \Delta G_{82} & \Delta G_{81} & \Delta G_{80} & 1 & 0 & 0 & 0 & 0 \\ \Delta S_{83} & \Delta S_{82} & \Delta S_{81} & \Delta R_{83} & \Delta R_{82} & \Delta R_{81} & \Delta G_{83} & \Delta G_{82} & \Delta G_{81} & 0 & 1 & 0 & 0 & 0 \\ \Delta S_{84} & \Delta S_{83} & \Delta S_{82} & \Delta R_{84} & \Delta R_{83} & \Delta R_{82} & \Delta G_{84} & \Delta G_{83} & \Delta G_{82} & 0 & 0 & 1 & 0 & 0 \\ \Delta S_{85} & \Delta S_{84} & \Delta S_{83} & \Delta R_{85} & \Delta R_{84} & \Delta R_{83} & \Delta G_{85} & \Delta G_{84} & \Delta G_{83} & 0 & 0 & 0 & 1 & 0 \\ \Delta S_{86} & \Delta S_{85} & \Delta S_{84} & \Delta R_{86} & \Delta R_{85} & \Delta R_{84} & \Delta G_{86} & \Delta G_{85} & \Delta G_{84} & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The second step begins with the computation of the new weighting matrix,

$$\hat{\mathbf{\Phi}} = \text{Est.Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{N}\sum_{i=1}^{n} \mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i.$$

After multiplying and dividing by the implicit (1/*n*) in the outside matrices, we obtain the estimator,

$$\begin{aligned} \boldsymbol{\theta}_{GMM}' &= \left[\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1} \\ &\quad \times \left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{y}_i\right) \\ &= \left[\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{Z}_i\right)\mathbf{W}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{Z}_i\right)\mathbf{W}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{y}_i\right). \end{aligned}$$

The estimator of the asymptotic covariance matrix for the estimator is the inverse matrix in square brackets in the first line of the result.

The primary focus of interest in the study was not the estimator itself, but the lag length and whether certain lagged values of the independent variables appeared in each equation. These restrictions would be tested by using the GMM criterion function, which in this formulation would be

$$nq = \left(\sum_{i=1}^{n}\hat{\mathbf{u}}_i'\mathbf{Z}_i\right)\mathbf{W}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\hat{\mathbf{u}}_i\right)$$

based on recomputing the residuals after GMM estimation. Note that the weighting matrix is not (necessarily) recomputed. For purposes of testing hypotheses, the same weighting matrix should be used.

At this point, we will consider the appropriate lag length, m. The specification can be reduced simply by redefining $\mathbf{X}$ to change the lag length. To test the specification, the weighting matrix must be kept constant for all restricted versions ($m = 2$ and $m = 1$) of the model.

The Dahlberg and Johansson data may be downloaded from the *Journal of Applied Econometrics* Web site—see Appendix Table F13.1. The authors provide the summary statistics for the raw data that are given in Table 13.1. Kroner, deflated by a municipality-specific price index, then converted to per capita values. Descriptive statistics for the raw data appear in Table 13.3.[27] Equations were estimated for all three variables, with maximum lag lengths of $m = 1$, 2, and 3. (The authors did not provide the actual estimates.) Estimation is done using the methods developed by Ahn and Schmidt (1995), Arellano and Bover (1995), and Holtz-Eakin, Newey, and Rosen (1988), as described. The estimates of the first specification provided are given in Table 13.4.

[27] The data provided on the Web site and used in our computations were further transformed by dividing by 100,000.

TABLE 13.3 Descriptive Statistics for Local Expenditure Data

Variable	*Mean*	*Std. Deviation*	*Minimum*	*Maximum*
Spending	18478.51	3174.36	12225.68	33883.25
Revenues	13422.56	3004.16	6228.54	29141.62
Grants	5236.03	1260.97	1570.64	12589.14

Table 13.5 contains estimates of the model parameters for each of the three equations, and for the three lag lengths, as well as the value of the GMM criterion function for each model estimated. The base case for each model has $m = 3$. There are three restrictions implied by each reduction in the lag length. The critical chi-squared value for three degrees of freedom is 7.81 for 95% significance, so at this level, we find that the two-level model is just barely accepted for the spending equation, but clearly appropriate for the other two—the difference between the two criteria is 7.62. Conditioned on $m = 2$, only the revenue model rejects the restriction of $m = 1$. As a final test, we might ask whether the data suggest that perhaps no lag structure at all is necessary. The GMM criterion value for the three equations with only the time dummy variables are 45.840, 57.908, and 62.042, respectively. Therefore, all three zero lag models are rejected.

Among the interests in this study were the appropriate critical values to use for the specification test of the moment restriction. With 16 degrees of freedom, the critical chi-squared value for 95% significance is 26.3, which would suggest that the revenues equation is misspecified. Using a bootstrap technique, the authors find that a more appropriate critical value leaves the specification intact. Finally, note that the three-equation model in the $m = 3$ columns of Table 13.5 imply a vector autoregression of the form

$$\mathbf{y}_t = \mathbf{\Gamma}_1\mathbf{y}_{t-1} + \mathbf{\Gamma}_2\mathbf{y}_{t-2} + \mathbf{\Gamma}_3\mathbf{y}_{t-3} + \mathbf{v}_t,$$

where $\mathbf{y}_t = (\Delta S_t, \Delta R_t, \Delta G_t)'$.

TABLE 13.4 Estimated Spending Equation

Variable	*Estimate*	*Standard Error*	*t Ratio*
Year 1983	−0.0036578	0.0002969	−12.32
Year 1984	−0.00049670	0.0004128	−1.20
Year 1985	0.00038085	0.0003094	1.23
Year 1986	0.00031469	0.0003282	0.96
Year 1987	0.00086878	0.0001480	5.87
Spending $(t - 1)$	1.15493	0.34409	3.36
Revenues $(t - 1)$	−1.23801	0.36171	−3.42
Grants $(t - 1)$	0.016310	0.82419	0.02
Spending $(t - 2)$	−0.0376625	0.22676	−0.17
Revenues $(t - 2)$	0.0770075	0.27179	0.28
Grants $(t - 2)$	1.55379	0.75841	2.05
Spending $(t - 3)$	−0.56441	0.21796	−2.59
Revenues $(t - 3)$	0.64978	0.26930	2.41
Grants $(t - 3)$	1.78918	0.69297	2.58

TABLE 13.5 Estimated Lag Equations for Spending, Revenue, and Grants

	Expenditure Model			*Revenue Model*			*Grant Model*		
	$m = 3$	$m = 2$	$m = 2$	$m = 3$	$m = 2$	$m = 1$	$m = 3$	$m = 2$	$m = 1$
S_{t-1}	1.155	0.8742	0.5562	−0.1715	−0.3117	−0.1242	−0.1675	−0.1461	−0.1958
S_{t-2}	−0.0377	0.2493	—	0.1621	0.0773	—	−0.0303	−0.0304	—
S_{t-3}	−0.5644	—	—	−0.1772	—	—	−0.0955	—	—
R_{t-1}	−0.2380	−0.8745	−0.5328	−0.0176	0.1863	−0.0245	0.1578	0.1453	0.2343
R_{t-2}	0.0770	−0.2776	—	−0.0309	0.1368	—	−0.0485	0.0175	—
R_{t-3}	0.6497	—	—	0.0034	—	—	0.0319	—	—
G_{t-1}	0.0163	−0.4203	0.1275	−0.3683	0.5425	0.0808	−0.2381	−0.2066	−0.0559
G_{t-2}	1.5538	0.1866	—	2.7152	2.4621	—	−0.0492	−0.0804	—
G_{t-3}	1.7892	—	—	0.0948	—	—	0.0598	—	—
nq	22.8287	30.4526	34.4986	30.5398	34.2590	53.2506	17.5810	20.5416	27.5927

13.7 SUMMARY AND CONCLUSIONS

The generalized method of moments provides an estimation framework that includes least squares, nonlinear least squares, instrumental variables, maximum likelihood, and a general class of estimators that extends beyond these. But it is more than just a theoretical umbrella. The GMM provides a method of formulating models and implied estimators without making strong distributional assumptions. Hall's model of household consumption is a useful example that shows how the optimization conditions of an underlying economic theory produce a set of distribution-free estimating equations. In this chapter, we first examined the classical method of moments. GMM as an estimator is an extension of this strategy that allows the analyst to use additional information beyond that necessary to identify the model, in an optimal fashion. After defining and establishing the properties of the estimator, we then turned to inference procedures. It is convenient that the GMM procedure provides counterparts to the familiar trio of test statistics: Wald, LM, and LR. In the final section, we specialized the GMM estimator for linear and nonlinear equations and multiple-equation models. We then developed an example that appears at many points in the recent applied literature, the dynamic panel data model with individual specific effects, and lagged values of the dependent variable.

Key Terms and Concepts

- Analog estimation
- Central limit theorem
- Criterion function
- Empirical moment equation
- Ergodic theorem
- Exactly identified cases
- Exponential family
- Generalized method of moments (GMM) estimator
- Instrumental variables
- Likelihood ratio statistic
- LM statistic
- Martingale difference series
- Maximum likelihood estimator
- Mean value theorem
- Method of moment generating functions
- Method of moments
- Method of moments estimators
- Minimum distance estimator (MDE)
- Moment equation
- Newey–West estimator
- Nonlinear instrumental variable estimator
- Order condition
- Orthogonality conditions

- Overidentified cases
- Overidentifying restrictions
- Population moment equation
- Probability limit
- Random sample
- Rank condition
- Slutsky theorem
- Specification test
- Sufficient statistic
- Taylor series
- Uncentered moment
- Wald statistic
- Weighted least squares
- Weighting matrix

Exercises

1. For the normal distribution $\mu_{2k} = \sigma^{2k}(2k)!/(k!2^k)$ and $\mu_{2k+1} = 0, k = 0, 1, \ldots$. Use this result to analyze the two estimators,

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} \quad \text{and} \quad b_2 = \frac{m_4}{m_2^2},$$

where $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$. The following result will be useful:

$$\text{Asy.Cov}[\sqrt{n}m_j, \sqrt{n}m_k] = \mu_{j+k} - \mu_j\mu_k + jk\mu_2\mu_{j-1}\mu_{k-1} - j\mu_{j-1}\mu_{k+1} - k\mu_{k-1}\mu_{j+1}.$$

Use the delta method to obtain the asymptotic variances and covariance of these two functions, assuming the data are drawn from a normal distribution with mean μ and variance σ^2. (*Hint:* Under the assumptions, the sample mean is a consistent estimator of μ, so for purposes of deriving asymptotic results, the difference between $\bar{x}$ and μ may be ignored. As such, no generality is lost by assuming the mean is zero, and proceeding from there.) Obtain $\mathbf{V}$, the 3×3 covariance matrix for the three moments, and then use the delta method to show that the covariance matrix for the two estimators is

$$\mathbf{JVJ}' = \begin{bmatrix} 6/n & 0 \\ 0 & 24/n \end{bmatrix},$$

where $\mathbf{J}$ is the 2×3 matrix of derivatives.

2. Using the results in Example 13.5, estimate the asymptotic covariance matrix of the method of moments estimators of P and λ based on m_1' and m_2'. [*Note:* You will need to use the data in Example C.1 to estimate $\mathbf{V}$.]
3. **Exponential Families of Distributions.** For each of the following distributions, determine whether it is an exponential family by examining the log-likelihood function. Then identify the sufficient statistics.
   a. Normal distribution with mean μ and variance σ^2.
   b. The Weibull distribution in Exercise 4 in Chapter 14.
   c. The mixture distribution in Exercise 3 in Chapter 14.
4. For the Wald distribution discussed in Example 13.3,

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}}\exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right], y > 0, \lambda > 0, \mu > 0,$$

we have the following results: $E[y] = \mu$, $\text{Var}[y] = \sigma^2 = \mu^3/\lambda$, $E[1/y] = 1/\mu + 1/\lambda$, $\text{Var}[1/y] = 1/(\lambda\mu) + 2/\lambda^2$, $E[y3] = \mu_3 = E[(y-\mu)^3/\sigma^3 = 3\mu^5/\lambda^2$.

   a. Derive the maximum likelihood estimators of μ and λ and an estimator of the asymptotic variances of the MLEs. (*Hint:* Expand the quadratic in the exponent and use the three terms in the derivation.)

b. Derive the method of moments estimators using the three different pairs of moments listed above, $E[y]$, $E[1/y]$ and $E[y3]$.

c. Using a random number generator, I generated a sample of 1,000 draws from the inverse Gaussian population with parameters μ and λ. I computed the following statistics:

	Mean	*Standard Deviation*
y	1.039892	1.438691
$1/y$	2.903571	2.976183
$y3 = (y - \mu)^3/\sigma^3$	4.158523	38.01372

[For the third variable, I used the known (to me) true values of the parameters.] Using the sample data, compute the maximum likelihood estimators of μ and λ and the estimates of the asymptotic standard errors. Compute the method of moments estimators using the means of $1/y$ and $y3$.

5. In the classical regression model with heteroscedasticity, which is more efficient, ordinary least squares or GMM? Obtain the two estimators and their respective asymptotic covariance matrices, then prove your assertion.

6. Consider the probit model analyzed in Chapter 17. The model states that for given vector of independent variables,

$$\text{Prob}[y_i = 1|\mathbf{x}_i] = \Phi[\mathbf{x}_i'\boldsymbol{\beta}], \quad \text{Prob}[y_i = 0|\mathbf{x}_i] = 1 - \text{Prob}[y_i = 1|\mathbf{x}_i].$$

Consider a GMM estimator based on the result that

$$E[y_i|\mathbf{x}_i] = \Phi(\mathbf{x}_i'\boldsymbol{\beta}).$$

This suggests that we might base estimation on the orthogonality conditions

$$E[(y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i] = \mathbf{0}.$$

Construct a GMM estimator based on these results. Note that this is not the nonlinear least squares estimator. Explain—what would the orthogonality conditions be for nonlinear least squares estimation of this model?

7. Consider GMM estimation of a regression model as shown at the beginning of Example 13.8. Let $\mathbf{W}_1$ be the optimal weighting matrix based on the moment equations. Let $\mathbf{W}_2$ be some other positive definite matrix. Compare the asymptotic covariance matrices of the two proposed estimators. Show conclusively that the asymptotic covariance matrix of the estimator based on $\mathbf{W}_1$ is not larger than that based on $\mathbf{W}_2$.

14
MAXIMUM LIKELIHOOD ESTIMATION

14.1 INTRODUCTION

The generalized method of moments discussed in Chapter 13 and the semiparametric, nonparametric, and Bayesian estimators discussed in Chapters 12 and 16 are becoming widely used by model builders. Nonetheless, the maximum likelihood estimator discussed in this chapter remains the preferred estimator in many more settings than the others listed. As such, we focus our discussion of generally applied estimation methods on this technique. Sections 14.2 through 14.6 present basic statistical results for estimation and hypothesis testing based on the maximum likelihood principle. Sections 14.7 and 14.8 present two extensions of the method, two-step estimation and pseudo maximum likelihood estimation. After establishing the general results for this method of estimation, we will then apply them to the more familiar setting of econometric models. The applications presented in Sections 14.9 and 14.10 apply the maximum likelihood method to most of the models in the preceding chapters and several others that illustrate different uses of the technique.

14.2 THE LIKELIHOOD FUNCTION AND IDENTIFICATION OF THE PARAMETERS

The probability density function, or pdf, for a random variable, y, conditioned on a set of parameters, $\boldsymbol{\theta}$, is denoted $f(y|\boldsymbol{\theta})$.[1] This function identifies the data-generating process that underlies an observed sample of data and, at the same time, provides a mathematical description of the data that the process will produce. The joint density of n *independent* and *identically distributed* (i.i.d.) observations from this process is the product of the individual densities,

$$f(y_1, \ldots, y_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y}). \tag{14-1}$$

This joint density is the likelihood function, defined as a function of the unknown parameter vector, $\boldsymbol{\theta}$, where $\mathbf{y}$ is used to indicate the collection of sample data. Note that we write the joint density as a function of the data conditioned on the parameters whereas when we form the likelihood function, we will write this function in reverse, as a function of the parameters, conditioned on the data. Though the two functions are the same, it is to be emphasized that the likelihood function is written in this fashion to highlight our interest in the parameters and the information about them that is contained in the

[1]Later we will extend this to the case of a random vector, $\mathbf{y}$, with a multivariate density, but at this point, that would complicate the notation without adding anything of substance to the discussion.

observed data. However, it is understood that the likelihood function is not meant to represent a probability density for the parameters as it is in Chapter 16. In this classical estimation framework, the parameters are assumed to be fixed constants that we hope to learn about from the data.

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}). \tag{14-2}$$

Again, to emphasize our interest in the parameters, given the observed data, we denote this function $L(\boldsymbol{\theta}|\mathbf{data}) = L(\boldsymbol{\theta}|\mathbf{y})$. The likelihood function and its logarithm, evaluated at $\boldsymbol{\theta}$, are sometimes denoted simply $L(\boldsymbol{\theta})$ and $\ln L(\boldsymbol{\theta})$, respectively, or, where no ambiguity can arise, just L or $\ln L$.

It will usually be necessary to generalize the concept of the likelihood function to allow the density to depend on other conditioning variables. To jump immediately to one of our central applications, suppose the disturbance in the classical linear regression model is normally distributed. Then, conditioned on its specific $\mathbf{x}_i$, y_i is normally distributed with mean $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ and variance σ^2. That means that the observed random variables are not i.i.d.; they have different means. Nonetheless, the observations are independent, and as we will examine in closer detail,

$$\ln L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}[\ln \sigma^2 + \ln(2\pi) + (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2/\sigma^2], \tag{14-3}$$

where $\mathbf{X}$ is the $n \times K$ matrix of data with ith row equal to $\mathbf{x}_i'$.

The rest of this chapter will be concerned with obtaining estimates of the parameters, $\boldsymbol{\theta}$, and testing hypotheses about them and about the data-generating process. Before we begin that study, we consider the question of whether estimation of the parameters is possible at all—the question of identification. Identification is an issue related to the formulation of the model. The issue of identification must be resolved before estimation can even be considered. The question posed is essentially this: Suppose we had an infinitely large sample—that is, for current purposes, all the information there is to be had about the parameters. Could we uniquely determine the values of $\boldsymbol{\theta}$ from such a sample? As will be clear shortly, the answer is sometimes no.

DEFINITION 14.1 Identification
*The parameter vector $\boldsymbol{\theta}$ is identified (**estimable**) if for any other parameter vector, $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, for some data $\mathbf{y}$, $L(\boldsymbol{\theta}^*|\mathbf{y}) \neq L(\boldsymbol{\theta}|\mathbf{y})$.*

This result will be crucial at several points in what follows. We consider two examples, the first of which will be very familiar to you by now.

Example 14.1 Identification of Parameters

For the regression model specified in (14-3), suppose that there is a nonzero vector $\mathbf{a}$ such that $\mathbf{x}_i'\mathbf{a} = 0$ for every $\mathbf{x}_i$. Then there is another parameter vector, $\boldsymbol{\gamma} = \boldsymbol{\beta} + \mathbf{a} \neq \boldsymbol{\beta}$ such that $\mathbf{x}_i'\boldsymbol{\beta} = \mathbf{x}_i'\boldsymbol{\gamma}$ for every $\mathbf{x}_i$. You can see in (14-3) that if this is the case, then the log-likelihood is the same whether it is evaluated at $\boldsymbol{\beta}$ or at $\boldsymbol{\gamma}$. As such, it is not possible to consider estimation

of $\boldsymbol{\beta}$ in this model because $\boldsymbol{\beta}$ cannot be distinguished from $\boldsymbol{\gamma}$. This is the case of perfect collinearity in the regression model, which we ruled out when we first proposed the linear regression model with "Assumption 2. Identifiability of the Model Parameters."

The preceding dealt with a necessary characteristic of the sample data. We now consider a model in which identification is secured by the specification of the parameters in the model. (We will study this model in detail in Chapter 17.) Consider a simple form of the regression model considered earlier, $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, where $\varepsilon_i | x_i$ has a normal distribution with zero mean and variance σ^2. To put the model in a context, consider a consumer's purchase of a large commodity such as a car where x_i is the consumer's income and y_i is the difference between what the consumer is willing to pay for the car, p_i^* (their reservation price) and the price tag on the car, p_i. Suppose rather than observing p_i^* or p_i, we observe only whether the consumer actually purchases the car, which, we assume, occurs when $y_i = p_i^* - p_i$ is positive. Collecting this information, our model states that they will purchase the car if $y_i > 0$ and not purchase it if $y_i \leq 0$. Let us form the likelihood function for the observed data, which are purchase (or not) and income. The random variable in this model is *purchase* or *not purchase*—there are only two outcomes. The probability of a purchase is

$$\begin{aligned}
\text{Prob(purchase} | \beta_1, \beta_2, \sigma, x_i) &= \text{Prob}(y_i > 0 | \beta_1, \beta_2, \sigma, x_i) \\
&= \text{Prob}(\beta_1 + \beta_2 x_i + \varepsilon_i > 0 | \beta_1, \beta_2, \sigma, x_i) \\
&= \text{Prob}[\varepsilon_i > -(\beta_1 + \beta_2 x_i) | \beta_1, \beta_2, \sigma, x_i] \\
&= \text{Prob}[\varepsilon_i/\sigma > -(\beta_1 + \beta_2 x_i)/\sigma | \beta_1, \beta_2, \sigma, x_i] \\
&= \text{Prob}[z_i > -(\beta_1 + \beta_2 x_i)/\sigma | \beta_1, \beta_2, \sigma, x_i],
\end{aligned}$$

where z_i has a standard normal distribution. The probability of not purchase is just one minus this probability. The likelihood function is

$$\prod_{i=\text{purchased}} [\text{Prob(purchase} | \beta_1, \beta_2, \sigma, x_i)] \prod_{i=\text{not purchased}} [1 - \text{Prob(purchase} | \beta_1, \beta_2, \sigma, x_i)].$$

We need go no further to see that the parameters of this model are not identified. If β_1, β_2, and σ are all multiplied by the same nonzero constant, regardless of what it is, then Prob(purchase) is unchanged, 1 − Prob(purchase) is also unchanged, and the likelihood function does not change. This model requires a normalization. The one usually used is $\sigma = 1$, but some authors have used $\beta_1 = 1$ or $\beta_2 = 1$, instead.[2]

14.3 EFFICIENT ESTIMATION: THE PRINCIPLE OF MAXIMUM LIKELIHOOD

The principle of **maximum likelihood** provides a means of choosing an asymptotically efficient estimator for a parameter or a set of parameters. The logic of the technique is easily illustrated in the setting of a discrete distribution. Consider a random sample of the following 10 observations from a Poisson distribution: 5, 0, 1, 1, 0, 3, 2, 3, 4, and 1. The density for each observation is

$$f(y_i | \theta) = \frac{e^{-\theta} \theta^{y_i}}{y_i!}.$$

[2]For examples, see Horowitz (1993) and Lewbel (2014).

Because the observations are independent, their joint density, which is the likelihood for this sample, is

$$f(y_1, y_2, \ldots, y_{10}|\theta) = \prod_{i=1}^{10} f(y_i|\theta) = \frac{e^{-10\theta}\theta^{\sum_{i=1}^{10} y_i}}{\prod_{i=1}^{10} y_i!} = \frac{e^{-10\theta}\theta^{20}}{207{,}360}.$$

The last result gives the probability of observing this particular sample, assuming that a Poisson distribution with as yet unknown parameter θ generated the data. What value of θ would make this sample most probable? Figure 14.1 plots this function for various values of θ. It has a single mode at $\theta = 2$, which would be the maximum likelihood estimate, or MLE, of θ.

Consider maximizing $L(\theta|\mathbf{y})$ with respect to θ. Because the log function is monotonically increasing and easier to work with, we usually maximize $\ln L(\theta|\mathbf{y})$ instead; in sampling from a Poisson population,

$$\ln L(\theta|\mathbf{y}) = -n\theta + \ln\theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!),$$

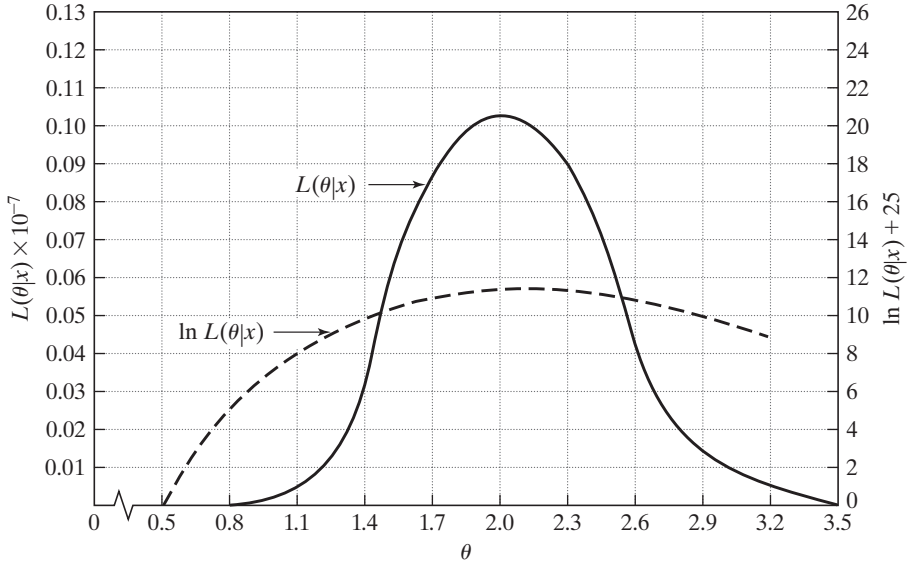
$$\frac{\partial \ln L(\theta|\mathbf{y})}{\partial\theta} = -n + \frac{1}{\theta}\sum_{i=1}^{n} y_i = 0 \Rightarrow \hat{\theta}_{ML} = \bar{y}_n.$$

For the assumed sample of observations,

$$\ln L(\theta|\mathbf{y}) = -10\theta + 20 \ln\theta - 12.242,$$

$$\frac{d \ln L(\theta|\mathbf{y})}{d\theta} = -10 + \frac{20}{\theta} = 0 \Rightarrow \hat{\theta} = 2,$$

FIGURE 14.1 Likelihood and Log-Likelihood Functions for a Poisson Distribution.



and

$$\frac{d^2 \ln L(\theta|\mathbf{y})}{d\theta^2} = \frac{-20}{\theta^2} < 0 \Rightarrow \text{this is a maximum.}$$

The solution is the same as before. Figure 14.1 also plots the log of $L(\theta|\mathbf{y})$ to illustrate the result.

The reference to the probability of observing the given sample is not exact in a continuous distribution, because a particular sample has probability zero. Nonetheless, the principle is the same. The values of the parameters that maximize $L(\boldsymbol{\theta}|\mathbf{data})$ or its log are the maximum likelihood estimates, denoted $\hat{\boldsymbol{\theta}}$. The logarithm is a monotonic function, so the values that maximize $L(\boldsymbol{\theta}|\mathbf{data})$ are the same as those that maximize $\ln L(\boldsymbol{\theta}|\mathbf{data})$. The necessary condition for maximizing $\ln L(\boldsymbol{\theta}|\mathbf{data})$ is

$$\frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{data})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{14-4}$$

This is called the **likelihood equation**. The general result then is that the MLE is a root of the likelihood equation. The application to the parameters of the data-generating process for a discrete random variable are suggestive that maximum likelihood is a good use of the data. It remains to establish this as a general principle. We turn to that issue in the next section.

Example 14.2 Log-Likelihood Function and Likelihood Equations for the Normal Distribution

In sampling from a normal distribution with mean μ and variance σ^2, the log-likelihood function and the likelihood equations for μ and σ^2 are

$$\ln L(\mu, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n}\left[\ln(2\pi) + \ln \sigma^2 + \frac{(y_i - \mu)^2}{\sigma^2}\right], \tag{14-5}$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0, \tag{14-6}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = 0. \tag{14-7}$$

To solve the likelihood equations, multiply (14-6) by σ^2 and solve for $\hat{\mu}$, then insert this solution in (14-7) and solve for σ^2. The solutions are

$$\hat{\mu}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_n \quad \text{and} \quad \hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2. \tag{14-8}$$

14.4 PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

Maximum likelihood estimators (MLEs) are most attractive because of their large-sample or asymptotic properties.

DEFINITION 14.2 Asymptotic Efficiency
An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed (CAN), and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.[3]

If certain regularity conditions are met, the MLE will have these properties. The finite sample properties are sometimes less than optimal. For example, the MLE may be biased; the MLE of σ^2 in Example 14.2 is biased downward. The occasional statement that the properties of the MLE are *only* optimal in large samples is not true, however. It can be shown that when sampling is from an exponential family of distributions (see Definition 13.1), there will exist sufficient statistics. If so, MLEs will be functions of them, which means that when minimum variance unbiased estimators exist, they will be MLEs.[4] Most applications in econometrics do not involve exponential families, so the appeal of the MLE remains primarily based on its asymptotic properties.

We use the following notation: $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator; $\boldsymbol{\theta}_0$ denotes the true value of the parameter vector; $\boldsymbol{\theta}$ denotes another possible value of the parameter vector, not the MLE and not necessarily the true values. Expectation based on the true values of the parameters is denoted $E_0[.]$. If we assume that the regularity conditions discussed momentarily are met by $f(\mathbf{x}, \boldsymbol{\theta}_0)$, then we have the following theorem.

THEOREM 14.1 Properties of an MLE
Under regularity, the MLE has the following asymptotic properties:

M1. ***Consistency:*** $\operatorname{plim} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

M2. ***Asymptotic normality:*** $\hat{\boldsymbol{\theta}} \overset{a}{\sim} N[\boldsymbol{\theta}_0, \{\mathbf{I}(\boldsymbol{\theta}_0)\}^{-1}]$, *where*

$$\mathbf{I}(\boldsymbol{\theta}_0) = -E_0[\partial^2 \ln L/\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0'].$$

M3. ***Asymptotic efficiency:*** $\hat{\boldsymbol{\theta}}$ *is asymptotically efficient and achieves the Cramér–Rao lower bound for consistent estimators, given in M2 and Theorem C.2.*

M4. ***Invariance:*** *The maximum likelihood estimator of* $\boldsymbol{\gamma}_0 = \mathbf{c}(\boldsymbol{\theta}_0)$ *is* $\mathbf{c}(\hat{\boldsymbol{\theta}})$ *if* $\mathbf{c}(\boldsymbol{\theta}_0)$ *is a continuous and continuously differentiable function.*

14.4.1 REGULARITY CONDITIONS

To sketch proofs of these results, we first obtain some useful properties of probability density functions. We assume that $(y_1, \ldots, y_n)$ is a random sample from the population with density function $f(y_i|\boldsymbol{\theta}_0)$ and that the following **regularity conditions** hold.[5]

[3] *Not larger* is defined in the sense of (A-118): The covariance matrix of the less efficient estimator equals that of the efficient estimator plus a nonnegative definite matrix.

[4] See Stuart and Ord (1989).

[5] Our statement of these is informal. A more rigorous treatment may be found in Stuart and Ord (1989) or Davidson and MacKinnon (2004).

DEFINITION 14.3 Regularity Conditions

R1. *The first three derivatives of* $\ln f(y_i|\boldsymbol{\theta})$ *with respect to* $\boldsymbol{\theta}$ *are continuous and finite for almost all* y_i *and for all* $\boldsymbol{\theta}$*. This condition ensures the existence of a certain Taylor series approximation to and the finite variance of the derivatives of* $\ln L$*.*

R2. *The conditions necessary to obtain the expectations of the first and second derivatives of* $\ln f(y_i|\boldsymbol{\theta})$ *are met.*

R3. *For all values of* $\boldsymbol{\theta}$, $|\partial^3 \ln f(y_i|\boldsymbol{\theta})/\partial\theta_j\partial\theta_k\partial\theta_l|$ *is less than a function that has a finite expectation. This condition will allow us to truncate the Taylor series.*

With these regularity conditions, we will obtain the following fundamental characteristics of $f(y_i|\boldsymbol{\theta})$: D1 is simply a consequence of the definition of the likelihood function. D2 leads to the moment condition which defines the maximum likelihood estimator. On the one hand, the MLE is found as the maximizer of a function, which mandates finding the vector that equates the gradient to zero. On the other hand, D2 is a more fundamental relationship that places the MLE in the class of generalized method of moments estimators. D3 produces what is known as the **information matrix equality**. This relationship shows how to obtain the asymptotic covariance matrix of the MLE.

14.4.2 PROPERTIES OF REGULAR DENSITIES

Densities that are *regular* by Definition 14.3 have three properties that are used in establishing the properties of maximum likelihood estimators:

THEOREM 14.2 Moments of the Derivatives of the Log Likelihood

D1. $\ln f(y_i|\boldsymbol{\theta})$, $\mathbf{g}_i = \partial \ln f(y_i|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, *and* $\mathbf{H}_i = \partial^2 \ln f(y_i|\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, $i = 1, \ldots, n$, *are all random samples of random variables. This statement follows from our assumption of random sampling. The notation* $\mathbf{g}_i(\boldsymbol{\theta}_0)$ *and* $\mathbf{H}_i(\boldsymbol{\theta}_0)$ *indicates the derivative evaluated at* $\boldsymbol{\theta}_0$*. Condition D1 is simply a consequence of the definition of the density.*

D2. $E_0[\mathbf{g}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$.

D3. $\text{Var}[\mathbf{g}_i(\boldsymbol{\theta}_0)] = -E[\mathbf{H}_i(\boldsymbol{\theta}_0)]$.

For the moment, we allow the range of y_i to depend on the parameters; $A(\boldsymbol{\theta}_0) \leq y_i \leq B(\boldsymbol{\theta}_0)$. (Consider, for example, finding the maximum likelihood estimator of θ_0 for a continuous uniform distribution with range $[0, \theta_0]$.) (In the following, the single integral $\int \ldots dy_i$ will be used to indicate the multiple integration over all the elements of a multivariate of y_i if that is necessary.) By definition,

$$\int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} f(y_i|\boldsymbol{\theta}_0)\, dy_i = 1.$$

Now, differentiate this expression with respect to $\boldsymbol{\theta}_0$. Leibnitz's theorem gives

$$\frac{\partial \int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} f(y_i|\boldsymbol{\theta}_0)\, dy_i}{\partial \boldsymbol{\theta}_0} = \int_{A(\boldsymbol{\theta}_0)}^{B(\boldsymbol{\theta}_0)} \frac{\partial f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\, dy_i + f(B(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)\frac{\partial B(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} - f(A(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)\frac{\partial A(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} = \mathbf{0}.$$

If the second and third terms go to zero, then we may interchange the operations of differentiation and integration. The necessary condition is that $\lim_{y_i \downarrow A(\boldsymbol{\theta}_0)} f(y_i|\boldsymbol{\theta}_0) = \lim_{y_i \uparrow B(\boldsymbol{\theta}_0)} f(y_i|\boldsymbol{\theta}_0) = 0$. (*Note:* The uniform distribution suggested earlier violates this condition.) Sufficient conditions are that the range of the observed random variable, y_i, does not depend on the parameters, which means that $\partial A(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}_0 = \partial B(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}_0 = \mathbf{0}$ or that the density is zero at the terminal points. This condition, then, is regularity condition R2. The latter is usually assumed, and we will assume it in what follows. So,

$$\begin{aligned}\frac{\partial \int f(y_i|\boldsymbol{\theta}_0)\, dy_i}{\partial \boldsymbol{\theta}_0} &= \int \frac{\partial f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\, dy_i = \int \frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} f(y_i|\boldsymbol{\theta}_0)\, dy_i \\ &= E_0\left[\frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right] = \mathbf{0}.\end{aligned}$$

This proves D2.

Because we may interchange the operations of integration and differentiation, we differentiate under the integral once again to obtain

$$\int \left[\frac{\partial^2 \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} f(y_i|\boldsymbol{\theta}_0) + \frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\frac{\partial f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\right] dy_i = \mathbf{0}.$$

But

$$\frac{\partial f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} = f(y_i|\boldsymbol{\theta}_0)\frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'},$$

and the integral of a sum is the sum of integrals. Therefore,

$$-\int \left[\frac{\partial^2 \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right] f(y_i|\boldsymbol{\theta}_0) dy_i = \int \left[\frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\right] f(y_i|\boldsymbol{\theta}_0) dy_i.$$

The left-hand side of the equation is the negative of the expected second derivatives matrix. The right-hand side is the expected square (outer product) of the first derivative vector. But because this vector has expected value $\mathbf{0}$ (we just showed this), the right-hand side is the variance of the first derivative vector, which proves D3,

$$\text{Var}_0\left[\frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right] = E_0\left[\left(\frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\right)\right] = -E\left[\frac{\partial^2 \ln f(y_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right].$$

14.4.3 THE LIKELIHOOD EQUATION

The log-likelihood function is

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}).$$

The first derivative vector, or **score vector**, is

$$\mathbf{g} = \frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{g}_i. \tag{14-9}$$

Because we are just adding terms, it follows from D1 and D2 that at $\boldsymbol{\theta}_0$,

$$E_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] = E_0[\mathbf{g}_0] = \mathbf{0}, \tag{14-10}$$

which is the **likelihood equation** mentioned earlier.

14.4.4 THE INFORMATION MATRIX EQUALITY

The Hessian of the log likelihood is

$$\mathbf{H} = \frac{\partial^2 \ln L(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = \sum_{i=1}^{n} \mathbf{H}_i.$$

Evaluating once again at $\boldsymbol{\theta}_0$, by taking

$$E_0[\mathbf{g}_0\mathbf{g}_0'] = E_0\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{g}_{0i}\mathbf{g}_{0j}'\right],$$

and, because of D1, dropping terms with unequal subscripts, we obtain

$$E_0[\mathbf{g}_0\mathbf{g}_0'] = E_0\left[\sum_{i=1}^{n} \mathbf{g}_{0i}\mathbf{g}_{0i}'\right] = E_0\left[\sum_{i=1}^{n} (-\mathbf{H}_{0i})\right] = -E_0[\mathbf{H}_0],$$

so that

$$\mathrm{Var}_0\left[\frac{\partial \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0}\right] = E_0\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0'}\right)\right] = -E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0|\mathbf{y})}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right]. \tag{14-11}$$

This very useful result is known as the **information matrix equality**. It states that the variance of the first derivative of ln L equals the negative of the second derivative.

14.4.5 ASYMPTOTIC PROPERTIES OF THE MAXIMUM LIKELIHOOD ESTIMATOR

We can now sketch a derivation of the asymptotic properties of the MLE. Formal proofs of these results require some fairly intricate mathematics. Two widely cited derivations are those of Cramér (1948) and Amemiya (1985). To suggest the flavor of the exercise, we will sketch an analysis provided by Stuart and Ord (1989) for a simple case, and indicate where it will be necessary to extend the derivation if it were to be fully general.

14.4.5.a Consistency

We assume that $f(y_i|\boldsymbol{\theta}_0)$ is a possibly multivariate density that at this point does not depend on covariates, $\mathbf{x}_i$. Thus, this is the i.i.d., random sampling case. Because $\hat{\boldsymbol{\theta}}$ is the MLE, in any finite sample, for any $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}$ (including the true $\boldsymbol{\theta}_0$) it must be true that

$$\ln L(\hat{\boldsymbol{\theta}}) \geq \ln L(\boldsymbol{\theta}). \tag{14-12}$$

Consider, then, the random variable $L(\boldsymbol{\theta})/L(\boldsymbol{\theta}_0)$. Because the log function is strictly concave, from Jensen's Inequality (Theorem D.13.), we have

$$E_0\left[\ln\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}\right] < \ln E_0\left[\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}\right]. \tag{14-13}$$

The expectation on the right-hand side is exactly equal to one, as

$$E_0\left[\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}\right] = \int\left(\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}\right)L(\boldsymbol{\theta}_0)d\mathbf{y} = 1 \tag{14-14}$$

is simply the integral of a joint density. So, the right-hand side of (14-13) equals zero. Divide the left-hand side of (14-13) by n to produce

$$E_0[1/n \ln L(\boldsymbol{\theta})] - E_0[1/n \ln L(\boldsymbol{\theta}_0)] < 0.$$

This produces a central result:

THEOREM 14.3 Likelihood Inequality

$E_0[(1/n) \ln L(\boldsymbol{\theta}_0)] > E_0[(1/n) \ln L(\boldsymbol{\theta})]$ *for any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ (*including* $\hat{\boldsymbol{\theta}}$).

In words, *the expected value of the log likelihood is maximized at the true value of the parameters*.

For any $\boldsymbol{\theta}$, including $\hat{\boldsymbol{\theta}}$,

$$[(1/n) \ln L(\boldsymbol{\theta})] = (1/n)\sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta})$$

is the sample mean of n i.i.d. random variables, with expectation $E_0[(1/n) \ln L(\boldsymbol{\theta})]$. Because the sampling is i.i.d. by the regularity conditions, we can invoke the Khinchine theorem, D.5; the sample mean converges in probability to the population mean. Using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, it follows from Theorem 14.3 that as $n \to \infty$, $\lim \text{Prob}\{[(1/n) \ln L(\hat{\boldsymbol{\theta}})] < [(1/n) \ln L(\boldsymbol{\theta}_0)]\} = 1$ if $\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_0$. But $\hat{\boldsymbol{\theta}}$ is the MLE, so for every n, $(1/n) \ln L(\hat{\boldsymbol{\theta}}) \geq (1/n) \ln L(\boldsymbol{\theta}_0)$. The only way these can both be true is if $(1/n)$ times the sample log likelihood evaluated at the MLE converges to the population expectation of $(1/n)$ times the log likelihood evaluated at the true parameters. There remains one final step. Does $(1/n) \ln L(\hat{\boldsymbol{\theta}}) \to (1/n) \ln L(\boldsymbol{\theta}_0)$ imply that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$? If there is a single parameter and the likelihood function is one to one, then clearly so. For more general cases, this requires a further characterization of the likelihood function. If the likelihood is strictly continuous and twice differentiable, which we assumed in the regularity conditions, and if the parameters of the model are identified, which we assumed at the beginning of this discussion, then yes, it does, so we have the result.

This is a heuristic proof. As noted, formal presentations appear in more advanced treatises than this one. We should also note we have assumed at several points that sample means converge to their population expectations. This is likely to be true for the sorts of applications usually encountered in econometrics, but a fully general set of results would look more closely at this condition. Second, we have assumed i.i.d. sampling in the preceding—that is, the density for $\mathbf{y}_i$ does not depend on any other variables, $\mathbf{x}_i$. This will almost never be true in practice. Assumptions about the behavior

of these variables will enter the proofs as well. For example, in assessing the large sample behavior of the least squares estimator, we have invoked an assumption that the data are well behaved. The same sort of consideration will apply here as well. We will return to this issue shortly. With all this in place, we have property M1, plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

14.4.5.b Asymptotic Normality

At the maximum likelihood estimator, the gradient of the log likelihood equals zero (by definition), so $\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. (This is the sample statistic, not the expectation.) Expand this set of equations in a Taylor series around the true parameters $\boldsymbol{\theta}_0$. We will use the mean value theorem to truncate the Taylor series for each element of $\mathbf{g}(\hat{\boldsymbol{\theta}})$ at the second term,

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{g}(\boldsymbol{\theta}_0) + \mathbf{H}(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}.$$

The K rows of the Hessian are each evaluated at a point $\bar{\boldsymbol{\theta}}_k$ that is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ [$\bar{\boldsymbol{\theta}}_k = w_k\hat{\boldsymbol{\theta}} + (1 - w_k)\boldsymbol{\theta}_0$ for some $0 < w_k < 1$]. (Although the vectors $\bar{\boldsymbol{\theta}}_k$ are different, they all converge to $\boldsymbol{\theta}_0$.) We then rearrange this function and multiply the result by $\sqrt{n}$ to obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = [-\mathbf{H}(\bar{\boldsymbol{\theta}})]^{-1}[\sqrt{n}\mathbf{g}(\boldsymbol{\theta}_0)].$$

Because $\text{plim}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}$, $\text{plim}(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) = \mathbf{0}$ as well. The second derivatives are continuous functions. Therefore, if the limiting distribution exists, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} [-\mathbf{H}(\boldsymbol{\theta}_0)]^{-1}[\sqrt{n}\mathbf{g}(\boldsymbol{\theta}_0)].$$

By dividing $\mathbf{H}(\boldsymbol{\theta}_0)$ and $\mathbf{g}(\boldsymbol{\theta}_0)$ by n, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \left[-\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]^{-1}[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)]. \tag{14-15}$$

We may apply the Lindeberg–Levy central limit theorem (D.18) to $[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)]$, because it is $\sqrt{n}$ times the mean of a random sample; we have invoked D1 again. The limiting variance of

$$[\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0)] \text{ is } -E_0[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)], \text{ so}$$

$$\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0) \xrightarrow{d} N\left\{\mathbf{0}, -E_0\left[\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}.$$

By virtue of Theorem D.2, $\text{plim}[-(1/n)\mathbf{H}(\boldsymbol{\theta}_0)] = -E_0[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)]$. This result is a constant matrix, so we can combine results to obtain

$$\left[-\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]^{-1}\sqrt{n}\bar{\mathbf{g}}(\boldsymbol{\theta}_0) \xrightarrow{d} N\left[\mathbf{0}, \left\{-E_0\left[\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\left\{-E_0\left[\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}\left\{-E_0\left[\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\right],$$

or

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\left[\mathbf{0}, \left\{-E_0\left[\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}_0)\right]\right\}^{-1}\right],$$

which gives the asymptotic distribution of the MLE,

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N[\boldsymbol{\theta}_0, \{\mathbf{I}(\boldsymbol{\theta}_0)\}^{-1}].$$

This last step completes M2.

Example 14.3 Information Matrix for the Normal Distribution

For the likelihood function in Example 14.2, the second derivatives are

$$\frac{\partial^2 \ln L}{\partial \mu^2} = \frac{-n}{\sigma^2},$$
$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2,$$
$$\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} = \frac{-1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu).$$

For the **asymptotic variance** of the maximum likelihood estimator, we need the expectations of these derivatives. The first is nonstochastic, and the third has expectation 0, as $E[y_i] = \mu$. That leaves the second, which you can verify has expectation $-n/(2\sigma^4)$ because each of the n terms $(y_i - \mu)^2$ has expected value σ^2. Collecting these in the information matrix, reversing the sign, and inverting the matrix gives the asymptotic covariance matrix for the maximum likelihood estimators,

$$\left\{-E_0\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}_0 \, \partial \boldsymbol{\theta}_0'}\right]\right\}^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}.$$

14.4.5.c Asymptotic Efficiency

Theorem C.2 provides the lower bound for the variance of an unbiased estimator. Because the asymptotic variance of the MLE achieves this bound, it seems natural to extend the result directly. There is, however, a loose end in that the MLE is almost never unbiased. As such, we need an asymptotic version of the bound, which was provided by Cramér (1948) and Rao (1945) (hence the name):

THEOREM 14.4 Cramér–Rao Lower Bound
Assuming that the density of y_i satisfies the regularity conditions R1–R3, the asymptotic variance of a consistent and asymptotically normally distributed estimator of the parameter vector $\boldsymbol{\theta}_0$ will always be at least as large as

$$[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = \left(-E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \, \partial \boldsymbol{\theta}_0'}\right]\right)^{-1} = \left(E_0\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\right)'\right]\right)^{-1}.$$

The asymptotic variance of the MLE is, in fact, equal to the Cramér–Rao lower bound for the variance of a consistent, asymptotically normally distributed estimator, so this completes the argument.[6]

14.4.5.d Invariance

Last, the invariance property, M4, is a mathematical result of the method of computing MLEs; it is not a statistical result as such. More formally, the MLE is invariant to

[6]A result reported by LeCam (1953) and recounted in Amemiya (1985, p. 124) suggests that, in principle, there do exist CAN functions of the data with smaller variances than the MLE. But the finding is a narrow result with no practical implications. For practical purposes, the statement may be taken as given.

one-to-one transformations of $\boldsymbol{\theta}$. Any transformation that is not one to one either renders the model inestimable if it is one to many or imposes restrictions if it is many to one. Some theoretical aspects of this feature are discussed in Davidson and MacKinnon (2004, pp. 446, 539–540). For the practitioner, the result can be extremely useful. For example, when a parameter appears in a likelihood function in the form $1/\theta_j$, it is usually worthwhile to reparameterize the model in terms of $\gamma_j = 1/\theta_j$. In an important application, Olsen (1978) used this result to great advantage. (See Section 19.3.3.) Suppose that the normal log likelihood in Example 14.2 is parameterized in terms of the **precision parameter**, $\theta^2 = 1/\sigma^2$. The log likelihood becomes

$$\ln L(\mu, \theta^2) = -(n/2)\ln(2\pi) + (n/2)\ln\theta^2 - \frac{\theta^2}{2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

The MLE for μ is clearly still $\bar{x}$. But the likelihood equation for θ^2 is now

$$\partial \ln L(\mu, \theta^2)/\partial\theta^2 = \frac{1}{2}\left[n/\theta^2 - \sum_{i=1}^{n}(y_i - \mu)^2 \right] = 0,$$

which has solution $\hat{\theta}^2 = n/\sum_{i=1}^{n}(y_i - \hat{\mu})^2 = 1/\hat{\sigma}^2$, as expected. There is a second implication. If it is desired to analyze a function of an MLE, then the function of $\hat{\boldsymbol{\theta}}$ will, itself, be the MLE.

14.4.5.e Conclusion

These four properties explain the prevalence of the maximum likelihood technique in econometrics. The second greatly facilitates hypothesis testing and the construction of interval estimates. The third is a particularly powerful result. The MLE has the minimum variance achievable by a consistent and asymptotically normally distributed estimator.

14.4.6 ESTIMATING THE ASYMPTOTIC VARIANCE OF THE MAXIMUM LIKELIHOOD ESTIMATOR

The asymptotic covariance matrix of the maximum likelihood estimator is a matrix of parameters that must be estimated (i.e., it is a function of the $\boldsymbol{\theta}_0$ that is being estimated). If the form of the expected values of the second derivatives of the log likelihood is known, then

$$[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = \left\{ -E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0\,\partial\boldsymbol{\theta}_0'} \right] \right\}^{-1} \tag{14-16}$$

can be evaluated at $\hat{\boldsymbol{\theta}}$ to estimate the covariance matrix for the MLE. This estimator will rarely be available. The second derivatives of the log likelihood will almost always be complicated nonlinear functions of the data whose exact expected values will be unknown. There are, however, two alternatives. A second estimator is

$$[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})]^{-1} = \left(-\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\,\partial\hat{\boldsymbol{\theta}}'} \right)^{-1}. \tag{14-17}$$

This estimator is computed simply by evaluating the actual (not expected) second derivatives matrix of the log-likelihood function at the maximum likelihood estimates. It is straightforward to show that this amounts to estimating the expected second derivatives

of the density with the sample mean of this quantity. Theorem D.4 and Result (D-5) can be used to justify the computation. The only shortcoming of this estimator is that the second derivatives can be complicated to derive and program for a computer. A third estimator based on result D3 in Theorem 14.2, that the expected second derivatives matrix is the covariance matrix of the first derivatives vector, is

$$[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})]^{-1} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'\right]^{-1} = [\hat{\mathbf{G}}'\hat{\mathbf{G}}]^{-1}, \tag{14-18}$$

where $\hat{\mathbf{g}}_i = \dfrac{\partial \ln f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}}$, and $\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_n]'$ is an $n \times K$ matrix with ith row equal to the transpose of the ith vector of derivatives in the terms of the log-likelihood function. For a single parameter, this estimator is just the reciprocal of the sum of squares of the first derivatives. This estimator is extremely convenient, in most cases, because it does not require any computations beyond those required to solve the likelihood equation. It has the added virtue that it is always nonnegative definite. For some extremely complicated log-likelihood functions, sometimes because of rounding error, the *observed* Hessian can be indefinite, even at the maximum of the function. The estimator in (14-18) is known as the **BHHH estimator**[7] and the **outer product of gradients estimator (OPG)**.

None of the three estimators given here is preferable to the others on statistical grounds; all are asymptotically equivalent. In most cases, the BHHH estimator will be the easiest to compute. One caution is in order. As the following example illustrates, these estimators can give different results in a finite sample. This is an unavoidable finite sample problem that can, in some cases, lead to different statistical conclusions. The example is a case in point. Using the usual procedures, we would reject the hypothesis that $\beta = 0$ if either of the first two variance estimators were used, but not if the third were used. The estimator in (14-16) is usually unavailable, as the exact expectation of the Hessian is rarely known. Available evidence suggests that in small or moderate-sized samples, (14-17) (the Hessian) is preferable.

Example 14.4 Variance Estimators for an MLE

The sample data in Example C.1 are generated by a model of the form

$$f(y_i, x_i/\beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)},$$

where y = income and x = education. To find the maximum likelihood estimate of β, we maximize

$$\ln L(\beta) = -\sum_{i=1}^{n} \ln(\beta + x_i) - \sum_{i=1}^{n} \frac{y_i}{\beta + x_i}.$$

The likelihood equation is

$$\frac{\partial \ln L(\beta)}{\partial \beta} = -\sum_{i=1}^{n} \frac{1}{\beta + x_i} + \sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^2} = 0, \tag{14-19}$$

[7] It appears to have been advocated first in the econometrics literature in Berndt et al. (1974).

which has the solution $\hat{\beta} = 15.602727$. To compute the asymptotic variance of the MLE, we require

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta^2} = \sum_{i=1}^{n} \frac{1}{(\beta + x_i)^2} - 2\sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^3}. \tag{14-20}$$

Because the function $E(y_i) = \beta + x_i$ is known, the exact form of the expected value in (14-20) is known. Inserting $\hat{\beta} + x_i$ for y_i in (14-20) and taking the negative of the reciprocal yields the first variance estimate, 44.2546. Simply inserting $\hat{\beta} = 15.602727$ in (14-20) and taking the negative of the reciprocal gives the second estimate, 46.16337. Finally, by computing the reciprocal of the sum of squares of first derivatives of the densities evaluated at $\hat{\beta}$,

$$\widehat{[\boldsymbol{I}(\hat{\boldsymbol{\beta}})]}^{-1} = \frac{1}{\sum_{i=1}^{n}[-1/(\hat{\beta} + x_i) + y_i/(\hat{\beta} + x_i)^2]^2},$$

we obtain the BHHH estimate, 100.5116.

14.5 CONDITIONAL LIKELIHOODS AND ECONOMETRIC MODELS

All of the preceding results form the statistical underpinnings of the technique of maximum likelihood estimation. But, for our purposes, a crucial element is missing. We have done the analysis in terms of the density of an observed random variable and a vector of parameters, $f(y_i|\boldsymbol{\alpha})$. But econometric models will involve exogenous or predetermined variables, $\mathbf{x}_i$, so the results must be extended. A workable approach is to treat this modeling framework the same as the one in Chapter 4, where we considered the large sample properties of the linear regression model. Thus, we will allow $\mathbf{x}_i$ to denote a mix of random variables and constants that enter the conditional density of y_i. By partitioning the joint density of y_i and $\mathbf{x}_i$ into the product of the conditional and the marginal, the log-likelihood function may be written

$$\ln L(\boldsymbol{\alpha}|\mathbf{data}) = \sum_{i=1}^{n} \ln f(y_i, \mathbf{x}_i|\boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\alpha}) + \sum_{i=1}^{n} \ln g(\mathbf{x}_i|\boldsymbol{\alpha}),$$

where any nonstochastic elements in $\mathbf{x}_i$ such as a time trend or dummy variable are being carried as constants. To proceed, we will assume as we did before that the process generating $\mathbf{x}_i$ takes place outside the model of interest. For present purposes, that means that the parameters that appear in $g(\mathbf{x}_i|\boldsymbol{\alpha})$ do not overlap with those that appear in $f(y_i|\mathbf{x}_i, \boldsymbol{\alpha})$. Thus, we partition $\boldsymbol{\alpha}$ into $[\boldsymbol{\theta}, \boldsymbol{\delta}]$ so that the log-likelihood function may be written

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\delta}|\mathbf{data}) = \sum_{i=1}^{n} \ln f(y_i, \mathbf{x}_i|\boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} \ln g(\mathbf{x}_i|\boldsymbol{\delta}).$$

As long as $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ have no elements in common and no restrictions connect them (such as $\theta + \delta = 1$), then the two parts of the log likelihood may be analyzed separately. In most cases, the marginal distribution of $\mathbf{x}_i$ will be of secondary (or no) interest.

Asymptotic results for the maximum conditional likelihood estimator must now account for the presence of $\mathbf{x}_i$ in the functions and derivatives of $\ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$. We will proceed under the assumption of well-behaved data so that sample averages such as

$$(1/n) \ln L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

and its gradient with respect to $\boldsymbol{\theta}$ will converge in probability to their population expectations. We will also need to invoke central limit theorems to establish the asymptotic normality of the gradient of the log likelihood, so as to be able to characterize the MLE itself. We will leave it to more advanced treatises such as Amemiya (1985) and Newey and McFadden (1994) to establish specific conditions and fine points that must be assumed to claim the "usual" properties for maximum likelihood estimators. For present purposes (and the vast bulk of empirical applications), the following minimal assumptions should suffice:

- **Parameter space.** Parameter spaces that have gaps and nonconvexities in them will generally disable these procedures. An estimation problem that produces this failure is that of "estimating" a parameter that can take only one among a discrete set of values. For example, this set of procedures does not include "estimating" the timing of a structural change in a model. The likelihood function must be a continuous function of a convex parameter space. We allow unbounded parameter spaces, such as $\sigma > 0$ in the regression model, for example.
- **Identifiability.** Estimation must be feasible. This is the subject of Definition 14.1 concerning identification and the surrounding discussion.
- **Well-behaved data.** Laws of large numbers apply to sample means involving the data and some form of central limit theorem (generally Lyapounov) can be applied to the gradient. Ergodic stationarity is broad enough to encompass any situation that is likely to arise in practice, though it is probably more general than we need for most applications, because we will not encounter dependent observations specifically until later in the book. The definitions in Chapter 4 are assumed to hold generally.

With these in place, analysis is essentially the same in character as that we used in the linear regression model in Chapter 4 and follows precisely along the lines of Section 12.5.

14.6 HYPOTHESIS AND SPECIFICATION TESTS AND FIT MEASURES

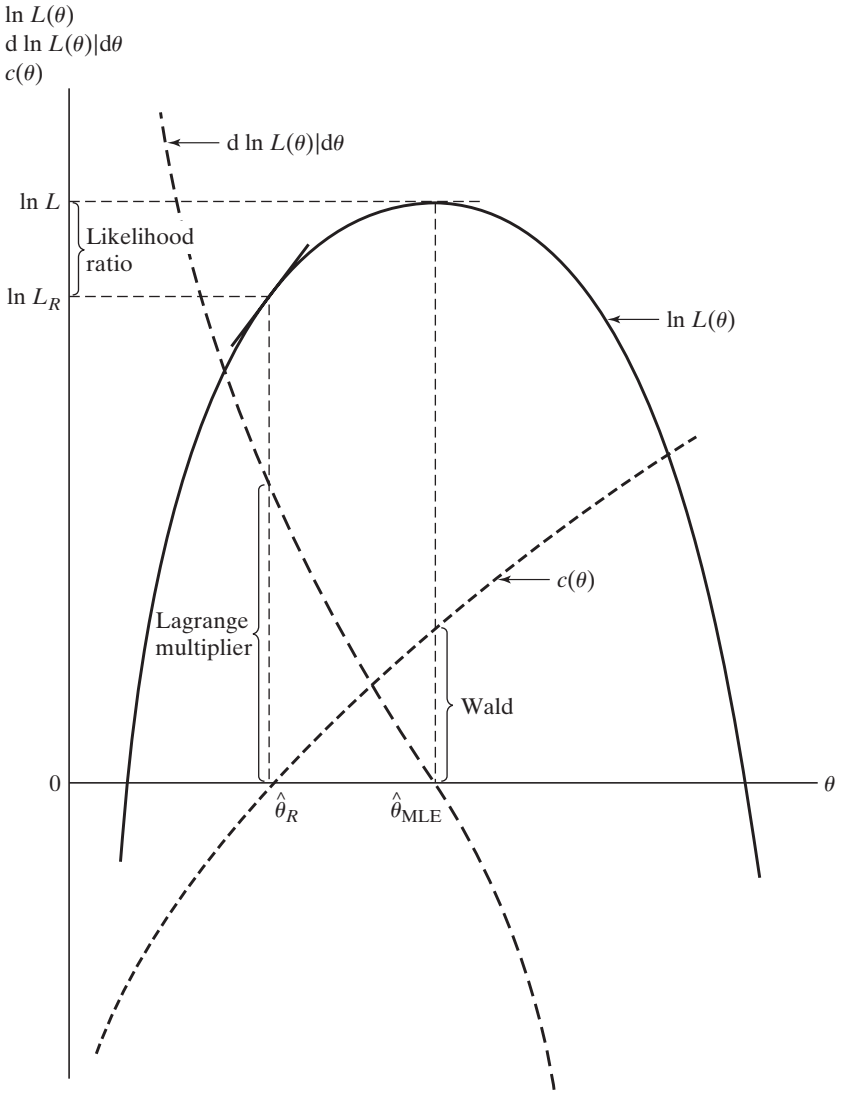
The next several sections will discuss the most commonly used test procedures: the likelihood ratio, Wald, and Lagrange multiplier tests.[8] We consider maximum likelihood estimation of a parameter θ and a test of the hypothesis H_0: $c(\theta) = 0$. The logic of the tests can be seen in Figure 14.2.[9] The figure plots the log-likelihood function $\ln L(\theta)$, its derivative with respect to θ, $d \ln L(\theta)/d\theta$, and the constraint $c(\theta)$. There are three approaches to testing the hypothesis suggested in the figure:

- **Likelihood ratio test.** If the restriction $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference, $\ln L_U - \ln L_R$, where L_U is the value of the likelihood function at the unconstrained value of θ and L_R is the value of the likelihood function at the restricted estimate.

[8] Extensive discussion of these procedures is given in Godfrey (1988).

[9] See Buse (1982). Note that the scale of the vertical axis would be different for each curve. As such, the points of intersection have no significance.

FIGURE 14.2 Three Bases for Hypothesis Tests.



- **Wald test.** If the restriction is valid, then $c(\hat{\theta}_{\text{MLE}})$ should be close to zero because the MLE is consistent. Therefore, the test is based on $c(\hat{\theta}_{\text{MLE}})$. We reject the hypothesis if this value is significantly different from zero.
- **Lagrange multiplier test.** If the restriction is valid, then the restricted estimator should be near the point that maximizes the log likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log likelihood at the point where the function is maximized subject to the restriction.

These three tests are asymptotically equivalent under the null hypothesis, but they can behave rather differently in a small sample. Unfortunately, their small-sample properties are unknown, except in a few special cases. As a consequence, the choice among them is typically made on the basis of ease of computation. The likelihood ratio test requires calculation of both restricted and unrestricted estimators. If both are simple to compute, then this way to proceed is convenient. The Wald test requires only the unrestricted estimator, and the Lagrange multiplier test requires only the restricted estimator. In some problems, one of these estimators may be much easier to compute than the other. For example, a linear model is simple to estimate but becomes nonlinear and cumbersome if a nonlinear constraint is imposed. In this case, the Wald statistic might be preferable. Alternatively, restrictions sometimes amount to the removal of nonlinearities, which would make the Lagrange multiplier test the simpler procedure.

14.6.1 THE LIKELIHOOD RATIO TEST

Let $\boldsymbol{\theta}$ be a vector of parameters to be estimated, and let H_0 specify some sort of restriction on these parameters. Let $\hat{\boldsymbol{\theta}}_U$ be the maximum likelihood estimator of $\boldsymbol{\theta}$ obtained without regard to the constraints, and let $\hat{\boldsymbol{\theta}}_R$ be the constrained maximum likelihood estimator. If $\hat{L}_U$ and $\hat{L}_R$ are the likelihood functions evaluated at these two estimates, then the **likelihood ratio** is

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}. \tag{14-21}$$

This function must be between zero and one. Both likelihoods are positive, and $\hat{L}_R$ cannot be larger than $\hat{L}_U$. (A restricted optimum is never superior to an unrestricted one.) If λ is too small, then doubt is cast on the restrictions.

An example from a discrete distribution helps fix these ideas. In estimating from a sample of 10 from a Poisson population at the beginning of Section 14.3, we found the MLE of the parameter θ to be 2. At this value, the likelihood, which is the probability of observing the sample we did, is 0.104×10^{-7}. Are these data consistent with H_0: $\theta = 1.8$? $L_R = 0.936 \times 10^{-8}$, which is, as expected, smaller. This particular sample is somewhat less probable under the hypothesis.

The formal test procedure is based on the following result.

THEOREM 14.5 Limiting Distribution of the Likelihood Ratio Test Statistic
Under regularity and under H_0, the limiting distribution of $-2 \ln \lambda$ is chi squared, with degrees of freedom equal to the number of restrictions imposed.

The null hypothesis is rejected if this value exceeds the appropriate critical value from the chi-squared tables. Thus, for the Poisson example,

$$-2 \ln \lambda = -2 \ln\left(\frac{0.0936}{0.104}\right) = 0.21072.$$

This chi-squared statistic with one degree of freedom is not significant at any conventional level, so we would not reject the hypothesis that $\theta = 1.8$ on the basis of this test.[10]

It is tempting to use the likelihood ratio test to test a simple null hypothesis against a simple alternative. For example, we might be interested in the Poisson setting in testing H_0: $\theta = 1.8$ against H_1: $\theta = 2.2$. But the test cannot be used in this fashion. The degrees of freedom of the chi-squared statistic for the likelihood ratio test equals the reduction in the number of dimensions in the parameter space that results from imposing the restrictions. In testing a simple null hypothesis against a simple alternative, this value is zero.[11] Second, one sometimes encounters an attempt to test one distributional assumption against another with a likelihood ratio test; for example, a certain model will be estimated assuming a normal distribution and then assuming a t distribution. The ratio of the two likelihoods is then compared to determine which distribution is preferred. This comparison is also inappropriate. The parameter spaces, and hence the likelihood functions of the two cases, are unrelated.

14.6.2 THE WALD TEST

A practical shortcoming of the likelihood ratio test is that it usually requires estimation of both the restricted and unrestricted parameter vectors. In complex models, one or the other of these estimates may be very difficult to compute. Fortunately, there are two alternative testing procedures, the Wald test and the Lagrange multiplier test, that circumvent this problem. Both tests are based on an estimator that is asymptotically normally distributed.

These two tests are based on the distribution of the full rank quadratic form considered in Section B.11.6. Specifically,

$$\text{If } \mathbf{x} \sim N_J[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \text{ then } (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \text{chi-squared}[J]. \qquad \textbf{(14-22)}$$

In the setting of a hypothesis test, under the hypothesis that $E(\mathbf{x}) = \boldsymbol{\mu}$, the quadratic form has the chi-squared distribution. If the hypothesis that $E(\mathbf{x}) = \boldsymbol{\mu}$ is false, however, then the quadratic form just given will, on average, be larger than it would be if the hypothesis were true.[12] This condition forms the basis for the test statistics discussed in this and the next section.

Let $\hat{\boldsymbol{\theta}}$ be the vector of parameter estimates obtained without restrictions. We hypothesize a set of restrictions,

$$H_0: \mathbf{c}(\boldsymbol{\theta}) = \mathbf{q}.$$

If the restrictions are valid, then at least approximately $\hat{\boldsymbol{\theta}}$ should satisfy them. If the hypothesis is erroneous, however, then $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}$ should be farther from $\mathbf{0}$ than would be explained by sampling variability alone. The device we use to formalize this idea is the Wald test.

[10]Of course, our use of the large-sample result in a sample of 10 might be questionable.

[11]Note that because both likelihoods are restricted in this instance, there is nothing to prevent $-2 \ln \lambda$ from being negative.

[12]If the mean is not μ, then the statistic in (14-22) will have a **noncentral chi-squared distribution**. This distribution has the same basic shape as the central chi-squared distribution, with the same degrees of freedom, but lies to the right of it. Thus, a random draw from the noncentral distribution will tend, on average, to be larger than a random observation from the central distribution.

THEOREM 14.6 Limiting Distribution of the Wald Test Statistic
The Wald statistic is

$$W = [\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}]'(\text{Asy.Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}])^{-1}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}].$$

Under H_0, W has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions [i.e., the number of equations in $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q} = 0$].
A derivation of the limiting distribution of the Wald statistic appears in Theorem 5.1.

This test is analogous to the chi-squared statistic in (14-22) if $\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}$ is normally distributed with the hypothesized mean of $\mathbf{0}$. A large value of W leads to rejection of the hypothesis. Note, finally, that W only requires computation of the unrestricted model. One must still compute the covariance matrix appearing in the preceding quadratic form. This result is the variance of a possibly nonlinear function, which we treated earlier.

$$\text{Est.Asy.Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}] = \hat{\mathbf{C}}\,\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}]\hat{\mathbf{C}}',$$
$$\hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'}\right]. \tag{14-23}$$

That is, $\mathbf{C}$ is the $J \times K$ matrix whose jth row is the derivatives of the jth constraint with respect to the K elements of $\boldsymbol{\theta}$. A common application occurs in testing a set of linear restrictions.

For testing a set of linear restrictions $\mathbf{R}\boldsymbol{\theta} = \mathbf{q}$, the Wald test would be based on

$$H_0\text{: } \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{R}\boldsymbol{\theta} - \mathbf{q} = \mathbf{0},$$
$$\hat{\mathbf{C}} = \left[\frac{\partial \mathbf{c}(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'}\right] = \mathbf{R}, \tag{14-24}$$
$$\text{Est.Asy.Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{q}] = \mathbf{R}\,\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}]\mathbf{R},$$

and

$$W = [\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{q}]'[\mathbf{R}\,\text{Est.Asy.Var}(\hat{\boldsymbol{\theta}})\mathbf{R}']^{-1}[\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{q}].$$

The degrees of freedom is the number of rows in $\mathbf{R}$.

If $\mathbf{c}(\boldsymbol{\theta}) = \mathbf{q}$ is a single restriction, then the Wald test will be the same as the test based on the confidence interval developed previously. If the test is

$$H_0\text{: } \theta = \theta_0 \quad \text{versus} \quad H_1\text{: } \theta \neq \theta_0,$$

then the earlier test is based on

$$z = \frac{|\hat{\theta} - \theta_0|}{s(\hat{\theta})}, \tag{14-25}$$

where $s(\hat{\theta})$ is the estimated asymptotic standard error. The test statistic is compared to the appropriate value from the standard normal table. The Wald test will be based on

$$W = [(\hat{\theta} - \theta_0) - 0](\text{Asy.Var}[(\hat{\theta} - \theta_0) - 0])^{-1}[(\hat{\theta} - \theta_0) - 0] = \frac{(\hat{\theta} - \theta_0)^2}{\text{Asy.Var}[\hat{\theta}]} = z^2. \tag{14-26}$$

Here W has a limiting chi-squared distribution with one degree of freedom, which is the distribution of the square of the standard normal test statistic in (14-25).

To summarize, the Wald test is based on measuring the extent to which the unrestricted estimates fail to satisfy the hypothesized restrictions. There are two shortcomings of the Wald test. First, it is a pure significance test against the null hypothesis, not necessarily for a specific alternative hypothesis. As such, its power may be limited in some settings. In fact, the test statistic tends to be rather large in applications. The second shortcoming is not shared by either of the other test statistics discussed here. The Wald statistic is not invariant to the formulation of the restrictions. For example, for a test of the hypothesis that a function $\theta = \beta/(1 - \gamma)$ equals a specific value q there are two approaches one might choose. A Wald test based directly on $\theta - q = 0$ would use a statistic based on the variance of this nonlinear function. An alternative approach would be to analyze the linear restriction $\beta - q(1 - \gamma) = 0$, which is an equivalent, but linear, restriction. The Wald statistics for these two tests could be different and might lead to different inferences. These two shortcomings have been widely viewed as compelling arguments against use of the Wald test. But, in its favor, the Wald test does not rely on a strong distributional assumption, as do the likelihood ratio and Lagrange multiplier tests. The recent econometrics literature is replete with applications that are based on distribution free estimation procedures, such as the GMM method. As such, in recent years, the Wald test has enjoyed a redemption of sorts.

14.6.3 THE LAGRANGE MULTIPLIER TEST

The third test procedure is the **Lagrange multiplier (LM)** or **efficient score** (or just **score**) **test**. It is based on the restricted model instead of the unrestricted model. Suppose that we maximize the log likelihood subject to the set of constraints $\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}$. Let $\boldsymbol{\lambda}$ be a vector of Lagrange multipliers and define the Lagrangean function

$$\ln L^*(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) + \boldsymbol{\lambda}'(\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q}).$$

The solution to the constrained maximization problem is the joint solution of

$$\begin{aligned} \frac{\partial \ln L^*}{\partial \boldsymbol{\theta}} &= \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0}, \\ \frac{\partial \ln L^*}{\partial \boldsymbol{\lambda}} &= \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}, \end{aligned} \tag{14-27}$$

where $\mathbf{C}'$ is the transpose of the derivatives matrix in the second line of (14-23). If the restrictions are valid, then imposing them will not lead to a significant difference in the maximized value of the likelihood function. In the first-order conditions, the meaning is that the second term in the derivative vector will be small. In particular, $\boldsymbol{\lambda}$ will be small. We could test this directly, that is, test H_0: $\boldsymbol{\lambda} = \mathbf{0}$, which leads to the Lagrange multiplier test. There is an equivalent simpler formulation, however. At the restricted maximum, the derivatives of the log-likelihood function are

$$\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R} = -\mathbf{C}'\hat{\boldsymbol{\lambda}} = \hat{\mathbf{g}}_R. \tag{14-28}$$

If the restrictions are valid, at least within of sampling variability, then $\hat{\mathbf{g}}_R = \mathbf{0}$. That is, the derivatives of the log likelihood evaluated at the restricted parameter vector will be approximately zero. The vector of first derivatives of the log likelihood is the vector of efficient scores. Because the test is based on this vector, it is called the score test as well as the Lagrange multiplier test. The variance of the first derivative vector is the information matrix, which we have used to compute the asymptotic covariance matrix of the MLE. The test statistic is based on reasoning analogous to that underlying the Wald test statistic.

THEOREM 14.7 Limiting Distribution of the Lagrange Multiplier Statistic
The Lagrange multiplier test statistic is

$$\text{LM} = \left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R}\right)' [\mathbf{I}(\hat{\boldsymbol{\theta}}_R)]^{-1} \left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R}\right).$$

Under the null hypothesis, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. All terms are computed at the restricted estimator.

The LM statistic has a useful form. Let $\hat{\mathbf{g}}_{iR}$ denote the ith term in the gradient of the log-likelihood function. Then $\hat{\mathbf{g}}_R = \sum_{i=1}^{n} \hat{\mathbf{g}}_{iR} = \hat{\mathbf{G}}_R' \mathbf{i}$, where $\hat{\mathbf{G}}_R$ is the $n \times K$ matrix with ith row equal to $\hat{\mathbf{g}}_{iR}'$ and $\mathbf{i}$ is a column of 1s. If we use the BHHH (outer product of gradients) estimator in (14-18) to estimate the Hessian, then $[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})]^{-1} = [\hat{\mathbf{G}}_R' \hat{\mathbf{G}}_R]^{-1}$, and

$$\text{LM} = \mathbf{i}' \hat{\mathbf{G}}_R [\hat{\mathbf{G}}_R' \hat{\mathbf{G}}_R]^{-1} \hat{\mathbf{G}}_R' \mathbf{i}.$$

Now, because $\mathbf{i}'\mathbf{i}$ equals n, $\text{LM} = n(\mathbf{i}' \hat{G}_R [\hat{\mathbf{G}}_R' \hat{\mathbf{G}}_R]^{-1} \hat{\mathbf{G}}_R' \mathbf{i}/n) = nR_i^2$, which is n times the uncentered squared multiple correlation coefficient in a linear regression of a column of 1s on the derivatives of the log-likelihood function computed at the restricted estimator. We will encounter this result in various forms at several points in the book.

14.6.4 AN APPLICATION OF THE LIKELIHOOD-BASED TEST PROCEDURES

Consider, again, the data in Example C.1. In Example 14.4, the parameter β in the model

$$f(y_i | x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)} \tag{14-29}$$

was estimated by maximum likelihood. For convenience, let $\alpha_i = 1/(\beta + x_i)$. This exponential density is a restricted form of a more general gamma distribution,

$$f(y_i | x_i, \beta, \rho) = \frac{\alpha_i^\rho}{\Gamma(\rho)} y_i^{\rho - 1} e^{-y_i \alpha_i}. \tag{14-30}$$

The restriction is $\rho = 1$.[13] We consider testing the hypothesis

$$H_0: \rho = 1 \quad \text{versus} \quad H_1: \rho \neq 1$$

using the various procedures described previously. The log likelihood and its derivatives are

$$\ln L(\beta, \rho) = \rho \sum_{i=1}^{n} \ln \alpha_i - n \ln \Gamma(\rho) + (\rho - 1) \sum_{i=1}^{n} \ln y_i - \sum_{i=1}^{n} y_i \alpha_i,$$

$$\frac{\partial \ln L}{\partial \beta} = -\rho \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} y_i \alpha_i^2, \qquad \frac{\partial \ln L}{\partial \rho} = \sum_{i=1}^{n} \ln \alpha_i - n\Psi(\rho) + \sum_{i=1}^{n} \ln y_i, \quad \textbf{(14-31)}$$

$$\frac{\partial^2 \ln L}{\partial \beta^2} = \rho \sum_{i=1}^{n} \alpha_i^2 - 2 \sum_{i=1}^{n} y_i \alpha_i^3, \qquad \frac{\partial^2 \ln L}{\partial \rho^2} = -n\Psi'(\rho), \quad \frac{\partial^2 \ln L}{\partial \beta \partial \rho} = -\sum_{i=1}^{n} \alpha_i.$$

[Recall that $\Psi(\rho) = d \ln \Gamma(\rho)/d\rho$ and $\Psi'(\rho) = d^2 \ln \Gamma(\rho)/d\rho^2$.] Unrestricted maximum likelihood estimates of β and ρ are obtained by equating the two first derivatives to zero. The restricted maximum likelihood estimate of β is obtained by equating $\partial \ln L/\partial \beta$ to zero while fixing ρ at one. The results are shown in Table 14.1. Three estimators are available for the asymptotic covariance matrix of the estimators of $\boldsymbol{\theta} = (\beta, \rho)'$. Using the actual Hessian as in (14-17), we compute $\mathbf{V} = [-\Sigma_i \partial^2 \ln f(y_i|x_i, \beta, \rho)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}']^{-1}$ at the maximum likelihood estimates. For this model, it is easy to show that $E[y_i|x_i] = \rho(\beta + x_i)$ (either by direct integration or, more simply, by using the result that $E[\partial \ln L/\partial \beta] = 0$ to deduce it). Therefore, we can also use the expected Hessian as in (14-16) to compute $\mathbf{V}_E = \{-\Sigma_i E[\partial^2 \ln f(y_i|x_i, \beta, \rho)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}']\}^{-1}$. Finally, by using the sums of squares and cross products of the first derivatives, we obtain the BHHH estimator in (14-18), $\mathbf{V}_B = [\Sigma_i (\partial \ln f(y_i|x_i, \beta, \rho)/\partial \boldsymbol{\theta})(\partial \ln f(y_i|x_i, \beta, \rho)/\partial \boldsymbol{\theta}')]^{-1}$. Results in Table 14.1 are based on $\mathbf{V}$.

The three estimators of the asymptotic covariance matrix produce notably different results:

$$\mathbf{V} = \begin{bmatrix} 5.499 & -1.653 \\ -1.653 & 0.6309 \end{bmatrix}, \quad \mathbf{V}_E = \begin{bmatrix} 4.900 & -1.473 \\ -1.473 & 0.5768 \end{bmatrix}, \quad \mathbf{V}_B = \begin{bmatrix} 13.370 & -4.322 \\ -4.322 & 1.537 \end{bmatrix}.$$

TABLE 14.1 Maximum Likelihood Estimates

Quantity	*Unrestricted Estimate*[a]	*Restricted Estimate*
β	−4.7185 (2.345)	15.6027 (6.794)
ρ	3.1509 (0.794)	1.0000 (0.000)
$\ln L$	−82.91605	−88.4363
$\partial \ln L/\partial \beta$	0.0000	0.0000
$\partial \ln L/\partial \rho$	0.0000	7.9145
$\partial^2 \ln L/\partial \beta^2$	−0.8557	−0.0217
$\partial^2 \ln L/\partial \rho^2$	−7.4592	−32.8987
$\partial^2 \ln L/\partial \beta \partial \rho$	−2.2420	−0.66891

[a]Estimated asymptotic standard errors based on $\mathbf{V}$ are given in parentheses.

[13]The gamma function $\Gamma(\rho)$ and the gamma distribution are described in Sections B.4.5 and E2.3.

Given the small sample size, the differences are to be expected. Nonetheless, the striking difference of the BHHH estimator is typical of its erratic performance in small samples.

- **Confidence interval test:** A 95% confidence interval for ρ based on the unrestricted estimates is $3.1509 \pm 1.96\sqrt{0.6309} = [1.5941, 4.7076]$. This interval does not contain $\rho = 1$, so the hypothesis is rejected.
- **Likelihood ratio test:** The LR statistic is $\lambda = -2[-88.43626 - (-82.91604)] = 11.0404$. The table value for the test, with one degree of freedom, is 3.842. The computed value is larger than this critical value, so the hypothesis is again rejected.
- **Wald test:** The Wald test is based on the unrestricted estimates. For this restriction, $c(\boldsymbol{\theta}) - q = \rho - 1$, $dc(\hat{\rho})/d\hat{\rho} = 1$, $\text{Est.Asy.Var}[c(\hat{\rho}) - q] = \text{Est.Asy.Var}[\hat{\rho}] = 0.6309$, so $W = (3.1517 - 1)^2/[0.6309] = 7.3384$. The critical value is the same as the previous one. Hence, H_0 is once again rejected. Note that the Wald statistic is the square of the corresponding test statistic that would be used in the confidence interval test, $|3.1509 - 1|/\sqrt{0.6309} = 2.73335$.
- **Lagrange multiplier test:** The Lagrange multiplier test is based on the restricted estimators. The estimated asymptotic covariance matrix of the derivatives used to compute the statistic can be any of the three estimators discussed earlier. The BHHH estimator, $\mathbf{V}_B$, is the empirical estimator of the variance of the gradient and is the one usually used in practice. This computation produces

$$\text{LM} = [0.0000 \quad 7.9145] \begin{bmatrix} 0.00995 & 0.26776 \\ 0.26776 & 11.199 \end{bmatrix}^{-1} \begin{bmatrix} 0.0000 \\ 7.9145 \end{bmatrix} = 15.687.$$

The conclusion is the same as before. Note that the same computation done using V rather than $\mathbf{V}_B$ produces a value of 5.1162. As before, we observe substantial small-sample variation produced by the different estimators.

The latter three test statistics have substantially different values. It is possible to reach different conclusions, depending on which one is used. For example, if the test had been carried out at the 1% level of significance instead of 5% and LM had been computed using $\mathbf{V}$, then the critical value from the chi-squared statistic would have been 6.635 and the hypothesis would not have been rejected by the LM test. Asymptotically, all three tests are equivalent. But, in a finite sample such as this one, differences are to be expected.[14] Unfortunately, there is no clear rule for how to proceed in such a case, which highlights the problem of relying on a particular significance level and drawing a firm reject or accept conclusion based on sample evidence.

14.6.5 COMPARING MODELS AND COMPUTING MODEL FIT

The test statistics described in Sections 14.6.1–14.6.3 are available for assessing the validity of restrictions on the parameters in a model. When the models are nested, any of the three mentioned testing procedures can be used. For nonnested models, the computation is a comparison of one model to another based on an estimation criterion to discern which is to be preferred. Two common measures that are based on the same logic as the adjusted R-squared for the linear model are

[14]For further discussion of this problem, see Berndt and Savin (1977).

$$\begin{aligned}\textbf{Akaike information criterion (AIC)} &= -2 \ln L + 2K,\\ \textbf{Bayes (Schwarz) information criterion (BIC)} &= -2 \ln L + K \ln n,\end{aligned}$$

where K is the number of parameters in the model. Choosing a model based on the lowest AIC is logically the same as using $\overline{R}^2$ in the linear model, nonstatistical, albeit widely accepted.

The AIC and BIC are information criteria, not fit measures as such. This does leave open the question of how to assess the "fit" of the model. Only the case of a linear least squares regression in a model with a constant term produces an R^2, which measures the proportion of variation explained by the regression. The ambiguity in R^2 as a fit measure arose immediately when we moved from the linear regression model to the generalized regression model in Chapter 9. The problem is yet more acute in the context of the models we consider in this chapter. For example, the estimators of the models for count data in Example 14.10 make no use of the "variation" in the dependent variable and there is no obvious measure of "explained variation."

A measure of fit that was originally proposed for discrete choice models in McFadden (1974), but surprisingly has gained wide currency throughout the empirical literature is the **likelihood ratio index**, which has come to be known as the **Pseudo R^2**. It is computed as

$$\text{Pseudo } R^2 = 1 - (\ln L)/(\ln L_0),$$

where $\ln L$ is the log likelihood for the model estimated and $\ln L_0$ is the log likelihood for the same model with only a constant term. The statistic does resemble the R^2 in a linear regression. The choice of name for this statistic is unfortunate, however, because even in the discrete choice context for which it was proposed, it has no connection to the fit of the model to the data. In discrete choice settings in which log likelihoods must be negative, the pseudo R^2 must be between zero and one and rises as variables are added to the model. It can obviously be zero, but is usually bounded below one. In the linear model with normally distributed disturbances, the maximized log likelihood is

$$\ln L = (-n/2)[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)].$$

With a small amount of manipulation, we find that the pseudo R^2 for the linear regression model is

$$\text{Pseudo } R^2 = \frac{-\ln(1 - R^2)}{1 + \ln 2\pi + \ln s_y^2},$$

while the *true* R^2 is $1 - \mathbf{e}'\mathbf{e}/\mathbf{e}_0'\mathbf{e}_0$. Because s_y^2 can vary independently of R^2—multiplying $\mathbf{y}$ by any scalar, A, leaves R^2 unchanged but multiplies s_y^2 by A^2—although the upper limit is one, there is no lower limit on this measure. It can even be negative. This same problem arises in any model that uses information on the scale of a dependent variable, such as the tobit model (Chapter 19). The computation makes even less sense as a fit measure in multinomial models such as the ordered probit model (Chapter 18) or the multinomial logit model. For discrete choice models, a variety of such measures are discussed in Chapter 17. For limited dependent variable and many loglinear models, some other measure that is related to a correlation between a prediction and the actual value would be more useable. Nonetheless, the measure has gained currency in the

contemporary literature.[15] Notwithstanding the general contempt for the likelihood ratio index, practitioners are often interested in comparing models based on some idea of the fit of the model the data. Constructing such a measure will be specific to the context, so we will return to the issue in the discussion of specific applications such as the binary choice in Chapter 17.

14.6.6 VUONG'S TEST AND THE KULLBACK–LEIBLER INFORMATION CRITERION

Vuong's (1989) approach to testing **nonnested models** is also based on the likelihood ratio statistic. The logic of the test is similar to that which motivates the likelihood ratio test in general. Suppose that $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$ and $g(y_i|\mathbf{Z}_i, \boldsymbol{\gamma})$ are two competing models for the density of the random variable y_i, with f being the null model, H_0, and g being the alternative, H_1. For instance, in Example 5.7, both densities are (by assumption now) normal, y_i is consumption, C_t, $\mathbf{Z}_i$ is $[1, Y_t, Y_{t-1}, C_{t-1}]$, $\boldsymbol{\theta}$ is $(\beta_1, \beta_2, \beta_3, 0, \sigma^2)$, $\boldsymbol{\gamma}$ is $(\gamma_1, \gamma_2, 0, \gamma_3, \omega^2)$, and σ^2 and ω^2 are the respective conditional variances of the disturbances, ε_{0t} and ε_{1t}. The crucial element of Vuong's analysis is that it need not be the case that either competing model is *true*; they may both be incorrect. What we want to do is attempt to use the data to determine which competitor is closer to the truth, that is, closer to the correct (unknown) model.

We assume that observations in the sample (disturbances) are conditionally independent. Let $L_{i,0}$ denote the ith contribution to the likelihood function under the null hypothesis. Thus, the log-likelihood function under the null hypothesis is $\Sigma_i \ln L_{i,0}$. Define $L_{i,1}$ likewise for the alternative model. Now, let m_i equal $\ln L_{i,1} - \ln L_{i,0}$. If we were using the familiar likelihood ratio test, then, the likelihood ratio statistic would be simply $LR = 2\Sigma_i m_i = 2n\overline{m}$ when $L_{i,0}$ and $L_{i,1}$ are computed at the respective maximum likelihood estimators. When the competing models are nested—H_0 is a restriction on H_1—we know that $\Sigma_i m_i \geq 0$. The restrictions of the null hypothesis will never increase the likelihood function. (In the linear regression model with normally distributed disturbances that we have examined so far, the log likelihood and these results are all based on the sum of squared residuals. And, as we have seen, imposing restrictions never reduces the sum of squares.) The limiting distribution of the LR statistic under the assumption of the null hypothesis is chi squared with degrees of freedom equal to the reduction in the number of dimensions of the parameter space of the alternative hypothesis that results from imposing the restrictions.

Vuong's analysis is concerned with nonnested models for which $\Sigma_i m_i$ need not be positive. Formalizing the test requires us to look more closely at what is meant by the *right* model (and provides a convenient departure point for the discussion in the next two sections). In the context of nonnested models, Vuong allows for the possibility that neither model is *true* in the absolute sense. We maintain the classical assumption that there does exist a true model, $h(y_i|\mathbf{Z}_i, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is the true parameter vector, but possibly neither hypothesized model is that true model. The **Kullback–Leibler Information Criterion (KLIC)** measures the distance between the true model (distribution) and a

[15]The software package *Stata* reports the pseudo R^2 with every model fit by MLE, but at the same time, admonishes its users not to interpret it as anything meaningful. See, for example, www.stata.com/support/faqs/stat/pseudor2.html. Cameron and Trivedi (2005) document the pseudo R^2 at length and then give similar cautions about it and urge their readers to seek a more meaningful measure of the correlation between model predictions and the outcome variable of interest. Wooldridge (2010, p. 575) dismisses it summarily, and argues that partial effects are more important.

hypothesized model in terms of the likelihood function. Loosely, the KLIC is the log-likelihood function under the hypothesis of the true model minus the log-likelihood function for the (misspecified) hypothesized model under the assumption of the true model. Formally, for the model of the null hypothesis,

$$\text{KLIC} = E[\ln h(y_i|\mathbf{Z}_i, \boldsymbol{\alpha})|h \text{ is true}] - E[\ln f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})|h \text{ is true}].$$

The first term on the right-hand side is what we would estimate with $(1/n)$ ln L if we maximized the log likelihood for the true model, $h(y_i|\mathbf{Z}_i, \boldsymbol{\alpha})$. The second term is what is estimated by $(1/n)$ ln L assuming (incorrectly) that $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$ is the correct model. Notice that $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$ is written in terms of a parameter vector, $\boldsymbol{\theta}$. Because $\boldsymbol{\alpha}$ is the true parameter vector, it is perhaps ambiguous what is meant by the parameterization, $\boldsymbol{\theta}$. Vuong (p. 310) calls this the "pseudotrue" parameter vector. It is the vector of constants that the estimator converges to when one uses the estimator implied by $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$. In Example 5.7, if H_0 gives the correct model, this formulation assumes that the least squares estimator in H_1 would converge to some vector of pseudo-true parameters. But these are not the parameters of the correct model—they would be the slopes in the population linear projection of C_t on $[1, Y_t, C_{t-1}]$.

Suppose the true model is $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, with normally distributed disturbances and $\mathbf{y} = \mathbf{Z\delta} + \mathbf{w}$ is the proposed competing model. The KLIC would be the expected log-likelihood function for the true model minus the expected log-likelihood function for the second model, still assuming that the first one is the truth. By construction, the KLIC is positive. We will now say that one model is better than another if it is closer to the truth based on the KLIC. If we take the difference of the two KLICs for two models, the true log-likelihood function falls out, and we are left with

$$\text{KLIC}_1 - \text{KLIC}_0 = E[\ln f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})|h \text{ is true}] - E[\ln g(y_i|\mathbf{Z}_i, \boldsymbol{\gamma})|h \text{ is true}].$$

To compute this using a sample, we would simply compute the likelihood ratio statistic, $n\overline{m}$ (without multiplying by 2) again. Thus, this provides an interpretation of the LR statistic. But, in this context, the statistic can be negative—we don't know which competing model is closer to the truth.

Vuong's general result for nonnested models (his Theorem 5.1) describes the behavior of the statistic

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}} = \sqrt{n}(\overline{m}/s_m), \quad m_i = \ln L_{i,1} - \ln L_{i,0}.$$

He finds:

1. Under the hypothesis that the models are "equivalent," $V \xrightarrow{D} N[0,1]$.
2. Under the hypothesis that $f(y_i|\mathbf{Z}_i, \boldsymbol{\theta})$ is "better," $V \xrightarrow{A.S.} +\infty$.
3. Under the hypothesis that $g(y_i|\mathbf{Z}_i, \boldsymbol{\gamma})$ is "better," $V \xrightarrow{A.S.} -\infty$.

This test is directional. Large positive values favor the null model while large negative values favor the alternative. The intermediate values (e.g., between −1.96 and +1.96 for 95% significance) are an inconclusive region. An application appears in Example 14.8.

14.7 TWO-STEP MAXIMUM LIKELIHOOD ESTIMATION

The applied literature contains a large and increasing number of applications in which elements of one model are embedded in another, which produces what are known as "two-step" estimation problems.[16] There are two parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The first appears in the second model, but the second does not appear in the first model. In such a situation, there are two ways to proceed. **Full information maximum likelihood (FIML)** estimation would involve forming the joint distribution $f(y_1, y_2|\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ of the two random variables and then maximizing the full log-likelihood function,

$$\ln L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \ln f(y_{i1}, y_{i2}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

A two-step procedure for this kind of model could be used by estimating the parameters of model 1 first by maximizing

$$\ln L_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \ln f_1(y_{i1}|\mathbf{x}_{i1}, \boldsymbol{\theta}_1)$$

and then maximizing the marginal likelihood function for y_2 while embedding the consistent estimator of $\boldsymbol{\theta}_1$, treating it as given. The second step involves maximizing

$$\ln L_2(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \ln f_2(y_{i2}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2).$$

There are at least two reasons one might proceed in this fashion. First, it may be straightforward to formulate the two separate log likelihoods, but very complicated to derive the joint distribution. This situation frequently arises when the two variables being modeled are from different kinds of populations, such as one discrete and one continuous (which is a very common case in this framework). The second reason is that maximizing the separate log likelihoods may be fairly straightforward, but maximizing the joint log likelihood may be numerically complicated or difficult.[17] The results given here can be found in an important reference on the subject, Murphy and Topel (2002, first published in 1985).

Suppose, then, that our model consists of the two marginal distributions, $f_1(y_1|\mathbf{x}_1, \boldsymbol{\theta}_1)$ and $f_2(y_2|\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Estimation proceeds in two steps.

1. Estimate $\boldsymbol{\theta}_1$ by maximum likelihood in model 1. Let $\hat{\mathbf{V}}_1$ be n times any of the estimators of the asymptotic covariance matrix of this estimator that were discussed in Section 14.4.6.
2. Estimate $\boldsymbol{\theta}_2$ by maximum likelihood in model 2, with $\hat{\boldsymbol{\theta}}_1$ inserted in place of $\boldsymbol{\theta}_1$ as if it were known. Let $\hat{\mathbf{V}}_2$ be n times any appropriate estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_2$.

[16] Among the best known of these is Heckman's (1979) model of sample selection discussed in Example 1.1 and in Chapter 19.

[17] There is a third possible motivation. If either model is misspecified, then the FIML estimates of both models will be inconsistent. But if only the second is misspecified, at least the first will be estimated consistently. Of course, this result is only "half a loaf," but it may be better than none.

The argument for consistency of $\hat{\boldsymbol{\theta}}_2$ is essentially that if $\boldsymbol{\theta}_1$ *were* known, then all our results for MLEs would apply for estimation of $\boldsymbol{\theta}_2$, and because plim $\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1$, asymptotically, this line of reasoning is correct. (See point 3 of Theorem D.16.) But the same line of reasoning is not sufficient to justify using $(1/n)\mathbf{V}_2$ as the estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_2$. Some correction is necessary to account for an estimate of $\boldsymbol{\theta}_1$ being used in estimation of $\boldsymbol{\theta}_2$. The essential result is the following:

THEOREM 14.8 Asymptotic Distribution of the Two-Step MLE [Murphy and Topel (2002)]

If the standard regularity conditions are met for both log-likelihood functions, then the second-step maximum likelihood estimator of $\boldsymbol{\theta}_2$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\mathbf{V}_2^* = \frac{1}{n}[\mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}' - \mathbf{R}\mathbf{V}_1\mathbf{C}' - \mathbf{C}\mathbf{V}_1\mathbf{R}']\mathbf{V}_2],$$

where

$$\mathbf{V}_1 = \text{Asy.Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)] \text{ based on } \ln L_1,$$

$$\mathbf{V}_2 = \text{Asy.Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)] \text{ based on } \ln L_2 | \boldsymbol{\theta}_1,$$

$$\mathbf{C} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_1'}\right)\right], \quad \mathbf{R} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_1}{\partial \boldsymbol{\theta}_1'}\right)\right].$$

The correction of the asymptotic covariance matrix at the second step requires some additional computation. Matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ are estimated by the respective uncorrected covariance matrices. Typically, the BHHH estimators,

$$\hat{\mathbf{V}}_1 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)\right]^{-1}$$

and

$$\hat{\mathbf{V}}_2 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2'}\right)\right]^{-1}$$

are used. The matrices $\mathbf{R}$ and $\mathbf{C}$ are obtained by summing the individual observations on the cross products of the derivatives. These are estimated with

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)$$

and

$$\hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1'}\right).$$

A derivation of this useful result is instructive. We will rely on (14-11) and the results of Section 14.4.5.B where the asymptotic normality of the maximum likelihood estimator is developed. The first-step MLE of $\boldsymbol{\theta}_1$ is defined by

$$\frac{1}{n}\frac{\partial \ln L_1(\hat{\boldsymbol{\theta}}_1)}{\hat{\boldsymbol{\theta}}_1} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f_1(y_{i1}|\mathbf{x}_{i1}, \hat{\boldsymbol{\theta}}_1)}{\partial \hat{\boldsymbol{\theta}}_1} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_{i1}(\hat{\boldsymbol{\theta}}_1) = \bar{\mathbf{g}}_1(\hat{\boldsymbol{\theta}}_1) = \mathbf{0}.$$

Using the results in that section, we obtained the asymptotic distribution from (14-15),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} [-\mathbf{H}_{11}^{(1)}(\theta_1)]^{-1}\sqrt{n}\bar{\mathbf{g}}_1(\theta_1),$$

where the expression means that the limiting distribution of the two random vectors is the same,

and

$$\mathbf{H}_{11}^{(1)} = E\left[\frac{1}{n}\frac{\partial^2 \ln L_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'}\right].$$

The second-step MLE of $\boldsymbol{\theta}_2$ is defined by

$$\frac{1}{n}\frac{\partial \ln L_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \hat{\boldsymbol{\theta}}_2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f_2(y_{i2}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \hat{\boldsymbol{\theta}}_2} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_{i2}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \hat{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \mathbf{0}.$$

Expand the derivative vector, $\bar{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$, in a linear Taylor series as usual, and use the results in Section 14.4.5.b once again,

$$\bar{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \left[\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right](\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) + \left[\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right](\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + o(1/n) = \mathbf{0},$$

where

$$\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = E\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_1'}\right] \text{ and } \mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = E\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2'}\right].$$

To obtain the asymptotic distribution, we use the same device as before,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) \xrightarrow{d} \left[-\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right]^{-1}\sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \theta_2) + \left[-\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right]^{-1}\left[\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right]\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1).$$

For convenience, denote $\mathbf{H}_{22}^{(2)} = \mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\mathbf{H}_{21}^{(2)} = \mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $\mathbf{H}_{11}^{(1)} = \mathbf{H}_{11}^{(1)}(\boldsymbol{\theta}_1)$. Now substitute the first-step estimator of $\boldsymbol{\theta}_1$ in this expression to obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) \xrightarrow{d} \left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\sqrt{n}\bar{\mathbf{g}}_1(\boldsymbol{\theta}_1).$$

Consistency and asymptotic normality of the two estimators follow from our earlier results. To obtain the asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_2$ we will obtain the limiting variance of the random vector in the preceding expression. The joint normal distribution of the two first derivative vectors has zero means and

$$\text{Var}\begin{bmatrix}\sqrt{n}\bar{\mathbf{g}}_1(\boldsymbol{\theta}_1)\\ \sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_2,\boldsymbol{\theta}_1)\end{bmatrix}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}.$$

Then, the asymptotic covariance matrix we seek is

$$\begin{aligned}\text{Var}\big[\sqrt{n}(\hat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_2)\big] = {}& \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\boldsymbol{\Sigma}_{22}\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ &+\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\boldsymbol{\Sigma}_{11}\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]'\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ &+\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\boldsymbol{\Sigma}_{21}\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]'\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ &+\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\boldsymbol{\Sigma}_{12}\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}.\end{aligned}$$

As we found earlier, the variance of the first derivative vector of the log likelihood is the negative of the expected second derivative matrix [see (14-11)]. Therefore $\boldsymbol{\Sigma}_{22}=[-\mathbf{H}_{22}^{(2)}]$ and $\boldsymbol{\Sigma}_{11}=[-\mathbf{H}_{11}^{(1)}]$. Making the substitution we obtain

$$\begin{aligned}\text{Var}\big[\sqrt{n}(\hat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_2)\big] = {}& \big[-\mathbf{H}_{22}^{(2)}\big]^{-1}+\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]'\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ &+\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\boldsymbol{\Sigma}_{21}\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]'\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\\ &+\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}\big[\mathbf{H}_{21}^{(2)}\big]\big[-\mathbf{H}_{11}^{(1)}\big]^{-1}\boldsymbol{\Sigma}_{12}\big[-\mathbf{H}_{22}^{(2)}\big]^{-1}.\end{aligned}$$

From (14-15), $[-\mathbf{H}_{11}^{(1)}]^{-1}$ and $[-\mathbf{H}_{22}^{(2)}]^{-1}$ are the $\mathbf{V}_1$ and $\mathbf{V}_2$ that appear in Theorem 14.8, which further reduces the expression to

$$\begin{aligned}&\text{Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_2)]\\ &\quad=\mathbf{V}_2+\mathbf{V}_2\big[\mathbf{H}_{21}^{(2)}\big]\mathbf{V}_1\big[\mathbf{H}_{21}^{(2)}\big]'\mathbf{V}_2-\mathbf{V}_2\boldsymbol{\Sigma}_{21}\mathbf{V}_1\big[\mathbf{H}_{21}^{(2)}\big]'\mathbf{V}_2-\mathbf{V}_2\big[\mathbf{H}_{21}^{(2)}\big]\mathbf{V}_1\boldsymbol{\Sigma}_{12}\mathbf{V}_2.\end{aligned}$$

Two remaining terms are $\mathbf{H}_{21}^{(2)}$, which is the $E[\partial^2 \ln L_2(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_1']$, which is being estimated by $-\mathbf{C}$ in the statement of the theorem [note (14-11) again for the change of sign] and $\boldsymbol{\Sigma}_{21}$, which is the covariance of the two first derivative vectors. This is being estimated by $\mathbf{R}$ in Theorem 14.8. Making these last two substitutions produces

$$\text{Var}\big[\sqrt{n}(\hat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_2)\big]=\mathbf{V}_2+\mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2-\mathbf{V}_2\mathbf{R}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2-\mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{R}'\mathbf{V}_2,$$

which completes the derivation.

Example 14.5 Two-Step ML Estimation

A common application of the two-step method is accounting for the variation in a constructed regressor in a second-step model. In this instance, the constructed variable is often an estimate of an expected value of a variable that is likely to be endogenous in the second-step model. In this example, we will construct a rudimentary model that illustrates the computations.

In Riphahn, Wambach, and Million (RWM, 2003), the authors studied whether individuals' use of the German health care system was at least partly explained by whether or not they had purchased a particular type of supplementary health insurance. We have used their data set, German Socioeconomic Panel (GSOEP), at several points. (See, Example 7.6.) One of the variables of interest in the study is *DocVis*, the number of times an individual visits the doctor during the survey year. RWM considered the possibility that the presence of supplementary (*Addon*) insurance had an influence on the number of visits. Our simple model is as follows: The model for the number of visits is a Poisson regression (see Section 18.4.1). This is a loglinear model that we will specify as

$$E[DocVis\,|\,\mathbf{x}_2, P_{Addon}]=\mu(\mathbf{x}_2'\boldsymbol{\beta},\gamma,\mathbf{x}_1'\boldsymbol{\alpha})=\exp[\mathbf{x}_2'\boldsymbol{\beta}+\gamma\Lambda(\mathbf{x}_1'\boldsymbol{\alpha})].$$

The model contains the dummy variable equal to 1 if the individual has *Addon* insurance and 0 otherwise, which is likely to be endogenous in this equation. But, an estimate of $E[Addon \mid \mathbf{x}_1]$ from a **logistic probability model** (see Section 17.2) for whether the individual has insurance,

$$\Lambda(\mathbf{x}_1'\boldsymbol{\alpha}) = \frac{\exp(\mathbf{x}_1'\boldsymbol{\alpha})}{1 + \exp(\mathbf{x}_1'\boldsymbol{\alpha})} = \text{Prob[Individual has purchased } \textit{Addon} \text{ insurance} \mid \mathbf{x}_1].$$

For purposes of the exercise, we will specify

$$(y_1 = \textit{Addon})\ \mathbf{x}_1 = (\textit{constant}, \textit{Age}, \textit{Education}, \textit{Married}, \textit{Kids}),$$
$$(y_2 = \textit{DocVis})\ \mathbf{x}_2 = (\textit{constant}, \textit{Age}, \textit{Education}, \textit{Income}, \textit{Female}).$$

As before, to sidestep issues related to the panel data nature of the data set, we will use the 4,483 observations in the 1988 wave of the data set, and drop the two observations for which *Income* is zero.

The log likelihood for the logistic probability model is

$$\ln L_1(\boldsymbol{\alpha}) = \Sigma_i\{(1 - y_{i1}) \ln[1 - \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})] + y_{i1} \ln \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})\}.$$

The derivatives of this log likelihood are

$$\mathbf{g}_{i1}(\boldsymbol{\alpha}) = \partial \ln f_1(y_{i1} \mid \mathbf{x}_{i1}, \boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = [y_{i1} - \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]\mathbf{x}_{i1}.$$

We will maximize this log likelihood with respect to $\boldsymbol{\alpha}$ and then compute $\mathbf{V}_1$ using the BHHH estimator, as in Theorem 14.8. We will also use $\mathbf{g}_{i1}(\boldsymbol{\alpha})$ in computing $\mathbf{R}$.

The log likelihood for the Poisson regression model is

$$\ln L_2 = \Sigma_i[-\mu(\mathbf{x}_{i2}'\beta, \gamma, \mathbf{x}_{i1}'\boldsymbol{\alpha}) + y_{i2} \ln \mu(\mathbf{x}_{i2}'\beta, \gamma, \mathbf{x}_{i1}'\boldsymbol{\alpha}) - \ln y_{i2}'].$$

The derivatives of this log likelihood are

$$\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) = \partial \ln f_2(y_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial(\boldsymbol{\beta}', \gamma)' = [y_{i2} - \mu(\mathbf{x}_{i2}'\boldsymbol{\beta}, \gamma, \mathbf{x}_{i1}'\boldsymbol{\alpha})][\mathbf{x}_{i2}', \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]'$$

$$\mathbf{g}_{i1}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) = \partial \ln f_2(y_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = [y_{i2} - \mu(\mathbf{x}_{i2}'\beta, \gamma, \mathbf{x}_{i1}'\boldsymbol{\alpha})]\gamma\, \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})[1 - \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]\mathbf{x}_{i1}.$$

We will use $\mathbf{g}_{i2}^{(2)}$ for computing $\mathbf{V}_2$ and in computing $\mathbf{R}$ and $\mathbf{C}$ and $\mathbf{g}_{i1}^{(2)}$ in computing $\mathbf{C}$. In particular,

$$\mathbf{V}_1 = [(1/n)\Sigma_i \mathbf{g}_{i1}(\boldsymbol{\alpha})\mathbf{g}_{i1}(\boldsymbol{\alpha})']^{-1},$$
$$\mathbf{V}_2 = [(1/n)\Sigma_i \mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})']^{-1},$$
$$\mathbf{C} = [(1/n)\Sigma_i \mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})\mathbf{g}_{i1}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})'],$$
$$\mathbf{R} = [(1/n)\Sigma_i \mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})\mathbf{g}_{i1}(\boldsymbol{\alpha})'].$$

Table 14.2 presents the two-step maximum likelihood estimates of the model parameters and estimated standard errors. For the first-step logistic model, the standard errors marked $\mathbf{H}_1$ vs. $\mathbf{V}_1$ compares the values computed using the negative inverse of the second derivatives matrix ($\mathbf{H}_1$) vs. the outer products of the first derivatives ($\mathbf{V}_1$). As expected with a sample this large, the difference is minor. The latter were used in computing the corrected covariance matrix at the second step. In the Poisson model, the comparison of $\mathbf{V}_2$ to $\mathbf{V}_2^*$ shows distinctly that accounting for the presence of $\hat{\boldsymbol{\alpha}}$ in the constructed regressor has a substantial impact on the standard errors, even in this relatively large sample. Note that the effect of the correction is to double the standard errors on the coefficients for the variables that the equations have in common, but it is quite minor for *Income* and *Female*, which are unique to the second-step model.

TABLE 14.2 Estimated Logistic and Poisson Models

	Logistic Model for Addon			*Poisson Model for DocVis*		
	Coefficient	***Standard Error (H_1)***	***Standard Error (V_1)***	***Coefficient***	***Standard Error (V_2)***	***Standard Error (V_2^*)***
Constant	−6.19246	0.60228	0.58287	0.77808	0.04884	0.09319
Age	0.01486	0.00912	0.00924	0.01752	0.00044	0.00111
Education	0.16091	0.03003	0.03326	−0.03858	0.00462	0.00980
Married	0.22206	0.23584	0.23523			
Kids	−0.10822	0.21591	0.21993			
Income				−0.80298	0.02339	0.02719
Female				0.16409	0.00601	0.00770
$\Lambda(\mathbf{x}_1'\boldsymbol{\alpha})$				3.91140	0.77283	1.87014

The covariance of the two gradients, $\mathbf{R}$, may converge to zero in a particular application. When the first- and second-step estimates are based on different samples, $\mathbf{R}$ is exactly zero. For example, in our earlier application, $\mathbf{R}$ is based on two residuals,

$$\mathbf{g}_{i1} = \{Addon_i - E[Addon_i|\mathbf{x}_{i1}]\} \text{ and } \mathbf{g}_{i2}^{(2)} = \{DocVis_i - E[DocVis_i|\mathbf{x}_{i2}, \boldsymbol{\Lambda}_{i1}]\}.$$

The two residuals may well be uncorrelated. This assumption would be checked on a model-by-model basis, but in such an instance, the third and fourth terms in $\mathbf{V}_2$ vanish asymptotically and what remains is the simpler alternative, $\mathbf{V}_2^{**} = (1/n)[\mathbf{V}_2 + \mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2]$. (In our application, the sample correlation between $\mathbf{g}_{i1}$ and $\mathbf{g}_{i2}^{(2)}$ is only 0.015658 and the elements of the estimate of $\mathbf{R}$ are only about 0.01 times the corresponding elements of $\mathbf{C}$—essentially about 99 percent of the correction in $\mathbf{V}_2^*$ is accounted for by $\mathbf{C}$.)

It has been suggested that this set of procedures might be more complicated than necessary.[18] There are two alternative approaches one might take. First, under general circumstances, the asymptotic covariance matrix of the second-step estimator could be approximated using the bootstrapping procedure that will be discussed in Section 15.4. We would note, however, if this approach is taken, then it is essential that both steps be "bootstrapped." Otherwise, taking $\hat{\boldsymbol{\theta}}_1$ as given and fixed, we will end up estimating $(1/n)\mathbf{V}_2$, not the appropriate covariance matrix. The point of the exercise is to account for the variation in $\hat{\boldsymbol{\theta}}_1$. The second possibility is to fit the full model at once. That is, use a one-step, full information maximum likelihood estimator and estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ simultaneously. Of course, this is usually the procedure we sought to avoid in the first place. And with modern software, this two-step method is often quite straightforward. Nonetheless, this is occasionally a possibility. Once again, Heckman's (1979) famous sample selection model provides an illuminating case. The two-step and full information estimators for Heckman's model are developed in Section 19.4.3.

[18] For example, Cameron and Trivedi (2005, p. 202).

14.8 PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION AND ROBUST ASYMPTOTIC COVARIANCE MATRICES

Maximum likelihood estimation requires complete specification of the distribution of the observed random variable(s). If the correct distribution is something other than what we assume, then the likelihood function is misspecified and the desirable properties of the MLE might not hold. This section considers a set of results on an estimation approach that is robust to some kinds of model misspecification. For example, we have found that if the conditional mean function is $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, then certain estimators, such as least squares, are "robust" to specifying the wrong distribution of the disturbances. That is, LS is MLE if the disturbances are normally distributed, but we can still claim some desirable properties for LS, including consistency, even if the disturbances are not normally distributed. This section will discuss some results that relate to what happens if we maximize the wrong log-likelihood function, and for those cases in which the estimator is consistent despite this, how to compute an appropriate asymptotic covariance matrix for it.[19]

14.8.1 A ROBUST COVARIANCE MATRIX ESTIMATOR FOR THE MLE

A heteroscedasticity robust covariance matrix for the least squares estimator was considered in Section 4.5.2. Based on the general result

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \Sigma_i \mathbf{x}_i \varepsilon_i, \tag{14-32}$$

a robust estimator of the asymptotic covariance matrix for $\mathbf{b}$ would be the White estimator,

$$\text{Est.Asy.Var}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1} [\Sigma_i (\mathbf{x}_i e_i)(\mathbf{x}_i e_i)'](\mathbf{X}'\mathbf{X})^{-1}.$$

If $\text{Var}[\varepsilon_i|\mathbf{x}_i] = \sigma^2$ and $\text{Cov}[\varepsilon_i, \varepsilon_j|\mathbf{X}] = 0$, then we can simplify the calculation to $\text{Est.Asy.Var}[\mathbf{b}] = s^2(\mathbf{X}'\mathbf{X})^{-1}$. But the first form is appropriate in either case—it is robust, at least, to heteroscedasticity. This estimator is not robust to correlation across observations, as in a time series (considered in Chapter 20) or to clustered data (considered in the next section). The variance estimator is robust to omitted variables in the sense that $\mathbf{b}$ estimates something consistently, $\boldsymbol{\gamma}$, though generally not $\boldsymbol{\beta}$, and the variance estimator appropriately estimates the asymptotic variance of $\mathbf{b}$ around $\boldsymbol{\gamma}$. The variance estimator might be similarly robust to endogeneity of one or more variables in $\mathbf{X}$, though, again, the estimator, $\mathbf{b}$, itself does not estimate $\boldsymbol{\beta}$. This point is important for the present context. The variance estimator may still be appropriate for the asymptotic covariance matrix for $\mathbf{b}$, but $\mathbf{b}$ estimates something other than $\boldsymbol{\beta}$.

Similar considerations arise in maximum likelihood estimation. The properties of the maximum likelihood estimator are derived from (14-15). The empirical counterpart to (14-32) is

$$\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0 \approx \left[-\frac{1}{n} \sum_{i=1}^{n} \mathbf{H}_i(\boldsymbol{\theta}_0) \right]^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta}_0) \right), \tag{14-33}$$

[19]Important references on this subject are White (1982a); Gourieroux, Monfort, and Trognon (1984); Huber (1967); and Amemiya (1985). A recent work with a large amount of discussion on the subject is Mittelhammer et al. (2000).

where $\mathbf{g}_i(\boldsymbol{\theta}_0) = \partial \ln f_i/\partial\boldsymbol{\theta}_0$, $\mathbf{H}_i(\boldsymbol{\theta}_0) = \partial^2 \ln f_i/\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0'$ and $\boldsymbol{\theta}_0 = \text{plim}\,\hat{\boldsymbol{\theta}}_{MLE}$. Note that $\boldsymbol{\theta}_0$ is the parameter vector that is estimated by maximizing $\ln L(\boldsymbol{\theta})$, though it might not be the target parameters of the model if the log likelihood is misspecified, the MLE may be inconsistent. Assuming that $\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i(\boldsymbol{\theta}_0) = \overline{\mathbf{H}}$, and the conditions needed for $\sqrt{n}\bar{\mathbf{g}} = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0)\right)$ to obey a central limit theorem are met, the appropriate estimator for the variance of the MLE around $\boldsymbol{\theta}_0$ would be

$$\text{Asy.Var}\left[\hat{\boldsymbol{\theta}}_{MLE}\right] = \left[-\overline{\mathbf{H}}\right]^{-1}\{\text{Asy.Var}[\bar{\mathbf{g}}]\}\left[-\overline{\mathbf{H}}\right]^{-1}. \tag{14-34}$$

The missing element is what to use for the asymptotic variance of $\bar{\mathbf{g}}$. If the information matrix equality (Property D3 in Theorem 14.2) holds, then $\text{Asy.Var}[\bar{\mathbf{g}}] = (-1/n)\overline{\mathbf{H}}$, and we get the familiar result $\text{Asy.Var}\left[\hat{\boldsymbol{\theta}}_{MLE}\right] = \frac{1}{n}\left[-\overline{\mathbf{H}}\right]^{-1}$. However, (14-34) applies whether or not the information matrix equality holds. We can estimate the variance of $\bar{\mathbf{g}}$ with

$$\text{Est.Asy.Var}[\bar{\mathbf{g}}] = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)'\right]. \tag{14-35}$$

The variance estimator for the MLE is then

$$\begin{aligned}&\text{Est.Asy.Var}\left[\hat{\boldsymbol{\theta}}_{MLE}\right]\\ &= \left[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)\right]^{-1}\left\{\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)'\right]\right\}\left[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i\left(\hat{\boldsymbol{\theta}}_{MLE}\right)\right]^{-1}.\end{aligned} \tag{14-36}$$

This is a robust covariance matrix for the maximum likelihood estimator.

If $\ln L(\boldsymbol{\theta}_0|\mathbf{y},\mathbf{X})$ is the appropriate conditional log likelihood, then the MLE is a consistent estimator of $\boldsymbol{\theta}_0$ and, because of the information matrix equality, the asymptotic variance of the MLE is $(1/n)$ times the bracketed term in (14-33). The issue of robustness would relate to the behavior of the estimator of $\boldsymbol{\theta}_0$ if the likelihood were misspecified. We assume that the function we are maximizing (we would now call it the *pseudo-log likelihood*) is regular enough that the maximizer that we compute converges to a parameter vector, $\boldsymbol{\beta}$. Then, by the results above, the asymptotic variance of the estimator is obtained without use of the information matrix equality. As in the case of least squares, there are two levels of robustness to be considered. To argue that the estimator, itself, is robust in this context, it must first be argued that the estimator is consistent for something that we want to estimate and that maximizing the wrong log likelihood nonetheless estimates the right parameter(s). If the model is not linear, this will generally be much more complicated to establish. For example, in the leading case, for a binary choice model, if one assumes that the probit model applies, and some other model applies, then the estimator is not robust to any of heteroscedasticity, omitted variables, autocorrelation, endogeneity, fixed or random effects, or the wrong distribution. (It is difficult to think of a model failure that the MLE is robust to.) Once the estimator, itself, is validated, then the robustness of the asymptotic covariance matrix is considered.[20]

[20]There is a trend in the current literature routinely to report "robust standard errors," based on (14-36) regardless of the likelihood function (which defines the model).

Example 14.6 A Regression with NonNormal Disturbances

If one believed that the regression disturbances were more widely dispersed than implied by the normal distribution, then the logistic or t distribution might provide an alternative specification. We consider the logistic. The model is

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, f(\varepsilon) = \frac{1}{\sigma}\frac{\exp(\varepsilon/\sigma)}{[1 + \exp(\varepsilon/\sigma)]^2} = \frac{1}{\sigma}\frac{\exp(w)}{[1 + \exp(w)]^2} = \frac{1}{\sigma}\Lambda(w)[1 - \Lambda(w)],$$

where $\wedge(w)$ is the logistic CDF. The logistic distribution is symmetric, as is the normal, but has a greater variance, $(\pi^2/3)\sigma^2$ compared to σ^2 for the normal, and greater kurtosis (tail thickness), 4.2 compared to 3.0 for the normal. Overall, the logistic distribution resembles a t distribution with 8 degrees of freedom, which has kurtosis 4.5 and variance $(4/3)\sigma^2$. The three densities for the standardized variable are shown in Figure 14.3.

The log-likelihood function is

$$\ln L(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \{-\ln \sigma + w_i - 2\ln[1 + \exp(w_i)]\}, w_i = (y_i - \mathrm{x}_i'\boldsymbol{\beta})/\sigma. \quad \textbf{(14-37)}$$

The terms in the gradient and Hessian are

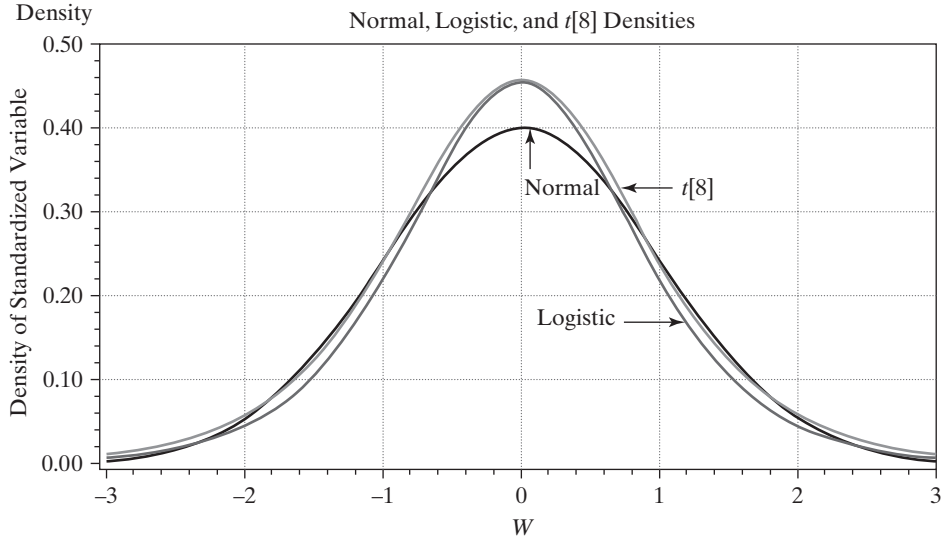
$$\mathbf{g}_i = \frac{-(1 - 2\Lambda(w_i))}{\sigma}\begin{pmatrix}\mathbf{x}_i\\ w_i\end{pmatrix} - \frac{1}{\sigma}\begin{pmatrix}\mathbf{0}\\ 1\end{pmatrix},$$

$$\mathbf{H}_i = \frac{-2\Lambda(w_i)(1 - \Lambda(w_i))}{\sigma^2}\begin{pmatrix}\mathbf{x}_i\\ w_i\end{pmatrix}\begin{pmatrix}\mathbf{x}_i\\ w_i\end{pmatrix}' + \frac{(1 - 2\Lambda(\mathrm{w}_i))}{\sigma^2}\begin{bmatrix}0 & \mathbf{x}_i\\ \mathbf{x}_i' & 2w_i\end{bmatrix} + \frac{1}{\sigma^2}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0}' & 1\end{bmatrix}.$$

The conventional estimator of the asymptotic covariance matrix of $\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\sigma}\end{pmatrix}$ would be $\left[-\sum_{i=1}^{n}\hat{\mathbf{H}}_i\right]^{-1}$.

The robust estimator would be

FIGURE 14.3 Standardized Normal, Logistic, and t[8] Densities.



$$\text{Est.Asy.Var}\left[\begin{matrix}\hat{\boldsymbol{\beta}}\\ \hat{\sigma}\end{matrix}\right] = \left[-\sum_{i=1}^{n}\hat{\mathbf{H}}_i\right]^{-1}\left[\sum_{i=1}^{n}\hat{\mathbf{g}}_i\hat{\mathbf{g}}_i'\right]\left[-\sum_{i=1}^{n}\hat{\mathbf{H}}_i\right]^{-1}.$$

The data in Appendix F14.1 are a panel of 247 dairy farms in Northern Spain, observed for 6 years, 1993–1998. The model is a simple Cobb–Douglas production function,

$$\ln y_{it} = \beta_0 + \beta_1 \ln x_{1,it} + \beta_2 \ln x_{2,it} + \beta_3 \ln x_{3,it} + \beta_4 \ln x_{4,it} + \varepsilon_{it},$$

where y_{it} is the log of milk production, $x_{1,it}$ is number of cows, $x_{2,it}$ is land in hectares, $x_{3,it}$ is labor, and $x_{4,it}$ is feed. The four inputs are transformed to logs, then to deviations from the means of the logs. We then estimated $\boldsymbol{\beta}$ and σ by maximizing the log likelihood for the logistic distribution. Results are shown in Table 14.3. Standard errors are computed using $\left[-\Sigma_i\Sigma_t\hat{\mathbf{H}}_{it}\right]^{-1}$. The robust standard errors shown in column (4) are based on (14-36). They are nearly identical to the uncorrected standard errors, which suggests that the departure of the logistic distribution from the true underlying model or the influence of heteroscedasticity are minor. Column (5) reports the cluster robust standard errors based on (14-38) discussed in the next section.

The departure of the data from the logistic distribution assumed in the likelihood function seems to be minor. The log likelihood does favor the logistic distribution; however, the models cannot be compared on this basis, because the test would have zero degrees of freedom—the models are not nested. The Vuong test examined in Section 14.6.6 might be helpful. The individual terms in the log likelihood are computed using (14-37). For the normal distribution, the term in the log likelihood would be $\ln f_{it} = -(1/2)[\ln 2\pi + \ln s^2 + (y_{it} - \mathbf{x}_{it}'\mathbf{b})^2/s^2]$ where $s^2 = \mathbf{e}'\mathbf{e}/n$. Using $d_{it} = (\ln f_{it}|\text{logistic} - \ln f_{it}|\text{normal})$, the test statistic is $V = \sqrt{n}\bar{d}/s_d = 1.682$. This slightly favors the logistic distribution, but is in the inconclusive region. We conclude that for these data, the normal and logistic models are essentially indistinguishable.

14.8.2 CLUSTER ESTIMATORS

Micro-level, or individual, data are often grouped or clustered. A model of production or economic success at the firm level might be based on a group of industries, with multiple

TABLE 14.3 Maximum Likelihood Estimates of a Production Function

Estimate	*(1) Least Squares*	*(2) MLE Logistic*	*(3) Standard Error*	*(4) Robust Std. Error*	*(5) Clustered Std.Error*
β_0	11.5775	11.5826	0.00353	0.00364	0.00751
β_1	0.59518	0.58696	0.01944	0.02124	0.03697
β_2	0.02305	0.02753	0.01086	0.01104	0.01924
β_3	0.02319	0.01858	0.01248	0.01226	0.02325
β_4	0.45176	0.45671	0.01069	0.01160	0.02071
σ	0.14012[a]	0.07807	0.00169	0.00164	0.00299
R^2	0.92555	0.95253[b]			
$\ln L$	809.676	821.197			

[a]MLE of $\sigma^2 = \mathbf{e}'\mathbf{e}/n$.
[b]R^2 is computed as the squared correlation between predicted and actual values.

firms in each industry. Analyses of student educational attainment might be based on samples of entire classes, or schools, or statewide averages of schools within school districts. And, of course, such "clustering" is the defining feature of a panel data set. We considered several of these types of applications in Section 4.5.3 and in our analysis of panel data in Chapter 11. The recent literature contains many studies of clustered data in which the analyst has estimated a pooled model but sought to accommodate the expected correlation across observations with a correction to the asymptotic covariance matrix. We used this approach in computing a robust covariance matrix for the pooled least squares estimator in a panel data model [see (11-3) and Examples 11.7 and 11.11].

For the normal linear regression model, the log likelihood that we maximize with the pooled least squares estimator is

$$\ln L = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2}\frac{(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta})^2}{\sigma^2}\right].$$

By multiplying and dividing by $(\sigma^2)^2$, the "cluster-robust" estimator in (11-3) can be written

$$\begin{aligned}\mathbf{W} &= \left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left[\sum_{i=1}^{n}(\mathbf{X}_i'\mathbf{e}_i)(\mathbf{e}_i'\mathbf{X}_i)\right]\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\\ &= \left(-\frac{1}{\hat\sigma^2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T_i}\frac{1}{\hat\sigma^2}\mathbf{x}_{it}e_{it}\right)\left(\sum_{t=1}^{T_i}\frac{1}{\hat\sigma^2}e_{it}\mathbf{x}_{it}'\right)\right]\left(-\frac{1}{\hat\sigma^2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1}.\end{aligned}$$

The terms in the second line are the first and second derivatives of $\ln f_{it}$ for the normal distribution mean $\mathbf{x}_{it}'\boldsymbol{\beta}$ and variance σ^2 shown in (14-3). A general form of the result is

$$W = \left(\sum_{i=1}^{n}\sum_{t=1}^{T_i}\frac{\partial^2\ln f_{it}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T_i}\frac{\partial\ln f_{it}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}}\right)\left(\sum_{t=1}^{T_i}\frac{\partial\ln f_{it}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}'}\right)\right]\left(\sum_{i=1}^{n}\sum_{t=1}^{T_i}\frac{\partial^2\ln f_{it}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}. \quad \textbf{(14-38)}$$

This form of the correction would account for unspecified correlation across the observations (the derivatives) within the groups. [The finite population correction in (11-4) is sometimes applied.]

Example 14.7 Cluster Robust Standard Errors

The dairy farm data used in Example 14.6 are a panel of 247 farms observed in 6 consecutive years. A correction of the standard errors for possible group effects would be natural. Column (5) of Table 14.3 shows the standard errors computed using (14-38). The corrected standard errors are nearly double the values in column (5). This suggests that although the distributional specification is reasonable, there does appear to be substantial correlation across the observations.We will examine this feature of the data further in Section 19.2.4 in the discussion of the stochastic production frontier model.

Consider the specification error that the estimator is intended to accommodate for the normal linear regression. Suppose that the observations in group i were multivariate normally distributed with disturbance mean vector zero and unrestricted $T_i \times T_i$ covariance matrix, $\boldsymbol{\Sigma}_i$. Then, the appropriate log-likelihood function would be

$$\ln L = \sum_{i=1}^{n}(-T_i/2\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}\boldsymbol{\varepsilon}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\varepsilon}_i),$$

where $\boldsymbol{\varepsilon}_i$ is the $T_i \times 1$ vector of disturbances for individual i. Therefore, by using pooled least squares, we have maximized the wrong likelihood function. Indeed, the $\boldsymbol{\beta}$ that maximizes this log-likelihood function is the GLS estimator (see Chapter 9), not the OLS estimator. But OLS and the cluster corrected estimator given earlier "work" in the sense that (1) the least squares estimator is consistent in spite of the misspecification and (2) the robust estimator does, indeed, estimate the appropriate asymptotic covariance matrix.

Now, consider the more general case. Suppose the data set consists of n multivariate observations, $[y_{i,1}, \ldots, y_{i,T_i}]$, $i = 1, \ldots, n$. Each cluster is a draw from joint density $f_i(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\theta})$. Once again, to preserve the generality of the result, we will allow the cluster sizes to differ. The appropriate log likelihood for the sample is

$$\ln L = \sum_{i=1}^{n} \ln f_i(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\theta}).$$

Instead of maximizing $\ln L$, we maximize a pseudo-log likelihood

$$\ln L_P = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln g(y_{it} | \mathbf{x}_{it}, \boldsymbol{\theta}),$$

where we make the possibly unreasonable assumption that the same parameter vector, $\boldsymbol{\theta}$, enters the pseudo-log likelihood as enters the correct one. Using our familiar first-order asymptotics, the **pseudo-maximum likelihood estimator** (MLE) will satisfy

$$(\hat{\boldsymbol{\theta}}_{P,ML} - \boldsymbol{\theta}) \approx \left(\frac{-1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{\partial^2 \ln f_{it}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \left(\frac{-1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{\partial \ln f_{it}}{\partial \boldsymbol{\theta}} \right) + (\boldsymbol{\theta} - \boldsymbol{\beta})$$

$$= \left(\frac{-1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \mathbf{H}_{it} \right)^{-1} \left(\sum_{i=1}^{n} a_i \bar{\mathbf{g}}_i \right) + (\boldsymbol{\theta} - \boldsymbol{\beta}),$$

where $a_i = T_i / \sum_{i=1}^{n} T_i$ and $\bar{\mathbf{g}}_i = (1/T_i) \sum_{t=1}^{T_i} \partial \ln f_{it} / \partial \boldsymbol{\theta}$. The trailing term in the expression is included to allow for the possibility that plim $\hat{\boldsymbol{\theta}}_{P,ML} = \boldsymbol{\beta}$, which may not equal $\boldsymbol{\theta}$.[21] Taking the expected outer product of this expression to estimate the asymptotic mean squared deviation will produce two terms—the cross term vanishes. The first will be the cluster-corrected matrix that is ubiquitous in the current literature. The second will be the squared error that may persist as n increases because the pseudo-MLE need not estimate the parameters of the model of interest.

We draw two conclusions. We can justify the cluster estimator based on this approximation. In general, it will estimate the expected squared variation of the pseudo-MLE around its probability limit. Whether it measures the variation around the appropriate parameters of the model hangs on whether the second term equals zero. In words, perhaps not surprisingly, this apparatus only works if the pseudo-MLE is consistent. Is that likely? Certainly not if the pooled model is ignoring unobservable fixed effects. Moreover, it will be inconsistent in most cases in which the misspecification is to ignore latent random effects as well. The pseudo-MLE is only consistent for random effects in a few special

[21] Note, for example, Cameron and Trivedi (2005, p. 842) specifically assume consistency in the generic model they describe.

cases, such as the linear model and Poisson and negative binomial models discussed in Chapter 18. It is not consistent in the probit and logit models in which this approach is often used. In the end, the cases in which the estimator are consistent are rarely, if ever, enumerated. The upshot is stated succinctly by Freedman (2006, p. 302): "The sandwich algorithm, under stringent regularity conditions, yields variances for the MLE that are asymptotically correct even when the specification—and hence the likelihood function—are incorrect. However, it is quite another thing to ignore bias. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter."

14.9 MAXIMUM LIKELIHOOD ESTIMATION OF LINEAR REGRESSION MODELS

We will now examine several applications of the MLE. We begin by developing the ML counterparts to most of the estimators for the classical and generalized regression models in Chapters 4 through 11. (Generally, the development for dynamic models becomes more involved than we are able to pursue here. The one exception we will consider is the standard model of autocorrelation.) We emphasize, in each of these cases, that we have already developed an efficient, generalized method of moments estimator that has the same asymptotic properties as the MLE under the assumption of normality. In more general cases, we will sometimes find that the GMM estimator is actually preferred to the MLE because of its robustness to failures of the distributional assumptions or its freedom from the necessity to make those assumptions in the first place. However, for the extensions of the classical model based on generalized least sqaures that are treated here, that is not the case. It might be argued that in these cases, the MLE is superfluous. There are occasions when the MLE will be preferred for other reasons, such as its invariance to transformation in nonlinear models and, possibly, its small sample behavior (although that is usually not the case). And, we will examine some nonlinear models in which there is no linear method of moments counterpart, so the MLE is the natural estimator. Finally, in each case, we will find some useful aspect of the estimator itself, including the development of algorithms such as Newton's method and the EM method for latent class models.

14.9.1 LINEAR REGRESSION MODEL WITH NORMALLY DISTRIBUTED DISTURBANCES

The linear regression model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i.$$

The likelihood function for a sample of n independent, identically, and normally distributed disturbances is

$$L = (2\pi\sigma^2)^{-n/2} e^{-\varepsilon'\varepsilon/(2\sigma^2)}.$$

The transformation from ε_i to y_i is $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, so the Jacobian for each observation, $|\partial\varepsilon_i/\partial y_i|$, is one.[22] Making the transformation, we find that the likelihood function for the n observations on the observed random variables is

$$L = (2\pi\sigma^2)^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}.$$

[22]See (B-41) in Section B.5. The analysis to follow is conditioned on $\mathbf{X}$. To avoid cluttering the notation, we will leave this aspect of the model implicit in the results. As noted earlier, we assume that the data-generating process for X does not involve $\boldsymbol{\beta}$ or σ^2 and that the data are well behaved as discussed in Chapter 4.

To maximize this function with respect to $\boldsymbol{\beta}$, it will be necessary to maximize the exponent or minimize the familiar sum of squares. Taking logs, we obtain the log-likelihood function for the classical regression model,

$$\begin{aligned}\ln L &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\ln 2\pi + \ln\sigma^2 + (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2/\sigma^2\right]. \end{aligned} \tag{14-39}$$

The necessary conditions for maximizing this log likelihood are

$$\begin{bmatrix}\dfrac{\partial \ln L}{\partial \boldsymbol{\beta}}\\ \dfrac{\partial \ln L}{\partial \sigma^2}\end{bmatrix} = \begin{bmatrix}\dfrac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2}\\ \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^4}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ 0\end{bmatrix}.$$

The values that satisfy these equations are

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b} \quad \text{and} \quad \hat{\sigma}^2_{\mathrm{ML}} = \frac{\mathbf{e}'\mathbf{e}}{n}.$$

The slope estimator is the familiar one, whereas the variance estimator differs from the least squares value by the divisor of n instead of $n - K$.[23]

The Cramér–Rao bound for the variance of an unbiased estimator is the negative inverse of the expectation of

$$\begin{bmatrix}\dfrac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \dfrac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\sigma^2}\\ \dfrac{\partial^2 \ln L}{\partial\sigma^2\partial\boldsymbol{\beta}'} & \dfrac{\partial^2 \ln L}{\partial(\sigma^2)^2}\end{bmatrix} = \begin{bmatrix}-\dfrac{\mathbf{X}'\mathbf{X}}{\sigma^2} & -\dfrac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sigma^4}\\ -\dfrac{\boldsymbol{\varepsilon}'\mathbf{X}}{\sigma^4} & \dfrac{n}{2\sigma^4} - \dfrac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^6}\end{bmatrix}.$$

In taking expected values, the off-diagonal term vanishes, leaving

$$[\mathbf{I}(\boldsymbol{\beta},\sigma^2)]^{-1} = \begin{bmatrix}\sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0}\\ \mathbf{0}' & 2\sigma^4/n\end{bmatrix}.$$

The least squares slope estimator is the maximum likelihood estimator for this model. Therefore, it inherits all the desirable *asymptotic* properties of maximum likelihood estimators.

We showed earlier that $s^2 = \mathbf{e}'\mathbf{e}/(n-K)$ is an unbiased estimator of σ^2. Therefore, the maximum likelihood estimator is biased toward zero,

$$E\left[\hat{\sigma}^2_{\mathrm{ML}}\right] = \frac{n-K}{n}\sigma^2 = \left(1 - \frac{K}{n}\right)\sigma^2 < \sigma^2. \tag{14-40}$$

Despite its small-sample bias, the maximum likelihood estimator of σ^2 has the same desirable asymptotic properties. We see in (14-40) that s^2 and $\hat{\sigma}^2$ differ only by a factor $-K/n$, which

[23] As a general rule, maximum likelihood estimators do not make corrections for degrees of freedom.

vanishes in large samples. It is instructive to formalize the asymptotic equivalence of the two. From (14-40), we know that

$$\sqrt{n}(\hat{\sigma}^2_{\mathrm{ML}} - \sigma^2) \xrightarrow{d} N[0, 2\sigma^4].$$

It follows that

$$z_n = \left(1 - \frac{K}{n}\right)\sqrt{n}\left(\hat{\sigma}^2_{\mathrm{ML}} - \sigma^2\right) + \frac{K}{\sqrt{n}}\sigma^2 \xrightarrow{d} \left(1 - \frac{K}{n}\right)N[0, 2\sigma^4] + \frac{K}{\sqrt{n}}\sigma^2.$$

But $K/\sqrt{n}$ and K/n vanish as $n \to \infty$, so the limiting distribution of z_n is also $N[0, 2\sigma^4]$. Because $z_n = \sqrt{n}(s^2 - \sigma^2)$, we have shown that the asymptotic distribution of s^2 is the same as that of the maximum likelihood estimator.

14.9.2 SOME LINEAR MODELS WITH NONNORMAL DISTURBANCES

The log-likelihood function for a linear regression model with normally distributed disturbances is

$$\begin{aligned} \ln L_N(\boldsymbol{\beta}, \sigma) &= \sum_{i=1}^{n}\{-\ln \sigma - (1/2)\ln 2\pi - (1/2)w_i^2\}, \\ w_i &= (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma, \sigma > 0. \end{aligned} \tag{14-41}$$

Example 14.6 considers maximum likelihood estimation of a linear regression model with logistically distributed disturbances. The appeal of the logistic distribution is its greater degree of kurtosis—its tails are thicker than those of the normal distribution. The log-likelihood function is

$$\begin{aligned} \ln L_L(\boldsymbol{\beta}, \sigma) &= \sum_{i=1}^{n}\{-\ln \sigma + w_i - 2\ln[1 + \exp(w_i)]\}, \\ w_i &= (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma, \sigma > 0. \end{aligned} \tag{14-42}$$

The logistic specification fixes the shape of the distribution, as suggested earlier, similar to a $t[8]$ distribution. The t distribution with an unrestricted degrees of freedom parameter (a special case of the generalized hyperbolic distribution) allows greater flexibility in this regard. The t distribution arises as the distribution of a sum of δ squares of normally distributed variables. But the degrees of freedom parameter need not be integer valued. We allow δ to be a free parameter, though greater than 4 for the first four moments to be finite. The density of a standardized t distributed random variable with degrees of freedom parameter δ is

$$f(w|\delta, \sigma) = \frac{\Gamma[(\delta + 1)/2]}{\Gamma(\delta/2)\Gamma(1/2)\sqrt{\pi\delta}}\frac{1}{\sigma}\left[1 + \frac{w^2}{\delta}\right]^{-(\delta+1)/2}.$$

The log-likelihood function is

$$\begin{aligned} \ln L_t(\boldsymbol{\beta}, \sigma, \delta) &= \sum_{i=1}^{n}\begin{pmatrix} -\ln \sigma + \ln \Gamma[(\delta + 1)/2] - \ln \Gamma(\delta/2) \\ -\ln \Gamma(1/2)] - (1/2)\ln \pi - (1/2)\ln \delta \\ -[(\delta + 1)/2]\ln(1 + w_i^2/\delta) \end{pmatrix}, \\ w_i &= (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma, \sigma > 0, \delta > 4. \end{aligned} \tag{14-43}$$

The centerpiece of the stochastic frontier model (Example 12.2 and Section 19.2.4) is a skewed distribution, the skew normal distribution,

$$f(w|\lambda,\sigma) = \frac{2}{\sigma\sqrt{2\pi}}\exp[-(1/2)w^2]\Phi(-\lambda w),\ \lambda \geq 0,$$

where $\Phi(z)$ is the CDF of the standard normal distribution. If the skewness parameter, λ, equals zero, this returns the standard normal distribution. The skew normal distribution arises as the distribution of $\varepsilon = \sigma_v v_i - \sigma_u|u_i|$, where v_i and u_i are standard normal variables, $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_v^2 + \sigma_u^2$. [Note that σ^2 is not the variance of ε. The variance $|u_i|$ is $(\pi - 2)/\pi$, not 1.] The log-likelihood function is

$$\ln L_{SN}(\boldsymbol{\beta},\sigma,\lambda) = \sum_{i=1}^{n}\{-\ln\sigma - (1/2)\ln(\pi/2) - (1/2)w_i^2 + \ln\Phi(-\lambda w_i)\},\ w_i = (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma. \quad \textbf{(14-44)}$$

Example 14.8 Logistic, t, and Skew Normal Disturbances

Table 14.4 shows the maximum likelihood estimates for the four models. There are only small differences in the slope estimators, as might be expected, at least for the first three, because the differences are in the spread of the distribution, not its shape. The skew normal density has a nonzero mean, $E[\sigma_u|u_i|) = (2/\pi)^{1/2}\sigma_u$, so the constant term has been adjusted. As noted, it is not possible directly to test the normal as a restriction on the logistic, as they have the same number of parameters. The Vuong test does not distinguish them. The *t* distribution would seem to be amenable to a direct specification test; however, the "restriction" on the *t* distribution that produces the normal is $\delta \to \infty$ which is not useable. However, we can exploit the invariance of the maximum likelihood estimator (property M4 in Table 14.1). The maximum likelihood estimator of $1/\delta$ is $1/\hat{\delta}_{MLE} = 0.101797 = \hat{\gamma}$. We can use the delta method to obtain a standard error. The estimated standard error will be $(1/\hat{\delta}_{MLE})^2(2.54296) = 0.026342$. A Wald test of H_0: $\hat{\gamma} = 0$ would test the normal versus the *t* distribution. The result is $[(0.101797 - 0)/0.026342]^2 = 14.934$, which is larger than the critical value of 3.84, so the hypothesis of normality is rejected. [There is a subtle problem with this test. The value $\gamma = 0$ is on the boundary of the parameter space, not the interior. As such, the chi-squared statistic does not have its usual properties. This issue is explored in Kodde and Palm (1988) and Coelli (1995), who suggest that an appropriate critical value for a single restriction would be 2.706, rather than 3.84.[24] The same consideration applies to the test of $\lambda = 0$ below.] We note, because the log-likelihood function could have been parameterized in terms of γ to begin with, we should be able to use a likelihood ratio test to test the same hypothesis. By the invariance result, the log likelihood in terms of γ would not change, so the test statistic will be $\lambda_{LR} = -2(809.676 - 822.192) = 25.032$. This produces the same conclusion. The normal distribution is nested within the skew normal, by $\lambda = 0$ or $\sigma_u = 0$. We can test the first of these with a likelihood ratio test; $\lambda_{LR} = -2(809.676 - 822.688) = 26.024$. The Wald statistic based on the derived estimate of σ_u would be $(0.15573/0.00279)^2 = 3115.56$.[25] The conclusion is the same for both cases. As noted, the *t* and logistic are essentially indistinguishable. The

[24]The critical value is found by solving for c in $.05 = (1/2)\text{Prob}(\chi^2[1] \geq c]$. For a chi-squared variable with one degree of freedom, the 90th percentile is 2.706.

[25]Greene and McKenzie (2015) show that for the stochastic frontier model examined here, the LM test for the hypothesis that $\sigma_u = 0$ can be based on the OLS residuals; the chi-squared statistic with one degree of freedom is $(n/6)(m_3/s^3)^2$ where m_3 is the third moment of the residuals and s^2 equals $\mathbf{e}'\mathbf{e}/n$. The value for this data set is 21.665.

remaining question, then, is whether the respecification of the model favors skewness or kurtosis. We do not have a direct statistical test available. The OLS estimator of $\boldsymbol{\beta}$ is consistent regardless, so some information might be contained in the residuals. Figure 14.4 compares the OLS residuals to the normal distribution with the same mean (zero) and standard deviation (0.14012). The figure does suggest the presence of skewness, not excess spread. Given the nature of the production function application, skewness is central to this model, so the findings so far might be expected. The development of the stochastic production frontier model is continued in Section 19.2.4.

14.9.3 HYPOTHESIS TESTS FOR REGRESSION MODELS

The standard test statistic for assessing the validity of a set of linear restrictions, $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$, the linear model with normally distributed disturbances is the F ratio,

$$F[J, n - K] = \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{J}. \tag{14-45}$$

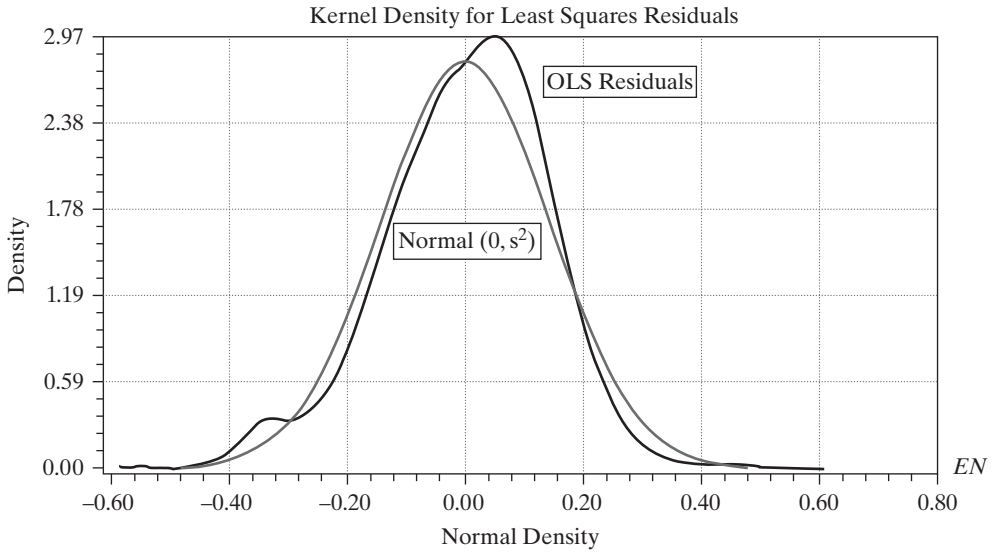
TABLE 14.4 Maximum Likelihood Estimates

(Estimated standard errors in parentheses)

Estimate	*OLS/MLE Normal*	*MLE Logistic*	*MLE t Frac. D.F.*	*MLE Skew Normal*
β_0	11.5775	11.5826	11.5813	11.6966[c]
	(0.00365)	(0.00353)	(0.00363)	(0.00447)
β_1	0.59518	0.58696	0.59042	0.58369
	(0.01958)	(0.01944)	(0.01803)	(0.01887)
β_2	0.02305	0.02753	0.02576	0.03555
	(0.01122)	(0.01086)	(0.01096)	(0.01113)
β_3	0.02319	0.01858	0.01971	0.02256
	(0.01303)	(0.01248)	(0.01299)	(0.01281)
β_4	0.45176	0.45671	0.45220	0.44948
	(0.01078)	(0.01069)	(0.00989)	(0.01035)
σ	0.14012[a]	0.07807	0.12519	0.13988[d]
	(0.00275)	(0.00169)	(0.00404)	(0.00279)
δ			9.82350	
			(2.54296)	
λ				1.50164
				(0.08748)
σ_u				0.15573[e]
				(0.00279)
R^2	0.92555	0.95253[b]	0.95254[b]	0.95250[b]
ln L	809.676	821.197	822.192	822.688

[a]MLE of $\sigma = \mathbf{e}'\mathbf{e}/n$.
[b]R^2 is computed as the squared correlation between predicted and actual values.
[c]Nonzero mean disturbance. Adjustment to β_0 is $\sigma_u(2/\pi)^{1/2} = -0.04447$.
[d]Reported $\sigma_\varepsilon = [\sigma_v^2 + \sigma_u^2(\pi - 2)/\pi]^{1/2}$. Estimated $\sigma_v = 0.10371$ (0.00418).
[e]σ_u is derived. $\sigma_u = \sigma\lambda/(1 + \lambda^2)^{1/2}$. Est.Cov$(\hat{\sigma}, \hat{\lambda}) = 2.3853\text{e}-7$. Standard error is computed using the delta method.

FIGURE 14.4 Distribution of Least Squares Residuals.



With normally distributed disturbances, the F test is valid in any sample size. The more general form of the statistic,

$$F[J, n - K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n - K)}, \tag{14-46}$$

is useable in large samples when the disturbances are homoscedastic even if the disturbances are not normally distributed and with nonlinear restrictions of the general form $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{0}$. In the linear regression setting with linear restrictions, the Wald statistic, $\mathbf{c}(\mathbf{b})'\{\text{Asy.Var}[\mathbf{c}(\mathbf{b})]\}^{-1}\mathbf{c}(\mathbf{b})$, equals $J \times F[J, n - K]$, so the large-sample validity extends beyond normal linear model. (See Sections 5.3.1 and 5.3.2.)

In this section, we will reconsider the Wald statistic and examine two related statistics, the likelihood ratio and Lagrange multiplier statistics. These statistics are both based on the likelihood function and, like the Wald statistic, are generally valid only asymptotically. No simplicity is gained by restricting ourselves to linear restrictions at this point, so we will consider general hypotheses of the form

$$H_0\text{: } \mathbf{c}(\boldsymbol{\beta}) = \mathbf{0},$$
$$H_1\text{: } \mathbf{c}(\boldsymbol{\beta}) \neq \mathbf{0}.$$

The Wald statistic for testing this hypothesis and its limiting distribution under H_0 would be

$$W = \mathbf{c}(\mathbf{b})'\{\mathbf{G}(\mathbf{b})[\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{G}(\mathbf{b})'\}^{-1}\mathbf{c}(\mathbf{b}) \xrightarrow{d} \chi^2[J],$$

where $\mathbf{G}(\mathbf{b}) = [\partial\mathbf{c}(\mathbf{b})/\partial\mathbf{b}']$.

The Wald statistic is based on the asymptotic distribution of the estimator. The covariance matrix can be replaced with any valid estimator of the asymptotic covariance. Also, for the same reason, the same distributional result applies to estimators based on the nonnormal distributions in Example 14.7, and indeed, for any estimator in any model setting in which $\hat{\boldsymbol{\beta}} \xrightarrow{a} N[\boldsymbol{\beta}, \mathbf{V}]$. The general result, then, is

$$W = \mathbf{c}(\hat{\boldsymbol{\beta}})'\{\mathbf{G}(\hat{\boldsymbol{\beta}})[\text{Asy.Var}(\hat{\boldsymbol{\beta}})]G(\hat{\boldsymbol{\beta}})'\}^{-1}\mathbf{c}(\hat{\boldsymbol{\beta}}) \xrightarrow{d} \chi^2[J]. \tag{14-47}$$

The Wald statistic is robust in that it relies on the large sample distribution of the estimator, not on the specific distribution that underlies the likelihood function. The Wald test will be the statistic of choice in a variety of settings, not only the likelihood-based one considered here.

The **likelihood ratio (LR) test** is carried out by comparing the values of the log-likelihood function with and without the restrictions imposed. We leave aside for the present how the restricted estimator $\mathbf{b}^*$ is computed (except for the linear model, which we saw earlier). The test statistic and its limiting distribution under H_0 are

$$LR = -2[\ln L_* - \ln L] \xrightarrow{d} \chi^2[J]. \tag{14-48}$$

This result is general for any nested models fit by maximum likelihood. The log likelihood for the normal/linear regression model is given in (14-39). The first-order conditions imply that regardless of how the slopes are computed, the estimator of σ^2 without restrictions on $\boldsymbol{\beta}$ will be $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})/n$ and likewise for a restricted estimator $\hat{\sigma}_*^2 = (\mathbf{y} - \mathbf{Xb}_*)'(\mathbf{y} - \mathbf{Xb}_*)/n = \mathbf{e}_*'\mathbf{e}_*/n$. Evaluated at the maximum likelihood estimator, the **concentrated log likelihood**[26] will be

$$\ln L_c = -\frac{n}{2}[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)]$$

and likewise for the restricted case. If we insert these in the definition of LR, then we obtain

$$LR = n \ln[\mathbf{e}_*'\mathbf{e}_*/\mathbf{e}'\mathbf{e}] = n(\ln \hat{\sigma}_*^2 - \ln \hat{\sigma}^2) = n \ln(\hat{\sigma}_*^2/\hat{\sigma}^2). \tag{14-49}$$

(Note, this is a specific result that applies to the linear or nonlinear regression model with normally distributed disturbances.)

The **Lagrange multiplier (LM) test** is based on the gradient of the log-likelihood function. The principle of the test is that if the hypothesis is valid, then at the restricted estimator, the derivatives of the log-likelihood function should be close to zero. There are two ways to carry out the LM test. The log-likelihood function can be maximized subject to a set of restrictions by using

$$\ln L_{\text{LM}} = -\frac{n}{2}\left[\ln 2\pi + \ln \sigma^2 + \frac{[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]/n}{\sigma^2}\right] + \boldsymbol{\lambda}'\mathbf{c}(\boldsymbol{\beta}).$$

[26] See Section E4.3.

The first-order conditions for a solution are

$$\begin{bmatrix} \dfrac{\partial \ln L_{\text{LM}}}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ln L_{\text{LM}}}{\partial \sigma^2} \\ \dfrac{\partial \ln L_{LM}}{\partial \boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} + \mathbf{G}(\boldsymbol{\beta})'\boldsymbol{\lambda} \\ \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \\ \mathbf{c}(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \end{bmatrix}. \tag{14-50}$$

The solutions to these equations give the restricted least squares estimator, $\mathbf{b}_*$; the usual variance estimator, now $\mathbf{e}_*'\mathbf{e}_*/n$; and the Lagrange multipliers. There are now two ways to compute the test statistic. In the setting of the classical linear regression model, when we actually compute the Lagrange multipliers, a convenient way to proceed is to test the hypothesis that the multipliers equal zero. For this model, the solution for $\boldsymbol{\lambda}_*$ is $\boldsymbol{\lambda}_* = [\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}']^{-1}(\mathbf{G}\mathbf{b} - \mathbf{q})$. This equation is a linear function of the unrestricted least squares estimator. If we carry out a Wald test of the hypothesis that $\boldsymbol{\lambda}_*$ equals $\mathbf{0}$, then the statistic will be

$$\text{LM} = \boldsymbol{\lambda}_*'\{\text{Est.Var}[\boldsymbol{\lambda}_*]\}^{-1}\boldsymbol{\lambda}_* = (\mathbf{G}\mathbf{b} - \mathbf{q})'[\mathbf{G}s_*^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}']^{-1}(\mathbf{G}\mathbf{b} - \mathbf{q}). \tag{14-51}$$

The disturbance variance estimator, s_*^2, based on the restricted slopes is $\mathbf{e}_*'\mathbf{e}_*/n$.

An alternative way to compute the LM statistic for the linear regression model produces an interesting result. In most situations, we maximize the log-likelihood function without actually computing the vector of Lagrange multipliers. (The restrictions are usually imposed some other way.) An alternative way to compute the statistic is based on the (general) result that under the hypothesis being tested,

$$E[\partial \ln L/\partial \boldsymbol{\beta}] = E[(1/\sigma^2)\mathbf{X}'\boldsymbol{\varepsilon}] = \mathbf{0}$$

and

$$\text{Asy.Var}[\partial \ln L/\partial \boldsymbol{\beta}] = -E[\partial^2 \ln L/\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}']^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.^{27} \tag{14-52}$$

We can test the hypothesis that at the restricted estimator, the derivatives are equal to zero. The statistic would be

$$\text{LM} = \frac{\mathbf{e}_*'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_*}{\mathbf{e}_*'\mathbf{e}_*/n} = nR_*^2. \tag{14-53}$$

In this form, the LM statistic is n times the coefficient of determination in a regression of the residuals $e_{i*} = (y_i - \mathbf{x}_i'\mathbf{b}_*)$ on the full set of regressors. Finally, for more general models and contexts, the same principle for the LM test produces

$$\begin{aligned} LM &= \left[\bar{\mathbf{g}}\left(\hat{\boldsymbol{\theta}}_R\right)\right]'\left[\text{Est.Asy.Var}\left(\bar{\mathbf{g}}\left(\hat{\boldsymbol{\theta}}_R\right)\right)\right]^{-1}\left[\bar{\mathbf{g}}\left(\hat{\boldsymbol{\theta}}_R\right)\right] \\ &= \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right)\right]'\left[\frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right)\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right)'\right\}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right)\right] \\ &= \mathbf{i}'\hat{\mathbf{G}}(\hat{\mathbf{G}}'\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}'\mathbf{i}, \end{aligned} \tag{14-54}$$

where $\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}_R\right) = \dfrac{\partial \ln f_i\left(\hat{\boldsymbol{\theta}}_R\right)}{\partial \hat{\boldsymbol{\theta}}_R}$, $\mathbf{i}$ is a column of ones, and $\mathbf{g}_i(\hat{\boldsymbol{\theta}}_R)'$ is the ith row of $\hat{\mathbf{G}}$.

[27] This makes use of the fact that the Hessian is block diagonal.

There is evidence that the asymptotic results for these statistics are problematic in small or moderately sized samples.[28] The true distributions of all three statistics involve the data and the unknown parameters and, as suggested by the algebra, converge to the F distribution *from above*. The implication is that the critical values from the chi-squared distribution are likely to be too small; that is, using the limiting chi-squared distribution in small samples is likely to exaggerate the significance of empirical results. Thus, in applications, the more conservative F statistic (or t for one restriction) may be preferable unless one's data are plentiful.

Example 14.9 Testing for Constant Returns to Scale

The Cobb–Douglas production function estimated in Examples 14.6 and 14.7 has returns to scale parameter $\gamma = \Sigma_k \partial \ln y / \partial \ln x_k = \beta_1 + \beta_2 + \beta_3 + \beta_4$. The hypothesis of constant returns to scale, $\gamma = 1$, is routinely tested in this setting. We will carry out this test using the three procedures defined earlier. The estimation results are shown in Table 14.5. For the likelihood ratio test, the chi-squared statistic equals $-2(794.624 - 822.688) = 56.129$. The critical value for a test statistic with one degree of freedom is 3.84, so the hypothesis will be rejected on this basis. For the Wald statistic, based on the unrestricted results, $\mathbf{c}(\boldsymbol{\beta}) = [(\beta_1 + \beta_2 + \beta_3 + \beta_4) - 1]$ and $\mathbf{G} = [1, 1, 1, 1]$. The part of the asymptotic covariance matrix needed for the test is shown with Table 4.5. The statistic is

$$W = \mathbf{c}'(\hat{\boldsymbol{\beta}}_U)[\mathbf{GVG}']^{-1}\mathbf{c}(\hat{\boldsymbol{\beta}}_U) = 57.312.$$

TABLE 14.5 Testing for Constant Returns to Scale in a Production Function

(Estimated standard errors in parentheses)

Estimate	*Stochastic Frontier Unrestricted*		*Stochastic Frontier Constant Returns to Scale*	
β_0[a]	11.7014	(0.00447)	11.7022[a]	(.00457)
β_1	0.58369	(0.01887)	0.55979	(.01903)
β_2	0.03555	(0.01113)	0.00812	(.01075)
β_3	0.02256	(0.01281)	−0.04367	(.00959)
β_4	0.44948	(0.01035)	0.47575	(.00997)
σ[b]	0.13988	(0.00279)	0.18962	(.00011)
λ	1.50164	(0.08748)	1.47082	(.08576)
σ_u[c]	0.15573[d]	(0.00279)	0.15681	(0.00289)
ln L	822.688		794.624	

[a] Unadjusted for nonzero mean of ε.
[b] Reported $\sigma_\varepsilon = [\sigma_v^2 + \sigma_u^2(\pi - 2)/\pi]^{1/2}$. Estimated $\sigma_v = 0.10371$ (0.00418).
[c] σ_u is derived. $\sigma_u = \sigma\lambda/(1 + \lambda^2)^{1/2}$. Est.Cov$(\hat{\sigma}, \hat{\lambda}) = 2.3853\text{e}-7$.
Standard error is computed using the delta method.

Estimated Asy.Var[b1,b2,b3,b4] (e − n = times 10^{-n}.)

0.0003562			
−0.0001079	0.0001238		
−5.576e − 5	9.193e − 6	0.0001642	
−0.0001542	1.810e − 5	−1.235e − 5	0.0001071

[28] See, for example, Davidson and MacKinnon (2004, pp. 424–428).

For the LM test, we need the derivatives of the log-likelihood function. For the particular terms,

$$\begin{aligned} \boldsymbol{g}_\beta &= \partial \ln f_i / \partial(\mathbf{x}_i'\boldsymbol{\beta}) &&= (1/\sigma)[w_i + \lambda A_i],\ A_i = \phi(-\lambda w_i)/\Phi(-\lambda w_i), \\ g_\sigma &= \partial \ln f_i / \partial \sigma &&= (1/\sigma)[-1 + w_i^2 + \lambda w_i A_i], \\ g_\lambda &= \partial \ln f_i / \partial \lambda &&= -w_i A_i. \end{aligned}$$

The calculation is in (14-48); LM = 56.398. The test results are nearly identical for the three approaches.

14.10 THE GENERALIZED REGRESSION MODEL

For the generalized regression model of Section 9.1,

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n, \\ E[\boldsymbol{\varepsilon} \mid \mathbf{X}] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] &= \sigma^2 \boldsymbol{\Omega}, \end{aligned}$$

and as before, we first assume that $\boldsymbol{\Omega}$ is a matrix of known constants. If the disturbances are multivariate normally distributed, then the log-likelihood function for the sample is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\ln|\boldsymbol{\Omega}|. \quad \textbf{(14-55)}$$

It might seem that simply using OLS and a heteroscedasticity robust covariance matrix (see Section 4.5) would be a preferred approach that does not rely on an assumption of normality. There are at least two situations in which GLS, and possibly MLE, might be justified. First, if there is known information about the disturbance variances, this simplicity is a minor virtue that wastes sample information. The grouped data application in Example 14.11 is such a case. Second, there are settings in which the variance itself is of interest, such as models of production risk [Asche and Tvertas (1999)] and in the heteroscedastic stochastic frontier model, which is generally based on the model in Section 14.10.3.[29]

14.10.1 GLS WITH KNOWN Ω

Because $\boldsymbol{\Omega}$ is a matrix of known constants, the maximum likelihood estimator of $\boldsymbol{\beta}$ is the vector that minimizes the **generalized sum of squares**, $S_*(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ (hence the name *generalized least squares*). The necessary conditions for maximizing L are

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} = \frac{\mathbf{X}_*'(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta})}{\sigma^2} = \mathbf{0}, \\ \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{n}{2\sigma^2}\left[\frac{(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta})'(\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta})}{n\sigma^2} - 1\right] = 0, \end{aligned} \quad \textbf{(14-56)}$$

[29] Just and Pope (1978, 1979).

where $\mathbf{X}_* = \boldsymbol{\Omega}^{-1/2}\mathbf{X}$ and $\mathbf{y}_* = \boldsymbol{\Omega}^{-1/2}\mathbf{y}$. The solutions are the OLS estimators using the transformed data,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{ML}} &= (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* &&= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y},\\ \hat{\sigma}^2_{\mathrm{ML}} &= \frac{(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}})'(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\beta}})}{n} &&= \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n},\end{aligned} \tag{14-57}$$

which implies that with normally distributed disturbances, generalized least squares is also maximum likelihood. The maximum likelihood estimator of σ^2 is biased. An unbiased estimator is the one in (9-20). The conclusion is that when $\boldsymbol{\Omega}$ is known, the maximum likelihood estimator is generalized least squares.

14.10.2 ITERATED FEASIBLE GLS WITH ESTIMATED $\boldsymbol{\Omega}$

When $\boldsymbol{\Omega}$ is unknown and must be estimated, then it is necessary to maximize the log likelihood in (14-55) with respect to the full set of parameters $[\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega}]$ simultaneously. Because an unrestricted $\boldsymbol{\Omega}$ contains $n(n + 1)/2 - 1$ free parameters, it is clear that some restriction will have to be placed on the structure of $\boldsymbol{\Omega}$ for estimation to proceed. We will examine applications in which $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$ for some smaller vector of parameters in the next several sections. We note only a few general results at this point.

1. For a given value of $\boldsymbol{\theta}$ the estimator of $\boldsymbol{\beta}$ would be GLS and the estimator of σ^2 would be the estimator in (14-57).
2. The likelihood equations for $\boldsymbol{\theta}$ will generally be complicated functions of $\boldsymbol{\beta}$ and σ^2, so joint estimation will be necessary. However, in many cases, for given values of $\boldsymbol{\beta}$ and σ^2, the estimator of $\boldsymbol{\theta}$ is straightforward. For example, in the model of (9-21), the iterated estimator of θ when $\boldsymbol{\beta}$ and σ^2 and a prior value of $\boldsymbol{\theta}$ are given is the prior value plus the slope in the regression of $(e_i^2/\hat{\sigma}_i^2 - 1)$ on $\mathbf{z}_i$.

The second step suggests a sort of back-and-forth iteration for this model that will work in many situations—starting with, say, OLS, iterating back and forth between 1 and 2 until convergence will produce the joint maximum likelihood estimator. Oberhofer and Kmenta (1974) showed that under some fairly weak requirements, most importantly that $\boldsymbol{\theta}$ not involve σ^2 or any of the parameters in $\boldsymbol{\beta}$, this procedure would produce the maximum likelihood estimator. The asymptotic covariance matrix of this estimator is the same as the GLS estimator. This is the same whether $\boldsymbol{\Omega}$ is known or estimated, which means that if $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ have no parameters in common, then exact knowledge of $\boldsymbol{\Omega}$ brings no gain in asymptotic efficiency in the estimation of $\boldsymbol{\beta}$ over estimation of $\boldsymbol{\beta}$ with a consistent estimator of $\boldsymbol{\Omega}$.

14.10.3 MULTIPLICATIVE HETEROSCEDASTICITY

Harvey's (1976) model of multiplicative heteroscedasticity is a very flexible, general model that includes many useful formulations as special cases. The general formulation is

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}_i'\boldsymbol{\alpha}). \tag{14-58}$$

A model with heteroscedasticity of the form $\sigma_i^2 = \sigma^2 \prod_{m=1}^{M} z_{im}^{\alpha_m}$ results if the logs of the variables are placed in $\mathbf{z}_i$. The groupwise heteroscedasticity model described in Section 9.7.2 is produced by making $\mathbf{z}_i$ a set of group dummy variables (one must be omitted). In this

case, σ^2 is the disturbance variance for the base group whereas for the other groups $\sigma_g^2 = \sigma^2 \exp(\alpha_g)$.

Let $\mathbf{z}_i$ include a constant term so that $\mathbf{z}_i' = [1, \mathbf{q}_i']$, where $\mathbf{q}_i$ is the original set of variables, and let $\boldsymbol{\gamma}' = [\ln \sigma^2, \boldsymbol{\alpha}']$. Then, the model is simply $\sigma_i^2 = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$. Once the full parameter vector is estimated, $\exp(\gamma_1)$ provides the estimator of σ^2. (This estimator uses the invariance result for maximum likelihood estimation. See Section 14.4.5.D) The log likelihood is

$$\begin{aligned}\ln L &= -\frac{1}{2}\sum_{i=1}^{n}\left[\ln \sigma_i^2 + \ln(2\pi) - \frac{\varepsilon_i^2}{\sigma_i^2}\right] \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\mathbf{z}_i'\boldsymbol{\gamma} + \ln(2\pi) + \frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right].\end{aligned} \tag{14-59}$$

The likelihood equations are

$$\begin{aligned}\frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\mathbf{x}_i\frac{\varepsilon_i}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}, \\ \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} &= \frac{1}{2}\sum_{i=1}^{n}\mathbf{z}_i\left(\frac{\varepsilon_i^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1\right) = \mathbf{0}.\end{aligned} \tag{14-60}$$

14.10.4 THE METHOD OF SCORING

For this model, the **method of scoring** turns out to be a particularly convenient way to maximize the log-likelihood function. The terms in the Hessian are

$$\frac{\partial^2 \ln L}{\partial\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}\partial\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}'} = -\sum_{i=1}^{n}\frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\begin{pmatrix}\mathbf{x}_i\\ \varepsilon_i\mathbf{z}_i\end{pmatrix}\begin{pmatrix}\mathbf{x}_i\\ \varepsilon_i\mathbf{z}_i\end{pmatrix}'. \tag{14-61}$$

The expected value of $\partial^2 \ln L/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}'$ is $\mathbf{0}$ because $E[\varepsilon_i|\mathbf{x}_i, \mathbf{z}_i] = 0$. The expected value of the fraction in $\partial^2 \ln L/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'$ is $E[\varepsilon_i^2/\sigma_i^2|\mathbf{x}_i, \mathbf{z}_i] = 1$. Let $\boldsymbol{\delta} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$. Then

$$-E\left(\frac{\partial^2 \ln L}{\partial\boldsymbol{\delta}\,\partial\boldsymbol{\delta}'}\right) = \begin{bmatrix}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} & \mathbf{0}\\ \mathbf{0}' & \frac{1}{2}\mathbf{Z}'\mathbf{Z}\end{bmatrix} = -\overline{\mathbf{H}}. \tag{14-62}$$

The method of scoring is an algorithm for finding an iterative solution to the likelihood equations. The iteration is

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \overline{\mathbf{H}}^{-1}\mathbf{g}_t,$$

where $\boldsymbol{\delta}_t$ (i.e., $\boldsymbol{\beta}_t$, $\boldsymbol{\gamma}_t$, and $\boldsymbol{\Omega}_t$) is the estimate at iteration t, $\mathbf{g}_t$ is the two-part vector of first derivatives $[\partial \ln L/\partial\boldsymbol{\beta}_t', \partial \ln L/\partial\boldsymbol{\gamma}_t']'$, and $\overline{\mathbf{H}}$ is partitioned likewise. [**Newton's method** uses the actual second derivatives in (14-61) rather than their expectations in (14-62). The scoring method exploits the convenience of the zero expectation of the off-diagonal block (cross derivative) in (14-62).] Because $\overline{\mathbf{H}}$ is block diagonal, the iteration can be written as separate equations,

$$\begin{aligned}\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + (\mathbf{X}\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\boldsymbol{\varepsilon}_t) \\ &= \boldsymbol{\beta}_t + (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_t^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_t) \\ &= (\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_t^{-1}\mathbf{y} \text{ (of course).}\end{aligned} \tag{14-63}$$

Therefore, the updated coefficient vector $\boldsymbol{\beta}_{t+1}$ is computed by FGLS using the previously computed estimate of $\boldsymbol{\gamma}$ to compute $\boldsymbol{\Omega}$. We use the same approach for $\boldsymbol{\gamma}$:

$$\begin{aligned}\boldsymbol{\gamma}_{t+1} &= \boldsymbol{\gamma}_t + [2(\mathbf{Z}'\mathbf{Z})^{-1}]\left[\frac{1}{2}\sum_{i=1}^{n}\mathbf{z}_i\left(\frac{\varepsilon_{i(t)}^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma}_t)} - 1\right)\right] \\ &= \boldsymbol{\gamma}_t + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{h}_t.\end{aligned} \tag{14-64}$$

The 2 and $\frac{1}{2}$ cancel. The updated value of $\boldsymbol{\gamma}$ is computed by adding the vector of coefficients in the least squares regression of $[\varepsilon_i^2/\exp(\mathbf{z}_i'\boldsymbol{\gamma}) - 1]$ on $\mathbf{z}_i$ to the old one. Note that the correction is $2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\partial \ln L/\partial\boldsymbol{\gamma})$, so convergence occurs when the derivative is zero.

The remaining detail is to determine the starting value for the iteration. Any consistent estimator will do. The simplest procedure is to use OLS for $\boldsymbol{\beta}$ and the slopes in a regression of the logs of the squares of the least squares residuals on $\mathbf{z}_i$ for $\boldsymbol{\gamma}$. Harvey (1976) shows that this method will produce an inconsistent estimator of $\gamma_1 = \ln \sigma^2$, but the inconsistency can be corrected just by adding 1.2704 to the value obtained. Thereafter, the iteration is simply:

1. Estimate the disturbance variance σ_i^2 with $\exp(\mathbf{z}_i'\boldsymbol{\gamma})$.
2. Compute $\boldsymbol{\beta}_{t+1}$ by FGLS.[30]
3. Update $\boldsymbol{\gamma}_t$ using the regression described in the preceding paragraph.
4. Compute $\mathbf{d}_{t+1} = [\boldsymbol{\beta}_{t+1}, \boldsymbol{\gamma}_{t+1}] - [\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t]$. If $\mathbf{d}_{t+1}$ is large, then return to step 1.

If $\mathbf{d}_{t+1}$ at step 4 is sufficiently small, then exit the iteration. The asymptotic covariance matrix is simply $-\mathbf{H}^{-1}$, which is block diagonal with blocks

$$\begin{aligned}\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{\text{ML}}] &= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}, \\ \text{Asy.Var}[\boldsymbol{\gamma}_{\text{ML}}] &= 2(\mathbf{Z}'\mathbf{Z})^{-1}.\end{aligned} \tag{14-65}$$

If desired, then $\hat{\sigma}^2 = \exp(\hat{\gamma}_1)$ can be computed. The asymptotic variance would be $[\exp(\gamma_1)]^2(\text{Asy.Var}[\hat{\gamma}_{1,\text{ML}}])$.

Testing the null hypothesis of homoscedasticity in this model,

$$\text{H}_0\text{: } \boldsymbol{\alpha} = \mathbf{0}$$

in (14-58), is particularly simple. The Wald test will be carried out by testing the hypothesis that the last M elements of $\boldsymbol{\gamma}$ are zero. Thus, the statistic will be

$$\lambda_{WALD} = \hat{\boldsymbol{\alpha}}'\left\{[\mathbf{0} \quad \mathbf{I}][2(\mathbf{Z}'\mathbf{Z})]^{-1}\begin{bmatrix}\mathbf{0}' \\ \mathbf{I}\end{bmatrix}\right\}^{-1}\hat{\boldsymbol{\alpha}}.$$

Because the first column in $\mathbf{Z}$ is a constant term, this reduces to

$$\lambda_{WALD} = \frac{1}{2}\hat{\boldsymbol{\alpha}}'(\mathbf{Z}_1'\mathbf{M}^0\mathbf{Z}_1)^{-1}\hat{\boldsymbol{\alpha}},$$

where $\mathbf{Z}_1$ is the last M columns of $\mathbf{Z}$, not including the column of ones, and $\mathbf{M}^0$ creates deviations from means. The likelihood ratio statistic is computed based on (14-59).

[30] The two-step estimator obtained by stopping here would be fully efficient if the starting value for γ were consistent, but it would not be the maximum likelihood estimator.

Under both the null hypothesis (homoscedastic—using OLS) and the alternative (heteroscedastic—using MLE), the third term in $\ln L$ reduces to $-n/2$. Therefore, the statistic is simply

$$\lambda_{LR} = 2(\ln L_1 - \ln L_0) = \sum_{i=1}^{n}\left[\ln s^2 - \ln \hat{\sigma}_i^2\right] = \sum_{i=1}^{n}\ln\left(\frac{s^2}{\hat{\sigma}_i^2}\right),$$

where $s^2 = \mathbf{e}'\mathbf{e}/n$ using the OLS residuals. To compute the LM statistic, we will use the expected Hessian in (14-62). Under the null hypothesis, the part of the derivative vector in (14-60) that corresponds to $\boldsymbol{\beta}$ is $(1/s^2)\mathbf{X}'\mathbf{e} = \mathbf{0}$. Therefore, using (14-60), the LM statistic is

$$\lambda_{LM} = \left[\frac{1}{2}\sum_{i=1}^{n}\left(\frac{e_i^2}{s^2} - 1\right)\begin{pmatrix}1\\ \mathbf{z}_{i1}\end{pmatrix}\right]'\left[\frac{1}{2}(\mathbf{Z}'\mathbf{Z})\right]^{-1}\left[\frac{1}{2}\sum_{i=1}^{n}\left(\frac{e_i^2}{s^2} - 1\right)\begin{pmatrix}1\\ \mathbf{z}_{i1}\end{pmatrix}\right].$$

The first element in the derivative vector is zero because $\sum_i e_i^2 = ns^2$. Therefore, the expression reduces to

$$\lambda_{LM} = \frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{e_i^2}{s^2} - 1\right)\mathbf{z}_{i1}\right]'(\mathbf{Z}_1'\mathbf{M}^0\mathbf{Z}_1)^{-1}\left[\sum_{i=1}^{n}\left(\frac{e_i^2}{s^2} - 1\right)\mathbf{z}_{i1}\right].$$

This is one-half times the explained sum of squares in the linear regression of the variable $h_i = (e_i^2/s^2 - 1)$ on $\mathbf{Z}$, which is the Breusch–Pagan/Godfrey LM statistic from Section 9.5.2.

Example 14.10 Multiplicative Heteroscedasticity

In Example 6.4, we fit a cost function for the U.S. airline industry of the form

$$\ln C_{it} = \beta_1 + \beta_2 \ln Q_{it} + \beta_3[\ln Q_{it}]^2 + \beta_4 \ln P_{fuel,i,t} + \beta_5\, Loadfactor_{i,t} + \varepsilon_{i,t},$$

where C_{it} is total cost, Q_{it} is output, and $P_{fuel,i,t}$ is the price of fuel, and the 90 observations in the data set are for six firms observed for 15 years. (The model also included dummy variables for firm and year, which we will omit for simplicity.) In Example 9.4, we fit a revised model in which the load factor appears in the variance of $\varepsilon_{i,t}$ rather than in the regression function. The model is

$$\sigma_{i,t}^2 = \sigma^2 \exp(\alpha\, Loadfactor_{i,t}) = \exp(\gamma_1 + \gamma_2\, Loadfactor_{i,t}).$$

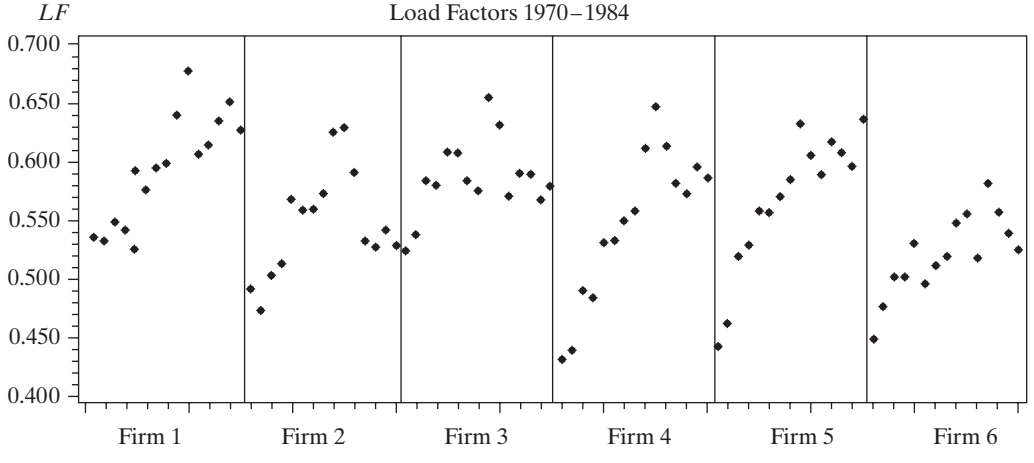
Estimates were obtained by iterating the weighted least squares procedure using weights $W_{i,t} = \exp(-c_1 - c_2\, Loadfactor_{i,t})$. The estimates of γ_1 and γ_2 were obtained at each iteration by regressing the logs of the squared residuals on a constant and $Loadfactor_{it}$. It was noted at the end of the example [and is evident in (14-61)] that these would be the wrong weights to use for iterated weighted least squares if we wish to compute the MLE. Table 14.6 reproduces the results from Example 9.4 and adds the MLEs produced using Harvey's method. The MLE of γ_2 is substantially different from the earlier result. The Wald statistic for testing the homoscedasticity restriction ($\alpha = 0$) is $(9.78076/2.839)^2 = 11.869$, which is greater than 3.84, so the null hypothesis would be rejected. The likelihood ratio statistic is $-2(54.2747 - 57.3122) = 6.075$, which produces the same conclusion. However, the LM statistic is 2.96, which conflicts. This is a finite sample result that is not uncommon. Figure 14.5 shows the pattern of load factors over the period observed. The variances of log costs would vary correspondingly. The increasing load factors in this period would have been a mixed benefit.

TABLE 14.6 Multiplicative Heteroscedasticity Model

	Constant	ln *Q*	$\ln^2 Q$	ln *Pf*	R^{2a}	*Sum of Squares*
OLS[b]	9.13823	0.92615	0.02915	0.41006	0.986167	1.57748
Std. Error	(0.24507)	(0.03231)	(0.01230)	(0.01881)		
Het. Robust S.E.	(0.22595)	(0.03013)	(0.01135)	(0.01752)		
Cluster Robust S.E.	(0.33493)	(0.10235)	(0.04084)	(0.02477)		
Two-step	9.2463	0.92136	0.02445	0.40352	0.9861187	1.612938
Std. Error	(0.21896)	(0.03303)	(0.01141)	(0.01697)		
Iterated[c]	9.2774	0.91609	0.02164	0.40174	0.9860708	1.645693
Std. Error	(0.20977)	(0.03299)	(0.01102)	(0.01633)		
MLE[d]	9.2611	0.91931	0.02328	0.40266	0.9860099	1.626301
Std. Error	(0.2099)	(0.03229)	(0.01099)	(0.01630)		

[a]Squared correlation between actual and fitted values.
[b]ln $L_{OLS} = 54.2747$, ln $L_{ML} = 57.3122$.
[c]Values of c_2 by iteration: 8.25434, 11.6225, 11.7070, 11.7106, 11.7110,
[d]Estimate of γ_2 is 9.78076 (2.83945).

FIGURE 14.5 Load Factors for Six Airlines, 1970–1984.



Example 14.11 Maximum Likelihood Estimation of Gasoline Demand

In Example 9.3, we examined a two-step FGLS estimator for the OECD gasoline demand. The model is a groupwise heteroscedastic specification. In (14-58), z_{it} would be a set of country specific dummy variables. The results from Example 9.3 are shown in Table 14.7 in results (1) and (2). The maximum likelihood estimates are shown in column (3). The parameter estimates are similar, as might be expected. It appears that the standard errors of the coefficients are quite a bit smaller using MLE compared to the two-step FGLS. However, the two estimators are essentially the same. They differ numerically, as expected. However, the asymptotic properties of the two estimators are the same.

TABLE 14.7 Estimated Gasoline Consumption Equations

	(1) OLS		(2) FGLS		(3) MLE	
	Coefficient	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*
ln Income	0.66225	0.07277	0.57507	0.02927	0.45404	0.02211
ln Price	−0.32170	0.07277	−0.27967	0.03519	−0.30461	0.02578
ln Cars/Cap	−0.64048	0.03876	−0.56540	0.01613	−0.47002	0.01275

14.11 NONLINEAR REGRESSION MODELS AND QUASI-MAXIMUM LIKELIHOOD ESTIMATION

In Chapter 7, we considered nonlinear regression models in which the nonlinearity in the parameters appeared entirely on the right-hand side of the equation. Maximum likelihood is often used when the disturbance in a regression, or the dependent variable, more generally, is not normally distributed. If the distribution departs from normality, a likelihood-based approach may provide a useful, efficient way to proceed with estimation and inference. The exponential regression model provides an application.

Example 14.12 Identification in a Loglinear Regression Model

In Example 7.6, we estimated an exponential regression model, of the form

$$E[\mathit{Income}|\mathit{Age}, \mathit{Education}, \mathit{Female}] = \exp(\gamma_1^* + \gamma_2\,\mathit{Age} + \gamma_3\,\mathit{Education} + \gamma_4\,\mathit{Female}).$$

This loglinear conditional mean is consistent with several different distributions, including the lognormal, Weibull, gamma, and exponential models. In each of these cases, the conditional mean function is of the form

$$\begin{aligned} E[\mathit{Income}|\mathbf{x}] &= g(\theta)\exp(\gamma_1 + \mathbf{x}'\boldsymbol{\gamma}_2) \\ &= \exp(\gamma_1^* + \mathbf{x}'\boldsymbol{\gamma}_2), \end{aligned}$$

where θ is an additional parameter of the distribution and $\gamma_1^* = \ln g(\theta) + \gamma_1$. Two implications are:

1. Nonlinear least squares (NLS) is robust at least to some failures of the distributional assumption. The nonlinear least squares estimator of $\boldsymbol{\gamma}_2$ will be consistent and asymptotically normally distributed in all cases for which $E[\mathit{Income}|\mathbf{x}] = \exp(\gamma_1^* + \mathbf{x}'\boldsymbol{\gamma}_2)$.
2. The NLS estimator cannot produce a consistent estimator of γ_1; plim $c_1 = \gamma_1^*$, which varies depending on the correct distribution. In the conditional mean function, any pair of values (θ, γ_1) for which $\gamma_1^* = \ln g(\theta) + \gamma_1$ is the same will lead to the same sum of squares. This is a form of multicollinearity; the pseudoregressor for θ is $\partial E[\mathit{Income}|\mathbf{x}]/\partial\theta = \exp(\gamma_1^* + \mathbf{x}'\boldsymbol{\gamma}_2)[g'(\theta)/g(\theta)]$ while that for γ_1 is $\partial E[\mathit{Income}|\mathbf{x}]/\partial\gamma_1 = \exp(\gamma_1^* + \mathbf{x}'\boldsymbol{\gamma}_2)$. The first is a constant multiple of the second. NLS cannot provide separate estimates of θ and γ_1 while MLE can—see the example to follow. Second, NLS might be less efficient than MLE because it does not use the information about the distribution of the dependent variable. This second consideration is uncertain. For estimation of $\boldsymbol{\gamma}_2$, the NLS estimator is less efficient for not using the distributional information. However, that shortcoming might be offset because the NLS estimator does not attempt to compute an independent estimator of the additional parameter, θ.

To illustrate, we reconsider the estimator in Example 7.6. The gamma regression model specifies

$$f(y|\mathbf{x}) = \frac{1}{\Gamma(\theta)\mu(\mathbf{x})^{\theta}} \exp[-y/\mu(\mathbf{x})]y^{\theta-1}, y > 0, \theta > 0, \mu(\mathbf{x}) = \exp(\gamma_1 + \mathbf{x}'\boldsymbol{\gamma}_2).$$

The conditional mean function for this model is

$$E[y|\mathbf{x}] = \theta/\mu(\mathbf{x}) = \theta \exp(\gamma_1 + \mathbf{x}'\gamma_2) = \exp(\gamma_1^* + \mathbf{x}'\gamma_2).$$

Table 14.8 presents estimates of θ and $(\gamma_1, \boldsymbol{\gamma}_2)$. Estimated standard errors appear in parentheses. The estimates in columns (1), (2), and (4) are all computed using nonlinear least squares. In (1), an attempt was made to estimate θ and γ_1 separately. The estimator converged on two values. However, the estimated standard errors are essentially infinite. The convergence to anything at all is due to rounding error in the computer. The results in column (2) are for γ_1^* and γ_2. The sums of squares for these two estimates as well as for those in (4) are all 112.19688, indicating that the three results merely show three different sets of results for which γ_1^* is the same. The full maximum likelihood estimates are presented in column (3). Note that an estimate of θ is obtained here because the assumed gamma distribution provides another independent moment equation for this parameter; $\partial \ln L/\partial\theta = -n \ln \Psi(\theta) + \Sigma_i(\ln y_i - \ln \mu(\mathbf{x})) = 0$, while the normal equations for the sum of squares provide the same equations for θ and γ_1.

14.11.1 MAXIMUM LIKELIHOOD ESTIMATION

The standard approach to modeling counts of events begins with the Poisson regression model,

$$\text{Prob}[Y = y_i|\mathbf{x}_i] = \frac{\exp(-\lambda_i)\lambda_i^{yi}}{y_i!}, \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), y_i = 0, 1, \ldots,$$

which has **loglinear conditional mean** function $E[y_i|\mathbf{x}_i] = \lambda_i$. (The Poisson regression model and other specifications for data on counts are discussed at length in Chapter 18. We

TABLE 14.8 Estimated Gamma Regression Model

	(1) NLS	*(2) Constrained NLS*	*(3) MLE*	*(4) NLS/MLE*
Constant	1.22468	−1.69331	−3.36826	−3.36380
	(47722.5)[a]	(0.04408)	(0.05048)	(0.04408)
Age	0.00207	0.00207	0.00153	0.00207
	(0.00061)[b]	(0.00061)	(0.00061)	(0.00061)
Education	0.04792	0.04792	0.04975	0.04792
	(0.00247)[b]	(0.00247)	(0.00286)	(0.00247)
Female	−0.00658	−0.00658	0.00696	−0.00658
	(0.01373)[b]	(0.01373)	(0.01322)	(0.08677)
θ	0.62699	—	5.31474	5.31474[c]
	(29921.3)[a]	—	(0.10894)	(0.00000)

[a]Reported value is not meaningful; this is rounding error. See text for description.
[b]Standard errors are the same as in column (2).
[c]Fixed at this value.

introduce the topic here to begin development of the MLE in a fairly straightforward, typical nonlinear setting.) Appendix Table F7.1 presents the Riphahn et al. (2003) data, which we will use to analyze a count variable, *DocVis*, the number of visits to physicans in the survey year. We are using the 1988 wave of the panel, with 4,483 observations. The histogram in Figure 14.6 shows a distinct spike at zero followed by rapidly declining frequencies. While the Poisson distribution, which is typically hump shaped, can accommodate this configuration if λ_i is less than one, the shape is nonetheless somewhat "non-Poisson."[31]

The geometric distribution,

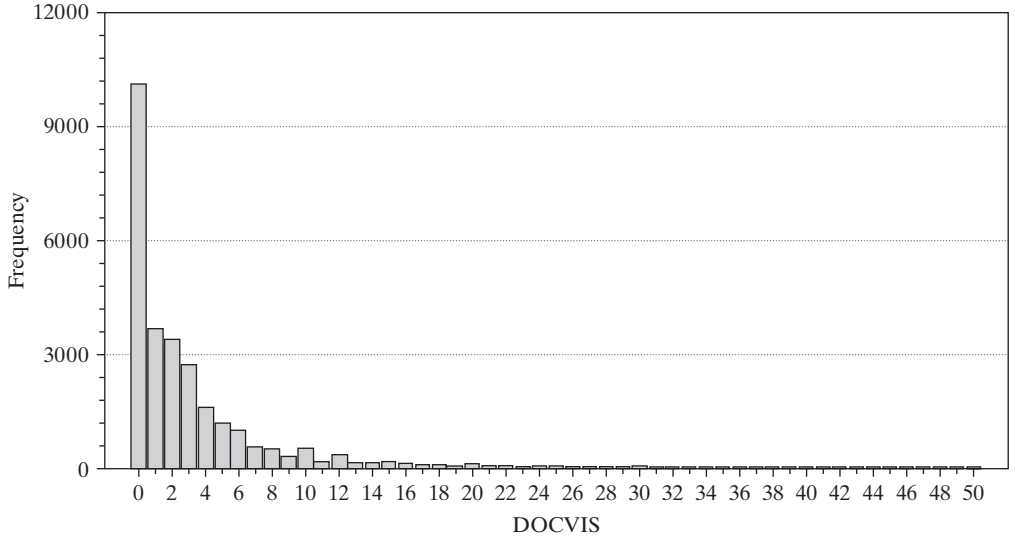
$$f(y_i|\mathbf{x}_i) = \theta_i(1 - \theta_i)^{y_i}, \theta_i = 1/(1 + \lambda_i), \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), y_i = 0, 1, \ldots,$$

is a convenient specification that produces the effect shown in Figure 14.4. (Note that, formally, the specification is used to model the number of failures before the first success in successive independent trials each with success probability θ_i, so in fact, it is misspecified as a model for counts. The model does provide a convenient and useful illustration, however. Moreover, it will turn out that the specification can deliver a consistent estimator of the parameters of interest even if the Poisson is the right model.) The conditional mean function is also $E[y_i|\mathbf{x}_i] = \lambda_i$. The partial effects in the model are $\partial E[y_i|\mathbf{x}_i]/\partial\mathbf{x}_i = \lambda_i\boldsymbol{\beta}$, so this is a distinctly nonlinear regression model. We will construct a maximum likelihood estimator, then compare the MLE to the **nonlinear least squares** and (mis-specified) linear least squares estimates.

The log-likelihood function is

$$\ln L = \sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \theta_i + y_i \ln(1 - \theta_i).$$

FIGURE 14.6 Histogram for Doctor Visits.



[31]So-called Hurdle and Zero Inflation models (discussed in Chapter 18) are often used for this situation.

The likelihood equations are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left(\frac{1}{\theta_i} - \frac{y_i}{1-\theta_i}\right)\frac{d\theta_i}{d\lambda_i}\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

Because

$$\frac{d\theta_i}{d\lambda_i}\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} = \left(\frac{-1}{(1+\lambda_i)^2}\right)\lambda_i\mathbf{x}_i = -\theta_i(1-\theta_i)\mathbf{x}_i,$$

the likelihood equations simplify to

$$\begin{aligned}\frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}(\theta_i y_i - (1-\theta_i))\mathbf{x}_i \\ &= \sum_{i=1}^{n}(\theta_i(1+y_i) - 1)\mathbf{x}_i.\end{aligned}$$

To estimate the asymptotic covariance matrix, we can use any of the estimators of $\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{\text{MLE}}]$ discussed earlier. The BHHH estimator would be

$$\begin{aligned}\text{Est.Asy.Var}_{\text{BHHH}}[\hat{\boldsymbol{\beta}}_{\text{MLE}}] &= \left[\sum_{i=1}^{n}\left(\frac{\partial \ln f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)\left(\frac{\partial \ln f(y_i,|\mathbf{x}_i,\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)'\right]^{-1} \\ &= \left[\sum_{i=1}^{n}(\hat{\theta}_i(1+y_i)-1)^2\mathbf{x}_i\mathbf{x}_i'\right] \\ &= [\hat{\mathbf{G}}'\hat{\mathbf{G}}]^{-1}.\end{aligned}$$

The negative inverse of the second derivatives matrix evaluated at the MLE is

$$\left[-\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}\,\partial \hat{\boldsymbol{\beta}}'}\right]^{-1} = \left[\sum_{i=1}^{n}(1+y_i)\hat{\theta}_i(1-\hat{\theta}_i)\mathbf{x}_i\mathbf{x}_i'\right]^{-1} = [-\hat{\mathbf{H}}]^{-1}.$$

As noted earlier, $E[y_i|\mathbf{x}_i] = \lambda_i = (1-\theta_i)/\theta_i$ is known, so we can also use the negative inverse of the expected second derivatives matrix,

$$\left[-E\left(\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}\,\partial \hat{\boldsymbol{\beta}}}\right)\right]^{-1} = \left[\sum_{i=1}^{n}(1-\hat{\theta}_i)\mathbf{x}_i\mathbf{x}_i'\right]^{-1} = \{-E[\hat{\mathbf{H}}]\}^{-1}.$$

Finally, although we are confident in the form of the conditional mean function, but uncertain about the distribution, it might make sense to use the robust estimator in (14-36),

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = [-\hat{\mathbf{H}}]^{-1}[\hat{\mathbf{G}}'\hat{\mathbf{G}}][-\hat{\mathbf{H}}]^{-1}.$$

To compute the estimates of the parameters, either Newton's method, $\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^t - [\hat{\mathbf{H}}^t]^{-1}\hat{\mathbf{g}}^t$, or the method of scoring, $\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^t - \{E[\hat{\mathbf{H}}^t]\}^{-1}\hat{\mathbf{g}}^t$, can be used, where $\mathbf{H}$ and $\mathbf{g}$ are the second and first derivatives that will be evaluated at the current estimates of the parameters. Like many models of this sort, there is a convenient set of starting values, assuming the model contains a constant term. Because $E[y_i|\mathbf{x}_i] = \lambda_i$, if we start the slope parameters at zero, then a natural starting value for the constant term is the log of $\bar{y}$.

14.11.2 QUASI-MAXIMUM LIKELIHOOD ESTIMATION

If one is confident in the form of the conditional mean function (and that is the function of interest), but less sure about the appropriate distribution, one might seek a robust approach. That is precisely the situation that arose in the preceding example. Given that *DocVis* is a nonnegative count, the exponential mean function makes sense. But we gave equal plausibility to a Poisson model, a geometric model, and a semiparametric approach based on nonlinear least squares. The conditional mean function is correctly specified, but each of these three approaches has a significant shortcoming. The Poisson model imposes an "equidispersion" (variance equal to the mean) that is likely to be transparently inconsistent with the data; the geometric model is manifestly an inappropriate specification, and the nonlinear least squares estimator ignores all information in the sample save for the form of the conditional mean function. A **quasi-MLE**(QMLE) approach based on linear exponential forms provides a somewhat robust approach in this sort of circumstance.

The exponential family of distributions is defined in Definition 13.1. For a random variable, y with density $f(y|\boldsymbol{\theta})$, the exponential family of distributions is

$$\ln f(y|\boldsymbol{\theta}) = a(y) + b(\boldsymbol{\theta}) + \Sigma_k c_k(y)s_k(\boldsymbol{\theta}).$$

Many familiar distributions are in this class, including the normal, logistic, Bernoulli, Poisson, gamma, exponential, Weibull, and others. Based on this framework, Gourieroux, Monfort, and Trognon (1984) proposed the class of conditional linear exponential families,

$$\ln f(y|\mu(\mathbf{x}, \boldsymbol{\beta})) = a(y) + b(\mu(\mathbf{x}, \boldsymbol{\beta})) + ys(\mu(\mathbf{x}, \boldsymbol{\beta})),$$

where the conditional mean function is $E[y|\mathbf{x}, \boldsymbol{\beta}] = \mu(\mathbf{x}, \boldsymbol{\beta})$. The usefulness of this class of specifications is that maximizing the implied log likelihood produces a consistent estimator of $\boldsymbol{\beta}$ even if the true distribution of $y|\mathbf{x}$ is not $f(y|\mu(\mathbf{x}, \boldsymbol{\beta}))$, so long as the mean is correctly specified.

Example 14.13 examines a count variable, $DocVis$ = the number of doctor visits. The assumed conditional mean function is $E[y_i|\mathbf{x}_i] = \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, but we are uncertain of the distribution. Two candidates are considered, geometric with $f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \theta_i(1 - \theta_i)^{y_i}$ with $\theta_i = 1/(1 + \lambda_i)$, and Poisson with $f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = \exp(-\lambda_i)\lambda_i{}^{y_i}/\Gamma(y_i + 1)$. Both of these distributions are in the LEF family; for the geometric, $\ln f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = \ln[\theta_i/(1 - \theta_i)] + y_i \ln \theta_i$, and for the Poisson, $\ln f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = -\lambda_i + y_i \ln \lambda_i - \ln \Gamma(y_i + 1)$. Because both are LEFs involving the same mean function, either log likelihood will produce a consistent estimator of the same $\boldsymbol{\beta}$.

The conditional variance is unspecified so far. In the two cases considered, the variance is a simple function of the mean. For the geometric distribution, $\text{Var}[y|\mathbf{x}] = \lambda(1 + \lambda)$; for the Poisson, $\text{Var}[y|\mathbf{x}] = E[y|\mathbf{x}] = \lambda$. This relationship will hold in general for linear exponential families. For another example, the Bernoulli distribution for a binary or fractional variable, $f(y|\mathbf{x}) = P_i^{y_i}(1 - P_i)^{(1-y_i)}$, where $P_i = \lambda_i/(1 + \lambda_i)$ has conditional variance, $\lambda_i/[1 + \lambda_i]^2 = P_i(1 - P_i)$. The other models examined below, gamma, Weibull, and negative binomial, all behave likewise. The conventional estimator of the asymptotic variance based on the information matrix, (14-16) or (14-17), would apply if the distribution of the LEF were the actual distribution of y_i. However, because the variance has not actually been specified, this may not be the case. Thus, the heteroscedasticity makes the robust variance matrix estimator in (14-36) a logical choice.

An apparently minor extension is needed to accommodate distributions that have an additional parameter, typically a shape parameter, such as the Weibull distribution,

$$f(y_i|\mathbf{x}_i) = \frac{\theta}{\lambda_i}\left(\frac{y_i}{\lambda_i}\right)^{\theta-1}\exp\left[-\left(\frac{y_i}{\lambda_i}\right)^{\theta}\right],$$

for which $E[y_i|\mathbf{x}_i] = \lambda_i\,\Gamma(1 + 1/\theta)$, or the gamma distribution,

$$f(y_i|\mathbf{x}_i) = \frac{y_i^{\theta-1}}{\lambda_i^{\theta}\Gamma(\theta)}\exp\left(-\frac{y_i}{\lambda_i}\right),$$

for which $E[y_i|\mathbf{x}_i] = \lambda_i\theta$. These random variables satisfy the assumptions of the LEF models, but the more detailed specifications create complications both for estimation and inference. First, for these models, the mean, λ_i, is no longer correctly specified. In the cases shown, there is a scaling parameter. If $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ as is typical, and $\boldsymbol{\beta}$ contains a constant term, then the constant term is offset by the log of that scaling term. For the Weibull model, the constant term is offset by $\ln\Gamma(1 + 1/\theta)$ while for the gamma model, the offset is $\ln\theta$. These would seem to be innocuous; however, if the conditional mean itself or partial effects of the mean are the objects of estimation, this is a potentially serious shortcoming. These two models noted are, like the candidates noted earlier, also heteroscedastic; for the gamma, $\text{Var}[y|\mathbf{x}] = \theta\lambda^2$, while for the Weibull, $\text{Var}[y|x] = \lambda\{\Gamma(1 + 2/\theta) - \Gamma^2(1 + 1/\theta)\}$. The robust estimator of the asymptotic covariance matrix in (14-36) for the QMLEs would still be preferred.

These four distributions noted and the others listed below are all members of the LEF, which would suggest that any of them could form the basis of a quasi-MLE for $(y|\mathbf{x})$. The distributions listed are, in principle, for binary (Bernoulli), count (Poisson, geometric), and continuous (gamma, Weibull, normal) random variables. The LEF approach should work best if the random variable studied is of the type that is natural for the form of the distribution used, or at least closest to it. Thus, in the example below, we have modeled the count variable using the geometric and Poisson. One could use the Bernoulli framework for a binary or fractional variable as the basis for the quasi-MLE. Given the results thus far, the Bernoulli LEF could also be used for a continuous variable, but the gamma or Weibull distribution would be a better choice. In general, the support of the observed variable should match that of the variable that underlies the candidate distribution, for example, the nonnegative integers in Example 14.13. [Continuity is not essential; the Poisson (exponential) LEF would work for a continuous (discrete) nonnegative variable.]

If interest centers on estimation of $\boldsymbol{\beta}$, our results would seem to imply that several of these distributions would suffice as the vehicle for estimation in a given situation. But intuition should suggest (no doubt correctly) that some choices should be better than others. On the other hand, why not just use nonlinear least squares (GMM) in all cases if only the conditional mean has been specified? The argument so far does not distinguish any of these estimators; they are all consistent. The criterion function chosen implies a weighting of the observations, and it would seem that some weighting schemes would be better (more efficient) than others, based on the same logic that makes generalized least squares better than ordinary least squares.

The preceding efficiency argument is somewhat ambiguous. It remains a question why one would use this approach instead of nonlinear least squares. The leading application

of these methods [and the focus of Gourieroux et al. (1984) who developed them] is about modeling counts such as our doctor visits variable, in the presence of unmeasured heterogeneity. Consider that in the canonical model for counts, the Poisson regression, there is no explicit place in the specification for unmeasured heterogeneity. The entire specification builds off the conditional mean, $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, and the marginal Poisson distribution. A natural way to extend the Poisson regression specification is $\lambda_i|\varepsilon_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i)$. The conditional mean function is $\Lambda_i = \mathrm{E}[\exp(\varepsilon_i)]\lambda_i$. If the model contains a constant term, then nothing is lost by assuming that $\mathrm{E}[\exp(\varepsilon_i)] = 1$, so $\Lambda_i = \lambda_i$. Left unspecified are the variance of $y_i|\mathbf{x}_i$ and the distribution of ε_i. We assume that ε_i is a conventional disturbance, exogenous to the rest of the model. Thus, the conditional (on $\mathbf{x}$) mean is correctly specified by λ_i which implies that the Poisson QMLE is a robust estimator for this model with only vaguely specified heterogeneity—it is exogenous and has mean 1.

The marginal distribution is $f(y_i|\mathbf{x}_i) = \int_{\varepsilon_i} f(y_i|\mathbf{x}_i, \varepsilon_i)g(\varepsilon_i)d\varepsilon_i$. If $\exp(\varepsilon_i)$ has a gamma distribution with mean 1, $G(\theta, \theta)$, this produces the negative binomial type 2 regression model,

$$f(y_i|\mathbf{x}_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1)\Gamma(\theta)}\left(\frac{\lambda_i}{\lambda_i + \theta}\right)^{y_i}\left(\frac{\theta}{\lambda_i + \theta}\right)^{\theta}, y_i = 0, 1, \ldots.$$

This random variable has mean λ_i and variance $= \lambda_i[1 + \lambda_i/\theta]$. The negative binomial density is a member of the LEF. The advantage of this formulation for count data is that the Poisson quasi-log likelihood will produce a consistent estimator of $\boldsymbol{\beta}$ regardless of the distribution of ε_i as long as ε_i is exogenous, homoscedastic (with respect to $\mathbf{x}_i$), and is parameterized free of $\boldsymbol{\beta}$.

To conclude, the QMLE would seem to be a competitor to the GMM estimator for certain kinds of models. In the leading application, it is a robust estimator that follows the form of the random variable while nonlinear least squares does not.

Example 14.13 Geometric Regression Model for Doctor Visits

In Example 7.6, we considered nonlinear least squares estimation of a loglinear model for the number of doctor visits variable shown in Figure 14.6. (41 observations for which DocVis $>$ 50 out of 27,326 in total are omitted from the figure). The data are drawn from the Riphahn et al. (2003) data set in Appendix Table F7.1. We will continue that analysis here by fitting a more detailed model for the count variable *DocVis*. The conditional mean analyzed here is

$$\ln E[DocVis_{it}|\mathbf{x}_{it}] = \beta_1 + \beta_2\, Age_{it} + \beta_3\, Educ_{it} + \beta_4\, Income_{it} + \beta_5\, Kids_{it}.$$

(This differs slightly from the model in Example 11.16.) For this exercise, with an eye toward the fixed effects model in Example 14.13, we have specified a model that does not contain any time-invariant variables, such as *Female.* (Also, for this application, we will use the entire sample.) Sample means for the variables in the model are given in Table 14.9. Note, these data are a panel. In this exercise, we are ignoring that fact, and fitting a pooled model. We will turn to panel data treatments in the next section, and revisit this application.

We used Newton's method for the optimization, with starting values as suggested earlier. The five iterations are shown in Table 14.9.

Convergence based on the LM criterion, $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$, is achieved after the fourth iteration. Note that the derivatives at this point are extremely small, albeit not absolutely zero. Table 14.10 presents the quasi-maximum likelihood estimates of the parameters. Several sets

TABLE 14.9 Newton Iterations

Start values:	0.11580e+1	0.00000	0.00000	0.00000	0.00000
1st derivatives	0.00000	−0.61777e+5	0.73202e+4	0.42575e+4	0.16464e+4
Parameters:	0.11580e+1	0.00000	0.00000	0.00000	0.00000
Iteration 1 F =	0.6287e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.1907e+4			
1st derivatives	0.48616e+3	−0.22449e+5	0.57162e+4	−0.17112e+3	−0.16521e+3
Parameters:	0.11186e+1	0.1762e−1	−0.50263e−1	−0.46274e−1	−0.15609
Iteration 2 F =	0.6192e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.1258e+2			
1st derivatives	−0.31284e+1	−0.15595e+3	−0.37197e+2	−0.10630e+1	−0.77186
Parameters:	0.10922e+1	0.17981e−1	−0.47303e−1	−0.46739e−1	−0.15683
Iteration 3 F =	0.6192e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.6759e−3			
1st derivatives	−0.18417e−3	−0.99368e−2	−0.21992e−2	−0.59354e−4	−0.25994e−4
Parameters:	0.10918e+1	0.17988e−1	−0.47274e−1	−0.46751e−1	−0.15686
Iteration 4 F =	0.6192e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.1831e−8			
1st derivatives	−0.35727e−11	0.86745e−10	−0.26302e−10	−0.61006e−11	−0.15620e−11
Parameters:	0.10918e+1	0.17988e−1	−0.47274e−1	−0.46751e−1	−0.15686
Iteration 5 F =	0.6192e+5	$\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ = 0.177e−12			

of standard errors are presented. The three sets based on different estimators of the information matrix are presented first. The fourth set is based on the cluster corrected covariance matrix discussed in Section 14.8.4. Because this is actually an (unbalanced) panel data set, we anticipate correlation across observations. Not surprisingly, the standard errors rise substantially. The partial effects listed next are computed in two ways. The *average partial effect* is computed by averaging $\lambda_i\boldsymbol{\beta}$ across the individuals in the sample. The *partial effect* is computed for the average individual by computing λ at the means of the data. The next-to-last column contains the ordinary least squares coefficients. In this model, there is no reason to expect ordinary least squares to provide a consistent estimator of $\boldsymbol{\beta}$. The question might arise, What does ordinary least squares estimate? The answer is the slopes of the linear projection of DocVis on x_{it}. The resemblance of the OLS coefficients to the estimated partial effects is more than coincidental, and suggests an answer to the question.

The analysis in Table 14.11 suggests three competing approaches to modeling DocVis. The results for the geometric regression model are given first in Table 14.10. At the beginning of this section, we noted that the more conventional approach to modeling a count variable such as DocVis is with the Poisson regression model. The quasi-log-likelihood function and its derivatives are even simpler than the geometric model

$$\ln L = \sum_{i=1}^{n} y_i \ln\lambda_i - \lambda_i - \ln y_i!,$$

$$\partial \ln L/\partial\boldsymbol{\beta} = \sum_{i=1}^{n} (y_i - \lambda_i)\mathbf{x}_i,$$

$$\partial^2 \ln L/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = \sum_{i=1}^{n} - \lambda_i \mathbf{x}_i\mathbf{x}_i'.$$

A third approach might be a semiparametric, nonlinear regression model,

$$y_{it} = \exp(x_{it}'\boldsymbol{\beta}) + \varepsilon_{it}.$$

TABLE 14.10 Estimated Geometric Regression Model Dependent Variable: DocVis: Mean = 3.18352, Standard Deviation = 5.68969, n = 27,326

		Std. Err.	*Std. Err.*	*Std. Err.*	*Std. Err.*		*PE*		*Var.*
Variable	***Estimate***	***H***	***E[H]***	***BHHH***	***Cluster***	***APE***	***Mean***	***OLS***	***Mean***
Constant	1.0918	0.0524	0.0524	0.0354	0.1083	—	—	2.656	
Age	0.0180	0.0007	0.0007	0.0005	0.0013	0.0572	0.057	0.061	43.52
Education	−0.0473	0.0033	0.0033	0.0023	0.0067	−0.150	−0.144	−0.121	11.32
Income	−0.4684	0.0411	0.0423	0.0278	0.0727	−1.490	−1.424	−1.621	0.352
Kids	−0.1569	0.0156	0.0155	0.0103	0.0306	−0.487	−0.477	−0.517	0.403

TABLE 14.11 Estimates of Three Models for DocVis

	Geometric Model			*Poisson Model*			*Nonlinear Reg.*		
Variable	***Estimate***	***Std. Err.***	***APE***	***Estimate***	***Std. Err.***	***APE***	***Estimate***	***Std. Err.***	***APE***
Constant	1.0918	0.1083		0.10480	0.1137		0.9802	0.1814	
Age	0.0180	0.0013	0.057	0.0184	0.0013	0.060	0.0187	0.0020	0.060
Education	−0.0473	0.0067	−0.150	−0.0433	0.0070	−0.138	−0.0361	0.0123	−0.115
Income	−0.4684	0.0727	−1.490	−0.5207	0.0822	−1.658	−0.5919	0.1283	−1.884
Kids	−0.1569	0.0306	−0.487	−0.1609	0.0312	−0.500	−0.1693	0.0488	−0.539

Without the distributional assumption, nonlinear least squares is robust, but inefficient compared to the QMLE. But the distributional assumption can be dropped altogether, and the model fit as a simple exponential regression. Note the similarity of the Poisson QMLE and the NLS estimator. For the QMLE, the likelihood equations, $\Sigma_{i=1}^{n}(y_i - \lambda_i)\mathbf{x}_i = \mathbf{0}$, imply that at the solution, the residuals, $(y_i - \lambda_i)$, are orthogonal to the actual regressors, $\mathbf{x}_i$. The NLS normal equations, $\Sigma_{i=1}^{n}(y_i - \lambda_i)\lambda_i\mathbf{x}_i = \Sigma_{i=1}^{n}(y_i - \lambda_i)\mathbf{x}_i^0 = \mathbf{0}$ will imply that at the solutions, the residuals are orthogonal to the pseudo-regressors, $\lambda_i\mathbf{x}_i$.

Table 14.11 presents the three sets of estimates. It is not obvious how to choose among the alternatives. Of the three, the Poisson model is used most often by far. The Poisson and geometric models are not nested, so we cannot use a simple parametric test to choose between them. However, these two models will surely fit the conditions for the Vuong test described in Section 14.6.6. To implement the test, we first computed

$$V_{it} = \ln f_{it}|\,\text{geometric} - \ln f_{it}|\,\text{Poisson}$$

using the respective QMLEs of the parameters. The test statistic given in Section 14.6.6 is then

$$V = \frac{(\sqrt{n})\overline{V}}{s_V}.$$

This statistic converges to standard normal under the underlying assumptions. A large positive value favors the geometric model. The computed sample value is 37.885, which strongly favors the geometric model over the Poisson. Figure 14.6 suggests an explanation

for this finding. The very large mass at $DocVis = 0$ is distinctly non-Poisson. This would motivate an extended model such as the negative binomial model, or more likely a two-part model such as the hurdle model examined in Section 18.4.8. The geometric model would likely provide a better fit to a data set such as this one. The three approaches do display a substantive difference. The average partial effects in Table 14.11 differ noticeably for the three specifications.

14.12 SYSTEMS OF REGRESSION EQUATIONS

The general form of the seemingly unrelated regression (SUR) model is given in (10-1) through (10-3),

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, M, \\ E[\boldsymbol{\varepsilon}_i|\mathbf{X}_1, \ldots, \mathbf{X}_M] &= 0, \\ E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j'|\mathbf{X}_1, \ldots, \mathbf{X}_M] &= \sigma_{ij}\mathbf{I}. \end{aligned} \tag{14-66}$$

FGLS estimation of this model is examined in detail in Section 10.2.3. We will now add the assumption of normally distributed disturbances to the model and develop the maximum likelihood estimators. This suggests a general approach for multiple equation systems. Given the covariance structure defined in (14-66), the joint normality assumption applies to the vector of M disturbances observed at time t, which we write as

$$\boldsymbol{\varepsilon}_t|\mathbf{X}_1, \ldots, \mathbf{X}_M \sim \mathrm{N}[\mathbf{0}, \boldsymbol{\Sigma}], t = 1, \ldots, T. \tag{14-67}$$

14.12.1 THE POOLED MODEL

The pooled model, in which all coefficient vectors are equal, provides a convenient starting point. With the assumption of equal coefficient vectors, the regression model becomes

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, i = 1, \ldots, M, t = 1, \ldots, T, \\ E[\boldsymbol{\varepsilon}_{it}|\mathbf{X}_1, \ldots, \mathbf{X}_M] &= 0, \\ E[\varepsilon_{it}\varepsilon_{js}|\mathbf{X}_1, \ldots, \mathbf{X}_M] &= \sigma_{ij} \quad \text{if} \quad t = s, \quad \text{and} \quad 0 \quad \text{if} \quad t \neq s. \end{aligned} \tag{14-68}$$

This is a model of heteroscedasticity and cross-sectional correlation. With multivariate normality, the log likelihood is

$$\ln L = \sum_{t=1}^{T}\left[-\frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t\right]. \tag{14-69}$$

As we saw earlier, the efficient estimator for this model is GLS, as shown in (10-22). Because the elements of $\boldsymbol{\Sigma}$ must be estimated, the FGLS estimator based on (10-23) and (10-13) is used.

The maximum likelihood estimator of $\boldsymbol{\beta}$, given $\boldsymbol{\Sigma}$, is GLS, based on (10-22). The maximum likelihood estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\sigma}_{ij} = \frac{(\mathbf{y}_i' - \mathbf{X}_i\hat{\hat{\boldsymbol{\beta}}}_{ML})'(\mathbf{y}_j - \mathbf{X}_j\hat{\hat{\boldsymbol{\beta}}}_{ML})}{T} = \frac{\hat{\hat{\boldsymbol{\varepsilon}}}_i'\hat{\hat{\boldsymbol{\varepsilon}}}_j}{T}, \tag{14-70}$$

based on the MLE of $\boldsymbol{\beta}$. If each MLE requires the other, how can we proceed to obtain both? The answer is provided by **Oberhofer and Kmenta** (1974), who show that for certain

models, including this one, one can iterate back and forth between the two estimators. Thus, the MLEs are obtained by iterating to convergence between (14-70) and

$$\hat{\hat{\boldsymbol{\beta}}} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}]. \quad \textbf{(14-71)}$$

The process may begin with the (consistent) ordinary least squares estimator, then (14-70), and so on. The computations are simple, using basic matrix algebra. Hypothesis tests about $\boldsymbol{\beta}$ may be done using the familiar Wald statistic. The appropriate estimator of the asymptotic covariance matrix is the inverse matrix in brackets in (10-22).

For testing the hypothesis that the off-diagonal elements of $\boldsymbol{\Sigma}$ are zero—that is, that there is no correlation across groups—there are three approaches. The likelihood ratio test is based on the statistic

$$\lambda_{\text{LR}} = T(\ln|\hat{\boldsymbol{\Sigma}}_{heteroscedastic}| - \ln|\hat{\boldsymbol{\Sigma}}_{general}|) = T\left(\sum_{i=1}^{M}\ln\hat{\sigma}_i^2 - \ln|\hat{\boldsymbol{\Sigma}}|\right), \quad \textbf{(14-72)}$$

where $\hat{\sigma}_i^2$ are the estimates of σ_i^2 obtained from the maximum likelihood estimates of the groupwise heteroscedastic model and $\hat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimator in the unrestricted model.[32] The large-sample distribution of the statistic is chi squared with $M(M-1)/2$ degrees of freedom. The Lagrange multiplier test developed by Breusch and Pagan (1980) provides an alternative. The general form of the statistic is

$$\lambda_{\text{LM}} = T\sum_{i=2}^{m}\sum_{j=1}^{i-1}r_{ij}^2, \quad \textbf{(14-73)}$$

where r_{ij}^2 is the ijth residual correlation coefficient. If every equation had a different parameter vector, then equation-specific ordinary least squares would be efficient (and ML) and we would compute r_{ij} from the OLS residuals (assuming that there are sufficient observations for the computation). Here, however, we are assuming only a single-parameter vector. Therefore, the appropriate basis for computing the correlations is the residuals from the iterated estimator in the groupwise heteroscedastic model, that is, the same residuals used to compute $\hat{\sigma}_i^2$. (An asymptotically valid approximation to the test can be based on the FGLS residuals instead.) Note that this is not a procedure for testing all the way down to the homoscedastic regression model. That case involves different LM and LR statistics based on the groupwise heteroscedasticity model. If either the LR statistic in (14-72) or the LM statistic in (14-73) is smaller than the critical value from the table, the conclusion, based on this test, is that the appropriate model is the groupwise heteroscedastic model.

14.12.2 THE SUR MODEL

The Oberhofer–Kmenta (1974) conditions are met for the seemingly unrelated regressions model, so maximum likelihood estimates can be obtained by iterating the FGLS procedure. We note, once again, that this procedure presumes the use of (10-11) for estimation of σ_{ij} at each iteration. Maximum likelihood enjoys no advantages over FGLS in its asymptotic properties.[33] Whether it would be preferable in a small sample is an open question whose answer will depend on the particular data set.

[32] *Note:* The excess variation produced by the restrictive model is used to construct the test.

[33] Jensen (1995) considers some variation on the computation of the asymptotic covariance matrix for the estimator that allows for the possibility that the normality assumption might be violated.

Example 14.14 ML Estimates of a Seemingly Unrelated Regressions Model

Although a bit dated, the Grunfeld data used in Application 11.2 have withstood the test of time and are still a standard data set used to demonstrate the SUR model. The data in Appendix Table F10.4 are for 10 firms and 20 years (1935–1954). For the purpose of this illustration, we will use the first four firms.[34]

The model is an investment equation,

$$I_{it} = \beta_{1i} + \beta_{2i}F_{it} + \beta_{3i}C_{it} + \varepsilon_{it}, t = 1, \ldots, 20, i = 1, \ldots, 10,$$

where

I_{it} = real gross investment for firm i in year t,
F_{it} = real value of the firm-shares outstanding,
C_{it} = real value of the capital stock.

The OLS estimates for the four equations are shown in the left panel of Table 14.12. The correlation matrix for the four OLS residual vectors is

$$\mathbf{R_e} = \begin{bmatrix} 1 & -0.261 & 0.279 & -0.273 \\ -0.261 & 1 & 0.428 & 0.338 \\ 0.279 & 0.428 & 1 & -0.0679 \\ -0.273 & 0.338 & -0.0679 & 1 \end{bmatrix}.$$

Before turning to the FGLS and MLE estimates, we carry out the LM test against the null hypothesis that the regressions are actually unrelated. We leave as an exercise to show that the LM statistic in (14-73) can be computed as

$$\lambda_{LM} = (T/2)[\text{trace}(\mathbf{R}_e'\mathbf{R}_e) - M] = 10.451.$$

The 95% critical value from the chi-squared distribution with 6 degrees of freedom is 12.59, so at this point, it appears that the null hypothesis is not rejected. We will proceed in spite of this finding.

TABLE 14.12 Estimated Investment Equations

		OLS		*FGLS*		*MLE*	
Firm	***Variable***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
	Constant	−149.78	97.58	−160.68	90.41	−179.41	86.66
1	*F*	0.1192	0.02382	0.1205	0.02187	0.1248	0.02086
	C	0.3714	0.03418	0.3800	0.03311	0.3802	0.03266
	Constant	−49.19	136.52	21.16	116.18	36.46	106.18
2	*F*	0.1749	0.06841	0.1304	0.05737	0.1244	0.05191
	C	0.3896	0.1312	0.4485	0.1225	0.4367	0.1171
	Constant	−9.956	28.92	−19.72	26.58	−24.10	25.80
3	*F*	0.02655	0.01435	0.03464	0.01279	0.03808	0.01217
	C	0.1517	0.02370	0.1368	0.02249	0.1311	0.02223
	Constant	−6.190	12.45	0.9366	11.59	2.581	11.54
4	*F*	0.07795	0.01841	0.06785	0.01705	0.06564	0.01698
	C	0.3157	0.02656	0.3146	0.02606	0.3137	0.02617

[34]The data are downloaded from the Web site for Baltagi (2005) at www.wiley.com/legacy/wileychi/baltagi/supp/Grunfeld.fil. See also Kleiber and Zeileis (2010).

The next step is to compute the covariance matrix for the OLS residuals using

$$\mathbf{W} = (1/T)\,\mathbf{E}'\mathbf{E} = \begin{bmatrix} \mathbf{7160.29} & -1967.05 & 607.533 & -282.756 \\ -1967.05 & \mathbf{7904.66} & 978.45 & 367.84 \\ 607.533 & 978.45 & \mathbf{660.829} & -21.3757 \\ -282.756 & 367.84 & -21.3757 & \mathbf{149.872} \end{bmatrix},$$

where $\mathbf{E}$ is the 20×4 matrix of OLS residuals. Stacking the data in the partitioned matrices,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_4 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \end{bmatrix},$$

we now compute $\hat{\mathbf{\Omega}} = \mathbf{W} \otimes \mathbf{I}_{20}$ and the FGLS estimates,

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}.$$

The estimated asymptotic covariance matrix for the FGLS estimates is the bracketed inverse matrix. These results are shown in the center panel in Table 14.12. To compute the MLE, we will take advantage of the Oberhofer and Kmenta (1974) result and iterate the FGLS estimator. Using the FGLS coefficient vector, we recompute the residuals, then recompute $\mathbf{W}$, then reestimate $\boldsymbol{\beta}$. The iteration is repeated until the estimated parameter vector converges. We use as our convergence measure the following criterion based on the change in the estimated parameter from iteration $(s-1)$ to iteration $(\mathbf{s})$:

$$\delta = [\hat{\boldsymbol{\beta}}(s) - \hat{\boldsymbol{\beta}}(s-1)]'[\mathbf{X}'[\hat{\mathbf{\Omega}}(s)]^{-1}\mathbf{X}][\hat{\boldsymbol{\beta}}(s) - \hat{\boldsymbol{\beta}}(s-1)].$$

The sequence of values of this criterion function are: 0.21922, 0.16318, 0.00662, 0.00037, 0.00002367825, 0.000001563348, 0.1041980×10^{-6}. We exit the iterations after iteration 7. The ML estimates are shown in the right panel of Table 14.12. We then carry out the likelihood ratio test of the null hypothesis of a diagonal covariance matrix. The maximum likelihood estimate of $\mathbf{\Sigma}$ is

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} \mathbf{7235.46} & -2455.13 & 615.167 & -325.413 \\ -2455.13 & \mathbf{8146.41} & 1288.66 & 427.011 \\ 615.167 & 1288.66 & \mathbf{702.268} & 2.51786 \\ -325.413 & 427.011 & 2.51786 & \mathbf{153.889} \end{bmatrix}.$$

The estimate for the constrained model is the diagonal matrix formed from the diagonals of $\mathbf{W}$ shown earlier for the OLS results. (The estimates are shown in boldface in the preceding matrix, $\mathbf{W}$.) The test statistic is then

$$\text{LR} = T(\ln|\text{diag}(\mathbf{W})| - \ln|\hat{\mathbf{\Sigma}}|) = 18.55.$$

Recall that the critical value is 12.59. The results contradict the LM statistic. The hypothesis of diagonal covariance matrix is now rejected.

Note that aside from the constants, the four sets of coefficient estimates are fairly similar. Because of the constants, there seems little doubt that the pooling restriction will be rejected. To find out, we compute the Wald statistic based on the MLE results. For testing

$$H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \boldsymbol{\beta}_4,$$

we can formulate the hypothesis as

$$H_0: \boldsymbol{\beta}_1 - \boldsymbol{\beta}_4 = \mathbf{0}, \boldsymbol{\beta}_2 - \boldsymbol{\beta}_4 = \mathbf{0}, \boldsymbol{\beta}_3 - \boldsymbol{\beta}_4 = \mathbf{0}.$$

The Wald statistic is

$$\lambda_W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{RVR}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) = 2190.96,$$

where $\mathbf{R} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0} & \mathbf{0} & -\mathbf{I}_3 \\ \mathbf{0} & \mathbf{I}_3 & \mathbf{0} & -\mathbf{I}_3 \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_3 & -\mathbf{I}_3 \end{bmatrix}$, $\mathbf{q} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$, and $\mathbf{V} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}$. Under the null hypothesis, the Wald statistic has a limiting chi-squared distribution with 9 degrees of freedom. The critical value is 16.92, so, as expected, the hypothesis is rejected. It may be that the difference is due to the different constant terms. To test the hypothesis that the four pairs of slope coefficients are equal, we replaced the $\mathbf{I}_3$ in $\mathbf{R}$ with $[[\mathbf{0}, \mathbf{I}_2]]$, the $\mathbf{0}$'s with 2×3 zero matrices, and $\mathbf{q}$ with a 6×1 zero vector. The resulting chi-squared statistic equals 229.005. The critical value is 12.59, so this hypothesis is rejected as well.

14.13 SIMULTANEOUS EQUATIONS MODELS

In Chapter 10, we noted two approaches to maximum likelihood estimation of the equation
system,

$$\mathbf{y}_t'\boldsymbol{\Gamma} + \mathbf{x}_t'\mathbf{B} = \boldsymbol{\varepsilon}_t',$$
$$\boldsymbol{\varepsilon}_t | \mathbf{X} \sim \mathrm{N}[\mathbf{0}, \boldsymbol{\Sigma}]: \qquad \textbf{(14-73)}$$

full information maximum likelihood (FIML) and limited information maximum likelihood (LIML). The FIML approach simultaneously estimates all model parameters. The FIML estimator for a linear equation system is extremely complicated both theoretically and practically. However, its asymptotic properties are identical to three-stage least squares (3SLS), which is straightforward and a standard feature of modern econometric software. (See Section 10.4.5.) Thus, the additional assumption of normality in the system brings no theoretical or practical advantage.

The LIML estimator is a single-equation approach that estimates the parameters of the model one equation at a time. We examined two approaches to computing the LIML estimator, both straightforward, when the equations are linear. The least variance ratio approach shown in Section 10.4.4 is based on some basic matrix algebra calculations—the only unconventional calculation involves the characteristic roots of an asymmetric matrix (or obtaining the matrix square root of a symmetric matrix). The more direct approach in Section (8.4.3) provides some useful results for interpreting the model.

The leading application of LIML estimation is for an equation that contains one endogenous variable. (This is the application in most of Chapter 8.) Let that be the first equation in (14-73),

$$y_1\gamma_{11} + y_2\gamma_{21} + \mathbf{x}_1'\mathbf{b}_1 = \varepsilon_1.$$

Normalize the equation, so the coefficient on y_1 is 1 and the other variables apprear on the right-hand side. Then,

$$y_1 = y_2\delta_1 + \mathbf{x}_1'\boldsymbol{\beta}_1 + w_1. \tag{14-74}$$

This is the structural form for the first equation that contains a single included endogenous variable. The reduced form for the entire system is $\mathbf{y}' = \mathbf{x}'(-\mathbf{B}\boldsymbol{\Gamma}^{-1}) + \mathbf{v}'$. [See Section 10.4.2 and (10-36).] The second equation in the reduced form is

$$y_2 = \mathbf{x}'\boldsymbol{\pi}_2 + u_2. \tag{14-75}$$

Note that the structural equation for y_1 involves only some of the exogenous variables in the system while the reduced form involves all of them including at least one that is not contained in $\mathbf{x}_1$. As we developed in Section 10.4.3, there must be exogenous variables in the system that are excluded from the y_1 equation—this is the order condition for identification. The disturbances in the two equations are linear functions of the disturbances in (14-73), so with normality, the disturbances in (14-74) and (14-75) are joint normal.

The two-equation system (14-74,14-75) is precisely the same as the one we examined in Section 8.4.3,

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \varepsilon \tag{14-76}$$

$$x_2 = \mathbf{z}'\boldsymbol{\gamma} + u, \tag{14-77}$$

where y_2 in (14-74) is the x_2 in (14-76) and $\mathbf{z} = (\mathbf{x}_1, \ldots)$. Equation (14-77) is the reduced form equation for y_2. This formalizes the results for an equation in a simultaneous equations model that contains one endogenous variable. The estimator is actually based on two equations, the structural equation of interest and the reduced form for the endogenous variable that appears in that equation. The log-likelihood function for the LIML estimator for this (actually) two-equation system is shown in (8-17). In the typical equation, (14-76) and (14-77) might well be the recursive structure. This construction of the model underscores the point that in a model that contains an endogenous variable, there is a second equation that "explains" the endogeneity.

For the practitioner, a useful result is that the asymptotic variance of the two-stage least squares (2SLS) estimator is the same as that of the LIML estimator. This would generally render the LIML estimator, with its additional normality assumption, moot. The exception would be the invariance of the LIML estimator to normalization of the equation (i.e., which variable appears on the left of the equals sign). This turns out to be useful in the context of analysis in the presence of weak instruments. (See Section 8.7.) More generally, the LIML and FIML estimators have been supplanted in the literature by much simpler GMM estimators, 2SLS, 3SLS, and extensions that accommodate heteroscedasticity. Interest remains in these estimators, but largely as a component of the ongoing theoretical research.

14.14 PANEL DATA APPLICATIONS

Application of panel data methods to the linear panel data models we have considered so far is a fairly marginal extension. For the random effects linear model, considered in the following Section 14.14.1, the MLE of $\boldsymbol{\beta}$ is, as always, FGLS given the MLEs of the variance parameters. The latter produce a fairly substantial complication, as we shall see. This extension does provide a convenient, interesting application to see the payoff

to the invariance property of the MLE—we will reparameterize a fairly complicated log-likelihood function to turn it into a simple one. Where the method of maximum likelihood becomes essential is in analysis of fixed and random effects in nonlinear models. We will develop two general methods for handling these situations in generic terms in Sections 14.14.3 and 14.14.4, then apply them in several models later in the book.

14.14.1 ML ESTIMATION OF THE LINEAR RANDOM EFFECTS MODEL

The contribution of the ith individual to the log likelihood for the random effects model [(11-28) to (11-32)] with normally distributed disturbances is

$$\begin{aligned}\ln L_i(\boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_u^2) &= -\frac{1}{2}\left[T_i \ln 2\pi + \ln|\boldsymbol{\Omega}_i| + (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right] \\ &= -\frac{1}{2}\left[T_i \ln 2\pi + \ln|\boldsymbol{\Omega}_i| + \boldsymbol{\varepsilon}_i'\boldsymbol{\Omega}_i^{-1}\boldsymbol{\varepsilon}_i\right],\end{aligned} \tag{14-78}$$

where

$$\boldsymbol{\Omega}_i = \sigma_\varepsilon^2\mathbf{I}_{Ti} + \sigma_u^2\mathbf{i}\mathbf{i}',$$

and $\mathbf{i}$ denotes a $T_i \times 1$ column of ones. Note that the $\boldsymbol{\Omega}_i$ varies over i because it is $T_i \times T_i$. Baltagi (2013) presents a convenient and compact estimator for this model that involves iteration between an estimator of $\phi^2 = [\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + T\sigma_u^2)]$, based on sums of squared residuals, and $(\alpha, \boldsymbol{\beta}, \sigma_\varepsilon^2)$ (α is the constant term) using FGLS. Unfortunately, the convenience and compactness come unraveled in the unbalanced case. We consider, instead, what Baltagi labels a "brute force" approach, that is, direct maximization of the log-likelihood function in (14-78). (See, Baltagi, pp. 169–170.)

Using (A-66), we find that

$$\boldsymbol{\Omega}_i^{-1} = \frac{1}{\sigma_\varepsilon^2}\left[\mathbf{I}_{T_i} - \frac{\sigma_u^2}{\sigma_\varepsilon^2 + T_i\sigma_u^2}\mathbf{i}\mathbf{i}'\right].$$

We will also need the determinant of $\boldsymbol{\Omega}_i$. To obtain this, we will use the product of its characteristic roots. First, write

$$|\boldsymbol{\Omega}_i| = (\sigma_\varepsilon^2)^{T_i}|\mathbf{I} + \gamma\mathbf{i}\mathbf{i}'|,$$

where $\gamma = \sigma_u^2/\sigma_\varepsilon^2$. To find the characteristic roots of the matrix, use the definition

$$[\mathbf{I} + \gamma\mathbf{i}\mathbf{i}']\mathbf{c} = \lambda\mathbf{c},$$

where $\mathbf{c}$ is a characteristic vector and λ is the associated characteristic root. The equation implies that $\gamma\mathbf{i}\mathbf{i}'\mathbf{c} = (\lambda - 1)\mathbf{c}$. Premultiply by $\mathbf{i}'$ to obtain $\gamma(\mathbf{i}'\mathbf{i})(\mathbf{i}'\mathbf{c}) = (\lambda - 1)(\mathbf{i}'\mathbf{c})$. Any vector $\mathbf{c}$ with elements that sum to zero will satisfy this equality. There will be $T_i - 1$ such vectors and the associated characteristic roots will be $(\lambda - 1) = 0$ or $\lambda = 1$. For the remaining root, divide by the nonzero $(\mathbf{i}'\mathbf{c})$ and note that $\mathbf{i}'\mathbf{i} = T_i$, so the last root is $T_i\gamma = \lambda - 1$ or $\lambda = (1 + T_i\gamma)$.[35] It follows that the log of the determinant is

$$\ln|\boldsymbol{\Omega}_i| = T_i \ln \sigma_\varepsilon^2 + \ln(1 + T_i\gamma).$$

[35]By this derivation, we have established a useful general result. The characteristic roots of a $T \times T$ matrix of the form $\mathbf{A} = (\mathbf{I} + a\mathbf{b}\mathbf{b}')$ are 1 with multiplicity $(T - 1)$ and $a\mathbf{b}'\mathbf{b}$ with multiplicity 1. The proof follows precisely along the lines of our earlier derivation.

Expanding the parts and multiplying out the third term gives the log-likelihood function

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} \ln L_i \\ &= -\frac{1}{2}\left[(\ln 2\pi + \ln \sigma_\varepsilon^2)\sum_{i=1}^{n} T_i + \sum_{i=1}^{n} \ln(1 + T_i\gamma) \right] - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i - \frac{\sigma_u^2(T_i\bar{\varepsilon}_i)^2}{\sigma_\varepsilon^2 + T_i\sigma_u^2} \right].\end{aligned}$$

Note that in the third term, we can write $\sigma_\varepsilon^2 + T_i\sigma_u^2 = \sigma_\varepsilon^2(1 + T_i\gamma)$ and $\sigma_u^2 = \sigma_\varepsilon^2\gamma$. After inserting these, two appearances of σ_ε^2 in the square brackets will cancel, leaving

$$\ln L = -\frac{1}{2}\sum_{i=1}^{n}\left(T_i(\ln 2\pi + \ln \sigma_\varepsilon^2) + \ln(1 + T_i\gamma) + \frac{1}{\sigma_\varepsilon^2}\left[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i - \frac{\gamma(T_i\bar{\varepsilon}_i)^2}{1 + T_i\gamma} \right] \right).$$

Now, let $\theta = 1/\sigma_\varepsilon^2$, $R_i = 1 + T_i\gamma$, and $Q_i = \gamma/R_i$. The individual contribution to the log likelihood becomes

$$\ln L_i = -\frac{1}{2}[\theta(\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i - Q_i(T_i\bar{\varepsilon}_i)^2) + \ln R_i - T_i \ln \theta + T_i \ln 2\pi].$$

The likelihood equations are

$$\begin{aligned}\frac{\partial \ln L_i}{\partial \boldsymbol{\beta}} &= \theta\left[\sum_{t=1}^{T_i} \mathbf{x}_{it}\varepsilon_{it} \right] - \theta\left[Q_i\left(\sum_{t=1}^{T_i} \mathbf{x}_{it} \right)\left(\sum_{t=1}^{Ti} \varepsilon_{it} \right) \right], \\ \frac{\partial \ln L_i}{\partial \theta} &= -\frac{1}{2}\left[\left(\sum_{t=1}^{T_i} \varepsilon_{it}^2 \right) - Q_i\left(\sum_{t=1}^{T_i} \varepsilon_{it} \right)^2 - \frac{T_i}{\theta} \right], \\ \frac{\partial \ln L_i}{\partial \gamma} &= \frac{1}{2}\left[\theta\left(\frac{1}{R_i^2}\left(\sum_{t=1}^{T_i} \varepsilon_{it} \right)^2 \right) - \frac{T_i}{R_i} \right].\end{aligned}$$

These will be sufficient for programming an optimization algorithm such as DFP or BFGS. (See Section E3.3.) We could continue to derive the second derivatives for computing the asymptotic covariance matrix, but this is unnecessary. For $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, we know that because this is a generalized regression model, the appropriate asymptotic covariance matrix is

$$\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{\text{MLE}}] = \left[\sum_{i=1}^{n} \mathbf{X}_i'\hat{\boldsymbol{\Omega}}_i^{-1}\mathbf{X}_i \right]^{-1}.$$

(See Section 11.5.2.) We also know that the MLEs of the variance components estimators will be asymptotically uncorrelated with the MLE of $\boldsymbol{\beta}$. In principle, we could continue to estimate the asymptotic variances of the MLEs of σ_ε^2 and σ_u^2. It would be necessary to derive these from the estimators of θ and γ, which one would typically do in any event. However, statistical inference about the disturbance variance, σ_ε^2, in a regression model, is typically of no interest. On the other hand, one might want to test the hypothesis that σ_u^2 equals zero, or $\gamma = 0$. Breusch and Pagan's (1979) LM statistic in (11-42) extended to the unbalanced panel case considered here would be

$$LM = \frac{\left(\sum_{i=1}^{N} T_i\right)^2}{\left[2\sum_{i=1}^{N} T_i(T_i - 1)\right]}\left[\frac{\sum_{i=1}^{N}(T_i\bar{e}_i)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T_i} e_{it}^2} - 1\right]^2$$

$$= \frac{\left(\sum_{i=1}^{N} T_i\right)^2}{\left[2\sum_{i=1}^{N} T_i(T_i - 1)\right]}\left[\frac{\sum_{i=1}^{N}[(T_i\bar{e}_i)^2 - \mathbf{e}_i'\mathbf{e}_i]}{\sum_{i=1}^{N}\mathbf{e}_i'\mathbf{e}_i}\right]^2.$$

Example 14.15 Maximum Likelihood and FGLS Estimates of A Wage Equation

Example 11.11 presented FGLS estimates of a wage equation using Cornwell and Rupert's panel data. We have reestimated the wage equation using maximum likelihood instead of FGLS. The parameter estimates appear in Table 14.13, with the FGLS and pooled OLS estimates. The estimates of the variance components are shown in the table as well. The similarity of the MLEs and FGLS slope estimates is to be expected given the large sample size. The difference in the estimates of σ_u is perhaps surprising. The estimator is not based on a simple sum of squares, however, so this kind of variation is common. The LM statistic for testing for the presence of the common effects is 3,497.02, which is far larger than the critical value of 3.84. With the MLE, we can also use an LR test to test for random effects against the null hypothesis of no effects. The chi-squared statistic based on the two log likelihoods is 3,662.25, which leads to the same conclusion.

TABLE 14.13 Wage Equation Estimated by FGLS and MLE

Variable	*Least Squares Estimate*	*Clustered Std. Error*	*Random Effects FGLS*	*Standard Error*	*Random Effects MLE*	*Standard Error*
Constant	5.25112	0.12355	4.04144	0.08330	3.12622	0.17761
Exp	0. 04010	0.00408	0.08748	0.00225	0.10721	0.00248
ExpSq	−0.00067	0.00009	−0.00076	0.00005	−0.00051	0.00005
Wks	0.00422	0.00154	0.00096	0.00059	0.00084	0.00060
Occ	−0.14001	0.02724	−0.04322	0.01299	−0.02512	0.01378
Ind	0.04679	0.02366	0.00378	0.01373	0.01380	0.01529
South	−0.05564	0.02616	−0.00825	0.02246	0.00577	0.03159
SMSA	0.15167	0.02410	−0.02840	0.01616	−0.04748	0.01896
MS	0.04845	0.04094	−0.07090	0.01793	−0.04138	0.01899
Union	0.09263	0.02367	0.05835	0.01350	0.03873	0.01481
Ed	0.05670	0.00556	0.10707	0.00511	0.13562	0.01267
Fem	−0.36779	0.04557	−0.30938	0.04554	−0.17562	0.11310
Blk	−0.16694	0.04433	−0.21950	0.05252	−0.26121	0.13747
θ					42.5265	
γ					29.9705	
σ_ε	0.34936		0.15206		0.15335	
σ_u	0.00000		0.31453		0.83949	

14.14.2 NESTED RANDOM EFFECTS

Consider a data set on test scores for multiple school districts in a state. To establish a notation for this complex model, we define a four-level unbalanced structure,

$$\begin{aligned} Z_{ijkt} &= \text{test score for student } t, \text{ teacher } k, \text{ school } j, \text{ district } i, \\ L &= \text{school districts, } i = 1, \ldots, L, \\ M_i &= \text{schools in each district, } j = 1, \ldots, M_i, \\ N_{ij} &= \text{teachers in each school, } k = 1, \ldots, N_{ij}, \\ T_{ijk} &= \text{students in each class, } t = 1, \ldots, T_{ijk}. \end{aligned}$$

Thus, from the outset, we allow the model to be unbalanced at all levels. In general terms, then, the random effects regression model would be

$$y_{ijkt} = \mathbf{x}'_{ijkt}\boldsymbol{\beta} + u_{ijk} + v_{ij} + w_i + \varepsilon_{ijkt}.$$

Strict exogeneity of the regressors is assumed at all levels. All parts of the disturbance are also assumed to be uncorrelated. (A normality assumption will be added later as well.) From the structure of the disturbances, we can see that the overall covariance matrix, $\boldsymbol{\Omega}$, is block diagonal over i, with each diagonal block itself block diagonal in turn over j, each of these is block diagonal over k, and, at the lowest level, the blocks, for example, for the class in our example, have the form for the random effects model that we saw earlier.

Generalized least squares has been well worked out for the balanced case.[36] Define the following to be constructed from the variance components, σ_ε^2, σ_u^2, σ_v^2, and σ_w^2:

$$\begin{aligned} \sigma_1^2 &= T\sigma_u^2 + \sigma_\varepsilon^2, \\ \sigma_2^2 &= NT\sigma_v^2 + T\sigma_u^2 + \sigma_\varepsilon^2 = \sigma_1^2 + NT\sigma_v^2, \\ \sigma_3^2 &= MNT\sigma_w^2 + NT\sigma_v^2 + T\sigma_u^2 + \sigma_\varepsilon^2 = \sigma_2^2 + MNT\sigma_w^2. \end{aligned}$$

Then, full generalized least squares is equivalent to OLS regression of

$$\widetilde{y}_{ijkt} = y_{ijkt} - \left(1 - \frac{\sigma_\varepsilon}{\sigma_1}\right)\bar{y}_{ijk\cdot} - \left(\frac{\sigma_\varepsilon}{\sigma_1} - \frac{\sigma_\varepsilon}{\sigma_2}\right)\bar{y}_{ij\cdot\cdot} - \left(\frac{\sigma_\varepsilon}{\sigma_2} - \frac{\sigma_\varepsilon}{\sigma_3}\right)\bar{y}_{i\cdot\cdot\cdot}$$

on the same transformation of $\mathbf{x}_{ijkt}$. FGLS estimates are obtained by three groupwise between estimators and the within estimator for the innermost grouping.

The counterparts for the unbalanced case can be derived, but the degree of complexity rises dramatically.[37] As Antwiler (2001) shows, however, if one is willing to assume normality of the distributions, then the log likelihood is very tractable. (We note an intersection of practicality with nonrobustness.) Define the variance ratios

$$\rho_u = \frac{\sigma_u^2}{\sigma_\varepsilon^2}, \rho_v = \frac{\sigma_v^2}{\sigma_\varepsilon^2}, \rho_w = \frac{\sigma_w^2}{\sigma_\varepsilon^2}.$$

Construct the following intermediate results

$$\theta_{ijk} = 1 + T_{ijk}\rho_u, \phi_{ij} = \sum_{k=1}^{N_{ij}} \frac{T_{ijk}}{\theta_{ijk}}, \theta_{ij} = 1 + \phi_{ij}\rho_v, \phi_i = \sum_{j=1}^{M_i} \frac{\phi_{ij}}{\theta_{ij}}, \theta_i = 1 + \rho_w\phi_i$$

[36] See, for example, Baltagi, Song, and Jung (2001), who also provide results for the three-level unbalanced case.

[37] See Baltagi et al. (2001).

and sums of squares of the disturbances $e_{ijkt} = y_{ijkt} - \mathbf{x}'_{ijkt}\boldsymbol{\beta}$,

$$A_{ijk} = \sum_{t=1}^{T_{ijk}} e_{ijkt}^2,$$

$$B_{ijk} = \sum_{t=1}^{T_{ijk}} e_{ijkt},\ B_{ij} = \sum_{k=1}^{N_{ij}} \frac{B_{ijk}}{\theta_{ijk}},\ B_i = \sum_{j=1}^{M_i} \frac{B_{ij}}{\theta_{ij}}.$$

The log likelihood is

$$\ln L = -\frac{1}{2} H \ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2}\Bigg[\sum_{i=1}^{L}\Bigg\{\ln\theta_i + \sum_{j=1}^{M_i}\Bigg\{\ln\theta_{ij} + \sum_{k=1}^{N_{ij}}{}',$$
$$\Bigg\{\ln\theta_{ijk} + \frac{A_{ijk}}{\sigma_\varepsilon^2} - \frac{\rho_u}{\theta_{ijk}}\frac{B_{ijk}^2}{\sigma_\varepsilon^2}\Bigg\} - \frac{\rho_v}{\theta_{ij}}\frac{B_{ij}^2}{\sigma_\varepsilon^2}\Bigg\} - \frac{\rho_w}{\theta_i}\frac{B_i^2}{\sigma_\varepsilon^2}\Bigg\}\Bigg],$$

where H is the total number of observations. (For three levels, $L = 1$ and $\rho_w = 0$.) Antwiler (2001) provides the first derivatives of the log-likelihood function needed to maximize $\ln L$. However, he does suggest that the complexity of the results might make numerical differentiation attractive. On the other hand, he finds the second derivatives of the function intractable and resorts to numerical second derivatives in his application. The complex part of the Hessian is the cross derivatives between $\boldsymbol{\beta}$ and the variance parameters, and the lower-right part for the variance parameters themselves. However, these are not needed. As in any generalized regression model, the variance estimators and the slope estimators are asymptotically uncorrelated. As such, one need only invert the part of the matrix with respect to $\boldsymbol{\beta}$ to get the appropriate asymptotic covariance matrix. The relevant block is

$$-\frac{\partial^2 \ln L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}}\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\mathbf{x}'_{ijkt} - \frac{\rho_w}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\Bigg(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\Bigg)\Bigg(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\Bigg)$$
$$-\frac{\rho_v}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\Bigg(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\Bigg(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\Bigg)\Bigg)\Bigg(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\Bigg(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\Bigg)\Bigg) \qquad \textbf{(14-79)}$$
$$-\frac{\rho_u}{\sigma_\varepsilon^2}\sum_{i=1}^{L}\Bigg(\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\Bigg(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\Bigg(\sum_{t=1}^{T_{ijk}}\mathbf{x}_{ijkt}\Bigg)\Bigg)\Bigg)\Bigg(\sum_{j=1}^{M_i}\frac{1}{\theta_{ij}}\Bigg(\sum_{k=1}^{N_{ij}}\frac{1}{\theta_{ijk}}\Bigg(\sum_{t=1}^{T_{ijk}}\mathbf{x}'_{ijkt}\Bigg)\Bigg)\Bigg).$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ is FGLS based on the maximum likelihood estimators of the variance parameters. Thus, expression (14-79) provides the appropriate covariance matrix for the GLS or maximum likelihood estimator. The difference will be in how the variance components are computed. Baltagi et al. (2001) suggest a variety of methods for the three-level model. For more than three levels, the MLE becomes more attractive.

Example 14.16 Statewide Productivity

Munnell (1990) analyzed the productivity of public capital at the state level using a Cobb–Douglas production function. We will use the data from that study to estimate a three-level log linear regression model,

$$\ln gsp_{jkt} = \alpha + \beta_1 \ln pc_{jkt} + \beta_2 \ln hwy_{jkt} + \beta_3 \ln water_{jkt}$$
$$+ \beta_4 \ln util_{jkt} + \beta_5 \ln emp_{jkt} + \beta_6 unemp_{jkt} + \varepsilon_{jkt} + u_{jk} + v_j,$$
$$j = 1, \ldots, 9;\ t = 1, \ldots, 17,\ k = 1, \ldots, N_j,$$

where the variables in the model are

$$
\begin{aligned}
gsp &= \text{gross state product,}\\
p\_cap &= \text{public capital} = \text{hwy} + \text{water} + \text{util,}\\
hwy &= \text{highway capital,}\\
water &= \text{water utility capital,}\\
util &= \text{utility capital,}\\
pc &= \text{private capital,}\\
emp &= \text{employment (labor),}\\
unemp &= \text{unemployment rate,}
\end{aligned}
$$

and we have defined $M = 9$ regions each consisting of a group of the 48 contiguous states:

$$
\begin{aligned}
Gulf &= \text{AL, FL, LA, MS,}\\
Midwest &= \text{IL, IN, KY, MI, MN, OH, WI,}\\
Mid\ Atlantic &= \text{DE, MD, NJ, NY, PA, VA,}\\
Mountain &= \text{CO, ID, MT, ND, SD, WY,}\\
New\ England &= \text{CT, ME, MA, NH, RI, VT,}\\
South &= \text{GA, NC, SC, TN, WV,}\\
Southwest &= \text{AZ, NV, NM, TX, UT,}\\
Tornado\ Alley &= \text{AR, IA, KS, MO, NE, OK}\\
West\ Coast &= \text{CA, OR, WA.}
\end{aligned}
$$

We have 17 years of data, 1970 to 1986, for each state."[38] The two- and three-level random effects models were estimated by maximum likelihood. The two-level model was also fit by FGLS, using the methods developed in Section 11.5.3.

Table 14.14 presents the estimates of the production function using pooled OLS, OLS for the fixed effects model, and both FGLS and maximum likelihood for the random effects models. Overall, the estimates are similar, though the OLS estimates do stand somewhat apart. This suggests, as one might suspect, that there are omitted effects in the pooled model. The *F* statistic for testing the significance of the fixed effects is 76.712 with 47 and 762 degrees of freedom. The critical value from the table is 1.379, so on this basis, one would reject the hypothesis of no common effects. Note, as well, the extremely large differences between the conventional OLS standard errors and the robust (cluster) corrected values. The three- or four-fold differences strongly suggest that there are latent effects at least at the regional level. It remains to consider which approach, fixed or random effects, is preferred. The Hausman test for fixed vs. random effects produces a chi-squared value of 18.987. The critical value is 12.592. This would imply that the fixed effects model would be the preferred specification. When we repeat the calculation of the Hausman statistic using the three-level estimates in the last column of Table 14.14, the statistic falls slightly to 15.327. Finally, note the similarity of all three sets of random effects estimates. In fact, under the hypothesis of mean independence, all three are consistent estimators. It is tempting at this point to carry out a likelihood ratio test of the hypothesis of the two-level model against the broader alternative three-level model. The test statistic would be twice the difference of the log-likelihoods, which is 2.46. For one degree of freedom, the critical chi squared with one degree of freedom is 3.84, so on this basis, we would not reject the hypothesis of the two-level model. We note, however, that there is a problem with this testing procedure. The hypothesis that a variance is zero is not well defined for the likelihood ratio test—the parameter under the null hypothesis is on the boundary of the parameter space ($\sigma_v^2 \geq 0$). In this instance, the familiar distribution theory does not apply. The results of Kodde and Palm (1988) in Example 14.8 can be used instead of the standard test.

[38]The data were downloaded from the Web site for Baltagi (2005) at www.wiley.com/legacy/wileychi/baltagi3e/. See Appendix Table F10.1.

TABLE 14.14 Estimated Statewide Production Function

	OLS		*Fixed Effects*	*Random Effects FGLS*	*Random Effects ML*	*Nested Random Effects*
	Estimate	*Std. Err.*[a]	*Estimate (Std. Err.)*	*Estimate (Std. Err.)*	*Estimate (Std. Err.)*	*Estimate (Std. Err.)*
α	1.9260	0.05250		2.1608	2.1759	2.1348
		(0.2143)		(0.1380)	(0.1477)	(0.1514)
β_1	0.3120	0.01109	0.2350	0.2755	0.2703	0.2724
		(0.04678)	(0.02621)	(0.01972)	(0.02110)	(0.02141)
β_2	0.05888	0.01541	0.07675	0.06167	0.06268	0.06645
		(0.05078)	(0.03124)	(0.02168)	(0.02269)	(0.02287)
β_3	0.1186	0.01236	0.0786	0.07572	0.07545	0.07392
		(0.03450)	(0.0150)	(0.01381)	(0.01397)	(0.01399)
β_4	0.00856	0.01235	−0.11478	−0.09672	−0.1004	−0.1004
		(0.04062)	(0.01814)	(0.01683)	(0.01730)	(0.01698)
β_5	0.5497	0.01554	0.8011	0.7450	0.7542	0.7539
		(0.06770)	(0.02976)	(0.02482)	(0.02664)	(0.02613)
β_6	−0.00727	0.001384	−0.005179	−0.005963	−0.005809	−0.005878
		(0.002946)	(0.000980)	(0.0008814)	(0.0009014)	(0.0009002)
σ_ε	0.085422		0.03676493	0.0367649	0.0366974	0.0366964
σ_u				0.0771064	0.0875682	0.0791243
σ_v						0.0386299
ln L	853.1372		1565.501		1429.075	1430.30576

[a]Robust (cluster) standard errors in parentheses. The covariance matrix is multiplied by a degrees of freedom correction, $nT/(nT - k) = 816/810$.

14.14.3 CLUSTERING OVER MORE THAN ONE LEVEL

Given the complexity of (14-79), one might prefer simply to use OLS in spite of its inefficiency. As might be expected, the standard errors will be biased owing to the correlation across observations; there is evidence that the bias is downward.[39] In that event, the robust estimator in (11-4) would be the natural alternative. In the example given earlier, the nesting structure was obvious. In other cases, such as our application in Example 11.16, that might not be true. In Example 14.16 and in the application in Baltagi (2013), statewide observations are grouped into regions based on intuition. The impact of an incorrect grouping is unclear. Both OLS and FGLS would remain consistent—both are equivalent to GLS with the wrong weights, which we considered earlier. However, the impact on the asymptotic covariance matrix for the estimator remains to be analyzed.

The nested structure of the data would call the clustering computation in (11-4) into question. If the grouping is done only on the innermost level (on teachers in our example), then the assumption that the clusters are independent is incorrect (teachers in the same school in our example). A two- or more level grouping might be called for in this case. For two levels, as in clusters within stratified data (such as panels on firms within industries) or panel data on individuals within neighborhoods), a reasonably

[39]See Moulton (1986).

compact procedure can be constructed. [See, e.g., Cameron and Miller (2015).] The pseudo-log-likelihood function is

$$\ln L = \sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{i=1}^{N_{cs}} \ln f(y_{ics}|\mathbf{x}_{ics}, \boldsymbol{\theta}), \tag{14-80}$$

where there are S strata, $s = 1, \ldots, S$, C_s clusters in stratum s, $c = 1, \ldots, C_s$ and N_{cs} individual observations in cluster c in stratum s, $i = 1, \ldots, N_{cs}$. We emphasize, this is not the true log likelihood for the sample; the assumed clustering and stratification of the data imply that observations are correlated. Let

$$\begin{aligned}
\mathbf{g}_{ics} &= \frac{\partial \ln f(y_{ics}|\mathbf{x}_{ics}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \mathbf{g}_{cs} = \sum_{i=1}^{N_{cs}} \mathbf{g}_{ics}, \mathbf{g}_s = \sum_{c=1}^{C_s} \mathbf{g}_{cs}, \\
\mathbf{G}_s &= \left(\sum_{c=1}^{C_s} \mathbf{g}_{cs}\mathbf{g}_{cs}'\right) - \frac{1}{C_s}\mathbf{g}_s\mathbf{g}_s', \qquad \mathbf{G} = \sum_{s=1}^{S}\mathbf{G}_s, \\
\mathbf{H} &= \sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{i=1}^{N_{cs}} \frac{\partial^2 \ln f(y_{ics}|\mathbf{x}_{ics}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{i=1}^{N_{cs}} \mathbf{H}_{ics}.
\end{aligned} \tag{14-81}$$

Then, the corrected covariance matrix for the pseudo-MLE would be

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}] = [-\hat{\mathbf{H}}]^{-1}[\hat{\mathbf{G}}][-\hat{\mathbf{H}}]^{-1} \tag{14-82}$$

For a linear model estimated using least squares, we would use $\mathbf{g}_{ics} = (e_{ics}/s^2)\mathbf{x}_{ics}$ and $\mathbf{H}_{ics} = (1/s^2)\mathbf{x}_{ics}\mathbf{x}_{ics}'$. The appearances of s^2 would cancel out in the final result. One last consideration concerns some finite population corrections. The terms in $\mathbf{G}$ might be weighted by a factor $w_s = (1 - C_s/C*)$ if stratum s consists of a finite set of $C*$ clusters of which C_s is a significant proportion, times the within cluster correction, $C_s/(C_s - 1)$, that appears in (11-4), and finally, times $(n - 1)/(n - K)$, where n is the full sample size and K is the number of parameters estimated.

14.14.4 RANDOM EFFECTS IN NONLINEAR MODELS: MLE USING QUADRATURE

Example 14.13 describes a nonlinear model for panel data, the geometric regression model,

$$\begin{aligned}
\text{Prob}[Y_{it} = y_{it}|\mathbf{x}_{it}] &= \theta_{it}(1 - \theta_{it})^{y_{it}}, y_{it} = 0, 1, \ldots, i = 1, \ldots, n, t = 1, \ldots, T_i, \\
\theta_{it} &= 1/(1 + \lambda_{it}), \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}).
\end{aligned}$$

As noted, this is a panel data model, although as stated, it has none of the features we have used for the panel data in the linear case. It is a regression model,

$$E[y_{it}|\mathbf{x}_{it}] = \lambda_{it},$$

which implies that

$$y_{it} = \lambda_{it} + \varepsilon_{it}.$$

This is simply a tautology that defines the deviation of y_{it} from its conditional mean. It might seem natural at this point to introduce a common fixed or random effect, as we did earlier in the linear case, as in

$$y_{it} = \lambda_{it} + \varepsilon_{it} + c_i.$$

However, the difficulty in this specification is that whereas ε_{it} is defined residually just as the difference between y_{it} and its mean, c_i is a freely varying random variable. Without extremely complex constraints on how c_i varies, the model as stated cannot prevent y_{it} from being negative. When building the specification for a nonlinear model, greater care must be taken to preserve the internal consistency of the specification. A frequent approach in **index function models** such as this one is to introduce the common effect in the conditional mean function. The random effects geometric regression model, for example, might appear

$$\text{Prob}[Y_{it} = y_{it}|\mathbf{x}_{it}] = \theta_{it}(1 - \theta_{it})^{y_{it}}, y_{it} = 0, 1, \ldots; i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$\theta_{it} = 1/(1 + \lambda_{it}), \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i),$$
$$f(u_i) = \text{the specification of the distribution of random effects over individuals.}$$

By this specification, it is now appropriate to state the model specification as

$$\text{Prob}[Y_{it} = y_{it}, |\mathbf{x}_{it}, u_i] = \theta_{it}(1 - \theta_{it})^{y_{it}}.$$

That is, our statement of the probability is now conditioned on both the observed data and the unobserved random effect. The random common effect can then vary freely and the inherent characteristics of the model are preserved.

Two questions now arise:

- How does one obtain maximum likelihood estimates of the parameters of the model? We will pursue that question now.
- If we ignore the individual heterogeneity and simply estimate the pooled model, will we obtain consistent estimators of the model parameters? The answer is sometimes, but usually not. The favorable cases are the simple loglinear models such as the geometric and Poisson models that we consider in this chapter. The unfavorable cases are most of the other common applications in the literature, including, notably, models for binary choice, censored regressions, two-part models, sample selection, and, generally, nonlinear models that do not have simple exponential means.[40]

We will now develop a maximum likelihood estimator for a nonlinear random effects model. To set up the methodology for applications later in the book, we will do this in a generic specification, then return to the specific application of the geometric regression model in Example 14.13. Assume, then, that the panel data model defines the probability distribution of a random variable, y_{it}, conditioned on a data vector, $\mathbf{x}_{it}$, and an unobserved common random effect, u_i. As always, there are T_i observations in the group, and the data on $\mathbf{x}_{it}$ and now u_i are assumed to be strictly exogenously determined. Our model for one individual is, then,

$$p(y_{it}|\mathbf{x}_{it}, u_i) = f(y_{it}|\mathbf{x}_{it}, u_i, \boldsymbol{\theta}),$$

where $p(y_{it}|\mathbf{x}_{it}, u_i)$ indicates that we are defining a conditional density while $f(y_{it}|\mathbf{x}_{it}, u_i, \theta)$ defines the functional form and emphasizes the vector of parameters to be estimated. We are also going to assume that, but for the common u_i, observations within a group would be independent—the dependence of observations in the group arises through the

[40] *Note:* This is the crucial issue in the consideration of robust covariance matrix estimation in Section 14.8. See, as well, Freedman (2006).

presence of the common u_i. The joint density of the T_i observations on y_{it} given u_i under these assumptions would be

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i} | \mathbf{X}_i, u_i) = \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}),$$

because conditioned on u_i, the observations are independent. But because u_i is part of the observation on the group, to construct the log likelihood, we will require the joint density,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}, u_i, | \mathbf{X}_i) = \left[\prod_{t=1}^{T_i} f(y_{it}, | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i).$$

The likelihood function is the joint density for the observed random variables. Because u_i is an unobserved random effect, to construct the likelihood function, we will then have to integrate it out of the joint density. Thus,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i} | \mathbf{X}_i) = \int_{u_i} \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i) du_i.$$

The contribution to the log-likelihood function of group i is, then,

$$\ln L_i = \ln \int_{u_i} \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i) du_i.$$

There are two practical problems to be solved to implement this estimator. First, it will be rare that the integral will exist in closed form. (It does when the density of y_{it} is normal with linear conditional mean and the random effect is normal, because, as we have seen, this is the random effects linear model.) As such, the practical complication that arises is how the integrals are to be computed. Second, it remains to specify the distribution of u_i over which the integration is taken. The distribution of the common effect is part of the model specification. Several approaches for this model have now appeared in the literature. The one we will develop here extends the random effects model with normally distributed effects that we have analyzed in the previous section. The technique is **Butler and Moffitt's method** (1982). It was originally proposed for extending the random effects model to a binary choice setting (see Chapter 17), but, as we shall see presently, it is straightforward to extend it to a wide range of other models. The computations center on a technique for approximating integrals known as **Gauss–Hermite quadrature**.

We assume that u_i is normally distributed with mean zero and variance σ_u^2. Thus,

$$f(u_i) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{u_i^2}{2\sigma_u^2} \right).$$

With this assumption, the ith term in the log likelihood is

$$\ln L_i = \ln \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{u_i^2}{2\sigma_u^2} \right) du_i.$$

To put this function in a form that will be convenient for us later, we now let $w_i = u_i/(\sigma_u\sqrt{2})$ so that $u_i = \sigma_u\sqrt{2}w_i = \phi w_i$ and the Jacobian of the transformation

from u_i to w_i is $du_i = \phi dw_i$. Now, we make the change of variable in the integral to produce the function

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi w_i, \boldsymbol{\theta}) \right] \exp(-w_i^2) dw_i.$$

For the moment, let

$$g(w_i) = \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi w_i, \boldsymbol{\theta}).$$

Then, the function we are manipulating is

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} g(w_i) \exp(-w_i^2) dw_i.$$

The payoff to all this manipulation is that integrals of this form can be computed very accurately by Gauss–Hermite quadrature. Gauss–Hermite quadrature replaces the integration with a weighted sum of the functions evaluated at a specific set of points. For the general case, this is

$$\int_{-\infty}^{\infty} g(w_i) \exp(-w_i^2) dw_i \approx \sum_{h=1}^{H} z_h g(v_h),$$

where z_h is the weight and v_h is the node. Tables of the nodes and weights are found in popular sources such as Abramovitz and Stegun (1971). For example, the nodes and weights for a four-point quadrature are

$$v_h = \pm 0.52464762327529002 \quad \text{and} \quad \pm 1.6506801238857849,$$
$$z_h = 0.80491409000549996 \quad \text{and} \quad 0.081312835447250001.$$

In practice, it is common to use eight or more points, up to a practical limit of about 96. Assembling all of the parts, we obtain the approximation to the contribution to the log likelihood,

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\theta}) \right].$$

The Hermite approximation to the log-likelihood function is

$$\ln L = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\theta}) \right]. \tag{14-83}$$

This function is now to be maximized with respect to $\boldsymbol{\theta}$ and ϕ. Maximization is a complex problem. However, it has been automated in contemporary software for some models, notably the binary choice models mentioned earlier, and is in fact quite straightforward to implement in many other models as well. The first and second derivatives of the log-likelihood function are correspondingly complex but still computable using quadrature. The estimate of σ_u and an appropriate standard error are obtained from $\hat{\phi}$ using the result $\phi = \sigma_u \sqrt{2}$. The hypothesis of no cross-period correlation can be tested with a likelihood ratio test.

Example 14.17 Random Effects Geometric Regression Model

We will use the preceding to construct a random effects model for the *DocVis* count variable analyzed in Example 14.10. Using (14-90), the approximate log-likelihood function will be

$$\ln L_H = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} \theta_{it}(1 - \theta_{it})^{y_{it}} \right],$$

$$\theta_{it} = 1/(1 + \lambda_{it}),\ \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \phi v_h).$$

The derivatives of the log likelihood are approximated as well. The following is the general result—development is left as an exercise:

$$\frac{\partial \log L}{\partial \begin{pmatrix} \beta \\ \phi \end{pmatrix}} = \sum_{i=1}^{n} \frac{1}{L_i} \frac{\partial L_i}{\partial \begin{pmatrix} \beta \\ \phi \end{pmatrix}}$$

$$\approx \sum_{i=1}^{n} \frac{\left\{ \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\beta}) \right] \left[\sum_{t=1}^{T_i} \frac{\partial \ln f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\beta})}{\partial \begin{pmatrix} \beta \\ \phi \end{pmatrix}} \right] \right\}}{\left\{ \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} z_h \left[\prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\beta}) \right] \right\}}.$$

It remains only to specialize this to our geometric regression model. For this case, the density is given earlier. The missing components of the preceding derivatives are the partial derivatives with respect to $\boldsymbol{\beta}$ and ϕ that were obtained in Section 14.14.4. The necessary result is

$$\frac{\partial \ln f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\beta})}{\partial \begin{pmatrix} \beta \\ \phi \end{pmatrix}} = [\theta_{it}(1 + y_{it}) - 1] \begin{pmatrix} \mathbf{x}_{it} \\ v_h \end{pmatrix}.$$

Maximum likelihood estimates of the parameters of the random effects geometric regression model are given in Example 14.13 with the fixed effects estimates for this model.

14.14.5 FIXED EFFECTS IN NONLINEAR MODELS: THE INCIDENTAL PARAMETERS PROBLEM

Using the same modeling framework that we used in the previous section, we now define a fixed effects model as an index function model with a group-specific constant term. As before, the model is the assumed density for a random variable,

$$p(y_{it} | d_{it}, \mathbf{x}_{it}) = f(y_{it} | \alpha_i d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where d_{it} is a dummy variable that takes the value one in every period for individual i and zero otherwise. (In more involved models, such as the censored regression model we examine in Chapter 19, there might be other parameters, such as a variance. For now, it is convenient to omit them—the development can be extended to add them later.) For convenience, we have redefined $\mathbf{x}_{it}$ to be the nonconstant variables in the model.[41] The

[41] In estimating a fixed effects linear regression model in Section 11.4, we found that it was not possible to analyze models with time-invariant variables. The same limitation applies in the nonlinear case, for essentially the same reasons. The time-invariant effects are absorbed in the constant term. In estimation, the columns of the derivatives matrix corresponding to time-invariant variables will be transformed to columns of zeros when we compute derivatives of the log-likelihood function.

parameters to be estimated are the K elements of $\boldsymbol{\beta}$ and the n individual constant terms, α_i. The log-likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln f(y_{it}|\,\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where $f(.)$ is the probability density function of the observed outcome, for example, the geometric regression model that we used in our previous example. It will be convenient to let

$$z_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} \text{ so that } p(y_{it}|\,d_{it}, \mathbf{x}_{it}) = f(y_{it}|\,z_{it}).$$

In the fixed effects linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means that eliminated the person-specific constants from the equation. (See Section 11.4.1.) In a few cases of nonlinear models, it is also possible to eliminate the fixed effects from the likelihood function, although in general not by taking deviations from means. One example is the **exponential regression model** that is used in duration modeling, for example for lifetimes of electronic components and electrical equipment such as light bulbs,

$$f(y_{it}|\,\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}) = \theta_{it}\exp(-\theta_{it}y_{it}),\ \theta_{it} = \exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}),\ y_{it} \geq 0.$$

It will be convenient to write $\theta_{it} = \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) = \gamma_i\Delta_{it}$. We are exploiting the invariance property of the MLE—estimating $\gamma_i = \exp(\alpha_i)$ is the same as estimating α_i. The log likelihood is

$$\begin{aligned}\ln L &= \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln \theta_{it} - \theta_{it}y_{it} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln(\gamma_i\Delta_{it}) - (\gamma_i\Delta_{it})y_{it}.\end{aligned} \tag{14-84}$$

The MLE will be found by equating the $n + K$ partial derivatives with respect to γ_i and $\boldsymbol{\beta}$ to zero. For each constant term,

$$\frac{\partial \ln L}{\partial \gamma_i} = \sum_{t=1}^{T_i}\left(\frac{1}{\gamma_i} - \Delta_{it}y_{it}\right).$$

Equating this to zero provides a solution for γ_i in terms of the data and $\boldsymbol{\beta}$,

$$\gamma_i = \frac{T_i}{\sum_{t=1}^{T_i}\Delta_{it}y_{it}}. \tag{14-85}$$

[Note the analogous result for the linear model in (11-16b).] Inserting this solution back in the log-likelihood function in (14-84), we obtain the concentrated log likelihood,

$$\ln L_C = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[\ln\left(\frac{T_i\Delta_{it}}{\sum_{s=1}^{T_i}\Delta_{is}y_{is}}\right) - \left(\frac{T_i\Delta_{it}}{\sum_{s=1}^{T_i}\Delta_{is}y_{is}}\right)y_{it}\right], \tag{14-86}$$

which is now only a function of $\boldsymbol{\beta}$. This function can now be maximized with respect to $\boldsymbol{\beta}$ alone. The MLEs for α_i are then found as the logs of the results of (14-92). Note, once again, we have eliminated the constants from the estimation problem, but not by computing deviations from group means. That is specific to the linear model.

The concentrated log likelihood is only obtainable in only a small handful of cases, including the linear model, the exponential model (as just shown), the Poisson regression model, and a few others. Lancaster (2000) lists some of these and discusses the underlying methodological issues. In most cases, if one desires to estimate the parameters of a fixed effects model, it will be necessary to actually compute the possibly huge number of constant terms, α_i, at the same time as the main parameters, $\boldsymbol{\beta}$. This has widely been viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception.[42] The likelihood equations for the general fixed effects, index function model are

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T_i} \frac{\partial \ln f(y_{it}|\, z_{it})}{\partial z_{it}} \frac{\partial z_{it}}{\partial \alpha_i} = \sum_{t=1}^{T_i} g_{it} = g_{i.} = 0,$$

and

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{\partial \ln f(y_{it}|\, z_{it})}{\partial z_{it}} \frac{\partial z_{it}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} g_{it}\mathbf{x}_{it} = \mathbf{0}.$$

The second derivatives matrix is

$$\frac{\partial^2 \ln L}{\partial \alpha_i^2} = \sum_{t=1}^{T_i} \frac{\partial^2 \ln f(y_{it}|\, z_{it})}{\partial z_{it}^2} = \sum_{t=1}^{T_i} h_{it} = h_{i.} < 0,$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \alpha_i} = \sum_{t=1}^{T_i} h_{it}\mathbf{x}_{it},$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} h_{it}\mathbf{x}_{it}\mathbf{x}_{it}' = \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}'},$$

where $\mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}'}$ is a negative definite matrix. The likelihood equations are a large system, but the solution turns out to be surprisingly straightforward.[43]

By using the formula for the partitioned inverse, we find that the $K \times K$ submatrix of the inverse of the Hessian that corresponds to $\boldsymbol{\beta}$, which would provide the asymptotic covariance matrix for the MLE, is

$$\mathbf{H}^{\boldsymbol{\beta}\boldsymbol{\beta}'} = \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it}\mathbf{x}_{it}\mathbf{x}_{it}' - \frac{1}{h_{i.}} \left(\sum_{t=1}^{T_i} h_{it}\mathbf{x}_{it} \right) \left(\sum_{t=1}^{T_i} h_{it}\mathbf{x}_{it}' \right) \right] \right\}^{-1},$$

$$= \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right] \right\}^{-1}, \quad \text{where} \quad \bar{\mathbf{x}}_i = \frac{\sum_{t=1}^{T_i} h_{it}\mathbf{x}_{it}}{h_{i.}}.$$

Note the striking similarity to the result we had in (11-4) for the fixed effects model in the linear case.[44] By assembling the Hessian as a partitioned matrix for $\boldsymbol{\beta}$ and the full vector of constant terms, then using (A-66b) and the preceding definitions to isolate one diagonal element, we find

$$\mathbf{H}^{\alpha_i\alpha_i} = \frac{1}{h_{i.}} + \bar{\mathbf{x}}_i' \mathbf{H}^{\boldsymbol{\beta}\boldsymbol{\beta}'} \bar{\mathbf{x}}_i.$$

[42] See, for example, Maddala (1987), p. 317.

[43] See Greene (2004a).

[44] A similar result is noted briefly in Chamberlain (1984).

Once again, the result has the same format as its counterpart in the linear model. In principle, the negatives of these would be the estimators of the asymptotic variances of the maximum likelihood estimators. (Asymptotic properties in this model are problematic, as we consider shortly.)

All of these can be computed quite easily once the parameter estimates are in hand, so that in fact, practical estimation of the model is not really the obstacle. [This must be qualified, however. Consider the likelihood equation for one of the constants in the geometric regression model. This would be

$$\sum_{t=1}^{T_i}[\theta_{it}(1 + y_{it}) - 1] = 0.$$

Suppose y_{it} equals zero in every period for individual i. Then, the solution occurs where $\boldsymbol{\Sigma}_i(\theta_{it} - 1) = 0$. But θ_{it} is between zero and one, so the sum must be negative and cannot equal zero. The likelihood equation has no solution with finite coefficients. Such groups would have to be removed from the sample to fit this model.]

It is shown in Greene (2004a) that, in spite of the potentially large number of parameters in the model, Newton's method can be used with the following iteration, which uses only the $K \times K$ matrix computed earlier and a few $K \times 1$ vectors:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(s+1)} &= \hat{\boldsymbol{\beta}}^{(s)} - \left\{\sum_{i=1}^{n}\left[\sum_{t=1}^{T_i} h_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\right]\right\}^{-1}\left\{\sum_{i=1}^{n}\left[\sum_{t=1}^{T_i} g_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\right]\right\} \\ &= \hat{\boldsymbol{\beta}}^{(s)} + \boldsymbol{\Delta}_{\beta}^{(s)},\end{aligned}$$

and

$$\hat{\alpha}_i^{(s+1)} = \hat{\alpha}_i^{(s)} - [(g_{i.}/h_{i.}) + \bar{\mathbf{x}}_i'\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(s)}].^{45}$$

This is a large amount of computation involving many summations, but it is linear in the number of parameters and does not involve any $n \times n$ matrices.

In addition to the theoretical virtues and shortcomings (yet to be addressed) of this model, we note the practical aspect of estimation of what are possibly a huge number of parameters, $n + K$. In the fixed effects case, n is not limited, and could be in the thousands in a typical application. In Examples 14.15 and 14.16, n is 7,293. Two large applications of the method described here are Kingdon and Cassen's (2007) study, in which they fit a fixed effects probit model with well over 140,000 dummy variable coefficients, and Fernandez-Val's (2009) study, which analyzes a model with 500,000 groups.

The problems with the fixed effects estimator are statistical, not practical.[46] The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. In this setting, not only is T_i fixed, it is also likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). There is, as well, a small sample (small T_i) bias in the slope estimators. This is the **incidental parameters problem**.[47] The source of the

[45]Similar results appear in Prentice and Gloeckler (1978) who attribute it to Rao (1973) and Chamberlain (1980, 1984).

[46]See Vytlacil, Aakvik, and Heckman (2005), Chamberlain (1980, 1984), Newey (1994), Bover and Arellano (1997), Chen (1998), and Fernandez-Val (2009) for some extensions of parametric and semiparametric forms of the binary choice models with fixed effects.

[47]See Neyman and Scott (1948) and Lancaster (2000).

problem appears to arise from estimating $n + K$ parameters with n multivariate observations—the number of parameters estimated grows with the sample size. The precise implication of the incidental parameters problem differs from one model to the next. In general, the slope estimators in the fixed effects model do converge to a parameter vector, but not to $\boldsymbol{\beta}$. In the most familiar cases, binary choice models such as probit and logit, the small T bias in the coefficient estimators appears to be proportional (e.g., 100% when $T = 2$), and away from zero, and to diminish monotonically with T, becoming essentially negligible as T reaches 15 or 20. In other cases involving continuous variables, the slope coefficients appear not to be biased at all, but the impact is on variance and scale parameters. The linear fixed effects model noted in Footnote 12 in Chapter 11 is an example; the stochastic frontier model (Section 19.2) is another. Yet, in models for truncated variables (Section 19.2), the incidental parameters bias appears to affect both the slopes (biased toward zero) and the variance parameters (also attenuated). We will examine the incidental parameters problem in more detail in Section 15.5.2.

Example 14.18 Fixed and Random Effects Geometric Regression

Example 14.13 presents pooled estimates for a geometric regression model,

$$f(y_{it}|\mathbf{x}_{it}) = \theta_{it}(1 - \theta_{it})^{y_{it}},\ \theta_{it} = 1/(1 + \lambda_{it}),\ \lambda_{it} = \exp(c_i + \mathbf{x}_{it}'\boldsymbol{\beta}),\ y_{it} = 0, 1, \ldots.$$

We will now reestimate the model under the assumptions of the random and fixed effects specifications. The methods of the preceding two sections are applied directly—no modification of the procedures was required. Table 14.15 presents the three sets of maximum likelihood estimates. The estimates vary considerably. The average group size is about five. This implies that the fixed effects estimator may well be subject to a small sample bias. Save for the coefficient on *Kids*, the fixed effects and random effects estimates are quite similar. On the other hand, the two panel models give similar results to the pooled model except for the *Income* coefficient. On this basis, it is difficult to see, based solely on the results, which should be the preferred model. The model is nonlinear to begin with, so the pooled model, which might otherwise be preferred on the basis of computational ease, now has no redeeming virtues. None of the three models is robust to misspecification. Unlike the linear model, in this and other nonlinear models, the fixed effects *estimator* is inconsistent when T is small in both random and fixed effects cases. The random effects estimator is consistent in the random effects model, but, as usual, not in the fixed effects model. The pooled estimator is inconsistent in both random and fixed effects cases (which calls into question the virtue of the robust covariance matrix). It might be tempting to use a Hausman specification test (see Section 11.5.5); however, the conditions that underlie the test are not met—unlike the linear model where the fixed effects estimator is consistent in both cases, here it is inconsistent in both cases. For better or worse, that leaves the analyst with the need to choose the model based on the underlying theory.

TABLE 14.15 Panel Data Estimates of a Geometric Regression for *DOCVIS*

	Pooled		*Random Effects*[a]		*Fixed Effects*	
Variable	***Estimate***	***Std. Err.***[b]	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
Constant	1.09189	0.10828	0.39936	0.09530		
Age	0.01799	0.00130	0.02209	0.00122	0.04845	0.00351
Education	−0.04725	0.00671	−0.04506	0.00626	−0.05434	0.03721
Income	−0.46836	0.07265	−0.19569	0.06106	−0.18760	0.09134
Kids	−0.15684	0.03055	−0.12434	0.02336	−0.00253	0.03687

[a]Estimated $\sigma_u = 0.95441$.
[b]Standard errors corrected for clusters in the panel.

14.15 LATENT CLASS AND FINITE MIXTURE MODELS

In this final application of maximum likelihood estimation, rather than explore a particular model, we will develop a technique that has been used in many different settings. The latent class modeling framework specifies that the distribution of the observed data is a mixture of a finite number of underlying populations. The model can be motivated in several ways:

- In the classic application of the technique, the observed data are drawn from a mixture of distinct underlying populations. Consider, for example, a historical or fossilized record of the intersection (or collision) of two populations.[48] The anthropological record consists of measurements on some variable that would differ imperfectly, but substantively, between the populations. However, the analyst has no definitive marker for which subpopulation an observation is drawn from. Given a sample of observations, they are interested in two statistical problems: (1) estimate the parameters of the underlying populations (models) and (2) classify the observations in hand as having originated in which population. The technique has seen a number of recent applications in health econometrics. For example, in a study of obesity, Greene, Harris, Hollingsworth, and Maitra (2008) speculated that their ordered choice model (see Chapter 19) might systematically vary in a sample that contained (it was believed) some individuals who have a genetic predisposition toward obesity and most that did not. In another application, Lambert (1992) studied the number of defective outcomes in a production process. When a "zero defectives" condition is observed, it could indicate either regime 1, "the process is under control," or regime 2, "the process is not under control but just happens to produce a zero observation."
- In a narrower sense, one might view parameter heterogeneity in a population as a form of discrete mixing. We have modeled parameter heterogeneity using continuous distributions in Section 11.10. The "finite mixture" approach takes the distribution of parameters across individuals to be discrete. (Of course, this is another way to interpret the first point.)
- The finite mixing approach is a means by which a distribution (model) can be constructed from a mixture of underlying distributions. Quandt and Ramsey's mixture of normals model in Example 13.4 is a case in which a nonnormal distribution is created by mixing two normal distributions with different parameters.

14.15.1 A FINITE MIXTURE MODEL

To lay the foundation for the more fully developed model that follows, we revisit the mixture of normals model from Example 13.4. Consider a population that consists of a latent mixture of two underlying normal distributions. Neglecting for the moment that it is unknown which applies to a given individual, we have, for individual i, one of the following:

[48]The first application of these methods was Pearson's (1894) analysis of 1,000 measures of the "forehead breadth to body length" of two intermingled species of crabs in the Bay of Naples.

$$f(y_i|\,class_i = 1) = \mathrm{N}[\mu_1, \sigma_1^2] = \frac{\exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}},$$

or

$$f(y_i|\,class_i = 2) = \mathrm{N}[\mu_2, \sigma_2^2] = \frac{\exp[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}}. \tag{14-87}$$

The contribution to the likelihood function is $f(y_i|\,class_i = 1)$ for an individual in class 1 and $f(y_i|\,class = 2)$ for an individual in class 2. Assume that there is a true proportion $\lambda = \mathrm{Prob}(class_i = 1)$ of individuals in the population that are in class 1, and $(1 - \lambda)$ in class 2. Then, the unconditional (marginal) density for individual i is

$$\begin{aligned} f(y_i) &= \lambda f(y_i|\,class_i = 1) + (1 - \lambda) f(y_i|\,class_i = 2) \\ &= E_{classes}\, f(y_i|\,class_i). \end{aligned} \tag{14-88}$$

The parameters to be estimated are $\lambda, \mu_1, \mu_2, \sigma_1$, and σ_2. Combining terms, the log likelihood for a sample of n individual observations would be

$$\ln L = \sum_{i=1}^{n} \ln\left(\frac{\lambda \exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}} + \frac{(1 - \lambda)\exp[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}}\right). \tag{14-89}$$

This is the mixture density that we saw in Example 13.4. We suggested the method of moments as an estimator of the five parameters in that example. However, this appears to be a straightforward problem in maximum likelihood estimation.

Example 14.19 A Normal Mixture Model for Grade Point Averages

Appendix Table F14.1 contains a data set of 32 observations used by Spector and Mazzeo (1980) to study whether a new method of teaching economics, the Personalized System of Instruction (PSI), significantly influenced performance in later economics courses. Variables in the data set include

GPA = the student's grade point average,
GRADE = dummy variable for whether the student's grade in Intermediate Macroeconomics was higher than in the principles course,
PSI = dummy variable for whether the individual participated in the PSI,
TUCE = the student's score on a pretest in economics.

We will use these data to develop a finite mixture normal model for the distribution of grade point averages.

We begin by computing maximum likelihood estimates of the parameters in (14-89). To estimate the parameters using an iterative method, it is necessary to devise a set of starting values. It might seem natural to use the simple values from a one-class model, $\bar{y}$ and s_y, and a value such as 1/2 for λ. However, the optimizer will immediately stop on these values, as the derivatives will be zero at this point. Rather, it is common to use some value near these—perturbing them slightly (a few percent), just to get the iterations started. Table 14.16 contains the estimates for this two-class finite mixture model. The estimates for the one-class model are the sample mean and standard deviation of *GPA*. [Because these are the MLEs, $\hat{\sigma}^2 = (1/n)\Sigma_{i=1}^{32}(GPA_i - \overline{GPA})^2$.] The means and standard deviations of the two classes are noticeably different—the model appears to be revealing a distinct splitting of the data into two classes. (Whether two is the appropriate number of classes is considered in Section 14.15.5.) It is tempting at this point to identify the two classes with some other covariate, either in

the data set or not, such as *PSI*. However, at this point, there is no basis for doing so—the classes are "latent." As the analysis continues, however, we will want to investigate whether any observed data help predict the class membership.

14.15.2 MODELING THE CLASS PROBABILITIES

The development thus far has assumed that the analyst has no information about class membership. Estimation of the prior probabilities (λ in the preceding example) is part of the estimation problem. There may be some, albeit imperfect, information about class membership in the sample as well. For our earlier example of grade point averages, we also know the individual's score on a test of economic literacy (*TUCE*). Use of this information might sharpen the estimates of the class probabilities. The mixture of normals distribution, for example, might be formulated

$$f(y_i|\mathbf{z}_i) = \left(\frac{\text{Prob}(class = 1|\mathbf{z}_i) \exp\left[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}} + \frac{[1 - \text{Prob}(class = 1|\mathbf{z}_i)] \exp\left[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2\right]}{\sigma_2\sqrt{2\pi}} \right),$$

where $\mathbf{z}_i$ is the vector of variables that help explain the class probabilities. To make the mixture model amenable to estimation, it is necessary to parameterize the probabilities. The logit probability model is a common device. [See Section 17.2. For applications, see Greene (2005, Section 2.3.3) and references cited.] For the two-class case, this might appear as follows:

$$\text{Prob}(class = 1|\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})},$$
$$\text{Prob}(class = 2|\mathbf{z}_i) = 1 - \text{Prob}(class = 1|\mathbf{z}_i). \qquad \textbf{(14-90)}$$

(The more general J class case is shown in Section 14.15.6.) The log likelihood for the mixture of two normal densities becomes

$$\ln L = \sum_{i=1}^{n} \ln L_i = \sum_{i=1}^{n} \ln \left(\left(\frac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})} \right) \frac{\exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}} + \left(\frac{1}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})} \right) \frac{\exp[-\frac{1}{2}(y_i - \mu_2)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}} \right). \qquad \textbf{(14-91)}$$

The log likelihood is now maximized with respect to μ_1, σ_1, μ_2, σ_2, and $\boldsymbol{\theta}$. If $\mathbf{z}_i$ contains a constant term and some other observed variables, then the earlier model returns if the coefficients on those other variables all equal zero. In this case, it follows that

TABLE 14.16 Estimated Normal Mixture Model

	One Class		*Latent Class* 1		*Latent Class* 2	
Parameter	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
$\boldsymbol{\mu}$	3.1172	0.08251	3.64187	0.3452	2.8894	0.2514
$\boldsymbol{\sigma}$	0.4594	0.04070	0.2524	0.2625	0.3218	0.1095
Probability	1.0000	0.0000	0.3028	0.3497	0.6972	0.3497
ln L	−20.51274		−19.63654			

$\lambda = \ln[\theta/(1 - \theta)]$. (This device is usually used to ensure that $0 < \lambda < 1$ in the earlier model.)

14.15.3 LATENT CLASS REGRESSION MODELS

To complete the construction of the latent class model, we note that the means (and, in principle, the variances) in the original model could be conditioned on observed data as well. For our normal mixture models, we might make the marginal mean, μ_j, a conditional mean,

$$\mu_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j.$$

In the data of Example 14.17, we also observe an indicator of whether the individual has participated in a special program designed to enhance the economics program (PSI). We might modify the model,

$$f(y_i|\, class_i = 1, PSI_i) = N[\mu_{i1}, \sigma_1^2] = \frac{\exp[-\frac{1}{2}(y_i - \beta_{1,1} - \beta_{2,1}PSI_i)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}},$$

and similarly for $f(y_i|\, class_i = 2, PSI_i)$. The model is now a **latent class linear regression model**.

More generally, as we will see shortly, the latent class, or **finite mixture model** for a variable y_i can be formulated as

$$f(y_i|\, class_i = j, \mathbf{x}_i) = h_j(y_i, \mathbf{x}_i, \boldsymbol{\gamma}_j),$$

where h_j denotes the density conditioned on class j—indexed by j to indicate, for example, the jth parameter vector $\boldsymbol{\gamma}_j = (\boldsymbol{\beta}_j, \sigma_j)$ and so on. The marginal class probabilities are

$$\text{Prob}(class_i = j|\mathbf{z}_i) = p_j(j, \mathbf{z}_i, \boldsymbol{\theta}).$$

The methodology can be applied to any model for y_i. In the example in Section 14.15.6, we will model a binary dependent variable with a probit model. The methodology has been applied in many other settings, such as stochastic frontier models [Orea and Kumbhakar (2004), Greene (2004)], Poisson regression models [Wedel et al. (1993)], and a wide variety of count, discrete choice, and limited dependent variable models [McLachlan and Peel (2000), Greene (2007b)].

Example 14.20 Latent Class Regression Model for Grade Point Averages

Combining 14.15.2 and 14.15.3, we have a latent class model for grade point averages,

$$f(GPA_i|\, class_i = j, PSI_i) = \frac{\exp[-\frac{1}{2}(y_i - \beta_{1j} - \beta_{2j}PSI_i)^2/\sigma_j^2]}{\sigma_j\sqrt{2\pi}}, j = 1, 2,$$

$$\text{Prob}(class_i = 1|\, TUCE_i) = \frac{\exp(\theta_1 + \theta_2 TUCE_i)}{1 + \exp(\theta_1 + \theta_2 TUCE_i)},$$

$$\text{Prob}(class_i = 2|\, TUCE_i) = 1 - \text{Prob}(class = 1|\, TUCE_i).$$

The log likelihood is now

$$\ln L = \sum_{i=1}^{n} \ln \left(\begin{array}{l} \left(\dfrac{\exp(\theta_1 + \theta_2 TUCE_i)}{1 + \exp(\theta_1 + \theta_2 TUCE_i)}\right) \dfrac{\exp[-\frac{1}{2}(y_i - \beta_{1,1} - \beta_{2,1}PSI_i)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}} \\ + \left(\dfrac{1}{1 + \exp(\theta_1 + \theta_2 TUCE_i)}\right) \dfrac{\exp[-\frac{1}{2}(y_i - \beta_{1,2} - \beta_{2,2}PSI_i)^2/\sigma_2^2]}{\sigma_2\sqrt{2\pi}} \end{array} \right).$$

Maximum likelihood estimates of the parameters are given in Table 14.17.

TABLE 14.17 Estimated Latent Class Linear Regression Model for GPA

	One Class		*Latent Class 1*		*Latent Class 2*	
Parameter	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
β_1	3.1011	0.1117	3.3928	0.1733	2.7926	0.04988
β_2	0.03675	0.1689	−0.1074	0.2006	−0.5703	0.07553
σ	0.4443	0.0003086	0.3812	0.09337	0.1119	0.04487
θ_1	0.0000	0.0000	−6.8392	3.07867	0.0000	0.0000
θ_2	0.0000	0.0000	0.3518	0.1601	0.0000	0.0000
$P(\text{class}\,\vert\, TUCE)$	1.0000		0.7063		0.2937	
$\ln L$	−20.48752		−13.39966			

14.15.4 PREDICTING CLASS MEMBERSHIP AND $\boldsymbol{\beta}_i$

The model in (14-91) now characterizes two random variables, y_i, the outcome variable of interest, and $class_i$, the indicator of which class the individual resides in. We have a joint distribution, $f(y_i, class_i)$, which we are modeling in terms of the conditional density, $f(y_i|class_i)$ in (14-87), and the marginal density of $class_i$ in (14-90). We have initially assumed the latter to be a simple Bernoulli distribution with $\text{Prob}(class_i = 1) = \lambda$, but then modified in the previous section to equal $\text{Prob}(class_i = 1|\mathbf{z}_i) = \Lambda(\mathbf{z}_i'\boldsymbol{\theta})$. These can be viewed as the prior probabilities in a Bayesian sense. If we wish to make a prediction as to which class the individual came from, using all the information that we have on that individual, then the prior probability is going to waste some information; it wastes the information on the observed outcome. The posterior, or conditional (on the remaining data) probability,

$$\text{Prob}(class_i = 1|\mathbf{z}_i\, y_i) = \frac{f(y_i, class = 1|\mathbf{z}_i)}{f(y_i)},$$

will be based on more information than the marginal probabilities. We have the elements that we need to compute this conditional probability. Use Bayes's theorem to write this as

$$\text{Prob}(class_i = 1|\mathbf{z}_i, y_i) = \frac{f(y_i|\text{class}_\text{i} = 1, \mathbf{z}_i)\ \text{Prob}(class_i = 1|\mathbf{z}_i)}{f(y_i|class_i = 1, \mathbf{z}_i)\ \text{Prob}(class_i = 1|\mathbf{z}_i) + f(y_i|class_i = 2, \mathbf{z}_i)\ \text{Prob}(class_i = 2|\mathbf{z}_i)}.$$

The denominator is L_i (not $\ln L_i$) from (14-91). The numerator is the first term in L_i. To continue our mixture of two normals example, the conditional (posterior) probability is

$$\text{Prob}(class_i = 1|\mathbf{z}_i, y_i) = \frac{\left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})}\right)\dfrac{\exp[-\frac{1}{2}(y_i - \mu_1)^2/\sigma_1^2]}{\sigma_1\sqrt{2\pi}}}{L_i},$$

while the unconditional probability is in (14-90). The conditional probability for the second class is computed using the other two marginal densities in the numerator (or by subtraction from one). Note that the conditional probabilities are functions of the data even if the unconditional ones are not. To come to the problem suggested at the outset,

then, the natural predictor of $class_i$ is the class associated with the largest estimated posterior probability.

In random parameter settings, we have also been interested in predicting $E[\boldsymbol{\beta}_i | y_i, \mathbf{X}_i]$. There are two candidates for the latent class model. Having made the best guess as to which specific class an individual resides in, a natural estimator of β_i would be the β_j associated with that class. A preferable estimator that uses more information would be the posterior expected value,

$$\hat{E}[\boldsymbol{\beta}_i | y_i, \mathbf{X}_i, z_i] = \sum_{j=1}^{J} \hat{\pi}_{ij}(\hat{\boldsymbol{\Theta}}, \mathbf{z}_i)\hat{\boldsymbol{\beta}}_j.$$

Example 14.21 Predicting Class Probabilities

Table 14.18 lists the observations sorted by GPA. The predictions of class membership reflect what one might guess from the coefficients in the table of coefficients. Class 2 members on average have lower GPAs than in class 1. The listing in Table 14.18 shows this clustering. It

TABLE 14.18 Estimated Latent Class Probabilities

GPA	*TUCE*	*PSI*	*CLASS*	P_1	P_1*	P_2	P_2*
2.06	22	1	2	0.7109	0.0116	0.2891	0.9884
2.39	19	1	2	0.4612	0.0467	0.5388	0.9533
2.63	20	0	2	0.5489	0.1217	0.4511	0.8783
2.66	20	0	2	0.5489	0.1020	0.4511	0.8980
2.67	24	1	1	0.8325	0.9992	0.1675	0.0008
2.74	19	0	2	0.4612	0.0608	0.5388	0.9392
2.75	25	0	2	0.8760	0.3499	0.1240	0.6501
2.76	17	0	2	0.2975	0.0317	0.7025	0.9683
2.83	19	0	2	0.4612	0.0821	0.5388	0.9179
2.83	27	1	1	0.9345	1.0000	0.0655	0.0000
2.86	17	0	2	0.2975	0.0532	0.7025	0.9468
2.87	21	0	2	0.6336	0.2013	0.3664	0.7987
2.89	14	1	1	0.1285	1.0000	0.8715	0.0000
2.89	22	0	2	0.7109	0.3065	0.2891	0.6935
2.92	12	0	2	0.0680	0.0186	0.9320	0.9814
3.03	25	0	1	0.8760	0.9260	0.1240	0.0740
3.10	21	1	1	0.6336	1.0000	0.3664	0.0000
3.12	23	1	1	0.7775	1.0000	0.2225	0.0000
3.16	25	1	1	0.8760	1.0000	0.1240	0.0000
3.26	25	0	1	0.8760	0.9999	0.1240	0.0001
3.28	24	0	1	0.8325	0.9999	0.1675	0.0001
3.32	23	0	1	0.7775	1.0000	0.2225	0.0000
3.39	17	1	1	0.2975	1.0000	0.7025	0.0000
3.51	26	1	1	0.9094	1.0000	0.0906	0.0000
3.53	26	0	1	0.9094	1.0000	0.0906	0.0000
3.54	24	1	1	0.8325	1.0000	0.1675	0.0000
3.57	23	0	1	0.7775	1.0000	0.2225	0.0000
3.62	28	1	1	0.9530	1.0000	0.0470	0.0000
3.65	21	1	1	0.6336	1.0000	0.3664	0.0000
3.92	29	0	1	0.9665	1.0000	0.0335	0.0000
4.00	21	0	1	0.6336	1.0000	0.3664	0.0000
4.00	23	1	1	0.7775	1.0000	0.2225	0.0000

also suggests how the latent class model is using the sample information. If the results in Table 14.16—just estimating the means, constant class probabilities—are used to produce the same table, when sorted, the highest 10 GPAs are in class 1 and the remainder are in class 2. The more elaborate model is adding information on *TUCE* to the computation. A low *TUCE* score can push a high GPA individual into class 2. (Of course, this is largely what multiple linear regression does as well.)

14.15.5 DETERMINING THE NUMBER OF CLASSES

There is an unsolved inference issue remaining in the specification of the model. The number of classes has been taken as a known parameter—two in our main example thus far, three in the following application. Ideally, one would like to determine the appropriate number of classes statistically. However, J is not a parameter in the model. A likelihood ratio test, for example, will not provide a valid result. Consider the original model in Example 14.17. The model has two classes and five parameters in total. It would seem natural to test down to a one-class model that contains only the mean and variance using the LR test. However, the number of restrictions here is actually ambiguous. If $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, then the mixing probability is irrelevant—the two class densities are the same, and it is a one-class model. Thus, the number of restrictions needed to get from the two-class model to the one-class model is ambiguous. It is neither two nor three. One strategy that has been suggested is to test upward, adding classes until the marginal class insignificantly changes the log likelihood or one of the information criteria such as the AIC or BIC (see Section 14.6.5). Unfortunately, this approach is likewise problematic because the estimates from any specification that is too short are inconsistent. The alternative would be to test down from a specification known to be too large. Heckman and Singer (1984b) discuss this possibility and note that when the number of classes becomes larger than appropriate, the estimator should break down. In our Example 14.15, if we expand to four classes, the optimizer breaks down, and it is no longer possible to compute the estimates. A five-class model does produce estimates, but some are nonsensical. This does provide at least the directions to seek a viable strategy. The authoritative treatise on finite mixture models by McLachlan and Peel (2000, Chapter 6) contains extensive discussion of this issue.

14.15.6 A PANEL DATA APPLICATION

The latent class model is a useful framework for applications in panel data. The class probabilities partly play the role of common random effects, as we will now explore. The latent class model can be interpreted as a random parameters model with a discrete distribution of the parameters.

Suppose that $\boldsymbol{\beta}_j$ is generated from a discrete distribution with J outcomes, or classes, so that the distribution of $\boldsymbol{\beta}_j$ is over these classes. Thus, the model states that an individual belongs to one of the J latent classes, indexed by the parameter vector, but it is unknown from the sample data exactly which one. We will use the sample data to estimate the parameter vectors, the parameters of the underlying probability distribution and the probabilities of class membership. The corresponding model formulation is now

$$f(y_{it}|\mathbf{x}_{it}, \mathbf{z}_i, \boldsymbol{\Delta}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ \ldots, \boldsymbol{\beta}_J) = \sum_{j=1}^{J} p_{ij}(\mathbf{z}_i, \boldsymbol{\Delta}) f(y_{it}|\, class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j), \qquad \textbf{(14-92)}$$

where it remains to parameterize the class probabilities, p_{ij}, and the structural model, $f(y_{it}|\, class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)$. The parameter matrix, $\boldsymbol{\Delta}$, contains the parameters of the discrete

probability distribution. It has J rows, one for each class, and M columns, for the M variables in $\mathbf{z}_i$. At a minimum, $M = 1$ and $\mathbf{z}_i$ contains a constant term if the class probabilities are fixed parameters as in Example 14.17. Finally, to accommodate the panel data nature of the sampling situation, we suppose that conditioned on $\boldsymbol{\beta}_j$, that is, on membership in class j, which is fixed over time, the observations on y_{it} are independent. Therefore, for a group of T_i observations, the joint density is

$$f(y_{i1}, y_{i2}, \ldots, y_{t,T_i} | \, class = j, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) = \prod_{t=1}^{T_i} f(y_{it} | \, class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j).$$

The log-likelihood function for a panel of data is

$$\ln L = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{J} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i) \prod_{t=1}^{T_i} f(y_{it} | \, class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j) \right]. \tag{14-93}$$

The class probabilities must be constrained be in (0,1) and to sum to 1. The approach that is usually used is to reparameterize them as a set of logit probabilities, as we did in the preceding examples. Then,

$$p_{ij}(\mathbf{z}_i, \boldsymbol{\Delta}) = \frac{\exp(\theta_{ij})}{\sum_{j=1}^{J} \exp(\theta_{ij})}, j = 1, \ldots, J, \theta_{ij} = \mathbf{z}_i' \boldsymbol{\delta}_j, \theta_{iJ} = 0 \; (\boldsymbol{\delta}_J = \mathbf{0}). \tag{14-94}$$

(See Section 18.2.2 for development of this model for the set of probabilities.) Note the restriction on θ_{ij}. This is an identification restriction. Without it, the same set of probabilities will arise if an arbitrary vector is added to every $\boldsymbol{\delta}_j$. The resulting log likelihood is a continuous function of the parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ and $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_J$. For all its apparent complexity, estimation of this model by direct maximization of the log likelihood is not especially difficult.[49] The number of classes that can be identified is likely to be relatively small (on the order of 5 or 10 at most), however, which has been viewed as a drawback of the approach. In general, the more complex the model for y_{it}, the more difficult it becomes to expand the number of classes. Also, as might be expected, the less rich the data set in terms of cross-group variation, the more difficult it is to estimate latent class models.

Estimation produces values for the structural parameters, $(\boldsymbol{\beta}_j, \boldsymbol{\delta}_j), j = 1, \ldots, J$. With these in hand, we can compute the prior class probabilities, p_{ij}, using (14-94). For prediction purposes, we are also interested in the posterior (to the data) class probabilities, which we can compute using Bayes' theorem [see (14-93)]. The conditional probability is

$$\begin{aligned} &\text{Prob}(class = j | \, \text{observation } i) \\ &\quad = \frac{f(\text{observation } i | \, class = j) \text{Prob}(class \; j)}{\sum_{j=1}^{\text{J}} f(\text{observation } i | \, class = j) \; \text{Prob}(class \; j)} \\ &\quad = \frac{f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} | \, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) p_{ij}(\mathbf{z}_j, \boldsymbol{\Delta})}{\sum_{j=1}^{J} f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} | \, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \boldsymbol{\beta}_j) p_{ij}(\mathbf{z}_j, \boldsymbol{\Delta})} \\ &\quad = w_{ij}. \end{aligned} \tag{14-95}$$

[49] See Section E.3 and Greene (2001, 2007b). The EM algorithm discussed in Section E.3.7 is especially well suited for estimating the parameters of latent class models. See McLachlan and Peel (2000).

The set of probabilities, $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iJ})$, gives the posterior density over the distribution of values of $\boldsymbol{\beta}$, that is, $[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J]$. For a particular model and allowing for grouping within a panel data set, the posterior probability for class j is found as

$$\begin{aligned}\text{Prob}(class = j|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i) &= \frac{p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i)\prod_{t=1}^{T_i} f(y_{it}|class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)}{\sum_{j=1}^{J} p_{ij}(\boldsymbol{\Delta}, \mathbf{z}_i)\prod_{t=1}^{T_i} f(y_{it}|class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)} \\ &= \frac{\left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\Delta}_j)}{\Sigma_{m=1}^{J}\exp(\mathbf{z}_i'\boldsymbol{\Delta}_m)}\right)\prod_{t=1}^{T_i} f(y_{it}|class = j, \mathbf{x}_{it}, \boldsymbol{\beta}_j)}{\sum_{j=1}^{J}\left(\dfrac{\exp(\mathbf{z}_i'\boldsymbol{\Delta}_j)}{\Sigma_{m=1}^{J}\exp(\mathbf{z}_i'\boldsymbol{\Delta}_m)}\right)\prod_{t=1}^{T_i} f(y_{it}|class = m, \mathbf{x}_{it}, \boldsymbol{\beta}_m)}.\end{aligned} \tag{14-96}$$

Example 14.22 A Latent Class Two-Part Model for Health Care Utilization

Jones and Bago D'Uva (2009) examined health care utilization in Europe using 8 waves of the ECHP panel data set. The variable of interest was numbers of visits to the physician. They examined two outcomes, visits to general practitioners and visits to specialists. The modeling framework was the latent class model in (14-92). The class-specific model was a two-part, negative binomial "hurdle" model for counts,

$$\text{Prob}(y_{it} = 0|\mathbf{x}_{it}, \boldsymbol{\beta}_{1j}) = \frac{1}{1 + \lambda_{1it,j}}, \lambda_{1it,j} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}_{1j})$$

$$\text{Prob}(y_{it}|y_{it} > 0, \mathbf{x}_{it}, \boldsymbol{\beta}_{2j}, \alpha_j) = \frac{(\alpha_j\lambda_{2it,j} + 1)^{-1/\alpha_j}\,\Gamma(y_{it} + 1/\alpha_j)[1 + (\lambda_{2it,j}^{-1}/\alpha_j)]^{-y_{it}}}{\Gamma(1/\alpha_j)\Gamma(y_{it} + 1)[1 - (\alpha_j\lambda_{2it,j} + 1)^{-1/\alpha_j}]},$$

$$\lambda_{2it,j} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}_{2j}), \alpha_j > 0.$$

[This is their equation (2) with $k = 0$.] The first equation is a participation equation, for whether the number of doctor visits equals 0 or some positive value. The second equation is the intensity equation that predicts the number of visits, given that the number of visits is positive. The count model is a *negative binomial model*. This is an extension of the Poisson regression model. The Poisson model is a limiting case when $\alpha_j \to 0$. The hurdle and count equations involve different coefficient vectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, so that the determinants of care have different effects on the two stages. Interpretation of this model is complicated by the results that variables appear in both equations, and that the conditional mean function is complex. The simple conditional mean, if there were no hurdle effects, would be $E[y_{it}|\mathbf{x}_{it}] = \lambda_{2it}$. However, with the hurdle effects,

$$E[y_{it}|\mathbf{x}_{it}] = \text{Prob}(y_{\text{it}} > 0|\mathbf{x}_{it}) \times \text{E}[y_{it}|y_{it} > 0, \mathbf{x}_{it}].$$

The authors examined the two components of this result separately. (The elasticity of the mean would be the sum of these two elasticities.) The mixture model involves two classes (as typical in this literature) A sampling of their results appears in Table 14.19 below. (The results are extracted from their Table 8.) Note that separate tables are given for "Low Users" and "High Users." The results in Section 14.15.4 are used to classify individuals into class 1 and class 2. It is then discovered that the average usage of those individuals classified as in class 1 is far lower than the average use of those in class 2.

TABLE 14.19 Country-Specific Estimated Income Coefficients and Elasticities for GP Visits

		Low Users		*High Users*	
Country		***Coefficient***	***Elasticity***	***Coefficient***	***Elasticity***
Austria	$P(y > 0)$	−0.051	−0.012	−0.109	−0.005
	$E[y \mid y > 0]$	0.012	0.009	0.039	0.035
Denmark	$P(y > 0)$	0.083	0.033	0.261	0.023
	$E[y \mid y > 0]$	0.042	0.021	−0.030	−0.024
The Netherlands	$P(y > 0)$	0.082	0.035	0.094	0.009
	$E[y \mid y > 0]$	−0.037	−0.019	−0.085	−0.068

Example 14.23 Latent Class Models for Health Care Utilization

In Examples 7.6 and 11.21, we proposed an exponential regression model,

$$y_{it} = DocVis_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) + \varepsilon_{it},$$

for the variable *DocVis*, the number of visits to the doctor, in the German health care data. (See Example 11.20 for details.) The regression results for the specification,

$$\mathbf{x}_{it} = (1, Age_{it}, Education_{it}, Income_{it}, Kids_{it}),$$

are repeated (in parentheses) in Table 14.20 for convenience. The nonlinear least squares estimator is only semiparametric; it makes no assumption about the distribution of $DocVis_{it}$ or about ε_{it}. We do see striking increases in the standard errors when the "cluster robust" asymptotic covariance matrix is used. (The estimates are given in parentheses.) The analysis at this point assumes that the nonlinear least squares estimator remains consistent in the presence of the cross-observation correlation. Given the way the model is specified, that is, only in terms of the conditional mean function, this is probably reasonable. The extension would imply a nonlinear generalized regression as opposed to a nonlinear ordinary regression.

TABLE 14.20 Panel Data Estimates of a Geometric Regression for DOCVIS

	Pooled		*Random Effects*[a]		*Fixed Effects*	
Variable	***Estimate***	***Std. Err.***[b]	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
Constant	1.09189 (0.98017)[c]	0.10828 (0.18137)	0.39936	0.09530		
Age	0.01799 (0.01873)	0.00130 (0.00198)	0.02209	0.00122	0.04845	0.00351
Education	−0.04725 (−0.03609)	0.00671 (0.01287)	−0.04506	0.00626	−0.05434	0.03721
Income	−0.46836 (−0.59189)	0.07265 (0.12827)	−0.19569	0.06106	−0.18760	0.09134
Kids	−0.15684 (−0.16930)	0.03055 (0.04882)	−0.12434	0.02336	−0.00253	0.03687

[a]Estimated $\sigma_u = 0.95441$.
[b]Standard errors corrected for clusters in the panel.
[c]Nonlinear least squares results in parentheses.

In Example 14.13, we narrowed this model by assuming that the observations on doctor visits were generated by a geometric distribution,

$$f(y_i|\mathbf{x}_i) = \theta_i(1-\theta_i)^{y_i}, \theta_i = 1/(1+\lambda_i), \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), y_i = 0, 1, \ldots.$$

The conditional mean is still $\exp(\mathbf{x}_{it}'\boldsymbol{\beta})$, but this specification adds the structure of a particular distribution for outcomes. The pooled model was estimated in Example 14.13. Examples 14.17 and 14.18 added the panel data assumptions of random, then fixed effects, to the model. The model is now

$$f(y_{it}|\mathbf{x}_{it}) = \theta_{it}(1-\theta_{it})^{y_{it}}, \theta_{it} = 1/(1+\lambda_{it}), \lambda_{it} = \exp(c_i + \mathbf{x}_{it}'\boldsymbol{\beta}), y_{it} = 0, 1, \ldots.$$

The pooled, random effects and fixed effects estimates appear in Table 14.17. The pooled estimates, where the standard errors are corrected for the panel data grouping, are comparable to the nonlinear least squares estimates with the robust standard errors. The parameter estimates are similar—both are consistent and this is a very large sample. The smaller standard errors seen for the MLE are the product of the more detailed specification. We will now relax the specification by assuming a two-class finite mixture model. We also specify that the class probabilities are functions of gender and marital status. For the latent class specification,

$$\text{Prob}(class_i = 1|\mathbf{z}_i) = \Lambda(\theta_1 + \theta_2 Female_i + \theta_3 Married_i).$$

The model structure is the geometric regression as before. Estimates of the parameters of the latent class model are shown in Table 14.21. See Section E3.7 for discussion of estimation methods.

Deb and Trivedi (2002) and Bago D'Uva and Jones (2009) suggested that a meaningful distinction between groups of health care system users would be between *infrequent* and *frequent* users. To investigate whether our latent class model is picking up this distinction in the data, we used (14-96) to predict the class memberships (class 1 or 2). We then linearly regressed $DocVis_{it}$ on a constant and a dummy variable for class 2. The results are

$$DocVis_{it} = 5.8034\ (0.0465) - 4.7801\ (0.06282)Class2_i + e_{it},$$

TABLE 14.21 Estimated Latent Class Geometric Regression Model for DocVis

	One Class		*Latent Class 1*		*Latent Class 2*	
Parameter	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
β_1	1.0918	0.1082	1.6423	0.05351	−0.3344	0.09288
β_2	0.0180	0.0013	0.01691	0.0007324	0.02649	0.001248
β_3	−0.0473	0.0067	−0.04473	0.003451	−0.06502	0.005739
β_4	−0.4687	0.0726	−0.4567	0.04688	0.01395	0.06964
β_5	−0.1569	0.0306	−0.1177	0.01611	−0.1388	0.02738
θ_1	0.000	0.000	−0.4280	0.06938	0.0000	0.0000
θ_2	0.000	0.000	0.8255	0.06322	0.0000	0.0000
θ_3	0.000	0.000	−0.07829	0.07143	0.0000	0.0000
Prob|z	1.0000		0.47697		0.52303	
ln L	−61917.97		−58708.63			

where estimated standard errors are in parentheses. The linear regression suggests that the class membership dummy variable is strongly segregating the observations into frequent and infrequent users. The information in the regression is summarized in the descriptive statistics in Table 14.22.

Finally, we did a specification search for the number of classes. Table 14.23 reports the log likelihoods and AICs for models with 1 to 8 classes. The lowest value of the AIC occurs with 7 classes, although the marginal improvement ends near to $J = 4$. The rightmost 8 columns show the averages of the conditional probabilities, which equal the unconditional probabilities. Note that when $J = 8$, three of the classes (2, 5, and 6) have extremely small probabilities. This suggests that the model might be overspecified. We will see another indicator in the next section.

14.15.7 A SEMIPARAMETRIC RANDOM EFFECTS MODEL

Heckman and Singer (1984a,b) suggested a semiparametric maximum likelihood approach to modeling latent heterogeneity in a duration model (Section 19.5) for unemployment spells. The methodology applies equally well to other settings, such as the one we are examining here. Their method can be applied as a finite mixture model in which only the constant term varies across classes. The log likelihood in this case would be

$$\ln L = \sum_{i=1}^{n} \ln \sum_{j=1}^{J} \pi_j \left(\prod_{t=1}^{T_i} f(y_{it} | \alpha_j + \mathbf{x}_{it}'\boldsymbol{\beta}) \right). \tag{14-97}$$

TABLE 14.22 Descriptive Statistics for Doctor Visits

Class	*Mean*	*Standard Deviation*
All, $n = 27{,}326$	3.18352	5.68979
Class 1, $n = 12{,}349$	5.80347	7.47579
Class 2, $n = 14{,}977$	1.02330	1.63076

TABLE 14.23 Specification Search for Number of Latent Classes

J	*ln L*	*AIC*	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
1	−61917.77	1.23845	1.0000							
2	−58708.48	1.17443	0.4770	0.5230						
3	−58036.15	1.16114	0.2045	0.6052	0.1903					
4	−57953.02	1.15944	0.1443	0.5594	0.2407	0.0601				
5	−57866.34	1.15806	0.0708	0.0475	0.4107	0.3731	0.0979			
6	−57829.96	1.15749	0.0475	0.0112	0.2790	0.1680	0.4380	0.0734		
7	−57808.50	1.15723	0.0841	0.0809	0.0512	0.3738	0.0668	0.0666	0.2757	
8	−57808.07	1.15738	0.0641	0.0038	0.4434	0.3102	0.0029	0.0002	0.1115	0.0640

This is a restricted form of (14-93). The specification is a random effects model in which the heterogeneity has a discrete, multinomial distribution with unconditional mixing probabilities.

Example 14.24 Semiparametric Random Effects Model

Estimates of a random effects geometric regression model are given in Table 14.17. The random effect (random constant term) is assumed to be normally distributed; the estimated standard deviation is 0.95441. Tables 14.24 and 14.25 present estimates of the semiparametric random effects model. The estimated constant terms and class probabilities are shown in Table 14.24. We fit mixture models for 2 through 7 classes. The AIC stopped falling at $J = 7$. The results for 6 and 7 are shown in the table. Note in the 7 class model, the estimated standard errors for the constants for classes 2 and 4 are essentially infinite—the values shown are the result of rounding error. As Heckman and Singer noted, this should be taken as evidence of overfitting the data. The remaining coefficients for the parametric parts of the model are shown in Table 14.25. The two approaches to fitting the random effects model produce similar results. The coefficients on the regressors and their estimated standard errors are very similar. The random effects in the normal model are estimated to have a mean of 0.39936 and standard deviation of 0.95441. The multinomial distribution in the mixture model has estimated mean 0.27770 and standard deviation 1.2333. Figure 14.7 shows a comparison of the two estimated distributions.[50]

TABLE 14.24 Heckman and Singer Semiparametric Random Effects Model

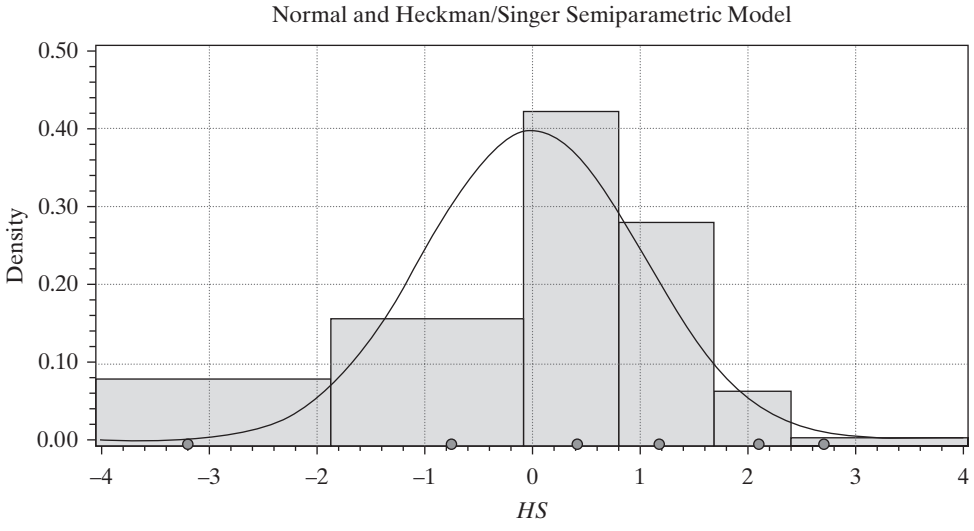
Class	α	*Std. Err.*	*P(class)*	α	*Std. Err.*	*P(class)*
1	−3.17815	0.28542	0.07394	−0.72948	0.16886	0.16825
2	−0.72948	0.15847	0.16825	1.23774	358561.2	0.04030
3	0.38886	0.11867	0.41734	0.38886	0.15112	0.41734
4	1.23774	0.12295	0.28452	1.23774	59175.41	0.24421
5	2.11958	0.28568	0.05183	2.11958	0.41549	0.05183
6	2.69846	0.98622	0.00412	2.69846	1.17124	0.00412
7				−3.17815	0.28863	0.07394

TABLE 14.25 Estimated Random Effects Exponential Count Data Model

	Finite Mixture Model		*Normal Random Effects Model*	
	Estimate	*Std. Err.*	*Estimate*	*Std. Err .*
Constant	$\bar{\hat{\alpha}} = 0.277697$		0.39936	0.09530
Age	0.02136	0.00115	0.02209	0.00122
Educ.	−0.03877	0.00607	−0.04506	0.00626
Income	−0.23729	0.05972	−0.19569	0.06106
Kids	−0.12611	0.02280	−0.12434	0.02336
	$s_\alpha = 1.23333$		$\sigma_u = 0.95441$	

[50]The multinomial distribution has interior boundaries at the midpoints between the estimated constants. The mass points have heights equal to the probabilities. The rectangles sum to slightly more than one—about 1.15. The figure is only a sketch of an implied approximation to the normal distribution in the parametric model.

FIGURE 14.7 Estimated Distributions of Random Effects.



14.16 SUMMARY AND CONCLUSIONS

This chapter has presented the theory and several applications of maximum likelihood estimation, which is the most frequently used estimation technique in econometrics after least squares. The maximum likelihood estimators are consistent, asymptotically normally distributed, and efficient among estimators that have these properties. The drawback to the technique is that it requires a fully parametric, detailed specification of the data-generating process. As such, it is vulnerable to misspecification problems. Chapter 13 considered GMM estimation techniques that are less parametric, but more robust to variation in the underlying data-generating process. Together, ML and GMM estimation account for the large majority of empirical estimation in econometrics.

Key Terms and Concepts

- Asymptotic variance
- BHHH estimator
- Butler and Moffitt's method
- Concentrated log likelihood
- Efficient score
- Finite mixture model
- Gauss–Hermite quadrature
- Generalized sum of squares
- Incidental parameters problem
- Index function model
- Information matrix equality
- Kullback–Leibler information criterion (KLIC)
- Lagrange multiplier statistic
- Lagrange multiplier (LM) test
- Latent class linear regression model
- Likelihood equation
- Likelihood ratio
- Likelihood ratio index
- Likelihood ratio statistic
- Likelihood ratio (LR) test
- Limited Information Maximum Likelihood
- Logistic probability model
- Loglinear conditional mean
- Maximum likelihood
- Method of scoring
- Newton's method
- Noncentral chi-squared distribution

- Nonlinear least squares
- Nonnested models
- Oberhofer–Kmenta estimator
- Outer product of gradients estimator (OPG)
- Precision parameter
- Pseudo-log-likelihood function
- Pseudo-MLE
- Quasi-MLE
- Random effects
- Regularity conditions
- Score test
- Score vector
- Vuong test

Exercises

1. Assume that the distribution of x is $f(x) = 1/\theta, 0 \leq x \leq \theta$. In random sampling from this distribution, prove that the sample maximum is a consistent estimator of θ. *Note:* You can prove that the maximum is the maximum likelihood estimator of θ. But the usual properties do not apply here. Why not? (*Hint:* Attempt to verify that the expected first derivative of the log likelihood with respect to θ is zero.)
2. In random sampling from the exponential distribution $f(x) = (1/\theta)e^{-x/\theta}, x \geq 0$, $\theta > 0$, find the maximum likelihood estimator of θ and obtain the asymptotic distribution of this estimator.
3. **Mixture distribution.** Suppose that the joint distribution of the two random variables x and y is

$$f(x, y) = \frac{\theta e^{-(\beta+\theta)y}(\beta y)^x}{x!}, \quad \beta, \theta > 0, y \geq 0, x = 0, 1, 2, \ldots.$$

   a. Find the maximum likelihood estimators of β and θ and their asymptotic joint distribution.
   b. Find the maximum likelihood estimator of $\theta/(\beta + \theta)$ and its asymptotic distribution.
   c. Prove that $f(x)$ is of the form

$$f(x) = \gamma(1 - \gamma)^x, x = 0, 1, 2, \ldots,$$

   and find the maximum likelihood estimator of γ and its asymptotic distribution.
   d. Prove that $f(y|x)$ is of the form

$$f(y|x) = \frac{\lambda e^{-\lambda y}(\lambda y)^x}{x!}, \quad y \geq 0, \lambda > 0.$$

   Prove that $f(y|x)$ integrates to 1. Find the maximum likelihood estimator of λ and its asymptotic distribution. (*Hint:* In the conditional distribution, just carry the x's along as constants.)
   e. Prove that

$$f(y) = \theta e^{-\theta y}, \quad y \geq 0, \quad \theta > 0.$$

   Find the maximum likelihood estimator of θ and its asymptotic variance.
   f. Prove that

$$f(x|y) = \frac{e^{-\beta y}(\beta y)^x}{x!}, \quad x = 0, 1, 2, \ldots, \beta > 0.$$

   Based on this distribution, what is the maximum likelihood estimator of β?

4. Suppose that x has the Weibull distribution

$$f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, \quad x \geq 0, \alpha, \beta > 0.$$

   a. Obtain the log-likelihood function for a random sample of n observations.
   b. Obtain the likelihood equations for maximum likelihood estimation of α and β. Note that the first provides an explicit solution for α in terms of the data and β. But, after inserting this in the second, we obtain only an implicit solution for β. How would you obtain the maximum likelihood estimators?
   c. Obtain the second derivatives matrix of the log likelihood with respect to α and β. The exact expectations of the elements involving β involve the derivatives of the gamma function and are quite messy analytically. Of course, your exact result provides an empirical estimator. How would you estimate the asymptotic covariance matrix for your estimators in part b?
   d. Prove that $\alpha\beta$ Cov[ln x, x^{β}] = 1. (*Hint:* The expected first derivatives of the log-likelihood function are zero.)
5. The following data were generated by the Weibull distribution of Exercise 4:

1.3043	0.49254	1.2742	1.4019	0.32556	0.29965	0.26423
1.0878	1.9461	0.47615	3.6454	0.15344	1.2357	0.96381
0.33453	1.1227	2.0296	1.2797	0.96080	2.0070	

   a. Obtain the maximum likelihood estimates of α and β, and estimate the asymptotic covariance matrix for the estimates.
   b. Carry out a Wald test of the hypothesis that $\beta = 1$.
   c. Obtain the maximum likelihood estimate of α under the hypothesis that $\beta = 1$.
   d. Using the results of parts a and c, carry out a likelihood ratio test of the hypothesis that $\beta = 1$.
   e. Carry out a Lagrange multiplier test of the hypothesis that $\beta = 1$.
6. **Limited Information Maximum Likelihood Estimation**. Consider a bivariate distribution for x and y that is a function of two parameters, α and β. The joint density is $f(x, y|\alpha, \beta)$. We consider maximum likelihood estimation of the two parameters. The full information maximum likelihood estimator is the now familiar maximum likelihood estimator of the two parameters. Now, suppose that we can factor the joint distribution as done in Exercise 3, but in this case, we have $f(x, y|\alpha, \beta) = f(y|x, \alpha, \beta)f(x|\alpha)$. That is, the conditional density for y is a function of both parameters, but the marginal distribution for x involves only α.
   a. Write down the general form for the log-likelihood function using the joint density.
   b. Because the joint density equals the product of the conditional times the marginal, the log-likelihood function can be written equivalently in terms of the factored density. Write this down, in general terms.
   c. The parameter α can be estimated by itself using only the data on x and the log likelihood formed using the marginal density for x. It can also be estimated with β by using the full log-likelihood function and data on both y and x. Show this.
   d. Show that the first estimator in part c has a larger asymptotic variance than the second one. This is the difference between a limited information maximum likelihood estimator and a full information maximum likelihood estimator.
   e. Show that if $\partial^2 \ln f(y|x, \alpha, \beta)/\partial\alpha\partial\beta = 0$, then the result in part d is no longer true.

7. Show that the likelihood inequality in Theorem 14.3 holds for the Poisson distribution used in Section 14.3 by showing that $E[(1/n) \ln L(\theta|y)]$ is uniquely maximized at $\theta = \theta_0$. (*Hint:* First show that the expectation is $-\theta + \theta_0 \ln \theta - E_0[\ln y_i]$.) Show that the likelihood inequality in Theorem 14.3 holds for the normal distribution.
8. For random sampling from the classical regression model in (14-3), reparameterize the likelihood function in terms of $\eta = 1/\sigma$ and $\boldsymbol{\delta} = (1/\sigma)\boldsymbol{\beta}$. Find the maximum likelihood estimators of η and $\boldsymbol{\delta}$ and obtain the asymptotic covariance matrix of the estimators of these parameters.
9. Consider sampling from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_M)$ and covariance matrix $\sigma^2\mathbf{I}$. The log-likelihood function is

$$\ln L = \frac{-nM}{2}\ln(2\pi) - \frac{nM}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})'(\mathbf{y}_i - \boldsymbol{\mu}).$$

Show that the maximum likelihood estimators of the parameters are $\hat{\mu}_m = \bar{y}_m$, and

$$\hat{\sigma}^2_{\text{ML}} = \frac{\sum_{i=1}^{n}\sum_{m=1}^{M}(y_{im} - \bar{y}_m)^2}{nM} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n}(y_{im} - \bar{y}_m)^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}^2_m.$$

Derive the second derivatives matrix and show that the asymptotic covariance matrix for the maximum likelihood estimators is

$$\left\{-E\left[\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right\}^{-1} = \begin{bmatrix}\sigma^2\mathbf{I}/n & \mathbf{0} \\ \mathbf{0} & 2\sigma^4/(nM)\end{bmatrix}.$$

Suppose that we wished to test the hypothesis that the means of the M distributions were all equal to a particular value μ^0. Show that the Wald statistic would be

$$W = (\bar{\mathbf{y}} - \mu^0\mathbf{i})'\left(\frac{\hat{\sigma}^2}{n}\mathbf{I}\right)^{-1}(\bar{\mathbf{y}} - \mu^0\mathbf{i}) = \left(\frac{n}{s^2}\right)(\bar{\mathbf{y}} - \mu^0\mathbf{i})'(\bar{\mathbf{y}} - \mu^0\mathbf{i}),$$

where $\bar{\mathbf{y}}$ is the vector of sample means.

Applications

1. **Binary Choice.** This application will be based on the health care data analyzed in Example 14.13 and several others. Details on obtaining the data are given in Appendix F Table 7.1. We consider analysis of a dependent variable, y_{it}, that takes values 1 and 0 with probabilities $F(\mathbf{x}_i'\boldsymbol{\beta})$ and $1 - F(\mathbf{x}_i'\boldsymbol{\beta})$, where F is a function that defines a probability. The dependent variable, *yit*, is constructed from the count variable *DocVis*, which is the number of visits to the doctor in the given year. Construct the binary variable

$$y_{it} = 1 \text{ if } DocVis > 0, 0 \text{ otherwise.}$$

We will build a model for the probability that y_{it} equals one. The independent variables of interest will be

$$\mathbf{x}_{it} = (1, age_{it}, educ_{it}, female_t, married_{it}, hsat_{it}).$$

a. According to the model, the theoretical density for y_{it} is

$$f(y_{it}|\mathbf{x}_{it}) = F(\mathbf{x}_{it}'\boldsymbol{\beta}) \text{ for } y_{it} = 1 \text{ and } 1 - F(\mathbf{x}_{it}'\boldsymbol{\beta}) \text{ for } y_{it} = 0.$$

We will assume that a "logit model" (see Section 17.2) is appropriate, so that

$$F(\mathbf{x}_{it}'\boldsymbol{\beta}) = \Lambda(\mathbf{x}_{it}'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta})}{1 - \exp(\mathbf{x}_{it}'\boldsymbol{\beta})}.$$

Show that for the two outcomes, the probabilities may be combined into the density function

$$f(y_{it}|\mathbf{x}_{it}) = g(y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}) = \Lambda[(2y_{it} - 1)\mathbf{x}_{it}'\boldsymbol{\beta}].$$

Now, use this result to construct the log-likelihood function for a sample of data on $(y_{it}, \mathbf{x}_{it})$. (*Note:* We will be ignoring the panel aspect of the data set. Build the model as if this were a cross section.)

b. Derive the likelihood equations for estimation of $\boldsymbol{\beta}$.

c. Derive the second derivatives matrix of the log-likelihood function. (*Hint:* The following will prove useful in the derivation: $d\Lambda(t)/dt = \Lambda(t)[1 - \Lambda(t)]$.)

d. Show how to use Newton's method to estimate the parameters of the model.

e. Does the method of scoring differ from Newton's method? Derive the negative of the expectation of the second derivatives matrix.

f. Obtain maximum likelihood estimates of the parameters for the data and variables noted. Report your results, estimates, standard errors, and so on, as well as the value of the log likelihood.

g. Test the hypothesis that the coefficients on female and marital status are zero. Show how to do the test using Wald, LM, and LR tests, and then carry out the tests.

h. Test the hypothesis that all the coefficients in the model save for the constant term are equal to zero.

2. The geometric distribution used in Examples 14.13, 14.17, 14.18, and 14.22 would not be the typical choice for modeling a count such as DocVis. The Poisson model suggested at the beginning of Section 14.11.1 would be the more natural choice (at least at the first step in an analysis). Redo the calculations in Exercises 14.13 and 14.17 using a Poisson model rather than a geometric model. Do the results change very much? It is difficult to tell from the coefficient estimates. Compute the partial effects for the Poisson model and compare them to the partial effects shown in Table 14.11.

3. (This application will require an optimizer. Maximization of a user-supplied function is provided by commands in *Stata*, *R*, *SAS*, *EViews* or *NLOGIT*.) Use the following pseudo-code to generate a random sample of 1,000 observations on y from a mixed normals population:

```
Set the seed of the random number generator at any specific value.
Generate two sets of 1,000 random draws from normal populations
  with standard deviations 1. For the means, use 1 for y1 and
  5 for y2.
Generate a set of 1,000 random draws, c, from uniform(0,1)
  population.
```

```
For each observation, if c < .3, y = y1; if c ≥ .3, use y = y2.
```

The log-likelihood function for the mixture of two normals is given in (14-89). (The first step sets the seed at a particular value so that you can replicate your calculation of the data sets.)

a. Find the values that maximize the log-likelihood function. As starting values, use the sample mean of y (the same value) and sample standard deviation of y (again, same value) and 0.5 for π.
b. You should have observed the iterations in part a never get started. Try again using $0.9\bar{y}$, $.9s_y$, $1.1\bar{y}$, $1.1s_y$, and 0.5. This should be much more satisfactory.
c. Experiment with the estimator by generating y_1 and y_2 with more similar means, such as 1 and 3, or 1 and 2.

15

SIMULATION-BASED ESTIMATION AND INFERENCE AND RANDOM PARAMETER MODELS

15.1 INTRODUCTION

Simulation-based methods have become increasingly popular in econometrics. They are extremely computer intensive, but steady improvements in recent years in computation hardware and software have reduced that cost enormously. The payoff has been in the form of methods for solving estimation and inference problems that have previously been unsolvable in analytic form. The methods are used for two main functions. First, **simulation**-based methods are used to infer the characteristics of random variables, including estimators, functions of estimators, test statistics, and so on, by sampling from their distributions. Second, simulation is used in constructing estimators that involve complicated integrals that do not exist in a closed form that can be evaluated. In such cases, when the integral can be written in the form of an expectation, simulation methods can be used to evaluate it to within acceptable degrees of approximation by estimating the expectation as the mean of a random sample. The technique of maximum simulated likelihood (MSL) is essentially a classical sampling theory counterpart to the hierarchical Bayesian estimator considered in Chapter 16. Since the celebrated paper of Berry, Levinsohn, and Pakes (1995), and the review by McFadden and Train (2000), maximum simulated likelihood estimation has been used in a large and growing number of studies.

The following are three examples from earlier chapters that have relied on simulation methods.

Example 15.1 Inferring the Sampling Distribution of the Least Squares Estimator

In Example 4.1, we demonstrated the idea of a sampling distribution by drawing several thousand samples from a population and computing a least squares coefficient with each sample. We then examined the distribution of the sample of linear regression coefficients. A histogram suggested that the distribution appeared to be normal and centered over the true population value of the coefficient.

Example 15.2 Bootstrapping the Variance of the LAD Estimator

In Example 4.3, we compared the asymptotic variance of the least absolute deviations (LAD) estimator to that of the ordinary least squares (OLS) estimator. The form of the asymptotic variance of the LAD estimator is not known except in the special case of normally distributed disturbances. We relied, instead, on a random sampling method to approximate features of the sampling distribution of the LAD estimator. We used a device (bootstrapping) that allowed us to draw a sample of observations from the population that produces the estimator. With that random sample, by computing the corresponding sample statistics, we can infer characteristics of the distribution such as its variance and its 2.5th and 97.5th percentiles, which can be used to construct a confidence interval.

Example 15.3 Least Simulated Sum of Squares

Familiar estimation and inference methods, such as least squares and maximum likelihood, rely on closed form expressions that can be evaluated exactly [at least in principle—likelihood equations such as (14-4) may require an iterative solution]. Model building and analysis often require evaluation of expressions that cannot be computed directly. Familiar examples include expectations that involve integrals with no closed form such as the random effects nonlinear regression model presented in Section 14.14.4. The estimation problem posed there involved nonlinear least squares estimation of the parameters of

$$E[y_{it} | \mathbf{x}_{it}, u_i] = h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i).$$

Minimizing the sum of squares,

$$S(\boldsymbol{\beta}) = \sum_i \sum_t [y_{it} - h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)]^2,$$

is not feasible because u_i is not observed. In this formulation,

$$E[y | \mathbf{x}_{it}] = E_u E[y_{it} | \mathbf{x}_{it}, u_i] = \int_u E[y_{it} | \mathbf{x}_{it}, u_i] f(u_i) \mathrm{d}u_i,$$

so the feasible estimation problem would involve the sum of squares,

$$S^*(\boldsymbol{\beta}) = \sum_i \sum_t \left[y_{it} - \int_u h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) f(u_i) \mathrm{d}u_i \right]^2.$$

When the function is linear and u_i is normally distributed, this is a simple problem—it reduces to ordinary least squares. If either condition is not met, then the integral generally remains in the estimation problem. Although the integral,

$$E_u[h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)] = \int_u h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) f(u_i) \mathrm{d}u_i,$$

cannot be computed, if a large sample of R observations from the population of u_i, that is, u_{ir}, $r = 1, \ldots R$, were observed, then by virtue of the law of large numbers, we could rely on

$$\begin{aligned} \text{plim}(1/R) \sum_r h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_{ir}) &= E_u E[y_{it} | \mathbf{x}_{it}, u_i] \\ &= \int_u h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) f(u_i) \mathrm{d}u_i. \end{aligned} \quad \textbf{(15-1)}$$

We are suppressing the extra parameter, σ_u, which would become part of the estimation problem. A convenient way to formulate the problem is to write $u_i = \sigma_u v_i$ where v_i has zero mean and variance one. By using this device, integrals can be replaced with sums that are feasible to compute. Our "simulated sum of squares" becomes

$$S_{\text{simulated}}(\boldsymbol{\beta}) = \sum_i \sum_t \left[y_{it} - (1/R) \sum_r h(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u v_{ir}) \right]^2, \quad \textbf{(15-2)}$$

which can be minimized by conventional methods. As long as (15-1) holds, then

$$\frac{1}{nT} \sum_i \sum_t \left[y_{it} - (1/R) \sum_r h(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u v_{ir}) \right]^2 \xrightarrow{p} \frac{1}{nT} \sum_i \sum_t \left[y_{it} - \int_v h(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u v_i) f(v_i) \mathrm{d}v_i \right]^2 \quad \textbf{(15-3)}$$

and it follows that with sufficiently increasing R, the $\boldsymbol{\beta}$ that minimizes the left-hand side converges (in nT) to the same parameter vector that minimizes the probability limit of the right-hand side. We are thus able to substitute a computer simulation for the intractable computation on the right-hand side of the expression.

This chapter will describe some of the common applications of simulation methods in econometrics. We begin in Section 15.2 with the essential tool at the heart of all the computations, random number generation. Section 15.3 describes simulation-based inference using the method of Krinsky and Robb as an alternative to the delta method (see Section 4.4.4). The method of bootstrapping for inferring the features of the distribution of an estimator is described in Section 15.4. In Section 15.5, we will use a Monte Carlo study to learn about the behavior of a test statistic and the behavior of the fixed effects estimator in some nonlinear models. Sections 15.6 through 15.9 present simulation-based estimation methods. The essential ingredient of this entire set of results is the computation of integrals. Section 15.6.1 describes an application of a simulation-based estimator, a nonlinear random effects model. Section 15.6.2 discusses methods of integration. Then, the methods are applied to the estimation of the random effects model. Sections 15.7 through 15.9 describe several techniques and applications, including maximum simulated likelihood estimation for random parameter and hierarchical models. A third major (perhaps *the* major) application of simulation-based estimation in the current literature is Bayesian analysis using Markov Chain Monte Carlo (MCMC or MC^2) methods. Bayesian methods are discussed separately in Chapter 16. Sections 15.10 and 15.11 consider two remaining aspects of modeling parameter heterogeneity, estimation of individual specific parameters, and a comparison of modeling with continuous distributions to less parametric modeling with discrete distributions using latent class models.

15.2 RANDOM NUMBER GENERATION

All the techniques we will consider here rely on samples of observations from an underlying population. We will sometimes call these *random samples*, though it will emerge shortly that they are never actually random. One of the important aspects of this entire body of research is the need to be able to replicate one's computations. If the samples of draws used in any kind of simulation-based analysis were truly random, then that would be impossible. Although the samples we consider here will appear to be random, they are, in fact, deterministic—the samples can be replicated. For this reason, the sampling methods described in this section are more often labeled *pseudo–random number generators*. (This does raise an intriguing question: Is it possible to generate truly random draws from a population with a computer? The answer for practical purposes is no.) This section will begin with a description of some of the mechanical aspects of random number generation. We will then detail the methods of generating particular kinds of random samples.[1]

15.2.1 GENERATING PSEUDO-RANDOM NUMBERS

Data are generated internally in a computer using **pseudo–random number generators**. These computer programs generate sequences of values that appear to be strings of draws from a specified probability distribution. There are many types of random

[1]See Train (2009, Chapter 3) for extensive further discussion.

number generators, but most take advantage of the inherent inaccuracy of the digital representation of real numbers. The method of generation is usually by the following steps:

1. Set a **seed**.
2. Update the seed by $\text{seed}_j = \text{seed}_{j-1} \times s$ value.
3. $x_j = \text{seed}_j \times x$ value.
4. Transform x_j if necessary, and then move x_j to desired place in memory.
5. Return to step 2, or exit if no additional values are needed.

Random number generators produce sequences of values that resemble strings of random draws from the specified distribution. In fact, the sequence of values produced by the preceding method is not truly random at all; it is a deterministic **Markov chain** of values. The set of 32 bits in the random value only appear random when subjected to certain tests.[2] Because the series is, in fact, deterministic, at any point that this type of generator produces a value it has produced before, it must thereafter replicate the entire sequence. Because modern digital computers typically use 32-bit double words to represent numbers, it follows that the longest string of values that this kind of generator can produce is $2^{32} - 1$ (about 4.3 billion). This length is the **period** of a random number generator. (A generator with a shorter period than this would be inefficient, because it is possible to achieve this period with some fairly simple algorithms.) Some improvements in the periodicity of a generator can be achieved by the method of **shuffling**. By this method, a set of, say, 128 values is maintained in an array. The random draw is used to select one of these 128 positions from which the draw is taken and then the value in the array is replaced with a draw from the generator. The period of the generator can also be increased by combining several generators.[3] The most popular random number generator in current use is the **Mersenne Twister**,[4] which has a period of about $2^{20,000}$.

The deterministic nature of pseudo–random number generators is both a flaw and a virtue. Many Monte Carlo studies require billions of draws, so the finite period of any generator represents a nontrivial consideration. On the other hand, being able to reproduce a sequence of values just by resetting the seed to its initial value allows the researcher to replicate a study.[5] The seed itself can be a problem. It is known that certain seeds in particular generators will produce shorter series or series that do not pass randomness tests. For example, *congruential* generators of the sort just discussed should be started from odd seeds.

15.2.2 SAMPLING FROM A STANDARD UNIFORM POPULATION

The output of the generator described in Section 15.2.1 will be a pseudo-draw from the $U[0, 1]$ population. (In principle, the draw should be from the closed interval $[0, 1]$. However, the actual draw produced by the generator will be strictly between zero and one with probability just slightly below one. In the application described, the draw will be constructed from the sequence of 32 bits in a double word. All

[2]See Press et al. (1986).

[3]See L'Ecuyer (1998), Gentle (2002, 2003), and Greene (2007b).

[4]See Matsumoto, and Nishimura (1998).

[5]Readers of empirical studies are often interested in replicating the computations. In Monte Carlo studies, at least in principle, data can be replicated efficiently merely by providing the random number generator and the seed.

but two of the $2^{31} - 1$ strings of bits will produce a value in (0, 1). The practical result is consistent with the theoretical one, that the probabilities attached to the terminal points are zero also.) When sampling from a standard uniform, U [0, 1] population, the sequence is a kind of difference equation, because given the initial seed, x_j is ultimately a function of x_{j-1}. In most cases, the result at step 3 is a pseudo-draw from the continuous uniform distribution in the range zero to one, which can then be transformed to a draw from another distribution by using the fundamental probability transformation.

15.2.3 SAMPLING FROM CONTINUOUS DISTRIBUTIONS

One is usually interested in obtaining a sequence of draws, $x_1, \ldots, x_R$, from some particular population such as the normal with mean μ and variance σ^2. A sequence of draws from $U[0, 1]$, $u_1, \ldots, u_R$, produced by the random number generator, is an intermediate step. These will be transformed into draws from the desired population. A common approach is to use the **fundamental probability transformation**. For continuous distributions, this is done by treating the draw, $u_r = F_r$, as if F_r was $F(x_r)$, where $F(.)$ is the cdf of x. For example, if we desire draws from the exponential distribution with known θ, then $F(x) = 1 - \exp(-\theta x)$. The inverse transform is $x = (-1/\theta)\ln(1 - F)$. For example, for a draw of $u = 0.4$ with $\theta = 5$, the associated x would be $(-1/5)\ln(1 - 0.4) = 0.1022$. For the logistic population with cdf $F(x) = \Lambda(x) = \exp(x)/[1 + \exp(x)]$, the inverse transformation is $x = \ln[F/(1 - F)]$. There are many references, for example, Evans, Hastings, and Peacock (2010) and Gentle (2003), that contain tables of inverse transformations that can be used to construct random number generators.

One of the most common applications is the draws from the standard normal distribution. This is complicated because there is no closed form for $\Phi^{-1}(F)$. There are several ways to proceed. A well-known approximation to the inverse function is given in Abramovitz and Stegun (1971),

$$\Phi^{-1}(F) = x \approx T - \frac{c_0 + c_1 T + c_2 T^2}{1 + d_1 T + d_2 T^2 + d_3 T^3},$$

where $T = [\ln(1/H^2)]^{1/2}$ and $H = F$ if $F > 0.5$ and $1 - F$ otherwise. The sign is then reversed if $F < 0.5$. A second method is to transform the $U[0, 1]$ values directly to a standard normal value. The Box–Muller (1958) method is $z = (-2 \ln u_1)^{1/2} \cos(2\pi u_2)$, where u_1 and u_2 are two independent $U[0, 1]$ draws. A second $N[0, 1]$ draw can be obtained from the same two values by replacing cos with sin in the transformation. The Marsaglia–Bray (1964) generator is $z_i = x_i[-(2/v)\ln v]^{1/2}$, where $x_i = 2u_i - 1$, u_i is a random draw from $U[0, 1]$ and $v = u_1^2 + u_2^2$, $i = 1, 2$. The pair of draws is rejected and redrawn if $v \geq 1$.

Sequences of draws from the standard normal distribution can easily be transformed into draws from other distributions by making use of the results in Section B.4. For example, the square of a standard normal draw will be a draw from chi-squared [1], and the sum of K chi-squared [1] is chi-squared $[K]$. From this relationship, it is possible to produce samples from the chi-squared $[K]$, *t[n]*, and *F[K,n]* distributions.

A related problem is obtaining draws from the truncated normal distribution. The random variable with truncated normal distribution is obtained from one with a normal distribution by discarding the part of the range above a value U and below a value L. The density of the resulting random variable is that of a normal distribution restricted to the range $[L, U]$. The truncated normal density is

$$f(x|\, L \leq x \leq U) = \frac{f(x)}{\text{Prob}[L \leq x \leq U]} = \frac{(1/\sigma)\phi[(x-\mu)/\sigma]}{\Phi[(U-\mu)/\sigma] - \Phi[(L-\mu)/\sigma]},$$

where $\phi(t) = (2\pi)^{-1/2}\exp(-t^2/2)$ and $\Phi(t)$ is the cdf. An obviously inefficient (albeit effective) method of drawing values from the truncated normal $[\mu, \sigma^2]$ distribution in the range $[L,U]$ is simply to draw F from the $U[0, 1]$ distribution and transform it first to a standard normal variate as discussed previously and then to the $N[\mu, \sigma^2]$ variate by using $x = \mu + \sigma\Phi^{-1}(F)$. Finally, the value x is retained if it falls in the range $[L,U]$ and discarded otherwise. This rejection method will require, on average, $1/\{\Phi[(U-\mu)/\sigma] - \Phi[(L-\mu)/\sigma]\}$ draws per observation, which could be substantial. A direct transformation that requires only one draw is as follows: Let $P_j = \Phi[(j-\mu)/\sigma], j = L, U$. Then

$$x = \mu + \sigma\Phi^{-1}[P_L + F \times (P_U - P_L)]. \qquad \textbf{(15-4)}$$

15.2.4 SAMPLING FROM A MULTIVARIATE NORMAL POPULATION

Many applications, including the method of Krinsky and Robb in Section 15.3, involve draws from a multivariate normal distribution with specified mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. To sample from this K-variate distribution, we begin with a draw, $\mathbf{z}$, from the K-variate standard normal distribution. This is done by first computing K independent standard normal draws, $z_1, \ldots, z_K$, using the method of the previous section and stacking them in the vector $\mathbf{z}$. Let $\mathbf{C}$ be a square root of $\boldsymbol{\Sigma}$ such that $\mathbf{CC}' = \boldsymbol{\Sigma}$. The desired draw is then $\mathbf{x} = \boldsymbol{\mu} + \mathbf{Cz}$, which will have covariance matrix $E[(\mathbf{x} - \boldsymbol{\mu}), (\mathbf{x} - \boldsymbol{\mu})'] = \mathbf{C}E[\mathbf{zz}']\mathbf{C}' = \mathbf{CIC}' = \boldsymbol{\Sigma}$. For the square root matrix, the usual choice is the **Cholesky decomposition**, in which $\mathbf{C}$ is a lower triangular matrix. (See Section A.6.11.) For example, suppose we wish to sample from the bivariate normal distribution with mean vector $\boldsymbol{\mu}$, unit variances, and correlation coefficient ρ. Then,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}.$$

The transformation of two draws z_1 and z_2 is $x_1 = \mu_1 + z_1$ and $x_2 = \mu_2 + [\rho z_1 + (1-\rho^2)^{1/2} z_2]$. Section 15.3 and Example 15.4 following show a more involved application.

15.2.5 SAMPLING FROM DISCRETE POPULATIONS

There is generally no inverse transformation available for discrete distributions, such as the Poisson. An inefficient, though usually unavoidable, method for some distributions is to draw the F and then search sequentially for the smallest value that has cdf equal to or greater than F. For example, a generator for the Poisson distribution is constructed as follows. The pdf is $\text{Prob}[x = j] = p_j = \exp(-\mu)\mu^j/j!$ where μ is the mean of the random

variable. The generator will use the recursion $p_j = p_{j-1} \times \mu/j, j = 1, \ldots$ beginning with $p_0 = \exp(-\mu)$. An algorithm that requires only a single random draw is as follows:

Initialize $c = \exp(-\mu), p = c, x = 0$;
Draw F from $U[0, 1]$;
Deliver x; $*$ exit with draw x if $c > F$;
Iterate: set $x = x + 1, p = p \times \mu/x, c = c + p$;
Return to $*$.

This method is based explicitly on the pdf and cdf of the distribution. Other methods are suggested by Knuth (1997) and Press et al. (2007).

The most common application of random sampling from a discrete distribution is, fortunately, also the simplest. The method of bootstrapping, and countless other applications involve random samples of draws from the **discrete uniform distribution**, $\text{Prob}(x = j) = 1/n, j = 1, \ldots, n$. In the bootstrapping application, we are going to draw random samples of observations from the sequence of integers $1, \ldots, n$, where each value must be equally likely. In principle, the random draw could be obtained by partitioning the unit interval into n equal parts, $[0, a_1), [a_1, a_2), \ldots, [a_{n-2}, a_{n-1}), [a_{n-1}, 1]; a_j = j/n, j = 1, \ldots, n - 1$. Then, random draw F delivers $x = j$ if F falls into interval j. This would entail a search, which could be time consuming. However, a simple method that will be much faster is simply to deliver $x =$ the integer part of $(n \times F + 1.0)$. (Once again, we are making use of the practical result that F will equal exactly 1.0—and x will equal $n + 1$—with ignorable probability.)

15.3 SIMULATION-BASED STATISTICAL INFERENCE: THE METHOD OF KRINSKY AND ROBB

Most of the theoretical development in this text has concerned the statistical properties of estimators—that is, the characteristics of sampling distributions such as the mean (probability limits), variance (asymptotic variance), and quantiles (such as the boundaries for confidence intervals). In cases in which these properties cannot be derived explicitly, it is often possible to infer them by using random sampling methods to draw samples from the population that produced an estimator and deduce the characteristics from the features of such a random sample. In Example 4.4, we computed a set of least squares regression coefficients, $b_1, \ldots, b_K$, and then examined the behavior of a nonlinear function $c_k = b_k/(1 - b_m)$ using the delta method. In some cases, the asymptotic properties of nonlinear functions such as these are difficult to derive directly from the theoretical distribution of the parameters. The sampling methods described here can be used for that purpose. A second common application is learning about the behavior of test statistics. For example, in Sections 5.3.3 and 14.6.3 [see (14-53)], we defined a Lagrange multiplier statistic for testing the hypothesis that certain coefficients are zero in a linear regression model. Under the assumption that the disturbances are normally distributed, the statistic has a limiting chi-squared distribution, which implies that the analyst knows what critical value to employ if he uses this statistic. Whether the statistic has this distribution if the disturbances are not normally distributed is unknown. Monte Carlo methods can be helpful in determining if the guidance of the chi-squared result

is useful in more general cases. Finally, in Section 14.7, we defined a two-step maximum likelihood estimator. Computation of the asymptotic variance of such an estimator can be challenging. Monte Carlo methods, in particular, bootstrapping methods, can be used as an effective substitute for the intractable derivation of the appropriate asymptotic distribution of an estimator. This and the next two sections will detail these three procedures and develop applications to illustrate their use.

The method of Krinsky and Robb is suggested as a way to estimate the asymptotic covariance matrix of $\mathbf{c} = \mathbf{f}(\mathbf{b})$, where $\mathbf{b}$ is an estimated parameter vector with asymptotic covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{f}(\mathbf{b})$ defines a set of possibly nonlinear functions of $\mathbf{b}$. We assume that $\mathbf{f}(\mathbf{b})$ is a set of continuous and continuously differentiable functions that do not involve the sample size and whose derivatives do not equal zero at $\boldsymbol{\beta} = \text{plim}\,\mathbf{b}$. (These are the conditions underlying the Slutsky theorem in Section D.2.3.) In Section 4.6, we used the delta method to estimate the asymptotic covariance matrix of $\mathbf{c}$; $\text{Est.Asy.Var}[\mathbf{c}] = \mathbf{GSG}'$, where $\mathbf{S}$ is the estimate of $\boldsymbol{\Sigma}$ and $\mathbf{G}$ is the matrix of partial derivatives, $\mathbf{G} = \partial\mathbf{f}(\mathbf{b})/\partial\mathbf{b}'$. The recent literature contains some occasional skepticism about the accuracy of the delta method. The method of Krinsky and Robb (1986, 1990, 1991) is often suggested as an alternative. In a study of the behavior of estimated elasticities based on a translog model, the authors (1986) advocated an alternative approach based on Monte Carlo methods and the law of large numbers. We have consistently estimated $\boldsymbol{\beta}$ and $(\sigma^2/n)\mathbf{Q}^{-1}$, the mean and variance of the asymptotic normal distribution of the estimator $\mathbf{b}$, with $\mathbf{b}$ and $s^2(\mathbf{X}'\mathbf{X})^{-1}$. It follows that we could estimate the mean and variance of the distribution of a function of $\mathbf{b}$ by drawing a random sample of observations from the asymptotic normal population generating $\mathbf{b}$, and using the empirical mean and variance of the sample of functions to estimate the parameters of the distribution of the function. The quantiles of the sample of draws, for example, the 0.025th and 0.975th quantiles, can be used to estimate the boundaries of a confidence interval of the functions. The multivariate normal sample would be drawn using the method described in Section 15.2.4.

Krinsky and Robb (1986) reported huge differences in the standard errors produced by the delta method compared to the simulation-based estimator. In a subsequent paper (1990), they reported that the entire difference could be attributed to a bug in the software they used—upon redoing the computations, their estimates were essentially the same with the two methods. It is difficult to draw a conclusion about the effectiveness of the delta method based on the received results—it does seem at this juncture that the delta method remains an effective device that can often be employed with a hand calculator as opposed to the much more computation-intensive Krinsky and Robb (1986) technique. Unfortunately, the results of any comparison will depend on the data, the model, and the functions being computed. The amount of nonlinearity in the sense of the complexity of the functions seems not to be the answer. Krinsky and Robb's case was motivated by the extreme complexity of the elasticities in a translog model. In another study, Hole (2006) examines a similarly complex problem and finds that the delta method still appears to be the more accurate procedure.

Example 15.4 Long-Run Elasticities

A dynamic version of the demand for gasoline model is estimated in Example 4.7. The model is

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(Income/Pop)_t + \beta_4 \ln P_{nc,t} + \beta_5 \ln P_{uc,t} + \gamma \ln (G/Pop)_{t-1} + \varepsilon_t.$$

In this model, the short-run price and income elasticities are β_2 and β_3. The long-run elasticities are $\phi_2 = \beta_2/(1 - \gamma)$ and $\phi_3 = \beta_3/(1 - \gamma)$, respectively. To estimate the long-run elasticities, we estimated the parameters by least squares and then computed these two nonlinear functions of the estimates. Estimates of the full set of model parameters and the estimated asymptotic covariance matrix are given in Example 4.7. The delta method was used to estimate the asymptotic standard errors for the estimates of ϕ_2 and ϕ_3. The three estimates of the specific parameters and the 3×3 submatrix of the estimated asymptotic covariance matrix are

$$Est.\begin{pmatrix}\beta_2\\ \beta_3\\ \gamma\end{pmatrix} = \begin{pmatrix}b_2\\ b_3\\ c\end{pmatrix} = \begin{pmatrix}-0.069532\\ 0.164047\\ 0.830971\end{pmatrix},$$

$$Est.\ Asy.\ Var\begin{pmatrix}b_2\\ b_3\\ c\end{pmatrix} = \begin{pmatrix}0.00021705 & 1.61265e{-}5 & -0.0001109\\ 1.61265e{-}5 & 0.0030279 & -0.0021881\\ -0.0001109 & -0.0021881 & 0.0020943\end{pmatrix}.$$

The method suggested by Krinsky and Robb would use a random number generator to draw a large trivariate sample, $(b_2, b_3, c)_r$, $r = 1, \ldots, R$, from the normal distribution with this mean vector and covariance matrix, and then compute the sample of observations on f_2 and f_3 and obtain the empirical mean and variance and the 0.025 and 0.975 quantiles from the sample. The method of drawing such a sample is shown in Section 15.2.4. We will require the square root of the covariance matrix. The Cholesky matrix is

$$\mathbf{C} = \begin{pmatrix}0.0147326 & 0 & 0\\ 0.00109461 & 0.0550155 & 0\\ -0.0075275 & -0.0396227 & 0.0216259\end{pmatrix}$$

The sample is drawn by obtaining vectors of three random draws from the standard normal population, $\mathbf{v}_r = (v_1, v_2, v_3)'_r$, $r = 1, \ldots, R$. The draws needed for the estimation are then obtained by computing $\mathbf{b}_r = \mathbf{b} + \mathbf{C}\mathbf{v}_r$, where $\mathbf{b}$ is the set of least squares estimates. We then compute the sample of estimated long-run elasticities, $f_{2r} = b_{2r}/(1 - c_r)$ and $f_{3r} = b_{3r}/(1 - c_r)$. The mean and standard deviation of the sample observations constitute the estimates of the functions and asymptotic standard errors.

Table 15.1 shows the results of these computations based on 1,000 draws from the underlying distribution. The estimates from Example 4.4 using the delta method are shown as well. The two sets of estimates are in quite reasonable agreement. For a 95% confidence interval for ϕ_2 based on the estimates, the t distribution with $51 - 6 = 45$ degrees of freedom and the delta method would be $-0.411358 \pm 2.014(0.152296)$. The result for ϕ_3 would be $0.970522 \pm 2.014(0.162386)$. These are shown in Table 15.2 with the same computation

TABLE 15.1 Simulation Results

	Regression Estimate		*Simulated Values*	
	Estimate	***Std.Err.***	***Mean***	***Std.Dev.***
β_2	−0.069532	0.0147327	−0.068791	0.0138485
β_3	0.164047	0.0550265	0.162634	0.0558856
γ	0.830971	0.0457635	0.831083	0.0460514
ϕ_2	−0.411358	0.152296	−0.453815	0.219110
ϕ_3	0.970522	0.162386	0.950042	0.199458

TABLE 15.2 Estimated Confidence Intervals

	ϕ_2		ϕ_3	
	Lower	***Upper***	***Lower***	***Upper***
Delta Method	−0.718098	−0.104618	0.643460	1.297585
Krinsky and Robb	−0.895125	−0.012505	0.548313	1.351772
Sample Quantiles	−0.983866	−0.209776	0.539668	1.321617

using the Krinsky and Robb estimated standard errors. The table also shows the empirical estimates of these quantiles computed using the 26th and 975th values in the samples. There is reasonable agreement in the estimates, though a considerable amount of sample variability is also evident, even in a sample as large as 1,000.

We note, finally, that it is generally not possible to replicate results such as these across software platforms because they use different random number generators. Within a given platform, replicability can be obtained by setting the seed for the random number generator.

15.4 BOOTSTRAPPING STANDARD ERRORS AND CONFIDENCE INTERVALS

The technique of bootstrapping is used to obtain a description of the sampling properties of empirical estimators using the sample data themselves, rather than broad theoretical results.[6] Suppose that $\hat{\boldsymbol{\theta}}_n$ is an estimator of a parameter vector $\boldsymbol{\theta}$ based on a sample, $\mathbf{Z} = [(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)]$. An approximation to the statistical properties of $\hat{\boldsymbol{\theta}}_n$ can be obtained by studying a sample of bootstrap estimators $\hat{\boldsymbol{\theta}}(b)_m$, $b = 1, \ldots, B$, obtained by sampling m observations, with replacement, from $\mathbf{Z}$ and recomputing $\hat{\boldsymbol{\theta}}$ with each sample. After a total of B times, the desired sampling characteristic is computed from

$$\hat{\boldsymbol{\Theta}} = [\hat{\boldsymbol{\theta}}(1)_m, \hat{\boldsymbol{\theta}}(2)_m, \ldots, \hat{\boldsymbol{\theta}}(B)_m].$$

The most common application of bootstrapping for consistent estimators when n is reasonably large is approximating the asymptotic covariance matrix of the estimator $\hat{\boldsymbol{\theta}}_n$ with

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}_n] = \frac{1}{B-1}\sum_{b=1}^{B}[\hat{\boldsymbol{\theta}}(b)_m - \bar{\hat{\boldsymbol{\theta}}}_B][\hat{\boldsymbol{\theta}}(b)_m - \bar{\hat{\boldsymbol{\theta}}}_B]', \tag{15-5}$$

where $\bar{\hat{\boldsymbol{\theta}}}_B$ is the average of the B bootstrapped estimates of $\boldsymbol{\theta}$. There are few theoretical prescriptions for the number of replications, B. Andrews and Buchinsky (2000) and Cameron and Trivedi (2005, pp. 361–362) make some suggestions for particular applications; Davidson and MacKinnon (2006) recommend at least 399. Several hundred is the norm; we have used 1,000 in our application to follow.[7] An application to the least absolute deviations estimator in the linear model is shown in the following example and in Chapter 4.

[6] See Efron (1979), Efron and Tibshirani (1994), and Davidson and Hinkley (1997), Brownstone and Kazimi (1998), Horowitz (2001), MacKinnon (2002), and Davidson and MacKinnon (2006).

[7] For applications, see, for example, Veall (1987, 1992), Vinod (1993), and Vinod and Raj (1994). Extensive surveys of uses and methods in econometrics appear in Cameron and Trivedi (2005), Horowitz (2001), and Davidson and MacKinnon (2006).

15.4.1 TYPES OF BOOTSTRAPS

The preceding is known as a **paired bootstrap**. The pairing is the joint sampling of y_i and $\mathbf{x}_i$. An alternative approach in a regression context would be to sample the observations on $\mathbf{x}_i$ once and then with each $\mathbf{x}_i$ sampled, generate the accompanying y_i by randomly generating the disturbance, then $\hat{y}_i(b) = \mathbf{x}_i(b)'\hat{\boldsymbol{\theta}}_n + \hat{\varepsilon}_i(b)$. This would be a **parametric bootstrap** in that in order to simulate the disturbances, we need either to know (or assume) the data-generating process that produces ε_i. In other contexts, such as in discrete choice modeling in Chapter 17, one would bootstrap sample the exogenous data in the model and then generate the dependent variable by this method using the appropriate underlying DGP. This is the approach used in 15.5.2 and in Greene (2004b) in a study of the incidental parameters problem in several limited dependent variable models. The obvious disadvantage of the parametric bootstrap is that one cannot learn of the influence of an unknown DGP for ε by assuming it is known. For example, if the bootstrap is being used to accommodate unknown heteroscedasticity in the model, then a parametric bootstrap that assumes homoscedasticity would defeat the purpose. The more natural application would be a **nonparametric bootstrap**, in which both $\mathbf{x}_i$ and y_i, and, implicitly, ε_i, are sampled simultaneously.

Example 15.5 Bootstrapping the Variance of the Median

There are few cases in which an exact expression for the sampling variance of the median is known. Example 15.7 examines the case of the median of a sample of 500 observations from the t distribution with 10 degrees of freedom. This is one of those cases in which there is no exact formula for the asymptotic variance of the median. However, we can use the bootstrap technique to estimate one empirically. In one run of the experiment, we obtained a sample of 500 observations for which we computed the median, −0.00786. We drew 100 samples of 500 with replacement from this sample of 500 and recomputed the median with each of these samples. The empirical square root of the mean squared deviation around this estimate of −0.00786 was 0.056. In contrast, consider the same calculation for the mean. The sample mean is −0.07247. The sample standard deviation is 1.08469, so the standard error of the mean is 0.04657. (The bootstrap estimate of the standard error of the mean was 0.052.) This agrees with our expectation in that the sample mean should generally be a more efficient estimator of the mean of the distribution in a large sample. There is another approach we might take in this situation. Consider the regression model $y_i = \alpha + \varepsilon_i$, where ε_i has a symmetric distribution with finite variance. The least absolute deviations estimator of the coefficient in this model is an estimator of the median (which equals the mean) of the distribution. So, this presents another estimator. Once again, the bootstrap estimator must be used to estimate the asymptotic variance of the estimator. Using the same data, we fit this regression model using the LAD estimator. The coefficient estimate is −0.05397 with a bootstrap estimated standard error of 0.05872. The estimated standard error agrees with the earlier one. The difference in the estimated coefficient stems from the different computations—the regression estimate is the solution to a linear programming problem while the earlier estimate is the actual sample median.

15.4.2 BIAS REDUCTION WITH BOOTSTRAP ESTIMATORS

The bootstrap estimation procedure has also been suggested as a method of reducing bias. In principle, we would compute $\hat{\boldsymbol{\theta}}_n - \text{bias}(\hat{\boldsymbol{\theta}}_n) = \hat{\boldsymbol{\theta}}_n - \{E[\hat{\boldsymbol{\theta}}_n] - \boldsymbol{\theta}\}$. Because neither $\boldsymbol{\theta}$ nor the exact expectation of $\hat{\boldsymbol{\theta}}_n$ is known, we estimate the first with the mean of the bootstrap replications and the second with the estimator itself. The revised estimator is

$$\hat{\boldsymbol{\theta}}_{n,B} = \hat{\boldsymbol{\theta}}_n - \left[\frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\theta}}(b)_m - \hat{\boldsymbol{\theta}}_n\right] = 2\hat{\boldsymbol{\theta}}_n - \bar{\hat{\boldsymbol{\theta}}}_B. \tag{15-6}$$

[Efron and Tibshirani (1994, p. 138) provide justification for what appears to be the wrong sign on the correction.] Davidson and MacKinnon (2006) argue that the smaller bias of the corrected estimator is offset by an increased variance compared to the uncorrected estimator.[8] The authors offer some other cautions for practitioners contemplating use of this technique. First, perhaps obviously, the extension of the method to samples with dependent observations presents some obstacles. For time-series data, the technique makes little sense—none of the bootstrapped samples will be a time series, so the properties of the resulting estimators will not satisfy the underlying assumptions needed to make the technique appropriate.

15.4.3 BOOTSTRAPPING CONFIDENCE INTERVALS

A second common application of bootstrapping methods is the computation of confidence intervals for parameters. This calculation will be useful when the underlying data-generating process is unknown, and the bootstrap method is being used to obtain appropriate standard errors for estimated parameters. A natural approach to bootstrapping confidence intervals for parameters would be to compute the estimated asymptotic covariance matrix using (15-5) and then form confidence intervals in the usual fashion. An improvement in terms of the bias of the estimator is provided by the **percentile method**.[9] By this technique, during each bootstrap replication, we compute

$$t_k^*(b) = \frac{\hat{\theta}_k(b) - \hat{\theta}_{n,k}}{se.(\hat{\theta}_{n,k})}, \tag{15-7}$$

where "k" indicates the kth parameter in the model, and $\hat{\theta}_{n,k}$, $s.e.(\hat{\theta}_{n,k})$ and $\hat{\theta}_k(b)$ are the original estimator and estimated standard error from the full sample and the bootstrap replicate. Then, with all B replicates in hand, the bootstrap confidence interval is

$$\hat{\theta}_{n,k} + t^*_{k[\alpha/2]} se.(\hat{\theta}_{n,k}) \text{ to } \hat{\theta}_{n,k} + t^*_{k[1-\alpha/2]} s.e.(\hat{\theta}_{n,k}). \tag{15-8}$$

(Note that $t^*_{k[\alpha/2]}$ is negative, which explains the plus sign in the left term.) For example, in our next application, next, we compute the estimator and the asymptotic covariance matrix using the full sample. We compute 1,000 bootstrap replications, and compute the t ratio in (15-7) for the education coefficient in each of the 1,000 replicates. After the bootstrap samples are accumulated, we sorted the results from (15-7), and the 25th and 975th largest values provide the values of t^*.

15.4.4 BOOTSTRAPPING WITH PANEL DATA: THE BLOCK BOOTSTRAP

Example 15.6 demonstrates the computation of a confidence interval for a coefficient using the bootstrap. The application uses the Cornwell and Rupert panel data set used in Example 11.4 and several later applications. There are 595 groups of seven observations in the data set. Bootstrapping with panel data requires an additional element in the computations. The bootstrap replications are based on sampling over i, not t. Thus, the bootstrap sample consists of n blocks of T (or T_i) observations—the ith group as a whole is sampled. This produces, then, a **block bootstrap** sample.

[8] See, as well, Cameron and Trivedi (2005).

[9] See Cameron and Trivedi (2005, p. 364).

Example 15.6 Block Bootstrapping Standard Errors and Confidence Intervals in a Panel

Example 11.4 presents least squares estimates and robust standard errors for the labor supply equation using Cornwell and Rupert's panel data set. There are 595 individuals and seven periods in the data set. As seen in the results in Table 11.1 (reproduced below), using a clustering correction in a robust covariance matrix for the least squares estimator produces substantial changes in the estimated standard errors. Table 15.3 reproduces the least squares coefficients and the standard errors associated with the conventional $s^2 (\mathbf{X}'\mathbf{X})^{-1}$ and the robust standard errors using the clustering correction in column (3). The block bootstrapped standard errors using 1,000 bootstrap replications are shown in column (4). The ability of the bootstrapping procedure to detect and mimic the effect of the clustering that is evident in columns (3) and (4). Note, as well, the resemblance to the naïve bootstrap estimates in column (5) and the conventional, uncorrected standard errors in column (2).

We also computed a confidence interval for the coefficient on *Ed* using the conventional, symmetric approach, $b_{Ed} \pm 1.96 s(b_{Ed})$, and the percentile method in (15-7) and (15-8). For the conventional estimator, we use $0.05670 \pm 1.96(0.00556) = [0.04580, 0.06760]$. For the bootstrap confidence interval method, we first computed and sorted the 1,000 t statistics based on (15-7). The 25th and 975th values were -2.148 and $+1.966$. The confidence interval is [0.04476, 0.06802].

Figure 15.1 shows a kernel density estimator of the distribution of the t statistics computed using (15-7) with the (approximate) standard normal density.

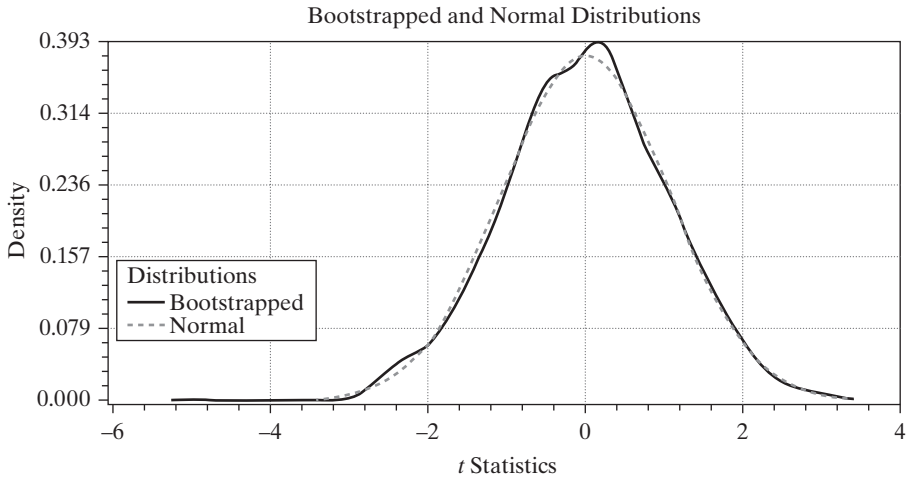
15.5 MONTE CARLO STUDIES

Simulated data generated by the methods of the preceding sections have various uses in econometrics. One of the more common applications is the analysis of the properties of estimators or in obtaining comparisons of the properties of estimators. For example,

TABLE 15.3 Bootstrap Estimates of Standard Errors for a Wage Equation

Variable	*(1) Least Squares Estimate*	*(2) Least Squares Standard Error*	*(3) Cluster Robust Standard Error*	*(4) Block Bootstrap Standard Error*	*(5) Simple Bootstrap Standard Error*
Constant	5.25112	0.07129	0.12355	0.12421	0.07761
Wks	0.00422	0.00108	0.00154	0.00159	0.00115
South	−0.05564	0.01253	0.02616	0.02557	0.01284
SMSA	0.15167	0.01207	0.02410	0.02383	0.01200
MS	0.04845	0.02057	0.04094	0.04208	0.02010
Exp	0.04010	0.00216	0.00408	0.00418	0.00213
Exp^2	−0.00067	0.00005	0.00009	0.00009	0.00005
Occ	−0.14001	0.01466	0.02724	0.02733	0.01539
Ind	0.04679	0.01179	0.02366	0.02350	0.01183
Union	0.09263	0.01280	0.02367	0.02390	0.01203
Ed	0.05670	0.00261	0.00556	0.00576	0.00273
Fem	−0.36779	0.02510	0.04557	0.04562	0.02390
Blk	−0.16694	0.02204	0.04433	0.04663	0.02103

FIGURE 15.1 Distributions of Test Statistics.



in time-series settings, most of the known results for characterizing the sampling distributions of estimators are asymptotic, large-sample results. But the typical time series is not very long, and descriptions that rely on T, the number of observations, going to infinity may not be very accurate. Exact finite-sample properties are usually intractable, however, which leaves the analyst with only the choice of learning about the behavior of the estimators experimentally.

In the typical application, one would either compare the properties of two or more estimators while holding the sampling conditions fixed or study how the properties of an estimator are affected by changing conditions such as the sample size or the value of an underlying parameter.

Example 15.7 Monte Carlo Study of the Mean Versus the Median

In Example D.8, we compared the asymptotic distributions of the sample mean and the sample median in random sampling from the normal distribution. The basic result is that both estimators are consistent, but the mean is asymptotically more efficient by a factor of

$$\frac{\text{Asy.Var[Median]}}{\text{Asy.Var[Mean]}} = \frac{\pi}{2} = 1.5708.$$

This result is useful, but it does not tell which is the better estimator in small samples, nor does it suggest how the estimators would behave in some other distribution. It is known that the mean is affected by outlying observations whereas the median is not. The effect is averaged out in large samples, but the small-sample behavior might be very different. To investigate the issue, we constructed the following experiment: We sampled 500 observations from the t distribution with d degrees of freedom by sampling $d + 1$ values from the standard normal distribution and then computing

$$t_{ir} = \frac{z_{ir,d+1}}{\sqrt{\frac{1}{d}\sum_{l=1}^{d} z_{ir,l}^2}}, \quad i = 1, \ldots, 500, \quad r = 1, \ldots, 100.$$

The t distribution with a low value of d was chosen because it has very thick tails and because large outlying values have high probability. For each value of d, we generated $R = 100$ replications. For each of the 100 replications, we obtained the mean and median. Because both are unbiased, we compared the mean squared errors around the true expectations using

$$M_d = \frac{(1/R)\sum_{r=1}^{R}(\text{median}_r - 0)^2}{(1/R)\sum_{r=1}^{R}(\bar{x}_r - 0)^2}.$$

We obtained ratios of 0.6761, 1.2779, and 1.3765 for $d = 3$, 6, and 10, respectively. (You might want to repeat this experiment with different degrees of freedom.) These results agree with what intuition would suggest. As the degrees of freedom parameter increases, which brings the distribution closer to the normal distribution, the sample mean becomes more efficient—the ratio should approach its limiting value of 1.5708 as d increases. What might be surprising is the apparent overwhelming advantage of the median when the distribution is very nonnormal even in a sample as large as 500.

The preceding is a very small application of the technique. In a typical study, there are many more parameters to be varied and more dimensions upon which the results are to be studied. One of the practical problems in this setting is how to organize the results. There is a tendency in Monte Carlo work to proliferate tables indiscriminately. It is incumbent on the analyst to collect the results in a fashion that is useful to the reader. For example, this requires some judgment on how finely one should vary the parameters of interest. One useful possibility that will often mimic the thought process of the reader is to collect the results of bivariate tables in carefully designed contour plots.

There are any number of situations in which Monte Carlo simulation offers the only method of learning about finite-sample properties of estimators. Still, there are a number of problems with Monte Carlo studies. To achieve any level of generality, the number of parameters that must be varied and hence the amount of information that must be distilled can become enormous. Second, they are limited by the design of the experiments, so the results they produce are rarely generalizable. For our example, we may have learned something about the t distribution, but the results that would apply in other distributions remain to be described. And, unfortunately, real data will rarely conform to any specific distribution, so no matter how many other distributions we analyze, our results would still only be suggestive. In more general terms, this problem of **specificity** [Hendry (1984)] limits most Monte Carlo studies to quite narrow ranges of applicability. There are very few that have proved general enough to have provided a widely cited result.

15.5.1 A MONTE CARLO STUDY: BEHAVIOR OF A TEST STATISTIC

Monte Carlo methods are often used to study the behavior of test statistics when their true properties are uncertain. This is often the case with Lagrange multiplier statistics. For example, Baltagi (2005) reports on the development of several new test statistics for panel data models such as a test for serial correlation. Examining the behavior of a test statistic is fairly straightforward. We are interested in two characteristics: the true **size of the test**—that is, the probability that it rejects the null hypothesis when that hypothesis is actually true (the probability of a type 1 error) and the **power of the test**—that is the probability that it will correctly reject a false null hypothesis (one minus the probability of a type 2 error). As we will see, the power of a test is a function of the alternative against which the null is tested.

To illustrate a Monte Carlo study of a test statistic, we consider how a familiar procedure behaves when the model assumptions are incorrect. Consider the linear regression model

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i, \quad \varepsilon_i | (x_i, z_i) \sim \mathrm{N}[0, \sigma^2].$$

The Lagrange multiplier statistic for testing the null hypothesis that γ equals zero for this model is

$$\mathrm{LM} = \mathbf{e}_0'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_0/(\mathbf{e}_0'\mathbf{e}_0/n),$$

where $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \mathbf{z})$ and $\mathbf{e}_0$ is the vector of least squares residuals obtained from the regression of $\mathbf{y}$ on the constant and $\mathbf{x}$ (and not $\mathbf{z}$). [See (14-53).] Under the assumptions of the preceding model, the large sample distribution of the LM statistic is chi squared with one degree of freedom. Thus, our testing procedure is to compute LM and then reject the null hypothesis $\gamma = 0$ if LM is greater than the critical value. We will use a nominal size of 0.05, so the critical value is 3.84. The theory for the statistic is well developed when the specification of the model is correct.[10] We are interested in two specification errors. First, how does the statistic behave if the normality assumption is not met? Because the LM statistic is based on the likelihood function, if some distribution other than the normal governs ε_i, then the LM statistic would not be based on the OLS estimator. We will examine the behavior of the statistic under the true specification that ε_i comes from a t distribution with five degrees of freedom. Second, how does the statistic behave if the homoscedasticity assumption is not met? The statistic is entirely wrong if the disturbances are heteroscedastic. We will examine the case in which the conditional variance is $\mathrm{Var}[\varepsilon_i | x_i, z_i] = \sigma^2[\exp(0.2x_i)]^2$.

The design of the experiment is as follows: We will base the analysis on a sample of 50 observations. We draw 50 observations on x_i and z_i from independent N[0, 1] populations at the outset of each cycle. For each of 1,000 replications, we draw a sample of 50 ε_i's according to the assumed specification. The LM statistic is computed and the proportion of the computed statistics that exceed 3.84 is recorded. The experiment is repeated for $\gamma = 0$ to ascertain the true size of the test and for values of γ including $-1, \ldots, -0.2, -0.1, 0, 0.1, 0.2, \ldots, 1.0$ to assess the power of the test. The cycle of tests is repeated for the two scenarios, the $t[5]$ distribution and the model with heteroscedasticity.

Table 15.4 lists the results of the experiment. The "Normal" column in each panel shows the expected results for the LM statistic under the model assumptions for which it is appropriate. The size of the test appears to be in line with the theoretical results. Comparing the first and third columns in each panel, it appears that the presence of heteroscedasticity seems not to degrade the power of the statistic. But the different distributional assumption does. Figure 15.2 plots the values in the table, and displays the characteristic form of the power function for a test statistic.

15.5.2 A MONTE CARLO STUDY: THE INCIDENTAL PARAMETERS PROBLEM

Section 14.14.5 examines the maximum likelihood estimator of a panel data model with fixed effects,

$$f(y_{it} | \mathbf{x}_{it}) = g(y_{it}, \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i, \boldsymbol{\theta}),$$

[10] See, for example, Godfrey (1988).

TABLE 15.4 Power Functions for LM Test

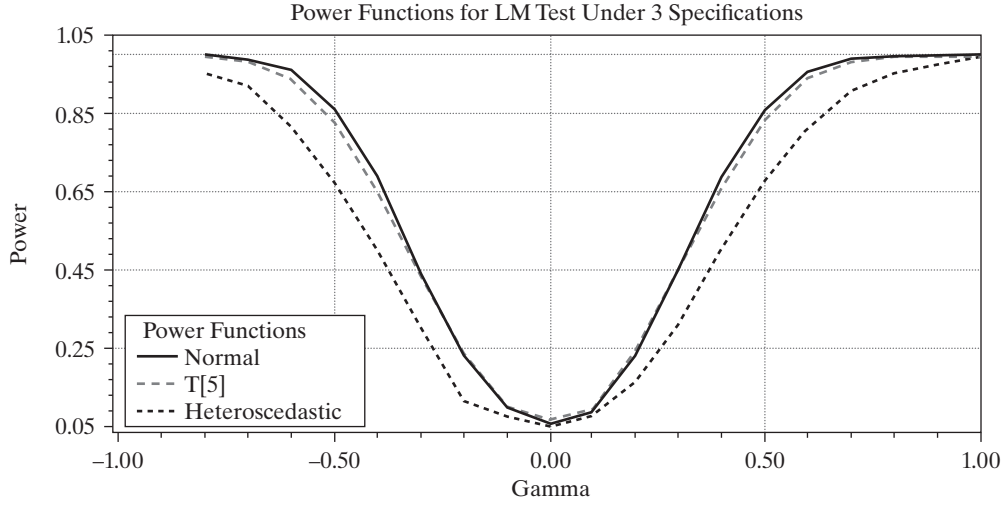
	Model				*Model*		
γ	***Normal***	***t[5]***	***Het.***	γ	***Normal***	***t[5]***	***Het.***
−1.0	1.000	0.993	1.000	0.1	0.090	0.083	0.098
−0.9	1.000	0.984	1.000	0.2	0.235	0.169	0.249
−0.8	0.999	0.953	0.996	0.3	0.464	0.320	0.457
−0.7	0.989	0.921	0.985	0.4	0.691	0.508	0.666
−0.6	0.961	0.822	0.940	0.5	0.859	0.680	0.835
−0.5	0.863	0.677	0.832	0.6	0.957	0.816	0.944
−0.4	0.686	0.500	0.651	0.7	0.989	0.911	0.984
−0.3	0.451	0.312	0.442	0.8	0.998	0.956	0.995
−0.2	0.236	0.177	0.239	0.9	1.000	0.976	0.998
−0.1	0.103	0.080	0.107	1.0	1.000	0.994	1.000
0.0	0.059	0.052	0.071				

where the individual effects may be correlated with x_{it}. The extra parameter vector $\boldsymbol{\theta}$ represents M other parameters that might appear in the model, such as the disturbance variance, σ_ε^2, in a linear regression model with normally distributed disturbance. The development there considers the mechanical problem of maximizing the log likelihood

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln g(y_{it}, \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i, \boldsymbol{\theta})$$

with respect to the $n + K + M$ parameters $(\alpha_1, \ldots, \alpha_n, \boldsymbol{\beta}, \boldsymbol{\theta})$. A statistical problem with this estimator that was suggested was that there is a phenomenon labeled the **incidental**

FIGURE 15.2 Power Functions.



parameters problem.[11] With the exception of a very small number of specific models (such as the Poisson regression model in Section 18.4.1), the *brute force*, unconditional maximum likelihood estimator of the parameters in this model is inconsistent. The result is straightforward to visualize with respect to the individual effects. Suppose that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ were actually known. Then, each α_i would be estimated with T_i observations. Because T_i is assumed to be fixed (and small), there is no asymptotic result to provide consistency for the MLE of α_i. But $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are estimated with $\Sigma_i\, T_i = N$ observations, so their large sample behavior is less transparent. One known result concerns the logit model for binary choice (see Sections 17.2–17.4). Kalbfleisch and Sprott (1970), Andersen (1973), Hsiao (1996), and Abrevaya (1997) have established that in the binary logit model, if $T_i = 2$, then plim $\hat{\boldsymbol{\beta}}_{\text{MLE}} = 2\boldsymbol{\beta}$. Two other cases are known with certainty. In the linear regression model with fixed effects and normally distributed disturbances, the slope estimator, $\mathbf{b}_{\text{LSDV}}$, is unbiased and consistent, however, the MLE of the variance, σ^2, converges to $(T - 1)\sigma^2/T$. (The degrees of freedom correction will adjust for this, but the MLE does not correct for degrees of freedom.) Finally, in the Poisson regression model (Section 18.4.7.b), the unconditional MLE is consistent.[12] Almost nothing else is known with certainty—that is, as a firm theoretical result—about the behavior of the maximum likelihood estimator in the presence of fixed effects. The literature appears to take as given the qualitative wisdom of Hsiao and Abrevaya, that the FE/MLE is inconsistent when T is small and fixed. (The implication that the severity of the inconsistency declines as T increases makes sense, but, again, remains to be shown analytically.)

The result for the two-period binary logit model is a standard result for discrete choice estimation. Several authors, all using Monte Carlo methods, have pursued the result for the logit model for larger values of T.[13] Greene (2004) analyzed the incidental parameters problem for other discrete choice models using Monte Carlo methods. We will examine part of that study.

The current studies are preceded by a small study in Heckman (1981) which examined the behavior of the fixed effects MLE in the following experiment:

$$
\begin{aligned}
&z_{it} = 0.1t + 0.5z_{i,t-1} + u_{it},\ z_{i0} = 5 + 10.0u_{i0},\\
&u_{it} \sim U[-0.5, 0.5],\ i = 1, \ldots, 100,\ t = 0, \ldots, 8,\\
&Y_{it} = \sigma_t\tau_i + \beta z_{it} + \varepsilon_{it},\ \tau_i \sim \text{N}[0, 1],\ \varepsilon_{it} \sim \text{N}[0, 1],\\
&y_{it} = 1 \text{ if } Y_{it} > 0,\ 0 \text{ otherwise.}
\end{aligned}
$$

Heckman attempted to learn something about the behavior of the MLE for the probit model with $T = 8$. He used values of $\beta = -1.0, -0.1$, and 1.0 and $\sigma_\tau = 0.5, 1.0$, and 3.0. The mean values of the maximum likelihood estimates of β for the nine cases are as follows:

	$\beta = -1.0$	$\beta = -0.1$	$\beta = 1.0$
$\sigma_\tau = 0.5$	−0.96	−0.10	0.93
$\sigma_\tau = 1.0$	−0.95	−0.09	0.91
$\sigma_\tau = 3.0$	−0.96	−0.10	0.90.

[11]See Neyman and Scott (1948), Lancaster (2000).

[12]See Cameron and Trivedi (1988).

[13]See, for example, Katz (2001).

The findings here disagree with the received wisdom. Where there appears to be a bias (i.e., excluding the center column), it seems to be quite small, and toward, not away from, zero.

The Heckman study used a very small sample and, moreover, analyzed the fixed effects estimator in a random effects model. (*Note:* τ_i is independent of z_{it}.) Greene (2004a), using the same parameter values, number of replications, and sample design, found persistent biases away from zero on the order of 15 to 20%. Numerous authors have extended the logit result for $T = 2$ with larger values of T, and likewise persistently found biases away from zero that diminish with increases in T. Greene (2004a) redid the experiment for the logit model and then replicated it for the probit and ordered probit models. The experiment is designed as follows: All models are based on the same index function

$$\begin{aligned} &w_{it} = \alpha_i + \beta x_{it} + \delta d_{it}, \quad \text{where } \beta = \delta = 1, \\ &x_{it} \sim \text{N}[0, 1], d_{it} = \mathbf{1}[x_{it} + h_{it} > 0], \qquad \text{where } h_{it} \sim \text{N}[0, 1], \\ &\alpha_i = \sqrt{T}\,\bar{x}_i + v_i, v_i \sim \text{N}[0, 1]. \end{aligned}$$

The regressors d_{it} and x_{it} are constructed to be correlated. The random term h_{it} is used to produce independent variation in d_{it}. There is, however, no within group correlation in x_{it} or d_{it} built into the data generator. (Other experiments suggested that the marginal distribution of x_{it} mattered little to the outcome of the experiment.) The correlations between the variables are approximately 0.7 between x_{it} and d_{it}, 0.4 between α_i and x_{it}, and 0.2 between α_i and d_{it}. The individual effect is produced from independent variation, v_i as well as the group mean of x_{it}. The latter is scaled by $\sqrt{T}$ to maintain the unit variances of the two parts—without the scaling, the covariance between α_i and x_{it} falls to zero as T increases and $\bar{x}_i$ converges to its mean of zero. Thus, the data generator for the index function satisfies the assumptions of the fixed effects model. The sample used for the results below contains $n = 1{,}000$ individuals. The data-generating processes for the discrete dependent variables are as follows:

$$\begin{aligned} &\textit{probit}: && y_{it} = \mathbf{1}[w_{it} + \varepsilon_{it} > 0], \varepsilon_{it} \sim \text{N}[0, 1], \\ &\textit{ordered probit}: && y_{it} = \mathbf{1}[w_{it} + \varepsilon_{it} > 0] + \mathbf{1}[w_{it} + \varepsilon_{it} > 3], \varepsilon_{it} \sim \mathbf{N}[0, 1], \\ &\textit{logit}: && y_{it} = \mathbf{1}[w_{it} + v_{it} > 0], v_{it} = \log[u_{it}/(1 - u_{it})], u_{it} \sim \text{U}[0, 1]. \end{aligned}$$

(The three discrete dependent variables are described in Chapters 17 and 18.)

Table 15.5 reports the results of computing the MLE with 200 replications. Models were fit with $T = 2, 3, 5, 8, 10$, and 20. (*Note:* This includes Heckman's experiment.) Each model specification and group size (T) is fit 200 times with random draws for ε_{it} or u_{it}. The data on the regressors were drawn at the beginning of each experiment (that is, for each T) and held constant for the replications. The table contains the average estimate of the coefficient and, for the binary choice models, the partial effects. The coefficients for the probit and logit models with $T = 2$ correspond to the received result, a 100% bias. The remaining values show, as intuition would suggest, that the bias decreases with increasing T. The benchmark case of $T = 8$ appears to be less benign than Heckman's results suggested. One encouraging finding for the model builder is that the biases in the estimated marginal effects appears to be somewhat less than for the coefficients. Greene (2004b) extends this analysis to some other models, including the tobit and truncated regression

TABLE 15.5 Means of Empirical Sampling Distributions, **N = 1,000** Individuals Based on 200 Replications

	Logit		*Probit*		*Ord. Probit*
Periods	***Coefficient***	***Partial Effect***[a]	***Coefficient***	***Partial Effect***[a]	***Coefficient***
$T = 2$					
β	2.020	1.676	2.083	1.474	2.328
δ	2.027	1.660	1.938	1.388	2.605
$T = 3$					
β	1.698	1.523	1.821	1.392	1.592
δ	1.668	1.477	1.777	1.354	1.806
$T = 5$					
β	1.379	1.319	1.589	1.406	1.305
δ	1.323	1.254	1.407	1.231	1.415
$T = 8$					
β	1.217	1.191	1.328	1.241	1.166
δ	1.156	1.128	1.243	1.152	1.220
$T = 10$					
β	1.161	1.140	1.247	1.190	1.131
δ	1.135	1.111	1.169	1.110	1.158
$T = 20$					
β	1.069	1.034	1.108	1.088	1.058
δ	1.062	1.052	1.068	1.047	1.068

[a]Average ratio of estimated partial effect to true partial effect.

models discussed in Chapter 19. The results there suggest that the conventional wisdom for the tobit model may not be correct—the incidental parameters (IP) problem seems to appear in the estimator of σ^2 in the tobit model, not in the estimators of the slopes.[14] This is consistent with the linear regression model, but not with the binary choice models.

15.6 SIMULATION-BASED ESTIMATION

Sections 15.3 through 15.5 developed a set of tools for inference about model parameters using simulation methods. This section will describe methods for using simulation as part of the estimation process. The modeling framework arises when integrals that cannot be computed directly appear in the estimation criterion function (sum of squares, log likelihood, and so on). To begin the development, in Section 15.6.1, we will construct a nonlinear model with random effects. Section 15.6.2 will describe how simulation is used to evaluate integrals for maximum likelihood estimation. Section 15.6.3 will develop an application, the random effects regression model.

[14]Research on the incidental parameters problem in discrete choice models, such as Fernandez-Val (2009), focuses on the slopes in the models. However, in all cases examined, the incidental parameters problem shows up as a proportional bias, which would seem to relate to an implicit scaling. The IP problem in the linear regression affects only the estimator of the disturbance variance.

15.6.1 RANDOM EFFECTS IN A NONLINEAR MODEL

In Example 11.20, we considered a nonlinear regression model for the number of doctor visits in the German Socioeconomic Panel. The basic form of the nonlinear regression model is

$$E[y_{it}|\mathbf{x}_{it}] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}), t = 1, \ldots, T_i, i = 1, \ldots, n.$$

In order to accommodate unobserved heterogeneity in the panel data, we extended the model to include a random effect,

$$E[y_{it}|\mathbf{x}_{it}, u_i] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i), \tag{15-9}$$

where u_i is an unobserved random effect with zero mean and constant variance, possibly normally distributed—we will turn to that shortly. We will now go a step further and specify a particular probability distribution for y_{it}. Because doctor visits is a count, the Poisson regression model would be a natural choice,

$$p(y_{it}|\mathbf{x}_{it}, u_i) = \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i). \tag{15-10}$$

Conditioned on $\mathbf{x}_{it}$ and u_i, the T_i observations for individual i are independent. That is, by conditioning on u_i, we treat them as data, the same as $\mathbf{x}_{it}$. Thus, the T_i observations are independent when they are conditioned on $\mathbf{x}_{it}$ and u_i. The joint density for the T_i observations for individual i is the product,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i, u_i) = \prod_{t=1}^{T_i}\frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i), t = 1, \ldots, T_i. \tag{15-11}$$

In principle at this point, the log-likelihood function to be maximized would be

$$\ln L = \sum_{i=1}^{n}\ln\left[\prod_{t=1}^{T_i}\frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}\right], \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i). \tag{15-12}$$

But it is not possible to maximize this log likelihood because the unobserved $u_i, i = 1, \ldots, n$, appears in it. The joint distribution of $(y_{i1}, y_{i2}, \ldots, y_{i,Ti}, u_i)$ is equal to the marginal distribution of u_i times the conditional distribution of $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,Ti})$ given u_i,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}, u_i|\mathbf{X}_i) = p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i, u_i)f(u_i),$$

where $f(u_i)$ is the marginal density for u_i. Now, we can obtain the marginal distribution of $(y_{i1}, y_{i2}, \ldots, y_{i,Ti})$ without u_i by

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i) = \int_{u_i} p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i, u_i)f(u_i)du_i.$$

For the specific application, with the Poisson conditional distributions for $y_{it}|u_i$ and a normal distribution for the random effect,

$$p(y_{i1}, y_{i2}, \ldots, y_{i,T_i}|\mathbf{X}_i) = \int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i}\frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}\right]\frac{1}{\sigma}\phi\left(\frac{u_i}{\sigma}\right)du_i, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i).$$

The log-likelihood function will now be

$$\ln L = \sum_{i=1}^{n} \ln \left\{ \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!} \right] \frac{1}{\sigma} \phi\left(\frac{u_i}{\sigma}\right) du_i \right\}, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i). \quad \textbf{(15-13)}$$

The optimization problem is now free of the unobserved u_i, but that complication has been traded for another one, the integral that remains in the function.

To complete this part of the derivation, we will simplify the log-likelihood function slightly in a way that will make it fit more naturally into the derivations to follow. Make the change of variable $u_i = \sigma w_i$, where w_i has mean zero and standard deviation one. Then, the Jacobian is $du_i = \sigma dw_i$, and the limits of integration for w_i are the same as for u_i. Making the substitution and multiplying by the Jacobian, the log-likelihood function becomes

$$\ln L = \sum_{i=1}^{n} \ln \left\{ \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!} \right] \phi(w_i) dw_i \right\}, \mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i). \quad \textbf{(15-14)}$$

The log likelihood is then maximized over $(\boldsymbol{\beta}, \sigma)$. The purpose of the simplification is to parameterize the model so that the distribution of the variable that is being integrated out has no parameters of its own. Thus, in (15-14), w_i is normally distributed with mean zero and variance one.

In the next section, we will turn to how to compute the integrals. Section 14.14.4 analyzes this model and suggests the **Gauss–Hermite quadrature** method for computing the integrals. In this section, we will derive a method based on simulation, **Monte Carlo integration**.[15]

15.6.2 MONTE CARLO INTEGRATION

Integrals often appear in econometric estimators in *open form*, that is, in a form for which there is no specific closed form function that is equivalent to them. For example, the integral, $\int_0^t \theta \exp(-\theta w) dw = 1 - \exp(-\theta t)$, is in closed form. The integral in (15-14) is in open form. There are various devices available for approximating open form integrals—Gauss–Hermite and Gauss–Laguerre quadrature noted in Section 14.14.4 and in Appendix E2.4 are two. The technique of Monte Carlo integration can often be used when the integral is in the form

$$h(y) = \int_w g(y|w) f(w) dw = E_w[g(y|w)],$$

where $f(w)$ is the density of w and and w is a random variable that can be simulated.[16]

If $w_1, w_2, \ldots, w_n$ are a random sample of observations on the random variable w and $g(w)$ is a function of w with finite mean and variance, then by the law of large numbers [Theorem D.4 and the corollary in (D-5)],

[15]The term *Monte Carlo* is in reference to the casino at Monte Carlo, where random number generation is a crucial element of the business.

[16]There are some necessary conditions on w and $g(y|w)$ that will be met in the applications that interest us here. Some details appear in Cameron and Trivedi (2005) and Train (2009).

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n} g(w_i) = E[g(w)].$$

The function in (15-14) is in this form,

$$\int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}]}{y_{it}!}\right]\phi(w_i)dw_i$$
$$= E_{w_i}[g(y_{i1}, y_{i2}, \ldots, y_{iT_i}|\, w_i, \mathbf{X}_i, \boldsymbol{\beta}, \sigma)],$$

where

$$g(y_{i1}, y_{i2}, \ldots, y_{iT_i}|\, w_i, \mathbf{X}_i, \boldsymbol{\beta}, \sigma) = \prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}]}{y_{it}!}$$

and w_i is a random variable with standard normal distribution. It follows, then, that

$$\text{plim}\ \frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}]}{y_{it}!}$$
$$= \int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}]}{y_{it}!}\right]\phi(w_i)dw_i. \quad \textbf{(5-15)}$$

This suggests the strategy for computing the integral. We can use the methods developed in Section 15.2 to produce the necessary set of random draws on w_i from the standard normal distribution and then compute the approximation to the integral according to (15-15).

Example 15.8 Fractional Moments of the Truncated Normal Distribution

The following function appeared in Greene's (1990) study of the stochastic frontier model:

$$h(M, \varepsilon) = \frac{\displaystyle\int_0^{\infty} z^M \tfrac{1}{\sigma}\phi\left[\tfrac{z - (-\varepsilon - \theta\sigma^2)}{\sigma}\right]dz}{\displaystyle\int_0^{\infty} \tfrac{1}{\sigma}\phi\left[\tfrac{z - (-\varepsilon - \theta\sigma^2)}{\sigma}\right]dz}.$$

The integral only exists in closed form for integer values of M. However, the weighting function that appears in the integral is of the form

$$f(z|z > 0) = \frac{f(z)}{\text{Prob}[z > 0]} = \frac{\tfrac{1}{\sigma}\phi\left(\tfrac{z - \mu}{\sigma}\right)}{\displaystyle\int_0^{\infty}\tfrac{1}{\sigma}\phi\left(\tfrac{z - \mu}{\sigma}\right)dz}.$$

This is a truncated normal distribution. It is the distribution of a normally distributed variable z with mean μ and standard deviation σ, conditioned on z being greater than zero. The integral is equal to the expected value of z^M given that z is greater than zero when z is normally distributed with mean $\mu = -\varepsilon - \theta\sigma^2$ and variance σ^2.

The truncated normal distribution is examined in Section 19.2. The function $h(M, \varepsilon)$ is the expected value of z^M when z is the truncation of a normal random variable with mean μ and standard deviation σ. To evaluate the integral by Monte Carlo integration, we would require a sample $z_1, \ldots, z_R$ from this distribution. We have the results we need in (15-4) with $L = 0$, so $P_L = \Phi[0 - (-\varepsilon - \theta\sigma^2)/\sigma] = \Phi(\varepsilon/\sigma + \theta\sigma)$ and $U = +\infty$ so $P_U = 1$. Then, a draw on z is obtained by

$$z = \mu + \sigma\Phi^{-1}[P_L + F(1 - P_L)],$$

where F is the primitive draw from $U[0, 1]$. Finally, the integral is approximated by the simple average of the draws,

$$h(M, \varepsilon) \approx \frac{1}{R}\sum_{r=1}^{R} z[\varepsilon, \theta, \sigma, F_r]^M.$$

This is an application of Monte Carlo integration. In certain cases, an integral can be approximated by computing the sample average of a set of function values. The approach taken here was to interpret the integral as an expected value. Our basic statistical result for the behavior of sample means implies that, with a large enough sample, we can approximate the integral as closely as we like. The general approach is widely applicable in Bayesian econometrics and classical statistics and econometrics as well.[17]

15.6.2a Halton Sequences and Random Draws for Simulation-Based Integration

Monte Carlo integration is used to evaluate the expectation

$$E[g(x)] = \int_x g(x)f(x)dx,$$

where $f(x)$ is the density of the random variable x and $g(x)$ is a smooth function. The Monte Carlo approximation is

$$E[g(x)] \approx \frac{1}{R}\sum_{r=1}^{R} g(x_r).$$

Convergence of the approximation to the expectation is based on the law of large numbers—a random sample of draws on $g(x)$ will converge in probability to its expectation. The standard approach to simulation-based integration is to use random draws from the specified distribution. Conventional simulation-based estimation uses a random number generator to produce the draws from a specified distribution. The central component of this approach is drawn from the standard continuous uniform distribution, $U[0, 1]$. Draws from other distributions are obtained from these draws by using transformations. In particular, for a draw from the normal distribution, where u_i is one draw from $U[0, 1]$, $v_i = \Phi^{-1}(u_i)$. Given that the initial draws satisfy the necessary assumptions, the central issue for purposes of specifying the simulation is the number of draws. Good performance in this connection requires large numbers of draws. Results differ on the number needed in a given application, but the general finding is that when simulation is done in this fashion, the number is large (hundreds or thousands). A consequence of this is that for large-scale problems, the amount of computation time in simulation-based estimation can be extremely large. Numerous methods have been devised for reducing the numbers of draws needed to obtain a satisfactory approximation. One such method is to introduce some autocorrelation into the draws—a small amount of negative correlation across the draws will reduce the variance of the simulation. **Antithetic draws**, whereby each draw in a sequence is included with its mirror image (w_i and $-w_i$ for normally distributed draws, w_i and $1 - w_i$ for uniform, for example), is one such method.[18]

[17]See Geweke (1986, 1988, 1989, 2005) for discussion and applications. A number of other references are given in Poirier (1995, p. 654) and Koop (2003). See, as well, Train (2009).

[18]See Geweke (1988) and Train (2009, Chapter 9).

Procedures have been devised in the numerical analysis literature for taking intelligent draws from the uniform distribution, rather than random ones.[19] An emerging literature has documented dramatic speed gains with no degradation in simulation performance through the use of a smaller number of **Halton draws** or other constructed, nonrandom sequences instead of a large number of random draws. These procedures appear to vastly reduce the number of draws needed for estimation (sometimes by a factor of 90% or more) and reduce the simulation error associated with a given number of draws. In one application of the method to be discussed here, Bhat (1999) found that 100 Halton draws produced lower simulation error than 1,000 random numbers.

A Halton sequence is generated as follows: Let r be a prime number. Expand the sequence of integers $g = 1, 2, \ldots$ in terms of the base r as

$$g = \sum_{i=0}^{I} b_i r^i \text{ where, by construction, } 0 \leq b_i \leq r - 1 \text{ and } r^I \leq g < r^{I+1}.$$

The Halton sequence of values that corresponds to this series is

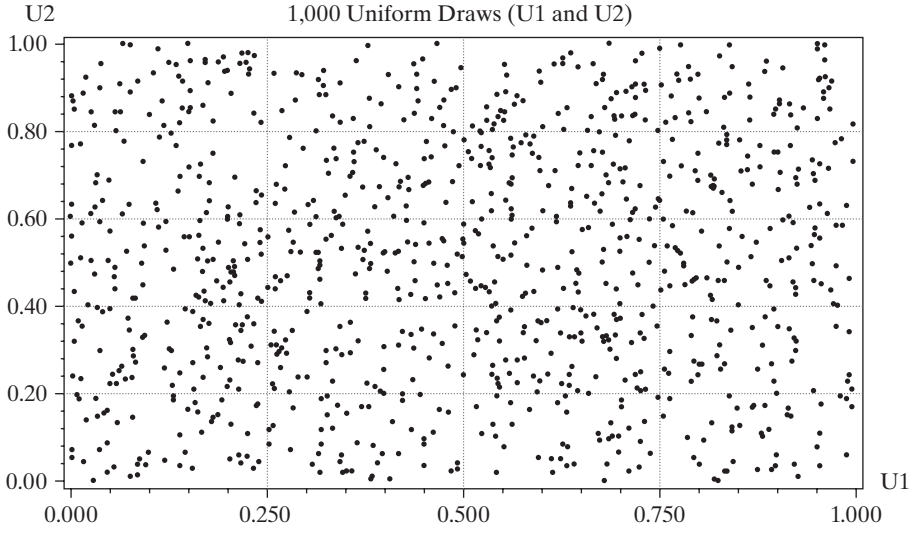
$$H(g) = \sum_{i=0}^{I} b_i r^{-i-1}.$$

For example, using base 5, the integer 37 has $b_0 = 2$, $b_1 = 2$, and $b_2 = 1$. Then

$$H_5(37) = 2 \times 5^{-1} + 2 \times 5^{-2} + 1 \times 5^{-3} = 0.488.$$

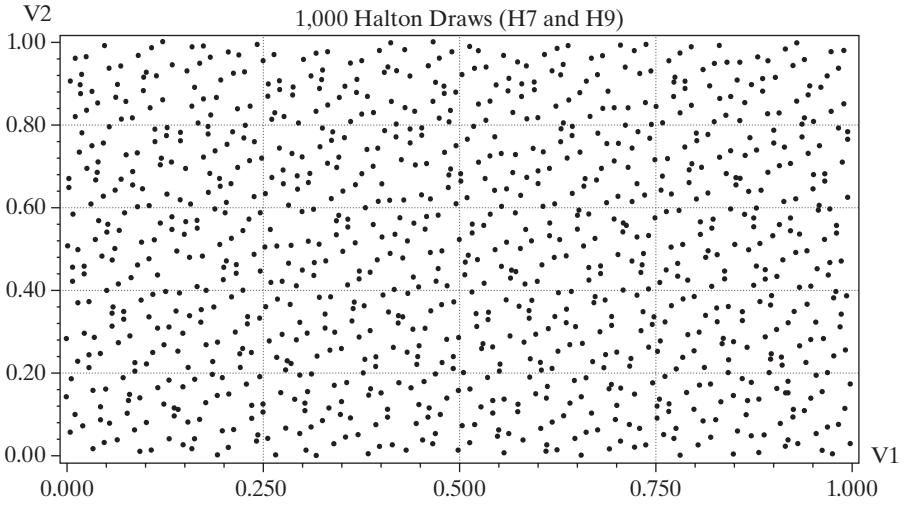
The sequence of Halton values is efficiently spread over the unit interval. The sequence is not random as the sequence of pseudo-random numbers is; it is a well-defined deterministic sequence. But randomness is not the key to obtaining accurate approximations to integrals. Uniform coverage of the support of the random variable is the central requirement. The large numbers of random draws are required to obtain smooth and dense coverage of the unit interval. Figures 15.3 and 15.4 show two sequences

FIGURE 15.3 Bivariate Distribution of Random Uniform Draws.



[19]See Train (1999, 2009) and Bhat (1999) for extensive discussion and further references.

FIGURE 15.4 Bivariate Distribution of Halton (7) and Halton (9).



of 1,000 Halton draws and two sequences of 1,000 pseudo-random draws. The Halton draws are based on $r = 7$ and $r = 9$. The clumping evident in the first figure is the feature (among others) that mandates large samples for simulations.

Example 15.9 Estimating the Lognormal Mean

We are interested in estimating the mean of a standard lognormally distributed variable. Formally, this result is

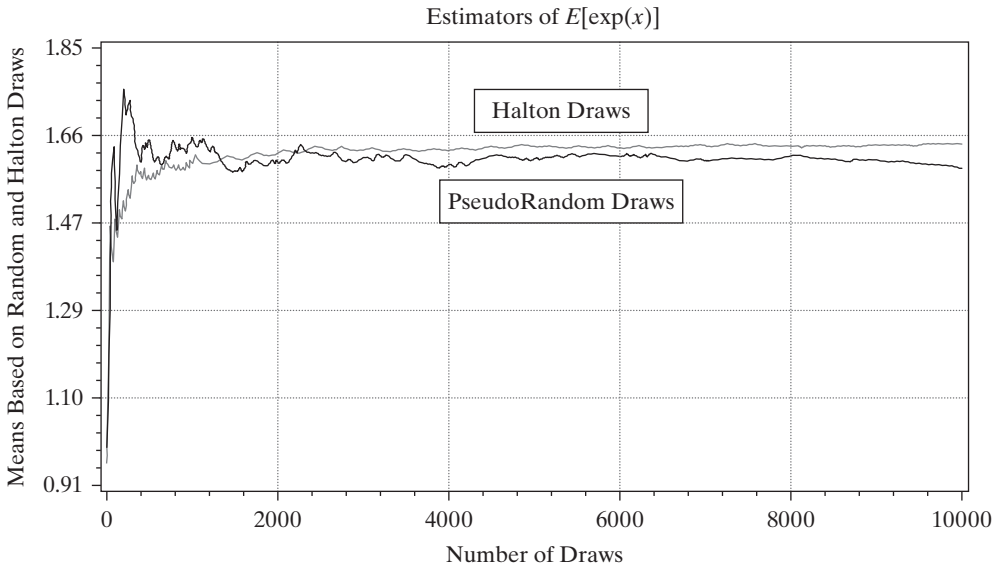
$$E[y] = \int_{-\infty}^{\infty} \exp(x) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] dx = 1.649.$$

To use simulation for the estimation, we will average n draws on $y = \exp(x)$ where x is drawn from the standard normal distribution. To examine the behavior of the Halton sequence as compared to that of a set of pseudo-random draws, we did the following experiment. Let $x_{i,t}$ = the sequence of values for a standard normally distributed variable. We draw $t = 1, \ldots, 10{,}000$ draws. For $i = 1$, we used a random number generator. For $i = 2$, we used the sequence of the first 10,000 Halton draws using $r = 7$. The Halton draws were converted to standard normal using the inverse normal transformation. To finish preparation of the data, we transformed $x_{i,t}$ to $y_{i,t} = \exp(x_{i,t})$. Then, for $n = 100, 110, \ldots, 10{,}000$, we averaged the first n observations in the sample. Figure 15.5 plots the evolution of the sample means as a function of the sample size. The lower trace is the sequence of Halton-based means. The greater stability of the Halton estimator is clearly evident in the figure.

15.6.2.b Computing Multivariate Normal Probabilities Using the GHK Simulator

The computation of bivariate normal probabilities is typically done using quadrature and requires a large amount of computing effort. Quadrature methods have been developed for trivariate probabilities as well, but the amount of computing effort needed at this level is enormous. For integrals of level greater than three, satisfactory (in terms of speed and accuracy) direct approximations remain to be developed. Our work thus far does

FIGURE 15.5 Estimates of $E[\exp(x)]$ Based on Random Draws and Halton Sequences, by Sample Size.



suggest an alternative approach. Suppose that $\mathbf{x}$ has a K-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. (No generality is sacrificed by the assumption of a zero mean, because we could just subtract a nonzero mean from the random vector wherever it appears in any result.) We wish to compute the K-variate probability, $\text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, a_K < x_K < b_K]$. The Monte Carlo integration technique is well suited for this problem. As a first approach, consider sampling R observations, $\mathbf{x}_r$, $r = 1, \ldots, R$, from this multivariate normal distribution, using the method described in Section 15.2.4. Now, define

$$d_r = \mathbf{1}[a_1 < x_{r1} < b_1, a_2 < x_{r2} < b_2, \ldots, a_K < x_{rK} < b_K].$$

(That is, $d_r = 1$ if the condition is true and 0 otherwise.) Based on our earlier results, it follows that

$$\text{plim}\, \bar{d} = \text{plim}\, \frac{1}{R}\sum_{r=1}^{R} d_r = \text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, a_K < x_K < b_K].^{20}$$

This method is valid in principle, but in practice it has proved to be unsatisfactory for several reasons. For large-order problems, it requires an enormous number of draws from the distribution to give reasonable accuracy. Also, even with large numbers of draws, it appears to be problematic when the desired tail area is very small. Nonetheless, the idea is sound, and recent research has built on this idea to produce some quite accurate and efficient simulation methods for this computation. A survey of the methods is given in McFadden and Ruud (1994).[21]

[20]This method was suggested by Lerman and Manski (1981).

[21]A symposium on the topic of simulation methods appears in *Review of Economic Statistics*, Vol. 76, November 1994. See, especially, McFadden and Ruud (1994), Stern (1994), Geweke, Keane, and Runkle (1994), and Breslaw (1994). See, as well, Gourieroux and Monfort (1996).

Among the simulation methods examined in the survey, the **GHK smooth recursive simulator** appears to be the most accurate.[22] The method is surprisingly simple. The general approach uses

$$\text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ \ldots, a_K < x_K < b_K] \approx \frac{1}{R}\sum_{r=1}^{R}\prod_{k=1}^{K} Q_{rk},$$

where Q_{rk} are easily computed univariate probabilities. The probabilities Q_{rk} are computed according to the following recursion: We first factor $\boldsymbol{\Sigma}$ using the **Cholesky factorization** $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$, where $\mathbf{C}$ is a lower triangular matrix (see Section A.6.11). The elements of $\mathbf{C}$ are l_{km}, where $l_{km} = 0$ if $m > k$. Then we begin the recursion with

$$Q_{r1} = \Phi(b_1/l_{11}) - \Phi(a_1/l_{11}).$$

Note that $l_{11} = \sigma_{11}$, so this is just the marginal probability, $\text{Prob}[a_1 < x_1 < b_1]$. Now, using (15-4), we generate a random observation ε_{r1} from the truncated standard normal distribution in the range

$$A_{r1} \text{ to } B_{r1} = a_1/l_{11} \text{ to } b_1/l_{11}.$$

(*Note:* The range is standardized because $l_{11} = \sigma_{11}$.) For steps $k = 2, \ldots, K$, compute

$$A_{rk} = \left[a_k - \sum_{m=1}^{k-1} l_{km}\varepsilon_{rm}\right] \Big/ l_{kk},$$

$$B_{rk} = \left[b_k - \sum_{m=1}^{k-1} l_{km}\varepsilon_{rm}\right] \Big/ l_{kk}.$$

Then,

$$Q_{rk} = \Phi(B_{rk}) - \Phi(A_{rk}).$$

Finally, in preparation for the next step in the recursion, we generate a random draw from the truncated standard normal distribution in the range A_{rk} to B_{rk}. This process is replicated R times, and the estimated probability is the sample average of the simulated probabilities.

The GHK simulator has been found to be impressively fast and accurate for fairly moderate numbers of replications. Its main usage has been in computing functions and derivatives for maximum likelihood estimation of models that involve multivariate normal integrals. We will revisit this in the context of the method of simulated moments when we examine the probit model in Chapter 17.

15.6.3 SIMULATION-BASED ESTIMATION OF RANDOM EFFECTS MODELS

In Section 15.6.2, (15-10), and (15-14), we developed a random effects specification for the Poisson regression model. For feasible estimation and inference, we replace the log-likelihood function,

$$\ln L = \sum_{i=1}^{n} \ln\left\{\int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}}{y_{it}!}\right]\phi(w_i)dw_i\right\},$$

with the simulated log-likelihood function,

[22] See Geweke (1989), Hajivassiliou (1990), and Keane (1994). Details on the properties of the simulator are given in Börsch-Supan and Hajivassiliou (1993).

$$\ln L_S = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}}{y_{it}!} \right\}. \quad \textbf{(15-16)}$$

We now consider how to estimate the parameters via maximum simulated likelihood. In spite of its complexity, the simulated log likelihood will be treated in the same way that other log likelihoods were handled in Chapter 14. That is, we treat $\ln L_S$ as a function of the unknown parameters conditioned on the data, $\ln L_S(\boldsymbol{\beta}, \sigma)$, and maximize the function using the methods described in Appendix E, such as the DFP or BFGS gradient methods. What is needed here to complete the derivation are expressions for the derivatives of the function. We note that the function is a sum of n terms; asymptotic results will be obtained in n; each observation can be viewed as one T_i-variate observation.

In order to develop a general set of results, it will be convenient to write each single density in the simulated function as

$$P_{itr}(\boldsymbol{\beta}, \sigma) = f(y_{it} | \mathbf{x}_{it}, w_{ir}, \boldsymbol{\beta}, \sigma) = P_{itr}(\boldsymbol{\theta}) = P_{itr}.$$

For our specific application in (15-16),

$$P_{itr} = \frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}]}{y_{it}!}.$$

The simulated log likelihod is, then,

$$\ln L_S = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta}) \right\}. \quad \textbf{(15-17)}$$

Continuing this shorthand, then, we will also define

$$P_{ir} = P_{ir}(\boldsymbol{\theta}) = \prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta}),$$

so that

$$\ln L_S = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} P_{ir}(\boldsymbol{\theta}) \right\}.$$

And, finally,

$$P_i = P_i(\boldsymbol{\theta}) = \frac{1}{R} \sum_{r=1}^{R} P_{ir},$$

so that

$$\ln L_S = \sum_{i=1}^{n} \ln P_i(\boldsymbol{\theta}). \quad \textbf{(15-18)}$$

With this general template, we will be able to accommodate richer specifications of the index function, now $\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_i$, and other models such as the linear regression, binary choice models, and so on, simply by changing the specification of P_{itr}.

The algorithm will use the usual procedure,

$$\hat{\boldsymbol{\theta}}^{(k)} = \hat{\boldsymbol{\theta}}^{(k-1)} + \textit{update vector},$$

starting from an initial value, $\hat{\boldsymbol{\theta}}^{(0)}$, and will exit when the update vector is sufficiently small. A natural initial value would be from a model with no random effects; that is, the pooled estimator for the linear or Poisson or other model with $\sigma = 0$. Thus, at entry to the iteration (update), we will compute

$$\ln \hat{L}_S^{(k-1)} = \sum_{i=1}^{n} \ln\left\{\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\frac{\exp[-\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}^{(k-1)} + \hat{\sigma}^{(k-1)}w_{ir})][\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}^{(k-1)} + \hat{\sigma}^{(k-1)}w_{ir})]^{y_{it}}]}{y_{it}!}\right\}.$$

To use a gradient method for the update, we will need the first derivatives of the function. Computation of an asymptotic covariance matrix may require the Hessian, so we will obtain this as well.

Before proceeding, we note two important aspects of the computation. First, a question remains about the number of draws, R, required for the maximum simulated likelihood estimator to be consistent. The approximated function,

$$\hat{E}_w[f(y|\mathbf{x}, w)] = \frac{1}{R}\sum_{r=1}^{R} f(y|\mathbf{x}, w_r),$$

is an unbiased estimator of $E_w[f(y|\mathbf{x}, w)]$. However, what appears in the simulated log likelihood is $\ln E_w[f(y|\mathbf{x}, w)]$, and the log of the estimator is a biased estimator of the log of its expectation. To maintain the asymptotic equivalence of the MSL estimator of $\boldsymbol{\theta}$ and the true MLE (if w were observed), it is necessary for the estimators of these terms in the log likelihood to converge to their expectations faster than the expectation of $\ln L$ converges to its expectation. The requirement is that $n^{1/2}/R \to 0$.[23] The estimator remains consistent if $n^{1/2}$ and R increase at the same rate; however, the asymptotic covariance matrix of the MSL estimator will then be larger than that of the true MLE. In practical terms, this suggests that the number of draws be on the order of $n^{.5+\delta}$ for some positive δ. [This does not state, however, what R should be for a given n; it only establishes the properties of the MSL estimator as n increases. For better or worse, researchers who have one sample of n observations often rely on the numerical stability of the estimator with respect to changes in R as their guide. Hajivassiliou (2000) gives some suggestions.] Note, as well, that the use of Halton sequences or any other autocorrelated sequences for the simulation, which is becoming more prevalent, interrupts this result. The appropriate counterpart to the Gourieroux and Monfort result for random sampling remains to be derived. One might suspect that the convergence result would persist, however. The usual standard is several hundred.

Second, it is essential that the same (pseudo- or Halton) draws be used every time the function or derivatives or any function involving these is computed for observation i. This can be achieved by creating the pool of draws for the entire sample before the optimization begins, and simply dipping into the same point in the pool each time a computation is required for observation i. Alternatively, if computer memory is an issue and the draws are re-created for each individual each time, the same practical result can be achieved by setting a preassigned seed for individual i, $seed(i) = s(i)$ for some simple monotonic function of i, and resetting the seed when draws for individual i are needed.

To obtain the derivatives, we begin with

$$\frac{\partial \ln L_S}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{(1/R)\sum_{r=1}^{R}\partial\left(\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})\right)/\partial\boldsymbol{\theta}}{(1/R)\sum_{r=1}^{R}\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})}. \qquad \textbf{(15-19)}$$

[23] See Gourieroux and Monfort (1996).

For the derivative term,

$$\begin{aligned}\partial\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} &= \left(\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})\right)\partial\left(\ln\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})\right)/\partial\boldsymbol{\theta}\\ &= \left(\prod_{t=1}^{T_i}P_{itr}(\boldsymbol{\theta})\right)\sum_{t=1}^{T_i}\partial\ln P_{itr}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\\ &= P_{ir}(\boldsymbol{\theta})\left(\sum_{t=1}^{T_i}\partial\ln P_{itr}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right) = P_{ir}(\boldsymbol{\theta})\sum_{t=1}^{T_i}\mathbf{g}_{itr}(\boldsymbol{\theta})\\ &= P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta}).\end{aligned} \tag{15-20}$$

Now, insert the result of (15-20) in (15-19) to obtain

$$\frac{\partial\ln L_S(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n}\frac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})}. \tag{15-21}$$

Define the weight $Q_{ir}(\boldsymbol{\theta}) = P_{ir}(\boldsymbol{\theta})/\Sigma_{r=1}^{R}P_{ir}(\boldsymbol{\theta})$ so that $0 < Q_{ir}(\boldsymbol{\theta}) < 1$ and $\Sigma_{r=1}^{R}Q_{ir}(\boldsymbol{\theta}) = 1$. Then,

$$\frac{\partial\ln L_S(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n}\sum_{r=1}^{R}Q_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\bar{\mathbf{g}}_i(\boldsymbol{\theta}). \tag{15-22}$$

To obtain the second derivatives, define $\mathbf{H}_{itr}(\boldsymbol{\theta}) = \partial^2\ln P_{itr}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ and let

$$\mathbf{H}_{ir}(\boldsymbol{\theta}) = \sum_{t=1}^{T_i}\mathbf{H}_{itr}(\boldsymbol{\theta})$$

and

$$\bar{\mathbf{H}}_i(\boldsymbol{\theta}) = \sum_{r=1}^{R}Q_{ir}(\boldsymbol{\theta})\mathbf{H}_{ir}(\boldsymbol{\theta}). \tag{15-23}$$

Then, working from (15-21), the second derivatives matrix breaks into three parts as follows:

$$\frac{\partial^2\ln L_s(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{i=1}^{n}\left[\begin{array}{l}\dfrac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{H}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})} +\\ \dfrac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})'}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})} - \left[\dfrac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})}\right]\left[\dfrac{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})\mathbf{g}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^{R}P_{ir}(\boldsymbol{\theta})}\right]'\end{array}\right].$$

We can now use (15-20) through (15-23) to combine these terms;

$$\frac{\partial^2\ln L_S}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{i=1}^{n}\left\{\bar{\mathbf{H}}_i(\boldsymbol{\theta}) + \sum_{r=1}^{R}Q_{ir}(\boldsymbol{\theta})[\mathbf{g}_{ir}(\boldsymbol{\theta}) - \bar{\mathbf{g}}_i(\boldsymbol{\theta})][\mathbf{g}_{ir}(\boldsymbol{\theta}) - \bar{\mathbf{g}}_i(\boldsymbol{\theta})]'\right\}. \tag{15-24}$$

An estimator of the asymptotic covariance matrix for the MSLE can be obtained by computing the negative inverse of this matrix.

Example 15.10 Poisson Regression Model with Random Effects

For the Poisson regression model, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma)'$ and

$$
\begin{aligned}
P_{itr}(\boldsymbol{\theta}) &= \frac{\exp[-\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}]}{y_{it}!} = \frac{\exp[-\mu_{itr}(\theta)]\mu_{itr}(\theta)^{y_{it}}}{y_{it}!} \\
\mathbf{g}_{itr}(\boldsymbol{\theta}) &= [y_{it} - \mu_{itr}(\boldsymbol{\theta})]\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix} \\
\mathbf{H}_{itr}(\boldsymbol{\theta}) &= -\mu_{itr}(\boldsymbol{\theta})\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}'.
\end{aligned}
\tag{15-25}
$$

Estimates of the random effects model parameters would be obtained by using these expressions in the preceding general template. We will apply these results in an application in Chapter 19 where the Poisson regression model is developed in greater detail.

Example 15.11 Maximum Simulated Likelihood Estimation of the Random Effects Linear Regression Model

The preceding method can also be used to estimate a linear regression model with random effects. We have already seen two ways to estimate this model, using two-step FGLS in Section 11.5.3 and by (closed form) maximum likelihood in Section 14.9.6.a. It might seem redundant to construct yet a third estimator for the model. However, this third approach will be the only feasible method when we generalize the model to have other random parameters in the next section. To use the simulation estimator, we define $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u, \sigma_\varepsilon)$. We will require

$$
\begin{aligned}
P_{itr}(\boldsymbol{\theta}) &= \frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\exp\left[-\frac{(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \sigma_u w_{ir})^2}{2\sigma_\varepsilon^2}\right], \\
\mathbf{g}_{itr}(\boldsymbol{\theta}) &= \begin{bmatrix}\left(\dfrac{(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \sigma_u w_{ir}}{\sigma_\varepsilon^2}\right)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix} \\ \dfrac{(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \sigma_u w_{ir})^2}{\sigma_\varepsilon^3} - \dfrac{1}{\sigma_\varepsilon}\end{bmatrix} = \begin{bmatrix}(\varepsilon_{itr}/\sigma_\varepsilon^2)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix} \\ (1/\sigma_\varepsilon)[(\varepsilon_{itr}^2/\sigma_\varepsilon^2) - 1]\end{bmatrix}, \\
\mathbf{H}_{itr}(\boldsymbol{\theta}) &= \begin{bmatrix}-(1/\sigma_\varepsilon^2)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix}' & -(2\varepsilon_{itr}/\sigma_\varepsilon^3)\begin{pmatrix}\mathbf{x}_{it}\\ w_{ir}\end{pmatrix} \\ -(2\varepsilon_{itr}/\sigma_\varepsilon^3)(\mathbf{x}_{it}'w_{ir}) & -(3\varepsilon_{itr}^2/\sigma_\varepsilon^4) + (1/\sigma_\varepsilon^2)\end{bmatrix}.
\end{aligned}
\tag{15-26}
$$

Note in the computation of the disturbance variance, σ_ε^2, we are using the sum of squared simulated residuals. However, the estimator of the variance of the heterogeneity, σ_u, is not being computed as a mean square. It is essentially the regression coefficient on w_{ir}. One surprising implication is that the actual estimate of σ_u can be negative. This is the same result that we have encountered in other situations. In no case is there a natural estimator of σ_u^2 that is based on a sum of squares. However, in this context, there is yet another surprising aspect of this calculation. In the simulated log-likelihood function, if every w_{ir} for every individual were changed to $-w_{ir}$ and σ_u is changed to $-\sigma_u$, then the exact same value of the function and all derivatives results. The implication is that the sign of σ_u is not identified in this setting. With no loss of generality, it is normalized to positive (+) to be consistent with the underlying theory that it is a standard deviation.

15.7 A RANDOM PARAMETERS LINEAR REGRESSION MODEL

We will slightly reinterpret the random effects model as

$$\begin{aligned} y_{it} &= \beta_{0i} + \mathbf{x}_{it1}'\boldsymbol{\beta}_1 + \varepsilon_{it}, \\ \beta_{0i} &= \beta_0 + u_i. \end{aligned} \tag{15-27}$$

This is equivalent to the random effects model, though in (15-27), we reinterpret it as a regression model with a randomly distributed constant term. In Section 11.10.1, we built a linear regression model that provided for parameter heterogeneity across individuals,

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}, \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta} + \mathbf{u}_i, \end{aligned} \tag{15-28}$$

where $\mathbf{u}_i$ has mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Gamma}$. In that development, we took a fixed effects approach in that no restriction was placed on the covariance between $\mathbf{u}_i$ and $\mathbf{x}_{it}$. Consistent with these assumptions, we constructed an estimator that involved n regressions of $\mathbf{y}_i$ on $\mathbf{X}_i$ to estimate $\boldsymbol{\beta}$ one unit at a time. Each estimator is consistent in T_i. (This is precisely the approach taken in the fixed effects model, where there are n unit specific constants and a common $\boldsymbol{\beta}$. The approach there is to estimate $\boldsymbol{\beta}$ first and then to regress $\mathbf{y}_i - \mathbf{X}_i, \mathbf{b}_{\text{LSDV}}$ on $\mathbf{d}_i$ to estimate α_i.) In the same way that assuming that u_i is uncorrelated with $\mathbf{x}_{it}$ in the fixed effects model provided a way to use FGLS to estimate the parameters of the random effects model, if we assume in (15-28) that $\mathbf{u}_i$ is uncorrelated with $\mathbf{X}_i$, we can extend the random effects model in Section 15.6.3 to a model in which some or all of the other coefficients in the regression model, not just the constant term, are randomly distributed. The theoretical proposition is that the model is now extended to allow individual heterogeneity in all coefficients.

To implement the extended model, we will begin with a simple formulation in which $\mathbf{u}_i$ has a diagonal covariance matrix—this specification is quite common in the literature. The implication is that the random parameters are uncorrelated; $\beta_{i,k}$ has mean β_k and variance γ_k^2. The model in (15-26) can modified to allow this case with a few minor changes in notation. Write

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_i, \tag{15-29}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the standard deviations $(\gamma_1, \gamma_2, \ldots, \gamma_K)$ of $(u_{i1}, \ldots, u_{iK})$ on the diagonal and $\mathbf{w}_i$ is now a random vector with zero means and unit standard deviations. Then, $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. The parameter vector in the model is now

$$\boldsymbol{\theta} = (\beta_1, \ldots, \beta_K, \gamma_1, \ldots, \gamma_K, \sigma_\varepsilon)'.$$

(In an application, some of the γ's might be fixed at zero to make the corresponding parameters nonrandom.) In order to extend the model, the disturbance in (15-26), $\varepsilon_{itr} = (y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} - \sigma_u w_{ir})$, becomes

$$\varepsilon_{itr} = y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{ir}). \tag{15-30}$$

Now, combine (15-17) and (15-29) with (15-30) to produce

$$\ln L_S = \sum_{i=1}^{n} \ln\left\{ \frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i} \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} \exp\left[\frac{(y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{ir}))^2}{2\sigma_\varepsilon^2} \right] \right\}. \tag{15-31}$$

In the derivatives in (15-26), the only change needed to accommodate this extended model is that the scalar w_{ir} becomes the vector $(w_{ir,1}x_{it1}, w_{ir,2}x_{it,2}, \ldots, w_{ir,K}x_{it,K})$. This is the element-by-element product of the regressors, $\mathbf{x}_{it}$, and the vector of random draws, $\mathbf{w}_{ir}$, which is the **Hadamard product**, **direct product**, or **Schur product** of the two vectors, usually denoted $\mathbf{x}_{it} \circ \mathbf{w}_{ir}$.

Although only a minor change in notation in the random effects template in (15-26), this formulation brings a substantial change in the formulation of the model. The integral in ln L is now a K dimensional integral. Maximum simulated likelihood estimation proceeds as before, with potentially much more computation as each draw now requires a K-variate vector of pseudo-random draws.

The random parameters model can now be extended to one with a full covariance matrix, $\boldsymbol{\Gamma}$ as we did with the fixed effects case. We will now let $\boldsymbol{\Gamma}$ in (15-29) be the Cholesky factorization of $\boldsymbol{\Gamma}$, so $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. (This was already the case for the simpler model with diagonal $\boldsymbol{\Gamma}$.) The implementation in (15-26) will be a bit complicated. The derivatives with respect to $\boldsymbol{\beta}$ are unchanged. For the derivatives with respect to $\boldsymbol{\Lambda}$, it is useful to assume for the moment that $\boldsymbol{\Lambda}$ is a full matrix, not a lower triangular one. Then, the scalar w_{ir} in the derivative expression becomes a $K^2 \times 1$ vector in which the $(k-1) \times K + l^{\text{th}}$ element is $x_{it,k} \times w_{ir,l}$. The full set of these is the **Kronecker product** of $\mathbf{x}_{it}$ and $\mathbf{w}_{ir}$, $\mathbf{x}_{it} \otimes \mathbf{w}_{ir}$. The necessary elements for maximization of the log-likelihood function are then obtained by discarding the elements for which $\boldsymbol{\Lambda}_{kl}$ are known to be zero—these correspond to $l > k$.

In (15-26), for the full model, for computing the MSL estimators, the derivatives with respect to $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$. are equated to zero. The result after some manipulation is

$$\frac{\partial \ln L_S}{\partial(\boldsymbol{\beta}, \boldsymbol{\Lambda})} = \sum_{i=1}^{n} \frac{1}{R} \sum_{r=1}^{R} \sum_{t=1}^{T_i} \frac{(y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{ir}))}{\sigma_\varepsilon^2} \begin{bmatrix} \mathbf{x}_{it} \\ \mathbf{x}_{it} \otimes \mathbf{w}_{ir} \end{bmatrix} = \mathbf{0}.$$

By multiplying this by σ_ε^2, we find, as usual, that σ_ε^2 is not needed for computation of the estimates of $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$. Thus, we can view the solution as the counterpart to least squares, which might call, instead, the least simulated sum of squares estimator. Once the simulated sum of squares is minimized with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$, then the solution for σ_ε^2 can be obtained via the likelihood equation,

$$\frac{\partial \ln L_S}{\partial \sigma_\varepsilon^2} = \sum_{i=1}^{n} \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\frac{-T_i}{2\sigma_\varepsilon^2} + \frac{\sum_{t=1}^{T_i} (y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{i,r}))^2}{2\sigma_\varepsilon^4} \right] \right\} = 0.$$

Multiply both sides of this equation by $-2\sigma_\varepsilon^4$ to obtain the equivalent condition

$$\frac{\partial \ln L_S}{\partial \sigma_\varepsilon^2} = \sum_{i=1}^{n} \left\{ \frac{1}{R} \sum_{r=1}^{R} T_i \left[-\sigma_\varepsilon^2 + \frac{\sum_{t=1}^{T_i} (y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{i,r}))^2}{T_i} \right] \right\} = 0.$$

By expanding this expression and manipulating it a bit, we find the solution for σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{n} F_i \frac{1}{R} \sum_{r=1}^{R} \hat{\sigma}_{\varepsilon,ir}^2, \text{ where } \hat{\sigma}_{\varepsilon,ir}^2 = \frac{\sum_{t=1}^{T_i} (y_{it} - \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{i,r}))^2}{T_i}$$

and $F_i = T_i / \Sigma_i T_i$ is a weight for each group that equals $1/n$ if T_i is the same for all i.

Example 15.12 Random Parameters Wage Equation

Estimates of the random effects log wage equation from the Cornwell and Rupert study in Examples 11.7 and 15.6 are shown in Table 15.6. The table presents estimates based on several assumptions. The encompassing model is

$$\ln Wage_{it} = \beta_{1,i} + \beta_{2,i}Wks_{i,t} + \cdots + \beta_{12,i}Fem_i + \beta_{13,i}Blk_i + \varepsilon_{it}, \quad \textbf{(15-32)}$$

$$\beta_{k,i} = \beta_k + \lambda_k w_{ik},\ w_{ik} \sim N[0, 1],\ k = 1, \ldots, 13. \quad \textbf{(15-33)}$$

TABLE 15.6 Estimated Wage Equations (Standard errors in parentheses)

Variable	*Pooled OLS*	*Feasible Two-Step GLS*	*Maximum Likelihood*	*Maximum Simulated Likelihood*[a]	*Random Parameters Max. Simulated Likelihood*[a]	
					β	λ
Wks	0.00422	0.00096	0.00084	0.00086	−0.00029	0.00614
	(0.00108)	(0.00059)	(0.00060)	(0.00047)	(0.00082)	(0.00042)
South	−0.05564	−0.00825	0.00577	0.00935	0.04941	0.20997
	(0.01253)	(0.02246)	(0.03159)	(0.00508)	(0.02002)	(0.01702)
SMSA	0.15167	−0.02840	−0.04748	−0.04913	−0.05486	0.01165
	(0.01207)	(0.01616)	(0.01896)	(0.00507)	(0.01747)	(0.02738)
MS	0.04845	−0.07090	−0.04138	−0.04142	−0.06358	0.02524
	(0.02057)	(0.01793)	(0.01899)	(0.00824)	(0.01896)	(0.03190)
Exp	0.04010	0.08748	0.10721	0.10668	0.09291	0.01803
	(0.00216)	(0.00225)	(0.00248)	(0.00096)	(0.00216)	(0.00092)
Exp^2	−0.00067	−0.00076	−0.00051	−0.00050	−0.00019	0.00008
	(0.00005)	(0.00005)	(0.00005)	(0.00002)	(0.00007	(0.00002)
Occ	−0.14001	−0.04322	−0.02512	−0.02437	−0.00963	0.02565
	(0.01466)	(0.01299)	(0.01378)	(0.00593)	(0.01331)	(0.01019)
Ind	0.04679	0.00378	0.01380	0.01610	0.00207	0.02575
	(0.01179)	(0.01373)	(0.01529)	(0.00490)	(0.01357)	(0.02420)
Union	0.09263	0.05835	0.03873	0.03724	0.05749	0.15260
	(0.01280)	(0.01350)	(0.01481)	(0.00509)	(0.01469)	(0.02022)
Ed	0.05670	0.10707	0.13562	0.13952	0.09356	0.00409
	(0.00261)	(0.00511)	(0.01267)	(0.01170)	(0.00359)	(0.00160)
Fem	−0.36779	−0.30938	−0.17562	−0.11694	−0.03864	0.28310
	(0.02510)	(0.04554)	(0.11310)	(0.01060)	(0.02467)	(0.00760)
Blk	−0.16694	−0.21950	−0.26121	−0.15184	−0.26864	0.02930
	(0.02204)	(0.05252)	(0.13747)	(0.00979)	(0.03156)	(0.03841)
Constant	5.25112	4.04144	3.12622	3.08362	3.81680	0.26347
	(0.07129)	(0.08330)	(0.17761)	(0.03276)	(0.06905)	(0.01628)
σ_u	0.00000	0.31453	0.83932	0.80926		
σ_ε	0.34936	0.15206	0.15334	0.15326	0.14354	
	LM = 3497.02				(0.00208)	
Ln L	−1523.254		307.873	309.173	365.313	

[a] Based on 500 Halton draws.

Under the assumption of homogeneity, that is, $\lambda_k = 0$, the pooled OLS estimator is consistent and efficient. As we saw in Chapter 11, under the random effects assumption, that is $\lambda_k = 0$ for $k = 2, \ldots, 13$ but $\lambda_1 \neq 0$, the OLS estimator is consistent, as are the next three estimators that explicitly account for the heterogeneity. To consider the full specification, write the model in the equivalent form

$$\begin{aligned}\ln Wage_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \left(\lambda_1 w_{i,1} + \sum_{k=2}^{13} \lambda_k w_{i,k} x_{it,k}\right) + \varepsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\beta} + W_{it} + \varepsilon_{it}.\end{aligned}$$

This is still a regression: $E[W_{it} + \varepsilon_{it}|\mathbf{X}] = 0$. (For the product terms, $E[\lambda_k w_{i,k} x_{it,k}|\mathbf{X}] = \lambda_k x_{it,k} E[w_{i,k}|x_{itk}] = 0$.) Therefore, even OLS remains consistent. The heterogeneity induces heteroscedasticity in W_{it} so the OLS estimator is inefficient and the conventional covariance matrix will be inappropriate. The random effects estimators of $\boldsymbol{\beta}$ in the center three columns of Table 15.6 are also consistent, by a similar logic. However, they likewise are inefficient. The result at work, which is specific to the linear regression model, is that we are estimating the mean parameters, β_k, and the variance parameters, λ_k and σ_ε, separately. Certainly, if λ_k is nonzero for $k = 2, \ldots, 13$, then the pooled and RE estimators that assume they are zero are all inconsistent. With $\boldsymbol{\beta}$ estimated consistently in an otherwise misspecified model, we would call the MLE and MSLE **pseudo-maximum likelihood estimators**.See Section 14.8.

Comparing the ML and MSL estimators of the random effects model, we find the estimates are similar, though in a few cases, noticeably different nonetheless. The estimates tend to differ most when the estimates themselves have large standard errors (small *t* ratios). This is partly due to the different methods of estimation in a finite sample of 595 observations. We could attribute at least some of the difference to the approximation error in the simulation compared to the exact evaluation of the (closed form) integral in the MLE. The full random parameters model is shown in the last two columns. Based on the likelihood ratio statistic of $2(365.312 - 309.173) = 112.28$ with 12 degrees of freedom, we would reject the hypothesis that $\lambda_2 = \lambda_3 = \cdots = \lambda_{13} = 0$. The 95% critical value with 12 degrees of freedom is 21.03. This random parameters formulation of the model suggests a need to reconsider the notion of *statistical significance* of the estimated parameters. In view of (15-33), it may be the case that the mean parameter might well be significantly different from zero while the corresponding standard deviation, λ, might be large as well, suggesting that a large proportion of the population remains statistically close to zero. Consider the estimate of $\beta_{3,i}$, the coefficient on $South_{it}$. The estimate of the mean, β_3, is 0.04941, with an estimated standard error of 0.02002. This implies a confidence interval for this parameter of $0.04941 \pm 1.96(0.02002) = [0.01017, 0.08865]$. But this is only the location of the center of the distribution. With an estimate of λ_k of 0.20997, the random parameters model suggests that in the population, 95% of individuals have an effect of *South* within $0.04941 \pm 1.96(0.20997) = [-0.36213, 0.46095]$. This is still centered near zero but has a different interpretation from the simple confidence interval for β itself. Most of the population is less than two standard deviations from zero. This analysis suggests that it might be an interesting exercise to estimate β_i rather than just the parameters of the distribution. We will consider that estimation problem in Section 15.10.

The next example examines a random parameters model in which the covariance matrix of the random parameters is allowed to be a free, positive definite matrix. That is,

$$\begin{aligned}&y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it} \\ &\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i,\ E[\mathbf{u}_i|\mathbf{X}] = \mathbf{0},\ \mathrm{Var}[\mathbf{u}_i|\mathbf{X}] = \boldsymbol{\Gamma}.\end{aligned} \qquad \textbf{(15-34)}$$

This is the random effects counterpart to the fixed effects model in Section 11.10.1. Note that the difference in the specifications is the random effects assumption, $E[\mathbf{u}_i|\mathbf{X}] = \mathbf{0}$. We continue to use the Cholesky decomposition of $\mathbf{\Gamma}$ in the reparameterized model

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{\Lambda}\mathbf{w}_i, E[\mathbf{w}_i|\mathbf{X}] = \mathbf{0}, \text{Var}[\mathbf{w}_i|\mathbf{X}] = \mathbf{I}.$$

Example 15.13 Least Simulated Sum of Squares Estimates of a Production Function Model

In Example 11.22, we examined Munnell's production model for gross state product,

$$\ln gsp_{it} = \beta_1 + \beta_2 \ln pc_{it} + \beta_3 \ln hwy_{it} + \beta_4 \ln water_{it} + \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7 unemp_{it} + \varepsilon_{it}, i = 1, \ldots, 48; t = 1, \ldots, 17.$$

The panel consists of state-level data for 17 years. The model in Example 11.19 (and Munnell's) provides no means for parameter heterogeneity save for the constant term. We have reestimated the model using the Hildreth and Houck approach. The OLS, feasible GLS, and maximum likelihood estimates are given in Table 15.7. (The OLS and FGLS results are reproduced from Table 11.21.) The chi-squared statistic for testing the null hypothesis of parameter homogeneity is 25,556.26, with $7(47) = 329$ degrees of freedom. The critical value from the table is 372.299, so the hypothesis would be rejected. Unlike the other cases we have examined in this chapter, the FGLS estimates are very different from OLS in these estimates. The FGLS estimates correspond to a fixed effects view, as they do not assume that the variation in the coefficients is unrelated to the exogenous variables. The underlying standard deviations are computed using **G** as the covariance matrix. [For these

TABLE 15.7 Estimated Random Coefficients Models

	Least Squares		*Feasible GLS*		*Maximum Simulated Likelihood*	
Variable	***Estimate***	***Standard Error***	***Estimate***	***Standard Error***	***Estimate***	***Standard Error***
Constant	1.9260	0.05250	1.6533	1.08331	2.02319	0.03801
					(0.53228)	
ln pc	0.3120	0.01109	0.09409	0.05152	0.32049	0.00621
					(0.15871)	
ln hwy	0.05888	0.01541	0.1050	0.1736	0.01215	0.00909
					(0.19212)	
ln water	0.1186	0.01236	0.07672	0.06743	0.07612	0.00600
					(0.17484)	
ln util	0.00856	0.01235	−0.01489	0.09886	−0.04665	0.00850
					(0.78196)	
ln emp	0.5497	0.01554	0.9190	0.1044	0.67568	0.00984
					(0.82133)	
unemp	−0.00727	0.001384	−0.004706	0.002067	−0.00791	0.00093
					(0.02171)	
σ_ε	0.08542		0.2129		0.02360	
ln L	853.1372				1527.196	

data, subtracting the second matrix rendered **G** not positive definite so, in the table, the standard deviations are based on the estimates using only the first term in (11-88).] The increase in the standard errors is striking. This suggests that there is considerable variation in the parameters across states. We have used (11-89) to compute the estimates of the state-specific coefficients.

The rightmost two columns of Table 15.7 present the maximum simulated likelihood estimates of the random parameters production function model. They somewhat resemble the OLS estimates, more so than the FGLS estimates, which are computed by an entirely different method. The values in parentheses under the parameter estimates are the estimates of the standard deviations of the distribution of $\mathbf{u}_i$, the square roots of the diagonal elements of $\boldsymbol{\Gamma}$. These are obtained by computing the square roots of the diagonal elements of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. The estimate of $\boldsymbol{\Lambda}$ is shown here.

$$\hat{\boldsymbol{\Lambda}} = \begin{bmatrix} 0.53228 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.12511 & 0.09766 & 0 & 0 & 0 & 0 & 0 \\ 0.17529 & -0.07196 & 0.03169 & 0 & 0 & 0 & 0 \\ 0.03467 & 0.03306 & 0.15498 & 0.06522 & 0 & 0 & 0 \\ 0.16413 & -0.03030 & -0.08889 & 0.59745 & 0.46772 & 0 & 0 \\ 0.14750 & -0.02049 & 0.05248 & 0.67429 & 0.44158 & 0.00167 & 0 \\ 0.00427 & -0.00337 & 0.00181 & 0.01640 & 0.01277 & 0.00239 & 0.00083 \end{bmatrix}.$$

An estimate of the correlation matrix for the parameters might also be informative. This is also derived from $\hat{\boldsymbol{\Lambda}}$ by computing $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$ and then transforming the covariances to correlations by dividing by the products of the respective standard deviations (the values in parentheses in Table 15.7). The result is

$$\mathbf{R} = \begin{bmatrix} 1 & & & & & & \\ -0.7883 & 1 & & & & & \\ 0.9124 & -0.9497 & 1 & & & & \\ 0.1983 & -0.0400 & 0.2563 & 1 & & & \\ 0.2099 & -0.1893 & 0.1873 & 0.2186 & 1 & & \\ 0.1796 & -0.1569 & 0.1837 & 0.3938 & 0.9802 & 1 & \\ 0.1966 & -0.2504 & 0.2512 & 0.3654 & 0.9669 & 0.9812 & 1 \end{bmatrix}.$$

15.8 HIERARCHICAL LINEAR MODELS

Example 11.23 examined an application of a two-level model, or hierarchical model, for mortgage rates,

$$RM_{it} = \beta_{1i} + \beta_{2,i}J_{it} + \text{various terms relating to the mortgage} + \varepsilon_{it}.$$

The second-level equation is

$$\beta_{2,i} = \alpha_1 + \alpha_2 \text{GFA}_i + \alpha_3 \text{ one-year treasury rate} + \alpha_4 \text{ ten-year treasury rate} + \alpha_5 \text{ credit risk} + \alpha_6 \text{ prepayment risk} + \cdots + u_i.$$

Recent research in many fields has extended the idea of hierarchical modeling to the full set of parameters in the model. (Depending on the field studied, the reader may find

these labeled *hierarchical models*, **mixed models**, *random parameters models,* or *random effects models*. The last of these generalizes our notion of random effects.) A two-level formulation of the model in (15-34) might appear as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it},$$
$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{u}_i.$$

(A three-level model is shown in Example 15.14.) This model retains the earlier stochastic specification but adds the measurement equation to the generation of the random parameters. model of the previous section now becomes

$$y_{it} = \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Lambda}\mathbf{w}_i) + \varepsilon_{it},$$

which is essentially the same as our earlier model in (15-28) to (15-31) with the addition of the product (interaction) terms of the form $\delta_{kl}x_{itk}z_{il}$, which suggests how it might be estimated (simply by adding the interaction terms to the previous formulation). In the template in (15-26), the term $\sigma_u w_{ir}$ becomes $\mathbf{x}_{it}'(\boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Lambda}\mathbf{w}_i)$, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\delta}', \boldsymbol{\lambda}', \sigma_\varepsilon)'$, where $\boldsymbol{\delta}'$ is a row vector composed of the rows of $\boldsymbol{\Delta}$, and $\boldsymbol{\lambda}'$ is a row vector composed of the rows of $\boldsymbol{\Lambda}$. The scalar term w_{ir} in the derivatives is replaced by a column vector of terms contained in $(\mathbf{x}_{it} \otimes \mathbf{z}_i, \mathbf{x}_{it} \otimes \mathbf{w}_{ir})$.

The hierarchical model can be extended in several useful directions. Recent analyses have expanded the model to accommodate multilevel stratification in data sets such as those we considered in the treatment of nested random effects in Section 14.9.6.b. A three-level model would appear as in the next example that relates to home sales,

$$\begin{aligned} &y_{ijt} = \mathbf{x}_{ijt}'\boldsymbol{\beta}_{ij} + \varepsilon_{it}, t = site, j = neighborhood, i = community,\\ &\boldsymbol{\beta}_{ij} = \boldsymbol{\phi}_i + \boldsymbol{\Delta}\mathbf{z}_{ij} + \mathbf{u}_{ij}\\ &\boldsymbol{\phi}_i = \boldsymbol{\pi} + \boldsymbol{\Phi}\mathbf{r}_i + \mathbf{v}_i. \end{aligned} \qquad \textbf{(15-35)}$$

Example 15.14 Hierarchical Linear Model of Home Prices

Beron, Murdoch, and Thayer (1999) used a hedonic pricing model to analyze the sale prices of 76,343 homes in four California counties: Los Angeles, San Bernardino, Riverside, and Orange. The data set is stratified into 2,185 census tracts and 131 school districts. Home prices are modeled using a three-level random parameters pricing model. (We will change their notation somewhat to make roles of the components of the model more obvious.) Let *site* denote the specific location (sale), *nei* denote the neighborhood, and *com* denote the community, the highest level of aggregation. The pricing equation is

$$\ln Price_{site,\,nei,\,com} = \pi^0_{nei,\,com} + \sum_{k=1}^{K}\pi^k_{nei,\,com}x_{k,site,nei,com} + \varepsilon_{site,nei,com},$$
$$\pi^k_{nei,\,com} = \beta^{0,k}_{com} + \sum_{l=1}^{L}\beta^{l,k}_{com}z_{k,nei,\,com} + r^k_{nei,com},\ k = 0,\ \ldots,K,$$
$$\beta^{l,k}_{com} = \gamma^{0,l,k} + \sum_{m=1}^{M}\gamma^{m,l,k}e_{m,com} + u^{l,k}_{com},\ l = 1,\ \ldots,L.$$

There are K level-one variables, x_k, and a constant in the main equation, L level-two variables, z_l, and a constant in the second-level equations, and M level-three variables, e_m, and a constant in the third-level equations. The variables in the model are as follows. The level-one variables define the hedonic pricing model,

x = house size, number of bathrooms, lot size, presence of central heating, presence of air conditioning, presence of a pool, quality of the view, age of the house, distance to the nearest beach.

Levels two and three are measured at the neighborhood and community levels,

z = percentage of the neighborhood below the poverty line, racial makeup of the neighborhood, percentage of residents over 65, average time to travel to work

and

e = FBI crime index, average achievement test score in school district, air quality measure, visibility index.

The model is estimated by maximum simulated likelihood.

The **hierarchical linear model** analyzed in this section is also called a *mixed model* and *random parameters model*. Although the three terms are usually used interchangeably, each highlights a different aspect of the structural model in (15-35). The hierarchical aspect of the model refers to the layering of coefficients that is built into stratified and panel data structures, such as in Example 15.4. The random parameters feature is a signature feature of the model that relates to the modeling of heterogeneity across units in the sample. Note that the model in (15-35) and Beron et al.'s application could be formulated without the random terms in the lower-level equations. This would then provide a convenient way to introduce interactions of variables in the linear regression model. The addition of the random component is motivated on precisely the same basis that u_i appears in the familiar random effects model in Section 11.5 and (15-39). It is important to bear in mind, in all these structures, strict mean independence is maintained between $\mathbf{u}_i$ and all other variables in the model. In most treatments, we go yet a step further and assume a particular distribution for u_i, typically joint normal. Finally, the mixed model aspect of the specification relates to the underlying integration that removes the heterogeneity, for example, in (15-13). The unconditional estimated model is a mixture of the underlying models, where the weights in the mixture are provided by the underlying density of the random component.

15.9 NONLINEAR RANDOM PARAMETER MODELS

Most of the preceding applications have used the linear regression model to illustrate and demonstrate the procedures. However, the template used to build the model has no intrinsic features that limit it to the linear regression. The initial description of the model and the first example were applied to a nonlinear model, the Poisson regression. We will examine a random parameters binary choice model in the next section as well. This random parameters model has been used in a wide variety of settings. One of the most common is the multinomial choice models that we will discuss in Chapter 18.

The simulation-based random parameters estimator/model is extremely flexible.[24] The simulation method, in addition to extending the reach of a wide variety of model classes, also allows great flexibility in terms of the model itself. For example, constraining a parameter to have only one sign is a perennial issue. Use of a lognormal specification of the parameter, $\beta_i = \exp(\beta + \sigma w_i)$, provides one method of restricting a random

[24]See Train and McFadden (2000) for discussion.

parameter to be consistent with a theoretical restriction. Researchers often find that the lognormal distribution produces unrealistically large values of the parameter. A model with parameters that vary in a restricted range that has found use is the random variable with symmetric about zero triangular distribution,

$$f(w) = \mathbf{1}[-a \leq w \leq 0](a + w)/a^2 + \mathbf{1}[0 < w \leq a](a - w)/a^2.$$

A draw from this distribution with $a = 1$ can be computed as

$$w = \mathbf{1}[u \leq .5][(2u)^{1/2} - 1] + \mathbf{1}[u > .5][1 - (2(1 - u))^{1/2}],$$

where u is the $U[0, 1]$ draw. Then, the parameter restricted to the range $\beta \pm \lambda$ is obtained as $\beta + \lambda w$. A further refinement to restrict the sign of the random coefficient is to force $\lambda = \beta$, so that β_i ranges from 0 to 2λ.[25] There is a large variety of methods for simulation that allow the model to be extended beyond the linear model and beyond the simple normal distribution for the random parameters.

Random parameters models have been implemented in several contemporary computer packages. The PROC MIXED package of routines in SAS uses a kind of generalized least squares for linear, Poisson, and binary choice models. The GLAMM program—Rabe-Hesketh, Skrondal, and Pickles (2005)—written for *Stata* uses quadrature methods for several models including linear, Poisson, and binary choice. The RPM and RPL procedures in *LIMDEP*/*NLOGIT* use the methods described here for linear, binary choice, censored data, multinomial, ordered choice, and several others. Finally, the *MLWin* package (www.bristol.ac.uk/cmm/software/mlwin/) is a large implementation of some of the models discussed here. *MLWin* uses MCMC methods with noninformative priors to carry out maximum simulated likelihood estimation.

15.10 INDIVIDUAL PARAMETER ESTIMATES

In our analysis of the various random parameters specifications, we have focused on estimation of the population parameters $\boldsymbol{\beta}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Lambda}$ in the model,

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Lambda}\mathbf{w}_i,$$

for example, in Example 15.13, where we estimated a model of production. At a few points, it is noted that it might be useful to estimate the individual specific $\boldsymbol{\beta}_i$. We did a similar exercise in analyzing the Hildreth/Houck/Swamy model in Example 11.19 in Section 11.11.1. The model is

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta} + \mathbf{u}_i,\end{aligned}$$

where no restriction is placed on the correlation between $\mathbf{u}_i$ and $\mathbf{X}_i$. In this fixed effects case, we obtained a feasible GLS estimator for the population mean, $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \hat{\mathbf{W}}_i \mathbf{b}_i,$$

where

$$\hat{\mathbf{W}}_i = \left\{ \sum_{i=1}^{n} [\hat{\boldsymbol{\Gamma}} + \hat{\sigma}_\varepsilon^2 (\mathbf{X}_i'\mathbf{X}_i)^{-1}]^{-1} \right\}^{-1} [\hat{\boldsymbol{\Gamma}} + \hat{\sigma}_\varepsilon^2 (\mathbf{X}_i'\mathbf{X}_i)^{-1}]^{-1}$$

[25] Discussion of this sort of model construction is given in Train and Sonnier (2003) and Train (2009).

and

$$\mathbf{b}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i.$$

For each group, we then proposed an estimator of $E[\boldsymbol{\beta}_i|\text{information in hand about group } i]$ as

$$\text{Est. } E[\boldsymbol{\beta}_i|\mathbf{y}_i, \mathbf{X}_i] = \mathbf{b}_i + \hat{\mathbf{Q}}_i(\hat{\boldsymbol{\beta}} - \mathbf{b}_i),$$

where

$$\hat{\mathbf{Q}}_i = \{[s_i^2(\mathbf{X}_i'\mathbf{X}_i)]^{-1} + \hat{\boldsymbol{\Gamma}}^{-1}\}^{-1}\hat{\boldsymbol{\Gamma}}^{-1}. \tag{15-36}$$

The estimator of $E[\boldsymbol{\beta}_i|\mathbf{y}_i, \mathbf{X}_i]$ is equal to the least squares estimator plus a proportion of the difference between $\hat{\boldsymbol{\beta}}$ and $\mathbf{b}_i$. (The matrix $\hat{\mathbf{Q}}_i$ is between $\mathbf{0}$ and $\mathbf{I}$. If there were a single column in $\mathbf{X}_i$, then $\hat{q}_i$ would equal $(1/\hat{\gamma})/\{(1/\hat{\gamma}) + [1/(s_i^2/\mathbf{x}_i'\mathbf{x}_i)]\}$.)

We can obtain an analogous result for the mixed models we have examined in this chapter.[26] From the initial model assumption, we have

$$f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i, \boldsymbol{\theta}),$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Lambda}\mathbf{w}_i \tag{15-37}$$

and $\boldsymbol{\theta}$ is any other parameters in the model, such as σ_ε in the linear regression model. For a panel, because we are conditioning on $\boldsymbol{\beta}_i$, that is, on $\mathbf{w}_i$, the T_i observations are independent, and it follows that

$$f(y_{i1}, y_{i2}, \ldots, y_{iTi}|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) = f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) = \Pi_t f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i, \boldsymbol{\theta}). \tag{15-38}$$

This is the contribution of group i to the likelihood function (not its log) for the sample, given $\boldsymbol{\beta}_i$; that is, note that the log of this term is what appears in the simulated log-likelihood function in (15-31) for the normal linear model and in (15-16) for the Poisson model. The marginal density for $\boldsymbol{\beta}_i$ is induced by the density of $\mathbf{w}_i$ in (15-37). For example, if $\mathbf{w}_i$ is joint normally distributed, then $f(\boldsymbol{\beta}_i) = N[\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i, \boldsymbol{\Lambda}\boldsymbol{\Lambda}']$. As we noted earlier in Section 15.9, some other distribution might apply. Write this generically as the marginal density of $\boldsymbol{\beta}_i$, $f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is the parameters of the underlying distribution of $\boldsymbol{\beta}_i$, for example $(\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Lambda})$ in (15-37). Then, the joint distribution of $\mathbf{y}_i$ and $\boldsymbol{\beta}_i$ is

$$f(\mathbf{y}_i, \boldsymbol{\beta}_i|\mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}) = f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta})f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega}).$$

We will now use Bayes' theorem to obtain $f(\boldsymbol{\beta}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega})$:

$$\begin{aligned} f(\boldsymbol{\beta}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}) &= \frac{f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta})f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega})}{f(\mathbf{y}_i|\mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega})} \\ &= \frac{f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta})f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega})}{\int_{\boldsymbol{\beta}_i} f(\mathbf{y}_i, \boldsymbol{\beta}_i|\mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega})d\boldsymbol{\beta}_i} \\ &= \frac{f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta})f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega})}{\int_{\boldsymbol{\beta}_i} f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta})f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega})d\boldsymbol{\beta}_i}. \end{aligned}$$

[26] See Revelt and Train (2000) and Train (2009).

The denominator of this ratio is the integral of the term that appears in the log-likelihood conditional on $\boldsymbol{\beta}_i$. We will return momentarily to computation of the integral. We now have the conditional distribution of $\boldsymbol{\beta}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}$. The conditional expectation of $\boldsymbol{\beta}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}$ is

$$E[\boldsymbol{\beta}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] = \frac{\int_{\boldsymbol{\beta}_i} \boldsymbol{\beta}_i f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega})}{\int_{\boldsymbol{\beta}_i} f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega}) d\boldsymbol{\beta}_i}.$$

Neither of these integrals will exist in closed form. However, using the methods already developed in this chapter, we can compute them by simulation. The simulation estimator will be

$$\begin{aligned}\text{Est.}E[\boldsymbol{\beta}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] &= \frac{(1/R)\sum_{r=1}^{R} \hat{\boldsymbol{\beta}}_{ir} \prod_{t=1}^{T_i} f(y_{it}|\mathbf{x}_{it}, \hat{\boldsymbol{\beta}}_{ir}, \hat{\boldsymbol{\theta}})}{(1/R)\sum_{r=1}^{R} \prod_{t=1}^{T_i} f(y_{it}|\mathbf{x}_{it}, \hat{\boldsymbol{\beta}}_{ir}, \hat{\boldsymbol{\theta}})} \\ &= \sum_{r=1}^{R} \hat{Q}_{ir} \hat{\boldsymbol{\beta}}_{ir},\end{aligned} \tag{15-39}$$

where $\hat{Q}_{ir}$ is defined in (15-20), (15-21), and

$$\hat{\boldsymbol{\beta}}_{ir} = \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Delta}}\mathbf{z}_i + \hat{\boldsymbol{\Lambda}}\mathbf{w}_{ir}.$$

This can be computed after the estimation of the population parameters. (It may be more efficient to do this computation during the iterations because everything needed to do the calculation will be in place and available while the iterations are proceeding.) For example, for the random parameters linear model, we will use

$$f(y_{it}|\mathbf{x}_{it}, \hat{\boldsymbol{\beta}}_{ir}, \hat{\boldsymbol{\theta}}) = \frac{1}{\hat{\sigma}_\varepsilon\sqrt{2\pi}} \exp\left[-\frac{(y_{it} - \mathbf{x}_{it}'(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Delta}}\mathbf{z}_i + \hat{\boldsymbol{\Lambda}}\mathbf{w}_{ir}))^2}{2\hat{\sigma}_\varepsilon^2}\right]. \tag{15-40}$$

We can also estimate the conditional variance of $\boldsymbol{\beta}_i$ by estimating first, one element at a time, $E[\beta_{i,k}^2|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}]$, then, again, one element at a time,

$$\begin{aligned}\text{Est.Var}[\beta_{\text{i, k}}|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] = &\{\text{Est.}E[\beta_{i,k}^2|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}]\} - \\ &\{\text{Est.}E[\beta_{i,k}|\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}]\}^2.\end{aligned} \tag{15-41}$$

With the estimates of the conditional mean and conditional variance in hand, we can then compute the limits of an interval that resembles a confidence interval as the mean plus and minus two estimated standard deviations. This will construct an interval that contains at least 95% of the conditional distribution of $\boldsymbol{\beta}_i$.

Some aspects worth noting about this computation are as follows:

- The preceding suggested interval is a classical (sampling-theory-based) counterpart to the highest posterior density interval that would be computed for $\boldsymbol{\beta}_i$ for a hierarchical Bayesian estimator.
- The conditional distribution from which $\boldsymbol{\beta}_i$ is drawn might not be symmetric or normal, so a symmetric interval of the mean plus and minus two standard deviations may pick up more or less than 95% of the actual distribution. This is likely to be a

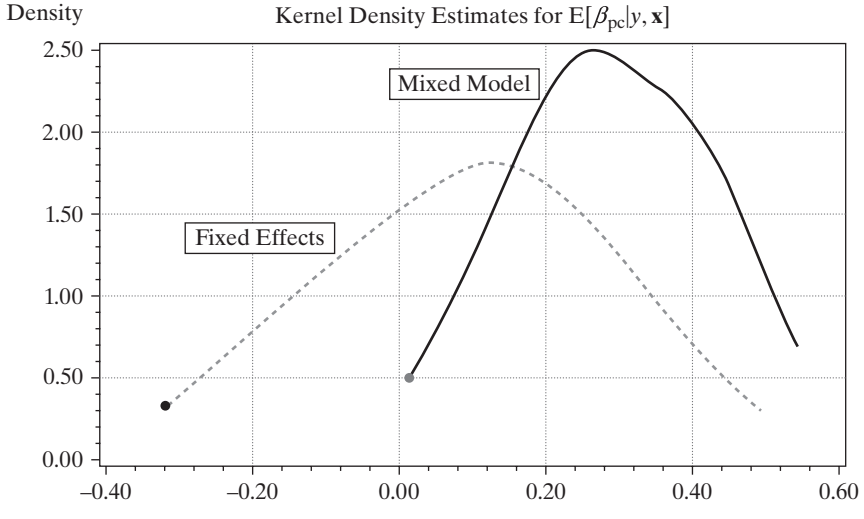
small effect. In any event, in any population, whether symmetric or not, the mean plus and minus two standard deviations will typically encompass at least 95% of the mass of the distribution.

- It has been suggested that this classical interval is too narrow because it does not account for the sampling variability of the parameter estimators used to construct it. But the suggested computation should be viewed as a *point estimate* of the interval, not an interval estimate as such. Accounting for the sampling variability of the estimators might well suggest that the endpoints of the interval should be somewhat farther apart. The Bayesian interval that produces the same estimation would be narrower because the estimator is posterior to, that is, applies only to the sample data.
- Perhaps surprisingly so, even if the analysis departs from normal marginal distributions $\boldsymbol{\beta}_i$, the sample distribution of the n estimated conditional means is not necessarily normal. Kernel estimators based on the n estimators, for example, can have a variety of shapes.
- A common misperception found in the Bayesian and classical literatures alike is that the preceding produces an estimator of $\boldsymbol{\beta}_i$. In fact, it is an estimator of conditional mean of the distribution from which $\boldsymbol{\beta}_i$ is an observation. By construction, for example, every individual with the same $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i)$ has the same prediction even though the $\mathbf{w}_i$ and any other stochastic elements of the model, such as ε_i, will differ across individuals.

Example 15.15 Individual State Estimates of a Private Capital Coefficient

Example 15.13 presents feasible GLS and maximum simulated likelihood estimates of Munnell's state production model. We have computed the estimates of $E[\beta_{2i} | \mathbf{y}_i, \mathbf{X}_i]$ for the 48 states in the sample using (15-36) for the fixed effects estimates and (15-39) for the random effects estimates. Figure 15.6 examines the estimated coefficients for private capital. Figure 15.6 displays kernel density estimates for the population distributions based on the fixed and random effects

FIGURE 15.6 Kernel Density Estimates of Parameter Distributions.



estimates computed using (15-36) and (15-39). The much narrower distribution corresponds to the random effects estimates. The substantial overall difference of the distributions is presumably due in large part to the difference between the fixed effects and random effects assumptions. One might suspect on this basis that the random effects assumption is restrictive.

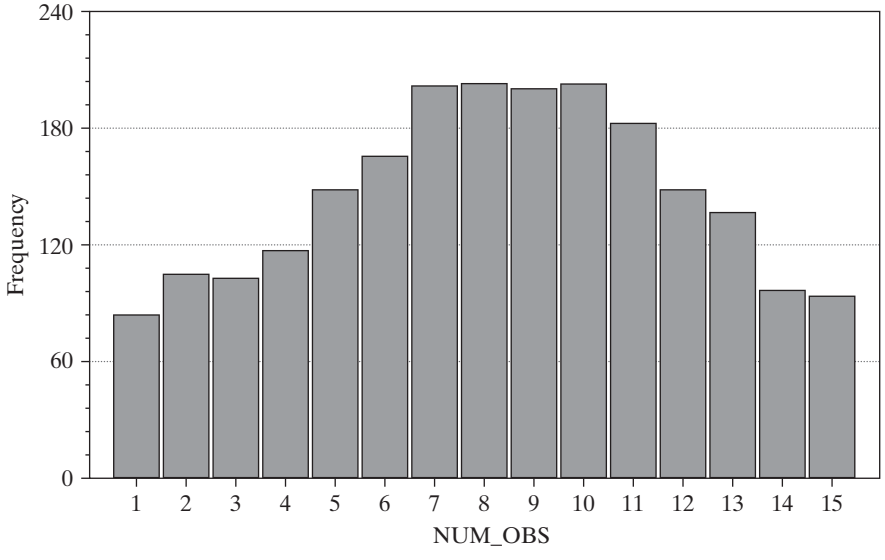
Example 15.16 Mixed Linear Model for Wages

Koop and Tobias (2004) analyzed a panel of 17,919 observations in their study of the relationship between wages and education, ability, and family characteristics. (See the end of chapter applications in Chapters 3, 5, and 11 and Appendix Table F3.2 for details on the location of the data.) The variables used in the analysis are:

		Mean	*Reported mean*
Person id	(time invariant)		
Education	(time varying)	12.68	12.68
Log of hourly wage	(time varying)	2.297	2.30
Potential experience	(time varying)	8.363	8.36
Time trend	(time varying)		
Ability	(time invariant)	0.0524	0.239
Mother's education	(time invariant)	11.47	12.56
Father's education	(time invariant)	11.71	13.17
Broken home dummy	(time invariant)	0.153	0.157
Number of siblings	(time invariant)	3.156	2.83

This is an unbalanced panel of 2,178 individuals. The means in the list are computed from the sample data. The authors report the second set of means based on a subsample of 14,170 observations whose parents have at least 9 years of education. Figure 15.7 shows a frequency count of the numbers of observations in the sample.
We will estimate the following hierarchical wage model:

FIGURE 15.7 Group Sizes for Wage Data Panel.



$$\ln Wage_{it} = \beta_{1,i} + \beta_{2,i}\, Education_{it} + \beta_3\, Experience_{it} + \beta_4\, Experience_{it}^2 + \beta_5\, Broken\ Home_i + \beta_6\, Siblings_i + \varepsilon_{it},$$

$$\beta_{1,i} = \alpha_{1,1} + \alpha_{1,2}\, Ability_i + \alpha_{1,3}\, Mother's\ education_i + \alpha_{1,4}\, Father's\ education_i + u_{1,i},$$

$$\beta_{2,i} = \exp(\alpha_{2,1} + \alpha_{2,2}\, Ability_i + \alpha_{2,3}\, Mother's\ education_i + \alpha_{2,4}\, Father's\ education_i + u_{2,i}).$$

We anticipate that the education effect will be nonnegative for everyone in the population, so we have built that effect into the model by using a lognormal specification for this coefficient. Estimates are computed using the maximum simulated likelihood method described in Sections 15.6.3 and 15.7. Estimates of the model parameters appear in Table 15.8. The four models in Table 15.8 are the pooled OLS estimates, the random effects model, and the random parameters models, first assuming that the random parameters are uncorrelated ($\Gamma_{21} = 0$) and then allowing free correlation ($\Gamma_{21} =$ nonzero). The differences between the conventional and the robust standard errors in the pooled model are fairly large, which suggests the presence of latent common effects. The formal estimates of the random effects model confirm this. There are only minor differences between the FGLS and the ML estimates of the random effects model. But the hypothesis of the pooled model is decisively rejected by the likelihood ratio test. The LM statistic [Section 11.5.5 and (11-42)] is 19,353.51, which is far larger than the critical value of 3.84. So, the hypothesis of the pooled model is firmly rejected. The likelihood ratio statistic based on the MLEs is 2(12300.51 − 8013.43) = 8,574.16, which produces the same conclusion. An alternative approach would be to test the hypothesis that $\sigma_u^2 = 0$ using a Wald statistic—the standard t test. The software used for this exercise reparameterizes the log likelihood in terms of $\theta_1 = \sigma_u^2/\sigma_\varepsilon^2$ and $\theta_2 = 1/\sigma_\varepsilon^2$. One approach, based on the delta method (see Section 4.4.4), would be to estimate σ_u^2 with the MLE of θ_1/θ_2. The

TABLE 15.8 Estimated Random Parameter Models

Variable	*Pooled OLS*	*RE/FGLS*	*RE/MLE*	*RE/MSL*	*Random Parameters*
Exp	0.09089	0.10272	0.10289	0.10277	0.10531
	(0.00431)	(0.00260)	(0.00261)	(0.00165)	(0.00165)
Exp^2	−0.00305	−0.00363	−0.00364	−0.00364	−0.00375
	(0.00025)	(0.00014)	(0.00014)	(0.000093)	(0.000093)
Broken Home	−0.05603	−0.06328	−0.06360	−0.05675	−0.04816
	(0.02178)	(0.02171)	(0.02252)	(0.00667)	(0.00665)
Siblings	−0.00202	−0.00664	−0.00675	−0.00841	−0.00125
	(0.00407)	(0.00384)	(0.00398)	(0.00116)	(0.00121)
Constant	0.69271	0.60995	0.61223	0.60346	*
	(0.05876)	(0.04665)	(0.04781)	(0.01744)	*
Education	0.08869	0.08954	0.08929	0.08982	*
	(0.00433)	(0.00337)	(0.00346)	(0.00123)	*
σ_ε	0.48079	0.328699	0.32913	0.32979	0.32949
σ_u	0.00000	0.350882	0.036580	0.37922	*
LM	19353.51				
ln *L*	−12300.51446		−8013.43044	−8042.97734	−7983.57355

* **Random Parameters**

$\hat{\beta}_{1,i} = 0.83417 + 0.02870\, Ability_i - 0.01355\, Mother's\ Ed_i + 0.00878\, Father's\ Ed_i + 0.30857 u_{1,i}$
(.04952) (0.01304) (0.00463) (0.00372)

$\hat{\beta}_{2,i} = \exp[-2.78412 + 0.05680\, Ability_i + 0.01960\, Mother's\ Ed_i - 0.00370\, Father's\ Ed_i + 0.10178\, u_{2,i}]$
(.05582) (0.01505) (0.00503) (0.00388)

asymptotic variance of this estimator would be estimated using Theorem 4.5. Alternatively, we might note that σ_ε^2 must be positive in this model, so it is sufficient simply to test the hypothesis that $\theta_1 = 0$. Our MLE of θ_1 is 9.23137 and the estimated asymptotic standard error is 0.10427. Following this logic, then, the test statistic is 88.57. This is far larger than the critical value of 1.96, so, once again, the hypothesis is rejected. We do note a problem with the LR and Wald tests: The hypothesis that $\sigma_u^2 = 0$ produces a nonstandard test under the null hypothesis because $\sigma_u^2 = 0$ is on the boundary of the parameter space. Our standard theory for likelihood ratio testing (see Chapter 14) requires the restricted parameters to be in the interior of the parameter space, not on the edge. The distribution of the test statistic under the null hypothesis is not the familiar chi squared.[27] The simple expedient in this complex situation is to use the LM statistic, which remains consistent with the earlier conclusion.

The fifth model in Table 15.8 presents the mixed model estimates. The mixed model allows Λ_{21} to be a free parameter. The implied estimators for σ_{u1}, σ_{u2}, and $\sigma_{u,21}$ are the elements of $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$, where $\hat{\mathbf{\Lambda}} = \begin{bmatrix} 0.30857 & 0.00000 \\ -0.06221 & 0.08056 \end{bmatrix}$. Then, $\hat{\sigma}_{u1} = \sqrt{\hat{\Lambda}_{11}^2} = 0.30857$ and $\hat{\sigma}_{u2} = \sqrt{\hat{\Lambda}_{21}^2 + \hat{\Lambda}_{22}^2} = 0.10178$.

Note that for both random parameters models, the estimate of σ_ε is relatively unchanged. The models decompose the variation across groups in the parameters differently, but the overall variation of the dependent variable is largely the same.

The interesting coefficient in the model is $\beta_{2,i}$. The coefficient on education in the model is $\beta_{2,i} = \exp(\alpha_{2,1} + \alpha_{2,2}\,\textit{Ability} + \alpha_{2,3}\,\textit{Mother's education} + \alpha_{2,4}\,\textit{Father's education} + u_{2,i})$. The raw coefficients are difficult to interpret. The expected value of β_{2i} equals $\exp(\boldsymbol{\alpha}_2'\mathbf{z}_i + \sigma_{u2}^2/2)$. The sample means for the three variables are 0.052374, 11.4719, and 11.7092, respectively. With these values, and $\sigma_{u2} = 0.10178$, the population mean value for the education coefficient is approximately 0.0727, which is in line with expectations. This is comparable to, though somewhat smaller than, the estimates for the pooled and random effects model. Of course, variation in this parameter across the sample individuals was the objective of this specification. Figure 15.8 plots a kernel density estimate for the estimated conditional means for the 2,178 sample individuals. The figure shows the range of variation in the sample estimates.

The authors of this study used Bayesian methods, but a very similar specification to ours to study heterogeneity in the returns to education. They proposed several specifications, including a latent class approach that we will consider momentarily. Their *massively* preferred specification[28] is similar to the one we used in our random parameters specification,

$$\begin{aligned} \ln \textit{Wage}_{it} &= \theta_{1,i} + \theta_{2,i}\,\textit{Education}_{it} + \boldsymbol{\gamma}'\mathbf{z}_{it} + \varepsilon_{it}, \\ \theta_{1,i} &= \theta_{1,0} + u_{1,i}, \\ \theta_{2,i} &= \theta_{2,0} + u_{2,i}. \end{aligned}$$

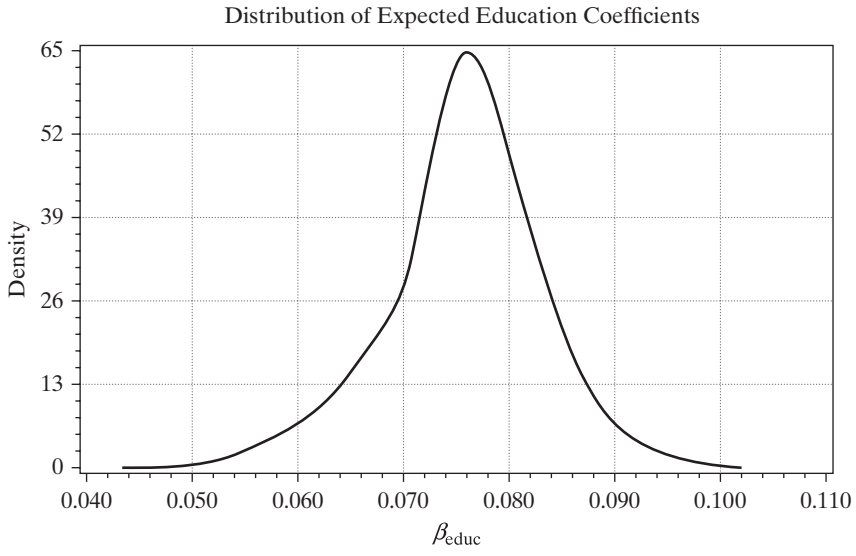
Among the preferred alternatives in their specification is a Heckman and Singer (1984) style (Section 14.15.7) latent class model with 10 classes. The specification would be

$$\begin{aligned} &\ln(\textit{Wage}_{it} \,|\, \textit{class} = j) = \theta_{1,j} + \theta_{2,j}\,\textit{Education}_{it} + \gamma' z_{it} + \varepsilon_{it}, \\ &\text{Prob}(\textit{class} = j) = \pi_j,\ j = 1, \ldots, 10. \end{aligned}$$

[27]This issue is confronted in Breusch and Pagan (1980) and Godfrey (1988) and analyzed at (great) length by Andrews (1998, 1999, 2000, 2001, 2002) and Andrews and Ploberger (1994, 1995).

[28]The model selection criterion used is the Bayesian information criterion, $2\ln f(\textbf{data}\,|\,\textbf{parameters}) - K \ln n$, where the first term would be the posterior density for the data, K is the number of parameters in the model, and n is the sample size. For frequentist methods such as those we use here, the first term would be twice the log likelihood. The authors report a BIC of −16,528 for their preferred model. The log likelihood for the 5 class latent class model reported below is −8053.676. With 22 free parameters (8 common parameters in the regression + 5(θ_1 and θ_2) + 4 free class probabilities), the BIC for our model is −16,275.45.

FIGURE 15.8 Kernel Density Estimate for Education Coefficient.



We fit this alternative model to explore the sensitivity of the returns coefficient to the specification. With 10 classes, the frequentist approach converged, but several of the classes were estimated to be extremely small—on the order of 0.1% of the population, and these segments produced nonsense values of θ_2 such as -5.0. Results for a finite mixture model with 5 classes are as follows (the other model coefficients are omitted):

Class	θ_{Ed}	π
1	0.09447	0.32211
2	0.05354	0.03644
3	0.09988	0.09619
4	0.07155	0.33285
5	0.05677	0.21241

The weighted average of these results is 0.07789. The numerous estimates of the returns to education computed in this example are in line with other studies, in this paper, elsewhere in the book, and in other studies. What we have found here is that the estimated returns, for example, by OLS in Table 15.8, are a bit lower when the model accounts for heterogeneity in the population.

15.11 MIXED MODELS AND LATENT CLASS MODELS

Sections 15.7 through 15.10 examined different approaches to modeling parameter heterogeneity. The fixed effects approach begun in Section 11.4 is extended to include the full set of regression coefficients in Section 11.10.1 where

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,$$
$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i,$$

and no restriction is placed on $E[\mathbf{u}_i|\mathbf{X}_i]$. Estimation produces a feasible GLS estimate of $\boldsymbol{\beta}$. Estimation of $\boldsymbol{\beta}$ begins with separate least squares estimation with each group, i—because of the correlation between $\mathbf{u}_i$ and $\mathbf{x}_{it}$, the pooled estimator is not consistent. The efficient estimator of $\boldsymbol{\beta}$ is then a mixture of the $\mathbf{b}_i$'s. We also examined an estimator of $\boldsymbol{\beta}_i$, using the optimal predictor from the conditional distributions, (15-39). The crucial assumption underlying the analysis is the possible correlation between $\mathbf{X}_i$ and $\mathbf{u}_i$. We also considered two modifications of this random coefficients model. First, a restriction of the model in which some coefficients are nonrandom provides a useful simplification. The familiar fixed effects model of Section 11.4 is such a case, in which only the constant term varies across individuals. Second, we considered a hierarchical form of the model

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{u}_i. \tag{15-42}$$

This approach is applied to an analysis of mortgage rates in Example 11.23.

A second approach to random parameters modeling builds from the crucial assumption added to (15-42) that $\mathbf{u}_i$ and $\mathbf{X}_i$ are uncorrelated. The general model is defined in terms of the conditional density of the random variable, $f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i, \boldsymbol{\theta})$, and the marginal density of the random coefficients, $f(\boldsymbol{\beta}_i|\mathbf{z}_i, \boldsymbol{\Omega})$, in which $\boldsymbol{\Omega}$ is the separate parameters of this distribution. This leads to the mixed models examined in this chapter. The random effects model that we examined in Section 11.5 and several other points is a special case in which only the constant term is random (like the fixed effects model). We also considered the specific case in which u_i is distributed normally with variance σ_u^2.

A third approach to modeling heterogeneity in parametric models is to use a discrete distribution, either as an approximation to an underlying continuous distribution, or as the model of the data-generating process in its own right. (See Section 14.15.) This model adds to the preceding a nonparametric specification of the variation in $\boldsymbol{\beta}_i$,

$$\text{Prob}(\boldsymbol{\beta}_i = \boldsymbol{\beta}_j|\mathbf{z}_i) = \pi_{ij}, j = 1, \ldots, J.$$

A somewhat richer, semiparametric form that mimics (15-42) is

$$\text{Prob}(\boldsymbol{\beta}_i = \boldsymbol{\beta}_j|\mathbf{z}_i) = \pi_j(\mathbf{z}_i, \boldsymbol{\Omega}), j = 1, \ldots, J.$$

We continue to assume that the process generating variation in $\boldsymbol{\beta}_i$ across individuals is independent of the process that produces $\mathbf{X}_i$—that is, in a broad sense, we retain the random effects approach. In the last example of this chapter, we will examine a comparison of mixed and finite mixture models for a nonlinear model.

Example 15.17 Maximum Simulated Likelihood Estimation of a Binary Choice Model

Bertschek and Lechner (1998) analyzed the innovations of a sample of German manufacturing firms. They used a probit model (Sections 17.2–17.4) to study firm innovations. The model is for $\text{Prob}(y_{it} = 1|\mathbf{x}_{it}, \boldsymbol{\beta}_i)$ where

$y_{it} = 1$ if firm i realized a product innovation in year t and 0 if not.

The independent variables in the model are

$x_{it,1} =$ constant,
$x_{it,2} =$ log of sales,

$x_{it,3}$ = relative size = ratio of employment in business unit to employment in the industry,
$x_{it,4}$ = ratio of industry imports to (industry sales + imports),
$x_{it,5}$ = ratio of industry foreign direct investment to (industry sales + imports),
$x_{it,6}$ = productivity = ratio of industry value added to industry employment,
$x_{it,7}$ = dummy variable indicating firm is in the raw materials sector,
$x_{it,8}$ = dummy variable indicating the firm is in the investment goods sector.

The sample consists of 1,270 German firms observed for five years, 1984–1988. (See Appendix Table F15.1.) The density that enters the log likelihood is

$$f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i) = \text{Prob}[y_{it}|\mathbf{x}_{it}'\boldsymbol{\beta}_i] = \Phi[(2y_{it} - 1)\mathbf{x}_{it}'\boldsymbol{\beta}_i], y_{it} = 0, 1,$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i, \mathbf{v}_i \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

To be consistent with Bertschek and Lechner (1998) we did not fit any firm-specific time-invariant components in the main equation for $\boldsymbol{\beta}_i$. Table 15.9 presents the estimated coefficients for the basic probit model in the first column. These are the values reported in the 1998 study. The estimates of the means, $\boldsymbol{\beta}$, are shown in the second column. There appear to be large differences in the parameter estimates, although this can be misleading as there is large variation across the firms in the posterior estimates. The third column presents the square roots of the implied diagonal elements of $\boldsymbol{\Sigma}$ computed as the diagonal elements of $\mathbf{CC}'$. These estimated standard deviations are for the underlying distribution of the parameter in the model—they are not estimates of the standard deviation of the sampling distribution of the estimator. That is shown for the mean parameter in the second column. The fourth column presents the sample means and standard deviations of the 1,270 estimated conditional estimates of the coefficients.

TABLE 15.9 Estimated Random Parameters Model

	Probit	*RP Mean*	*RP Std. Dev.*	*Empirical Distn.*
Constant	−1.96031	−3.43237	0.44947	−3.42768
	(0.37298)	(0.28187)	(0.02121)	(0.15151)
ln Sales	0.17711	0.31054	0.09014	0.31113
	(0.03580)	(0.02757)	(0.00242)	(0.06206)
Relative Size	1.07274	4.36456	3.91986	4.37532
	(0.26871)	(0.27058)	(0.23881)	(1.03431)
Import	1.13384	1.69975	0.93927	1.70413
	(0.24331)	(0.18440)	(0.07287)	(0.20289)
FDI	2.85318	2.91042	0.93468	2.91600
	(0.64233)	(0.47161)	(0.32610)	(0.15182)
Productivity	−2.34116	−4.05320	2.52542	−4.02747
	(1.11575)	(1.04683)	(0.21665)	(0.54492)
Raw materials	−0.27858	−0.42055	0.34962	−0.41966
	(0.12656)	(0.10694)	(0.06926)	(0.05948)
Investment	0.18796	0.30491	0.04672	0.30477
	(0.06287)	(0.04756)	(0.02812)	(0.00812)
ln L	−4114.05		−3524.66	

TABLE 15.10 Estimated Latent Class Model

	Class 1	*Class 2*	*Class 3*	*Posterior*
Constant	−2.32073	−2.70546	−8.96773	−3.77582
	(0.65898)	(0.73335)	(2.46099)	(2.14253)
ln Sales	0.32265	0.23337	0.57148	0.34283
	(0.06516)	(0.06790)	(0.19448)	(0.08919)
Relative Size	4.37802	0.71974	1.41997	2.57719
	(0.87099)	(0.29163)	(0.71765)	(1.29454)
Import	0.93572	2.25770	3.12177	1.80964
	(0.41140)	(0.50726)	(1.33320)	(0.74348)
FDI	2.19747	2.80487	8.37073	3.63157
	(1.58729)	(1.02824)	(2.09091)	(1.98176)
Productivity	−5.86238	−7.70385	−0.91043	−5.48219
	(1.53051)	(4.10134)	(1.46314)	(1.78348)
Raw Materials	−0.10978	−0.59866	0.85608	−0.07825
	(0.17459)	(0.37942)	(0.40407)	(0.36666)
Investment	0.13072	0.41353	0.46904	0.29184
	(0.11851)	(0.12388)	(0.23876)	(0.12462)
ln L = −3503.55				
Class Prob (Prior)	0.46950	0.33073	0.19977	
	(0.03762)	(0.03407)	(0.02629)	
Class Prob (Posterior)	0.46950	0.33073	0.19976	
	(0.39407)	(0.28906)	(0.32492)	
Pred. Count	649	366	255	

The latent class formulation developed in Section 14.15 provides an alternative approach for modeling latent parameter heterogeneity.[29] To illustrate the specification, we will reestimate the random parameters innovation model using a three-class latent class model. Estimates of the model parameters are presented in Table 15.10. The estimated conditional mean shown, which is comparable to the empirical means in the rightmost column in Table 15.9 for the random parameters model, are the sample average and standard deviation of the 1,270 firm-specific posterior mean parameter vectors. They are computed using $\hat{\boldsymbol{\beta}}_i = \boldsymbol{\Sigma}_{j=1}^{3} \hat{\pi}_{ij} \hat{\boldsymbol{\beta}}_j$, where $\hat{\pi}_{ij}$ is the conditional estimator of the class probabilities in (14-97). These estimates differ considerably from the probit model, but they are quite similar to the empirical means in Table 15.9. In each case, a confidence interval around the posterior mean contains the one-class pooled probit estimator. Finally, the (identical) prior and average of the sample posterior class probabilities are shown at the bottom of the table. The much larger empirical standard deviations reflect that the posterior estimates are based on aggregating the sample data and involve, as well, complicated functions of all the model parameters. The estimated numbers of class members are computed by assigning to each firm the predicted class associated with the highest posterior class probability.

[29]See Greene (2001) for a survey. For two examples, Nagin and Land (1993) employed the model to study age transitions through stages of criminal careers and Wang et al. (1998) and Wedel et al. (1993) used the Poisson regression model to study counts of patents.

15.12 SUMMARY AND CONCLUSIONS

This chapter has outlined several applications of simulation-assisted estimation and inference. The essential ingredient in any of these applications is a random number generator. We examined the most common method of generating what appear to be samples of random draws from a population—in fact, they are deterministic Markov chains that only appear to be random. Random number generators are used directly to obtain draws from the standard uniform distribution. The inverse probability transformation is then used to transform these to draws from other distributions. We examined several major applications involving random sampling:

- Random sampling, in the form of bootstrapping, allows us to infer the characteristics of the sampling distribution of an estimator, in particular its asymptotic variance. We used this result to examine the sampling variance of the median in random sampling from a nonnormal population. Bootstrapping is also a useful, robust method of constructing confidence intervals for parameters.
- Monte Carlo studies are used to examine the behavior of statistics when the precise sampling distribution of the statistic cannot be derived. We examined the behavior of a certain test statistic and of the maximum likelihood estimator in a fixed effects model.
- Many integrals that do not have closed forms can be transformed into expectations of random variables that can be sampled with a random number generator. This produces the technique of Monte Carlo integration. The technique of maximum simulated likelihood estimation allows the researcher to formulate likelihood functions (and other criteria such as moment equations) that involve expectations that can be integrated out of the function using Monte Carlo techniques. We used the method to fit random parameters models.

  The techniques suggested here open up a vast range of applications of Bayesian statistics and econometrics in which the characteristics of a posterior distribution are deduced from random samples from the distribution, rather than brute force derivation of the analytic form. Bayesian methods based on this principle are discussed in Chapter 16.

Key Terms and Concepts

- Antithetic draws
- Block bootstrap
- Cholesky decomposition
- Cholesky factorization
- Direct product
- Discrete uniform distribution
- Fundamental probability transformation
- Gauss–Hermite quadrature
- GHK smooth recursive simulator
- Hadamard product
- Halton draws
- Hierarchical linear model
- Incidental parameters problem
- Kronecker product
- Markov chain
- Mersenne Twister
- Mixed model
- Monte Carlo integration
- Nonparametric bootstrap
- Paired bootstrap
- Parametric bootstrap
- Percentile method
- Period
- Power of a test
- Pseudo maximum likelihood estimator
- Pseudo–random number generator
- Schur product
- Seed
- Simulation
- Size of a test
- Shuffling
- Specificity

Exercises

1. The exponential distribution has density $f(x) = \theta \exp(-\theta x)$. How would you obtain a random sample of observations from an exponential population?
2. The Weibull population has survival function $S(x) = \lambda p \exp(-(\lambda x)p)$. How would you obtain a random sample of observations from a Weibull population? (The survival function equals one minus the cdf.)
3. Derive the first-order conditions for nonlinear least squares estimation of the parameters in (15-2). How would you estimate the asymptotic covariance matrix for your estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$?

Applications

1. Does the Wald statistic reject the null hypothesis too often? Construct a Monte Carlo study of the behavior of the Wald statistic for testing the hypothesis that γ equals zero in the model of Section 15.5.1. Recall that the Wald statistic is the square of the t ratio on the parameter in question. The procedure of the test is to reject the null hypothesis if the Wald statistic is greater than 3.84, the critical value from the chi-squared distribution with one degree of freedom. Replicate the study in Section 15.5.1 that is for all three assumptions about the underlying data.
2. A regression model that describes income as a function of experience is

$$\ln Income_i = \beta_1 + \beta_2 Experience_i + \beta_3 Experience_i^2 + \varepsilon_i.$$

3. The model implies that ln *Income* is largest when $\partial \ln Income / \partial Experience$ equals zero. The value of *Experience* at which this occurs is where $\beta_4 + 2\beta_5 Experience = 0$, or $Experience* = -\beta_2/\beta_3$. Describe how to use the delta method to obtain a confidence interval for *Experience**. Now, describe how to use bootstrapping for this computation. A model of this sort using the Cornwell and Rupert data appears in Example 15.6. Using your proposals here, carry out the computations for that model using the Cornwell and Rupert data.

16
BAYESIAN ESTIMATION AND INFERENCE

16.1 INTRODUCTION

The preceding chapters (and those that follow this one) are focused primarily on parametric specifications and classical estimation methods. These elements of the econometric method present a bit of a methodological dilemma for the researcher. They appear to straightjacket the analyst into a fixed and immutable specification of the model. But in any analysis, there is uncertainty as to the magnitudes, sometimes the signs and, at the extreme, even the meaning of parameters. It is rare that the presentation of a set of empirical results has not been preceded by at least some exploratory analysis. Proponents of the Bayesian methodology argue that the process of *estimation* is not one of deducing the values of fixed parameters, but rather, in accordance with the scientific method, one of continually updating and sharpening our subjective beliefs about the state of the world. Of course, this adherence to a subjective approach to model building is not necessarily a virtue. If one holds that *models* and *parameters* represent objective truths that the analyst seeks to discover, then the subjectivity of Bayesian methods may be less than perfectly comfortable.

Contemporary applications of Bayesian methods typically advance little of this theological debate. The modern practice of Bayesian econometrics is much more pragmatic. As we will see in several of the following examples, Bayesian methods have produced some remarkably efficient solutions to difficult estimation problems. Researchers often choose the techniques on practical grounds, rather than in adherence to their philosophical basis; indeed, for some, the Bayesian estimator is merely an algorithm.[1]

Bayesian methods have have been employed by econometricians since well before Zellner's classic (1971) presentation of the methodology to economists, but until fairly recently, were more or less at the margin of the field. With recent advances in technique (notably the Gibbs sampler) and the advance of computer software and hardware that has made simulation-based estimation routine, Bayesian methods that rely heavily on both have become widespread throughout the social sciences. There are libraries of work on Bayesian econometrics, a rapidly expanding applied literature.[2] This chapter will introduce the vocabulary and techniques of Bayesian econometrics. Section 16.2

[1]For example, the Website of MLWin, a widely used program for random parameters modeling, www.bristol.ac.uk/cmm/software/mlwin/features/mcmc.html, states that their use of diffuse priors for Bayesian models produces approximations to maximum likelihood estimators. Train (2001) is an interesting application that compares Bayesian and classical estimators of a random parameters model. Another comparison appears in Example 16.7 below.

[2]Recent additions to the dozens of books on the subject include Gelman et al. (2004), Geweke (2005), Gill (2002), Koop (2003), Lancaster (2004), Congdon (2005), and Rossi et al. (2005). Readers with a historical bent will find Zellner (1971) and Leamer (1978) worthwhile reading. There are also many methodological surveys. Poirier and Tobias (2006) as well as Poirier (1988, 1995) sharply focus the nature of the methodological distinctions between the classical (frequentist) and Bayesian approaches.

lays out the essential foundation for the method. The canonical application, the linear regression model, is developed in Section 16.3. Section 16.4 continues the methodological development. The fundamental tool of contemporary Bayesian econometrics, the Gibbs sampler, is presented in Section 16.5. Three applications and several more limited examples are presented in Sections 16.6 through 16.8. Section 16.6 shows how to use the Gibbs sampler to estimate the parameters of a probit model without maximizing the likelihood function. This application also introduces the technique of data augmentation. Bayesian counterparts to the panel data random and fixed effects models are presented in Section 16.7. A hierarchical Bayesian treatment of the random parameters model is presented in Section 16.8 with a comparison to the classical treatment of the same model. Some conclusions are drawn in Section 16.9. The presentation here is nontechnical. A much more extensive entry-level presentation is given by Lancaster (2004). Intermediate-level presentations appear in Cameron and Trivedi (2005, Chapter 13), and Koop (2003). A more challenging treatment is offered in Geweke (2005). The other sources listed in footnote 2 are oriented to applications.

16.2 BAYES' THEOREM AND THE POSTERIOR DENSITY

The centerpiece of the Bayesian methodology is **Bayes' theorem**: for events A and B, the conditional probability of event A given that B has occurred is

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{16-1}$$

Paraphrased for our applications here, we would write

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})P(\text{parameters})}{P(\text{data})}.$$

In this setting, the data are viewed as constants whose distributions do not involve the parameters of interest. For the purpose of the study, we treat the data as only a fixed set of additional information to be used in updating our beliefs about the parameters. Note the similarity to (12-1). Thus, we write

$$\begin{aligned} P(\text{parameters} \mid \text{data}) &\propto P(\text{data} \mid \text{parameters})P(\text{parameters}) \\ &= \textbf{Likelihood function} \times \textbf{Prior density}. \end{aligned} \tag{16-2}$$

The symbol $\propto$ means "is proportional to." In the preceding equation, we have dropped the marginal density of the data, so what remains is not a proper density until it is scaled by what will be an inessential proportionality constant. The first term on the right is the joint distribution of the observed random variables **y**, given the parameters. As we shall analyze it here, this distribution is the normal distribution we have used in our previous analysis—see (12-1). The second term is the **prior beliefs** of the analyst. The left-hand side is the **posterior density** of the parameters, given the current body of data, or our revised beliefs about the distribution of the parameters after seeing the data. The posterior is a mixture of the prior information and the current information, that is, the data. Once obtained, this posterior density is available to be the **prior density** function

when the next body of data or other usable information becomes available. The principle involved, which appears nowhere in the classical analysis, is one of continual accretion of knowledge about the parameters.

Traditional Bayesian estimation is heavily parameterized. The prior density and the likelihood function are crucial elements of the analysis, and both must be fully specified for estimation to proceed. The Bayesian estimator is the mean of the posterior density of the parameters, a quantity that is usually obtained either by integration (when closed forms exist), approximation of integrals by numerical techniques, or by Monte Carlo methods, which are discussed in Section 15.6.2.

Example 16.1 Bayesian Estimation of a Probability

Consider estimation of the probability that a production process will produce a defective product. In case 1, suppose the sampling design is to choose $N = 25$ items from the production line and count the number of defectives. If the probability that any item is defective is a constant θ between zero and one, then the likelihood for the sample of data is

$$L(\theta \,|\, \textbf{data}) = \theta^D (1 - \theta)^{25 - D},$$

where D is the number of defectives, say, 8. The maximum likelihood estimator of θ will be $p = D/25 = 0.32$, and the asymptotic variance of the maximum likelihood estimator is estimated by $p(1 - p)/25 = 0.008704$.

Now, consider a Bayesian approach to the same analysis. The posterior density is obtained by the following reasoning:

$$p(\theta \,|\, \textbf{data}) = \frac{p(\theta, \textbf{data})}{p(\textbf{data})} = \frac{p(\theta, \textbf{data})}{\int_\theta p(\theta, \textbf{data}) d\theta} = \frac{p(\textbf{data} \,|\, \theta) p(\theta)}{p(\textbf{data})}$$
$$= \frac{\textit{Likelihood}(\textbf{data} \,|\, \theta) \times p(\theta)}{p(\textbf{data})},$$

where $p(\theta)$ is the prior density assumed for θ. [We have taken some license with the terminology, because the **likelihood function** is conventionally defined as $L(\theta \,|\, \textbf{data})$.] Inserting the results of the sample first drawn, we have the posterior density,

$$p(\theta \,|\, \textbf{data}) = \frac{\theta^D (1 - \theta)^{N - D} p(\theta)}{\int_\theta \theta^D (1 - \theta)^{N - D} p(\theta) d\theta}.$$

What follows depends on the assumed prior for θ. Suppose we begin with a noninformative prior that treats all *allowable* values of θ as equally likely. This would imply a uniform distribution over (0,1). Thus, $p(\theta) = 1$, $0 \leq \theta \leq 1$. The denominator with this assumption is a beta integral (see Section E2.3) with parameters $a = D + 1$ and $b = N - D + 1$, so the posterior density is

$$p(\theta \,|\, \textbf{data}) = \frac{\theta^D (1 - \theta)^{N - D}}{\left(\dfrac{\Gamma(D + 1)\Gamma(N - D + 1)}{\Gamma(D + 1 + N - D + 1)} \right)} = \frac{\Gamma(N + 2) \theta^D (1 - \theta)^{N - D}}{\Gamma(D + 1)\Gamma(N - D + 1)}.$$

This is the density of a random variable with a beta distribution with parameters $(\alpha, \beta) = (D + 1, N - D + 1)$. (See Section B.4.6.) The mean of this random variable is $(D + 1)/(N + 2) = 9/27 = 0.3333$ (as opposed to 0.32, the MLE). The posterior variance is $[(D + 1)/(N - D + 1)]/[(N + 3)(N + 2)^2] = 0.007936$ compared to 0.00874 for the MLE.

There is a loose end in this example. If the uniform prior were truly noninformative, that would mean that the only information we had was in the likelihood function. Why didn't the Bayesian estimator and the MLE coincide? The reason is that the uniform prior over [0,1] is not really noninformative. It did introduce the information that θ must fall in the unit interval. The prior mean is 0.5 and the prior variance is 1/12. The posterior mean is an average of the MLE and the prior mean. Another less than obvious aspect of this result is the smaller variance of the Bayesian estimator. The principle that lies behind this (aside from the fact that the prior did in fact introduce some certainty in the estimator) is that the Bayesian estimator is conditioned on the specific sample data. The theory behind the classical MLE implies that it averages over the entire population that generates the data. This will always introduce a greater degree of uncertainty in the classical estimator compared to its Bayesian counterpart.

16.3 BAYESIAN ANALYSIS OF THE CLASSICAL REGRESSION MODEL

The complexity of the algebra involved in Bayesian analysis is often extremely burdensome. For the linear regression model, however, many fairly straightforward results have been obtained. To provide some of the flavor of the techniques, we present the full derivation only for some simple cases. In the interest of brevity, and to avoid the burden of excessive algebra, we refer the reader to one of the several sources that present the full derivation of the more complex cases.[3]

The classical normal regression model we have analyzed thus far is constructed around the conditional multivariate normal distribution $N[\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}]$. The interpretation is different here. In the sampling theory setting, this distribution embodies the information about the observed sample data given the assumed distribution and the fixed, albeit unknown, parameters of the model. In the Bayesian setting, this function summarizes the information that a particular realization of the data provides about the assumed distribution of the model parameters. To underscore that idea, we rename this joint density the *likelihood for* $\boldsymbol{\beta}$ *and* σ^2 *given the data*, so

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{-[(1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]}. \tag{16-3}$$

For purposes of the following results, some reformulation is useful. Let $d = n - K$ (the degrees of freedom parameter), and substitute

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\boldsymbol{\beta} - \mathbf{b})$$

in the exponent. Expanding this produces

$$\left(-\frac{1}{2\sigma^2}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \left(-\frac{1}{2}ds^2\right)\left(\frac{1}{\sigma^2}\right) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)(\boldsymbol{\beta} - \mathbf{b}).$$

After a bit of manipulation (note that $n/2 = d/2 + K/2$), the likelihood may be written

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) = [2\pi]^{-d/2}[\sigma^2]^{-d/2} e^{-(d/2)(s^2/\sigma^2)} [2\pi]^{-K/2}[\sigma^2]^{-K/2} e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}.$$

This density embodies all that we have to learn about the parameters from the observed data. Because the data are taken to be constants in the joint density, we may multiply

[3]These sources include Judge et al. (1982, 1985), Maddala (1977a), Mittelhammer et al. (2000), and the canonical reference for econometricians, Zellner (1971). A remarkable feature of the current literature is the degree to which the analytical components have become ever simpler while the applications have become progressively more complex. This will become evident in Sections 16.5–16.7.

this joint density by the (very carefully chosen), inessential (because it does not involve $\boldsymbol{\beta}$ or σ^2) constant function of the observations,

$$A = \frac{\left(\frac{d}{2}s^2\right)^{(d/2)+1}}{\Gamma\left(\frac{d}{2}+1\right)}[2\pi]^{(d/2)}|\mathbf{X}'\mathbf{X}|^{-1/2}.$$

For convenience, let $v = d/2$. Then, multiplying $L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$ by A gives

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto \frac{[vs^2]^{v+1}}{\Gamma(v+1)}\left(\frac{1}{\sigma^2}\right)^{v} e^{-vs^2(1/\sigma^2)}[2\pi]^{-K/2}|\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \times e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}. \quad \textbf{(16-4)}$$

The likelihood function is proportional to the product of a gamma density for $z = 1/\sigma^2$ with parameters $\lambda = vs^2$ and $P = v + 1$ [see (B-39); this is an **inverted gamma distribution**] and a K-variate normal density for $\boldsymbol{\beta}|\sigma^2$ with mean vector $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The reason will be clear shortly.

16.3.1 ANALYSIS WITH A NONINFORMATIVE PRIOR

The departure point for the Bayesian analysis of the model is the specification of a **prior distribution**. This distribution gives the analyst's prior beliefs about the parameters of the model. One of two approaches is generally taken. If no prior information is known about the parameters, then we can specify a **noninformative prior** that reflects that. We do this by specifying a flat prior for the parameter in question:[4]

$$g(\text{parameter}) \propto \text{constant}.$$

There are different ways that one might characterize the lack of prior information. The implication of a flat prior is that within the range of valid values for the parameter, all intervals of equal length—hence, in principle, all values—are equally likely. The second possibility, an **informative prior**, is treated in the next section. The posterior density is the result of combining the likelihood function with the prior density. Because it pools the full set of information available to the analyst, once the data have been drawn, the posterior density would be interpreted the same way the prior density was before the data were obtained.

To begin, we analyze the case in which σ^2 is assumed to be known. This assumption is obviously unrealistic, and we do so only to establish a point of departure. Using Bayes' theorem, we construct the posterior density,

$$f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2) = \frac{L(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})g(\boldsymbol{\beta}|\sigma^2)}{f(\mathbf{y})} \propto L(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})g(\boldsymbol{\beta}|\sigma^2),$$

assuming that the distribution of $\mathbf{X}$ does not depend on $\boldsymbol{\beta}$ or σ^2. Because $g(\boldsymbol{\beta}|\sigma^2) \propto$ a constant, this density is the one in (16-4). For now, write

$$f(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X}) \propto h(\sigma^2)[2\pi]^{-K/2}|\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}, \quad \textbf{(16-5)}$$

[4]That this *improper* density might not integrate to one is only a minor difficulty. Any constant of integration would ultimately drop out of the final result. See Zellner (1971, pp. 41–53) for a discussion of noninformative priors.

where

$$h(\sigma^2) = \frac{[\nu s^2]^{\nu+1}}{\Gamma(\nu+1)}\left[\frac{1}{\sigma^2}\right]^{\nu} e^{-\nu s^2(1/\sigma^2)}. \tag{16-6}$$

For the present, we treat $h(\sigma^2)$ simply as a constant that involves σ^2, not as a probability density; (16-5) is conditional on σ^2. Thus, the posterior density $f(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})$ is proportional to a multivariate normal distribution with mean $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

This result is familiar, but it is interpreted differently in this setting. First, we have combined our prior information about $\boldsymbol{\beta}$ (in this case, no information) and the sample information to obtain a posterior distribution. Thus, on the basis of the sample data in hand, we obtain a distribution for $\boldsymbol{\beta}$ with mean $\mathbf{b}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The result is dominated by the sample information, as it should be if there is no prior information. In the absence of any prior information, the mean of the posterior distribution, which is a type of Bayesian point estimate, is the sampling theory estimator, $\mathbf{b}$.

To generalize the preceding to an unknown σ^2, we specify a noninformative prior distribution for $\ln \sigma$ over the entire real line.[5] By the change of variable formula, if $g(\ln \sigma)$ is constant, then $g(\sigma^2)$ is proportional to $1/\sigma^2$.[6] Assuming that $\boldsymbol{\beta}$ and σ^2 are independent, we now have the noninformative joint prior distribution,

$$g(\boldsymbol{\beta}, \sigma^2) = g_{\boldsymbol{\beta}}(\boldsymbol{\beta}) g_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}.$$

We can obtain the **joint posterior distribution** for $\boldsymbol{\beta}$ and σ^2 by using

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = L(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X}) g_{\sigma^2}(\sigma^2) \propto L(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X}) \times \frac{1}{\sigma^2}. \tag{16-7}$$

For the same reason as before, we multiply $g_{\sigma^2}(\sigma^2)$ by a well-chosen constant, this time $\nu s^2\Gamma(\nu+1)/\Gamma(\nu+2) = \nu s^2/(\nu+1)$. Multiplying (16-5) by this constant times $g_{\sigma^2}(\sigma^2)$ and inserting $h(\sigma^2)$ gives the joint posterior for $\boldsymbol{\beta}$ and σ^2, given $\mathbf{y}$ and $\mathbf{X}$,

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto \frac{[\nu s^2]^{\nu+2}}{\Gamma(\nu+2)}\left[\frac{1}{\sigma^2}\right]^{\nu+1} e^{-\nu s^2(1/\sigma^2)} [2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \times e^{-(1/2)(\boldsymbol{\beta}-b)'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}.$$

To obtain the marginal posterior distribution for $\boldsymbol{\beta}$, it is now necessary to integrate σ^2 out of the joint distribution (and vice versa to obtain the marginal distribution for σ^2). By collecting the terms, $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$ can be written as

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto A \times \left(\frac{1}{\sigma^2}\right)^{P-1} e^{-\lambda(1/\sigma^2)},$$

[5] See Zellner (1971) for justification of this prior distribution.

[6] Many treatments of this model use σ rather than σ^2 as the parameter of interest. The end results are identical. We have chosen this parameterization because it makes manipulation of the likelihood function with a gamma prior distribution especially convenient. See Zellner (1971, pp. 44–45) for discussion.

where

$$A = \frac{[vs^2]^{v+2}}{\Gamma(v+2)}[2\pi]^{-K/2}|(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2},$$
$$P = v + 2 + K/2 = (n-K)/2 + 2 + K/2 = (n+4)/2,$$

and

$$\lambda = vs^2 + \tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}).$$

The marginal posterior distribution for $\boldsymbol{\beta}$ is

$$\int_0^\infty f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})d\sigma^2 \propto A\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{P-1} e^{-\lambda(1/\sigma^2)}d\sigma^2.$$

To do the integration, we have to make a change of variable; $d(1/\sigma^2) = -(1/\sigma^2)^2 d\sigma^2$, so $d\sigma^2 = -(1/\sigma^2)^{-2}\, d(1/\sigma^2)$. Making the substitution—the sign of the integral changes twice, once for the Jacobian and back again because the integral from $\sigma^2 = 0$ to ∞ is the negative of the integral from $(1/\sigma^2) = 0$ to ∞—we obtain

$$\int_0^\infty f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})d\sigma^2 \propto A\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{P-3} e^{-\lambda(1/\sigma^2)}d\left(\frac{1}{\sigma^2}\right)$$
$$= A \times \frac{\Gamma(P-2)}{\lambda^{P-2}}.$$

Reinserting the expressions for A, P, and λ produces

$$f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) \propto \frac{[vs^2]^{v+2}\Gamma(v + K/2)}{\Gamma(v+2)}[2\pi]^{-K/2}|X'X|^{-1/2}}{\left[vs^2 + \tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})\right]^{v+K/2}}. \quad \textbf{(16-8)}$$

This density is proportional to a **multivariate *t* distribution**[7] and is a generalization of the familiar univariate distribution we have used at various points. This distribution has a degrees of freedom parameter, $d = n - K$, mean $\mathbf{b}$, and covariance matrix $(d/(d-2)) \times [s^2(\mathbf{X}'\mathbf{X})^{-1}]$. Each element of the K-element vector $\boldsymbol{\beta}$ has a marginal distribution that is the univariate t distribution with degrees of freedom $n - K$, mean b_k, and variance equal to the kth diagonal element of the covariance matrix given earlier. Once again, this is the same as our sampling theory result. The difference is a matter of interpretation. In the current context, the estimated distribution is for $\boldsymbol{\beta}$ and is centered at $\mathbf{b}$.

16.3.2 ESTIMATION WITH AN INFORMATIVE PRIOR DENSITY

Once we leave the simple case of noninformative priors, matters become quite complicated, both at a practical level and, methodologically, in terms of just where the prior comes from. The integration of σ^2 out of the posterior in (16-7) is complicated by itself. It is made much more so if the prior distributions of $\boldsymbol{\beta}$ and σ^2 are at all involved. Partly to offset these difficulties, researchers have used **conjugate priors**, which are ones

[7] See, for example, Judge et al. (1985) for details. The expression appears in Zellner (1971, p. 67). Note that the exponent in the denominator is $v + K/2 = n/2$.

that have the same form as the conditional density and are therefore amenable to the integration needed to obtain the marginal distributions.[8]

Example 16.2 Estimation with a Conjugate Prior

We continue Example 16.1, but we now assume a conjugate prior. For likelihood functions involving proportions, the beta prior is a common device, for reasons that will emerge shortly. The beta prior is

$$p(\theta) = \frac{\Gamma(\alpha + \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}.$$

Then the posterior density becomes

$$\frac{\theta^{D}(1-\theta)^{N-D}\dfrac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}}{\displaystyle\int_0^1 \theta^{D}(1-\theta)^{N-D}\frac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}d\theta} = \frac{\theta^{D+\alpha-1}(1-\theta)^{N-D+\beta-1}}{\displaystyle\int_0^1 \theta^{D+\alpha-1}(1-\theta)^{N-D+\beta-1}d\theta}.$$

The posterior density is, once again, a beta distribution, with parameters $(D + \alpha, N - D + \beta)$. The posterior mean is

$$E[\theta \,|\, \mathbf{data}] = \frac{D+\alpha}{N+\alpha+\beta}.$$

(Our previous choice of the uniform density was equivalent to $\alpha = \beta = 1$.) Suppose we choose a prior that conforms to a prior mean of 0.5, but with less mass near zero and one than in the center, such as $\alpha = \beta = 2$. Then the posterior mean would be $(8 + 2)/(25 + 3) = 0.33571$. (This is yet larger than the previous estimator. The reason is that the prior variance is now smaller than 1/12, so the prior mean, still 0.5, receives yet greater weight than it did in the previous example.)

Suppose that we assume that the prior beliefs about $\boldsymbol{\beta}$ may be summarized in a K-variate normal distribution with mean $\boldsymbol{\beta}_0$ and variance matrix $\boldsymbol{\Sigma}_0$. Once again, it is illuminating to begin with the case in which σ^2 is assumed to be known. Proceeding in exactly the same fashion as before, we would obtain the following result: The posterior density of $\boldsymbol{\beta}$ conditioned on σ^2 and the data will be normal with

$$\begin{aligned} E[\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}, \mathbf{X}] &= \{\boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}\{\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b}\} \\ &= \mathbf{F}\boldsymbol{\beta}_0 + (\mathbf{I} - \mathbf{F})\mathbf{b}, \end{aligned} \tag{16-9}$$

where

$$\begin{aligned} \mathbf{F} &= \{\boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}\boldsymbol{\Sigma}_0^{-1} \\ &= \{[\text{prior variance}]^{-1} + [\text{conditional variance}]^{-1}\}^{-1}[\text{prior variance}]^{-1}. \end{aligned} \tag{16-10}$$

This vector is a matrix weighted average of the prior and the least squares (sample) coefficient estimates, where the weights are the inverses of the prior and the conditional

[8]Our choice of noninformative prior for $\ln \sigma$ led to a convenient prior for σ^2 in our derivation of the posterior for $\boldsymbol{\beta}$. The idea that the prior can be specified arbitrarily in whatever form is mathematically convenient is very troubling; it is supposed to represent the accumulated prior belief about the parameter. On the other hand, it could be argued that the conjugate prior is the posterior of a previous analysis, which could justify its form. The issue of how priors should be specified is one of the focal points of the methodological debate. Non-Bayesians argue that it is disingenuous to claim the methodological high ground and then base the crucial prior density in a model purely on the basis of mathematical convenience. In a small sample, this assumed prior is going to dominate the results, whereas in a large one, the sampling theory estimates will dominate anyway.

covariance matrices.[9] The smaller the variance of the estimator, the larger its weight, which makes sense. Also, still taking σ^2 as known, we can write the variance of the posterior normal distribution as

$$\text{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] = \{\boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}. \tag{16-11}$$

Notice that the posterior variance combines the prior and conditional variances on the basis of their inverses.[10] We may interpret the noninformative prior as having infinite elements in $\boldsymbol{\Sigma}_0$. This assumption would reduce this case to the earlier one.

Once again, it is necessary to account for the unknown σ^2. If our prior over σ^2 is to be informative as well, then the resulting distribution can be extremely cumbersome. A conjugate prior for $\boldsymbol{\beta}$ and σ^2 that can be used is

$$g(\boldsymbol{\beta}, \sigma^2) = g_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2)g_{\sigma^2}(\sigma^2), \tag{16-12}$$

where $g_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2)$ is normal, with mean $\boldsymbol{\beta}^0$ and variance $\sigma^2\mathbf{A}$ and

$$g_{\sigma^2}(\sigma^2) = \frac{[m\sigma_0^2]^{m+1}}{\Gamma(m+1)}\left(\frac{1}{\sigma^2}\right)^m e^{-m\sigma_0^2(1/\sigma^2)}. \tag{16-13}$$

This distribution is an inverted gamma distribution. It implies that $1/\sigma^2$ has a gamma distribution. The prior mean for σ^2 is σ_0^2 and the prior variance is $\sigma_0^4/(m-1)$.[11] The product in (16-12) produces what is called a **normal-gamma prior**, which is the natural conjugate prior for this form of the model. By integrating out σ^2, we would obtain the prior marginal for $\boldsymbol{\beta}$ alone, which would be a multivariate t distribution.[12] Combining (16-12) with (16-13) produces the joint posterior distribution for $\boldsymbol{\beta}$ and σ^2. Finally, the marginal posterior distribution for $\boldsymbol{\beta}$ is obtained by integrating out σ^2. It has been shown that this posterior distribution is multivariate t with

$$E[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}] = \{[\overline{\sigma}^2\mathbf{A}]^{-1} + [\overline{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}\{[\overline{\sigma}^2\mathbf{A}]^{-1}\boldsymbol{\beta}_0 + [\overline{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b}\} \tag{16-14}$$

and

$$\text{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}] = \left(\frac{j}{j-2}\right)\{[\overline{\sigma}^2\mathbf{A}]^{-1} + [\overline{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1}, \tag{16-15}$$

where j is a degrees of freedom parameter and $\overline{\sigma}^2$ is the Bayesian estimate of σ^2. The prior degrees of freedom m is a parameter of the prior distribution for σ^2 that would have been determined at the outset. (See the following example.) Once again, it is clear that as the amount of data increases, the posterior density, and the estimates thereof, converge to the sampling theory results.

---

[9]Note that it will not follow that individual elements of the posterior mean vector lie between those of $\boldsymbol{\beta}_0$ and b. See Judge et al. (1985, pp. 109–110) and Chamberlain and Leamer (1976).

[10]Precisely this estimator was proposed by Theil and Goldberger (1961) as a way of combining a previously obtained estimate of a parameter and a current body of new data. They called their result a "mixed estimator." The term "mixed estimation" takes an entirely different meaning in the current literature, as we saw in Chapter 15.

[11]You can show this result by using gamma integrals. Note that the density is a function of $1/\sigma^2 = 1/x$ in the formula of (B-39), so to obtain $E[\sigma^2]$, we use the analog of $E[1/x] = \lambda/(P-1)$ and $E[(1/x)^2] = \lambda^2/[(P-1)(P-2)]$. In the density for $(1/\sigma^2)$, the counterparts to λ and P are $m\sigma_0^2$ and $m+1$.

[12]Full details of this (lengthy) derivation appear in Judge et al. (1985, pp. 106–110) and Zellner (1971).

TABLE 16.1 Estimates of the MPC

Years	*Estimated MPC*	*Variance of b*	*Degrees of Freedom*	*Estimated σ*
1940–1950	0.6848014	0.061878	9	24.954
1950–2000	0.92481	0.000065865	49	92.244

Example 16.3 Bayesian Estimate of the Marginal Propensity to Consume

In Example 3.2, an estimate of the marginal propensity to consume is obtained using 11 observations from 1940 to 1950, with the results shown in the top row of Table 16.1. [Referring to Example 3.2, the variance is (6,848.975/9)/12,300.182.] A classical 95% confidence interval for β based on these estimates is (0.1221, 1.2475). (The very wide interval probably results from the obviously poor specification of the model.) Based on noninformative priors for β and σ^2, we would estimate the posterior density for β to be univariate t with nine degrees of freedom, with mean 0.6848014 and variance (11/9)0.061878 = 0.075628. An HPD interval for β would coincide with the confidence interval. Using the fourth quarter (yearly) values of the 1950–2000 data used in Example 5.3, we obtain the new estimates that appear in the second row of the table.

We take the first estimate and its estimated distribution as our prior for β and obtain a posterior density for β based on an informative prior instead. We assume for this exercise that σ may be taken as known at the sample value of 24.954. Then,

$$\bar{b} = \left[\frac{1}{0.061878} + \frac{1}{0.000065865}\right]^{-1}\left[\frac{0.6848014}{0.061878} + \frac{0.92481}{0.000065865}\right] = 0.92455,$$

The weighted average is overwhelmingly dominated by the far more precise sample estimate from the larger sample. The posterior variance is the inverse in brackets, which is 0.000065795. This is close to the variance of the latter estimate. An HPD interval can be formed in the familiar fashion. It will be slightly narrower than the confidence interval, because the variance of the posterior distribution is slightly smaller than the variance of the sampling estimator. This reduction is the value of the prior information. (As we see here, the prior is not particularly informative.)

16.4 BAYESIAN INFERENCE

The posterior density is the Bayesian counterpart to the likelihood function. It embodies the information that is available to make inference about the econometric model. As we have seen, the mean and variance of the posterior distribution correspond to the classical (sampling theory) point estimator and asymptotic variance, although they are interpreted differently. Before we examine more intricate applications of Bayesian inference, it is useful to formalize some other components of the method, point and interval estimation and the Bayesian equivalent of testing a hypothesis.[13]

16.4.1 POINT ESTIMATION

The posterior density function embodies the prior and the likelihood and therefore contains all the researcher's information about the parameters. But for purposes of presenting

[13]We do not include prediction in this list. The Bayesian approach would treat the prediction problem as one of estimation in the same fashion as parameter estimation. The value to be forecasted is among the unknown elements of the model that would be characterized by a prior and would enter the posterior density in a symmetric fashion along with the other parameters.

results, the density is somewhat imprecise, and one normally prefers a point or interval estimate. The natural approach would be to use the mean of the posterior distribution as the estimator. For the noninformative prior, we use **b**, the **sampling theory** estimator.

One might ask at this point, why bother? These Bayesian point estimates are identical to the sampling theory estimates. All that has changed is our interpretation of the results. This situation is, however, exactly the way it should be. Remember that we entered the analysis with noninformative priors for $\boldsymbol{\beta}$ and σ^2. Therefore, the only information brought to bear on estimation is the sample data, and it would be peculiar if anything other than the sampling theory estimates emerged at the end. The results do change when our prior brings out of sample information into the estimates, as we shall see later.

The results will also change if we change our motivation for estimating $\boldsymbol{\beta}$. The parameter estimates have been treated thus far as if they were an end in themselves. But in some settings, parameter estimates are obtained so as to enable the analyst to make a decision. Consider then, a **loss function**, $H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, which quantifies the cost of basing a decision on an estimate $\hat{\boldsymbol{\beta}}$ when the parameter is $\boldsymbol{\beta}$. The expected, or average, loss is

$$E_{\boldsymbol{\beta}}[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})] = \int_{\boldsymbol{\beta}} H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) d\boldsymbol{\beta}, \tag{16-16}$$

where the weighting function, f, is the marginal posterior density. (The joint density for $\boldsymbol{\beta}$ and σ^2 would be used if the loss were defined over both.) The Bayesian point estimate is the parameter vector that minimizes the expected loss. If the loss function is a quadratic form in $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, then the mean of the posterior distribution is the *minimum expected loss* (MELO) estimator. The proof is simple. For this case,

$$E[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})|\mathbf{y}, \mathbf{X}] = E\left[\tfrac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|\mathbf{y}, \mathbf{X}\right].$$

To minimize this, we can use the result that

$$\begin{aligned} \partial E[H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})|\mathbf{y}, \mathbf{X}]/\partial\hat{\boldsymbol{\beta}} &= E[\partial H(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})/\partial\hat{\boldsymbol{\beta}}|\mathbf{y}, \mathbf{X}] \\ &= E[-\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|\mathbf{y}, \mathbf{X}]. \end{aligned}$$

The minimum is found by equating this derivative to $\mathbf{0}$, whence, because $-\mathbf{W}$ is irrelevant, $\hat{\boldsymbol{\beta}} = E[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}]$. This kind of loss function would state that errors in the positive and negative directions are equally bad, and large errors are much worse than small errors. If the loss function were a linear function instead, then the MELO estimator would be the median of the posterior distribution. These results are the same in the case of the noninformative prior that we have just examined.

16.4.2 INTERVAL ESTIMATION

The counterpart to a confidence interval in this setting is an interval of the posterior distribution that contains a specified probability. Clearly, it is desirable to have this interval be as narrow as possible. For a unimodal density, this corresponds to an interval within which the density function is higher than any points outside it, which justifies the term **highest posterior density (HPD) interval**. For the case we have analyzed, which involves a symmetric distribution, we would form the HPD interval for $\boldsymbol{\beta}$ around the least squares estimate **b**, with terminal values taken from the standard t tables. Section 4.8.3 shows the (classical) derivation of an HPD interval for an asymmetric distribution, in that case for a prediction of y when the regression models ln y.

16.4.3 HYPOTHESIS TESTING

The Bayesian methodology treats the classical approach to hypothesis testing with a large amount of skepticism. Two issues are especially problematic. First, a close examination of only the work we have done in Chapter 5 will show that because we are using consistent estimators, with a large enough sample, we will ultimately reject any (nested) hypothesis unless we adjust the significance level of the test downward as the sample size increases. Second, the all-or-nothing approach of either rejecting or not rejecting a hypothesis provides no method of simply sharpening our beliefs. Even the most committed of analysts might be reluctant to discard a strongly held prior based on a single sample of data, yet that is what the sampling methodology mandates. The Bayesian approach to hypothesis testing is much more appealing in this regard. Indeed, the approach might be more appropriately called *comparing hypotheses*, because it essentially involves only making an assessment of which of two hypotheses has a higher probability of being correct.

The Bayesian approach to hypothesis testing bears large similarity to Bayesian estimation.[14] We have formulated two hypotheses, a null, denoted H_0, and an alternative, denoted H_1. These need not be complementary, as in H_0: "statement A is true" versus H_1: "statement A is not true," because the intent of the procedure is not to reject one hypothesis in favor of the other. For simplicity, however, we will confine our attention to hypotheses about the parameters in the regression model, which often are complementary. Assume that before we begin our experimentation (i.e., data gathering, statistical analysis) we are able to assign **prior probabilities** $P(H_0)$ and $P(H_1)$ to the two hypotheses. The **prior odds ratio** is simply the ratio

$$\text{Odds}_{prior} = \frac{P(H_0)}{P(H_1)}. \tag{16-17}$$

For example, one's uncertainty about the sign of a parameter might be summarized in a prior odds over $H_0: \beta \geq 0$ versus $H_1: \beta < 0$ of $0.5/0.5 = 1$. After the sample evidence is gathered, the prior will be modified, so the posterior is, in general,

$$\text{Odds}_{posterior} = B_{01} \times \text{Odds}_{prior}.$$

The value B_{01} is called the **Bayes factor** for comparing the two hypotheses. It summarizes the effect of the sample data on the prior odds. The end result, $\text{Odds}_{posterior}$, is a new odds ratio that can be carried forward as the prior in a subsequent analysis.

The Bayes factor is computed by assessing the likelihoods of the data observed under the two hypotheses. We return to our first departure point, the likelihood of the data, given the parameters,

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \tag{16-18}$$

Based on our priors for the parameters, the expected, or average likelihood, assuming that hypothesis j is true ($j = 0, 1$), is

$$f(\mathbf{y} \mid \mathbf{X}, H_j) = E_{\boldsymbol{\beta},\sigma^2}[f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}, H_j)] = \int_{\sigma^2}\int_{\boldsymbol{\beta}} f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}, H_j) g(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2.$$

[14] For extensive discussion, see Zellner and Siow (1980) and Zellner (1985, pp. 275–305).

(This conditional density is also the **predictive density** for **y**.) Therefore, based on the observed data, we use Bayes's theorem to reassess the probability of H_j; the posterior probability is

$$P(H_j|\mathbf{y}, \mathbf{X}) = \frac{f(\mathbf{y}|\mathbf{X}, H_j)P(H_j)}{f(\mathbf{y})}.$$

The posterior odds ratio is $P(H_0|\mathbf{y}, \mathbf{X})/P(H_1|\mathbf{y}, \mathbf{X})$, so the Bayes factor is

$$B_{01} = \frac{f(\mathbf{y}|\mathbf{X}, H_0)}{f(y|\mathbf{X}, H_1)}.$$

Example 16.4 Posterior Odds for the Classical Regression Model

Zellner (1971) analyzes the setting in which there are two possible explanations for the variation in a dependent variable *y*:

$$\text{Model0: } y = \mathbf{x}_0'\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0$$

and

$$\text{Model1: } y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1.$$

We will briefly sketch his results. We form *informative priors* for $[\boldsymbol{\beta}, \sigma^2]_j$, $j = 0, 1$, as specified in (16-12) and (16-13), that is, multivariate normal and inverted gamma, respectively. Zellner then derives the Bayes factor for the posterior odds ratio. The derivation is lengthy and complicated, but for large n, with some simplifying assumptions, a useful formulation emerges. First, assume that the priors for σ_0^2 and σ_1^2 are the same. Second, assume that $[|\mathbf{A}_0^{-1}|/|\mathbf{A}_0^{-1} + \mathbf{X}_0'\mathbf{X}_0|]/[|\mathbf{A}_1^{-1}|/|\mathbf{A}_1^{-1} + \mathbf{X}_1'\mathbf{X}_1|] \rightarrow 1$. The first of these would be the usual situation, in which the uncertainty concerns the covariation between y_i and $\mathbf{x}_i$, not the amount of residual variation (lack of fit). The second concerns the relative amounts of information in the prior (**A**) versus the likelihood (**X′X**). These matrices are the inverses of the covariance matrices, or the **precision matrices**. [Note how these two matrices form the matrix weights in the computation of the posterior mean in (16-9).] Zellner (p. 310) discusses this assumption at some length. With these two assumptions, he shows that as n grows large,[15]

$$B_{01} \approx \left(\frac{s_0^2}{s_1^2}\right)^{-(n+m)/2} = \left(\frac{1 - R_0^2}{1 - R_1^2}\right)^{-(n+m)/2}.$$

Therefore, the result favors the model that provides the better fit using R^2 as the fit measure. If we stretch Zellner's analysis a bit by interpreting model 1 as *the model* and model 0 as "no model" (that is, the relevant part of $\boldsymbol{\beta}_0 = \mathbf{0}$, so $R_0^2 = 0$), then the ratio simplifies to

$$B_{01} = (1 - R_1^2)^{(n+m)/2}.$$

Thus, the better the fit of the regression, the lower the Bayes factor in favor of model 0 (no model), which makes intuitive sense.

Zellner and Siow (1980) have continued this analysis with noninformative priors for $\boldsymbol{\beta}$ and σ_j^2. Specifically, they use a flat prior for ln σ [see (16-7)] and a multivariate Cauchy prior (which has infinite variances) for $\boldsymbol{\beta}$. Their main result (3.10) is

[15] A ratio of exponentials that appears in Zellner's result (his equation 10.50) is omitted. To the order of approximation in the result, this ratio vanishes from the final result. (Personal correspondence from A. Zellner to the author.)

$$B_{01} = \frac{\frac{1}{2}\sqrt{\pi}}{\Gamma[(k+1)/2]}\left(\frac{n-K}{2}\right)^{k/2}(1-R^2)^{(n-K-1)/2}.$$

This result is very much like the previous one, with some slight differences due to degrees of freedom corrections and the several approximations used to reach the first one.

16.4.4 LARGE-SAMPLE RESULTS

Although all statistical results for Bayesian estimators are necessarily "finite sample" (they are conditioned on the sample data), it remains of interest to consider how the estimators behave in large samples.[16] Do Bayesian estimators "converge" to something? To do this exercise, it is useful to envision having a sample that is the entire population. Then, the posterior distribution would characterize this entire population, not a sample from it. It stands to reason in this case, at least intuitively, that the posterior distribution should coincide with the likelihood function. It will (as usual) save for the influence of the prior. But as the sample size grows, one should expect the likelihood function to overwhelm the prior. It will, unless the strength of the prior grows with the sample size (that is, for example, if the prior variance is of order $1/n$). An informative prior will still fade in its influence on the posterior unless it becomes *more* informative as the sample size grows.

The preceding suggests that the posterior mean will converge to the maximum likelihood estimator. The MLE is the parameter vector that is at the mode of the likelihood function. The Bayesian estimator is the **posterior mean**, not the mode, so a remaining question concerns the relationship between these two features. The **Bernstein–von Mises "theorem"** [See Cameron and Trivedi (2005, p. 433) and Train (2003, Chapter 12)] states that the posterior mean and the maximum likelihood estimator will coverge to the same probability limit and have the same limiting normal distribution. A form of central limit theorem is at work.

But for remaining philosophical questions, the results suggest that for large samples, the choice between Bayesian and frequentist methods can be one of computational efficiency. (This is the thrust of the application in Section 16.8. Note, as well, footnote 1 at the beginning of this chapter. In an infinite sample, the maintained uncertainty of the Bayesian estimation framework would have to arise from deeper questions about the model. For example, the mean of the entire population is its mean; there is no uncertainty about the parameter.)

16.5 POSTERIOR DISTRIBUTIONS AND THE GIBBS SAMPLER

The foregoing analysis has proceeded along a set of steps that includes formulating the likelihood function (the model), the prior density over the objects of estimation, and the posterior density. To complete the inference step, we then analytically derived the characteristics of the posterior density of interest, such as the mean or mode, and the

[16]The standard preamble in econometric studies, that the analysis to follow is "exact" as opposed to approximate or "large sample," refers to this aspect—the analysis is conditioned on and, by implication, applies only to the sample data in hand. Any inference outside the sample, for example, to hypothesized random samples is, like the sampling theory counterpart, approximate.

variance. The complicated element of any of this analysis is determining the moments of the posterior density, for example, the mean,

$$\hat{\theta} = E[\theta|\text{data}] = \int_{\theta} \theta\, p(\theta|\textbf{data}) d\theta. \tag{16-19}$$

There are relatively few applications for which integrals such as this can be derived in closed form. (This is one motivation for conjugate priors.) The modern approach to Bayesian inference takes a different strategy. The result in (16-19) is an expectation. Suppose it were possible to obtain a random sample, as large as desired, from the population defined by $p(\theta|\textbf{data})$. Then, using the same strategy we used throughout Chapter 15 for simulation-based estimation, we could use that sample's characteristics, such as mean, variance, quantiles, and so on, to infer the characteristics of the posterior distribution. Indeed, with an (essentially) infinite sample, we would be freed from having to limit our attention to a few simple features such as the mean and variance and we could view any features of the posterior distribution that we like. The (much less) complicated part of the analysis is the formulation of the posterior density.

It remains to determine how the sample is to be drawn from the posterior density. This element of the strategy is provided by a remarkable (and remarkably useful) result known as the **Gibbs sampler**.[17] The central result of the Gibbs sampler is as follows: We wish to draw a random sample from the joint population (x, y). The joint distribution of x and y is either unknown or intractable and it is not possible to sample from the joint distribution. However, assume that the conditional distributions $f(x|y)$ and $f(y|x)$ are known and simple enough that it is possible to draw univariate random samples from both of them. The following iteration will produce a bivariate random sample from the joint distribution:

Gibbs Sampler:

1. Begin the cycle with a value of x_0 that is in the right range of $x|y$,
2. Draw an observation $y_0|x_0$, from the known population $y|x$,
3. Draw an observation $x_t|y_{t-1}$, from the known population $x|y$,
4. Draw an observation $y_t|x_t$ from the known population of $y|x$.

Iteration of steps 3 and 4 for several thousand cycles will eventually produce a random sample from the joint distribution. (The first several thousand draws are discarded to avoid the influence of the initial conditions—this is called the **burn in**.) [Some technical details on the procedure appear in Cameron and Trivedi (Section 13.5).]

Example 16.5 Gibbs Sampling from the Normal Distribution

To illustrate the mechanical aspects of the Gibbs sampler, consider random sampling from the joint normal distribution. We consider the bivariate normal distribution first. Suppose we wished to draw a random sample from the population

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

As we have seen in Chapter 15, a direct approach is to use the fact that linear functions of normally distributed variables are normally distributed. [See (B-80).] Thus, we might

[17] See Casella and George (1992).

transform a series of independent normal draws $(u_1, u_2)'$ by the Cholesky decomposition of the covariance matrix,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_i = \begin{bmatrix} 1 & 0 \\ \theta_1 & \theta_2 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_i = \mathbf{L}\mathbf{u}_i,$$

where $\theta_1 = \rho$ and $\theta_2 = \sqrt{1 - \rho^2}$. The Gibbs sampler would take advantage of the result

$$x_1 | x_2 \sim N[\rho x_2, (1 - \rho^2)],$$

and

$$x_2 | x_1 \sim N[\rho x_1, (1 - \rho^2)].$$

To sample from a trivariate, or multivariate population, we can expand the Gibbs sequence in the natural fashion. For example, to sample from a trivariate population, we would use the Gibbs sequence

$$\begin{aligned} x_1 | x_2, x_3 &\sim N[\beta_{1,2}x_2 + \beta_{1,3}x_3, \Sigma_{1|2,3}], \\ x_2 | x_1, x_3 &\sim N[\beta_{2,1}x_1 + \beta_{2,3}x_3, \Sigma_{2|1,3}], \\ x_3 | x_1, x_2 &\sim N[\beta_{3,1}x_1 + \beta_{3,2}x_2, \Sigma_{3|1,2}], \end{aligned}$$

where the conditional means and variances are given in Theorem B.7. This defines a three-step cycle.

The availability of the Gibbs sampler frees the researcher from the necessity of deriving the analytical properties of the full, joint posterior distribution. Because the formulation of conditional priors is straightforward, and the derivation of the conditional posteriors is only slightly less so, this tool has facilitated a vast range of applications that previously were intractable. For an example, consider, once again, the classical normal regression model. From (16-7), the joint posterior for $(\boldsymbol{\beta}, \sigma^2)$ is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto{}& \frac{[\nu s^2]^{\nu+2}}{\Gamma(\nu + 2)} \left[\frac{1}{\sigma^2}\right]^{\nu+1} \exp(-\nu s^2/\sigma^2)[2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \\ &\times \exp(-(1/2)(\boldsymbol{\beta} - \mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta} - \mathbf{b}). \end{aligned}$$

If we wished to use a simulation approach to characterizing the posterior distribution, we would need to draw a $K + 1$ variate sample of observations from this intractable distribution. However, with the assumed priors, we found the conditional posterior for $\boldsymbol{\beta}$ in (16-5):

$$p(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) = N[\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

From (16-6), we can deduce that the conditional posterior for $\sigma^2 | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}$ is an inverted gamma distribution with parameters $m\sigma_0^2 = \nu\hat{\sigma}^2$ and $m = \nu$ in (16-13):

$$p(\sigma^2 | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) = \frac{[\nu\hat{\sigma}^2]^{\nu+1}}{\Gamma(\nu + 1)} \left[\frac{1}{\sigma^2}\right]^{\nu} \exp(-\nu\hat{\sigma}^2/\sigma^2), \quad \hat{\sigma}^2 = \frac{\Sigma_{i=1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{n - K}.$$

This sets up a Gibbs sampler for sampling from the joint posterior of $\boldsymbol{\beta}$ and σ^2. We would cycle between random draws from the multivariate normal for $\boldsymbol{\beta}$ and the inverted gamma distribution for σ^2 to obtain a $K + 1$ variate sample on $(\boldsymbol{\beta}, \sigma^2)$. [Of course, for this application, we do know the marginal posterior distribution for $\boldsymbol{\beta}$—see (16-8).]

The Gibbs sampler is not truly a random sampler; it is a Markov chain—each "draw" from the distribution is a function of the draw that precedes it. The random input at each cycle provides the randomness, which leads to the popular name for this strategy, **Markov chain Monte Carlo** or **MCMC** or **MC**^{2} (pick one) estimation. In its simplest form, it provides a remarkably efficient tool for studying the posterior distributions in very complicated models. The example in the next section shows a striking example of how to locate the MLE for a probit model without computing the likelihood function or its derivatives. In Section 16.8, we will examine an extension and refinement of the strategy, the Metropolis–Hasting algorithm.

In the next several sections, we will present some applications of Bayesian inference. In Section 16.9, we will return to some general issues in classical and Bayesian estimation and inference. At the end of the chapter, we will examine Koop and Tobias's (2004) Bayesian approach to the analysis of heterogeneity in a wage equation based on panel data. We used classical methods to analyze these data in Example 15.16.

16.6 APPLICATION: BINOMIAL PROBIT MODEL

Consider inference about the binomial probit model for a dependent variable that is generated as follows (see Sections 17.2–17.4):

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim \mathrm{N}[0, 1], \tag{16-20}$$

$$y_i = 1 \quad \text{if} \quad y_i^* > 0, \text{otherwise } y_i = 0. \tag{16-21}$$

(Theoretical motivation for the model appears in Section 17.3.) The data consist of $(\mathbf{y}, \mathbf{X}) = (y_i, \mathbf{x}_i), i = 1, \ldots, n$. The random variable y_i has a Bernoulli distribution with probabilities

$$\text{Prob}[y_i = 1 | \mathbf{x}_i] = \Phi(\mathbf{x}_i'\boldsymbol{\beta}),$$
$$\text{Prob}[y_i = 0 | \mathbf{x}_i] = 1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}).$$

The likelihood function for the observed data is

$$L(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} [\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}.$$

(Once again, we cheat a bit on the notation—the likelihood function is actually the joint density for the data, given $\mathbf{X}$ and $\boldsymbol{\beta}$.) Classical maximum likelihood estimation of $\boldsymbol{\beta}$ is developed in Section 17.3. To obtain the posterior mean (Bayesian estimator), we assume a noninformative, flat (improper) prior for $\boldsymbol{\beta}$,

$$p(\boldsymbol{\beta}) \propto 1.$$

The posterior density would be

$$p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = \frac{\prod_{i=1}^{n} [\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}(1)}{\int_{\boldsymbol{\beta}} \prod_{i=1}^{n} [\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}(1) d\boldsymbol{\beta}},$$

and the estimator would be the posterior mean,

$$\hat{\boldsymbol{\beta}} = E[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}] = \frac{\int_{\boldsymbol{\beta}} \boldsymbol{\beta} \prod_{i=1}^{n} [\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i} d\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}} \prod_{i=1}^{n} [\Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i} d\boldsymbol{\beta}}. \tag{16-22}$$

Evaluation of the integrals in (16-22) is hopelessly complicated, but a solution using the Gibbs sampler and a technique known as **data augmentation**, pioneered by Albert and Chib (1993a), is surprisingly simple. We begin by treating the unobserved y_i^*'s as unknowns to be estimated, along with $\boldsymbol{\beta}$. Thus, the $(K + n) \times 1$ parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{y}^*)$. We now construct a Gibbs sampler. Consider, first, $p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X})$. If y_i^* is known, then y_i is known [see (16-21)]. It follows that

$$p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X}) = p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{X}).$$

This posterior defines a linear regression model with normally distributed disturbances and known $\sigma^2 = 1$. It is precisely the model we saw in Section 16.3.1, and the posterior we need is in (16-5), with $\sigma^2 = 1$. So, based on our earlier results, it follows that

$$p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X}) = \mathrm{N}[\mathbf{b}^*, (\mathbf{X}'\mathbf{X})^{-1}], \tag{16-23}$$

where

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*.$$

For y_i^*, ignoring y_i for the moment, it would follow immediately from (16-20) that

$$p(y_i^*|\boldsymbol{\beta}, \mathbf{X}) = \mathrm{N}[\mathbf{x}_i'\boldsymbol{\beta}, 1].$$

However, y_i is informative about y_i^*. If y_i equals one, we know that $y_i^* > 0$ and if y_i equals zero, then $y_i^* \leq 0$. The implication is that conditioned on $\boldsymbol{\beta}$, $\mathbf{X}$, and $\mathbf{y}$, y_i^* has the truncated (above or below zero) normal distribution that is developed in Sections 19.2.1 and 19.2.2. The standard notation for this is

$$\begin{aligned} p(y_i^*|y_i = 1, \boldsymbol{\beta}, \mathbf{x}_i) &= \mathrm{N}^+[\mathbf{x}_i'\boldsymbol{\beta}, 1], \\ p(y_i^*|y_i = 0, \boldsymbol{\beta}, \mathbf{x}_i) &= \mathrm{N}^-[\mathbf{x}_i'\boldsymbol{\beta}, 1]. \end{aligned} \tag{16-24}$$

Results (16-23) and (16-24) set up the components for a Gibbs sampler that we can use to estimate the posterior means $E[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}]$ and $E[\mathbf{y}^*|\mathbf{y}, \mathbf{X}]$. The following is our algorithm:

Gibbs Sampler for the Binomial Probit Model

1. Compute $\mathbf{X}'\mathbf{X}$ once at the outset and obtain $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}' = (\mathbf{X}'\mathbf{X})^{-1}$ (Cholesky decomposition).
2. Start $\boldsymbol{\beta}$ at any value such as $\mathbf{0}$.
3. Result (15-4) shows how to transform a draw from U[0, 1] to a draw from the truncated normal with underlying mean μ and standard deviation σ. For this application, the draw is

$$\begin{aligned} y_{i,r}^*(r) &= \mathbf{x}_i'\boldsymbol{\beta}_{r-1} + \Phi^{-1}[1 - (1 - U)\Phi(\mathbf{x}_i'\boldsymbol{\beta}_{r-1})] \quad \text{if } y_i = 1, \\ y_{i,r}^*(r) &= \mathbf{x}_i'\boldsymbol{\beta}_{r-1} + \Phi^{-1}[U\Phi(-\mathbf{x}_i'\boldsymbol{\beta}_{r-1})] \quad \text{if } y_i = 0. \end{aligned}$$

   This step is used to draw the n observations on $y_{i,r}^*(r)$.

4. Section 15.2.4 shows how to draw an observation from the multivariate normal population. For this application, we use the results at step 3 to compute $\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*(r)$. We obtain a vector, $\mathbf{v}$, of K draws from the N[0, 1] population, then $\boldsymbol{\beta}(r) = \mathbf{b}^* + \mathbf{L}\mathbf{v}$.

The iteration cycles between steps 3 and 4. This should be repeated several thousand times, discarding the burn-in draws, then the estimator of $\boldsymbol{\beta}$ is the sample mean of the retained draws. The posterior variance is computed with the variance of the retained draws. Posterior estimates of y_i^* would typically not be useful.

Example 16.6 Gibbs Sampler for a Probit Model

In Examples 14.19 through 14.21, we examined Spector and Mazzeo's (1980) widely traveled data on a binary choice outcome. (The example used the data for a different model.) The binary probit model studied in the paper was

$$\text{Prob}(GRADE_i = 1 \,|\, \boldsymbol{\beta}, \mathbf{x}_i) = \Phi(\beta_1 + \beta_2 GPA_i + \beta_3 TUCE_i + \beta_4 PSI_i).$$

The variables are defined in Example 14.19. Their probit model is studied in Example 17.3. The sample contains 32 observations. Table 16.2 presents the maximum likelihood estimates and the posterior means and standard deviations for the probit model. For the Gibbs sampler, we used 5,000 draws, and discarded the first 1,000.

The results in Table 16.2 suggest the similarity of the posterior mean estimated with the Gibbs sampler to the maximum likelihood estimate. However, the sample is quite small, and the differences between the coefficients are still fairly substantial. For a striking example of the behavior of this procedure, we now revisit the German health care data examined in Example 14.23 and several other examples throughout the book. The probit model to be estimated is

$$\begin{aligned}\text{Prob}(\textit{Doctor visits}_{it} > 0) = \Phi(&\beta_1 + \beta_2\, \textit{Age}_{it} + \beta_3\, \textit{Education}_{it} + \beta_4\, \textit{Income}_{it} \\ &+ \beta_5\, \textit{Kids}_{it} + \beta_6\, \textit{Married}_{it} + \beta_7\, \textit{Female}_{it}).\end{aligned}$$

The sample contains data on 7,293 families and a total of 27,326 observations. We are pooling the data for this application. Table 16.3 presents the probit results for this model using the same procedure as before. (We used only 500 draws and discarded the first 100.)

The similarity is what one would expect given the large sample size. We note before proceeding to other applications, notwithstanding the striking similarity of the Gibbs sampler to the MLE, that this is not an efficient method of estimating the parameters of a probit model. The estimator requires generation of thousands of samples of potentially thousands of observations. We used only 500 replications to produce Table 16.3. The computations took about five minutes. Using Newton's method to maximize the log likelihood directly took less than five seconds. Unless one is wedded to the Bayesian paradigm, on strictly practical grounds, the MLE would be the preferred estimator.

TABLE 16.2 Probit Estimates for Grade Equation

	Maximum Likelihood		*Posterior Means and Std. Devs.*	
Variable	***Estimate***	***Std. Error***	***Posterior Mean***	***Posterior S.D.***
Constant	−7.4523	2.5425	−8.6286	2.7995
GPA	1.6258	0.6939	1.8754	0.7668
TUCE	0.0517	0.0839	0.0628	0.0869
PSI	1.4263	0.5950	1.6072	0.6257

TABLE 16.3 Probit Estimates for Doctor Visits Equation

	Maximum Likelihood		*Posterior Means and Std. Devs.*	
Variable	***Estimate***	***Std. Error***	***Posterior Mean***	***Posterior S.D.***
Constant	−0.124332	0.058146	−0.126287	0.054759
Age	0.011892	0.000796	0.011979	0.000801
Education	−0.014959	0.003575	−0.015142	0.003625
Income	−0.132595	0.046552	−0.126693	0.047979
Kids	−0.152114	0.018327	−0.151492	0.018400
Married	0.073518	0.020644	0.071977	0.020852
Female	0.355906	0.016017	0.355828	0.015913

This application of the Gibbs sampler demonstrates in an uncomplicated case how the algorithm can provide an alternative to actually maximizing the log likelihood. We do note that the similarity of the method to the EM algorithm in Section E.3.7 is not coincidental. Both procedures use an estimate of the unobserved, censored data, and both estimate $\boldsymbol{\beta}$ by using OLS using the predicted data.

16.7 PANEL DATA APPLICATION: INDIVIDUAL EFFECTS MODELS

We consider a panel data model with common individual effects,

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathrm{N}[0, \sigma_\varepsilon^2].$$

In the Bayesian framework, there is no need to distinguish between fixed and random effects. The classical distinction results from an asymmetric treatment of the data and the parameters. So, we will leave that unspecified for the moment. The implications will emerge later when we specify the prior densities over the model parameters.

The likelihood function for the sample under normality of ε_{it} is

$$p(\mathbf{y}|\alpha_1, \ldots, \alpha_n, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mathbf{X}) = \prod_{i=1}^{n}\prod_{t=1}^{T_i} \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} \exp\left(-\frac{(y_{it} - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta})^2}{2\sigma_\varepsilon^2}\right).$$

The remaining analysis hinges on the specification of the prior distributions. We will consider three cases. Each illustrates an aspect of the methodology.

First, group the full set of location (regression) parameters in one $(n + K) \times 1$ slope vector, $\boldsymbol{\gamma}$. Then, with the disturbance variance, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon^2) = (\boldsymbol{\gamma}, \sigma_\varepsilon^2)$. Define a conformable data matrix, $\mathbf{Z} = (\mathbf{D}, \mathbf{X})$, where $\mathbf{D}$ contains the n dummy variables so that we may write the model

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

in the familiar fashion for our common effects linear regression. (See Chapter 11.) We now assume the **uniform-inverse gamma prior** that we used in our earlier treatment of the linear model,

$$p(\boldsymbol{\gamma}, \sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2.$$

The resulting (marginal) posterior density for $\boldsymbol{\gamma}$ is precisely that in (16-8) (where now the slope vector includes the elements of $\boldsymbol{\alpha}$). The density is an $(n + K)$ variate t with mean equal to the OLS estimator and covariance matrix $[(\Sigma_i T_i - n - K)/(\Sigma_i T_i - n - K - 2)]s^2(\mathbf{Z}'\mathbf{Z})^{-1}$.

Because OLS in this model as stated means the within estimator, the implication is that with this noninformative prior over $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the model is equivalent to the fixed effects model. Note, again, this is not a consequence of any assumption about correlation between effects and included variables. That has remained unstated; though, by implication, we would allow correlation between $\mathbf{D}$ and $\mathbf{X}$.

Some observers are uncomfortable with the idea of a **uniform prior** over the entire real line.[18] Formally, our assumption of a uniform prior over the entire real line is an **improper prior** because it cannot have a positive density and integrate to one over the entire real line. As such, the posterior appears to be ill defined. However, note that the "improper" uniform prior will, in fact, fall out of the posterior, because it appears in both numerator and denominator. The practical solution for location parameters, such as a vector of regression slopes, is to assume a nearly flat, "almost uninformative" prior. The usual choice is a conjugate normal prior with an arbitrarily large variance. (It should be noted, of course, that as long as that variance is finite, even if it is large, the prior is informative. We return to this point in Section 16.9.)

Consider, then, the conventional normal-gamma prior over $(\boldsymbol{\gamma}, \sigma_\varepsilon^2)$ where the conditional (on σ_ε^2) prior normal density for the slope parameters has mean $\boldsymbol{\gamma}_0$ and covariance matrix $\sigma_\varepsilon^2\mathbf{A}$, where the $(n + K) \times (n + K)$ matrix, $\mathbf{A}$, is yet to be specified. [See the discussion after (16-13).] The marginal posterior mean and variance for $\boldsymbol{\gamma}$ for this set of assumptions are given in (16-14) and (16-15). We reach a point that presents two rather serious dilemmas for the researcher. The posterior was simple with our uniform, noninformative prior. Now, it is necessary actually to specify $\mathbf{A}$, which is potentially large. (In one of our main applications in this text, we are analyzing models with $n = 7{,}293$ constant terms and about $K = 7$ regressors.) It is hopelessly optimistic to expect to be able to specify all the variances and covariances in a matrix this large, unless we actually have the results of an earlier study (in which case we would also have a prior estimate of $\boldsymbol{\gamma}$). A practical solution that is frequently chosen is to specify $\mathbf{A}$ to be a diagonal matrix with extremely large diagonal elements, thus emulating a uniform prior without having to commit to one. The second practical issue then becomes dealing with the actual computation of the order $(n + K)$ inverse matrix in (16-14) and (16-15). Under the strategy chosen, to make $\mathbf{A}$ a multiple of the identity matrix, however, there are forms of partitioned inverse matrices that will allow solution to the actual computation.

Thus far, we have assumed that each α_i is generated by a different normal distribution, $-\boldsymbol{\gamma}_0$ and $\mathbf{A}$, however specified, have (potentially) different means and variances for the elements of $\boldsymbol{\alpha}$. The third specification we consider is one in which all α_i's in the model are assumed to be draws from the same population. To produce this specification, we use a **hierarchical prior** for the individual effects. The full model will be

$$
\begin{aligned}
y_{it} &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_{it} \sim \mathrm{N}[0, \sigma_\varepsilon^2], \\
p(\boldsymbol{\beta}|\sigma_\varepsilon^2) &= \mathrm{N}[\boldsymbol{\beta}_0, \sigma_\varepsilon^2\mathbf{A}], \\
p(\sigma_\varepsilon^2) &= \mathrm{Gamma}(\sigma_0^2, m), \\
p(\alpha_i) &= \mathrm{N}[\mu_\alpha, \tau_\alpha^2], \\
p(\mu_\alpha) &= \mathrm{N}[a, Q], \\
p(\tau_\alpha^2) &= \mathrm{Gamma}(\tau_0^2, \nu).
\end{aligned}
$$

[18] See, for example, Koop (2003, pp. 22–23), Zellner (1971, p. 20), and Cameron and Trivedi (2005, pp. 425–427).

We will not be able to derive the posterior density (joint or marginal) for the parameters of this model. However, it is possible to set up a Gibbs sampler that can be used to infer the characteristics of the posterior densities statistically. The sampler will be driven by conditional normal posteriors for the location parameters, $[\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma_\varepsilon^2, \mu_\alpha, \tau_\alpha^2]$, $[\alpha_i|\boldsymbol{\beta}, \sigma_\varepsilon^2, \mu_\alpha, \tau_\alpha^2]$, and $[\mu_\alpha|\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\varepsilon^2, \tau_\alpha^2]$ and conditional gamma densities for the scale (variance) parameters, $[\sigma_\varepsilon^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_\alpha, \tau_\alpha^2]$ and $[\tau_\alpha^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mu_\alpha]$.[19] The assumption of a common distribution for the individual effects and an independent prior for $\boldsymbol{\beta}$ produces a Bayesian counterpart to the random effects model.

16.8 HIERARCHICAL BAYES ESTIMATION OF A RANDOM PARAMETERS MODEL

We now consider a Bayesian approach to estimation of the random parameters model.[20] For an individual i, the conditional density for the dependent variable in period t is $f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i)$, where $\boldsymbol{\beta}_i$ is the individual specific $K \times 1$ parameter vector and $\mathbf{x}_{it}$ is individual specific data that enter the probability density.[21] For the sequence of T observations, assuming conditional (on $\boldsymbol{\beta}_i$) independence, person i's contribution to the likelihood for the sample is

$$f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_i) = \prod_{t=1}^{T} f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i), \tag{16-25}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})$ and $\mathbf{X}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}]$. We will suppose that $\boldsymbol{\beta}_i$ is distributed normally with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$. (This is the "hierarchical" aspect of the model.) The unconditional density would be the expected value over the possible values of $\boldsymbol{\beta}_i$,

$$f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{\beta}_i} \prod_{t=1}^{T} f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_i)\phi_K[\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}]d\boldsymbol{\beta}_i, \tag{16-26}$$

where $\phi_K[\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}]$ denotes the K variate normal prior density for $\boldsymbol{\beta}_i$ given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Maximum likelihood estimation of this model, which entails estimation of the deep parameters, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, then estimation of the individual specific parameters, $\boldsymbol{\beta}_i$ is considered in Sections 15.7 through 15.11. We now consider the Bayesian approach to estimation of the parameters of this model.

To approach this from a Bayesian viewpoint, we will assign noninformative prior densities to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. As is conventional, we assign a flat (noninformative) prior to $\boldsymbol{\beta}$.

[19]The procedure is developed at length by Koop (2003, pp. 152–153).

[20]Note that there is occasional confusion as to what is meant by *random parameters* in a random parameters (RP) model. In the Bayesian framework we discuss in this chapter, the "randomness" of the random parameters in the model arises from the uncertainty of the analyst. As developed at several points in this book (and in the literature), the randomness of the parameters in the RP model is a characterization of the heterogeneity of parameters across individuals. Consider, for example, in the Bayesian framework of this section, in the RP model, each vector $\boldsymbol{\beta}_i$ is a random vector with a distribution (defined hierarchically). In the classical framework, each $\boldsymbol{\beta}_i$ represents a single draw from a parent population.

[21]To avoid a layer of complication, we will embed the time-invariant effect $\Delta\mathbf{z}_i$ in $\mathbf{x}_{it}'\boldsymbol{\beta}$. A full treatment in the same fashion as the latent class model would be substantially more complicated in this setting (although it is quite straightforward in the maximum simulated likelihood approach discussed in Section 15.11).

The variance parameters are more involved. If it is assumed that the elements of $\boldsymbol{\beta}_i$ are conditionally independent, then each element of the (now) diagonal matrix $\boldsymbol{\Sigma}$ may be assigned the inverted gamma prior that we used in (16-13). A full matrix $\boldsymbol{\Sigma}$ is handled by assigning to $\boldsymbol{\Sigma}$ an **inverted Wishart** prior density with parameters scalar K and matrix $K \times \mathbf{I}$.[22] This produces the joint posterior density,

$$\Lambda(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} \,|\, \text{all data}) = \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T} f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\beta}_i) \phi_K[\boldsymbol{\beta}_i | \boldsymbol{\beta}, \boldsymbol{\Sigma}] \right\} \times p(\boldsymbol{\beta}, \boldsymbol{\Sigma}). \quad \textbf{(16-27)}$$

This gives the joint density of all the unknown parameters conditioned on the observed data. Our Bayesian estimators of the parameters will be the posterior means for these $(n + 1)K + K(K + 1)/2$ parameters. In principle, this requires integration of (16-27) with respect to the components. As one might guess at this point, that integration is hopelessly complex and not remotely feasible.

However, the techniques of Markov chain Monte Carlo (MCMC) simulation estimation (the Gibbs sampler) and the **Metropolis–Hastings algorithm** enable us to sample from the (only seemingly hopelessly complex) joint density $\Lambda(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} \,|\, \textbf{all data})$ in a remarkably simple fashion. Train (2001 and 2002, Chapter 12) describes how to use these results for this random parameters model.[23] The usefulness of this result for our current problem is that it is, indeed, possible to partition the joint distribution, and we can easily sample from the conditional distributions. We begin by partitioning the parameters into $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\boldsymbol{\delta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$. Train proposes the following strategy: To obtain a draw from $\boldsymbol{\gamma} | \boldsymbol{\delta}$, we will use the Gibbs sampler to obtain a draw from the distribution of $(\boldsymbol{\beta} | \boldsymbol{\Sigma}, \boldsymbol{\delta})$ and then one from the distribution of $(\boldsymbol{\Sigma} | \boldsymbol{\beta}, \boldsymbol{\delta})$. We will lay out this first, then turn to sampling from $\boldsymbol{\delta} | \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

Conditioned on $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ has a K-variate normal distribution with mean $\overline{\boldsymbol{\beta}} = (1/n) | \Sigma_{i=1}^{n} \boldsymbol{\beta}_i$ and covariance matrix $(1/n)\boldsymbol{\Sigma}$. To sample from this distribution we will first obtain the Cholesky factorization of $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$ where $\mathbf{L}$ is a lower triangular matrix. (See Section A.6.11.) Let $\mathbf{v}$ be a vector of K draws from the standard normal distribution. Then, $\overline{\boldsymbol{\beta}} + \mathbf{L}\mathbf{v}$ has mean vector $\overline{\boldsymbol{\beta}} + \mathbf{L} \times \mathbf{0} = \overline{\boldsymbol{\beta}}$ and covariance matrix $\mathbf{L}\mathbf{I}\mathbf{L}' = \boldsymbol{\Sigma}$, which is exactly what we need. So, this shows how to sample a draw from the conditional distribution $\boldsymbol{\beta}$.

To obtain a random draw from the distribution of $\boldsymbol{\Sigma} | \boldsymbol{\beta}, \boldsymbol{\delta}$, we will require a random draw from the inverted Wishart distribution. The marginal posterior distribution of $\boldsymbol{\Sigma} | \boldsymbol{\beta}, \boldsymbol{\delta}$ is inverted Wishart with parameters scalar $K + n$ and matrix $\mathbf{W} = (K\mathbf{I} + n\mathbf{V})$, where $\mathbf{V} = (1/n) \sum_{i=1}^{n} (\boldsymbol{\beta}_i - \overline{\boldsymbol{\beta}})(\boldsymbol{\beta}_i - \overline{\boldsymbol{\beta}})'$. Train (2001) suggests the following strategy for sampling a matrix from this distribution: Let $\mathbf{M}$ be the lower triangular Cholesky factor of $\mathbf{W}^{-1}$, so $\mathbf{M}\mathbf{M}' = \mathbf{W}^{-1}$. Obtain $K + n$ draws of $\mathbf{v}_k = K$ standard normal variates. Then, obtain $\mathbf{S} = \mathbf{M}\left(\sum_{k=1}^{K+n} \mathbf{v}_k \mathbf{v}_k' \right)\mathbf{M}'$. Then $\boldsymbol{\Sigma}^j = \mathbf{S}^{-1}$ is a draw from the inverted Wishart distribution. [This is fairly straightforward, as it involves only random sampling from the standard normal distribution. For a diagonal $\boldsymbol{\Sigma}$ matrix, that is, uncorrelated parameters

[22] The Wishart density is a multivariate counterpart to the chi-squared distribution. Discussion may be found in Zellner (1971, pp. 389–394) and Gelman (2003).

[23] Train describes the use of this method for mixed (random parameters) multinomial logit models. By writing the densities in generic form, we have extended his result to any general setting that involves a parameter vector in the fashion described above. The classical version of this appears in Section 15.11 for the binomial probit model and in Section 18.2.7 for the mixed logit model.

in $\boldsymbol{\beta}_i$, it simplifies a bit further. A draw for the nonzero kth diagonal element can be obtained using $(1 + n\mathbf{V}_{kk})/\sum_{k=1}^{K+n} v_{rk}^2$.]

The difficult step is sampling $\boldsymbol{\beta}_i$. For this step, we use the Metropolis–Hastings (M–H) algorithm suggested by Chib and Greenberg (1995, 1996) and Gelman et al. (2004). The procedure involves the following steps:

1. Given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ and "tuning constant" τ (to be described next), compute $\mathbf{d} = \tau\mathbf{L}\mathbf{v}$ where $\mathbf{L}$ is the Cholesky factorization of $\boldsymbol{\Sigma}$ and $\mathbf{v}$ is a vector of K independent standard normal draws.
2. Create a trial value $\boldsymbol{\beta}_{i1} = \boldsymbol{\beta}_{i0} + \mathbf{d}$ where $\boldsymbol{\beta}_{i0}$ is the previous value.
3. The posterior distribution for $\boldsymbol{\beta}_i$ is the likelihood that appears in (16-26) times the joint normal prior density, $\phi_K[\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}]$. Evaluate this posterior density at the trial value $\boldsymbol{\beta}_{i1}$ and the previous value $\boldsymbol{\beta}_{i0}$. Let

$$R_{10} = \frac{f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_{i1})\phi_K(\boldsymbol{\beta}_{i1}|\boldsymbol{\beta}, \boldsymbol{\Sigma})}{f(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}_{i0})\phi_K(\boldsymbol{\beta}_{i0}|\boldsymbol{\beta}, \boldsymbol{\Sigma})}.$$

4. Draw one observation, u, from the standard uniform distribution, $U[0, 1]$.
5. If $u < R_{10}$, then accept the trial (new) draw. Otherwise, reuse the old one.

This M–H iteration converges to a sequence of draws from the desired density. Overall, then, the algorithm uses the Gibbs sampler and the Metropolis–Hastings algorithm to produce the sequence of draws for all the parameters in the model. The sequence is repeated a large number of times to produce each draw from the joint posterior distribution. The entire sequence must then be repeated N times to produce the sample of N draws, which can then be analyzed, for example, by computing the posterior mean.

Some practical details remain. The tuning constant, τ, is used to control the iteration. A smaller τ increases the acceptance rate. But at the same time, a smaller τ makes new draws look more like old draws so this slows down the process. Gelman et al. (2004) suggest $\tau = 0.4$ for $K = 1$ and smaller values down to about 0.23 for higher dimensions, as will be typical. Each multivariate draw takes many runs of the MCMC sampler. The process must be started somewhere, though it does not matter much where. Nonetheless, a "burn-in" period is required to eliminate the influence of the starting value. Typical applications use several draws for this burn-in period for each run of the sampler. How many sample observations are needed for accurate estimation is not certain, though several hundred would be a minimum. This means that there is a huge amount of computation done by this estimator. However, the computations are fairly simple. The only complicated step is computation of the acceptance criterion at step 3 of the M–H iteration. Depending on the model, this may, like the rest of the calculations, be quite simple.

Example 16.7 Bayesian and Classical Estimation of Heterogeneity in the Returns to Education

Koop and Tobias (2004) study individual heterogeneity in the returns to education using a panel data set from the National Longitudinal Survey of Youth (NLSY). In a wage equation such as

$$\ln Wage_{it} = \theta_{1,i} + \theta_{2,i}\, Education_{it} + \gamma_1 Experience_{it} + \gamma_2 Experience_{it}^2 + \gamma_3 Time_{it} + \gamma_4 Unemp_{it} + \varepsilon_{it}, \quad \textbf{(16-28)}$$

individual heterogeneity appears in the intercept and in the returns to education. Received estimates of the returns to education, θ_2 here, computed using OLS, are biased due to the

endogeneity of *Education* in the equation. The missing variables would include ability and motivation. Instrumental variable approaches will mitigate the problem (and IV estimators are typically larger than OLS), but the authors are concerned that the results might be specific to the instrument used. They cite the example of using as an instrumental variable a dummy variable for presence of a college in the county of residence, by which the IV estimator will deliver the returns to education for those who attend college given that there is a college in their county, but not for others (the *local average treatment effect* rather than the *average treatment effect*). They propose a structural approach based on directly modeling the heterogeneity. They examine several models including random parameters (continuous variation) and latent class (discrete variation) specifications. They propose extensions of the familiar models by introducing covariates into the heterogeneity model (see Example 15.16) and by exploiting time variation in schooling as part of the identification strategy. Bayesian methods are used for the estimation and inference.[24]

Several models are considered. The one most preferred is the hierarchical linear model examined in Example 15.16:

$$\begin{aligned}\theta_{1,i} &= \theta_{1,0} + \lambda_{1,1}\textit{Ability}_i + \lambda_{1,2}\textit{Mother's Education}_i + \lambda_{1,3}\,\textit{Father's Education}_i + \\ &\quad + \lambda_{1,4}\,\textit{Broken Home}_i + \lambda_{1,5}\,\textit{Siblings}_i + u_{1,i},\\ \theta_{2,i} &= \theta_{2,0} + \lambda_{2,1}\,\textit{Ability}_i + \lambda_{2,2}\,\textit{Mother's Education}_i + \lambda_{2,3}\,\textit{Father's Education}_i \\ &\quad + \lambda_{2,4}\,\textit{Broken Home}_i + \lambda_{2,5}\,\textit{Siblings}_i + u_{2,i}. \end{aligned} \qquad \textbf{(16-29)}$$

The candidate models are framed as follows:

$y_{it}\,\vert\,\mathbf{x}_{it},\mathbf{z}_{it},\boldsymbol{\theta}_i,\boldsymbol{\gamma},\boldsymbol{\sigma}_\varepsilon^2$	$\sim N[\mathbf{x}_{it}'\boldsymbol{\theta}_i + \mathbf{z}_{it}'\boldsymbol{\gamma},\ \boldsymbol{\sigma}_\varepsilon^2]$	(main regression model),
$\boldsymbol{\gamma}\,\vert\,\boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma$	$\sim N[\boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma]$	(normal distribution for location parameters),
$\sigma_\varepsilon^{-2}\,\vert\, s_\varepsilon^{-2}, \eta_\varepsilon$	$\sim G(s_\varepsilon^{-2},\eta_\varepsilon)$	(gamma distribution for $1/\sigma_\varepsilon^2$),
$\boldsymbol{\theta}_i\,\vert\,\boldsymbol{\lambda}, \mathbf{w}_i$	$\sim f(\boldsymbol{\theta}_i\,\vert\,\boldsymbol{\lambda},\mathbf{w}_i)$	(varies by model, discrete or continuous),
$\boldsymbol{\lambda}\,\vert\,\underline{\boldsymbol{\lambda}}$	$\sim g(\underline{\boldsymbol{\lambda}})$	(varies by model).

The models for $\boldsymbol{\theta}_i\,\vert\,\boldsymbol{\lambda},\mathbf{w}_i$ are either discrete or continuous distributions, parameterized in terms of a vector of parameters, $\boldsymbol{\lambda}$ and a vector of time-invariant variables, w_i. [*Note,* for example, (16-29).] The model for the regression slopes, $\underline{\boldsymbol{\gamma}}$, and the regression variance, σ_ε^2, will be common to all the specifications. The models for the heterogeneity, $\boldsymbol{\theta}_i\,\vert\,\boldsymbol{\lambda},\mathbf{w}_i$ and for $\boldsymbol{\lambda}\,\vert\,\underline{\boldsymbol{\lambda}}$ will vary with the specification. The models considered are:

1. $\theta_{1,i} = \theta_{1,0}$ and $\theta_{2,i} = \theta_{2,0}$, no heterogeneity (−24,212),
2. $\boldsymbol{\theta}_i \sim N[\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_u]$, a simple random parameters model (−15,886),
3. $\theta_{1,i} \sim N[\theta_1,\sigma_{u1}^2]$, $\theta_{2,i} = \theta_{2,0}$, a random effects model (−16,501),
4. $\boldsymbol{\theta}_i = \boldsymbol{\theta}_g^0$ with probability π_g, a latent class model(−16,528),
5. $f(\boldsymbol{\theta}_i) = \Sigma_g\ \pi_g\ \phi(\boldsymbol{\theta}_i\,\vert\,\boldsymbol{\theta}_g^0, \Sigma_g)$, a finite mixture of normal (−15,898).

(The BIC values for model selection reported in the study are shown in parentheses. These are discussed further below.) The preferred model is model 2 with mean function $\boldsymbol{\theta}_0 + \boldsymbol{\Lambda}\mathbf{w}_i$. This is

[24]The authors note, "Although the length of our panel is rather short, this does not create a significant problem for us as we employ a Bayesian approach which provides exact finite sample results." It is not clear at this point what problem is caused by the short panel—actually, for most of the sample the panel is reasonably long (see Figure 15.7)—or how exact inference mitigates that problem. Likewise, "estimates of the individual-level parameters obtained from our hierarchical model incorporate not only information from the outcomes of that individual, but also incorporate information obtained from the other individuals in the sample." As the authors carefully note later, they do not actually compute individual specific estimates, but rather conditional means for individuals with specific characteristics. (Both from p. 828.)

(16-28) and (16-29). Model 4 could be also augmented with $\mathbf{w}_i$. This would be a latent class model with $prob(\theta_i = \theta_g^0) = \exp(\mathbf{w}_i'\boldsymbol{\lambda}_g)/\sum_{g=1}^{G}\exp(\mathbf{w}_i'\boldsymbol{\lambda}_g)$. This model is developed in Section 14.15.2. Estimates based on this latent class formulation are shown below.

The data set is an unbalanced panel of 2,178 individuals, altogether 17,919 person-year observations with T_i ranging from 1 to 15. (See Figure 15.7.) Means of the data are given in Example 15.16.[25] Most of the analysis is based on the full data set. However, models involving the time-invariant variables were estimated using 1,694 individuals (14,170 person-year observations) whose parents have at least 9 years of education. A Gibbs sampler is used with 11,000 repetitions; the first 1,000 are discarded as the burn-in. (The Gibbs sampler, priors, and other computational details are provided in an appendix in the paper.) Two devices are proposed to choose among the models. First, the posterior odds ratio in Section 16.4.3 is computed. With equal priors for the models, the posterior odds equals the likelihood ratio, which is computed for two models, A and B, as $\exp(\ln L^A - \ln L^B)$. The log likelihoods for models 1, 2, and 3 are −12,413, −8,046, and −8,153. Small differences in the log likelihoods always translate to huge differences in the posterior odds. For these cases, the posterior odds in favor of model 2 against model 3 are exp(107), which is overwhelming ("massive"). (The log likelihood for the version of model 2 in Example 15.16 is −7,983, which is also vastly better than the model 2 here by this criterion.) A second criterion is the Bayesian information criterion, which is $2\ln L - K\ln n$, where K is the number of parameters estimated and n is the number of individuals (2,170 or 1,694). The BICs for the five models are listed above with the model specifications. The model with no heterogeneity is clearly rejected. Among the others, Model 2, the random parameters specification, is preferred by a wide margin. Model 5, the mixture of two normal distributions with heterogeneous means, is second, followed by Model 3, the random effects model. Model 4, the latent class model, is clearly the least preferred.

Continuous Distribution of Heterogeneity

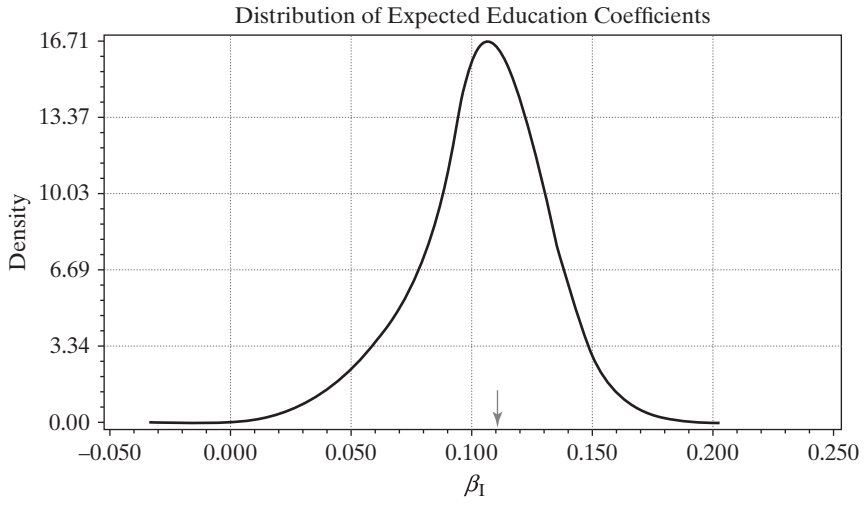
The main results for the study are based on the subsample and Model 2. The reported posterior means of the coefficient distributions of (16-29) are shown in the right panel in Table 16.4. (Results are extracted from Tables IV and V in the paper.) We re-estimated (16-28) and (16-29) using the methods of Sections 15.7 and 15.8. The estimates of the parameters

TABLE 16.4 Estimated Wage Equations

	Random Parameters Model		*Koop–Tobias Posterior Means*	
Variable	***Constant***	***Education***	***Constant***	***Education***
Exp	0.12621		0.126	
Exp^2	−0.00388		−0.004	
Time	−0.01787		−0.024	
Unemployment			−0.004	
Constant	0.39277	0.09578	0.797	0.070
Ability	−0.13177	0.01568	−0.073	0.0125
Mother's Educ	0.02864	−0.00167	0.021	−0.001
Father's Educ	0.00242	−0.00022	−0.022	0.002
Broken Home	0.12963	−0.01640	0.115	−0.015
Number Siblings	−0.08323	0.00659	−0.079	0.007

[25]The unemployment rate variable in (16-28) is not included in the JAE archive data set that we have used to partially replicate this study in Example 15.16 and here.

FIGURE 16.1 Random Parameters Estimate of Expected Returns.



of the model are shown in the left panel of Table 16.4. Overall, the mean return is about 11% (0.11). We did the same analysis with the classical results based on Section 15.10. The individual specific estimates are summarized in Figure 16.1 (which is nearly identical to the authors' Figure 3). The results are essentially the same as Koop and Tobias's. The differences are attributable to the different methodologies – the prior distributions will have at least some influence on the results – and to our omission of the unemployment rate from the main equation. The authors' reported results suggest that the impact of the unemployment rate on the results is minor, which would suggest that the differences in the estimated results primarily reflect the different approaches to the analysis. The similarity of the end results would be anticipated by the Bernstein–von Mises theorem. (See Section 16.4.4.)

Discrete Distribution of Heterogeneity

Model 4 in the study is a latent class model. The authors fit a model with $G = 10$ classes. The model is a Heckman and Singer style (Section 14.15.7) specification in that the coefficients on the time-varying variables are the same in all 10 classes. The class probabilities are specified as fixed constants. This provides a discrete distribution for the heterogeneity in $\boldsymbol{\theta}_i$. Model 4 was the least preferred model among the candidates.

We fit a 5 segment latent class model based on (16-28) and (16-29). The parameters on the time-varying variables in (16-28) are the same in all classes—only the constant terms and the education coefficients differ across the classes. The class probabilities are built on the time-invariant effects, ability, parent's education, etc. (The authors do not report a model with this form of heterogeneity.) The log likelihood for this extension of the model is

$$\ln L = \sum_{i=1}^{n} \ln \sum_{g=1}^{G} \pi_{ig}(\mathbf{w}_i) \left(\prod_{t=1}^{T_i} f(y_{it} | \theta_{0,g} + \theta_{1,g}\, Education_{it} + \mathbf{z}_{it}'\boldsymbol{\gamma}) \right) \qquad \textbf{(16-30)}$$

$$\pi_{ig}(\mathbf{w}_i) = \frac{\exp(\mathbf{w}_i'\boldsymbol{\lambda}_g)}{\sum_{g=1}^{G} \exp(\mathbf{w}_i'\boldsymbol{\lambda}_g)}.$$

Using the suggested subsample, the log likelihood for the model in (16-30) is 6235.02. When the time-invariant variables are not included in the class probabilities, the log likelihood falls to 6192.66. By a standard likelihood ratio test, the chi squared is 84.72, with 20 degrees of freedom (the 5 additional coefficients in G-1 of the class probabilities). The critical chi squared is 31.02. We computed $E[\theta_{1,i} | \textbf{data}]$ for each individual based on the estimated posterior class probabilities as

$$\hat{E}[\theta_{1,i}] = \Sigma_{g=1}^{G} \hat{\pi}_{ig}(\theta_{1,g} | \mathbf{w}_i, \textbf{data}_i)\hat{\theta}_{1,g}. \quad \textbf{(16-31)}$$

(See Section 14.15.4.) The overall estimate of returns to education is the sample average of these, 0.107. Figure 16.2 shows the results of this computation for the 1,694 individuals. We then used the method in Section 14.15.4. to estimate the class assignments and computed the means of the expected returns for the individuals assigned to each of the 5 classes. The results are shown in Table 16.5. Finally, because we now have a complete (estimated) assignment of the individuals, we constructed in Figure 16.3 a comparison of distributions of the expected coefficients in each of the 5 classes.

This analysis has examined the heterogeneity in the returns to education by a variety of model specifications. In the end, the results are quite consistent across the different models and based on the two methodologies.

16.9 SUMMARY AND CONCLUSIONS

This chapter has introduced the major elements of the Bayesian approach to estimation and inference. The contrast between Bayesian and classical, or frequentist, approaches to the analysis has been the subject of a decades-long dialogue among

FIGURE 16.2 Estimated Distribution of Expected Returns Based on Latent Class Model.

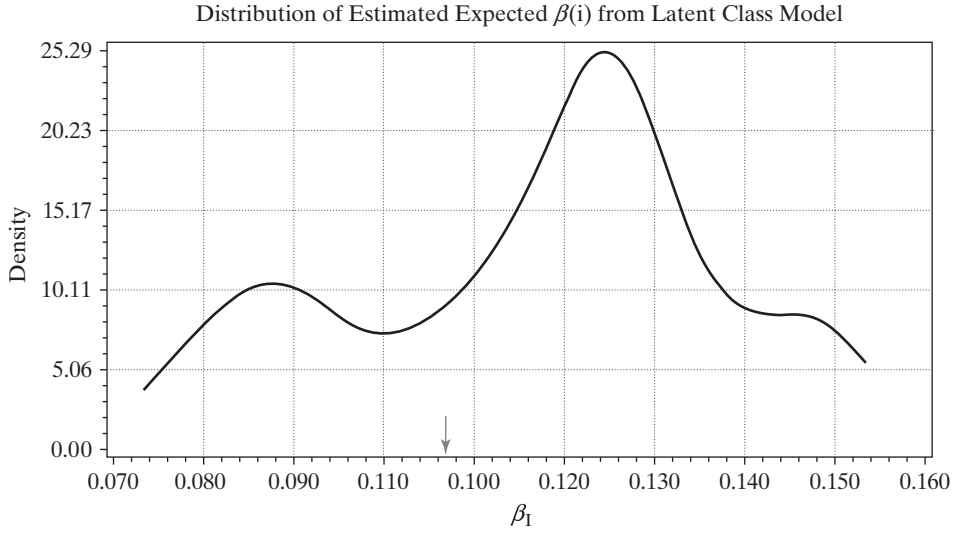
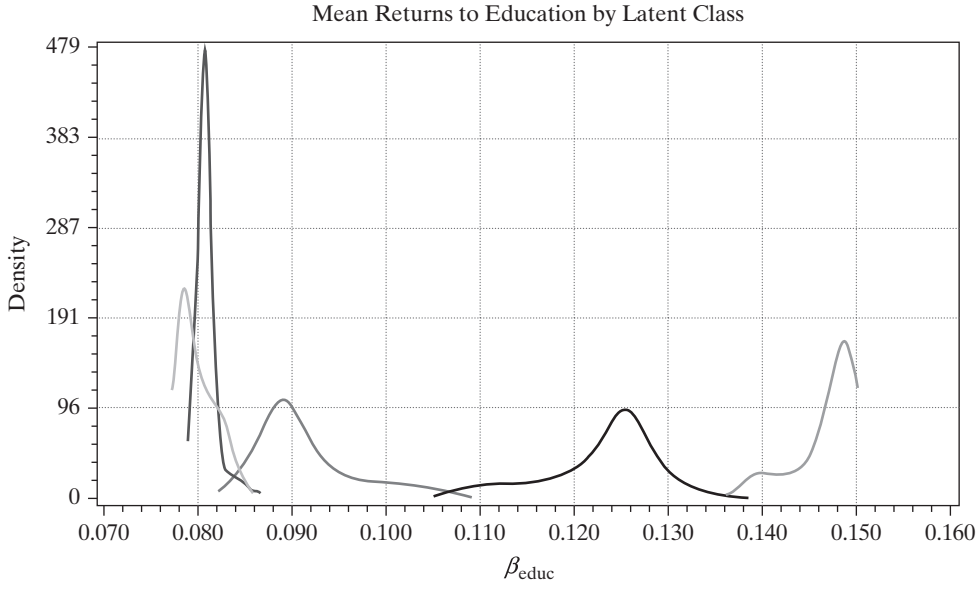


TABLE 16.5 Estimated Expected Returns to Schooling by Class

Class	*Mean*	n_g
1	0.147	167
2	0.092	608
3	0.080	189
4	0.123	640
5	0.083	90
Full Sample	0.107	1,694

FIGURE 16.3 Kernel Density Estimates of Expected Returns by Class.



practitioners and philosophers. As the frequency of applications of Bayesian methods has grown dramatically in the modern literature, however, the approach to the body of techniques has typically become more pragmatic. The Gibbs sampler and related techniques including the Metropolis–Hastings algorithm have enabled some remarkable simplifications of previously intractable problems. For example, recent developments in commercial software have produced a wide choice of mixed estimators which are various implementations of the maximum likelihood procedures and hierarchical Bayes procedures (such as the *Sawtooth and MLWin* programs). Unless one is dealing with a small sample, the choice between these can be based on convenience. There is little methodological difference. This returns us to the practical point noted earlier. The choice between the Bayesian approach and the sampling theory method in this application would not be based on a fundamental methodological criterion, but on purely practical considerations—the end result is largely the same.

This chapter concludes our survey of estimation and inference methods in econometrics. We will now turn to two major areas of applications, microeconometrics in Chapters 17–19. which is primarily oriented to cross-section and panel data applications, and time series and (broadly) macroeconometrics in Chapters 20 and 21.

Key Terms and Concepts

- Bayes factor
- Bayes' theorem
- Bernstein–von Mises theorem
- Burn in
- Conjugate prior
- Data augmentation
- Gibbs sampler
- Hierarchical prior
- Highest posterior density (HPD) interval
- Improper prior
- Informative prior
- Inverted gamma distribution
- Inverted Wishart
- Joint posterior distribution
- Likelihood function
- Loss function
- Markov chain Monte Carlo (MCMC)
- Metropolis–Hastings algorithm
- Multivariate *t* distribution
- Noninformative prior
- Normal-gamma prior
- Posterior density
- Posterior mean
- Precision matrix
- Predictive density
- Prior beliefs
- Prior density
- Prior distribution
- Prior odds ratio
- Prior probabilities
- Sampling theory
- Uniform-inverse gamma prior
- Uniform prior

Exercise

1. Suppose the distribution of $y_i|\lambda$ is Poisson,

$$f(y_i|\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} = \frac{\exp(-\lambda)\lambda^{y_i}}{\Gamma(y_i + 1)}, \quad y_i = 0, 1, \ldots, \lambda > 0.$$

We will obtain a sample of observations, $y_i, \ldots, y_n$. Suppose our prior for λ is the inverted gamma, which will imply

$$p(\lambda) \propto \frac{1}{\lambda}.$$

a. Construct the likelihood function, $p(y_1, \ldots, y_n|\lambda)$.
b. Construct the posterior density,

$$p(\lambda|y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n|\lambda)p(\lambda)}{\int_0^\infty p(y_1, \ldots, y_n|\lambda)p(\lambda)d\lambda}.$$

c. Prove that the Bayesian estimator of λ is the posterior mean, $E[\lambda|y_1, \ldots, y_n] = \bar{y}$.
d. Prove that the posterior variance is $\text{Var}[\lambda|y_l, \ldots, y_n] = \bar{y}/n$.
(*Hint:* You will make heavy use of gamma integrals in solving this problem. Also, you will find it convenient to use $\Sigma_i y_i = n\bar{y}$.)

Applications

1. Consider a model for the mix of male and female children in families. Let K_i denote the family size (number of children), $K_i = 1, \ldots$. Let F_i denote the number of female children, $F_i = 0, \ldots, K_i$. Suppose the density for the number of female children in a family with K_i children is binomial with constant success probability θ:

$$p(F_i | K_i, \theta) = \binom{K_i}{F_i} \theta^{F_i}(1 - \theta)^{K_i - F_i}.$$

We are interested in analyzing the "probability," θ. Suppose the (conjugate) prior over θ is a beta distribution with parameters a and b:

$$p(\theta) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1}.$$

Your sample of 25 observations is given here:

K_i	2	1	1	5	5	4	4	5	1	2	4	4	2	4	3	2	3	2	3	5	3	2	5	4	1
F_i	1	1	1	3	2	3	2	4	0	2	3	1	1	3	2	1	3	1	2	4	2	1	1	4	1

   a. Compute the classical maximum likelihood estimate of θ.
   b. Form the posterior density for θ given (K_i, F_i), $i = 1, \ldots, 25$ conditioned on a and b.
   c. Using your sample of data, compute the posterior mean assuming $a = b = 1$.
   d. Using your sample of data, compute the posterior mean assuming $a = b = 2$.
   e. Using your sample of data, compute the posterior mean assuming $a = 1$ and $b = 2$.

17
BINARY OUTCOMES AND DISCRETE CHOICES

17.1 INTRODUCTION

This is the first of three chapters that will survey models used in **microeconometrics**. The analysis of individual choice that is the focus of this field is fundamentally about modeling discrete outcomes such as purchase decisions, whether or not to buy insurance, voting behavior, choice among a set of alternative brands, travel modes or places to live, and responses to survey questions about the strength of preferences or about self-assessed health or well-being. In these and any number of other cases, the *dependent variable* is not a quantitative measure of some economic outcome, but rather an indicator of whether or not some outcome has occurred. It follows that the regression methods we have used up to this point are largely inappropriate. We turn, instead, to modeling probabilities and using econometric tools to make probabilistic statements about the occurrence of these events. We will also examine models for counts of occurrences. These are closer to familiar regression models, but are, once again, about discrete outcomes of behavioral choices. As such, in this setting as well, we will be modeling probabilities of events, rather than conditional mean functions.

The models used in this area of study are inherently (and intrinsically) nonlinear. We have developed some of the elements of nonlinear modeling in Chapters 7 and 14. Those elements are combined in whole in the study of discrete choices. This chapter will focus on binary choices, where *the model* is the probability of an event. Many general treatments of nonlinear modeling in econometrics, in fact, focus on only this segment of the field. This is reasonable. Nearly the full set of results used more broadly, for specification, estimation, inference, and analysis can be developed and understood in this particular application. We will take that approach here. Several of the parts of nonlinear modeling will be developed in detail in this chapter, then invoked or extended in straightforward ways in the chapters to follow.

The models that are analyzed in this and Chapter 18 are built on a platform of preferences of decision makers. We take a **random utility** view of the choices that are observed. The decision maker is faced with a situation or set of alternatives and reveals something about his or her underlying preferences by the choice that he or she makes. The choice(s) made will be affected by observable influences—this is, for example, the ultimate objective of advertising—and by unobservable characteristics of the chooser. The blend of these fundamental bases for individual choice is at the core of the broad range of models that we will examine here.[1]

[1]See Greene and Hensher (2010, Chapter 4) for a historical perspective on this approach to model specification.

This chapter and Chapter 18 will describe four broad frameworks for analysis. The first is the simplest:

Binary Choice: The individual faces two choices and makes that choice between the two that provides the greater utility. Many such settings involve the choice between taking an action and not taking that action, for example, the decision whether or not to purchase health insurance. In other cases, the decision might be between two distinctly different choices, such as the decision whether to travel to and from work via public or private transportation. In the binary choice case, the 0/1 outcome is merely a label for "no/yes"—the numerical values are a mathematical convenience. This chapter will present a lengthy survey of models and methods for binary choices.

The binary choice case naturally extends to cases of more than two outcomes. For one example, in our our travel mode case, the individual choosing private transport might choose between private transport as driver and private transport as passenger, or public transport by train or by bus. Such multinomial (many named) choices are *unordered*. Another case is one that is a constant staple of the online experience. Instead of being asked a binary choice, "Did you like our service?," the hapless surfer will be asked an *ordered* multinomial choice, "On a scale from 1 to 5, how much did you like our service?"

Multinomial Choice: The individual chooses among more than two choices, once again, making the choice that provides the greatest utility. At one level, this is a minor variation of the binary choice case—the latter is, of course, a special case of the former. But more elaborate models of multinomial choice allow a rich specification of consumer preferences. In the multinomial case, the observed response is again a label for the selected choice; it might be a brand, the name of a place, or the type of travel mode. Numerical assignments are not meaningful in this setting.

Ordered Choice: The individual reveals the strength of his or her preferences with respect to a single outcome. Familiar cases involve survey questions about strength of feelings about a particular commodity, such as a movie, or self-assessments of social outcomes such as health in general or self-assessed well-being. In the ordered choice setting, opinions are given meaningful numeric values, usually $0, 1, \ldots, J$ for some upper limit, J. For example, opinions might be labeled 0, 1, 2, 3, 4 to indicate the strength of preferences for a product, a movie, a candidate or a piece of legislation. But in this context, the numerical values are only a ranking, not a quantitative measure. Thus, a "1" is greater than a "0" only in a qualitative sense, not by one unit, and the difference between a "2" and a "1" is not the same as that between a "1" and a "0."

In these three cases, although the numerical outcomes are merely labels of some nonquantitative outcome, the analysis will nonetheless have a regresson-style motivation. Throughout, the models will be based on the idea that observed covariates are relevant in explaining the observed choices and in how changes in those attributes can help explain variation in choices. For example, in the binary outcome "did or did not purchase health insurance," a conditioning model suggests that covariates such as age, income, and family situation will help explain the choice. Chapter 18 will describe a range of models that have been developed around these considerations.

We will also be interested in a fourth application of discrete outcome models:

Event Counts: The observed outcome is a count of the number of occurrences. In many cases, this is similar to the preceding three settings in that the dependent variable

measures an individual choice, such as the number of visits to the physician or the hospital, the number of derogatory reports in one's credit history, the number of vehicles in a household's capital stock, or the number of visits to a particular recreation site. In other cases, the event count might be the outcome of some natural process, such as the occurrence rate of a disease in a population or the number of defects per unit of time in a production process. In these settings, we will be doing a more familiar sort of regression modeling. However, the models will still be constructed specifically to accommodate the discrete (and nonnegative) nature of the observed response variable and the modeling of probabilities of occurrences of events rather than some measure of the events themselves.

We will consider these four cases in turn. The four broad areas have many elements in common; however, there are also substantive differences between the particular models and analysis techniques used in each. This chapter will develop the first topic, models for binary choices. In each section, we will include several applications and present the single basic model that is the centerpiece of the methodology, and, finally, examine some recently developed extensions of the model. This chapter contains a very lengthy discussion of models for binary choices. This analysis is as long as it is because, first, the models discussed are used throughout microeconometrics—the central model of binary choice in this area is as ubiquitous as linear regression. Second, all the econometric issues and features that are encountered in the other areas will appear in the analysis of binary choice, where we can examine them in a fairly straightforward fashion.

It will emerge that, at least in econometric terms, the models for multinomial and ordered choice considered in Chapter 18 can be built from the two fundamental building blocks, the model of random utility and the translation of that model into a description of binary choices. There are relatively few new econometric issues that arise here. Chapter 18 will be largely devoted to suggesting different approaches to modeling choices among multiple alternatives and models for ordered choices. Once again, models of preference scales, such as movie or product ratings, or self-assessments of health or well-being, can be naturally built up from the fundamental model of random utility. Finally, Chapter 18 will develop the well-known Poisson regression model for counts of events. We will then extend the model to demonstrate some recent applications and innovations.

Chapters 17 and 18 are a lengthy but far from complete survey of topics in estimating **qualitative response (QR)** models. In general, because the outcome variable in the first three of these four cases is merely the name of an event, not the event itself, linear regression will be an inappropriate approach. In most cases, the method of estimation is maximum likelihood.[2] Therefore, readers interested in the mechanics of estimation may want to review the material in Appendices D and E before continuing. The various properties of maximum likelihood estimators are discussed in Chapter 14. We shall assume throughout these chapters that the necessary conditions behind the optimality properties of maximum likelihood estimators are met and, therefore, we will not derive or establish these properties specifically for the QR models. Detailed proofs for most of these models

[2]In the binary choice case, it is possible arbitrarily to assign two numerical values to the outcomes, typically 0 and 1, and "linearly regress" this constructed variable on the covariates. We will examine this strategy at some length with an eye to what information it reveals. The strategy would make little sense in the multinomial choice cases. Since the count data case is, in fact, a quantitative regression setting, the comparison of a linear regression approach to the intrinsically nonlinear regression approach is worth a close look.

can be found in surveys by Amemiya (1981), McFadden (1984), Maddala (1983), and Dhrymes (1984). Additional commentary on some of the issues of interest in the contemporary literature is given by Manski and McFadden (1981) and Maddala and Flores-Lagunes (2001). Agresti (2002) and Cameron and Trivedi (2005) contain numerous theoretical developments and applications. Greene (2008) and Greene and Hensher (2010) provide, among many others, general surveys of discrete choice models and methods.[3]

17.2 MODELS FOR BINARY OUTCOMES

For purposes of studying individual behavior, we will construct models that link a decision or outcome to a set of factors, at least in the spirit of regression. Our approach will be to analyze each of them in the general framework of probability models:

$$\text{Prob(event } j \text{ occurs}|\mathbf{x}) = \text{Prob}(Y = j|\mathbf{x}) = F(\text{relevant effects, parameters, } \mathbf{x}). \quad \textbf{(17-1)}$$

The study of qualitative choice focuses on appropriate specification, estimation, and use of models for the probabilities of events, where in most cases, the *event* is an individual's choice among a set of two or more alternatives. Henceforth, we will use the shorthand,

$$\text{Prob}(Y = 1|\mathbf{x}) = \text{Probability that event of interest occurs}|\mathbf{x},$$

and, naturally, $\text{Prob}(Y = 0|\mathbf{x}) = [1 - \text{Prob}(Y = 1|\mathbf{x})]$ is the probability that the event does not occur.

Example 17.1 Labor Force Participation Model

In Example 5.2, we estimated an earnings equation for the subsample of 428 married women who participated in the formal labor market taken from a full sample of 753 observations. The semilog earnings equation is of the form

$$\ln \textit{earnings} = \beta_1 + \beta_2\, \textit{age} + \beta_3\, \textit{age}^2 + \beta_4\, \textit{education} + \beta_5\, \textit{kids} + \varepsilon,$$

where *earnings* is *hourly wage* times *hours worked, education* is measured in years of schooling, and *kids* is a binary variable that equals one if there are children under 18 in the household. What of the other 325 individuals? The underlying labor supply model described a market in which labor force participation is the outcome of a market process whereby the demanders of labor services are willing to offer a wage based on expected marginal product, and individuals themselves make a decision whether or not to accept the offer depending on whether it exceeds their own reservation wage. The first of these depends on, among other things, education, while the second (we assume) depends on such variables as age, the presence of children in the household, other sources of income (husband's), and marginal tax rates on labor income. The sample we used to fit the earnings equation contains data on all these other variables. The models considered in this chapter would be appropriate for modeling the outcome $y = 1$ (*in the labor force,* 428 observations) or 0 (*not in the labor force*, 325 observations). For example, we would be interested how and how significantly the presence of children in the household (*kids*) affects the labor force participation.

Models for explaining a binary dependent variable are typically motivated in two contexts. The labor force participation model in Example 17.1 describes a process of individual choice between two alternatives in which the choice is influenced by

[3]There are dozens of book-length surveys of discrete choice models. Two others that are heavily oriented to an application of these methods are Train (2009) and Hensher, Rose, and Greene (2015).

observable effects (children, tax rates) and *unobservable* aspects of the preferences of the individual. The relationship between voting behavior and income is another example. In other cases, the **binary choice model** arises in a setting in which the nature of the observed data dictates the special treatment of a binary dependent variable model. In these cases, the analyst is essentially interested in a regression-like model of the sort considered in Chapters 2 through 7. With data on the variable of interest and a set of covariates, they are interested in specifying a relationship between the former and the latter, more or less along the lines of the models we have already studied. For example, in a model of the demand for tickets for sporting events, in which the variable of interest is number of tickets, it could happen that the observation consists only of whether the sports facility was filled to capacity (demand greater than or equal to capacity so $Y = 1$) or not ($Y = 0$). The event here is still qualitative, but now it is constructed as an indicator of a censoring (or not) of an underlying continuous variable, in this case, unobserved true demand. It will generally turn out that the models and techniques used in both cases (and, indeed, the underlying structure) are the same. Nonetheless, it is useful to examine both of them.

17.2.1 RANDOM UTILITY

An interpretation of data on individual choices is provided by a random utility model. Let U_a and U_b represent an individual's utility of two choices. For example, U_a might be the utility of rental housing and U_b that of home ownership. The observed choice between the two reveals which one provides the greater utility, but not the underlying unobservable utilities. Hence, the observed indicator equals 1 if $U_a > U_b$ and 0 if $U_a \leq U_b$. If we define, $U = U_a - U_b$, then $Y = \mathbf{1}(U > 0)$ [where $\mathbf{1}$ (condition) equals 1 if condition is true and 0 if it is false]. This is precisely the same as the censoring case noted earlier.

A common formulation is the linear random utility model,

$$U_a = \mathbf{w}'\boldsymbol{\beta}_a + \mathbf{z}_a'\boldsymbol{\gamma}_a + \varepsilon_a \quad \text{and} \quad U_b = \mathbf{w}'\boldsymbol{\beta}_b + \mathbf{z}_b'\boldsymbol{\gamma}_b + \varepsilon_b. \tag{17-2}$$

In (17-2), the observable (measurable) vector of **characteristics** of the individual is denoted $\mathbf{w}$; this might include gender, age, income, and other demographics. The vectors $\mathbf{z}_a$ and $\mathbf{z}_b$ denote features (**attributes**) of the two choices that might be choice specific. In a voting context, for example, the attributes might be indicators of the competing candidates' positions on important issues. The random terms, ε_a and ε_b, represent the stochastic elements that are specific to and known only by the individual, but not by the observer (analyst). To continue our voting example, ε_a might represent an intangible, general preference for candidate a, such as party affiliation.

The completion of the model for the determination of the observed outcome (choice) is the revelation of the ranking of the preferences by the choice the individual makes. Thus, if we denote by $Y = 1$ the consumer's choice of alternative a, we infer from $Y = 1$ that $U_a > U_b$. Because the outcome is ultimately driven by the random elements in the utility functions, we have

$$\begin{aligned}
\text{Prob}[Y = 1 \,|\, \mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] &= \text{Prob}[U_a > U_b] \\
&= \text{Prob}[(\mathbf{w}'\boldsymbol{\beta}_a + \mathbf{z}_a'\boldsymbol{\gamma}_a + \varepsilon_a) - (\mathbf{w}'\boldsymbol{\beta}_b + \mathbf{z}_b'\boldsymbol{\gamma}_b + \varepsilon_b) > 0 \,|\, \mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] \\
&= \text{Prob}[\{\mathbf{w}'(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) + (\mathbf{z}_a'\boldsymbol{\gamma}_a - \mathbf{z}_b'\boldsymbol{\gamma}_b)\} + (\varepsilon_a - \varepsilon_b) > 0 \,|\, \mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] \\
&= \text{Prob}[\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 \,|\, \mathbf{x}],
\end{aligned}$$

where $\mathbf{x}'\boldsymbol{\beta}$ collects all the observable elements of the difference of the two utility functions and ε denotes the difference between the two random elements.

Example 17.2 Structural Equations for a Binary Choice Model

Nakosteen and Zimmer (1980) analyzed a model of migration based on the following structure:[4] For a given individual, the market wage that can be earned at the present location is

$$y_p^* = \mathbf{w}_p'\boldsymbol{\beta}_p + \varepsilon_p.$$

Variables in the equation include age, sex, race, growth in employment, and growth in per capita income. If the individual migrates to a new location, then his or her market wage would be

$$y_m^* = \mathbf{w}_m'\boldsymbol{\beta}_m + \varepsilon_m.$$

Migration entails costs that are related both to the individual and to the labor market,

$$C^* = \mathbf{z}'\boldsymbol{\alpha} + u.$$

Costs of moving are related to whether the individual is self-employed and whether that person recently changed his or her industry of employment. They migrate if the benefit $y_m^* - y_p^*$ is greater than the cost, C^*. The net benefit of moving is

$$\begin{aligned} M^* &= y_m^* - y_p^* - C^* \\ &= \mathbf{w}_m'\boldsymbol{\beta}_m - \mathbf{w}_p'\boldsymbol{\beta}p - \mathbf{z}'\boldsymbol{\alpha} + (\varepsilon_m - \varepsilon_p - u) \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon. \end{aligned}$$

Because M^* is unobservable, we cannot treat this equation as an ordinary regression. The individual either moves or does not. After the fact, we observe only y_m^* if the individual has moved or y_p^* if he or she has not. But we do observe that $M = 1$ for a move and $M = 0$ for no move.

17.2.2 THE LATENT REGRESSION MODEL

Discrete dependent-variable models are often cast in the form of **index function models**. We view the outcome of a discrete choice as a reflection of an underlying regression. As an often-cited example, consider the decision to make a large purchase. The theory states that the consumer makes a marginal benefit/marginal cost calculation based on the utilities achieved by making the purchase and by not making the purchase (and by using the money for something else). We model the difference between perceived benefit and cost as an unobserved variable y^* such that

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

Note that this is the result of the *net utility* calculation in the previous section and in Example 17.2. We assume that ε has mean zero (there is a constant term in $\mathbf{x}$) and has either a logistic distribution with variance $\pi^2/3$ or a standard normal distribution with variance one, or some other specific distribution with known variance. We do not observe

[4] A number of other studies have also used variants of this basic formulation. Some important examples are Willis and Rosen (1979) and Robinson and Tomes (1982). The study by Tunali (1986) examined in Example 17.13 is another application. The now standard approach, in which participation equals one if wage offer $(\mathbf{x}_w'\boldsymbol{\beta}_w + \varepsilon_w)$ minus reservation wage $(\mathbf{x}_r'\boldsymbol{\beta}_r + \varepsilon_r)$ is positive, underlies Heckman (1979) and is also used in Fernandez and Rodriguez-Poo (1997). Brock and Durlauf (2000) describe a number of models and situations involving individual behavior that give rise to binary choice models. The Di Maria et al. (2010) study of the light bulb puzzle in Example 17.4 is another example of an elaborate structural random utility model that produces a binary outcome. This application is also closely related to Rubin's (1974, 1978) potential outcomes model discussed in Section 8.5.

the net benefit of the purchase (i.e., net utility), only whether it is made or not. Therefore, our observation is

$$\begin{aligned} y &= 1 \quad \text{if } y^* > 0, \\ y &= 0 \quad \text{if } y^* \leq 0. \end{aligned}$$

The statement in (17-3) is conveniently denoted $y = \mathbf{1}(y* > 0)$. In this formulation, $\mathbf{x}'\boldsymbol{\beta}$ is called the index function. The assumption of known variance of ε is an innocent normalization. Note, once again, the outcomes 0 and 1 are merely labels of the event. Now, suppose the variance of ε is, instead, an unrestricted parameter σ^2. The **latent regression** will be $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon*$, where now $\varepsilon*$ has variance one. But $(y^*/\sigma) = \mathbf{x}'(\boldsymbol{\beta}/\sigma) + \varepsilon$ is the same model with the same data. The observed data will be unchanged; y is still 0 or 1, *depending only on the sign* of y^*, not on its scale. This means that there is no information about σ in the sample data so σ cannot be estimated. The parameter vector $\boldsymbol{\beta}$ in this model is only "identified up to scale."[5] The assumption of zero for the threshold in (17-4) is likewise innocent if the model contains a constant term (and not if it does not).[6] Let a be a supposed nonzero threshold and α be the unknown constant term and, for the present, $\mathbf{x}$ and $\boldsymbol{\beta}$ contain the rest of the index not including the constant term. Then, the probability that y equals one is

$$\text{Prob}(y^* > a|\mathbf{x}) = \text{Prob}(\alpha + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > a|\mathbf{x}) = \text{Prob}[(\alpha - a) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0|\mathbf{x}]. \quad \textbf{(17-3)}$$

Because α is unknown, the difference $(\alpha - a)$ remains an unknown parameter. The end result is that if the model contains a constant term, it is unchanged by the choice of the threshold in (17-4). The choice of zero is a normalization with no significance. With the two normalizations, then,

$$\text{Prob}(y^* > 0|\mathbf{x}) = \text{Prob}(\varepsilon > -\mathbf{x}'\boldsymbol{\beta}|\mathbf{x}). \quad \textbf{(17-4)}$$

A remaining detail in the model is the choice of the specific distribution for ε. We will consider several. The overwhelming majority of applications are based either on the normal or the logistic distribution. If the distribution is symmetric, as are the normal and logistic, then

$$\text{Prob}(y^* > 0|\mathbf{x}) = \text{Prob}(\varepsilon < \mathbf{x}'\boldsymbol{\beta}|\mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}), \quad \textbf{(17-5)}$$

where $F(t)$ is the cdf of the random variable, ε. This provides an underlying structural model for the probability.

17.2.3 FUNCTIONAL FORM AND PROBABILITY

Consider the model of labor force participation suggested in Example 17.1. The respondent either participates in the formal labor market ($Y = 1$) or does not ($Y = 0$) in the period in which the survey is taken. We believe that a set of factors, such as age, marital status, education, and work experience, gathered in a vector $\mathbf{x}$, explain the decision, so that

$$\begin{aligned} \text{Prob}(Y = 1|\mathbf{x}) &= F(\mathbf{x}, \boldsymbol{\beta}) \\ \text{Prob}(Y = 0|\mathbf{x}) &= 1 - F(\mathbf{x}, \boldsymbol{\beta}). \end{aligned} \quad \textbf{(17-6)}$$

[5] In some treatments [e.g., Horowitz (1990) and Lewbel (2000)] it is more convenient to normalize one of the elements of $\boldsymbol{\beta}$ to equal 1 and leave σ free to vary. In the end, only $\boldsymbol{\beta}/\sigma$ is estimated, so this is inconsequential.

[6] Unless there is some compelling reason, binary choice models should not be estimated without constant terms.

The set of parameters $\boldsymbol{\beta}$ reflects the impact of changes in $\mathbf{x}$ on the probability. For example, among the factors that might interest us is the partial effect of having children in the household on the probability of labor force participation. The challenge at this point is to devise a suitable specification for the right-hand side of the equation.

Our requirement is a model that will produce predictions consistent with the underlying theory in (17-5) and (17-6). For a given regressor vector, we would expect

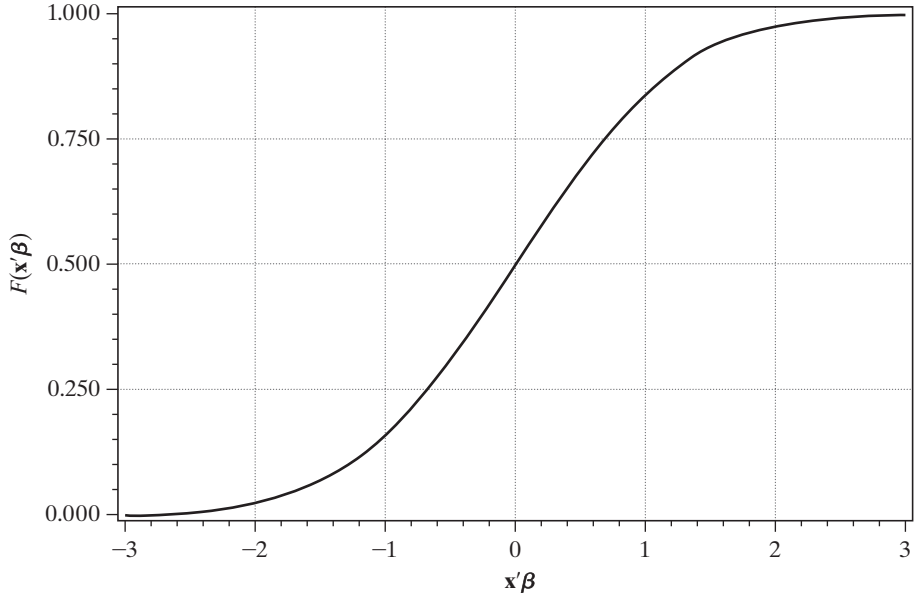
$$0 \leq \text{Prob}(Y = 1|\mathbf{x}) \leq 1, \tag{17-7}$$

$$\lim_{\mathbf{x}'\boldsymbol{\beta} \to -\infty} \text{Prob}(Y = 1|\mathbf{x}) = 0,$$
$$\lim_{\mathbf{x}'\boldsymbol{\beta} \to +\infty} \text{Prob}(Y = 1|\mathbf{x}) = 1. \tag{17-8}$$

See Figure 17.1. In principle, any proper, continuous probability distribution defined over the real line will suffice. The normal distribution has been used in many analyses, giving rise to the **probit model**,[7]

$$\text{Prob}(Y = 1|\mathbf{x}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(t)dt = \Phi(\mathbf{x}'\boldsymbol{\beta}). \tag{17-9}$$

FIGURE 17.1 Model for a Probability.



[7]The term "probit" derives from "probability unit," in turn from the use of inverse normal probability units in bioassay. See Finney (1971) and Greene and Hensher (2010, Ch. 4).

The function $\phi(t)$ is a commonly used notation for the standard normal density function and $\Phi(t)$ is the cdf. Partly because of its mathematical convenience, the logistic distribution,

$$\text{Prob}(Y = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} = \Lambda(\mathbf{x}'\boldsymbol{\beta}), \tag{17-10}$$

has also been used in many applications. We shall use the notation $\Lambda(.)$ to indicate the logistic distribution function. For this case, the density is $\Lambda(t)[1 - \Lambda(t)]$. This model is called the **logit** model for reasons we shall discuss below. Both of these distributions have the familiar bell shape of symmetric distributions and sigmoid shape shown in Figure 17.1. Other models which do not assume symmetry, such as the **Gumbel model** or Type I extreme value model,

$$\text{Prob}(Y = 1|\mathbf{x}) = \exp[-\exp(-\mathbf{x}'\boldsymbol{\beta})],$$

complementary log log model,

$$\text{Prob}(Y = 1|\mathbf{x}) = 1 - \exp[-\exp(\mathbf{x}'\boldsymbol{\beta})],$$

and the Burr model,[8]

$$\text{Prob}(Y = 1|\mathbf{x}) = \left[\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}\right]^{\gamma} = [\Lambda(\mathbf{x}'\boldsymbol{\beta})]^{\gamma},$$

have also been employed. Still other distributions have been suggested,[9] but the probit and logit models are by far the most common frameworks used in econometric applications.

The question of which distribution to use is a natural one. The logistic distribution is similar to the normal except in the tails, which are considerably heavier. (It more closely resembles a t distribution with seven degrees of freedom.) For intermediate values of $\mathbf{x}'\boldsymbol{\beta}$, the two distributions tend to give very similar probabilities. The logistic distribution tends to give larger probabilities to $Y = 1$ when $\mathbf{x}'\boldsymbol{\beta}$ is extremely small (and smaller probabilities to $Y = 1$ when $\mathbf{x}'\boldsymbol{\beta}$ is very large) than the normal distribution. It is difficult to provide practical generalities on this basis, however, as they would require knowledge of $\boldsymbol{\beta}$. We might expect different predictions from the two models, however, if the sample contains (1) very few responses (Y's equal to 1) or very few nonresponses (Y's equal to 0) and (2) very wide variation in an important independent variable, particularly if (1) is also true. There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds. Amemiya (1981) discusses a number of related issues, but as a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference. As seen in the following example, the symmetric and asymmetric distributions can give somewhat different results, and here, the guidance on how to choose is unfortunately sparse. On the other hand, for estimation of the quantities usually of interest (partial effects), in the sample sizes typical in modern

[8]Or Scobit model for a skewed logit model; see Nagler (1994).

[9]See, for example, Maddala (1983, pp. 27–32), Aldrich and Nelson (1984), and *Stata* (2014).

research, it turns out that the different functional forms tend to give comfortably similar results. The choice of which $F(.)$ to use is ultimately less important than the choice of $\mathbf{x}$ and $\mathbf{x}'\boldsymbol{\beta}$. We will examine this proposition in more detail below.

17.2.4 PARTIAL EFFECTS IN BINARY CHOICE MODELS

Most analyses will be directed at examining the relationships between the covariates, $\mathbf{x}$, and the probability of the event, $\text{Prob}(Y = 1|\mathbf{x}) = F(y|\mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta})$, typically, the partial effects. Whatever distribution is used, it is important to note that the parameters of the model ($\boldsymbol{\beta}$), like those of any nonlinear model, are not necessarily the partial effects we are accustomed to analyzing. In general, via the chain rule,

$$\frac{\partial F(y|\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{dF(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})}\right] \times \boldsymbol{\beta} = f(\mathbf{x}'\boldsymbol{\beta}) \times \boldsymbol{\beta}, \tag{17-11}$$

where $f(.)$ is the density function that corresponds to the distribution function, $F(.)$. For the normal distribution (probit model), this result is

$$\frac{\partial F(y|\mathbf{x})}{\partial \mathbf{x}} = \phi(\mathbf{x}'\boldsymbol{\beta}_{probit}) \times \boldsymbol{\beta}_{probit}.$$

For the logistic distribution,

$$\frac{d\Lambda(\mathbf{x}'\boldsymbol{\beta}_{logit})}{d(\mathbf{x}'\boldsymbol{\beta}_{logit})} = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_{logit})}{[1 + \exp(\mathbf{x}'\boldsymbol{\beta}_{logit})]^2} = \Lambda(\mathbf{x}'\boldsymbol{\beta}_{logit})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta}_{logit})],$$

so, in the logit model,

$$\frac{\partial F(y|\mathbf{x})}{\partial \mathbf{x}} = \Lambda(\mathbf{x}'\boldsymbol{\beta}_{logit})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta}_{logit})]\boldsymbol{\beta}_{logit}.$$

These values will vary with the values of $\mathbf{x}$. In index function models generally, the set of partial effects is a multiple of the coefficient vector.

As we will observe below in several applications, a common empirical regularity for estimates of probit and logit models is $\hat{\boldsymbol{\beta}}_{logit} \approx 1.6\hat{\boldsymbol{\beta}}_{probit}$. This might suggest quite a large difference between the two models, however, that would be misleading. As a general result, the partial effects produced by these two (and other) models will be nearly the same. Near the middle of the range of the probabilities, where $F(\mathbf{x}'\boldsymbol{\beta})$ is roughly 0.5, the logistic partial effects will be roughly $0.5(1 - 0.5)\boldsymbol{\beta}_{logit}$ while the probit partial effects will be roughly $0.4\boldsymbol{\beta}_{probit}$ (where 0.4 is the normal density at the point where the cdf equals 0.5). If the two partial effects are to be the same, then $0.25\boldsymbol{\beta}_{logit} = 0.4\boldsymbol{\beta}_{probit}$ or $\boldsymbol{\beta}_{logit} = 1.6\boldsymbol{\beta}_{probit}$. Observed estimates will vary around this general result. An example is shown in Table 17.1.

For computing partial effects one can evaluate the expressions at the sample means of the data, producing the partial effects at the averages (PEA),

$$PEA = \hat{\boldsymbol{\gamma}}(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}.$$

The means of the data do not always produce a realistic scenario for the computation. For example, the mean gender of 0.5 does not correspond to any individual in the sample. It is more common to evaluate the partial effects at every actual observation and use

the sample average of the individual partial effects, producing the **average partial effects (APE)**. The desired computation would be

$$APE = \hat{\bar{\gamma}} = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}. \tag{17-12}$$

It is usually, the "average partial effect," that is, the expected value of the partial effect, that is actually of interest. Let $\boldsymbol{\gamma}^0$ denote the population parameter. Then,

$$APE^0 = \boldsymbol{\gamma}^0 = E_x\left[\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}}\right]. \tag{17-13}$$

One might wonder whether the APE produces a different answer from the PEA. It is tempting to suggest that the difference is a small sample effect, but it is not, at least not entirely. Assume the parameters are known, and let the average partial effect for variable x_k be

$$\bar{\gamma}_k = APE_k = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial F(\mathbf{x}_i'\boldsymbol{\beta})}{\partial x_{ik}} = \frac{1}{n}\sum_{i=1}^{n} F'(\mathbf{x}_i'\boldsymbol{\beta})\beta_k = \frac{1}{n}\sum_{i=1}^{n}\gamma_k(\mathbf{x}_i).$$

We will compute this at the MLE, $\hat{\boldsymbol{\beta}}$. Now, expand this function in a second-order Taylor series around the point of sample means, $\bar{\mathbf{x}}$, to obtain

$$\begin{aligned}\bar{\gamma}_k &= \frac{1}{n}\sum_{i=1}^{n}\left[\gamma_k(\bar{\mathbf{x}}) + \sum_{m=1}^{k}\frac{\partial\gamma_k(\bar{\mathbf{x}})}{\partial\bar{x}_m}(x_{im}-\bar{x}_m) + \frac{1}{2}\sum_{l=1}^{K}\sum_{m=1}^{K}\frac{\partial^2\gamma_k(\bar{\mathbf{x}})}{\partial\bar{x}_l\partial\bar{x}_m}(x_{il}-\bar{x}_l)(x_{im}-\bar{x}_m) + \Delta_i(\bar{\mathbf{x}})\right] \\ &= \gamma_k(\bar{\mathbf{x}}) + \frac{1}{2}\sum_{l=1}^{K}\sum_{m=1}^{K} g_{lm}S_{lm} + \bar{\Delta}(\bar{\mathbf{x}}),\end{aligned}$$

where $\Delta(\bar{\mathbf{x}})$ is the remaining higher-order terms. The first of the four terms is the partial effect at the sample means. The second term is zero. The third is an average of functions of the variances and covariances of the data and the curvature of the probability function at the means. The final term is the remainder. Little can be said to characterize these two terms in any particular sample. In applications, the difference is usually relatively small.

Another complication for computing partial effects in a nonlinear model arises because $\mathbf{x}$ will often include dummy variables—for example, a labor force participation equation will often contain a dummy variable for marital status. It is not appropriate to apply (17-12) for the effect of a change in a dummy variable, or a change of state. The appropriate partial effect for a binary independent variable, say, d, would be

$$\text{PEA} = \text{Prob}[Y = 1|\bar{\mathbf{x}}_{(d)}, d = 1] - \text{Prob}[Y = 1|\bar{\mathbf{x}}_{(d)}, d = 0] \tag{17-14}$$

or

$$\text{APE} = \frac{1}{n}\sum_{i=1}^{n}[\text{Prob}(Y = 1|\mathbf{x}_{i,(d)}, d_i = 1) - \text{Prob}(Y = 1|\mathbf{x}_{i,(d)}, d_i = 0)],$$

where d denotes the other variables in the model excluding the dummy variable in question. Simply taking the derivative with respect to the binary variable as if it were continuous provides an approximation that is often surprisingly accurate. In Example 17.3, for the binary variable *PSI*, the average difference in the two probabilities for the probit model is 0.374, whereas the derivative approximation is $0.222 \times 1.426 = 0.317$. In a

larger sample, the differences are often very small. Nonetheless, the difference in the probabilities is the preferred computation, and is automated in standard software.

If the dummy variable in the choice model is a treatment as PSI is in the example below, then the APE would estimate the average treatment, ATE, for the population. But the average treatment on the treated, ATET, would require a change in the computation. If the treatment were exogenous (e.g., if students were carefully randomly assigned to PSI), then computing the APE over the subsample with $d_i = 1$, would be an appropriate estimator.[10] Any difference between ATE and ATET would then be attributable to systematic differences in $\mathbf{x}|d = 1$ and $\boldsymbol{x}|(d = 0 \, or \, d = 1)$. If the treatment were endogenous, then neither APE nor APE$|d = 1$ would be an appropriate estimator—indeed, the model itself would have to be extended. We will treat this case in Section 17.6.

17.2.5 ODDS RATIOS IN LOGIT MODELS

The odds *in favor* of an event is the ratio Prob($Y = 1$)/Prob($Y = 0$). For the logit model—the result is not meaningful for the other models considered—the odds "in favor of $Y = 1$" are

$$Odds = \frac{\text{Prob}(Y = 1|\mathbf{x})}{\text{Prob}(Y = 0|\mathbf{x})} = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})/[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]}{1/[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]} = \exp(\mathbf{x}'\boldsymbol{\beta}).$$

Consider the effect on the odds of the change of a dummy variable, d,

$$Odds\ Ratio = \frac{Odds(\mathbf{x}, d = 1)}{Odds(\mathbf{x}, d = 0)} = \frac{\left[\dfrac{\exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 1)/[1 + \exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 1)]}{1/[1 + \exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 1)]}\right]}{\left[\dfrac{\exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 0)/[1 + \exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 0)]}{1/[1 + \exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 0)]}\right]} = \exp(\delta).$$

Therefore, the change in the odds when a variable changes by one unit somewhat resembles a partial effect, though in fact it is not a derivative. "Odds ratios" are reported in many studies that are based on logit models. When the experiment of changing the variable in question, x_k, by one unit is meaningful, $\exp(\beta_k)$ for the respective coefficient reports the multiplicative change in the ratio. The proportional change would be $\exp(\delta) - 1$. [Received studies always report $\exp(\delta)$, not $\exp(\delta) - 1$.] If the experiment of a change in one unit is not meaningful, the odds ratio, like the simple partial effect, could be misleading. Note, in Example 17.8 (Table17.5) below, we have computed a partial effect for income of roughly −0.03. However, a change in income of a full unit in these data is not a meaningful experiment—the full range of values is about 1.0–3.0. The more useful calculation for a variable x_k is $\partial\text{Prob}(Y = 1|\mathbf{x})/\partial x_k \times dx_k$. In Example 17.8, for the income variable, $\text{d}x_k = 0.1$ would be more informative. A similar computation would be appropriate for the odds ratios, though it is unclear how that might be constructed independently of the specific change for a specific variable, in which case, the partial effect (or elasticity) might be more straightforward. The odds ratio is meaningful for a dummy variable, however. We examine an application in Example 17.11.

[10]Use of linear regression with binary dependent variables to estimate treatment effects in randomized trials is discussed in Department of Health and Human Services, Office of Adolescent Health, Evaluation Technical Assistance Brief No. 6, December 2014, www.hhs.gov/ash/oah-initiatives/assets/lpm-tabrief.pdf (accessed June 2016).

Example 17.3 Probability Models

The data listed in Appendix Table F14.1 were taken from a study by Spector and Mazzeo (1980), which examined whether a new method of teaching economics, the Personalized System of Instruction (*PSI*), significantly influenced performance in later economics courses. The "dependent variable" used in the application is *GRADE*, which indicates whether a student's grade in an intermediate macroeconomics course was higher than that in the principles course. The other variables are *GPA*, their grade point average; *TUCE*, the score on a pretest that indicates entering knowledge of the material; and *PSI*, the binary variable indicator of whether the student was exposed to the new teaching method. (Spector and Mazzeo's specific equation was somewhat different from the one estimated here.)

Table 17.1 presents five sets of parameter estimates. The coefficients and average partial effects were computed for four probability models: probit, logit, Gompertz, and complementary log log and for the linear regression of *GRADE* on the covariates. The last four sets of estimates are computed by maximizing the appropriate log-likelihood function. Inference is discussed in the next section, so standard errors are not presented here. The scale factor given in the last row is the average of the density function evaluated at the means of the variables. If one looked only at the coefficient estimates, then it would be natural to conclude that the five models had produced radically different estimates. But a comparison of the columns of average partial effects shows that this conclusion is clearly wrong. The models are very similar; in fact, the logit and probit models results are nearly identical.

The data used in this example are only moderately unbalanced between 0s and 1s for the dependent variable (21 and 11). As such, we might expect similar results for the probit and logit models.[11] One indicator is a comparison of the coefficients. In view of the different variances of the distributions, one for the normal and $\pi^2/3$ for the logistic, we might expect to obtain comparable estimates by multiplying the probit coefficients by $\pi/\sqrt{3} \approx 1.8$. Amemiya (1981) found, through trial and error, that scaling by 1.6 instead produced better results. This proportionality result is frequently cited. The result in (17-11) may help explain the finding. The index $\mathbf{x}'\boldsymbol{\beta}$ is not the random variable. The partial effect in the probit model for, say, x_k is $\phi(\mathbf{x}'\boldsymbol{\beta}_p)\beta_{pk}$, whereas that for the logit is $\Lambda(1 - \Lambda)\beta_{lk}$. (The subscripts p and l are for probit and logit.) Amemiya suggests that his approximation works best at the center of

TABLE 17.1 Estimated Probability Models

	Linear		*Logit*		*Probit*		*Comp. Log Log*		*Gompertz*	
Variable	***Coeff.***	***APE***	***Coeff.***	***APE***	***Coeff.***	***APE***	***Coeff.***	***APE***	***Coeff.***	***APE***
Constant	−1.498	–	−13.021	–	−7.452	–	−10.361	–	−7.141	–
GPA	0.464	0.464	2.826	0.363	1.626	0.361	2.293	0.413	1.584	0.319
TUCE	0.010	0.010	0.095	0.012	0.052	0.011	0.041	0.007	0.060	0.012
PSI[a]	0.379	0.379	2.379	0.358	1.426	0.374	1.562	0.312	1.616	0.411
Mean $f(\mathbf{x}'\boldsymbol{\beta})$		1.000		0.128		0.222		0.180		0.201

[a]Partial effects for PSI computed as average of [Prob(Grade $= 1|\mathbf{x}_{(PSI)}, PSI = 1)$ − Prob(Grade $= 1|\mathbf{x}_{(PSI)}, PSI = 0)$].

[11]One might be tempted in this case to suggest an asymmetric distribution for the model, such as the Gumbel distribution. However, the asymmetry in the model, to the extent that it is present at all, refers to the values of ε, not to the observed sample of values of the dependent variable.

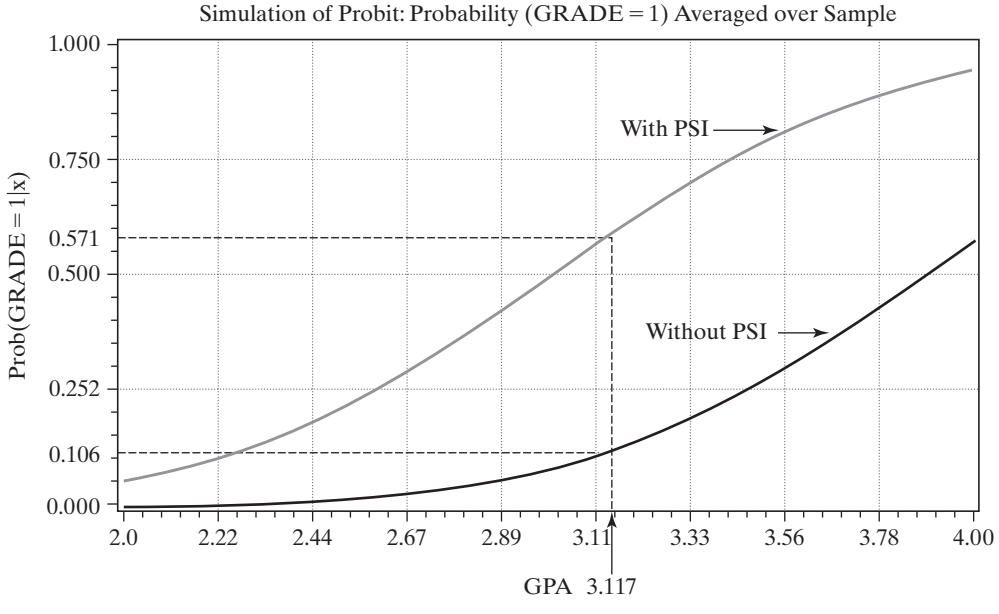
the distribution, where $F = 0.5$, or $\mathbf{x}'\boldsymbol{\beta} = 0$ for either distribution. Suppose it is. Then $\phi(0) = 0.3989$ and $\Lambda(0)[1 - \Lambda(0)] = 0.25$. If the partial effects are to be the same, then $0.3989\beta_{pk} = 0.25\beta_{lk}$, or $\beta_{lk} = 1.6\beta_{pk}$, which is the regularity observed by Amemiya. Note, though, that as we depart from the center of the distribution, the relationship will move away from 1.6. Because the logistic density descends more slowly than the normal, for unbalanced samples such as ours, the ratio of the logit coefficients to the probit coefficients will tend to be larger than 1.6. The ratios for the ones in Table 17.1 are closer to 1.7 than 1.6.

The computation of effects of dummy variables in binary choice settings is an important (one might argue, the most important) element of the analysis. One way to analyze the effect of a dummy variable on the whole distribution is to compute Prob($Y = 1$) over the range of $\mathbf{x}'\boldsymbol{\beta}$ (using the sample estimates) and with the two values of the binary variable. Using the coefficients from the probit model in Table 17.1, we have the following probabilities as a function of *GPA*, at the mean of TUCE (21.938):

$$PSI = 0\text{: Prob}(GRADE = 1) = \Phi[-7.452 + 1.626GPA + 0.052(21.938)],$$
$$PSI = 1\text{: Prob}(GRADE = 1) = \Phi[-7.452 + 1.626GPA + 0.052(21.938) + 1.426].$$

Figure 17.2 shows these two functions plotted over the range of *GPA* observed in the sample, 2.0 to 4.0. The partial effect of *PSI* is the difference between the two functions, which ranges from only about 0.06 at $GPA = 2$ to about 0.50 at *GPA* of 3.5. This effect shows that the probability that a student's grade will increase after exposure to *PSI* is far greater for students with high *GPAs* than for those with low *GPAs*. At the sample mean of *GPA* of 3.117, the effect of *PSI* on the probability is 0.465. The simple estimate of the partial effect at the mean is 0.468. But of course, this calculation does not show the wide range of differences displayed in Figure 17.2. The APE averages over the entire distribution, and equals 0.374. This latter figure is probably more representative of the desired effect. (In the typical application with a much larger sample, the differences in these results will usually be much smaller.)

FIGURE 17.2 Effect of *GPA* on Predicted Probabilities.

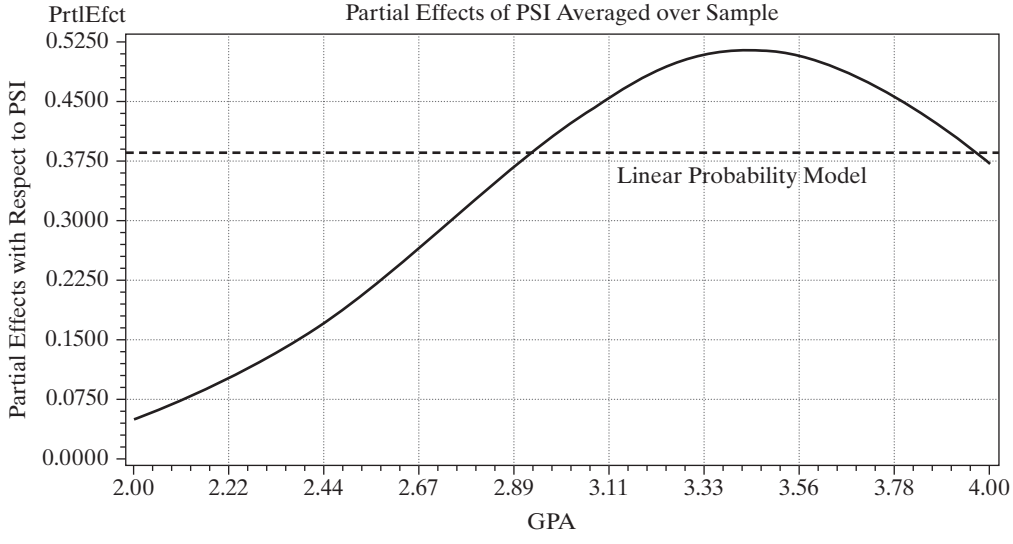


The odds ratio for the *PSI* variable is exp(2.379) = 10.6. This would imply that the odds of a grade increase for those who take the *PSI* are more than 10 times the odds for a student who does not. From Figure 17.2, for the average student, the odds ratio would appear to be about (0.571/0.429)/(0.106/0.894) = 11.1, which is essentially the same result. The partial effect of *PSI* for that student is 0.571 − 0.106 = 0.465. It is clear from Figure 17.2, however, that the partial effect of *PSI* varies greatly depending on the *GPA*. The odds ratio, being a constant, will mask that aspect of the results. The plot in Figure 17.2 is suggestive, but imprecise. A more direct analysis would examine the effect of *PSI* on the probability as it varies with *GPA*. Figure 17.3 shows that effect. The unsurprising conclusion is that the impact of *PSI* is greatest for students in the middle of the grade distribution, not at the low end, which might have been expected. We also see that the marginal benefit of *PSI* actually begins to diminish for the students with the highest *GPA*s, probably because they are most likely already to have *GRADE* = 1. [Figure 17.3 also shows the estimated effect from the linear probability, model (Section 17.2.6) which, like the odds ratio, oversimplifies the relationship.]

Example 17.4 The Light Bulb Puzzle: Examining Partial Effects

The *light bulb puzzle* refers to an observed sluggishness by consumers in adopting energy efficient and environmentally less harmful CFL (compact fluorescent light) bulbs in spite of their advantageous cost and environmental impacts. Di Maria, Ferreira, and Lazarova (2010) examined a survey of Irish energy consumers to learn about the underlying preferences that seem to be driving this puzzling outcome. The authors develop a model of utility maximization over consumption of conventional lighting and CFL lighting. Utility is derived from two sources, consumption of the lighting (in lumens) and environmental impact, *I*. Determination of the binary outcome, "adopt CFL," is based on maximizing utility from the two sources, subject to the costs of adoption, including effort. Individual heterogeneity enters the utility calculation (as a random component) through differences in environmental preferences, perceived costs, understanding of the technology, the costs of the effort in adoption, and differences in individual discount rates.

FIGURE 17.3 Effect of *PSI* on *GRADE* by *GPA*.



The empirical analysis is based on a survey of 1,500 Irish lighting consumers in the 2001 Urban Institute Ireland National Survey on Quality of Life. Inputs to the adoption model are in three components:

Environmental Interest:[12]

Support of Kyoto Protocol (1–4), Importance of Environment (1, 2, 3), Knowledge of Environment (0, 1).

Demographics:

Age, Gender, Marital Status, Family Size, Education (4 levels), Income

Housing Attributes:

Rural, Own/Rent, Detached or Semidetached Number of Rooms, House Built Before the 1960s.

The authors report coefficient estimates for probit models with standard errors and partial effects evaluated at the means of the data. Among the statistically significant results reported are partial effects of 0.098 for support of the Kyoto Protocal, 0.044 for the Importance of the Environment, and 0.115 for Knowledge of the Environment. Overall, about 30% of the sample are adopters. The environmental interest variables, therefore, are found to exert a very large influence. The mean values of these variables are 3.05, 2.51, and 0.85, respectively. Thus, starting from the base of 3.05, increased support for Kyoto increases the acceptance rate from about 0.30 to about 0.398, or roughly a third. For the *Importance* variable, the change from the average to the highest would be about 0.5, and the partial effect is 0.044, so the probability would increase by about 0.022 from a base of about 0.3, or about 7.3%, a much smaller increase. For the *Knowledge* variable, the partial effect is 0.115. Increasing this variable from 0 to 1 would increase the probability from 0.3 by about 0.115, or, again, by about one-third.

The average income in the sample is €22,987. The log of the mean is about 10. An increase in the log of income of one unit would take it to 11, or income of about €62,500, which is larger than the maximum in the sample. A more reasonable experiment might be to raise income by about 10%, in which case the log income rises by about 0.095.The partial effect for log income is 0.073. An increase in the log of income of 0.095 would be associated with an increase in the average probability of $0.095 \times 0.073 = 0.007$. This would correspond to a 2.3% increase in the probability, from 0.30 to 0.307.

The authors report an experiment with the marginal effects: "As robustness checks we first estimated the marginal effects associated with the coefficients in Table 5 at different levels of income (1st, 25th, 50th, 75th, and 99th percentile) and educational attainment. The marginal impacts discussed above increase monotonically with the level of income and education, but these increases are not statistically significant." That is, they examined the changes in the partial effect of education associated with changes in income. Superficially, this is an estimation of $\partial[\partial\text{Prob}(\text{Adopt} = 1)/\partial\text{Education}]/\partial\text{income}$. This is the analysis in Figure 17.3.

17.2.6 THE LINEAR PROBABILITY MODEL

The binary outcome suggests a regression model,

$$F(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta},$$

with

$$E[y|\mathbf{x}] = \{0 \times [1 - F(\mathbf{x}, \boldsymbol{\beta})]\} + \{1 \times [F(\mathbf{x}, \boldsymbol{\beta})]\} = F(\mathbf{x}, \boldsymbol{\beta}).$$

[12]The authors used a principal component for the three measures in one specification of the model, but the preferred specification used the three environmental variables separately.

This implies the regression model,

$$\begin{aligned} y &= E[y|\mathbf{x}] + (y - E[y|\mathbf{x}]) \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon. \end{aligned}$$

The **linear probability model (LPM)** has a number of shortcomings. A minor complication arises because ε is heteroscedastic in a way that depends on $\boldsymbol{\beta}$. Because $\mathbf{x}'\boldsymbol{\beta} + \varepsilon$ must equal 0 or 1, ε equals either $-\mathbf{x}'\boldsymbol{\beta}$ or $1 - \mathbf{x}'\boldsymbol{\beta}$, with probabilities $1 - F$ and F, respectively. Thus, you can easily show that in this model,

$$\text{Var}[\varepsilon|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}).$$

We could manage this complication with an FGLS estimator in the fashion of Chapter 9, though this only solves the estimation problem, not the theoretical one.[13] A more serious flaw is that without some ad hoc tinkering with the disturbances, we cannot be assured that the predictions from this model will truly look like probabilities. We cannot constrain $\mathbf{x}'\boldsymbol{\beta}$ to the 0–1 interval. Such a model produces both nonsense probabilities and negative variances. Five of the 32 observations in Example 17.3 predict negative probabilities. (This failure of the model to adhere to the basic assumptions of the theory is sometimes labeled "incoherence.")

In spite of the list of shortcomings, the LPM has been used in a number of recent studies. The principal motivation is that it appears to reliably reproduce the partial effects obtained from the formal models such as probit and logit—often only the signs and statistical significance are of interest. Proponents of the LPM argue that it produces a good approximation to the partial effects in the nonlinear models. The authors of the study in Example 17.5 state that they obtained similar results from a logit model (in the 2002 version, a probit model in the 2003 version). If that is always the case, and given the restrictiveness and incoherence of the linear specification, what is the LPM's advantage? Proponents point to two:

1. **Simplicity.** This is, of course, dubious because modern software requires merely the press of a different button or two for nonlinear models. The argument gains more currency in models that contain endogenous variables. We will return to this case below.
2. **Robustness.** The assumptions of normality or logisticality (?) are fragile while linearity is distribution free. This remains actually to be verified. Researchers disagree on the appropriateness of the LPM. For discussion, see Lewbel, Dong, and Yang (2012) and Angrist and Pischke (2009).

Example 17.5 Cheating in the Chicago School System—An LPM

Jacob and Levitt (2002, 2003) used a binary choice model to detect cheating by teachers on behalf of their students in the Chicago school system. The study developed a method of detecting whether test results had been altered. The model used to generate the final results

[13]There is a deeper peculiarity about this formulation. In the regression models we have examined up to this point, the disturbance, ε, is assumed to embody the independent variation of influences (other variables) that are generated outside the model. Because the disturbance in this model arises only tautologically through the need to have y on the LHS of the equation equal y on the RHS, there is no room in the linear probability model for left-out variables to explain some of the variation in y. For a given $\mathbf{x}$, ε cannot vary independently of $\mathbf{x}$. Although the least squares residuals, e_i, are algebraically orthogonal to $\mathbf{x}_i$, it is difficult to construct a statistical understanding of independence or uncorrelatedness of ε_i and $\mathbf{x}_i$.

in the study is an LPM for the variable "Indicator of classroom cheating." In one of the main results in the paper, the authors report (2002, p. 41): "[T]eachers are roughly 6 percentage points more likely to cheat for students who scored in the second quartile (between the 25th and 50th percentile) in the prior year, as compared to students scoring at the third or fourth quartiles." The coefficient on the relevant variable in the LPM is 0.057, or roughly 6%. This seems like a moderate result. However, only about 1% of the observations in their sample are actually classified as having cheated, overall. As such, if 1% is the baseline, the "6 percentage points" is actually a 600% increase! The moderate result is actually extreme. The result is not surprising, however. The linear probability model forces the probability function to have the same slope all the way from zero to one. It is clear from Figure 17.1, however, that in the extreme tails, such as $F(.) = 0.01$, the function will be much flatter than in the center of the distribution.[14] Unless the entire distribution of the data is confined to the extreme ends of the range, having to accommodate the middle of the distribution will make the LPM highly inaccurate in the tails.[15] An implication of this restriction is shown in Figure 17.3.

17.3 ESTIMATION AND INFERENCE FOR BINARY CHOICE MODELS

With the exception of the linear probability model, estimation of binary choice models is usually based on the method of maximum likelihood. Each observation is treated as a single draw from a Bernoulli distribution (binomial with one draw). The model with success probability $F(\mathbf{x}'\boldsymbol{\beta})$ and independent observations leads to the joint probability, or likelihood function,

$$\text{Prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \mathbf{X}) = \prod_{y_i=0} [1 - F(\mathbf{x}_i'\boldsymbol{\beta})] \prod_{y_i=1} F(\mathbf{x}_i'\boldsymbol{\beta}),$$

where $\mathbf{X}$ denotes $[\mathbf{x}_i]_{i=1,\ldots,n}$. The likelihood function for a sample of n observations can be conveniently written as

$$L(\boldsymbol{\beta} | \text{data}) = \prod_{i=1}^{n} [F(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i} [1 - F(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}. \tag{17-15}$$

Taking logs, we obtain

$$\ln L = \sum_{i=1}^{n} \{y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1 - y_i) \ln[1 - F(\mathbf{x}_i'\boldsymbol{\beta})]\}.^{16} \tag{17-16}$$

The likelihood equations are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[y_i \frac{f_i}{F_i} + (1 - y_i) \frac{-f_i}{(1 - F_i)} \right] \mathbf{x}_i = \mathbf{0}, \tag{17-17}$$

where f_i is the density, $dF_i/d(\mathbf{x}_i'\boldsymbol{\beta})$. [In (17-17) and later, we will use the subscript i to indicate that the function has an argument $\mathbf{x}_i'\boldsymbol{\beta}$.] The choice of a particular form for F_i leads to the empirical model.

Unless we are using the linear probability model, the likelihood equations in (17-17) will be nonlinear and require an iterative solution. All of the models we have seen thus

[14]This result appears in the 2002 (NBER) version of the paper, but not in the 2003 version.

[15]See Wooldridge (2010, pp. 562–564).

[16]If the distribution is symmetric, as the normal and logistic are, then $1 - F(\mathbf{x}'\boldsymbol{\beta}) = F(-\mathbf{x}'\boldsymbol{\beta})$. There is a further simplification. Let $q = 2y - 1$. Then $\ln L = \Sigma_i \ln F(q_i \mathbf{x}_i'\boldsymbol{\beta})$.

far are relatively straightforward to calibrate. For the logit model, by inserting (17-10) in (17-17), we get, after a bit of manipulation, the likelihood equations,

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \Lambda_i)\mathbf{x}_i = \mathbf{0}. \tag{17-18}$$

Note that if $\mathbf{x}_i$ contains a constant term, the first-order conditions imply that the average of the predicted probabilities must equal the proportion of ones in the sample.[17] This implication also bears some similarity to the least squares normal equations if we view the term $y_i - \Lambda_i$ as a residual.[18] For the probit model, the log likelihood is

$$\ln L = \sum_{y_i=0} \ln[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})] + \sum_{y_i=1} \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}). \tag{17-19}$$

The first-order conditions for maximizing $\ln L$ are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{y_i=0} \frac{-\phi_i}{1 - \Phi_i}\mathbf{x}_i + \sum_{y_i=1} \frac{\phi_i}{\Phi_i}\mathbf{x}_i = \sum_{y_i=0} \lambda_{0i}x_i + \sum_{y_i=1} \lambda_{1i}\mathbf{x}_i.$$

Using the device suggested in footnote 16, we can reduce this to

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{q_i \phi(q_i \mathbf{x}_i'\boldsymbol{\beta})}{\Phi(q_i \mathbf{x}_i'\boldsymbol{\beta})} \right] \mathbf{x}_i = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i = \mathbf{0}, \tag{17-20}$$

where $q_i = 2y_i - 1$.

The actual second derivatives for the logit model are quite simple:

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_i \Lambda_i(1 - \Lambda_i)\mathbf{x}_i\mathbf{x}_i'. \tag{17-21}$$

The second derivatives do not involve the random variable y_i, so Newton's method is also the **method of scoring** for the logit model. The Hessian is always negative definite, so the log likelihood is globally concave. Newton's method will usually converge to the maximum of the log likelihood in just a few iterations unless the data are especially badly conditioned. The computation is slightly more involved for the probit model. A useful simplification is obtained by using the variable $\lambda(y_i, \mathbf{x}_i'\boldsymbol{\beta}) = \lambda_i$ that is defined in (17-20). The second derivatives can be obtained using the result that for any z, $d\phi(z)/dz = -z\phi(z)$. Then, for the probit model,

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} - \lambda_i[\lambda_i + (q_i \mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i\mathbf{x}_i'. \tag{17-22}$$

This matrix is also negative definite for all values of $\boldsymbol{\beta}$. The proof is less obvious than for the logit model.[19] It suffices to note that the scalar part in the summation is $\text{Var}[\varepsilon | \varepsilon \leq \boldsymbol{\beta}'\mathbf{x}] - 1$ when $y = 1$ and $\text{Var}[\varepsilon | \varepsilon \geq -\boldsymbol{\beta}'\mathbf{x}] - 1$ when $y = 0$. The unconditional variance is one. Because truncation always reduces variance—see

[17]The same result holds for the linear probability model. Although regularly observed in practice, the result has not been proven for the probit model.

[18]The first derivative of the log likelihood with respect to the constant term produces the **generalized residual** in many settings. See, for example, Chesher, Lancaster, and Irish (1985) and the equivalent result for the tobit model in Section 19.3.2.

[19]See, for example, Amemiya (1985, pp. 273–274) and Maddala (1983, p. 63).

Theorem 18.2—in both cases, the variance is between zero and one, so the value is negative.[20]

The asymptotic covariance matrix for the maximum likelihood estimator can be estimated by using the negative inverse of the Hessian evaluated at the maximum likelihood estimates. There are also two other estimators available. The Berndt, Hall, Hall, and Hausman estimator [see (14-18) and Example 14.4] would be $(\mathbf{B})^{-1}$ where

$$\mathbf{B} = \sum_{i=1}^{n} g_i^2 \mathbf{x}_i \mathbf{x}_i',$$

where $g_i = (y_i - \Lambda_i)$ for the logit model [see (17-18)] and $g_i = \lambda_i$ for the probit model [see (17-20)]. The third estimator would be based on the expected value of the Hessian. As we saw earlier, the Hessian for the logit model does not involve y_i, so $\mathbf{H} = E[\mathbf{H}]$. But because λ_i is a function of y_i [see (17-20)], this result is not true for the probit model. Amemiya (1981) showed that for the probit model,

$$E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]_{\text{probit}} = \sum_{i=1}^{n} \lambda_{0i}\lambda_{1i}\mathbf{x}_i\mathbf{x}_i'. \tag{17-23}$$

Once again, the scalar part of the expression is always negative [note in (17-20) that λ_{0i} is always negative and λ_{i1} is always positive]. The estimator of the asymptotic covariance matrix for the maximum likelihood estimator is then the negative inverse of whatever matrix is used to estimate the expected Hessian. Because the actual Hessian is generally used for the iterations, this option is the usual choice. As we shall see later, though, for certain hypothesis tests, the BHHH estimator is a more convenient choice.

17.3.1 ROBUST COVARIANCE MATRIX ESTIMATION

The probit maximum likelihood estimator is often labeled a quasi-maximum likelihood estimator (QMLE) in view of the possibility that the normal probability model might be misspecified. White's (1982a) robust sandwich estimator for the asymptotic covariance matrix of the QMLE (see Section 14.11 for discussion),

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = [-\hat{\mathbf{H}}]^{-1}[\hat{\mathbf{B}}][-\hat{\mathbf{H}}]^{-1},$$

has been used in a number of studies based on the probit model.[21] (Indeed, it is ubiquitous in the contemporary literature.) If the probit model is correctly specified, then $\text{plim}(1/n)(\hat{\mathbf{B}}) = \text{plim}(1/n)(-\hat{\mathbf{H}})$ and either single matrix will suffice, so the robustness issue is moot. On the other hand, the probit (Q-) maximum likelihood estimator is *not* consistent in the presence of any form of heteroscedasticity, unmeasured heterogeneity, omitted variables (even if they are orthogonal to the included ones), nonlinearity of the functional form of the index, or an error in the distributional assumption [with some narrow exceptions as described by Ruud (1986)]. Thus, in almost any case, the sandwich estimator provides an appropriate asymptotic covariance matrix for an estimator that is biased in an unknown direction.[22] White raises this issue explicitly, although it seems to receive little attention in the literature: "It is the consistency of the QMLE for the parameters of interest in a wide range of situations which insures its usefulness as the

[20]See Johnson and Kotz (1993) and Heckman (1979). We will make repeated use of this result in Chapter 19.

[21]For example, Fernandez and Rodriguez-Poo (1997), Horowitz (1993), and Blundell, Laisney, and Lechner (1993).

[22]See Section 14.11 and Freedman (2006).

basis for robust estimation techniques" (1982a, p. 4). His very useful result is that, if the QMLE converges to a probability limit, then the sandwich estimator can be used under certain circumstances to estimate the asymptotic covariance matrix of that estimator. But there is no guarantee that the QMLE *will* converge to anything interesting or useful. Simply computing a robust covariance matrix for an otherwise inconsistent estimator does not give it redemption. Consequently, the virtue of a robust covariance matrix in this setting is unclear. It is true, however, that the robust estimator does appropriately estimate the asymptotic covariance for the parameter vector that is estimated by maximizing the log likelihood, whether that is $\boldsymbol{\beta}$ or something else. In practice, because the model is generally reasonably specified, the correction usually makes little difference.

Similar considerations apply to the cluster correction of the asymptotic covariance matrix for the MLE described in Section 14.8.2. For data with clustered structure, the estimator is

$$\mathbf{V} = \frac{C}{C-1}\left(-\sum_{c=1}^{C}\sum_{t=1}^{N_c}\frac{\partial^2 \ln f_{ct}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}\left[\sum_{c=1}^{C}\left(\sum_{t=1}^{N_c}\frac{\partial \ln f_{ct}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}}\right)\left(\sum_{t=1}^{N_c}\frac{\partial \ln f_{ct}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}'}\right)\right]\left(-\sum_{c=1}^{C}\sum_{t=1}^{N_c}\frac{\partial^2 \ln f_{ct}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}. \quad \textbf{(17-24)}$$

(The analogous form will apply for a panel data arrangement with n groups and T_i observations in group i.) The matrix provides an appropriate estimator for the asymptotic variance for the MLE. Whether the MLE, itself, estimates the parameter vector of interest when the observations are correlated (clustered) is a separate issue.

Example 17.6 Robust Covariance Matrices for Probit and LPM Estimators

In Example 7.6, we considered nonlinear least squares estimation of a loglinear model for the number of doctor visits variable shown in Figure 14.6. The data are drawn from the Riphahn et al. (2003) data set in Appendix Table F7.1. We will continue that analysis here by fitting a more detailed model for the binary variable *Doctor* = **1** (*DocVis* > 0). The index function for the model is

$$\text{Prob}(\textit{Doctor} = 1|\mathbf{x}_{it}] = F(\beta_1 + \beta_2\,\textit{Age}_{it} + \beta_3\,\textit{Educ}_{it} + \beta_4\,\textit{Income}_{it} + \beta_5\,\textit{Kids}_{it} + \beta_6\,\textit{Health Satisfaction}_{it} + \beta_7\,\textit{Marital Status}_{it}).$$

The data are an unbalanced panel of 27,326 household-years in 7,293 groups. We will examine the 3,377 observations in the 1994 wave, then the full data set. Descriptive statistics for the variables in the model are given in Table 17.2. (We will use these data in

TABLE 17.2 Descriptive Statistics for Binary Choice Model

	Full Panel: n = 27,326				*1994 Wave: n = 3,377*	
Variable	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*	*Mean*	*Standard Deviation*
Doctor	0.629	0.483	0	1	0.658	0.474
Age	43.526	11.330	25	64	42.627	11.586
Education	11.321	2.325	7	18	11.506	2.403
Income	0.352	0.177	0.0015	3.0671	0.445	0.217
Kids	0.403	0.490	0	1	0.388	0.487
Health Sat.	6.786	2.294	0	10	6.643	2.215
Married	0.759	0.428	0	1	0.710	0.454

several examples to follow.) Table 17.3 presents two sets of estimates for each of the probit model and the linear probability model. The 1994 wave of the panel is used for the top panel of results. The comparison is between the conventional standard errors and the robust standard errors. These would be the White estimator for the LPM and the robust estimator in (14-36) for the MLE. In both cases, there is essentially no difference in the estimated standard errors. This would be the typical result. The lower panel shows the impact of correcting the standard errors of the pooled estimator in a panel. The robust standard errors are based on (17-24). In this case, there is a tangible difference, though perhaps less than one might expect. The correction for clustering produces a 20% to 50% increase in the standard errors.

17.3.2 HYPOTHESIS TESTS

The full menu of procedures is available for testing hypotheses about the coefficients. The simplest method for a single restriction would be the usual t tests, using the standard errors from the estimated asymptotic covariance matrix for the MLE. Based on the asymptotic normal distribution of the estimator, we would use the standard normal table rather than the t table for critical points. (See the several previous examples.) For more involved restrictions, it is possible to use the Wald test. For a set of restrictions $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the statistic is

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'\{\mathbf{R}(\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}])\mathbf{R}'\}^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}).$$

TABLE 17.3 Estimates for Binary Choice Models

	Cross Section Estimates, 1994 Wave					
	Probit Model			*Linear Probability Model*		
Variable	*Coefficient*	*Standard Error*	*Robust Std. Error*	*Coefficient*	*Std. Error*	*Robust Std. Error*
Constant	1.69384	0.18199	0.18063	1.05062	0.05986	0.05840
Age	0.00448	0.00240	0.00238	0.00147	0.00080	0.00079
Education	−0.01205	0.01002	0.01002	−0.00448	0.00343	0.00351
Income	−0.09149	0.11187	0.11473	−0.02671	0.03842	0.04016
Kids	−0.24557	0.05514	0.05541	−0.08398	0.01874	0.01907
Health Sat.	−0.18503	0.01201	0.01187	−0.05800	0.00363	0.00319
Married	0.10571	0.06134	0.06131	0.03666	0.02055	0.02040
	Full Panel Data Pooled Estimates					
Variable	*Coefficient*	*Std. Error*	*Clustered Std. Error*	*Coefficient*	*Std. Error*	*Clustered Std. Error*
Constant	1.46973	0.06538	0.08687	0.99472	0.02246	0.02988
Age	0.00617	0.00082	0.00107	0.00213	0.00029	0.00037
Education	−0.01527	0.00360	0.00499	−0.00587	0.00127	0.00180
Income	−0.02838	0.04746	0.05727	−0.00285	0.01667	0.02031
Kids	−0.12993	0.01868	0.02354	−0.04508	0.00656	0.00837
Health Sat.	−0.17466	0.00396	0.00490	−0.05757	0.00126	0.00141
Married	0.06591	0.02103	0.02762	0.02363	0.00730	0.00958

For example, for testing the hypothesis that a subset of the coefficients, say, the last M, are zero, the Wald statistic uses $\mathbf{R} = [\mathbf{0}|\mathbf{I}_M]$ and $\mathbf{q} = \mathbf{0}$. Collecting terms, we find that the test statistic for this hypothesis is

$$W = \hat{\boldsymbol{\beta}}_M' \mathbf{V}_M^{-1} \hat{\boldsymbol{\beta}}_M, \tag{17-25}$$

where the subscript M indicates the subvector or submatrix corresponding to the M variables and $\mathbf{V}$ is the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

Likelihood ratio and Lagrange multiplier statistics can also be computed. The likelihood ratio statistic is

$$\text{LR} = -2[\ln \hat{L}_R - \ln \hat{L}_U],$$

where $\hat{L}_R$ and $\hat{L}_U$ are the likelihood functions evaluated at the restricted and unrestricted estimates, respectively.

A common test, which is similar to the F test that all the slopes in a regression are zero, is the likelihood ratio test that all the slope coefficients in the probit or logit model are zero. For this test, the constant term remains unrestricted. In this case, the restricted log likelihood is the same for both probit and logit models,

$$\ln L_0 = n[P \ln P + (1 - P) \ln(1 - P)], \tag{17-26}$$

where P is the proportion of the observations that have dependent variable equal to 1. These tests of models ML1 and ML2 are shown in Table 17.9 in Example 17.14.

It might be tempting to use the likelihood ratio test to choose between the probit and logit models. But there is no restriction involved and the test is not valid for this purpose. To underscore the point, there is nothing in its construction to prevent the chi-squared statistic for this "test" from being negative. Note, again, in Example 17.14, the log likelihood for the logit model is −1,991.13 while for the probit model (not shown) it is −1,990.36. This might suggest a preference for the probit model, but one could not carry out a test based on these results.

The **Lagrange multiplier test** statistic is $\text{LM} = \mathbf{g}'\mathbf{V}\mathbf{g}$, where $\mathbf{g}$ is the first derivatives of the *unrestricted* model evaluated at the *restricted* parameter vector and $\mathbf{V}$ is any of the estimators of the asymptotic covariance matrix of the maximum likelihood estimator, once again computed using the restricted estimates. Davidson and MacKinnon (1984) find evidence that $E[\mathbf{H}]$ is the best of the three estimators, which gives

$$\text{LM} = \left(\sum_{i=1}^{n} g_i \mathbf{x}_i\right)' \left[\sum_{i=1}^{n} E[-h_i]\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \left(\sum_{i=1}^{n} g_i \mathbf{x}_i\right), \tag{17-27}$$

where $E[-h_i]$ is defined in (17-21) for the logit model and in (17-23) for the probit model. One could use the robust estimator in Section 13.3.1 instead.

For the logit model, when the hypothesis is that all the slopes are zero, the LM statistic is

$$\text{LM} = nR^2,$$

where R^2 is the uncentered coefficient of determination in the regression of $(y_i - \bar{y})$ on $\mathbf{x}_i$ and $\bar{y}$ is the proportion of 1s in the sample. An alternative formulation based on the BHHH estimator, which we developed in Section 14.4.6 is also convenient. For any

of the models considered (probit, logit, Gumbel, etc.), the first derivative vector can be written as

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} g_i \mathbf{x}_i = \mathbf{X}'\mathbf{G}\mathbf{i},$$

where $\mathbf{G}(n \times n) = \text{diag}[g_1, g_2, \ldots, g_n]$ and $\mathbf{i}$ is an $n \times 1$ column of 1s. The BHHH estimator of the Hessian is $(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})$, so the LM statistic based on this estimator is

$$\text{LM} = n\left[\frac{1}{n}\mathbf{i}'(\mathbf{G}\mathbf{X})(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{G}')\mathbf{i}\right] = nR_{\mathbf{i}}^2, \tag{17-28}$$

where $R_{\mathbf{i}}^2$ is the uncentered coefficient of determination in a regression of a column of ones on the first derivatives of the logs of the individual probabilities.

All the statistics listed here are asymptotically equivalent and under the null hypothesis of the restricted model have limiting chi-squared distributions with degrees of freedom equal to the number of restrictions being tested.

Example 17.7 Testing for Structural Break in a Logit Model

The probit model in Example 17.6, based on Riphahn, Wambach, and Million (2003), is

$$\text{Prob}(DocVis_{it} > 0) = \Phi(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Education_{it} + \beta_4\, Income + \beta_5\, Kids_{it} + \beta_6\, HealthSat_{it} + \beta_7\, Married_{it}).$$

In the original study, the authors split the sample on the basis of gender and fit separate models for male- and female-headed households. We will use the preceding results to test for the appropriateness of the sample splitting. This test of the pooling hypothesis is a counterpart to the **Chow test** of structural change in the linear model developed in Section 6.6.2. Because we are not using least squares (in a linear model), we use the likelihood-based procedures rather than an F test as we did earlier. Estimates of the three models (based on the 1994 wave of the datra) are shown in Table 17.4. The chi-squared statistic for the likelihood ratio test is

$$\text{LR} = -2(-1{,}990.534 - (-1{,}117.587 - 840.246)) = 65.402.$$

The 95% critical value for seven degrees of freedom is 14.067. To carry out the Wald test for this hypothesis there are two numerically identical ways to proceed. First, using the estimates

TABLE 17.4 Estimated Models for Pooling Hypothesis

	Pooled Sample		*Male*		*Female*	
Variable	***Estimate***	***Std. Error***	***Estimate***	***Std. Error***	***Estimate***	***Std. Error***
Constant	1.69384	0.18199	1.51850	0.23388	1.80570	0.30341
Age	0.00448	0.00240	0.00509	0.00331	0.00031	0.00374
Education	−0.01205	0.01002	−0.01351	0.01309	0.00842	0.01645
Income	−0.09149	0.11187	0.09350	0.15627	−0.30374	0.16447
Kids	−0.24557	0.05514	−0.28068	0.07676	−0.26567	0.08357
Health Sat.	−0.18503	0.01201	−0.19514	0.01635	−0.16289	0.01797
Married	0.10571	0.06134	0.13027	0.08862	0.08212	0.08862
ln L	−1,990.534		−1,117.587		−840.246	
Sample Size	3,377		1,812		1,565	

for *Male* and *Female* samples separately, we can compute a chi-squared statistic to test the hypothesis that the difference of the two coefficients is zero. This would be

$$W = [\hat{\boldsymbol{\beta}}_{Male} - \hat{\boldsymbol{\beta}}_{Female}]'[\text{Est.Asy.Var}(\hat{\boldsymbol{\beta}}_{Male}) + \text{Est.Asy.Var}(\hat{\boldsymbol{\beta}}_{Female})]^{-1}[\hat{\boldsymbol{\beta}}_{Male} - \hat{\boldsymbol{\beta}}_{Female}] = 64.6942.$$

Another way to obtain the same result is to add to the pooled model the original seven variables now multiplied by the *Female* dummy variable. We use the augmented $\mathbf{X}$ matrix $\mathbf{X}^* = [\mathbf{X}, \textit{female} \times \mathbf{X}]$. The model with 14 variables is now estimated, and a test of the pooling hypothesis is done by testing the joint hypothesis that the coefficients on these seven additional variables are zero. The Lagrange multiplier test is carried out by using this augmented model as well. To apply (17-28), the necessary derivatives are in (17-18). For the probit model, the derivative matrix is simply $\mathbf{G}^* = \text{diag}[\lambda_i]$ from (17-20). For the LM test, the vector $\boldsymbol{\beta}$ that is used is the one for the restricted model. Thus, $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}'_{Pooled}, 0, 0, 0, 0, 0, 0, 0)'$. The estimated values that appear in $\mathbf{G}^*$ are simply those obtained from the pooled model. Then,

$$\text{LM} = \mathbf{i}'\mathbf{G}^*\mathbf{X}^*[(\mathbf{X}^{*\prime}\mathbf{G}^{*\prime})(\mathbf{G}^*\mathbf{X}^*)]^{-1}\mathbf{X}^{*\prime}\mathbf{G}^{*\prime}\mathbf{i} = 65.9686.$$

The pooling hypothesis is rejected by all three procedures.

17.3.3 INFERENCE FOR PARTIAL EFFECTS

The predicted probabilities, $F(\mathbf{x}'\hat{\boldsymbol{\beta}}) = \hat{F}$, and the estimated partial effects, $f(\mathbf{x}'\hat{\boldsymbol{\beta}}) \times \hat{\boldsymbol{\beta}} = \hat{f}\,\hat{\boldsymbol{\beta}}$, are nonlinear functions of the parameter estimates. We have three methods of computing asymptotic standard errors for these: the delta method, the method of Krinsky and Robb, and bootstrapping. All three methods can be found in applications in the received literature. Discussion of the various methods and some related issues appears in Dowd, Greene, and Norton (2014).

17.3.3.a The Delta Method

To compute standard errors, we can use the linear approximation approach discussed in Section 4.6. For the predicted probabilities,

$$\text{Est.Asy.Var}[\hat{F}] = [\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}]'\mathbf{V}[\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}],$$

where

$$\mathbf{V} = \text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}].$$

The estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ can be any of those described earlier. Let $z = \mathbf{x}'\hat{\boldsymbol{\beta}}$. Then the derivative vector is

$$[\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}] = [d\hat{F}/dz][\partial z/\partial\hat{\boldsymbol{\beta}}] = \hat{f}\mathbf{x}.$$

Combining terms gives

$$\text{Est.Asy.Var}[\hat{F}] = \hat{f}^2\mathbf{x}'\mathbf{V}\mathbf{x},$$

which depends on the particular $\mathbf{x}$ vector used. This result is also useful when a partial effect is computed for a dummy variable. In that case, the estimated effect is

$$\Delta\hat{F} = [\hat{F}|(d = 1)] - [\hat{F}|(d = 0)].$$

The estimator of the asymptotic variance would be

$$\text{Est.Asy.Var}[\Delta\hat{F}] = [\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}]'\mathbf{V}[\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}], \qquad \textbf{(17-29)}$$

where

$$[\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}] = \hat{f}_1 \times \begin{pmatrix}\overline{\mathbf{x}}_{(d)}\\ 1\end{pmatrix} - \hat{f}_0 \times \begin{pmatrix}\overline{\mathbf{x}}_{(d)}\\ 0\end{pmatrix}.$$

For the other partial effects, let $\hat{\boldsymbol{\gamma}}(\mathbf{x}) = \hat{f}(\mathbf{x}'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}$. Then

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}(\mathbf{x})] = \left[\frac{\partial \hat{\boldsymbol{\gamma}}(\mathbf{x})}{\partial \hat{\boldsymbol{\beta}}'}\right]\mathbf{V}\left[\frac{\partial \hat{\boldsymbol{\gamma}}(\mathbf{x})}{\partial \hat{\boldsymbol{\beta}}'}\right]'.$$

The matrix of derivatives (the Jacobian) is

$$\hat{f}(\mathbf{x}'\hat{\boldsymbol{\beta}})\left(\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}'}\right) + \hat{\boldsymbol{\beta}}\left(\frac{d\hat{f}(\mathbf{x})}{dz}\right)\left(\frac{\partial z}{\partial \hat{\boldsymbol{\beta}}'}\right) = \hat{f}(\mathbf{x})\mathbf{I} + \left(\frac{d\hat{f}(\mathbf{x})}{dz}\right)\hat{\boldsymbol{\beta}}\mathbf{x}'.$$

For the probit model, $df(z)/dz = -z\phi(z)$, so

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}(\mathbf{x})] = \{\phi(\mathbf{x}'\hat{\boldsymbol{\beta}})\}^2 \times [\mathbf{I} - (\mathbf{x}'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}\mathbf{x}']\mathbf{V}[\mathbf{I} - (\mathbf{x}'\hat{\boldsymbol{\beta}})\mathbf{x}\hat{\boldsymbol{\beta}}'].$$

For the logit model, $\hat{f}(\mathbf{x}'\hat{\beta}) = \hat{\Lambda}(\mathbf{x})[1 - \hat{\Lambda}(\mathbf{x})]$, so

$$\frac{d\hat{f}(\mathbf{x}'\hat{\boldsymbol{\beta}})}{dz} = [1 - 2\hat{\Lambda}(\mathbf{x})]\left(\frac{d\hat{\Lambda}(\mathbf{x})}{dz}\right) = [1 - 2\hat{\Lambda}(\mathbf{x})]\hat{\Lambda}(\mathbf{x})[1 - \hat{\Lambda}(\mathbf{x})].$$

Collecting terms, we obtain

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}(\mathbf{x})] = \{\hat{\Lambda}(\mathbf{x})[1 - \hat{\Lambda}(\mathbf{x})]\}^2[\mathbf{I} + [1 - 2\hat{\Lambda}(\mathbf{x})]\hat{\boldsymbol{\beta}}\mathbf{x}']\hat{\mathbf{V}}[\mathbf{I} + [1 - 2\Lambda(\mathbf{x})]\mathbf{x}\hat{\boldsymbol{\beta}}'].$$

As before, the value obtained will depend on the $\mathbf{x}$ vector used. A common application sets $\mathbf{x}$ at $\bar{\mathbf{x}}$, the means of the data.

The average partial effects would be computed as

$$\bar{\hat{\boldsymbol{\gamma}}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}{\partial \mathbf{x}_i} = \left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\right]\hat{\boldsymbol{\beta}}.$$

The preceding estimator appears to be the mean of a random sample. It would be if it were based on the true $\boldsymbol{\beta}$. But the n terms based on the same $\hat{\boldsymbol{\beta}}$ are correlated. The delta method must account for the asymptotic (co)variation of the terms in the sum of functions of $\hat{\boldsymbol{\beta}}$. To use the delta method to estimate the asymptotic standard errors for the average partial effects, $\widehat{APE}_k$, we would use

$$\begin{aligned}
\text{Est.Asy.Var}[\bar{\hat{\boldsymbol{\gamma}}}] &= \frac{1}{n^2}\,\text{Est.Asy.Var}\left[\sum_{i=1}^{n}\hat{\boldsymbol{\gamma}}_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Est.Asy.Cov}[\hat{\boldsymbol{\gamma}}_i, \hat{\boldsymbol{\gamma}}_j] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{G}_i(\hat{\boldsymbol{\beta}})\hat{\mathbf{V}}\mathbf{G}_j'(\hat{\boldsymbol{\beta}}) \\
&= \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{G}_i(\hat{\boldsymbol{\beta}})\right]\hat{\mathbf{V}}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{G}_j'(\hat{\boldsymbol{\beta}})\right],
\end{aligned}$$

where

$$\mathbf{G}_i(\hat{\boldsymbol{\beta}}) = \frac{\partial f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}'} = f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{I} + f'(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}\mathbf{x}_i'.$$

The estimator of the asymptotic covariance matrix for the APE is simply

$$\text{Est.Asy.Var}[\bar{\hat{\boldsymbol{\gamma}}}] = \overline{\mathbf{G}(\hat{\boldsymbol{\beta}})}\,\hat{\mathbf{V}}\,\overline{\mathbf{G}'(\hat{\boldsymbol{\beta}})}.$$

The appropriate covariance matrix is computed by making the same adjustment as in the partial effects—the derivative matrices are averaged over the observations rather than being computed at the means of the data.

17.3.3.b An Adjustment to the Delta Method

The delta method treats the data as *fixed in repeated samples*. If, instead, the APE were treated as a parameter to be estimated—that is, a feature of the population from which $(y_i, \mathbf{x}_i)$ are randomly drawn—then the asymptotic variance would account for the variation in $\mathbf{x}_i$ as well.[23] In the application, then, there are two sources of variation: the first is the sampling variation of the parameter estimator of $\boldsymbol{\beta}$ and the second is the sampling variability due to the variation in $\mathbf{x}$.[24] An appropriate asymptotic variance for the APE would be the sum of the two terms.[25]

Assume for the moment that $\boldsymbol{\beta}$ is known. Then, the APE is

$$\bar{\boldsymbol{\gamma}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial F(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \mathbf{x}_i} = \left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i'\boldsymbol{\beta})\right]\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}_i.$$

Based on the sample of observations on the partial effects, the natural estimator of the variance of each of the K estimated partial effects would be

$$\hat{\sigma}^2_{\gamma,k} = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(\gamma_k(\mathbf{x}_i) - \bar{\gamma}_k)\right]^2 = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(PE_{i,k} - APE_k)\right]^2.^{26}$$

The asymptotic variance of the partial effects estimator is intended to reflect the variation of the parameter estimator, $\hat{\boldsymbol{\beta}}$, whereas the preceding estimator generates the variation from the heterogeneity of the sample data while holding the parameter fixed at $\hat{\boldsymbol{\beta}}$. For example, for a logit model, $\hat{\gamma}_k(\mathbf{x}_i) = \hat{\beta}_k\Lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}})[1 - \Lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}})] = \hat{\boldsymbol{\beta}}_k\hat{\delta}_i$, and $\hat{\delta}_i$ is the same for all k. It follows that

$$\hat{\sigma}^2_{\gamma,k} = \hat{\beta}_k^2\left[\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\delta}_i - \bar{\hat{\delta}})^2\right] = \hat{\beta}_k^2 s^2_{\hat{\delta}}.$$

The delta method would use, instead, the kth diagonal element of

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}(\mathbf{x})] = \{\hat{\Lambda}(\mathbf{x})[1 - \hat{\Lambda}(\mathbf{x})]\}^2[\mathbf{I} + [1 - 2\hat{\Lambda}(\mathbf{x})]\hat{\boldsymbol{\beta}}\mathbf{x}']\hat{\mathbf{V}}[\mathbf{I} + [1 - 2\Lambda(\mathbf{x})]\mathbf{x}\hat{\boldsymbol{\beta}}'].$$

To account for the variation of the data as well, the variance estimator would be the sum of these two terms.

The impact of the adjustment is data dependent. In our experience, it is usually minor. (It is trivial in the example below.) We do note that the APEs are sometimes computed for specific configurations of $\mathbf{x}$, or specific values, or specific subsets of observations. In these cases, the appropriate adjustment, if any, is unclear.

---

[23]For example, see equation (17-13).

[24]The two sources of variation are the disturbances (the random part of the random utility model) and the variation of the observed sample of $\mathbf{x}_i$. This does raise a question as to the meaning of the standard errors, robust or otherwise, computed for the linear probability model.

[25]See Wooldridge (2010, p. 467 and 2011, pp. 184–186) for formal development of this result.

[26]See, for example, Contoyannis et al. (2004, p. 498), who reported computing the "sample standard deviation of the partial effects."

17.3.3.c The Method of Krinsky and Robb

The method of Krinsky and Robb was described in Section 15.3. For present purposes, we will apply the method as follows. The MLEs of the model parameters are $\hat{\boldsymbol{\beta}}$ and $\mathbf{V}$. We will draw a random sample of R draws from the multivariate normal population with this mean and variance. This is done by first computing the Cholesky decomposition of $\mathbf{V} = \mathbf{CC}'$ where $\mathbf{C}$ is a lower triangular matrix. With this in hand, we draw R standard multivariate normal vectors $\mathbf{w}_r$, then $\hat{\boldsymbol{\beta}}(r) = \hat{\boldsymbol{\beta}} + \mathbf{Cw}_r$. With each $\hat{\boldsymbol{\beta}}(r)$, we compute the partial effects, either APE or PEA, $\hat{\boldsymbol{\gamma}}(r)$. The estimator of the asymptotic variance is the empirical variance of this sample of R observations,

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}] = \frac{1}{R}\sum_{r=1}^{R} (\hat{\boldsymbol{\gamma}}(r) - \bar{\hat{\boldsymbol{\gamma}}})^2.$$

Note that Krinsky and Robb will accommodate the sampling variability of $\hat{\boldsymbol{\beta}}$ but not the sample variation in $\mathbf{x}_i$ considered in the preceding adjustment to the delta method.

17.3.3.d Bootstrapping

Bootstrapping is described in Section 15.4. It is essentially the same as Krinsky and Robb save that the sample of draws of $\hat{\boldsymbol{\beta}}$ (r) is obtained by repeatedly sampling n observations from the data with replacement and reestimating the model with each. In principle, bootstrapping will automatically account for the extra variation due to the data discussed in Section 17.3.2b.

Example 17.8 STANDARD ERRORS FOR PARTIAL EFFECTS

Table 17.5 shows estimates of a simple probit model,

$$\text{Prob}(DocVis_{it} > 0) = \Phi(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Education_{it} + \beta_4\, Income_{it} + \beta_5\, Kids_{it} + \beta_6\, HealthSat_{it} + \beta_7\, Married_{it}).$$

We report the average partial effects and the partial effects at the means. These results are based on the 1994 wave of the panel in Example 17.7. The sample size is 3,377. As noted earlier, the APEs and PEAs differ slightly, but not enough that one would draw a different conclusion about the population from one versus the other. In computing the standard errors for the APEs, we used the delta method without the adjustment in Section 17.3.2b. When that adjustment is made, the results are almost identical. The only change is the standard error for the coefficient on health satisfaction which changes from 0.00361 to 0.00362.

TABLE 17.5 Comparison of Estimators of Partial Effects

	Probit Model		*Average Partial Effects*		*Partial Effects at Means*	
Variable	*Coefficient*	*Std. Error*	*Avg. Partial Effect*	*Std. Error*	*Partial Effect at Means*	*Std. Error*
Constant	1.69384	0.18199				
Age	0.00448	0.00240	0.00150	0.00080	0.00161	0.00086
Education	−0.01205	0.01002	−0.00404	0.00336	−0.00433	0.00360
Income	−0.09149	0.11187	−0.03067	0.03749	−0.03290	0.04022
Kids	−0.24557	0.05514	−0.08358	0.01890	−0.08830	0.01982
Health Sat.	−0.18503	0.01201	−0.06202	0.00362	−0.06653	0.00426
Married	0.10571	0.06134	0.02086	0.02086	0.04801	0.02206

TABLE 17.6 Comparison of Methods for Computing Standard Errors for Average Partial Effects

Variable	*Avg. Partial Effect*	*Std. Error Delta Method*	*Std. Error Krinsky and Robb**	*Std. Error Bootstrap**
Age	0.00150	0.00080	0.00081	0.00080
Education	−0.00404	0.00336	0.00336	0.00372
Income	−0.03067	0.03749	0.03680	0.04065
Kids	−0.08358	0.01890	0.01839	0.02032
Health Sat.	−0.06202	0.00361	0.00384	0.00372
Married	0.02086	0.02086	0.01971	0.02248

*100 Replications.

Table 17.6 compares the three methods of computing standard errors for average partial effects. These results, in a moderate sized data set, in a typical application, are consistent with the theoretical proposition that any of the three methods should be useable. The choice could be based on convenience.

Example 17.9 Hypothesis Tests About Partial Effects

Table 17.7 presents the maximum likelihood estimates for the probit model,

$$\text{Prob}(DocVis_{it} > 0) = \Phi(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Education_{it} + \beta_4\, Income_{it} + \beta_5\, Kids_{it} + \beta_6\, Health + \beta_7\, Married_{it}).$$

(The column labeled "Interaction Model" is the estimates of the model in Example 17.14.) The *t* ratios listed are used for testing the hypothesis that the coefficient or partial effect is zero. The similarity of the *t* statistics for the coefficients and the partial effects is typical. The interpretation differs, however. Consider the test of the hypothesis that the coefficient on *Kids* is zero. The value of −4.45 leads to rejection of the null bypothesis. The same hypothesis about the average partial effect produces the same conclusion. The question is, what should be the conclusion if these tests conflict? If the *t* ratio on the APE for *Kids* were 0.45, then the tests would conflict. And, because

$$\text{APE}\,(Kids) = \beta_{kids} \times E[density\,|\,\mathbf{x}],$$

TABLE 17.7 Estimates for Binary Choice Models

	Cross Section Estimation, 1994 Wave						
	Probit Model				*Average Partial Effects*		
Variable	*Coefficient*	*Std. Error*	*t Ratio*	*(Interaction Model)*	*Estimate*	*Std. Error*	*t Ratio*
Constant	1.69384	0.18199	9.31	1.98542	–	–	–
Age	0.00448	0.00240	1.86	−0.00177	0.00150	0.00080	−1.86
Education	−0.01205	0.01002	−1.20	−0.03466	−0.00404	0.00336	−1.20
Income	−0.09149	0.11187	−0.82	−0.09903	−0.03067	0.03749	−0.82
Kids	−0.24557	0.05514	−4.45	−0.24976	−0.08358	0.01890	−4.42
Health Sat.	−0.18503	0.01201	−15.40	−0.18527	−0.06202	0.00362	−17.15
Married	0.10571	0.06134	1.72	−0.10598	0.03571	0.02086	1.71
Age × Educ.				0.00055			

the conflict would be fundamental. We have already rejected the hypothesis that β_{kids} equals zero, so the only way that the APE can equal zero is if the second term is zero. But the second term is positive by construction—the density must be positive. Worse, if the expected density were zero, then all the other APEs would be zero as well. The natural way out of the dilemma is to base tests about relevance of variables on the structural model, not on the partial effects. The implication runs in the direction from the structure to the partial effects, not the reverse. That leaves a question. Is there a use for the standard errors for the partial effects? Perhaps not for hypothesis tests, but for developing confidence intervals as in the next example.

Example 17.10 Confidence Intervals for Partial Effects

Continuing the development of Section 17.3.3, the usual approach could be taken for forming a confidence interval for the APE. For example, based on the results in Table 17.7, we would estimate the APE for Kids to be $-0.08358 \pm 1.96\,(0.0189) = [-0.12062 - 0.0465]$. As we noted in Example 17.3, the single estimate of the APE might not capture the interesting variation in the partial effect as other variables change. Figure 17.4 below reproduces the APE for PSI as it varies with GPA in the example of the performance in economics courses. We have added to Figure 17.3 confidence intervals for the APE of PSI for a set of values of GPA ranging from 2 to 4 to show a confidence region.

Example 17.11 Inference about Odds Ratios

The results in Table 17.8 are obtained for a logit model for *GRADE* in Example 17.3. (The coefficient estimates appear in Table 17.1.)

We are interested in the odds ratios for this model, which as we saw in Section 17.2.5, would be computed as $\exp(\hat{\beta}_k)$ for each estimate. Williams (2015) reports the following post-estimation results for this model using Version 11 (and later) of *Stata*. (Some detail has been omitted.)

FIGURE 17.4 Confidence Region for Average Partial Effect.

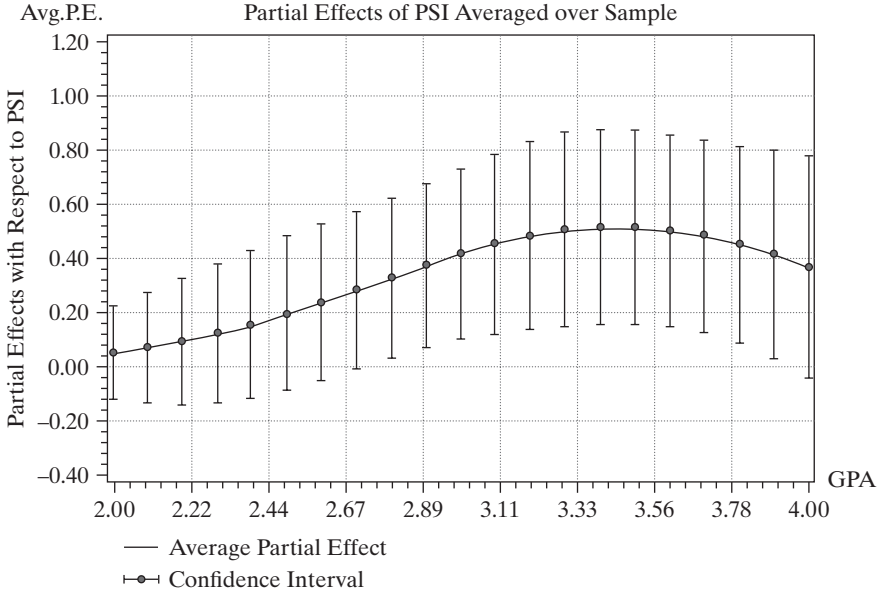


TABLE 17.8 Estimated Logit Model

Variable	*Coefficient*	*Std. Error*	*t Ratio*	*P Value*	*95% Confidence Lower*	*Interval Upper*
Constant	−13.0213	4.93132	−2.64	0.0083	−22.6866	−3.3561
GPA	2.82611	1.26294	2.24	0.0252	0.35079	5.3014
TUCE	0.09516	0.14155	0.67	0.5014	−0.18228	0.37260
PSI	2.37869	1.06456	2.23	0.0255	0.29218	4.46520

```
------------------------------------------------------------------------------
       grade |  Odds Ratio   Std. Err.      z    P>|z|     [95% conf. Interval]
-------------+----------------------------------------------------------------
         gpa |    16.87972   21.31809     2.24   0.035    1.420194    200.6239
        tuce |    1.098832   .1556859     0.67   0.501    .8333651    1.451502
         psi |    10.79073   11.48743     2.23   0.025    1.339344    86.93802
------------------------------------------------------------------------------
```

This result from a widely used software package provides context to consider what is reported and how to interpret it. The estimated odds ratios appear in the first column. To obtain the standard errors, we would use the delta method. The Jacobian for each coefficient is $d[\exp(\hat{\beta}_k)]/d\,\hat{\beta}_k = \exp(\hat{\beta}_k)$, so the standard error would just be the odds ratio times the original estimated standard error. Thus, $21.31809 = 16.87972 \times 1.26294$. But the z is not the ratio of the odds ratio to the estimated standard error. It is the z ratio for the original coefficient. On the other hand, it would make no sense to test the hypothesis that the odds ratio equals zero, because it must be positive. Perhaps the meaningful test would be against the value 1.0, but 2.24 is not equal to $(16.87972 - 1)/21.31898$ either. The 2.24 and the P value next to it are simply carried over from the original logit model. The implied test is that the odds ratio equals one—it is implied by the equality of the coefficient to zero. The confidence interval would typically be computed as we did in the previous example, but again, the values shown are not equal to $16.87972 \pm 1.96\ (21.31809)$. They are equal to exp(0.35079) to exp(5.30143) which is the confidence interval from the original coefficient. This is logical—we have estimated a 95% confidence interval for β, so these values do provide a 95% interval for the exponent. In Section 4.8.3, we considered whether this would be the shortest 95% confidence interval for a prediction of y from $\ln y$, which is what we have done here, and discovered that it is not. On the other hand, it is unclear what utility that is not provided by the coefficient would be provided by the confidence interval for the odds ratio. Finally, as noted earlier, the odds ratio is useful for the conceptual experiment of changing the variable by one unit. For the *GPA* which ranges from 2 to 4 and for *PSI* which is a dummy variable, these would seem appropriate. *TUCE* is a test score that ranges around 30. A unit change in *TUCE* might not be as interesting.

17.3.4 INTERACTION EFFECTS

Models with **interaction effects**, such as

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3\, Education_{it} + \beta_4\, Income_{it} + \beta_5\, Kids_{it} + \beta_6\, Health_{it} + \beta_7\, Married_{it} + \beta_8\, Age_{it} \times Education_{it}),$$

have attracted considerable attention in recent applications of binary choice models.[27] A practical issue concerns the computation of partial effects by standard computer packages. Write the model as

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + \beta_7 x_{7it} + \beta_8 x_{8it}).$$

[27] See, for example, Ai and Norton (2004) and Greene (2010).

Estimation of the model parameters is routine. Rote computation of partial effects using (17-11) will produce

$$PE_8 = \partial \text{Prob}(DocVis > 0)/\partial x_8 = \beta_8 \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})],$$

which is what common computer packages will dutifully report. The problem is that $x_8 = x_2 x_3$, and PE_8 in the previous equation is *not* the partial effect for x_8—there is no meaningful partial effect for x_8 because $x_8 = x_2 x_3$. Moreover, the partial effects for x_2 and x_3 will also be misreported by the rote computation. To revert back to our original specification,

$$\partial \text{Prob}(DocVis > 0|\mathbf{x})/\partial\, Age = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_2 + \beta_8\, Education),$$
$$\partial \text{Prob}(DocVis > 0|\mathbf{x})/\partial\, Education = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_3 + \beta_8\, Age),$$

and what is computed as $\partial \text{Prob}(DocVis > 0|\mathbf{x})/\partial(Age \times Education)$ is meaningless. The practical problem motivating Ai and Norton (2004) was that the computer package does not know that x_8 is $x_2 x_3$, so it computes a partial effect for x_8 as if it could vary *partially* from the other variables. The (now) obvious solution is for the analyst to force the correct computations of the relevant partial effects by whatever software he or she is using, perhaps by programming the computations themselves.[28]

The practical complication raises a theoretical question that is less clear cut. What is the *interaction effect* in the model? In a linear model based on the preceding, we would have

$$\partial^2 E[y|\mathbf{x}]/\partial x_2 \partial x_3 = \beta_8,$$

which is unambiguous. However, in this *nonlinear* binary choice model, the correct result is

$$\partial^2 E[y|\mathbf{x}]/\partial x_2 \partial x_3 = \{\Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]\}\beta_8 + \{\Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})][1 - 2\Lambda(\mathbf{x}'\boldsymbol{\beta})]\}(\beta_2 + \beta_8\, Education)(\beta_3 + \beta_8 Age).$$

Not only is β_8 not the interesting effect, but there is also a complicated additional term. Loosely, we can associate the first term as a *direct* effect—note that it is the naïve term PE_8 from earlier. The second part can be attributed to the fact that we are differentiating a nonlinear model—essentially, the second part of the partial effect results from the nonlinearity of the function. The existence of an interaction effect in this model is inescapable—notice that the second part is nonzero (generally) even if β_8 does equal zero. Whether this is intended to represent an interaction in some economic sense is unclear. In the absence of the product term in the model, probably not. We can see an implication of this in Figure 17.1. At the point where $\mathbf{x}'\boldsymbol{\beta} = 0$, where the probability equals one half, the probability function is linear. At that point, $(1 - 2\Lambda)$ will equal zero and the functional form effect will be zero as well. When $\mathbf{x}'\boldsymbol{\beta}$ departs from zero, the probability becomes nonlinear. (These same effects can be shown for the probit model—at $\mathbf{x}'\boldsymbol{\beta} = 0$, the second derivative of the probit probability is $-\mathbf{x}'\boldsymbol{\beta}\phi(\mathbf{x}'\boldsymbol{\beta}) = 0$.)

[28]The practical issue is now widely understood. Modern computer packages are able to understand model specifications stated in structural form. For our example, rather than compute x_8, the user would literally specifically the instruction to the software as $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_2{*}x_3$ (not computing x_8) and the computation of partial effects would be done accordingly.

We developed an extensive application of interaction effects in a nonlinear model in Example 7.6. In that application, using the same data for the numerical exercise, we analyzed a nonlinear regression $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. The results obtained in that study were general, and will apply to the application here, where the nonlinear regression is $E[y|\mathbf{x}] = \Lambda(\mathbf{x}'\boldsymbol{\beta})$ or $\Phi(\mathbf{x}'\boldsymbol{\beta})$.

Example 17.12 Interaction Effect

We added an interaction term, *Age* × *Education*, to the model in Example 17.9. The model is now

$$\text{Prob}(DocVis_{it} > 0) = \Phi(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Education_{it} + \beta_4\, Income_{it} + \beta_5\, Kids_{it} + \beta_6\, Health_{it} + \beta_7\, Married_{it} + \beta_8\, Age_{it} \times Education_{it}).$$

Estimates of the model parameters appear in Table 17.6. Estimation of the probit model produces an estimate of β_8 of 0.00055. It is not clear what this measures. From the correctly specified and estimated model (with the explicit interaction term), the estimated partial effect for education is $\phi(\mathbf{x}'\boldsymbol{\beta})(\beta_3 + \beta_8 Age) = -0.00392$. By fitting the model with x_8 instead of x_2 times x_3, we obtain the first term as the (erroneous) partial effect of education, −0.01162. This implies that the second term, $\phi(\mathbf{x}'\boldsymbol{\beta})\beta_8 Age$, is −0.00392 + 0.01162 = 0.00770. As noted, the naïve calculation produces a value that has little to do with the desired result.

17.4 MEASURING GOODNESS OF FIT FOR BINARY CHOICE MODELS

There have been many fit measures suggested for discrete response models.[29] The general intent is to devise a counterpart to the R^2 in linear regression. The R^2 for a linear model provides two useful measures. First, when computed as $1 - \mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$, it measures the success of the estimator at optimizing (minimizing) the fitting criterion, $\mathbf{e}'\mathbf{e}$. That is the interpretation of R^2 as the proportion of the variation of y that is explained by the model. Second, when computed as $Corr^2(y, \mathbf{x}'\mathbf{b})$, it measures the extent to which the predictions of the model are able to mimic the actual data. Fit measures for discrete choice models are based on the same two ideas. We will discuss several.

17.4.1 FIT MEASURES BASED ON THE FITTING CRITERION

Most applications of binary choice modeling use a maximum likelihood estimator. The log-likelihood function itself is the fitting criterion, so as a starting point for considering the performance of the estimator, $\ln L_{\text{MLE}} = \Sigma_{i=1}^{n}[(1 - y_i)\ln(1 - \hat{P}_i) + y_i \ln \hat{P}_i]$ is computed using the MLEs of the parameters. Following the first motivation for R^2, the hypothesis that all the slopes in the model are zero is often interesting. The log likelihood computed with only a constant term will be $\ln L_0 = n[P_0 \ln P_0 + P_1 \ln P_1]$ where n is the sample size and P_j is the sample proportion of zeros or ones. (*Note:* $\ln L_0$ is based only on the sample proportions, so it will be the same regardless of the model.) McFadden's (1974) "Pseudo R^2" or "likelihood ratio index" is

$$R^2_{Pseudo} = LRI = 1 - \frac{\ln L_{MLE}}{\ln L_0}.$$

[29] See, for example, Cragg and Uhler (1970), Amemiya (1981), Maddala (1983), McFadden (1974), Ben-Akiva and Lerman (1985), Kay and Little (1986), Veall and Zimmermann (1992), Zavoina and McKelvey (1975), Efron (1978), and Cramer (1999). A survey of techniques appears in Windmeijer (1995). See, as well, Long and Freese (2006, Sec. 3.5) for a catalog of fit measures for discrete dependent variable models.

This measure has an intuitive appeal in that it is bounded by zero and one and it increases when variables are added to the model.[30] If all the slope coefficients (but not the constant term) are zero, then R^2_{Pseudo} equals zero. Unlike R^2, there is no way to make R^2_{Pseudo} reach one. Moreover, the values between zero and one have no natural interpretation. If $P(\mathbf{x}_i'\boldsymbol{\beta})$ is a proper cdf, then even with many regressors the model cannot fit perfectly unless $\mathbf{x}_i'\boldsymbol{\beta}$ goes to $+\infty$ or $-\infty$. As a practical matter, it does happen. But when it does, it indicates a flaw in the model, not a good fit. If the range of one of the independent variables contains a value, say x^*, such that the sign of $(x - x*)$ predicts y perfectly and vice versa, then the model will become a perfect predictor. This result also holds in general if the sign of $\mathbf{x}'\boldsymbol{\beta}$ gives a perfect predictor for some vector $\boldsymbol{\beta}$. For example, one might mistakenly include as a regressor a dummy variable that is identical, or nearly so, to the dependent variable. In this case, the maximization procedure will break down precisely because $\mathbf{x}'\boldsymbol{\beta}$ is diverging during the iterations.[31]

Notwithstanding all of the preceding, this statistic is very commonly reported with empirical results, with references to "fit" and even "proportion of variation explained." A "degrees of freedom correction," $\overline{R}^2_{Pseudo} = 1 - \dfrac{\ln L_{MLE} - K}{\ln L_0}$, has been suggested, as well as some similar ad hoc "adjustments," such as the "Cox and Snell $R^2_{CS} = 1 - \exp(-(\ln \mathrm{L}_M - \ln \mathrm{L}_0)/n)$. We note, however, none of these are fit measures in the familiar sense, and they are not R^2-like measures of explained variation. As a final note, another shortcoming of these measures is that they are based on a particular estimation criterion. There are other estimators for binary choice models, as shown in Example 17.14.

The pseudo R^2 will be most useful for comparing one model to another. If the models are nested, then the log-likelihood function is the natural choice, as examined in the next section. For more general cases, researchers often use one of the information criteria, typically the Akaike Information Criterion,

$$AIC = -2 \ln L + 2K \qquad \text{or} \qquad AIC/n,$$

or Schwartz's Bayesian Information Criterion,

$$BIC = -2 \ln L + K \ln n \qquad \text{or} \qquad BIC/n.$$

In general, a lower IC value suggests a better model. In comparing nonnested models, some care is needed in interpreting this result, however.

17.4.2 FIT MEASURES BASED ON PREDICTED VALUES

Fit measures based on the predicted probabilities rather than the log likelihood have also been suggested. For example, Efron (1978) proposed a direct counterpart to R^2,

$$R^2_{Efron} = 1 - \frac{\Sigma_{i=1}^n (y_i - \hat{P}_i)^2}{\Sigma_{i=1}^n (y_i - \bar{y})^2}.$$

[30]The log likelihood for a binary choice model must be negative as it is a sum of logs of probabilities. The model with fewer variables is a restricted version of the larger model so it must have a smaller log likelihood. Thus, the log-likelihood function increases when variables are added to the model, and the LRI must be between zero and one. For models with continuous variables, the log likelihood can be positive, so these appealing results are not assured.

[31]See McKenzie (1998) for an application and discussion.

The ambiguity in this measure comes from treating $(y_i - \hat{P}_i)$ as a quantitative residual when the y_i is actually only a label of the outcome. Ben-Akiva and Lerman (1985) and Kay and Little (1986) suggested a fit measure that is keyed to the prediction rule,

$$R^2_{\text{BL}} = \frac{1}{n}\sum_{i=1}^{n}[y_i\hat{P}_i + (1 - y_i)(1 - \hat{P}_i)],$$

which can be written as a simple weighted average of the mean predicted probabilities of the two outcomes, $R^2_{\text{BL}} = P_0\hat{P}_0 + P_1\hat{P}_1$. A difficulty in this computation is that in unbalanced samples, the less frequent outcome will usually be predicted very badly by the standard procedure, and this measure does not pick up that point. Cramer (1999) and Tjur (2009) have suggested an alternative measure, the *coefficient of discrimination*, that directly considers this failure,

$$\begin{aligned}\lambda &= (\text{average } \hat{P}|\, y_i = 1) - (\text{average } \hat{P}|\, y_i = 0)\\ &= (\text{average}(1 - \hat{P})|\, y_i = 0) - (\text{average}(1 - \hat{P})|\, y_i = 1).\end{aligned}$$

This measure heavily penalizes the incorrect predictions, and because each proportion is taken within the subsample, it is not unduly influenced by the large proportionate size of the group of more frequent outcomes.

A useful summary of the predictive ability of the model is a 2×2 table of the hits and misses of a prediction rule such as

$$\hat{y} = 1 \quad \text{if } \hat{F} > F^* \text{ and 0 otherwise.} \qquad \textbf{(17-30)}$$

(In information theory, this is labeled a *confusion matrix*.) The usual threshold value is 0.5, on the basis that we should predict a one if the model says a one is more likely than a zero. Consider, for example, the naïve predictor

$$\hat{y} = 1 \quad \text{if } P > 0.5 \text{ and 0 otherwise,} \qquad \textbf{(17-31)}$$

where P is the simple proportion of ones in the sample. This rule will always predict correctly 100 P% of the observations, which means that the naïve model does not have zero fit. In fact, if the proportion of ones in the sample is very high, it is possible to construct examples in which the second model will generate more correct predictions than the first! Once again, this flaw is not in the model; it is a flaw in the fit measure.[32] The important element to bear in mind is that the coefficients of the estimated model are not chosen so as to maximize this (or any other) fit measure, as they are in the linear regression model where **b** maximizes R^2.

Another consideration is that 0.5, although the usual choice, may not be a very good value to use for the threshold. If the sample is **unbalanced**—that is, has many more ones than zeros, or vice versa—then by this prediction rule it might never predict a one (or zero). To consider an example, suppose that in a sample of 10,000 observations, only 1,000 have $Y = 1$. We know that the average predicted probability in the sample will be 0.10. As such, it may require an extreme configuration of regressors even to produce a $\hat{P}$ of 0.2, to say nothing of 0.5. In such a setting, the prediction rule may fail every time to predict when $Y = 1$. The obvious adjustment is to reduce F^*. Of course, this adjustment comes at a cost. If we reduce the threshold F^* so as to predict $y = 1$ more often, then we will increase the number of correct classifications of observations that do

[32] See Amemiya (1981).

have $y = 1$, but we will also increase the number of times that we *incorrectly* classify as ones observations that have $y = 0$.[33] In general, any prediction rule of the form in (17-30) will make two types of errors: It will incorrectly classify zeros as ones and ones as zeros. In practice, these errors need not be symmetric in the costs that result. For example, in a credit scoring model, incorrectly classifying an applicant as a bad risk is not the same as incorrectly classifying a bad risk as a good one.[34] Changing F^* will always reduce the probability of one type of error while increasing the probability of the other. There is no correct answer as to the best value to choose. It depends on the setting and on the criterion function upon which the prediction rule depends.

17.4.3 SUMMARY OF FIT MEASURES

The likelihood ratio index and various modifications of it are related to the likelihood ratio statistic for testing the hypothesis that the coefficient vector is zero. Cramer's measure is oriented more toward the relationship between the fitted probabilities and the actual values. It is usefully tied to the standard prediction rule $\hat{y} = \mathbf{1}[\hat{P} > 0.5]$. Whether these have a close relationship to any type of fit in the familiar sense is uncertain. In some cases, it appears so. But the maximum likelihood estimator, on which many of the fit measures are based, is not chosen so as to maximize a fitting criterion based on prediction of y as it is in the linear regression model (which maximizes R^2). It is chosen to maximize the joint density of the observed dependent variables. It remains an interesting question for research whether fitting y well or obtaining good parameter estimates is a preferable estimation criterion. Evidently, they need not be the same thing.

Example 17.13 Prediction with a Probit Model

Tunali (1986) estimated a probit model in a study of migration, subsequent remigration, and earnings for a large sample of observations of male members of households in Turkey. Among his results, he reports the confusion matrix shown here for a probit model: The estimated model is highly significant, with a likelihood ratio test of the hypothesis that the coefficients (16 of them) are zero based on a chi-squared value of 69 with 16 degrees of freedom.[35] The model predicts 491 of 690, or 71.2%, of the observations correctly, although the likelihood ratio index is only 0.083. A naïve model, which always predicts that $y = 0$ because $P < 0.5$, predicts 487 of 690, or 70.6%, of the observations correctly. This result is hardly suggestive of no fit. The maximum likelihood estimator produces several significant influences on the probability but makes only four more correct predictions than the naïve predictor.[36]

		Predicted	
	D = 0	**D = 1**	***Total***
Actual **D = 0**	471	16	487
D = 1	183	20	203
Total	654	36	690

[33]The technique of discriminant analysis is used to build a procedure around this consideration. In this setting, we consider not only the number of correct and incorrect classifications, but also the cost of each type of misclassification.

[34]See Boyes, Hoffman, and Low (1989).

[35]This view actually understates slightly the significance of his model, because the preceding predictions are based on a bivariate model. The likelihood ratio test fails to reject the hypothesis that a univariate model applies, however.

[36]It is also noteworthy that nearly all the correct predictions of the maximum likelihood estimator are the zeros. It hits only 10% of the ones in the sample.

Example 17.14 Fit Measures for a Logit Model

Table 17.9 presents estimates of a logit model for the specification in Example 17.12. Results ML1 are the MLEs for the full model. ML2 is a restricted version from which *Age*, *Education,* and *Health* are excluded. The variables removed are highly significant; the chi-squared statistic for the four restrictions is 2(2,137.06 − 1,991.13) = 291.86. The critical value for 95% from the chi-squared table with four degrees of freedom is 9.49, so the excluded variables significantly contribute to the likelihood for the data. We consider the fit of the model based on the measures suggested earlier. The results labeled NLS in Table 17.9 were computed by nonlinear least squares, rather than MLE. The criterion function is $SS(\mathbf{b}_{NLS}) = \Sigma_i(y_i - \Lambda(\boldsymbol{\beta}'\mathbf{x}_i)^2$. We are interested in how the fit obtained by this alternative estimator compares to that obtained by the MLE. Table 17.10 shows the various scalar fit measures. Note, first, the log likelihood strongly favors ML1. The nonlinear least squares estimates appear rather different from the MLEs but produce nearly the same log likelihood. However, the statistically significant coefficients, on *Kids*, *Health,* and *Married*, are actually almost the same, which would explain the finding. The information criteria favor ML1 as might be expected. The predictive influence of the excluded variables in ML2 is clear in the scalar measures, which generally rise from about 0.01 to 0.10. The Ben-Akiva and Lerman measure does not discriminate between the two specifications. Cramer and the others are essentially the same. Based on the confusion matrices, the count R^2 underscores the difficulty of summarizing the fit of the model to the data. The two models do essentially equally well, though, at predicting different outcomes. ML1 predicts the zeros much better than ML2, but at the cost of many more erroneous predictions of the observations with y equal to one. Overall, the results for this model are typical. The ambiguity of the overall picture suggests the difficulty of constructing a single scalar measure of fit for a binry choice model. The comparison between ML1 and ML2 provided by the Cramer or the other measures seems appropriate. However, it is unclear how to interpret the 0.10 value for the fit measures. It obviously does not reflect a "proportion of explained variation." Nor, however, does it (or the pseudo R^2) have any connection to the ability of the model to predict the outcome variable—the standard predictor obtains a 67.3% success rate. But the naïve predictor, Doctor = 1, will predict correctly 2,222/3,377 or 65.8% of the cases, so the full model improves the success rate from 65.8% to 67.3%

TABLE 17.9 Estimated Parameters for Logit Model for Prob (Doctor=1)
(Absolute values of z statistics in parentheses for model ML1)

	Maximum Likelihood ML1	***Maximum Likelihood ML2***	***Nonlinear Least Squares NLS***
Constant	3.18430 (4.00)	0.85360	2.98328
Age	−0.00097 (0.05)	0.00000	0.00294
Education	−0.05054 (0.18)	0.00000	−0.03707
Income	−0.15076 (0.81)	−0.52235	−0.09437
Kids	−0.41358 (4.50)	−0.57608	−0.42014
Health	−0.30957 (14.9)	0.00000	−0.30032
Married	0.17415 (1.71)	0.37995	0.17301
Age × *Education*	0.00072 (0.47)	0.00000	0.00028

TABLE 17.10 Fit Measures for Estimated Logit Models

	ML1	*ML2*	*NLS*
Based on the log likelihood			
Ln L_0	−2,169.27	−2,169.27	−2,169.27
Ln L_M	−1,991.13	−2,137.06	−1,991.41
Chi squared[df]	356.28[7]	64.41[3]	
Pseudo R^2	0.08212	0.01484	0.0819923
Adjusted Pseudo R^2	0.07889	0.01162	0.0787654
AIC	3,998.27	4,290.13	3,998.81
AIC/n	1.18397	1.27040	1.18413
BIC	4,047.26	4,339.12	4,047.81
BIC/n	1.19848	1.28491	1.19864
Based on the predicted outcomes			
Cramer R^2	0.09840	0.01867	0.09644
Cox-Snell R^2	0.10013	0.01889	0.09998
Efron R^2	0.09736	0.01827	0.09750
Ben-Akiva – Lerman R^2	0.54992	0.54992	0.54954
Count R^2	0.67338	0.65591	0.67516
Confusion Matrix	$\begin{bmatrix} 289 & 866 & 1155 \\ 237 & 1985 & 2222 \\ 526 & 2851 & 3377 \end{bmatrix}$	$\begin{bmatrix} 17 & 1138 & 1155 \\ 24 & 2198 & 2222 \\ 41 & 3336 & 3377 \end{bmatrix}$	$\begin{bmatrix} 285 & 870 & 1155 \\ 227 & 1995 & 2222 \\ 512 & 2865 & 3377 \end{bmatrix}$

17.5 SPECIFICATION ANALYSIS

In the linear regression model, we considered two important specification problems: the effect of omitted variables and the effect of heteroscedasticity. In the linear regression model, $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, when least squares estimates $\mathbf{b}_1$ are computed omitting $\mathbf{X}_2$,

$$E[\mathbf{b}_1] = \boldsymbol{\beta}_1 + [\mathbf{X}_1'\mathbf{X}_1]^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2,$$

unless $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal or $\boldsymbol{\beta}_2 = \mathbf{0}$, $\mathbf{b}_1$ is biased. If we ignore heteroscedasticity, then although the least squares estimator is still unbiased and consistent, it is inefficient and the usual estimate of its sampling covariance matrix is inappropriate. Yatchew and Griliches (1984) have examined these same issues in the setting of the probit and logit models. In the context of a binary choice model, they find the following:

1. If x_2 is omitted from a model containing x_1 and x_2, (i.e., $\boldsymbol{\beta}_2 \neq 0$) then

$$\text{plim } \hat{\boldsymbol{\beta}}_1 = c_1\boldsymbol{\beta}_1 + c_2\boldsymbol{\beta}_2,$$

where c_1 and c_2 are complicated functions of the unknown parameters. The implication is that even if the omitted variable is uncorrelated with the included one, the coefficient on the included variable will be inconsistent.

2. If the disturbances in the underlying model, $y = \mathbf{1}[(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon) > 0]$, are heteroscedastic, then the maximum likelihood estimators are inconsistent and

the covariance matrix is inappropriate. This is in contrast to the linear regression case, where heteroscedasticity only affects the estimated asympotic variance of the estimator.

In both of these cases (and others), the impact of the specification error on estimates of partial effects and predictions is less clear, but probably of greater interest.

Any of the three methods of hypothesis testing discussed here can be used to analyze these two specification problems. The Lagrange multiplier test has the advantage that it can be carried out using the estimates from the restricted model, which might bring a saving in computational effort for the test for heteroscedasticity.[37] To reiterate, the Lagrange multiplier statistic is computed as follows. Let the null hypothesis, H_0, be a specification of the model, and let H_1 be the alternative. For example, H_0 might specify that only variables $\mathbf{x}_1$ appear in the model, whereas H_1 might specify that $\mathbf{x}_2$ appears in the model as well. It is assumed that the null model is nested in the alternative. The statistic is

$$\text{LM} = \mathbf{g}_0' \mathbf{V}_0^{-1} \mathbf{g}_0,$$

where $\mathbf{g}_0$ is the vector of derivatives of the log likelihood as specified by H_1 but evaluated at the maximum likelihood estimator of the parameters assuming that H_0 is true, and $\mathbf{V}_0^{-1}$ is any of the consistent estimators of the asymptotic variance matrix of the maximum likelihood estimator under H_1, also computed using the maximum likelihood estimators based on H_0. The statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions.

17.5.1 OMITTED VARIABLES

The hypothesis to be tested is

$$H_0: y^* = \mathbf{x}_1' \boldsymbol{\beta}_1 + \varepsilon,$$
$$H_1: y^* = \mathbf{x}_1' \boldsymbol{\beta} + \mathbf{x}_2' \boldsymbol{\beta}_2 + \varepsilon,$$

so the test is of the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$. The Lagrange multiplier test would be carried out as follows:

1. Estimate the model in H_0 by maximum likelihood. The restricted coefficient vector is $[\hat{\boldsymbol{\beta}}_1, \mathbf{0}]$.
2. Let $\mathbf{x}$ be the compound vector, $[\mathbf{x}_1, \mathbf{x}_2]$.

The statistic is then computed according to (17-27) or (17-28). For a logit model, for example, the test is carried out as follows: (1) Fit the null model by ML; (2) Compute the fitted probabilities using the null model and the "residuals," $e_i = y_i - P_{i,0}$ arranged in diagonal matrix $\mathbf{E}$; (3) The LM statistic is $\mathbf{1}'\mathbf{E}\mathbf{X}(\mathbf{X}'\mathbf{E}^2\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}'\mathbf{1}$. As usual, this can be computed as n times an uncentered R^2, here in the regression of a column of ones on variables $e_i\mathbf{x}_i$. The likelihood ratio test is equally straightforward. Using the estimates of the two models, the statistic is simply $2(\ln L_1 - \ln L_0)$. The Wald statistic would be based on estimates of the alternative model and is computed as in (17-25).

[37] The results in this section are based on Davidson and MacKinnon (1984) and Engle (1984). A symposium on the subject of specification tests in discrete choice models is Blundell (1987).

17.5.2 HETEROSCEDASTICITY

We use the standard formulation analyzed by Harvey (1976)[38] (see Section 14.10.3), $\text{Var}[\varepsilon|\mathbf{z}] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2$. We will obtain results specifically for the probit model; the logit or other models are essentially the same.

The starting point is an extension of the binary choice model,

$$y* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, y = \mathbf{1}(y* > 0),$$
$$E[\varepsilon|\mathbf{x}, \mathbf{z}] = 0, \text{Var}[\varepsilon|\mathbf{x}, \mathbf{z}] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2.$$

There is an ambiguity in the formulation of the model. A nonlinear index function, probit model (with no suggestion of heteroscedasticity),

$$y^{**} = \frac{\mathbf{x}'\boldsymbol{\beta}}{\exp(\mathbf{z}'\boldsymbol{\gamma})} + \varepsilon, \quad y = \mathbf{1}(y** > 0), \varepsilon \sim \text{N}[0,1],$$

leads to the identical log likelihood and the identical estimated parameters. It is not possible to distinguish heteroscedasticity from this nonlinearity in the conditional mean function.[39] Unlike the linear regression model, in this binary choice context, the data contain no direct (identifying) information about scaling, or variation of the dependent variable. (Hence, the *observational equivalence* of the two specifications.) The (identical) signs of $y*$ and $y**$ are unaffected by the variance function. More broadly, the binary choice model creates an ambiguity in the distinction between heteroscedasticity and variation in the mean of the underlying regression.

The presence of heteroscedasticity requires some care in interpreting the coefficients. For a variable w_k that could be in $\mathbf{x}$ or $\mathbf{z}$ or both,

$$\frac{\partial \text{Prob}(y = 1|\mathbf{x}, \mathbf{z})}{\partial w_k} = \left\{\phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}}{\exp(\mathbf{z}'\boldsymbol{\gamma})}\right]\frac{1}{\exp(\mathbf{z}'\boldsymbol{\gamma})}\right\}(\beta_k - (\mathbf{x}'\boldsymbol{\beta})\gamma_k). \tag{17-32}$$

Only the first (second) term applies if w_k appears only in $\mathbf{x}$ ($\mathbf{z}$). This implies that the simple coefficient may differ greatly from the effect that is of interest in the estimated model. This effect is clearly visible in the next example.[40]

The log likelihood is

$$\ln L = \sum_{i=1}^{n}\left\{y_i \ln F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right) + (1 - y_i)\ln\left[1 - F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right)\right]\right\}. \tag{17-33}$$

---

[38]See Knapp and Seaks (1992) for an application. Other formulations are suggested by Fisher and Nagin (1981), Hausman and Wise (1978), Horowitz (1993), and Khan (2013).

[39]See Khan (2013) for extensive discussion of this observational equivalence. Manski (1988) notes this as well.

[40]Wooldridge (2010, pp. 602–603) develops the identification issue in terms of the *average structural function* [Blundell and Powell (2004)]; $\text{ASF}(\mathbf{x}) = E_{\mathbf{z}}[\Phi(\exp(-\mathbf{z}'\boldsymbol{\gamma})\mathbf{x}'\boldsymbol{\beta})]$. Under this interpretation, the partial effect is $\partial\text{ASF}(\mathbf{x})/\partial\mathbf{x} = E_{\mathbf{z}}[\phi(\exp(-\mathbf{z}'\boldsymbol{\gamma})\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}]$. The Average Structural Function treats $\mathbf{z}$ and $\mathbf{x}$ differently (even if they share variables). This computes the function for a fixed $\mathbf{x}$, averaging over the sample values of $\mathbf{z}$. The empirical estimator would be $\partial\hat{A}\hat{S}F(\mathbf{x})/\partial\mathbf{x} = (1/n)\Sigma_{i=1}^{n}\phi[\exp(-\mathbf{z}_i'\hat{\boldsymbol{\gamma}})\mathbf{x}'\hat{\boldsymbol{\beta}}]\hat{\boldsymbol{\beta}}$. The author suggests "the uncomfortable conclusion is that we have no convincing way of choosing" between (17-32) and this alternative result. Recent applications generally report (17-32), notwithstanding this alternative interpretation. One advantage of interpretation (17-32) is that it explicitly examines the effect of variation in $\mathbf{z}$ on the response probability, particularly in the typical case in which $\mathbf{z}$ and $\mathbf{x}$, have variables in common.

To be able to estimate all the parameters, $\mathbf{z}$ cannot have a constant term. The derivatives are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{f_i(y_i - F_i)}{F_i(1 - F_i)} \right] \exp(-\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{x}_i,$$
$$\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \left[\frac{f_i(y_i - F_i)}{F_i(1 - F_i)} \right] \exp(-\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i(-\mathbf{x}_i'\boldsymbol{\beta}). \quad \textbf{(17-34)}$$

If the model is estimated assuming that $\boldsymbol{\gamma} = \mathbf{0}$, then we can easily test for homoscedasticity. Let g_i equal the bracketed function in (17-34), $\mathbf{G} = \text{diag}(g_i)$ and

$$\mathbf{w}_i = \begin{bmatrix} \mathbf{x}_i \\ (-\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{z}_i \end{bmatrix}, \quad \textbf{(17-35)}$$

computed at the maximum likelihood estimator, assuming that $\boldsymbol{\gamma} = \mathbf{0}$. Then, the LM statistic is

$$\text{LM} = \mathbf{i}'\mathbf{GW}[(\mathbf{W}'\mathbf{G})(\mathbf{GW})]^{-1}\mathbf{W}'\mathbf{Gi} = nR^2,$$

where the regression is of a column of ones on $g_i\mathbf{w}_i$. Wald and likelihood ratio tests of the hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$ are also straightforward based on maximum likelihood estimates of the full model.

Davidson and MacKinnon (1981) carried out a Monte Carlo study to examine the true sizes and power functions of these tests. As might be expected, the test for omitted variables is relatively powerful. The test for heteroscedasticity may pick up some other form of misspecification, however, including perhaps the simple omission of $\mathbf{z}$ from the index function, so its power may be problematic. It is perhaps not surprising that the same problem arose earlier in our test for heteroscedasticity in the linear regression model. The problem in the binary choice context stems partly from the ambiguous interpretation of the role of $\mathbf{z}$ in the model discussed earlier.

Example 17.15 Specification Test in a Labor Force Participation Model

Using the data described in Example 17.1, we fit a probit model for labor force participation based on the following specification [see Wooldridge (2010, p. 580)]:[41]

$$\text{Prob}[LFP = 1] = F(\textit{Constant}, \textit{Other Income}, \textit{Education}, \textit{Experience}, \textit{Experience}^2, \textit{Age}, \textit{Kids Under 6}, \textit{Kids 6 to 18}).$$

For these data, $P = 428/753 = 0.568393$. The restricted (all slopes equal zero, free constant term) log likelihood is $325 \times \ln(325/753) + 428 \times \ln(428/753) = -514.8732$. The unrestricted log likelihood for the probit model is -401.3022. The chi-squared statistic is, therefore, 227.142. The critical value from the chi-squared distribution with seven degrees of freedom is 14.07, so the joint hypothesis that the coefficients on *Other Income, etc.* are all zero is rejected.

Consider the alternative hypothesis, that the constant term and the coefficients on *Other Income, etc.* are the same whether the individual resides in a city (CITY = 1) or not (CITY = 0), against the alternative that an altogether different equations apply for the two

[41] Other income is computed as family income minus the wife's hours times the wife's reported wage, divided by 1,000. This produces several small negative values. In the interest of comparability to the received application, we have left these values intact.

TABLE 17.11 Estimated Coefficients

		Homoscedastic		*Heteroscedasti*	
		Estimate (Std. Err.)	***Partial Effect****	***Estimate (Std. Err.)***	***Partial Effect****
Constant	β_1	0.27008 (0.5086)	–	0.25140 (0.4548)	–
Other Inc.	β_2	−0.01202 (0.0048)	−0.00362 (0.0014)	−0.01075 (0.0044)	−0.00362 (0.0014)
Education	β_3	0.13090 (0.0253)	0.39370 (0.0072)	0.11734 (0.0255)	0.03949 (0.0072)
Exper	β_4	0.12335 (0.0187)	0.02558 (0.0022)	0.11190 (0.0197)	0.02599 (0.0022)
$Exper^2$	β_5	−0.00189 (0.0006)		−0.00171 (0.0006)	
Age	β_6	−0.05285 (0.0085)	−0.01590 (0.0024)	−0.04774 (0.0089)	−0.01607 (0.0024)
Kids < 6	β_7	−0.86833 (0.1185)	−0.26115 (0.0131)	−0.77151 (0.1356)	−0.25968 (0.0318)
Kids 6–18	β_8	0.03600 (0.0438)	0.01083 (0.0319)	0.02800 (0.0390)	0.00943 (0.0130)
City	γ	0.00000		−0.17446 (0.1541)	0.00843 (0.0075)
$\ln L$		−401.302		−400.641	

*Average partial effects and estimated standard errors include both mean ($\boldsymbol{\beta}$) and variance ($\boldsymbol{\gamma}$) effects.

groups of women. To test this hypothesis, we would use a counterpart to the Chow test of Section 6.4.1 and Example 6.9. The restricted model in this instance would be based on the pooled data set of all 753 observations. The log likelihood for the pooled model—which has a constant term and the seven variables listed above—is −401.302. The log likelihoods for this model based on the 484 observations with CIT = 1 and the 269 observations with CIT = 0 are −255.552 and −142.727, respectively. The log likelihood for the unrestricted model with separate coefficient vectors is thus the sum, −398.279. The chi-squared statistic for testing the eight restrictions of the pooled model is twice the difference, 6.046. The 95% critical value from the chi-squared distribution with 8 degrees of freedom is 15.51, so at this significance level, the hypothesis that the constant terms and the other coefficients are all the same is not rejected.

Table 17.11 presents estimates of the probit model with a correction for heteroscedasticity of the form $\text{Var}[\varepsilon_i] = [\exp(\gamma CITY)]^2$. The three tests for homoscedasticity give

$$\text{LR} = 2[-400.641 - (-401.302)] = 1.322,$$
$$\text{LM} = 1.362 \text{ based on the BHHH estimator,}$$
$$\text{Wald} = (-1.13)^2 = 1.276.$$

The 95% critical value for one restrictions is 3.84 so the three tests are consistent in not rejecting the hypothesis that γ equals zero.

17.5.3 DISTRIBUTIONAL ASSUMPTIONS

One concern about the models suggested here is that the choice of the particular distribution is itself vulnerable to a specification error. For example, the problem arises if a probit model is analyzed when a logit model would be appropriate.[42] It might seem logical to test the hypothesis of the model along with the other specification analyses one might do. Alternatively, a more robust, less parametric specification might be attractive. The substantive difference between probit and logit coefficient estimates in the preceding examples (e.g., Example 17.3) is misleading. The difference masks the underlying scaling of

[42]See, for example, Ruud (1986).

the distributions. The partial effects generated by the models are typically almost identical. This is a widely observed result that suggests that concerns about biases in the coefficients due to the wrong distribution might be misplaced. The other element of the analysis is the predicted probabilities. Once again, the scaling of the coefficients by the different models disguises the typical similarity of the predicted probabilities of the different parametric models. A broader question concerns the specific distribution compared to a semi- or nonparametric alternative. Manski's (1988) maximum score estimator [and Horowitz's (1992) smoothed version], Klein and Spady's (1993) semiparametric (kernel function based), and Khan's (2013) heteroscedastic probit model are a few of the less heavily parameterized specifications that have been proposed for binary choice models. Frolich (2006) presents a comprehensive survey of nonparametric approaches to binary choice modeling, with an application to Portuguese female labor supply.

The linear probability model is not offered as a robust alternative specification for the choice model. Proponents of the linear probability model argue only that the linear regression delivers a reliable approximation to the partial effects of the underlying true probability model.[43] The robustness aspect is speculative. The approximation does appear to mimic the nonlinear results in many cases. In terms of the relevant computations, partial effects and predicted probabilities, the various candidates seem to behave similarly. An essential ingredient is often the curvature in the tails that allows predicted probabilities to mimic the features of unbalanced samples. From this standpoint, the linear model would seem to be the less robust specification. (See Example 17.5.) It is precisely this rigidity of the LPM (as well as the parametric models) that motivates the nonparametric approaches such as the local likelihood logit approach advocated by Frolich (2006).

Example 17.16 Distributional Assumptions

Table 17.12 presents estimates of the model in Example 17.36 based on the linear probability model and four alternative specifications. Only the estimated partial effects are shown in the table. The probit estimates match the authors' results. The correspondence of the various results is consistent with the earlier observations. Generally, the models produce similar results. The linear probability model does stand alone for two of the seven results, for the market share and productivity variables.

TABLE 17.12 Estimated Partial Effects in a Model of Innovation

	Linear	*Probit*	*Logit*	*Complementary Log Log*	*Gompertz*
Log Sales	0.05198	0.06573	0.06766	0.06457	0.06639
Share	0.09492	0.39812	0.43993	0.33011	0.49826
Imports	0.45284	0.42080	0.41101	0.43734	0.40304
FDI	1.07787	1.05890	1.08753	0.99556	1.12929
Productivity	−0.55012	−0.86887	−1.01060	−0.85039	−0.87471
Raw Material	−0.09861	−0.10569	−0.09635	−0.10626	−0.10615
Investment	0.07879	0.07045	0.06758	0.07704	0.06356

[43]Chung and Goldberger (1984), Stoker (1986, 1992), and Powell (1994) (among others) consider general cases in which $\boldsymbol{\beta}$ can be consistently estimated "up to scale" using ordinary least squares. For example, Stoker (1986) shows that if $\mathbf{x}$ is multivariate normally distributed, then the LPM would provide a consistent estimator of the slopes of the probability function under very general specifications.

17.5.4 CHOICE-BASED SAMPLING

In some studies, the mix of ones and zeros in the observed sample of the dependent variable is deliberately skewed in favor of one outcome or the other to achieve a more balanced sample than random sampling would produce.[44] The sampling is said to be **choice based**. In the studies noted, the dependent variable measured the occurrence of loan default, which is a relatively uncommon occurrence. To enrich the sample, observations with $y = 1$ (default) were oversampled. Intuition should suggest (correctly) that the bias in the sample should be transmitted to the parameter estimates, which will be estimated so as to mimic the sample, not the population, which is known to be different. Manski and Lerman (1977) derived the weighted exogenous sampling maximum likelihood (WESML) estimator for this situation. The estimator requires that the true population proportions, ω_1 and ω_0, be known. Let p_1 and p_0 be the sample proportions of ones and zeros. Then the estimator is obtained by maximizing a weighted log likelihood,

$$\ln L = \sum_{i=1}^{n} w_i \ln F(q_i \mathbf{x}_i'\boldsymbol{\beta}),$$

where $w_i = y_i(\omega_1/p_1) + (1 - y_i)(\omega_0/p_0)$. Note that w_i takes only two different values. The derivatives and the Hessian are likewise weighted. A final correction is needed after estimation; the appropriate estimator of the asymptotic covariance matrix is the sandwich estimator discussed in Section 17.3.1, $(-\mathbf{H})^{-1}(\mathbf{B})(-\mathbf{H})^{-1}$ (with weighted $\mathbf{B}$ and $\mathbf{H}$), instead of $\mathbf{B}$ or $\mathbf{H}$ alone. (The weights are not squared in computing $\mathbf{B}$.) WESML and the choice-based sampling estimator are not the free lunch they may appear to be. That which the biased sampling does, the weighting undoes. It is common for the end result to be very large standard errors, which might be viewed as unfortunate, insofar as the purpose of the biased sampling was to balance the data precisely to avoid this problem.

Example 17.17 Credit Scoring

In Example 7.12, we examined the spending patterns of a sample of 10,499 cardholders for a major credit card vendor. The sample of cardholders is a subsample of 13,444 applicants for the credit card. Applications for credit cards, then (1992) and now, are processed by a major nationwide processor, Fair Isaacs, Inc. The algorithm used by the processors is proprietary. However, conventional wisdom holds that a few variables are important in the process, such as *Age*, *Income*, *OwnRent* (whether the applicant owns hi or her home), *Self-Employed* (whether he or she is self-employed), and how long the applicant has lived at his or her current address. The number of major and minor derogatory reports (60-day and 30-day delinquencies) are also very influential variables in credit scoring. The probit model we will use to "model the model" is

$$\begin{aligned}\text{Prob}(\textit{Cardholder} = 1) &= \text{Prob}(C = 1 \,|\, \mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2\, \textit{Age} + \beta_3\, \textit{Income} + \beta_4\, \textit{OwnRent} \\ &\quad + \beta_5\, \textit{Months Living at Current Address} \\ &\quad + \beta_6\, \textit{Self-Employed} \\ &\quad + \beta_7\, \textit{Number of major derogatory reports} \\ &\quad + \beta_8\, \textit{Number of minor derogatory reports}).\end{aligned}$$

[44] For example, Boyes, Hoffman, and Low (1989) and Greene (1992).

TABLE 17.13 Estimated Card Application Equation (*t* ratios in parentheses)

	Unweighted			*Weighted*		
Variable	***Estimate***	***Std. Error***		***Estimate***	***Std. Error***	
Constant	0.31783	0.05094	(6.24)	−1.13089	0.04725	(−23.94)
Age	0.00184	0.00154	(1.20)	0.00156	0.00145	(1.07)
Income	0.00095	0.00025	(3.86)	0.00094	0.00024	(3.92)
OwnRent	0.18233	0.03061	(5.96)	0.23967	0.02968	(8.08)
CurrentAddress	0.02237	0.00120	(18.67)	0.02106	0.00109	(19.40)
SelfEmployed	−0.43625	0.05585	(−7.81)	−0.47650	0.05851	(−8.14)
Major Derogs	−0.69912	0.01920	(−36.42)	−0.64792	0.02525	(−25.66)
Minor Derogs	−0.04126	0.01865	(−2.21)	−0.04285	0.01778	(−2.41)

In the data set, 78.1% of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2%, so the sample is substantially choice based on this variable. The sample was deliberately skewed in favor of cardholders for purposes of the original study [Greene (1992)]. The weights to be applied for the WESML estimator are 0.232/0.781 = 0.297 for the observations with $C = 1$ and 0.768/0.219 = 3.507 for observations with $C = 0$. Table 17.13 presents the unweighted and weighted estimates for this application. The change in the estimates produced by the weighting is quite modest, save for the constant term. The results are consistent with the conventional wisdom that *Income* and *OwnRent* are two important variables in a credit application and self-employment receives a substantial negative weight. But as might be expected, the single most significant influence on cardholder status is major derogatory reports. Because lenders are strongly focused on default probability, past evidence of default behavior will be a major consideration.

17.6 TREATMENT EFFECTS AND ENDOGENOUS VARIABLES IN BINARY CHOICE MODELS

Consider the binary choice model with endogenous right-hand side variable, T,

$$y* = \mathbf{x}'\boldsymbol{\beta} + T\gamma + \varepsilon, y = \mathbf{1}(y* > 0), \text{Cov}(T, \varepsilon) \neq 0.$$

We examine the two leading cases:

1. T is an endogenous dummy variable that indicates some kind of treatment or program participation such as graduating from high school or college, receiving some kind of job training, purchasing health insurance, etc.[45]
2. T is an endogenous continuous variable. Because the model is not linear, conventional instrumental variable estimators such as two-stage least squares (2SLS) are not appropriate. We consider the alternative estimators based on the maximum likelihood estimator.

[45]Discussion appears in Angrist (2001) and Angrist and Pischke (2009, 2010).

17.6.1 ENDOGENOUS TREATMENT EFFECT

A structural model in which a treatment effect will be correlated with the unobservables is

$$T_i^* = \mathbf{z}_i'\boldsymbol{\alpha} + u_i,\ T_i = \mathbf{1}[T_i^* > 0],$$
$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + \varepsilon_i,\ y_i = \mathbf{1}[y_i^* > 0],$$
$$\begin{pmatrix}\varepsilon_i\\ u_i\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right].$$

The correlation between u and ε induces the endogeneity of T in the equation for y. We are interested in two effects: (1) the causal treatment effect of T on Prob(y $= 1|\mathbf{x}, T$), and (2) the partial effects of $\mathbf{x}$ and $\mathbf{z}$ on Prob($\mathbf{y} = 1|\mathbf{x}, \mathbf{z}, T$) in the presence of the endogenous treatment.

This **recursive model** is a bivariate probit model (Section 17.9.5). The log likelihood is constructed from the joint probabilities of the observed outcomes. The four possible outcomes and associated probabilities are obtained as the marginal probabilities for T times the conditional probabilities for $y|T$. Thus, $P(y = 1, T = 1) = P(y = 1|T = 1)P(T = 1)$. The marginal probability for $T = 1$ is just $\Phi(\mathbf{z}_i'\boldsymbol{\alpha})$, whereas the conditional probability is the bivariate normal probability divided by the marginal, $\Phi_2(\mathbf{x}_i'\boldsymbol{\beta} + \gamma, \mathbf{z}_i'\boldsymbol{\alpha}, \rho)/\Phi(\mathbf{z}_i'\boldsymbol{\alpha})$. The product returns the bivariate normal probability. The other three terms in the log likelihood are derived similarly. The four terms are

$$P(y = 1, T = 1|\mathbf{x}, \mathbf{z}) = \Phi_2(\mathbf{x}'\boldsymbol{\beta} + \gamma, \mathbf{z}'\boldsymbol{\alpha}, \rho),$$
$$P(y = 1, T = 0|\mathbf{x}, \mathbf{z}) = \Phi_2(\mathbf{x}'\boldsymbol{\beta} + \gamma, -\mathbf{z}'\boldsymbol{\alpha}, -\rho),$$
$$P(y = 0, T = 1|\mathbf{x}, \mathbf{z}) = \Phi_2[-(\mathbf{x}'\boldsymbol{\beta} + \gamma), \mathbf{z}'\boldsymbol{\alpha}, -\rho],$$
$$P(y = 0, T = 0|\mathbf{x}, \mathbf{z}) = \Phi_2[-(\mathbf{x}'\boldsymbol{\beta} + \gamma), -\mathbf{z}'\boldsymbol{\alpha}, \rho).$$

The log likelihood is then

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \rho) = \sum_{i=1}^{n} \ln \text{Prob}(y = y_i, T = T_i|\mathbf{x}, \mathbf{z}).$$

Estimation is discussed in Section 17.9.5. The model looks like a conventional simultaneous-equations model; the difference arises from the nonlinear transformation of (y^*,T^*) that produces the observed (y,T). One implication is that whereas for identification of a linear model of this form, there would have to be at least one variable in $\mathbf{z}$ that is not in $\mathbf{x}$, that is not the case here. The model is identified partly through the nonlinearity of the functional form. (See the commentary in Example 17.18.)

The *treatment effect (TE)* is derived from the marginal distribution of *y,*

$$\begin{aligned}\text{TE} &= \text{Prob}(y = 1|\mathbf{x}, T = 1) - \text{Prob}(y = 1|\mathbf{x}, T = 0)\\ &= \Phi(\mathbf{x}'\boldsymbol{\beta} + \gamma) - \Phi(\mathbf{x}'\boldsymbol{\beta}).\end{aligned}$$

The *average treatment effect (ATE)*, will be estimated by averaging the estimates of TE over the sample observations. The *treatment effect on the treated (ATET)* would be based on the conditional probability, Prob($y = 1|T = 1$),

$$\text{TET} = \Phi\left[\frac{(\mathbf{x}'\boldsymbol{\beta} + \gamma) - \rho(\mathbf{z}'\boldsymbol{\alpha})}{\sqrt{1 - \rho^2}}\right] - \Phi\left[\frac{(\mathbf{x}'\boldsymbol{\beta}) - \rho(\mathbf{z}'\boldsymbol{\alpha})}{\sqrt{1 - \rho^2}}\right].$$

The ATET is computed by averaging this quantity over the sample observations for which $T_i = 1$.[46]

To compute the average partial effects for the exogenous variables, we will require

$$\begin{aligned}&\text{Prob}(y = 1|\mathbf{x}, \mathbf{z}, T = 0)\,\text{Prob}(T = 0|\mathbf{z}) + \\ &\text{Prob}(y = 1|\mathbf{x}, \mathbf{z}, T = 1)\,\text{Prob}(T = 1|\mathbf{z}) \\ &\quad = \Phi_2(\mathbf{x}'\boldsymbol{\beta} + \gamma, \mathbf{z}'\boldsymbol{\alpha}, \rho) + \Phi_2(\mathbf{x}'\boldsymbol{\beta}, -\mathbf{z}'\boldsymbol{\alpha}, -\rho\end{aligned}$$

The partial effects for $\mathbf{x}$ and $\mathbf{z}$ are then

$$\frac{\partial\,\text{Prob}(y = 1|\mathbf{x}, \mathbf{z})}{\partial\begin{pmatrix}\mathbf{x}\\ \mathbf{z}\end{pmatrix}} = \frac{\partial[\Phi_2(\mathbf{x}'\boldsymbol{\beta} + \gamma, \mathbf{z}'\boldsymbol{\alpha}, \rho) + \Phi_2(\mathbf{x}'\boldsymbol{\beta}, -\mathbf{z}'\boldsymbol{\alpha}, -\rho)]}{\partial\begin{pmatrix}\mathbf{x}\\ \mathbf{z}\end{pmatrix}}.$$

Expressions for the derivatives appear in Section 17.9. This is a fairly intricate calculation. It is automated or conveniently computed in contemporary software, however. We can interpret $\partial\text{Prob}(y = 1|\mathbf{x}, \mathbf{z})/\partial\mathbf{x}$ as a *direct effect* and $\partial\text{Prob}(\text{y} = 1|\mathbf{x}, \mathbf{z})/\partial\mathbf{z}$ as an indirect effect on y that is transmitted through T. For variables that appear in both $\mathbf{x}$ and $\mathbf{z}$, the total effect is the sum of the two. The computations are illustrated in Example 17.19 below.

Example 17.18 An Incentive Program for Quality Medical Care

Scott, Schurer, Jensen, and Sivey (2009) examined an incentive program for Australian general practitioners to provide high quality care in diabetes management. The specific outcome of interest is ordering HbA1c tests as part of a diabetes consultation. The treatment of interest is participation in the incentive program.

A pay-for-performance program, the Practice Incentive Program (PIP) was superimposed on the Australian fee for service system in 1999 to encourage higher quality of care in chronic diseases including diabetes. Program participation by general practitioners (GPs) was voluntary. The quality of care outcome is whether the HbA1c test is administered. Analysis is conducted with a unique data set on GP consultations. The authors compare the average proportion of HbA1c tests ordered by GPs who have joined the incentive scheme with the average proportion of tests ordered by GPs who have not joined, while controlling for key sources of unobserved heterogeneity. A key assumption here is that HbA1c tests are undersupplied in the absence of the PIP scheme and therefore more frequent HbA1c testing is related to higher quality management. The endogenous nature of general practitioners' participation in the PIP is addressed by applying a bivariate probit model, using exclusion restrictions to aid identification of the causal parameters.

The GP will join the PIP if the utility from joining is positive. Utility depends on the additional income from joining the PIP, from the diabetes sign-on payment and negatively on the costs of accreditation and establishing the requisite IT systems. GPs will increase quality of care if the utility of doing so is positive, which partly depends on PIP membership. The bivariate probit model used is

$$\begin{aligned}Y_{ij}* &= \alpha_1 + \boldsymbol{\beta}_1'\mathbf{X}_{ij} + \beta_{PIP}\,PIP_{ij} + u_{1ij}\\ PIP_{ij}* &= \alpha_2 + \boldsymbol{\beta}_2'\mathbf{X}_{\text{ij}} + \boldsymbol{\pi}'\mathbf{I}_{ij} + u_{2ij},\end{aligned}$$

where $Y_{ij} = \mathbf{1}$ (GP j ordered an HbA1c test in recorded consultation i),
and $PIP_{ij} = \mathbf{1}$ (Practice in which GP_j works has joined the *PIP* program).

[46]See Jones (2007).

The authors calculate the marginal treatment effect of PIP using $ME_{PIP} = \beta_{PIP}\,\phi(\hat{\boldsymbol{\beta}}_1'\bar{\mathbf{x}})$.[47] Regarding the specification, they note "[a]lthough the model is formally identified by its nonlinear functional form, as long as the full rank condition of the data matrix is ensured (Heckman, 1978; Wilde, 2000), we introduce exclusion restrictions to aid identification of the causal parameter β_{PIP} (Maddala, 1983); Monfardini and Radice (2008). The row vector $\mathbf{I}_{ij}$ captures the variables in the PIP participation equation (5) but excluded from the outcome equation (4)."

Marginal effects for PIP status are reported (in Table II) for two treatment groups. For the first group, the estimated effect is roughly 0.2. In year 1 of the data set, before the PIP was introduced, the average proportion of HbA1c tests conducted was 13%. After the reform was introduced, the average diabetes patient therefore faced a probability of 32% of receiving an HbA1c test during an average encounter in a practice that has joined the PIP. The result from a univariate probit model that treated PIP as exogenous produced a corresponding value of only 0.028.

Example 17.19 Moral Hazard In German Health Care

Riphahn, Wambach, and Million (2003) examined health care utilization in a panel data set of German households. The main objective of the study was to consider evidence of moral hazard. The authors considered the joint determination of hospital and doctor visits in a bivariate count data model. The model assessed whether purchase of Add-on insurance was associated with heavier use of the health care system. All German households have some form of health insurance. In our data, roughly 89% have the compulsory public form. Some households, typically higher income, can opt, instead, for private insurance. The "Add-on" insurance, that is available to those who have the compulsory public insurance, provides coverage for additional benefits, such as certain prevention programs and additional dental coverage. We will construct a small model to suggest the computations of treatment effects in a recursive bivariate probit model. The structure for one of the two count variables is

$$Hospital* = \beta_1 + \beta_2\, Age + \beta_3\, Working + \beta_4\, Health + \gamma\, Addon + \varepsilon,$$
$$Addon* = \alpha_1 + \alpha_2\, Age + \alpha_3\, Education + \alpha_4\, Income + \alpha_5\, Married + \alpha_6\, Kids + \alpha_7\, Health + u.$$

Hospital is constructed as **1**(*Hospital Visits* $>$ 0) while *Add-On* $=$ **1**(*Household has Add-On Insurance*). Estimation is based, once again, on the 1994 wave of the data.

Estimation results are shown in Table 17.14. We find that the only significant determinant of hospital visitation is *Health* (measured as self-reported Health Satisfaction). The crucial parameter is γ, the coefficient on *Add-On*. The value of 0.04131 for APE(Add-On) is the estimated average treatment effect. We find, as did Riphahn, that the data do not appear to support the hypothesis of moral hazard. The *t* ratio on *Add-On* in the regression is only 0.16, far from significant. On the other hand, the estimated value, 0.04131, is not trivial. The mean value of *Hospital* is 0.091; 9.1% of this sample had at least one hospital visit in 1994. On average, if the subgroup of *Add-On* policy holders visited the hospital with 0.04 greater probability, this represents, using 0.091 as the base, an increase of 44% in the rate. That is actually quite large. For comparison purposes, the 2SLS estimates of this model are shown in the last column. (The authors of the application in Example 17.6 used 2SLS for estimation of their recursive bivariate probit model.) As might be expected, the 2SLS estimates provide a good approximation to the average partial effects of the exogenous variables. However, it produces an estimate for the causal *Add-On* effect that is three times as large as the FIML estimate, and has the wrong sign.

[47]The calculation of ME_{PIP} treats *PIP* as if it were continuous and differentiates the probability. This approximates $\Phi(\hat{\boldsymbol{\beta}}_1'\bar{\mathrm{x}} + \beta_{\mathrm{PIP}}) - \Phi(\hat{\boldsymbol{\beta}}_1'\bar{\mathrm{x}})$ as suggested earlier. The authors note: "An alternative is to calculate the difference in the probabilities of an HbA1c test in a consultation in which the practice participates in the PIP, and a practice that does not. Our method assumes the treatment indicator to be continuous to be able to use the delta method. We compared the two methods and the magnitude of the marginal effect is the same." (There is, in fact, no obstacle to using the delta method for the difference in the probabilities. See equation (17-29).) The authors computed the TE at the means of the data rather than averaging the TE values over the observations.

TABLE 17.14 Estimates of Recursive Bivariate Probit Model

	Add-On			*Hospital*				
Variable	***Estimate***	***Std. Error***	***t Ratio***	***Estimate***	***Std. Error***	***t Ratio***	***APE***	***2SLS***
Constant	−3.64543	0.42225	−8.63	−0.56009	0.18342	−3.05		0.24352
Health	0.00452	0.02552	0.18	−0.14258	0.01412	−10.10	−0.02195	−0.02505
Working				0.00728	0.07223	0.10	0.00112	0.00121
Add-On				0.23389	1.43618	0.16	0.04131	−0.11826
Age	0.00884	0.00568	1.56	0.00210	0.00292	0.72	0.00034*	0.00035
Education	0.07896	0.02030	3.89					
Income	0.48428	0.23142	2.09					
Married	−0.09885	0.13584	−0.73					
Kids	0.21025	0.13142	1.60					
ρ				−0.01363	0.60432	−0.02		
Log likelihood function	−1296.40433							
Estimation based on $N = 3377$, $K = 13$								

*Average Treatment Effect. Estimated ATET is 0.03861

17.6.2 ENDOGENOUS CONTINUOUS VARIABLE

If the endogenous variable in the recursive model is continuous, the structure is

$$T_i = \mathbf{z}_i'\boldsymbol{\alpha} + u_i,$$
$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + \varepsilon_i,\ y_i = \mathbf{1}[y_i^* > 0],$$
$$\begin{pmatrix}\varepsilon_i \\ u_i\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\sigma_u \\ \rho\sigma_u & \sigma_u^2\end{pmatrix}\right].$$

In the model for labor force participation in Example 17.15, Family income is endogenous.

17.6.2.a IV and GMM Estimation

The instrumental variable estimator described in Chapter 8 is based on moments of the data, variances, and covariances. In this binary choice setting, we are not using any form of least squares to estimate the parameters, so the IV method would appear not to apply. Generalized method of moments is a possibility. Starting from

$$E[\varepsilon_i | \mathbf{z}_i, \mathbf{x}_i] = 0,$$
$$E[T_i\mathbf{z}_i] \neq \mathbf{0},$$

a natural instrumental variable estimator would be based on the moment condition,

$$E\left[(y_i^* - \mathbf{x}_i'\boldsymbol{\beta} - \gamma T_i)\begin{pmatrix}\mathbf{x}_i \\ \mathbf{z}_i^*\end{pmatrix}\right] = \mathbf{0}.$$

(In this formulation, $\mathbf{z}_i*$ would contain only the variables in $\mathbf{z}_i$ not also contained in $\mathbf{x}$.) However, y_i^* is not observed, y_i is. The approach that was used in Avery et al. (1983), Butler and Chatterjee (1997), and Bertschek and Lechner (1998) is to assume that the instrumental variables are orthogonal to the residual, $[y - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i)]$; that is,

$$E\left[[y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i)]\begin{pmatrix}\mathbf{x}_i \\ \mathbf{z}_i^*\end{pmatrix}\right] = \mathbf{0}.$$

This form of the moment equation, based on observables, can form the basis of a straightforward two-step GMM estimator. (See Chapter 13 for details.)

17.6.2.b Partial ML Estimation

Simple probit estimation based on y_i and $(\mathbf{x}_i, T_i)$ will not consistently estimate $(\boldsymbol{\beta}, \gamma)$ because of the correlation between T_i and ε_i induced by the correlation between u_i and ε_i. The maximum likelihood estimator is based on the full specification of the model, including the bivariate normality assumption that underlies the endogeneity of T. One possibility is to use the partial reduced form obtained by inserting the first equation in the second. This becomes a probit model with probability $\text{Prob}(y_i = 1|\mathbf{x}_i, \mathbf{z}_i) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}* + \mathbf{z}_i^{*\prime}\boldsymbol{\alpha}^*)$. This will produce a consistent estimator of $\boldsymbol{\beta}^* = \boldsymbol{\beta}/(1 + \gamma^2\sigma_u^2 + 2\gamma\sigma_u\rho)^{1/2}$ and $\boldsymbol{\alpha}^* = \gamma\boldsymbol{\alpha}/(1 + \gamma^2\sigma_u^2 + 2\gamma\sigma_u\rho)^{1/2}$ as the coefficients on $\mathbf{x}_i$ and $\mathbf{z}_i$, respectively. (The procedure would estimate a mixture of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ for any variable that appears in both $\mathbf{x}_i$ and $\mathbf{z}_i$.) Newey (1987) suggested a minimum chi-squared estimator that does estimate all parameters. Linear regression of T_i on $\mathbf{z}_i$ produces estimates of $\boldsymbol{\alpha}$ and σ_u^2, which suggests a third possible estimator, based on a two-step MLE. But there is no method of moments estimator of ρ or γ produced by this procedure, so this estimator is incomplete.

17.6.2.c Full Information Maximum Likelihood Estimation

A more direct and actually simpler approach is full information maximum likelihood. The log likelihood is built up from the joint density of y_i and T_i, which we write as the product of the conditional and the marginal densities,

$$f(y_i, T_i) = f(y_i|T_i)f(T_i).$$

To derive the conditional distribution, we use results for the bivariate normal, and write

$$\varepsilon_i|u_i = [(\rho\sigma_u)/\sigma_u^2]u_i + v_i,$$

where v_i is normally distributed with $\text{Var}[v_i] = (1 - \rho^2)$. Inserting this in the second equation, we have

$$y_i^*|T_i = \mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)u_i + v_i.$$

Therefore,

$$\text{Prob}[y_i = 1|\mathbf{x}_i, T_i] = \Phi\left[\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)u_i}{\sqrt{1 - \rho^2}}\right]. \tag{17-36}$$

Inserting the expression for $u_i = (T_i - \mathbf{z}_i'\boldsymbol{\alpha})$, and using the normal density for the marginal distribution of T_i in the first equation, we obtain the log-likelihood function for the sample,

$$\ln L = \sum_{i=1}^{n}\left\{\ln\Phi\left[(2y_i - 1)\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)(T_i - \mathbf{z}_i'\boldsymbol{\alpha})}{\sqrt{1 - \rho^2}}\right)\right] + \ln\left[\frac{1}{\sigma_u}\phi\left(\frac{T_i - \mathbf{z}_i'\boldsymbol{\alpha}}{\sigma_u}\right)\right]\right\}. \tag{17-37}$$

Some convenience can be obtained by rewriting the log-likelihood function as

$$\ln L = \sum_{i=1}^{n}\ln\Phi[(2y_i - 1)(\mathbf{x}_i'\widetilde{\boldsymbol{\beta}} + \widetilde{\gamma}T_i + \tau[(T_i - \mathbf{z}_i'\boldsymbol{\alpha})/\sigma_u]] + \sum_{i=1}^{n}\ln\left[\frac{1}{\sigma_u}\phi[(T_i - \mathbf{z}_i'\boldsymbol{\alpha})/\sigma_u]\right],$$

where $\widetilde{\boldsymbol{\beta}} = \left(1/\sqrt{1-\rho^2}\right)\boldsymbol{\beta}$, $\widetilde{\gamma} = \left(1/\sqrt{1-\rho^2}\right)\gamma$ and $\tau = \left(\rho/\sqrt{1-\rho^2}\right)$. The delta method can be used to recover the original parameters and appropriate standard errors after estimation.[48]

Partial effects are derived from the first term in (17-37),

$$\frac{\partial \text{Prob}(y=1|\mathbf{x},T,\mathbf{z})}{\partial\begin{pmatrix}\mathbf{x}\\ T\\ \mathbf{z}\end{pmatrix}} = \frac{\partial\Phi\left(\dfrac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)(T_i - \mathbf{z}_i'\boldsymbol{\alpha})}{\sqrt{1-\rho^2}}\right)}{\partial\begin{pmatrix}\mathbf{x}\\ T\\ \mathbf{z}\end{pmatrix}}$$

$$= \phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)(T_i - \mathbf{z}_i'\boldsymbol{\alpha})}{\sqrt{1-\rho^2}}\right)\frac{1}{\sqrt{1-\rho^2}}\begin{pmatrix}\boldsymbol{\beta}\\ \gamma + \rho/\sigma_u\\ -(\rho/\sigma_u)\boldsymbol{\alpha}\end{pmatrix}.$$

17.6.2.d Residual Inclusion and Control Functions

A further simplification of the log-likelihood function is obtained by writing

$$\ln L = \sum_{i=1}^{n}\ln\Phi[(2y_i-1)(\mathbf{x}_i'\widetilde{\boldsymbol{\beta}} + \widetilde{\gamma}T_i + \tau\widetilde{u}_i] + \sum_{i=1}^{n}\ln\left[\frac{1}{\sigma_u}\phi(\widetilde{u}_i)\right],$$

$\widetilde{u}_i = (T_i - \mathbf{z}_i'\boldsymbol{\alpha})/\sigma_u$. This "residual inclusion" form suggests a two-step approach. The parameters in the linear regression, $\boldsymbol{\alpha}$ and σ_u, can be consistently estimated by a linear regression of T on $\mathbf{z}$. The scaled residual $\widetilde{u}_i = (T_i - \mathbf{z}_i'\mathbf{a})/s_u$ can now be computed and inserted into the log likelihood. Note that the second term in the log likelihood involves parameters that have already been estimated at the first step, so it can be ignored. The second-step log likelihood is, then,

$$\ln L = \sum_{i=1}^{n}\ln\Phi[(2y_i-1)(\mathbf{x}_i'\widetilde{\boldsymbol{\beta}} + \widetilde{\gamma}w_i + \tau\hat{\widetilde{u}}_i)].$$

This can be maximized using the methods developed in Section 17.3. The estimator of ρ can be recovered from $\rho = \tau/(1+\tau^2)^{1/2}$. Estimators of $\boldsymbol{\beta}$ and γ follow, and the delta method can be used to construct standard errors. Because this is a two-step estimator, the resulting estimator of the asymptotic covariance matrix would be adjusted using the Murphy and Topel (2002) results in Section 14.7. Bootstrapping the entire apparatus (i.e., both steps—see Section 15.4) would be an alternative way to estimate an asymptotic covariance matrix. The original (one-step) log likelihood is not very complicated, and full information estimation is fairly straightforward. The preceding demonstrates how the alternative two-step method would proceed and suggests how the residual inclusion method proceeds. The general approach of residual inclusion for nonlinear models with endogenous variables is explored in detail by Terza, Basu, and Rathouz (2008).

17.6.2.e A Control Function Estimator

In the residual inclusion estimator noted earlier the endogeneity of T in the probit model is mitigated by adding the estimated residual to the equation—in the presence

[48]Recent applications of this estimator have referred to it as *instrumental variable probit* estimation. The estimator is a full information maximum likelihood estimator.

of the residual, T is no longer correlated with ε. We took this approach in estimating a linear model in Section 8.4.2. Blundell and Powell (2004) label the foregoing the **control function** approach to accommodating the endogeneity. The residual inclusion estimator suggested here was proposed by Rivers and Vuong (1988). As noted, the estimator is fully parametric. They propose an alternative semiparametric approach that retains much of the functional form specification, but works around the specific distributional assumptions. Adapting their model to our earlier notation, their departure point is a general specification that produces, once again, a control function,

$$E[y_i|\mathbf{x}_i, T_i, u_i] = F(\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i, u_i).$$

Note that (17-36) satisfies the assumption; however, they reach this point without assuming either joint or marginal normality. The authors propose a three-step, semiparametric approach to estimating the structural parameters. In an application somewhat similar to Example 17.8, they apply the technique to a labor force participation model for British men in which a variable of interest is a dummy variable for education greater than 16 years, the endogenous variable in the participation equation, also of interest, is earned income of the spouse, and an instrumental variable is a welfare benefit entitlement. Their findings are rather more substantial than ours; they find that when the endogeneity of other family income is accommodated in the equation, the education coefficient increases by 40% and remains significant, but the coefficient on other income increases by more than tenfold.

Example 17.20 Labor Supply Model

In Examples 5.2, 17.1, and 17.15, we examined a labor supply model for married women using Mroz's (1987) data on labor supply. The wife's labor force participation equation suggested in Example 17.15 is

$$\text{Prob}[LFP = 1] = F(\textit{Constant}, \textit{Other Income}, \textit{Education}, \textit{Experience}, \textit{Experience}^2, \textit{Age}, \textit{Kids Under 6}, \textit{Kids 6 to 18}).$$

The *Other Income* (non-wife's) would likely be jointly determined with the LFP decision. We model this with

$$\textit{Other Income} = \alpha_1 + \alpha_2\,\textit{Husband's Age} + \alpha_3\,\textit{Husband's Education} + \alpha_4\,\textit{City} + \alpha_5\,\textit{Kids Under 6} + \alpha_6\,\textit{Kids 6 to 18} + u.$$

As before, we use the Mroz (1987) labor supply data described in Example 5.2. Table 17.15 reports the naïve single-equation and full information maximum likelihood estimates of the parameters of the two equations. The third set of results is the two-step estimator detailed in Section 17.6.2d. Standard errors for the maximum likelihood estimators are based on the derivatives of the log-likelihood function. Standard errors for the two-step estimator are computed using 50 bootstrap replications. (Both steps are computed for the bootstrap replications.)

Comparing the two sets of probit estimates, it appears that the (assumed) endogeneity of the *Other Income* is not substantially affecting the estimates. The results are nearly the same. There are two simple ways to test the hypothesis that ρ equals zero. The FIML estimator produces an estimated asymptotic standard error with the estimate of ρ, so a Wald test can be carried out. For the preceding results, the Wald statistic would be $(0.18777/0.13625)^2 = 1.378^2 = 1.899$. The critical value from the chi-squared table for one degree of freedom would be 3.84, so we would not reject the hypothesis of exogeneity. The second approach would use the likelihood ratio test. Under the null hypothesis of exogeneity, the probit model and the regression equation can be estimated independently. The log likelihood for the full model would be the sum of the two log likelihoods, which would be $-401.30 + (-2{,}844.103) = -3{,}245.405$. The

TABLE 17.15 Estimated Labor Supply Model

	Probit		*FIML*			*2-Step Control Function*	
Variable	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*	*APE*	*Estimate*	*Std. Err.*
LFP Equation for Wife							
Constant	0.27008	0.50859	0.21277	0.51736		0.21811	0.50719
Education	0.13090	0.02525	0.14571	0.02689	0.05693	0.14816	0.02900
Experience	0.12335	0.01872	0.12299	0.01851	0.04805	0.12521	0.01868
Experience^2	−0.00189	0.00060	−0.00192	0.00060	−0.00075	−0.00196	0.00053
Age	−0.05285	0.00848	−0.04878	0.00951	−0.01906	−0.04970	0.00914
Kids Under 6	−0.86833	0.11852	−0.83049	0.12684	−0.32447	−0.84568	0.13693
Kids 6–18	0.03600	0.04348	0.04781	0.04214	0.01868	0.04855	0.05240
Non-wife Inc.	−0.01202	0.00484	−0.02761	0.01254	−0.01079	−0.02798	0.01500
Residual						0.01795	0.01572
Non-wife Income Equation							
Constant			−10.6816	4.34481		−10.5492	
Hus. Age			0.23009	0.07089		0.22818	
Hus. Education			1.35361	0.12978		1.34613	
City			3.54202	0.91338		3.62319	
Kids Under 6			1.36755	0.67056		1.36403	
Kids 6–18			0.67856	0.36160		0.67573	
σ			10.5708	0.15966		10.61312	
ρ			0.18777	0.13625			
ln L	−401.302		−3244.556			−2844.103	

log likelihood for the combined model is −3,244.556. Twice the difference is 0.849, which is also well under the 3.84 critical value, so on this basis as well, we would not reject the null hypothesis that $\rho = 0$. As would now be expected, the three sets of estimates are nearly the same. The estimate of −0.02761 for the coefficient on *Other Income* implies that a $1,000 increase reduces the LFP by about 0.028. Because the participation rate is about 0.57, the $1,000 increase suggests a reduction in participation of about 4.9%. The mean value of other income is roughly $20,000, so the 5% increase in *Other Income* is associated with a 5% decrease in LFP, or an elasticity of about one.

17.6.3 ENDOGENOUS SAMPLING

We have encountered several instances of nonrandom sampling in the binary choice setting. In Example 17.17, we examined an application in credit scoring in which the balance in the sample of responses of the outcome variable, $C = 1$ for acceptance of an application and $C = 0$ for rejection, is different from the known proportions in the population. The sample was skewed in favor of observations with $C = 1$ to enrich the data set. A second type of nonrandom sampling arises in the analysis of nonresponse/attrition in the GSOEP in Example 17.29 below. Here, the observed sample is not random with respect to individuals' presence in the sample at different waves of the panel. The

first of these represents selection specifically on an observable outcome—the observed dependent variable. We construct a model for the second of these that relies on an assumption of selection on a set of certain observables—the variables that enter the probability weights. We will now examine a third form of nonrandom sample selection, based crucially on the *unobservables* in the two equations of a bivariate probit model.

We return to the banking application of Example 17.17. In that application, we examined a binary choice model,

$$\begin{aligned}\text{Prob}(\textit{Cardholder} = 1|\mathbf{x}) &= \text{Prob}(C = 1|\mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2\, \textit{Age} + \beta_3\, \textit{Income} + \beta_4\, \textit{OwnRent} \\ &\quad + \beta_5\, \textit{Months at Current Address} \\ &\quad + \beta_6\, \textit{Self-Employed} \\ &\quad + \beta_7\, \textit{Number of Major Derogatory Reports} \\ &\quad + \beta_8\, \textit{Number of Minor Derogatory Reports}).\end{aligned}$$

From the point of view of the lender, cardholder status is not the interesting outcome in the credit history, default is. The more interesting equation describes $\text{Prob}(\textit{Default} = 1|\mathbf{z}, C = 1)$. The natural approach, then, would be to construct a binary choice model for the interesting default variable using the historical data for a sample of cardholders. The problem with the approach is that the sample of cardholders is not randomly drawn from the full population—applicants are screened with an eye specifically toward whether or not they seem likely to default. In this application, and in general, there are three economic agents, the credit scorer (e.g., Fair Isaacs), the lender, and the borrower. Each of them has latent characteristics in the equations that determine their behavior. It is these latent characteristics that drive, in part, the application/scoring process and, ultimately, the consumer behavior.

A model that can accommodate these features is

$$\begin{aligned}&S^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad S = \mathbf{1}(S^* > 0), \\ &y^* = \mathbf{x}_1'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y = \mathbf{1}(y^* > 0), \\ &\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix}\Big|\,\mathbf{x}_1, \mathbf{x}_2 \Bigg) \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right], \\ &(y, \mathbf{x}_2) \text{ observed only when } S = 1,\end{aligned}$$

which contains an observation rule, $S = 1$, and a behavioral outcome, $y = 0$ or 1. The endogeneity of the sampling rule implies that

$$\text{Prob}(y = 1|S = 1, \mathbf{x}_2) \neq \Phi(\mathbf{x}_2'\boldsymbol{\beta}).$$

From properties of the bivariate normal distribution, the appropriate probability is

$$\text{Prob}(y = 1|S = 1, \mathbf{x}_1, \mathbf{x}_2) = \Phi\left[\frac{\mathbf{x}_2'\boldsymbol{\beta}_2 + \rho\mathbf{x}_1'\boldsymbol{\beta}_1}{\sqrt{1 - \rho^2}}\right].$$

If ρ is not zero, then in using the simple univariate probit model, we are omitting from our model any variables that are in $\mathbf{x}_1$ but not in $\mathbf{x}_2$, and in any case, the estimator is inconsistent by a factor $(1 - \rho^2)^{-1/2}$. To underscore the source of the bias, if ρ equals

zero, the conditional probability returns to the model that would be estimated with the selected sample. Thus, the bias arises because of the correlation of (i.e., the selection on) the unobservables, ε_1 and ε_2. This model was employed by Wynand and van Praag (1981) in the first application of Heckman's (1979) sample selection model in a nonlinear setting to insurance purchases by Boyes, Hoffman, and Lowe (1989) in a study of bank lending by Greene (1992) to the credit card application begun in Example 17.17 and continued in Example 17.21 and hundreds of applications since.

Given that the forms of the probabilities are known, the appropriate log-likelihood function for estimation of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and ρ is easily obtained. The log likelihood must be constructed for the joint or the marginal probabilities, not the conditional ones. For the selected observations, that is, $(y = 0, S = 1)$ or $(y = 1, S = 1)$, the relevant probability is simply

$$\text{Prob}(y = 0 \text{ or } 1 \,|\, S = 1) \times \text{Prob}(S = 1) = \Phi_2[(2y - 1)\mathbf{x}_2'\boldsymbol{\beta}_2, \mathbf{x}_1'\boldsymbol{\beta}_1, (2y - 1)\rho].$$

For the observations with $S = 0$, the probability that enters the likelihood function is simply $\text{Prob}(S = 0 \,|\, \mathbf{x}_1) = \Phi(-\mathbf{x}_1'\boldsymbol{\beta}_1)$. Estimation is then based on a simpler form of the bivariate probit log likelihood that we examined in Section 17.6.1. Partial effects and post-estimation analysis would follow the analysis for the bivariate probit model. The desired partial effects would differ by the application, whether one desires the partial effects from the conditional, joint, or marginal probability would vary. The necessary results are in Section 17.9.3.

Example 17.21 Cardholder Status and Default Behavior

In Example 17.9, we estimated a logit model for cardholder status,

$$\begin{aligned}\text{Prob}(\textit{Cardholder} = 1) &= \text{Prob}(C = 1 \,|\, \mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2 \textit{Age} + \beta_3 \textit{Income} + \beta_4 \textit{OwnRent} \\ &\quad + \beta_5 \textit{ Current Address} + \beta_6 \textit{ SelfEmployed} \\ &\quad + \beta_7 \textit{ Major Derogatory Reports} \\ &\quad + \beta_8 \textit{ Minor Derogatory Reports}),\end{aligned}$$

using a sample of 13,444 applications for a credit card. The complication in that example was that the sample was choice based. In the data set, 78.1% of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2%, so the sample is substantially choice based on this variable. The sample was deliberately skewed in favor of cardholders for purposes of the original study.[49] The weights to be applied for the WESML estimator are 0.232/0.781 = 0.297 for the observations with $C = 1$ and 0.768/0.219 = 3.507 for observations with $C = 0$. Of the 13,444 applicants in the sample, 10,499 were accepted (given the credit cards). The "default rate" in the sample is 996/10,499 or 9.48%. This is slightly less than the population rate at the time, 10.3%. For purposes of a less complicated numerical example, we will ignore the choice-based sampling nature of the data set for the present. An orthodox treatment of both the selection issue and the choice-based sampling treatment is left for the exercises [and pursued in Greene (1992).]

We have formulated the cardholder equation so that it probably resembles the policy of credit scorers, both then and now. A major derogatory report results when a credit account that is being monitored by the credit reporting agency is more than 60 days late in payment. A minor derogatory report is generated when an account is 30 days delinquent. Derogatory

[49] See Greene (1992).

TABLE 17.16 Estimated Joint Cardholder and Default Probability Models

	Endogenous Sample Model			*Uncorrelated Equations*		
Variable/Equation	***Estimate***	***Std. Error***	***(t)***	***Estimate***	***Std. Error***	
			Cardholder Equation			
Constant	0.30516	0.04781	(6.38)	0.31783	0.04790	(6.63)
Age	0.00226	0.00145	(1.56)	0.00184	0.00146	(1.26)
Current Address	0.00091	0.00024	(3.80)	0.00095	0.00024	(3.94)
OwnRent	0.18758	0.03030	(6.19)	0.18233	0.03048	(5.98)
Income	0.02231	0.00093	(23.87)	0.02237	0.00093	(23.95)
SelfEmployed	−0.43015	0.05357	(−8.03)	−0.43625	0.05413	(−8.06)
Major Derogatory	−0.69598	0.01871	(−37.20)	−0.69912	0.01839	(−38.01)
Minor Derogatory	−0.04717	0.01825	(−2.58)	−0.04126	0.01829	(−2.26)
			Default Equation			
Constant	−0.96043	0.04728	(−20.32)	−0.81528	0.04104	(−19.86)
Dependents	−0.04995	0.01415	(3.53)	0.04993	0.01442	(3.46)
Income	−0.01642	0.00122	(−13.41)	−0.01837	0.00119	(−15.41)
Expend/Income	−0.16918	0.14474	(−1.17)	−0.14172	0.14913	(−0.95)
Correlation	0.41947	0.11762	(3.57)	0.00000		
Log Likelihood		−8,660.90650			−8,670.78831	

reports are a major contributor to credit decisions. Contemporary credit processors such as Fair Isaacs place extremely heavy weight on the "credit score," a single variable that summarizes the credit history and credit-carrying capacity of an individual. We did not have access to credit scores at the time of this study. The selection equation was given earlier. The default equation is a behavioral model. There is no obvious standard for this part of the model. We have used three variables, *Dependents*, the number of dependents in the household, *Income*, and *Exp_Income*, which equals the ratio of the average credit card expenditure in the 12 months after the credit card was issued to average monthly income. Default status is measured for the first 12 months after the credit card was issued.

Estimation results are presented in Table 17.16. These are broadly consistent with the earlier results—the models with no correlation from Example 17.9 are repeated in Table 17.16. There are two tests we can employ for endogeneity of the selection. The estimate of ρ is 0.41947 with a standard error of 0.11762. The t ratio for the test that ρ equals zero is 3.57, by which we can reject the hypothesis. Alternatively, the likelihood ratio statistic based on the values in Table 17.16 is $2(8{,}670.78831 - 8{,}660.90650) = 19.76362$. This is larger than the critical value of 3.84, so the hypothesis of zero correlation is rejected. The results are as might be expected, with one counterintuitive result, that a larger credit burden, expenditure to income ratio, appears to be associated with lower default probabilities, though not significantly so.

17.7 PANEL DATA MODELS

Qualitative response models have been a growth industry in econometrics. The recent literature, particularly in the area of panel data analysis, has produced a number of new techniques. The availability of large, high-quality panel data sets on microeconomic

behavior has supported an interest in extending the models of Chapter 11 to binary (and other discrete) choice models. In this section, we will survey a few results from this rapidly growing literature.

The structural model for a possibly unbalanced panel of data would be

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$y_{it} = \mathbf{1}(y_{it}^* > 0). \tag{17-38}$$

Most of the interesting cases to be analyzed will start from our familiar common effects model,

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} + u_i, \quad i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,} \tag{17-39}$$

where, as before (see Sections 11.4 and 11.5), u_i is the unobserved, individual specific heterogeneity. Once again, we distinguish between *random* and *fixed* effects models by the relationship between u_i and $\mathbf{x}_{it}$. The assumption of *strict exogeneity*, that $f(u_i|\mathbf{X}_i)$ is not dependent on $\mathbf{X}_i$, produces the **random effects model**. Note that this places a restriction on the distribution of the heterogeneity. If that distribution is unrestricted, so that u_i and $\mathbf{x}_{it}$ may be correlated, then we have the **fixed effects model**. As before, the distinction does not relate to any intrinsic characteristic of the effect itself.

As we shall see shortly, this modeling framework is fraught with difficulties and unconventional estimation problems. Among them are the following: Estimation of the random effects model requires very strong assumptions about the heterogeneity; the fixed effects model relaxes these assumptions, but the natural estimator in this case encounters an **incidental parameters problem** that renders the maximum likelihood estimator inconsistent even when the model is correctly specified.

17.7.1 THE POOLED ESTIMATOR

To begin, it is useful to consider the pooled estimator that results if we simply ignore the heterogeneity, u_i, in (17-39) and fit the model as if the cross-section specification of Section 17.2.2 applies.[50] If the fixed effects model is appropriate, then results for omitted variables, including the Yatchew and Griliches (1984) result, apply. The pooled MLE that ignores fixed effects will be inconsistent—possibly wildly so. (*Note:* Because the estimator is ML, not least squares, converting the data to deviations from group means is not a solution—converting the binary dependent variable to deviations will produce a new variable with unknown properties.)

The random effects case is simpler. From (17-39), the marginal probability implied by the model is

$$\begin{aligned}\text{Prob}(y_{it} = 1|\mathbf{x}_{it}) &= \text{Prob}(v_{it} + u_i > -\mathbf{x}_{it}'\boldsymbol{\beta}) \\ &= F[\mathbf{x}_{it}'\boldsymbol{\beta}/(1 + \sigma_u^2)^{1/2}] \\ &= F(\mathbf{x}_{it}'\boldsymbol{\delta}).\end{aligned}$$

[50] We could begin the analysis by establishing the assumptions within which we can estimate the parameters of interest ($\boldsymbol{\beta}$) by treating the panel as a long cross section. The point of the exercise, however, is that those assumptions are unlikely to be met in any realistic application.

The implication is that based on the marginal distributions, we can consistently estimate $\boldsymbol{\delta}$ (but not $\boldsymbol{\beta}$ or σ_u separately) by pooled MLE.[51] This would be a pseudo MLE because the log-likelihood function is not the true log likelihood for the full set of observed data, but it is the correct product of the marginal distributions for $y_{it}|\mathbf{x}_{it}$. (This would be the binary choice case counterpart to consistent estimation of $\boldsymbol{\beta}$ in a linear random effects model by pooled ordinary least squares.) The implication, which is absent in the linear case, is that ignoring the random effects in a pooled model produces an attenuated (inconsistent—downward biased) estimate of $\boldsymbol{\beta}$; the scale factor that produces $\boldsymbol{\delta}$ is $1/(1 + \sigma_u^2)^{1/2}$, which is between zero and one. The implication for the partial effects is less clear. In the model specification, the partial effect is

$$PE(\mathbf{x}_{it}, u_i) = \partial \text{Prob}[y_{it} = 1|\mathbf{x}_{it}, u_i]/\partial \mathbf{x}_{it} = \boldsymbol{\beta} \times f(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i),$$

which is not computable. The useful result would be

$$E_u[PE(\mathbf{x}_{it}, u_i)] = \boldsymbol{\beta}\, E_u[f(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)].$$

Wooldridge (2010) shows that the end result, assuming normality of both v_{it} and u_i is $E_u[PE(\mathbf{x}_{it}, u_i)] = \boldsymbol{\delta}\phi(\mathbf{x}_{it}'\boldsymbol{\delta})$. Thus far, surprisingly, it would seem that simply pooling the data and using the simple MLE works. The estimated standard errors will be incorrect, so a correction such as the cluster estimator shown in Section 14.8.2 would be appropriate. Three considerations suggest that one might want to proceed to the full MLE in spite of these results: (1) The pooled estimator will be inefficient compared to the full MLE; (2) the pooled estimator does not produce an estimator of σ_u that might be of interest in its own right; and (3) the FIML estimator is available in contemporary software and is no more difficult to estimate than the pooled estimator. Note that the pooled estimator is not justified (over the FIML approach) on robustness considerations because the same normality and random effects assumptions that are needed to obtain the FIML estimator will be needed to obtain the preceding results for the pooled estimator.

17.7.2 RANDOM EFFECTS

A specification that has the same structure as the random effects model of Section 11.5 has been implemented by Butler and Moffitt (1982). We will sketch the derivation to suggest how random effects can be handled in discrete and limited dependent variable models such as this one. Full details on estimation and inference may be found in Butler and Moffitt (1982) and Greene (1995a). We will then examine some extensions of the Butler and Moffitt model.

The random effects model specifies

$$\varepsilon_{it} = v_{it} + u_i,$$

where v_{it} and u_i are independent random variables with

$$E[v_{it}|\mathbf{X}] = 0; \text{Cov}[v_{it}, v_{js}, |\mathbf{X}] = \text{Var}[v_{it}|\mathbf{X}] = 1, \quad \text{if } i = j \text{ and } t = s; 0 \text{ otherwise},$$

$$E[u_i|\mathbf{X}] = 0; \text{Cov}[u_i, u_j|\mathbf{X}] = \text{Var}[u_i|\mathbf{X}] = \sigma_u^2, \quad \text{if } i = j; 0 \text{ otherwise},$$

$$\text{Cov}[v_{it}, u_j|\mathbf{X}] = 0 \text{ for all } i, t, j,$$

[51] This result is explored at length in Wooldridge (2010).

and $\mathbf{X}$ indicates all the exogenous data in the sample, $\mathbf{x}_{it}$ for all i and t.[52] Then,

$$E[\varepsilon_{it}|\mathbf{X}] = 0,$$
$$\text{Var}[\varepsilon_{it}|\mathbf{X}] = \sigma_v^2 + \sigma_u^2 = 1 + \sigma_u^2,$$

and

$$\text{Corr}[\varepsilon_{it}, \varepsilon_{is}|\mathbf{X}] = \rho = \frac{\sigma_u^2}{1 + \sigma_u^2}.$$

The new free parameter is $\sigma_u^2 = \rho/(1 - \rho)$.

Recall that in the cross-section case, the marginal probability associated with an observation is

$$P(y_i|\mathbf{x}_i) = \int_{L_i}^{U_i} f(\varepsilon_i)d\varepsilon_i, (L_i, U_i) = (-\infty, -\mathbf{x}_i'\boldsymbol{\beta}) \quad \text{if } y_i = 0 \text{ and } (-\mathbf{x}_i'\boldsymbol{\beta}, +\infty) \quad \text{if } y_i = 1.$$

This simplifies to $\Phi[(2y_i - 1)\mathbf{x}_i'\boldsymbol{\beta}]$ for the normal distribution and $\Lambda[(2y_i - 1)\mathbf{x}_i'\boldsymbol{\beta}]$ for the logit model. In the fully general case with an unrestricted covariance matrix, the contribution of group i to the likelihood would be the joint probability for all T_i observations,

$$L_i = P(y_{i1}, \ldots, y_{iT_i}|\mathbf{X}) = \int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i})d\varepsilon_{i1}d\varepsilon_{i2} \ldots d\varepsilon_{iT_i}. \quad \textbf{(17-40)}$$

The integration of the joint density, as it stands, is impractical in most cases. The special nature of the random effects model allows a simplification, however. We can obtain the joint density of the v_{it}'s by integrating u_i out of the joint density of $(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}, u_i)$, which is

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}, u_i) = f(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}|u_i)f(u_i).$$

So,

$$f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) = \int_{-\infty}^{+\infty} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}|u_i)f(u_i)\, du_i.$$

The advantage of this form is that conditioned on u_i, the ε_{it}'s are independent, so

$$f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) = \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} f(\varepsilon_{it}|u_i)f(u_i)\, du_i.$$

Inserting this result in (17-40) produces

$$L_i = P(y_{i1}, \ldots, y_{iT_i}|\mathbf{X}) = \int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} \prod_{t=1}^{T_i} f(\varepsilon_{it}|u_i)f(u_i)\, du_i\, d\varepsilon_{i1}\, d\varepsilon_{i2} \ldots d\varepsilon_{iT_i}.$$

This may not look like much simplification, but in fact, it is. Because the ranges of integration are independent, we may change the order of integration:

$$L_i = P(y_{i1}, \ldots, y_{iT_i}|\mathbf{X}) = \int_{-\infty}^{+\infty} \left[\int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} \prod_{t=1}^{T_i} f(\varepsilon_{it}|u_i)\, d\varepsilon_{i1}\, d\varepsilon_{i2} \ldots d\varepsilon_{iT_i} \right] f(u_i)\, du_i.$$

[52] See Wooldridge (2010) for discussion of this strict exogeneity assumption.

Conditioned on the common u_i, the ε's are independent, so the term in square brackets is just the product of the individual probabilities. We can write this as

$$L_i = P(y_{i1}, \ldots, y_{iT_i}|\mathbf{X}) = \int_{-\infty}^{+\infty}\left[\prod_{t=1}^{T_i}\left(\int_{L_{it}}^{U_{it}} f(\varepsilon_{it}|u_i)d\varepsilon_{it}\right)\right]f(u_i)\,du_i. \quad \textbf{(17-41)}$$

Now, consider the individual densities in the product. Conditioned on u_i, these are the now-familiar probabilities for the individual observations, computed now at $\mathbf{x}_{it}'\boldsymbol{\beta} + u_i$. This produces a general form for random effects for the binary choice model. Collecting all the terms, we have reduced it to

$$L_i = P(y_{i1}, \ldots, y_{iT_i}|\mathbf{X}) = \int_{-\infty}^{+\infty}\left[\prod_{t=1}^{T_i}\text{Prob}(Y_{it} = y_{it}|\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)\right]f(u_i)\,du_i. \quad \textbf{(17-42)}$$

It remains to specify the distributions, but the important result thus far is that the entire computation requires only one-dimensional integration. The inner probabilities may be any of the models we have considered so far, such as probit, logit, Gumbel, and so on. The intricate part that remains is how to do the outer integration. **Butler and Moffitt's quadrature method** assuming that u_i is normally distributed is detailed in Section 14.14.4.

A number of authors have found the Butler and Moffitt formulation to be a satisfactory compromise between a fully unrestricted model and the cross-sectional variant that ignores the correlation altogether. An application that includes both group and time effects is Tauchen, Witte, and Griesinger's (1994) study of arrests and criminal behavior. The Butler and Moffitt approach has been criticized for the restriction of equal correlation across periods. But it does have a compelling virtue that the model can be efficiently estimated even with fairly large T_i, using conventional computational methods.[53]

A remaining problem with the Butler and Moffitt specification is its assumption of normality. In general, other distributions are problematic because of the difficulty of finding either a closed form for the integral or a satisfactory method of approximating the integral. An alternative approach that allows some flexibility is the method of **maximum simulated likelihood** (MSL), which was discussed in Section 15.6. The transformed likelihood we derived in (17-42) is an expectation,

$$\begin{aligned} L_i &= \int_{-\infty}^{+\infty}\left[\prod_{t=1}^{T_i}\text{Prob}(Y_{it} = y_{it}|\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)\right]f(u_i)\,du_i \\ &= E_{u_i}\left[\prod_{t=1}^{T_i}\text{Prob}(Y_{it} = y_{it}|\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)\right]. \end{aligned}$$

This expectation can be approximated by simulation rather than **quadrature**. First, let θ now denote the scale parameter in the distribution of u_i. This would be σ_u for a normal distribution, for example, or some other scaling for the logistic or uniform distribution. Then, write the term in the likelihood function as

$$L_i = E_{u_i}\left[\prod_{t=1}^{T_i}F(y_{it}, \mathbf{x}_{it}'\boldsymbol{\beta} + \theta u_i)\right] = E_u[h(u_i)].$$

Note that u_i is free of any unknown parameters. For example, for normally distributed u, by this transformation, θ is σ_u and now, $u \sim N[0, 1]$. The function is smooth, continuous, and

[53] See Greene (2007b).

continuously differentiable. If this expectation is finite, then the conditions of the law of large numbers should apply, which would mean that for a sample of observations $u_{i1}, \ldots, u_{iR}$,

$$\text{plim}\ \frac{1}{R}\sum_{r=1}^{R} h(u_{ir}) = E_u[h(u_i)].$$

This suggests, based on the results in Chapter 15, an alternative method of maximizing the log likelihood for the random effects model. A sample of person-specific draws from the population u_i can be generated with a random number generator. For the Butler and Moffitt model with normally distributed u_i, the simulated log-likelihood function is

$$\ln L_{Simulated} = \sum_{i=1}^{n} \ln\left\{\frac{1}{R}\sum_{r=1}^{R}\left[\prod_{t=1}^{T_i} F[(2y_{it}-1)(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u u_{ir})]\right]\right\}. \tag{17-43}$$

This function is maximized with respect to $\boldsymbol{\beta}$ and σ_u. Note that in the preceding, as in the quadrature approximated log likelihood, the model can be based on a probit, logit, or any other functional form desired.

For testing the hypothesis of the restricted, pooled model, a Lagrange multiplier approach that does not require estimation of the full random effects model will be attractive. Greene and McKenzie (2015) derived an LM test specifically for the random effects model. Let λ_{it} equal the derivative with respect to the constant term under H_0, defined in (17-20), and let $\tau_{it} = -(q_{it}\mathbf{x}_{it}'\boldsymbol{\beta})\lambda_{it} - \lambda_{it}^2$. Then,

$$\mathbf{g}_i = \begin{bmatrix} \displaystyle\sum_{t=1}^{T_i} \lambda_{it}\mathbf{x}_{it} \\ \displaystyle\frac{1}{2}\left(\sum_{t=1}^{T_i}\tau_{it}\right) + \frac{1}{2}\left(\sum_{t=1}^{T_i}\lambda_{it}\right)^2 \end{bmatrix}.$$

Finally, $\mathbf{g}_i'$ is the ith row of the $n \times (K+1)$ matrix $\mathbf{G}$. The LM statistic is $\text{LM} = \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i} = nR^2$ in the regression of a column of ones on $\mathbf{g}_i$. The first K elements of $\mathbf{i}'\mathbf{G}$ equal zero as they are the score of the log likelihood under H_0. Therefore, the LM statistic is the square of the $(K+1)$ element of $\mathbf{i}'\mathbf{G}$ times the last diagonal element of the matrix $(\mathbf{G}'\mathbf{G})^{-1}$. Wooldridge (2010) proposes an omnibus test of the null of the pooled model against the more general model that contains lagged values of $\mathbf{x}_{it}$ and/or y_{it}. The two steps of the test are: (1) Pooled probit estimation of the null model; and (2) Pooled probit estimation of the augmented model $\text{Prob}(y_{it} = 1) = \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma u_{i,t-1})$ based on observations $t = 2, \ldots, T_i$ where $u_{it} = (y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta})$. The test is a simple Wald, LM, or LR test of the hypothesis that γ equals zero.

We have examined two approaches to estimation of a probit model with random effects. GMM estimation is a third possibility. Avery, Hansen, and Hotz (1983), Bertschek and Lechner (1998), and Inkmann (2000) examine this approach; the latter two offer some comparison with the quadrature and simulation-based estimators considered here. (Our application in Example 17.36 will use the Bertschek and Lechner data.)

17.7.3 FIXED EFFECTS

The fixed effects model is

$$\begin{aligned} y_{it}^* &= \alpha_i d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i, \\ y_{it} &= \mathbf{1}(y_{it}^* > 0), \end{aligned} \tag{17-44}$$

where d_{it} is a dummy variable that takes the value one for individual i and zero otherwise. For convenience, we have redefined $\mathbf{x}_{it}$ to be the nonconstant variables in the model. The parameters to be estimated are the K elements of $\boldsymbol{\beta}$ and the n individual constant terms. Before we consider the several virtues and shortcomings of this model, we consider the practical aspects of estimation of what are possibly a huge number of parameters; $(n + K)$; n is not limited here, and could be in the thousands in a typical application. The log-likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln P(y_{it}|\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}), \tag{17-45}$$

where $P(.)$ is the probability of the observed outcome, for example, $\Phi[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$ for the probit model or $\Lambda[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$ for the logit model, where $q_{it} = 2y_{it} - 1$. What follows can be extended to any index function model, but for the present, we will confine our attention to symmetric distributions such as the normal and logistic, so that the probability can be conveniently written as $\text{Prob}(Y_{it} = y_{it}|\mathbf{x}_{it}) = P[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$. It will be convenient to let $z_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}$ so $(Y_{it} = y_{it}|\mathbf{x}_{it}) = P(q_{it}z_{it})$.

In our previous application of this model, in the linear regression case, we found that estimation of the parameters was simplified by a transformation of the data to deviations from group means, which eliminated the person-specific constants from the estimator. (See Section 11.4.1.) Save for the special case discussed later, that will not be possible here, so that if one desires to estimate the parameters of this model, it will be necessary actually to compute the possibly huge number of constant terms at the same time. This has been widely viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception.[54] The method for estimation of nonlinear fixed effects models such as the probit and logit models is detailed in Section 14.9.6.d.[55]

The problems with the fixed effects estimator are statistical, not practical. The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. But in this setting, not only is T_i fixed, it is likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). The estimator of $\boldsymbol{\beta}$ is a function of the estimators of α, which means that the MLE of $\boldsymbol{\beta}$ is not consistent either. This is the **incidental parameters problem**. [See Neyman and Scott (1948) and Lancaster (2000).] How serious this bias is remains a question in the literature. Two pieces of received wisdom are Hsiao's (1986) results for a binary logit model [with additional results in Abrevaya (1997)] and Heckman and MaCurdy's (1980) results for the probit model. Hsiao found that for $T_i = 2$, the bias in the MLE of $\boldsymbol{\beta}$ is 100%, which is extremely pessimistic. Heckman and MaCurdy found in a Monte Carlo study that in samples of $n = 100$ and $T = 8$, the bias appeared to be on the order of 10%, which is substantive, but certainly less severe than Hsiao's results suggest. No other theoretical results have been shown for other models, although in *very* few cases, it can be shown that there is no incidental parameters problem. (The Poisson model mentioned in Section 14.9.6.d

[54]See, for example, Maddala (1987), p. 317.

[55]Fernandez-Val (2009) reports using that method to fit a probit model for 500,000 groups.

is one of these special cases.) The available mix of theoretical results and Monte Carlo evidence suggests that for binary choice estimation of static models, plim $\hat{\boldsymbol{\beta}}_{FE} = S(T)\boldsymbol{\beta}$ where $S(2) = 2$, $S(T + 1) < S(T)$ and $\lim_{T \to \infty} S(T) = 1$.[56] The issue is much less clear for dynamic models—there is little small T wisdom, though the large T result appears to apply as well.

The fixed effects approach does have some appeal in that it does not require an assumption of orthogonality of the independent variables and the heterogeneity. An ongoing pursuit in the literature is concerned with the severity of the tradeoff of this virtue against the incidental parameters problem. Some commentary on this issue appears in Arellano (2001). Results of our own investigation appear in Section 15.5.2 and Greene (2004).

17.7.3.a A Conditional Fixed Effects Estimator

Why does the incidental parameters problem arise here and not in the linear regression model?[57] Recall that estimation in the regression model was based on the deviations from group means, not the original data as it is here. The result we exploited there was that although $f(y_{it}|\mathbf{X}_i)$ is a function of α_i, $f(y_{it}|\mathbf{X}_i, \bar{y}_i)$ is not a function of α_i, and we used the latter in estimation of $\boldsymbol{\beta}$. In that setting, $\bar{y}_i$ is a **minimal sufficient statistic** for α_i. Sufficient statistics are available for a few distributions that we will examine, but not for the probit model. They are available for the logit model, as we now examine.

A fixed effects binary logit model is

$$\text{Prob}(y_{it} = 1|\mathbf{x}_{it}) = \frac{e^{\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

The unconditional likelihood for the nT independent observations is

$$L = \prod_i \prod_t (F_{it})^{y_{it}} (1 - F_{it})^{1 - y_{it}}.$$

Chamberlain (1980) [following Rasch (1960) and Andersen (1970)] observed that the **conditional likelihood function**,

$$L^c = \prod_{i=1}^{n} \text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \middle| \sum_{t=1}^{T_i} y_{it} \right),$$

is free of the incidental parameters, α_i. The joint likelihood for each set of T_i observations conditioned on the number of ones in the set is

$$\text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \middle| \sum_{t=1}^{T_i} y_{it}, \mathbf{x}_i \right)$$

$$= \frac{\exp\left(\sum_{t=1}^{T_i} y_{it}\mathbf{x}_{it}'\beta \right)}{\sum_{\Sigma_t d_{it} = S_i} \exp\left(\sum_{t=1}^{T_i} d_{it}\mathbf{x}_{it}'\beta \right)}. \qquad \textbf{(17-46)}$$

[56] For example, Hahn and Newey (2002), Fernandez-Val (2009), Greene (2004), Katz (2001), Han (2002) and others.

[57] The incidental parameters problem *does* show up in ML estimation of the FE linear model, where Neyman and Scott (1948) discovered it, in estimation of σ_ε^2. The MLE of σ_ε^2 is $\boldsymbol{e}'\boldsymbol{e}/nT$, which converges to $[(T - 1)/T]\sigma_\varepsilon^2 < \sigma_\varepsilon^2$.

The function in the denominator is summed over the set of all $\binom{T_i}{S_i}$ different sequences of T_i zeros and ones that have the same sum as $S_i = \sum_{t=1}^{T_i} y_{it}$.[58]

Consider the example of $T_i = 2$. The unconditional likelihood is

$$L = \prod_i \text{Prob}(Y_{i1} = y_{i1}) \, \text{Prob}(Y_{i2} = y_{i2}).$$

For each pair of observations, we have these possibilities:

1. $y_{i1} = 0$ and $y_{i2} = 0$. $\text{Prob}(0, 0 \,|\, \text{sum} = 0) = 1$.
2. $y_{i1} = 1$ and $y_{i2} = 1$. $\text{Prob}(1, 1 \,|\, \text{sum} = 2) = 1$.

The ith term in L^c for either of these is just one, so they contribute nothing to the conditional likelihood function.[59] When we take logs, these terms (and these observations) will drop out. But suppose that $y_{i1} = 0$ and $y_{i2} = 1$. Then

$$\text{Prob}(0, 1 \,|\, \text{sum} = 1) = \frac{\text{Prob}(0, 1 \text{ and sum} = 1)}{\text{Prob}(\text{sum} = 1)} = \frac{\text{Prob}(0, 1)}{\text{Prob}(0, 1) + \text{Prob}(1, 0)}.$$

Therefore, for this pair of observations, the conditional probability is

$$\frac{\dfrac{1}{1 + e^{\alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta}}} \dfrac{e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}}}{\dfrac{1}{1 + e^{\alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta}}} \dfrac{e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}} + \dfrac{e^{\alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{i1}'\boldsymbol{\beta}}} \dfrac{1}{1 + e^{\alpha_i + \mathbf{x}_{i2}'\boldsymbol{\beta}}}} = \frac{e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}}{e^{\mathbf{x}_{i1}'\boldsymbol{\beta}} + e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}}.$$

By conditioning on the sum of the two observations, we have removed the heterogeneity. Therefore, we can construct the conditional likelihood function as the product of these terms for the pairs of observations for which the two observations are (0, 1). Pairs of observations with (1, 0) are included analogously. The product of the terms such as the preceding, for those observation sets for which the sum is not zero or T_i, constitutes the conditional likelihood. Maximization of the resulting function is straightforward and may be done by conventional methods.

As in the linear regression model, it is of some interest to test whether there is indeed heterogeneity. With homogeneity ($\alpha_i = \alpha$), there is no unusual problem, and the model can be estimated, as usual, as a logit model. It is not possible to test the hypothesis using the likelihood ratio test, however, because the two likelihoods are not comparable. (The conditional likelihood is based on a restricted data set.) None of the usual tests of restrictions can be used because the individual effects are never actually estimated.[60] Hausman's (1978) specification test is a natural one to use here, however. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood

[58] The enumeration of all these computations stands to be quite a burden—see Arellano (2000, p. 47) or Baltagi (2005, p. 235). In fact, using a recursion suggested by Krailo and Pike (1984), the computation even with T_i up to 100 is routine.

[59] In the probit model when we encounter this situation, the individual constant term cannot be estimated and the group is removed from the sample. The same effect is at work here.

[60] This produces a difficulty for this estimator that is shared by the semiparametric estimators discussed in the next section. Because the fixed effects are not estimated, it is not possible to compute probabilities or marginal effects with these estimated coefficients, and it is a bit ambiguous what one can do with the results of the computations. The brute force estimator that actually computes the individual effects might be preferable.

estimator (CMLE) and the usual maximum likelihood estimator are consistent, but Chamberlain's is inefficient. (It fails to use the information that $\alpha_i = \alpha$, and it may not use all the data.) Under the alternative hypothesis, the unconditional maximum likelihood estimator is inconsistent,[61] whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic,

$$\chi^2 = (\hat{\boldsymbol{\beta}}_{\text{CML}} - \hat{\boldsymbol{\beta}}_{\text{ML}})'(\text{Var[CML]} - \text{Var[ML]})^{-1}(\hat{\boldsymbol{\beta}}_{\text{CML}} - \hat{\boldsymbol{\beta}}_{\text{ML}}). \quad \textbf{(17-47)}$$

The estimated covariance matrices are those computed for the two maximum likelihood estimators. For the unconditional maximum likelihood estimator, the row and column corresponding to the constant term are dropped. A large value will cast doubt on the hypothesis of homogeneity. (There are K degrees of freedom for the test.) It is possible that the covariance matrix for the maximum likelihood estimator will be larger than that for the conditional maximum likelihood estimator. If so, then the difference matrix in brackets is assumed to be a zero matrix, and the chi-squared statistic is therefore zero.

Example 17.22 Binary Choice Models for Panel Data

In Example 17.6, we fit a pooled binary logit model $y = \mathbf{1}(DocVis > 0)$ using the German health care utilization data examined in appendix Table F7.1. The model is

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Income_{it} + \beta_4\, Kids_{it} + \beta_5\, Education_{it} + \beta_6\, Married_{it}).$$

No account of the panel nature of the data set was taken in that exercise. The sample contains a total of 27,326 observations on 7,293 families with T_i ranging from 1 to 7. Table 17.17 lists estimates of parameter estimates and estimated standard errors for probit and logit random and fixed effects models. There is a surprising amount of variation across the estimators. The coefficients are in bold to facilitate reading the table. It is generally difficult to compare across the estimators. The three estimators would be expected to produce very different estimates in any of the three specifications—recall, for example, the pooled estimator is inconsistent in either the fixed or random effects cases. The logit results include two fixed effects estimators. The line marked "U" is the unconditional (inconsistent) estimator. The one marked "C" is Chamberlain's consistent estimator. Note for all three fixed effects estimator it is necessary to drop from the sample any groups that have $DocVis_{it}$ equal to zero or one for every period. There were 3,046 such groups, which is about 42% of the sample. We also computed the probit random effects model in two ways, first by using the Butler and Moffitt method, then by using maximum simulated likelihood estimation. In this case, the estimators are very similar, as might be expected. The estimated correlation coefficient, ρ, is computed as $\sigma_u^2/(\sigma_\varepsilon^2 + \sigma_u^2)$. For the probit model, $\sigma_\varepsilon^2 = 1$. The MSL estimator computes $s_u = 0.9088376$, from which we obtained ρ. The estimated partial effects for the models are shown in Table 17.18. The average of the fixed effects constant terms is used to obtain a constant term for the unconditional fixed effects case. No estimator is available for the conditional fixed effects case. Once again there is a considerable amount of variation across the different estimators. On average, the fixed effects models tend to produce much larger values than the pooled or random effects models.

Example 17.23 Fixed Effects Logit Model: Magazine Prices Revisited

The fixed effects model does have some appeal, but the incidental parameters problem is a significant shortcoming of the unconditional probit and logit estimators. The conditional

[61] Hsiao (2003) derives the result explicitly for some particular cases.

TABLE 17.17 Estimated Parameters for Panel Data Binary Choice Models

			Variable					
Model	***Estimate***	**ln *L***	***Constant***	***Age***	***Income***	***Kids***	***Education***	***Married***
Logit	$\boldsymbol{\beta}$		**0.25112**	**0.02071**	**−0.18630**	**−0.22947**	**−0.04557**	**0.08530**
Pooled	St. Err.	−17673.09	0.09114	0.00129	0.07509	0.02954	0.00565	0.03328
	Rob.SE[a]		0.12827	0.00174	0.09160	0.03831	0.00808	0.04531
Logit R.E.	$\boldsymbol{\beta}$	−16277.04	**0.06447**	**0.03416**	**0.00237**	**−0.26127**	**−0.05786**	**0.02707**
$\rho = 0.41503$	St. Err.		0.16391	0.00225	0.11299	0.04589	0.01071	0.05328
Logit	$\boldsymbol{\beta}$	−9452.55		**0.10469**	**−0.05712**	**−0.08828**	**−0.11673**	**−0.05761**
F.E.(U)[b]	St. Err.			0.00726	0.17844	0.07447	0.06880	0.10619
Logit	$\boldsymbol{\beta}$	−6299.02		**0.08471**	**−0.04732**	**−0.07767**	**−0.09084**	**−0.05229**
F.E.(C)[c]	St. Err.			0.00650	0.15891	0.06228	0.05668	0.09304
Probit	$\boldsymbol{\beta}$	−17670.93	**0.15501**	**0.01283**	**−0.11666**	**−0.14118**	**−0.02811**	**0.05226**
Pooled	St. Err.		0.05652	0.00079	0.04635	0.01822	0.00350	0.02046
	Rob.SE[a]		0.07959	0.00107	0.05647	0.02361	0.00501	0.02790
Probit:RE[d]	$\boldsymbol{\beta}$	−16273.96	**0.03410**	**0.02014**	**−0.00267**	**−0.15377**	**−0.03371**	**0.01629**
$\rho = 0.44788$[e]	St. Err.		0.09635	0.00132	0.06670	0.02704	0.00629	0.03135
Probit:RE[f]	$\boldsymbol{\beta}$	−16274.06	**0.03447**	**0.02013**	**−0.00261**	**−0.15359**	**−0.03379**	**0.01749**
$\rho = 0.44768$	St. Err.		0.06337	0.00090	0.05212	0.02030	0.00394	0.02280
Probit	$\boldsymbol{\beta}$	−9453.47		**0.06249**	**−0.03155**	**−0.04818**	**−0.07222**	**−0.03298**
F.E.(U)	St. Err.			0.00432	0.10749	0.04457	0.04074	0.06364

[a]Robust, "cluster" corrected standard error.
[b]Unconditional fixed effects estimator.
[c]Conditional fixed effects estimator.
[d]Butler and Moffitt estimator.
[e]Probit LM statistic = 1011.43.
[f]Maximum simulated likelihood estimator.

TABLE 17.18 Estimated Partial Effects for Panel Data Binary Choice Models

Model	*Age*	*Income*	*Kids*	*Education*	*Married*
Logit, P[a]	0.00472	−0.04238	−0.05272	−0.01037	0.01951
Logit: RE,Q[b]	0.00705	0.00049	−0.05461	−0.01193	0.00560
Logit: F,U[c]	0.02570	−0.01402	−0.02167	−0.02865	−0.01404
Logit: F,C[d]	—	—	—	—	—
Probit, P[a]	0.00475	−0.04315	−0.05267	−0.01040	0.01942
Probit RE.Q[b]	0.00550	−0.00073	−0.04226	−0.00920	0.00445
Probit:RE,S[e]	0.00694	−0.00090	−0.05362	−0.01166	0.00605
Probit: F,U[c]	0.01312	−0.00662	−0.01012	−0.01516	−0.00688

[a]Pooled estimator.
[b]Butler and Moffitt estimator.
[c]Unconditional fixed effects estimator.
[d]Conditional fixed effects estimator. Partial effects not computed.
[e]Maximum simulated likelihood estimator.

MLE for the fixed effects logit model is a fairly common approach. A widely cited application of the model is Cecchetti's (1986) analysis of changes in newsstand prices of magazines. Cecchetti's model was

$$\text{Prob}(\textit{Price change in year t of magazine i}) = \Lambda(\alpha_j + \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where the variables in $\mathbf{x}_{it}$ are: (1) time since last price change, (2) inflation since last change, (3) previous fixed price change, (4) current inflation, (5) industry sales growth, and (6) sales volatility. The fixed effect in the model is indexed "*j*" rather than "*i*" as it is defined as a three-year interval for magazine *i*. Thus, a magazine that had been on the newstands for nine years would have three constants, not just one. In addition to estimating several specifications of the price change model, Cecchetti used the Hausman test in (17-47) to test for the existence of the common effects. Some of Cecchetti's results appear in Table 17.19.

Willis (2006) argued that Cecchetti's estimates were inconsistent and the Hausman test is invalid because right-hand-side variables (1), (2), and (6) are all functions of lagged dependent variables. This state dependence invalidates the use of the sum of the observations for the group as a sufficient statistic in the Chamberlain estimator and the Hausman tests. He proposes, instead, a method suggested by Heckman and Singer (1984b) to incorporate the unobserved heterogeneity in the *unconditional* likelihood function. The Heckman and Singer model can be formulated as a latent class model (see Section 14.15.7) in which the classes are defined by different constant terms—the remaining parameters in the model are constrained to be equal across classes. Willis fit the Heckman and Singer model with two classes to a restricted version of Cecchetti's model using variables (1), (2), and (5). The results in Table 17.19 show some of the results from Willis's Table I. (Willis reports that he could not reproduce Cecchetti's results—the ones in Cecchetti's second column would be the counterparts—because of some missing values. In fact, Willis's estimates are quite far from Cecchetti's results, so it will be difficult to compare them. Both are reported here.)

The two mass points reported by Willis are shown in Table 17.19. He reported that these two values (−1.94 and −29.15) correspond to class probabilities of 0.88 and 0.12, though it is difficult to make the translation based on the reported values. He does note that the change in the log likelihood in going from one mass point (pooled logit model) to two is marginal, only from −500.45 to −499.65. There is another anomaly in the results that is consistent with this

TABLE 17.19 Models for Magazine Price Changes (Standard errors in parentheses)

	Pooled	*Unconditional FE*	*Conditional FE Cecchetti*	*Conditional FE Willis*	*Heckman and Singer*
β_1	−1.10 (0.03)	−0.07 (0.03)	1.12 (3.66)	1.02 (0.28)	−0.09 (0.04)
β_2	6.93 (1.12)	8.83 (1.25)	11.57 (1.68)	19.20 (7.51)	8.23 (1.53)
β_5	−0.36 (0.98)	−1.14 (1.06)	5.85 (1.76)	7.60 (3.46)	−0.13 (1.14)
Constant 1	−1.90 (0.14)				−1.94 (0.20)
Constant 2					−29.15 (1.1e11)
ln L	−500.45	−473.18	−82.91	−83.72	−499.65
Sample size	1026	1026		543	1026

finding. The reported standard error for the second mass point is 1.1×10^{11}, or essentially $+\infty$. The finding is consistent with overfitting the latent class model. The results suggest that the better model is a one-class (pooled) model.

17.7.3.b Mundlak's Approach, Variable Addition, and Bias Reduction

Thus far, both the fixed effects (FE) and the random effects (RE) specifications present problems for modeling binary choice with panel data. The MLE of the FE model is inconsistent even when the model is properly specified—this is the incidental parameters problem. (And, like the linear model, the FE probit and logit models do not allow time-invariant regressors.) The random effects specification requires a strong, often unreasonable assumption that the effects and the regressors are uncorrelated. Of the two, the FE model is the more appealing, though with modern longitudinal data sets with many demographics, the problem of time-invariant variables would seem to be compelling. This would seem to recommend the conditional estimator in Section 17.4.4, save for yet another complication. With no estimates of the constant terms, neither probabilities nor partial effects can be computed with the results. We are left making inferences about ratios of coefficients. Two approaches have been suggested for finding a middle ground: Mundlak's (1978) approach that involves projecting the effects on the group means of the time-varying variables and recent developments such as Fernandez-Val's (2009) approach that involves correcting the bias in the FE MLE.

The Mundlak (1978) approach[62] augments (17-44) as follows:

$$\begin{aligned} y_{it}^{*} &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \\ \text{Prob}(y_{it} = 1 \,|\, \mathbf{x}_{it}) &= F(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}) \\ \alpha_i &= \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + u_i, \end{aligned}$$

where we have used $\bar{\mathbf{x}}_i$ generically for the group means of the time-varying variables in $\mathbf{x}_{it}$. The reduced form of the model is

$$\text{Prob}(y_{it} = 1 \,|\, \mathbf{X}_i) = F(\alpha + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i).$$

(Wooldridge and Chamberlain also suggest using all years of $\mathbf{x}_{it}$ rather than the group means. This raises a problem in unbalanced panels, however. We will ignore this possibility.) The projection of α_i on $\bar{\mathbf{x}}_i$ produces a random effects formulation. As in the

[62]See also Chamberlain (1984) and Wooldridge (2010).

linear model (see Sections 11.5.6 and 11.5.7), it also suggests a means of testing for fixed versus random effects. Because $\boldsymbol{\delta} = \mathbf{0}$ produces the pure random effects model, a joint Wald test of the null hypothesis that $\boldsymbol{\delta}$ equals zero can be used.

Example 17.24 Panel Data Random Effects Estimators

Example 17.22 presents several panel data estimators for the probit and logit models. Pooled, random effects, and fixed effects estimates are given for the probit model

$$\text{Prob}(DocVis_{it} > 0) = \Phi(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Income_{it} + \beta_4\, Kids_{it} + \beta_5\, Education_{it} + \beta_6\, Married_{it}).$$

We continue that analysis here by considering Mundlak's approach to the common effects model. Table 17.20 presents the random effects model from earlier, and the augmented estimator that contains the group means of the variables, all of which are time varying. The addition of the group means to the regression brings large changes to the estimates of the parameters, which might suggest the appropriateness of the fixed effects model. A formal test is carried by computing a Wald statistic for the null hypothesis that the last five coefficients in the augmented model equal zero. The chi-squared statistic equals 113.35 with 5 degrees of freedom. The critical value from the chi-squared table for 95% significance is 11.07, so the hypothesis that $\boldsymbol{\delta}$ equals zero, that is, the hypothesis of the random effects model (restrictions), is rejected. The two log likelihoods are −16,273.96 for the REM and −16,222/04 for the augmented REM. The LR statistic would be twice the difference, or 103.4. This produces the same conclusion. The FEM appears to be the preferred model.

A series of recent studies has sought to maintain the fixed effects specification while correcting the bias due to the incidental parameters problem. There are two broad approaches. Hahn and Kuersteiner (2004), Hahn and Newey (2005), and Fernandez-Val (2009) have developed an approximate, "large T" result for plim$(\hat{\boldsymbol{\beta}}_{FE,MLE} - \boldsymbol{\beta})$ that produces a direct correction to the estimator, itself. Fernandez-Val (2009) develops corrections for the estimated constant terms as well. Arellano and Hahn (2006, 2007) propose a modification of the log-likelihood function with, in turn, different first-order estimation equations, that produces an approximately unbiased estimator of $\boldsymbol{\beta}$. In a similar fashion to the second of these approaches, Carro (2007) modifies the first-order conditions (estimating equations) from the original log-likelihood function, once again to produce an approximately unbiased estimator of $\boldsymbol{\beta}$. [In general, given the overall approach of using a large T approximation, the payoff to these estimators is to reduce the bias of the FE, MLE from O$(1/T)$ to O$(1/T^2)$, which is a considerable reduction.] These estimators are not yet in widespread use. The received evidence suggests that in the

TABLE 17.20 Estimated Random Effects Models

	Basic Random Effects		*Mundlak Formulation*			
	Estimate	***Std. Error***	***Estimate***	***Std. Error***	***Mean***	***Std. Error***
Constant	0.03410	(0.09635)	0.37496	(0.10501)		
Age	0.02014	(0.00132)	0.05032	(0.00357)	−0.03656	(0.00384)
Income	−0.00267	(0.06770)	−0.02863	(0.09325)	−0.35365	(0.13991)
Kids	−0.15377	(0.02704)	−0.04195	(0.03752)	−0.22516	(0.05499)
Education	−0.03371	(0.00629)	−0.05450	(0.03307)	0.02391	(0.03374)
Married	0.01629	(0.03135)	−0.02661	(0.05180)	0.14689	(0.06606)

simple case we are considering here, the incidental parameters problem is a secondary concern when T reaches say 10 or so. For some modern public use data sets, such as the BHPS or GSOEP which are well beyond their 15th wave, the incidental parameters problem may not be too severe. However, most of the studies mentioned above are concerned with dynamic models (see Section 17.7.4), where the problem is possibly more severe than in the static case. Research in this area is ongoing.

17.7.4 DYNAMIC BINARY CHOICE MODELS

A random or fixed effects model that explicitly allows for lagged effects would be

$$y_{it} = \mathbf{1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma y_{i,t-1} + \varepsilon_{it} > 0).$$

Lagged effects, or **persistence**, in a binary choice setting can arise from three sources, serial correlation in ε_{it}, the heterogeneity, α_i, or true **state dependence** through the term $\gamma y_{i,t-1}$. Chiappori (1998) and Arellano (2001) suggest an application to the French automobile insurance market in which the incentives built into the pricing system are such that having an accident in one period should lower the probability of having one in the next (state dependence), but some drivers remain more likely to have accidents than others in every period, which would reflect the heterogeneity instead. State dependence is likely to be particularly important in the typical panel, which has only a few observations for each individual. Heckman (1981a) examined this issue at length. Among his findings were that the somewhat muted small sample bias in fixed effects models with $T = 8$ was made much worse when there was state dependence. A related problem is that with a relatively short panel, the **initial conditions**, y_{i0}, have a crucial impact on the entire path of outcomes. Modeling dynamic effects and initial conditions in binary choice models is more complex than in the linear model, and by comparison, there are relatively fewer firm results in the applied literature.[63]

The correlation between α_i and $y_{i,t-1}$ in the dynamic binary choice model makes $y_{i,t-1}$ endogenous. Thus, the estimators we have examined so far will not be consistent. Two familiar alternative approaches that have appeared in recent applications are due to Heckman (1981) and Wooldridge (2005), both of which build on the random effects specification. Heckman's approach provides a separate equation for the initial condition,

$$\begin{aligned}\text{Prob}(y_{i1} = 1 | \mathbf{x}_{i1}, \mathbf{z}_i, \alpha_i) &= \Phi(\mathbf{x}_{i1}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau} + \theta\alpha_i)\\ \text{Prob}(y_{it} = 1 | \mathbf{x}_{it}, y_{i,t-1}, \alpha_i) &= \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i), t = 2, \ldots, T_i,\end{aligned}$$

where $\mathbf{z}_i$ is a set of instruments observed at the first period that are not contained in $\mathbf{x}_{it}$. The conditional log likelihood is

$$\begin{aligned}\ln L | \boldsymbol{\alpha} &= \sum_{i=1}^{n} \ln\left\{\Phi[(2y_{i1} - 1)(\mathbf{x}_{i1}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau} + \theta\alpha_i)]\prod_{t=2}^{T_i}\Phi[(2y_{it} - 1)(\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i)]\right\}\\ &= \sum_{i=1}^{n} \ln L_i | \alpha_i.\end{aligned}$$

[63] A survey of some of these results is given by Hsiao (2003). Most of Hsiao (2003) is devoted to the linear regression model. A number of studies specifically focused on discrete choice models and panel data have appeared recently, including Beck, Epstein, Jackman, and O'Halloran (2001), Arellano (2001), and Greene (2001). Vella and Verbeek (1998) provide an application to the joint determination of wages and union membership. Other important references are Aguirregabiria and Mira (2010), Carro (2007), and Fernandez-Val (2009). Stewart (2006) and Arulampalam and Stewart (2007) provide several results for practitioners.

We now adopt the random effects approach and further assume that α_i is normally distributed with mean zero and variance σ_α^2. The random effects log-likelihood function can be maximized with respect to $(\boldsymbol{\delta}, \boldsymbol{\tau}, \theta, \boldsymbol{\beta}, \gamma, \sigma_\alpha)$ using either the Butler and Moffitt quadrature method or the maximum simulated likelihood method described in Section 17.4.2. Stewart and Arulampalam (2007) suggest a useful shortcut for formulating the Heckman model. Let $D_{it} = 1$ and $\gamma = \theta - 1$ in period 1 and 0 in every other period, $C_{it} = 1 - D_{it}$. Then, the two parts may be combined in

$$\ln L|\boldsymbol{\alpha} = \sum_{i=1}^{n} \ln \prod_{t=1}^{T_i} \{\Phi[(2y_{it} - 1)\langle C_{it}(\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i,t-1}) + D_{it}(\mathbf{x}_{it}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau}) + (1 + \lambda D_{it})\alpha_i \rangle]\}.$$

In this form, the model can be viewed as a random parameters (random constant term) model in which there is heteroscedasticity in the random part of the constant term.

Wooldridge's approach builds on the Mundlak device of the previous section. Starting from the same point, he suggests a model for the random effect conditioned on the initial value. Thus,

$$\alpha_i | y_{i1}, \mathbf{z}_i \sim N[\alpha_0 + \eta y_{i1} + \mathbf{z}_i'\boldsymbol{\tau}, \sigma_\alpha^2].$$

Assembling the parts, Wooldridge's model is a bit simpler than Heckman's,

$$\begin{aligned}&\text{Prob}(Y_{it} = y_{it}|\mathbf{x}_{it}, y_{i1}, u_i)\\ &\quad = \Phi[(2y_{it} - 1)(\alpha_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \eta y_{i1} + \mathbf{z}_i'\boldsymbol{\tau} + u_i)], t = 2, \ldots, T_i.\end{aligned}$$

The source of the instruments $\mathbf{z}_i$ is unclear. Wooldridge (2005) simplifies the model a bit by using, instead, a Mundlak approach, using the group means of the time-varying variables as $\mathbf{z}$. The resulting random effects formulation is

$$\begin{aligned}&\text{Prob}(Y_{it} = y_{it}|\mathbf{x}_{it}, y_{i1}, y_{i,t-1}, u_i)\\ &\quad = \Phi[(2y_{it} - 1)(\alpha_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \eta y_{i1} + \bar{\mathbf{x}}_i'\boldsymbol{\tau} + u_i)], t = 2, \ldots, T_i.\end{aligned}$$

Much of the contemporary literature has focused on methods of avoiding the strong parametric assumptions of the probit and logit models. Manski (1987) and Honore and Kyriazidou (2000) show that Manski's (1986) maximum score estimator can be applied to the differences of unequal pairs of observations in a two-period panel with fixed effects. However, the limitations of the maximum score estimator have motivated research on other approaches. An extension of lagged effects to a parametric model is Chamberlain (1985), Jones and Landwehr (1988), and Magnac (1997), who added state dependence to Chamberlain's fixed effects logit estimator. Unfortunately, once the identification issues are settled, the model is only operational if there are no other exogenous variables in it, which limits its usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well.

Dong and Lewbel (2010) have extended Lewbel's *special regressor* method to dynamic binary choice models and have devised an estimator based on an IV linear regression. Honore and Kyriazidou (2000) have combined the logic of the *conditional logit model* and Manski's maximum score estimator. They specify

$$\begin{aligned}&\text{Prob}(y_{i0} = 1|\mathbf{x}_i, \alpha_i) = p_0(\mathbf{x}_i, \alpha_i) \quad \text{where } \mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}),\\ &\text{Prob}(y_{it} = 1|\mathbf{x}_i, \alpha_i, y_{i0}, y_{i1}, \ldots, y_{i,t-1}) = F(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma y_{i,t-1}) \quad t = 1, \ldots, T.\end{aligned}$$

The analysis assumes a single regressor and focuses on the case of $T = 3$. The resulting estimator resembles Chamberlain's but relies on observations for which $\mathbf{x}_{it} = \mathbf{x}_{i,t-1}$,

which rules out direct time effects as well as, for practical purposes, any continuous variable. The restriction to a single regressor limits the generality of the technique as well. The need for observations with equal values of $\mathbf{x}_{it}$ is a considerable restriction, and the authors propose a kernel density estimator for the difference, $\mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, instead which does relax that restriction a bit. The end result is an estimator that converges (they conjecture) but to a nonnormal distribution and at a rate slower than $n^{-1/3}$.

Semiparametric estimators for dynamic models at this point in the development are still primarily of theoretical interest. Models that extend the parametric formulations to include state dependence have a much longer history, including Heckman (1978, 1981a, 1981b), Heckman and MaCurdy (1980), Jakubson (1988), Keane (1993), and Beck et al. (2001) to name a few.[64] In general, even without heterogeneity, dynamic models ultimately involve modeling the joint outcome $(y_{i0}, \ldots, y_{iT})$, which necessitates some treatment involving multivariate integration. Example 17.14 describes an application. Stewart (2006) provides another.

Example 17.25 A Dynamic Model for Labor Force Participation and Disability

Gannon (2005) modeled the relationship between labor force participation and disability in Ireland with a panel data set, *The Living in Ireland Survey 1995–2000*. The sample begins in 1995 with 7,254 individuals, but with attrition, shrinks to 3,670 in 2000. The dynamic probit model is

$$y_{it}^* = b_0 + b_1 y_{i,t-1} + b_2 D_{it} + b_3 D_{i,t-1} + b_4 z_{it} + \alpha_i + \varepsilon_{it}, \; y_{it} = \mathbf{1}(y_{it}^* > 0),$$

where y_{it} is the labor force participation indicator and D_{it} is an indicator of disability. The related covariates are gathered in z_{it}. The lagged value of D_{it} helps distinguish longer-term disabilities from those recently acquired. Unobserved time-invariant individual effects are captured by the common effect, α_i. The lagged dependent variable helps distinguish between the impact of the individual effect and the inertia of past participation. Variables in z_{it} include age, residence region, education, marital status, children, and unearned income.

The starting point of the analysis is a pooled probit model without the common effect (with standard errors corrected for the clustering at the individual level). The pooled model leaves two interesting questions:

1. Do the control variables adequately account for the unobserved characteristics?
2. Does past disability affect participation directly as in the model, or through some different channel that affects past participation?

The author adopts Wooldridge's (2005) (Mundlak) form of the random effects model we examined in Section 17.7.3.b and Example 17.24 to deal with the unobserved heterogeneity and the initial conditions problem. Thus, the initial value of y_{it} and the group means of time-varying variables are added to the random effects model,

$$y_{it}^* = b_1 y_{i,t-1} + b_2 D_{it} + b_3 D_{i,t-1} + b_4 z_{it} + \alpha_0 + \alpha_1 y_{i0} + \boldsymbol{\alpha}_2' \bar{\mathbf{x}}_i + a_i + \varepsilon_{it}, \; y_{it} = \mathbf{1}(y_{it}^* > 0).$$

The resulting model is now estimated using the Butler and Moffitt method for random effects.

Example 17.26 An Intertemporal Labor Force Participation Equation

Hyslop (1999) presents a model of the labor force participation of married women. The focus of the study is the high degree of persistence in the participation decision. Data used in the

[64] Beck et al. (2001) is a bit different from the others mentioned in that in their study of "state failure," they observe a large sample of countries (147) over a fairly large number of years, 40. As such, they are able to formulate their models in a way that makes the asymptotics with respect to T appropriate. They can analyze the data essentially in a time-series framework. Sepanski (2000) is another application that combines state dependence and the random coefficient specification of Akin, Guilkey, and Sickles (1979).

study were the years 1979–1985 of the *Panel Study of Income Dynamics*. A sample of 1,812 continuously married couples was studied. Exogenous variables that appeared in the model were measures of permanent and transitory income and fertility captured in yearly counts of the number of children from 0 to 2, 3 to 5, and 6 to 17 years old. Hyslop's formulation, in general terms, is

$$\begin{aligned}
&\text{(initial condition) } y_{i0} = \mathbf{1}(\mathbf{x}_{i0}'\boldsymbol{\beta}_0 + v_{i0} > 0),\\
&\text{(dynamic model) } y_{it} = \mathbf{1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i + v_{it} > 0)\\
&\text{(heterogeneity correlated with participation) } \alpha_i = \mathbf{z}_i'\boldsymbol{\delta} + \eta_i,\\
&\text{(stochastic specification)}\\
&\qquad \eta_i | \mathbf{X}_i \sim \text{N}[0, \sigma_\eta^2],\\
&\qquad v_{i0} | \mathbf{X} \sim \text{N}[0, \sigma_0^2],\\
&\qquad w_{it} | \mathbf{X}_i \sim \text{N}[0, \sigma_{\text{w}}^2],\\
&\qquad v_{it} = \rho v_{i,t-1} + w_{it},\ \sigma_\eta^2 + \sigma_w^2 = 1,\\
&\qquad \text{Corr}[v_{i0}, v_{it}] = \rho^t, \quad t = 1, \ldots, T-1.
\end{aligned}$$

The presence of the autocorrelation and state dependence in the model invalidate the simple maximum likelihood procedures we examined earlier. The appropriate likelihood function is constructed by formulating the probabilities as

$$\text{Prob}(y_{i0}, y_{i1}, \ldots) = \text{Prob}(y_{i0}) \times \text{Prob}(y_{i1} | y_{i0}) \times \cdots \times \text{Prob}(y_{iT} | y_{i,T-1}).$$

This still involves a $T = 7$ order normal integration, which is approximated in the study using a simulator similar to the GHK simulator discussed in 15.6.2.b. Among Hyslop's results are a comparison of the model fit by the simulator for the multivariate normal probabilities with the same model fit using the maximum simulated likelihood technique described in Section 15.6.

17.7.5 A SEMIPARAMETRIC MODEL FOR INDIVIDUAL HETEROGENEITY

The panel data analysis considered thus far has focused on modeling heterogeneity with the fixed and random effects specifications. Both assume that the heterogeneity is continuously distributed among individuals. The random effects model is fully parametric, requiring a full specification of the likelihood for estimation. The fixed effects model is essentially semiparametric. It requires no specific distributional assumption; however, it does require that the realizations of the latent heterogeneity be treated as parameters, either estimated in the unconditional fixed effects estimator or conditioned out of the likelihood function when possible. As noted in Example 17.23, Heckman and Singer's (1984b) model provides a less stringent specification based on a discrete distribution of the latent heterogeneity. A straightforward method of implementing their model is to cast it as a latent class model in which the classes are distinguished by different constant terms and the associated probabilities. The class probabilities are treated as parameters to be estimated with the model parameters.

Example 17.27 Semiparametric Models of Heterogeneity

We have extended the random effects and fixed effects logit models in Example 17.22 by fitting the Heckman and Singer (1984b) model. Table 17.21 shows the specification search and the results under different specifications. The first column of results shows the estimated fixed effects model from Example 17.22. The conditional estimates are shown in parentheses. Of the 7,293 groups in the sample, 3,056 are not used in estimation of the fixed effects models because the sum of $Doctor_{it}$ is either 0 or T_i for the group. The mean and standard deviation of the estimated underlying heterogeneity distribution are computed using the estimates of

TABLE 17.21 Estimated Heterogeneity Models

		Number of Classes				
	Fixed Effect	***1***	***2***	***3***	***4***	***5***
β_1	0.10475	0.02071	0.03033	0.03368	0.03408	0.03416
	(0.08476)					
β_2	−0.06097	−0.18592	0.02555	−0.00580	−0.00635	−0.01363
	(−0.05038)					
β_3	−0.08841	−0.22947	−0.24708	−0.26388	−0.26590	−0.26626
	(−0.07776)					
β_4	−0.11671	−0.04559	−0.05092	−0.05802	−0.05975	−0.05918
	(−0.09082)					
β_5	−0.05732	0.08529	0.04297	0.03794	0.02923	0.03070
	(−0.52072)					
α_1	−2.62334	0.25111	0.91764	1.71669	1.94536	2.76670
		(1.00000)	(0.62681)	(0.34838)	(0.29309)	(0.11633)
α_2			−1.47800	−2.23491	−1.76371	1.18323
			(0.37319)	(0.18412)	(0.21714)	(0.26468)
α_3				−0.28133	−0.03674	−1.96750
				(0.46749)	(0.46341)	(0.19573)
α_4					−4.03970	−0.25588
					(0.02636)	(0.40930)
α_5						−6.48191
						(0.01396)
Mean	−2.62334	0.25111	0.02361	0.05506	0.06369	0.05471
Std. Dev.	3.13415	0.00000	1.15866	1.40723	1.48707	1.62143
ln L	−9458.638	−17673.10	−16353.14	−16278.56	−16276.07	−16275.85
	(−6299.02)					
AIC/N	1.00349	1.29394	1.19748	1.19217	1.19213	1.19226

α_i for the remaining 4,237 groups. The remaining five columns in the table show the results for different numbers of latent classes in the Heckman and Singer model. The listed constant terms are the "mass points" of the underlying distributions. The associated class probabilities are shown in parentheses under them. The mean and standard deviation are derived from the 2-to-5 point discrete distributions shown. It is noteworthy that the mean of the distribution is relatively stable, but the standard deviation rises monotonically. The search for the best model would be based on the AIC. As noted in Section 14.15.5, using a likelihood ratio test in this context is dubious, as the number of degrees of freedom is ambiguous. Based on the AIC, the four-class model is the preferred specification.

17.7.6 MODELING PARAMETER HETEROGENEITY

In Section 11.10, we examined specifications that extend the underlying heterogeneity to all the parameters of the model. We have considered two approaches. The random parameters or mixed models discussed in Chapter 15 allow parameters to be distributed continuously across individuals. The latent class model in Section 14.15 specifies a discrete distribution instead. (The Heckman and Singer model in the previous section

applies this method to the constant term.) Most of the focus to this point, save for Example 14.17, has been on linear models.

The random effects model can be cast as a model with a random constant term,

$$y_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$y_{it} = \mathbf{1}(y_{it}^* > 0),$$

where $\alpha_i = \alpha + \sigma_u u_i$. This is simply a reinterpretation of the model we just analyzed. We might, however, now extend this formulation to the full parameter vector. The resulting structure is

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i,$$
$$y_{it} = \mathbf{1}(y_{it}^* > 0),$$

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_i$ and $\boldsymbol{\Gamma}$ is a nonnegative definite diagonal matrix—some of its diagonal elements could be zero for nonrandom parameters. The method of estimation is maximum simulated likelihood. The simulated log likelihood is now

$$\ln L_{Simulated} = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} F[q_{it}(\mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_{ir}))] \right] \right\}.$$

The simulation now involves R draws from the multivariate distribution of $\mathbf{u}$. Because the draws are uncorrelated—$\boldsymbol{\Gamma}$ is diagonal—this is essentially the same estimation problem as the random effects model considered previously. This model is estimated in Example 17.28. Example 17.28 also presents a similar model that assumes that the distribution of $\boldsymbol{\beta}_i$ is discrete rather than continuous.

Example 17.28 Parameter Heterogeneity in a Binary Choice Model

We have extended the logit model for doctor visits from Example 17.14 to allow the parameters to vary randomly across individuals. The random parameters logit model is

$$\text{Prob}(Doctor_{it} = 1) = \Lambda(\beta_{1i} + \beta_{2i}\, Age_{it} + \beta_{3i}\, Income_{it} + \beta_{4i}\, Kids_{it} + \beta_{5i}\, Educ_{it} + \beta_{6i}\, Married_{it}),$$

where the two models for the parameter variation we have employed are:

$$\text{Continuous:} \quad \beta_{ki} = \beta_k + \sigma_k u_{ki},\ u_{ki} \sim \text{N}[0, 1], k = 1, \ldots, 6, \text{Cov}[u_{ki}, u_{mi}] = 0,$$
$$\text{Discrete:} \quad \beta_{ki} = \beta_k^1 \text{ with probability } \pi_1,$$
$$\beta_k^2 \text{ with probability } \pi_2,$$
$$\beta_k^3 \text{ with probability } \pi_3.$$

We have chosen a three-class latent class model for the illustration. In an application, one might undertake a systematic search, such as in Example 17.27 to find a preferred specification. Table 17.22 presents the fixed parameter (pooled) logit model and the two random parameters versions. (There are infinite variations on these specifications that one might explore—see Chapter 15 for discussion— we have shown only the simplest to illustrate the models.[65])

Figure 17.5 shows the implied distribution for the coefficient on age. For the continuous distribution, we have simply plotted the normal density. For the discrete distribution, we first

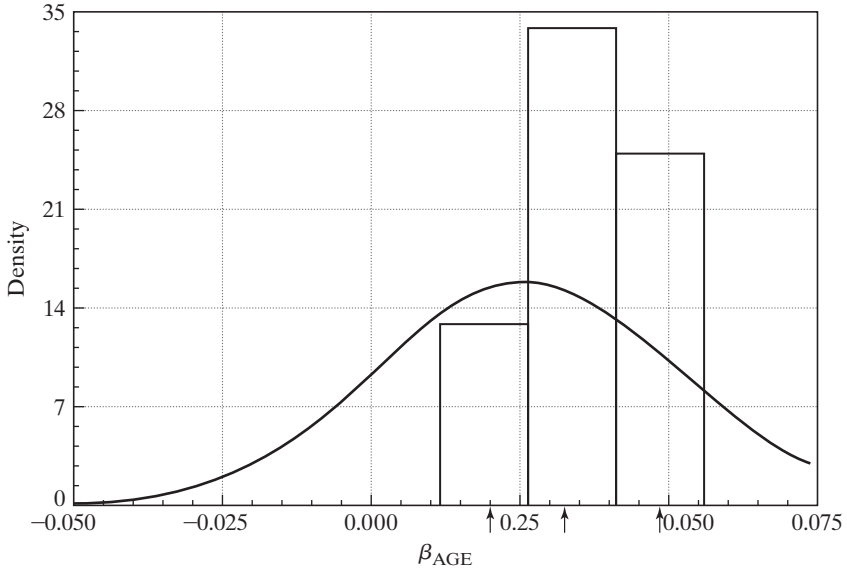
[65]Nonreplicability is an ongoing challenge in empirical work in economics. (See, for instance, Example 17.14.) The problem is particularly acute in analyses that involve simulation such as Monte Carlo studies and random parameter models. In the interest of replicability, we note that the random parameter estimates in Table 17.22 were computed with NLOGIT [Econometric Software (2007)] and are based on 50 Halton draws. We used the first six sequences (prime numbers 2, 3, 5, 7, 11, 13) and discarded the first 10 draws in each sequence.

TABLE 17.22 Estimated Heterogeneous Parameter Models

	Pooled	*Random Parameters*		*Latent Class*		
Variable	*Estimate: β*	*Estimate: β*	*Estimate: σ*	*Estimate: β*	*Estimate: β*	*Estimate: β*
Constant	0.25111	−0.03496	0.81651	0.96605	−0.18579	−1.52595
	(0.09114)	(0.07553)	(0.01654)	(0.43757)	(0.23907)	(0.43498)
Age	0.02071	0.02631	0.02533	0.04906	0.03225	0.01998
	(0.00129)	(0.00110)	(0.00042)	(0.00695)	(0.00315)	(0.00626)
Income	−0.18592	−0.00436	0.10737	−0.27917	−0.06863	0.45487
	(0.07506)	(0.06245)	(0.03828)	(0.37149)	(0.16748)	(0.31153)
Kids	−0.22947	−0.17461	0.55520	−0.28385	−0.28336	−0.11708
	(0.02954)	(0.02452)	(0.02387)	(0.14279)	(0.06640)	(0.12363)
Education	−0.04559	−0.04051	0.03792	−0.02530	−0.05734	−0.09385
	(0.00565)	(0.00475)	(0.00134)	(0.02777)	(0.01247)	(0.02797)
Married	0.08529	0.01462	0.07070	−0.10875	0.02533	0.23571
	(0.03329)	(0.027417)	(0.01736)	(0.17228)	(0.07593)	(0.14369)
Class	1.00000	1.00000		0.34833	0.46181	0.18986
Prob.	(0.00000)	(0.00000)		(0.03850)	(0.02806)	(0.02234)
ln *L*	−17673.10	−16271.72		−16265.59		

obtained the mean (0.0358) and standard deviation (0.0107). Notice that the distribution is tighter than the estimated continuous normal (mean, 0.026; standard deviation, 0.0253). To suggest the variation of the parameter (purely for purpose of the display, because the distribution is discrete), we placed the mass of the center interval, 0.461, between the midpoints of the intervals between the center mass point and the two extremes. With a width

FIGURE 17.5 Distribution of AGE Coefficient.



of 0.0145 the density is 0.461/0.0145 = 31.8. We used the same interval widths for the outer segments. This range of variation covers about five standard deviations of the distribution.

17.7.7 NONRESPONSE, ATTRITION, AND INVERSE PROBABILITY WEIGHTING

Missing observations is a common problem in the analysis of panel data. Nicoletti and Peracchi (2005) suggest several reasons that, for example, panels become unbalanced:

- Demographic events such as death;
- Movement out of the scope of the survey, such as institutionalization or emigration;
- Refusal to respond at subsequent waves;
- Absence of the person at the address;
- Other types of noncontact.

The GSOEP that we [from Riphahn, Wambach, and Million (2003)] have used in many examples in this text is one such data set. Jones, Koolman, and Rice (2006) (JKR) list several other applications, including the British Household Panel Survey (BHPS), the European Community Household Panel (ECHP), and the Panel Study of Income Dynamics (PSID).

If observations are missing completely at random (MCAR, see Section 4.7.4) then the problem of nonresponse can be ignored, though for estimation of dynamic models, either the analysis will have to be restricted to observations with uninterrupted sequences of observations, or some very strong assumptions and interpolation methods will have to be employed to fill the gaps. (See Section 4.7.4 for discussion of the terminology and issues in handling missing data.) The problem for estimation arises when observations are missing for reasons that are related to the outcome variable of interest. **Nonresponse bias** and a related problem, attrition bias (individuals leave permanently during the study), result when conventional estimators, such as least squares or the probit maximum likelihood estimator being used here are applied to samples in which observations are present or absent from the sample for reasons related to the outcome variable. It is a form of sample selection bias that we will examine further in Chapter 19.

Verbeek and Nijman (1992) have suggested a test for endogeneity of the sample response pattern. (We will adopt JKR's notation and terminology for this.) Let h denote the outcome of interest and $\mathbf{x}$ denote the relevant set of covariates. Let R denote the pattern of response. If nonresponse is (completely) random, then $E[h|\mathbf{x}, R] = E[h|\mathbf{x}]$. This suggests a variable addition test (neglecting other panel data effects); a pooled model that contains R in addition to $\mathbf{x}$ can provide the means for a simple test of endogeneity. JKR (and Verbeek and Nijman) suggest using the number of waves at which the individual is present as the measure of R. Thus, adding R to the pooled model, we can use a simple t test for the hypothesis.

Devising an estimator given that (non)response is nonignorable requires a more detailed understanding of the process generating the response pattern. The crucial issue is whether the sample selection is based *on unobservables* or *on observables*. **Selection on unobservables** results when, after conditioning on the relevant variables, $\mathbf{x}$, and other information, $\mathbf{z}$, the sampling mechanism is still nonrandom with respect to the disturbances in the models. Selection on unobservables is at the heart of the sample selectivity methodology pioneered by Heckman (1979) that we will study in Chapter 19. (Some applications of the role of unobservables in biased estimation are discussed in Chapter 8, where we examine sources of endogeneity in regression models.) If selection

is on observables and then conditioned on an appropriate specification involving the observable information, (**x**,**z**), a consistent estimator of the model parameters will be available by purging the estimator of the endogeneity of the sampling mechanism.

JKR adopt an **inverse probability weighted (IPW)** estimator devised by Robins, Rotnitsky, and Zhao (1995), Fitzgerald, Gottshalk, and Moffitt (1998), Moffitt, Fitzgerald, and Gottshalk (1999), and Wooldridge (2002). The estimator is based on the general MCAR assumption that $P(R = 1|h, \mathbf{x}, \mathbf{z}) = P(R = 1|\mathbf{x}, \mathbf{z})$. That is, the observable covariates convey all the information that determines the response pattern—the probability of nonresponse does not vary systematically with the outcome variable once the exogenous information is accounted for. Implementing this idea in an estimator would require that **x** and **z** be observable when $R = 0$, that is, the exogenous data be available for the nonresponders. This will typically not be the case; in an unbalanced panel, the entire observation is missing. Wooldridge (2002) proposed a somewhat stronger assumption that makes estimation feasible: $P(R = 1|h, \mathbf{x}, \mathbf{z}) = P(R = 1|\mathbf{z})$ where **z** is a set of covariates available at wave 1 (entry to the study). To compute Wooldridge's IPW estimator, we will begin with the sample of all individuals who are present at wave 1 of the study. (In our Example 17.17, based on the GSOEP data, not all individuals are present at the first wave.) At wave 1, $(\mathbf{x}_{i1}, \mathbf{z}_{i1})$ are observed for all individuals to be studied; $\mathbf{z}_{i1}$ contains information on observables that are not included in the outcome equation and that predict the response pattern at subsequent waves, including the response variable at the first wave. At wave 1, then, $P(R_{i1} = 1|\mathbf{x}_{i1}, \mathbf{z}_{i1}) = 1$. Wooldridge suggests using a probit model for $P(R_{it} = 1|\mathbf{x}_{i1}, \mathbf{z}_{i1})$, $t = 2, \ldots, T$ for the remaining waves to obtain predicted probabilities of response, $\hat{p}_{it}$. The IPW estimator then maximizes the weighted log likelihood,

$$\ln L_{IPW} = \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{R_{it}}{\hat{p}_{it}} \ln L_{it}.$$

Inference based on the weighted log-likelihood function can proceed as in Section 17.3. A remaining detail concerns whether the use of the predicted probabilities in the weighted log-likelihood function makes it necessary to correct the standard errors for two-step estimation. The case here is not an application of the two-step estimators we considered in Section 14.7, because the first step is not used to produce an estimated parameter vector in the second. Wooldridge (2002) shows that the standard errors computed without the adjustment are "conservative" in that they are larger than they would be with the adjustment.

Example 17.29 Nonresponse in the GSOEP Sample

Of the 7,293 individuals in the GSOEP data that we have used in several earlier examples, 3,874 were present at wave 1 (1984) of the sample. The pattern of the number of waves present by these 3,874 is shown in Figure 17.6. The waves are 1984–1988, 1991, and 1994. A dynamic model would be based on the 1,600 of those present at wave 1 who were also present for the next four waves. There is a substantial amount of nonresponse in these data. Not all individuals exit the sample with the first nonresponse, however, so the resulting panel remains unbalanced. The impression suggested by Figure 17.6 could be a bit misleading—the nonresponse pattern is quite different from simple attrition. For example, 364 of the 3,874 individuals who responded at wave 1 did not respond at wave 2 but returned to the sample at wave 3.

To employ the Verbeek and Nijman test, we used the entire sample of 27,326 household years of data. The pooled probit model for $DocVis > 0$ produced the results at the left in

Table 17.23. A *t* (Wald) test of the hypothesis that the coefficient on number of waves present is zero is strongly rejected, so we proceed to the inverse probability weighted estimator. For computing the inverse probability weights, we used the following specification:

x_{i1} = *constant*, *age*, *income*, *educ*, *kids*, *married*

z_{i1} = *female*, *handicapped dummy*, *percentage handicapped*, *university*, *working*, *blue collar*, *white collar*, *public servant*, y_{i1}

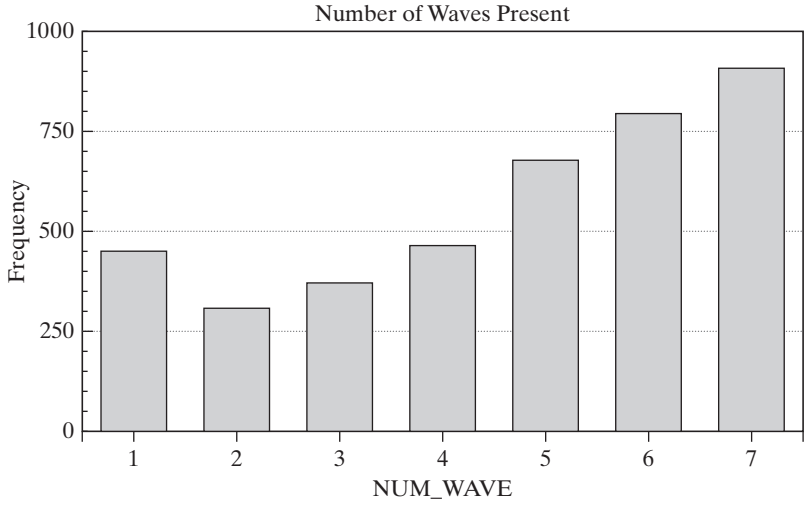
y_{i1} = *DoctorVisits* > 0 *in period* 1.

This first-year data vector is used as the observed explanatory variables in probit models for waves 2 to 7 for the 3,874 individuals who were present at wave 1. There are 3,874 observations for each of these probit models, because all were observed at wave, 1. Fitted probabilities for R_{it} are computed for waves 2 to 7, while $R_{i1} = 1$. The sample means of these probabilities, which equals the proportion of the 3,874 who responded at each wave, are 1.000, 0.730, 0.672, 0.626, 0.682, 0.568, and 0.386, respectively. Table 17.23 presents the estimated models for several specifications In each case, it appears that the weighting brings some moderate changes in the parameters and, uniformly, reductions in the standard errors.

TABLE 17.23 Inverse Probability Weighted Estimators

		Pooled Model		***Random Effects–Mundlak***		***Fixed Effects***	
Variable	***Endog. Test***	***Unwtd.***	***IPW***	***Unwtd.***	***IPW***	***Unwtd.***	***IPW***
Constant	0.26411	0.03369	−0.02373	0.09838	0.13237		
	(0.05893)	(0.07684)	(0.06385)	(0.16081)	(0.17019)		
Age	0.01369	0.01667	0.01831	0.05141	0.05656	0.06210	0.06841
	(0.00080)	(0.00107)	(0.00088)	(0.00422)	(0.00388)	(0.00506)	(0.00465)
Income	−0.12446	−0.17097	−0.22263	0.05794	0.01699	0.07880	0.03603
	(0.04636)	(0.05981)	(0.04801)	(0.11256)	(0.10580)	(0.12891)	(0.12193)
Education	−0.02925	−0.03614	−0.03513	−0.06456	−0.07058	−0.07752	−0.08574
	(0.00351)	(0.00449)	(0.00365)	(0.06104)	(0.05792)	(0.06582)	(0.06149)
Kids	−0.13130	−0.13077	−0.13277	−0.04961	−0.03427	−0.05776	−0.03546
	(0.01828)	(0.02303)	(0.01950)	(0.04500)	(0.04356)	(0.05296)	(0.05166)
Married	0.06759	0.06237	0.07015	−0.06582	−0.09235	−0.07939	−0.11283
	(0.02060)	(0.02616)	(0.02097)	(0.06596)	(0.06330)	(0.08146)	(0.07838)
Mean Age				−0.03056	−0.03401		
				(0.00479)	(0.00455)		
Mean Income				−0.66388	−0.78077		
				(0.18646)	(0.18866)		
Mean Education				0.02656	0.02899		
				(0.06160)	(0.05848)		
Mean Kids				−0.17524	−0.20615		
				(0.07266)	(0.07464)		
Mean Married				0.22346	0.25763		
				(0.08719)	(0.08433)		
Number of Waves	−0.02977						
	(0.00450)						
ρ				0.46538	0.48616		

FIGURE 17.6 Number of Waves Responded for Those Present at Wave 1.



17.9 SPATIAL BINARY CHOICE MODELS

Section 11.7 presented a model of spatial interaction among sample observations. In an application, Bell and Bockstael (2000) constructed a spatial hedonic regression model of house prices that were influenced by attributes and by neighborhood effects. We considered two frameworks for the regression model: spatial autoregression (SAR),

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \rho\Sigma_{j=1}^{n} w_{ij}y_j + \varepsilon_i, \text{ or, for all } n \text{ observations, } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon},$$

and spatial autocorrelation (SAC),

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \text{ where } \varepsilon_i = \rho\Sigma_{j=1}^{n} w_{ij}\varepsilon_j + u_i, \text{ or } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} = \rho\mathbf{W}\boldsymbol{\varepsilon} + \mathbf{u}.$$

Both cases produce a generalized regression model with full $n \times n$ covariance matrix when y is a continuous random variable. The model frameworks turn on the crucial spatial correlation parameter, ρ, and the specification of the contiguity matrix, $\mathbf{W}$, which defines the form of the spatial correlation. In Bell and Bockstael's application, in the sample of 1,000 home sales, the elements of $\mathbf{W}$ (in one of several specifications) are

$$W_{ij} = \frac{\mathbf{1}(\text{Home } i \text{ and } j \text{ are } < 600 \text{ meters apart})}{\text{Distance between homes } i \text{ and } j}; \; W_{ii} = 0.$$

(The rows of $\mathbf{W}$ are standardized.) Conditioned on the value of ρ, this produces a generalized regression model that is estimated by GMM or maximum likelihood.

We are interested in extending the idea of spatial interaction to a binary outcome.[66] Some received examples are:

- Garrett, Wagner, and Wheelock (2005) examined banks' choices of branch banking;
- McMillen (1992) examined factors associated with high (or low) crime rates in neighborhoods of Columbus, Ohio;

[66]Smirnov (2010) provides a survey of applications of spatial models to nonlinear regression settings.

- Pinske and Slade (2006) examined operation decisions (open/closed) for a panel of copper mines;
- Flores-Lagunes and Schnier (2012) extended Heckman's (1979) two-step estimator to include spatial effects in both the selection (probit) and regression steps. They apply the method to a sample of 320 observations on trawl fishing in which only 207 are fully reported (selected).
- Klier and McMillen (2008) analyzed county-wide data on auto supply plant location decisions in the U.S. Midwest. An industry that serviced the auto manufacturing centered around Detroit was earlier oriented west-east from Chicago to New York. During the mid-20th century, entry took place along an axis running from south to north (along with an historic internal migration in the U.S. that accompanied the decline of the coal industry). Klier and McMillen examined data on counties and whether an auto supplier was located in the county, a binary outcome.

The model framework is a binary choice model,

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, y_i = \mathbf{1}(y_i^* > 0).$$

The distribution for most applications will be the normal or logistic leading to a probit or logit model. A model of spatial autoregression would be

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \rho\Sigma_{j=1}^n w_{ij}y_j^* + \varepsilon_i, y_i = \mathbf{1}(y_i^* > 0).$$

Based on a random utility interpretation, it would be difficult to motivate spatial interaction based on the latent utilities.[67] The spatial autoregression model based on the observed outcomes instead would be

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \rho\Sigma_{j=1}^n w_{ij}y_j^* + \varepsilon_i, y_i = \mathbf{1}(y_i^* > 0).$$

This might seem more reasonable; however, this model is incoherent—it is not possible to insure that $\text{Prob}(y_i = 1 | \mathbf{x}_i)$ lies between zero and one. A spatial error model used in several applications is

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i; \varepsilon_i = \rho\Sigma_{j=1}^n w_{ij}\varepsilon_j + u_i, u_i \sim \text{N}[0, 1], y_i = \mathbf{1}(y_i^* > 0).$$

Pinske and Slade (1998, 2006) and McMillen (1992) use this framework to construct a GMM estimator based on the generalized residuals, λ_i, defined in (17-20). Solving for the reduced form,

$$\boldsymbol{\varepsilon} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u}.$$

The full covariance matrix for the n observations would be

$$\text{Var}[\boldsymbol{\varepsilon}] = \sigma_u^2[(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})]^{-1} = \sigma_u^2\mathbf{D}(\rho).$$

(Note that $\sigma_u^2 = 1$.) Then,

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \Sigma_{j=1}^n \mathbf{D}_{ij}(\rho)u_j, y_i = \mathbf{1}(y_i^* > 0).$$

[67] But Klier and McMillen (2008, p. 462) note, "The assumption that the latent variable depends on spatially lagged values of the latent variable may be disputable in some settings. In our example, we are assuming that the propensity to locate a new supplier plant in a county depends on the propensity to locate plants in nearby counties, and it does *not* depend simply on whether new plants have located nearby. The assumption is reasonable in this context because of the forward-looking nature of plant location decisions."

The marginal probability is

$$\text{Prob}(y_i = 1 \,|\, \mathbf{x}_i) = \text{Prob}(\mathbf{x}_i'\boldsymbol{\beta} + \Sigma_{j=1}^{n}\mathbf{D}_{ij}(\rho)u_j > 0)$$
$$= F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\Sigma_{j=1}^{n}[\mathbf{D}_{ij}(\rho)]^2}\right) = F[\mathbf{x}_i^*(\mathbf{D}, \rho)'].$$

This corresponds to the heteroscedastic probit model in Section 17.5.2. (The difference here is that the observations are all correlated.) We have seen two GMM approaches to estimation. Consistent with Bertschuk and Lechner's (1998) approach based on simple regression residuals, the GMM estimator would use $E\{\mathbf{z}_i \times [y_i - \Phi(\mathbf{x}_i*(\mathbf{D}, \rho)'\boldsymbol{\beta})]\} = \mathbf{0}$, where $\mathbf{z}_i$ is the set of instrumental variables. McMillen (1992) and Pinske and Slade (2006) use the generalized residuals, here $\lambda_i*(\mathbf{D}, \rho)$, defined in (17-20), instead,

$$E\left[\mathbf{z}_i \times \left\{\frac{(y_i - \Phi[\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta}])\phi[\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta}]}{\Phi[\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta}](1 - \Phi[\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta}])}\right\}\right] = E[\mathbf{z}_i \times \lambda(\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta})] = \mathbf{0}.$$

Pinske and Slade (2006) used a probit model while Klier and McMillen proposed a logit model. The estimation method is largely the same in both cases.

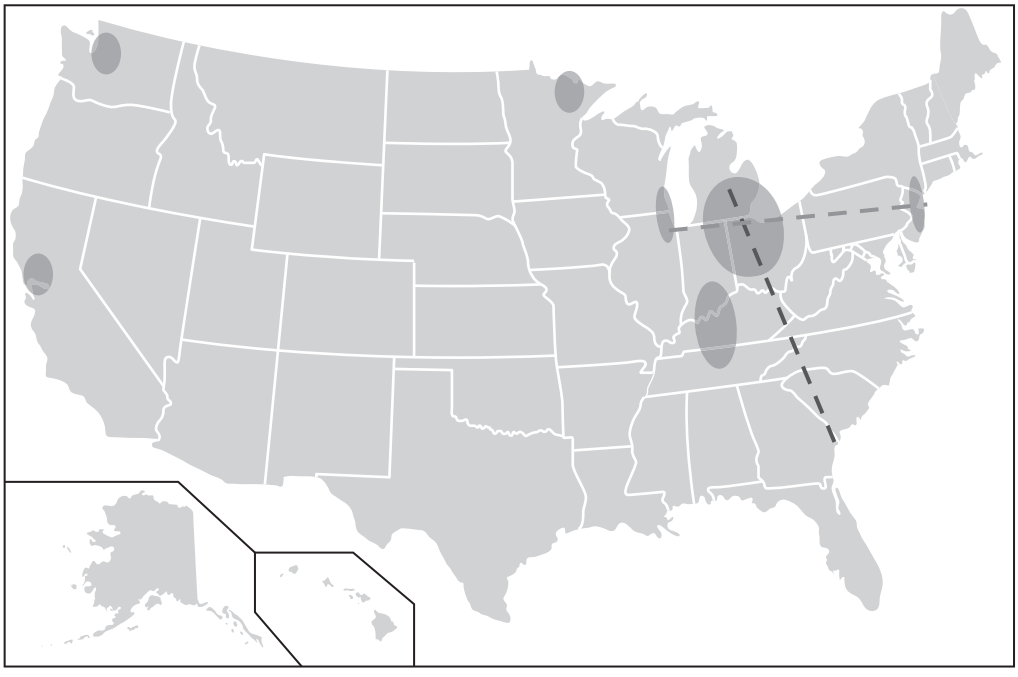
The preceding estimators use an approximation based on the marginal probability to form a feasible GMM estimator. Case (1992) suggests that if the contiguity pattern were compressed so that the data set consists of a finite number of neighborhoods, each with a small enough number of members, then the model could be handled directly by maximum likelihood. It would resemble a panel probit model in this case. Klier and McMillen used this approach to simplify their estimation procedure. Wang, Iglesias, and Wooldridge (2013) proposed a similar approach to an unrestricted model based on the principle of a partial likelihood. By using a spatial moving average for $\boldsymbol{\varepsilon}$, they show how to use pairs of observations to formulate a bivariate heteroscedastic probit model that identifies the spatial parameters.

Example 17.30 A Spatial Logit Model for Auto Supplier Locations

Klier and McMillen (2008) specified a binary logit model with spatial error correlation to model whether a county experienced a new auto supply location in 1991—2003. The data consist of 3,107 county observations. The weighting matrix is initially specified as $1/n_i$ where n_i = the number of counties that are contiguous to county i—share a common border. To speed up computation, the weighting matrix is further reduced so that counties are only contiguous if they are in the same census region. This produces a block diagonal **W** that greatly simplifies the estimation. Figure 17.7 [Based on Figure 2 from Klier and McMillen (2008)] illustrates clusters of U.S. counties that experienced entry of new auto suppliers. The east-west oriented line shows the existing focus of the industry. The north-south line (roughly oriented with historical U.S. Route 23) shows the focus of new plants in the years studied. Results for the spatial correlation model are compared to a pooled logit model. The estimated spatial autocorrelation coefficient, ρ, is moderately large (0.425 with a standard error of 0.180), however, the results are similar for the two specifications. For example, one of the central results, the coefficient on *Proportion Manufacturing Employment*, is 6.877 (1.039) in the pooled model and 5.307 (1.224) in the spatial model. The magnitudes of the coefficients are difficult to interpret and partial effects were not computed.[68] The signs are generally consistent with expectations.

[68]Wooldridge (2010) and Wang, Iglesias, and Wooldridge (2013) recommend analyzing Average Structural Functions (ASFs) for the heteroscedastic probit (logit) model considered here. Since the weighting matrix, W, does not involve any exogenous variables, the derivatives of the ASFs will be identical to the average partial effects. (See footnote 40 in Section 17.5.2.)

FIGURE 17.7 Counties with New Plants.



17.9 THE BIVARIATE PROBIT MODEL

In Chapter 10, we analyzed a number of different multiple-equation extensions of the linear and generalized regression model. A natural extension of the probit model would be to allow more than one equation, with correlated disturbances, in the same form as the seemingly unrelated regressions model. The general specification for a two-equation model would be

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad y_1 = \mathbf{1}(y_1^* > 0), \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y_2 = \mathbf{1}(y_2^* > 0), \\ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \end{aligned} \tag{17-48}$$

This bivariate probit model is interesting in its own right for modeling the joint determination of two variables, such as doctor and hospital visits in the next example. It also provides the framework for modeling in two common applications. In many cases, a treatment effect, or endogenous influence, takes place in a binary choice context. The **bivariate probit** model provides a specification for analyzing a case in which a probit model contains an endogenous binary variable in one of the equations. In Section 17.6.1 (Examples 17.18 and 17.19), we extended (17-48) to

$$
\begin{aligned}
&T^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \qquad T = \mathbf{1}(T^* > 0),\\
&y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma T + \varepsilon_2, \quad y = \mathbf{1}(y^* > 0),\\
&\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}\Big|\,\mathbf{x}_1, \mathbf{x}_2 \Bigg) \sim N\Bigg[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\Bigg].
\end{aligned}
\tag{17-49}
$$

This model extends the case in Section 17.6.2, where T^* rather than T appears on the right-hand side of the second equation. In Example 17.35, T denotes whether a liberal arts college supports a women's studies program on the campus while y is a binary indicator of whether the economics department provides a gender economics course. A second common application, in which the first equation is an endogenous sampling rule, is another variant of the bivariate probit model:

$$
\begin{aligned}
&S^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad S = 1 \text{ if } S^* > 0, 0 \quad \text{otherwise},\\
&y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \quad \text{otherwise},\\
&\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}\Big|\,\mathbf{x}_1, \mathbf{x}_2 \Bigg) \sim N\Bigg[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\Bigg],\\
&(y, \mathbf{x}_2) \text{ observed only when } S = 1.
\end{aligned}
\tag{17-50}
$$

In Example 17.21, we studied an application in which S is the result of a credit card application (or any sort of loan application) while y_2 is a binary indicator for whether the borrower defaults on the credit account (loan). This is a form of endogenous sampling (in this instance, sampling on unobservables) that has some commonality with the attrition problem that we encountered in Section 17.7.7.

In Section 17.10, we will extend (17-48) to more than two equations. This will allow direct treatment of multiple binary outcomes. It will also allow a more general panel data model for T periods than is provided by the random effects specification.

17.9.1 MAXIMUM LIKELIHOOD ESTIMATION

The bivariate normal cdf is

$$
\text{Prob}(X_1 < x_1, X_2 < x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} \phi_2(z_1, z_2, \rho)dz_1dz_2,
$$

which we denote $\Phi_2(x_1, x_2, \rho)$. The density is[69]

$$
\phi_2(x_1, x_2, \rho) = \frac{e^{-(1/2)(x_1^2 + x_2^2 - 2\rho x_1 x_2)/(1-\rho^2)}}{2\pi(1 - \rho^2)^{1/2}}.
$$

To construct the log likelihood, let $q_{i1} = 2y_{i1} - 1$ and $q_{i2} = 2y_{i2} - 1$. Thus, $q_{ij} = 1$ if $y_{ij} = 1$ and -1 if $y_{ij} = 0$ for $j = 1$ and 2. Now let

$$
z_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_j \quad \text{and} \quad w_{ij} = q_{ij}z_{ij}, \quad j = 1, 2,
$$

and

$$
\rho_{i^*} = q_{i1}q_{i2}\rho.
$$

[69] See Section B.9.

Note the notational convention. The subscript 2 is used to indicate the bivariate normal distribution in the density ϕ_2 and cdf Φ_2. In all other cases, the subscript 2 indicates the variables in the second equation. As before, $\phi(.)$ and $\Phi(.)$ without subscripts denote the univariate standard normal density and cdf.

The probabilities that enter the likelihood function are

$$\text{Prob}(Y_1 = y_{i1}, Y_2 = y_{i2} | \mathbf{x}_1, \mathbf{x}_2) = \Phi_2(w_{i1}, w_{i2}, \rho_{i*}),$$

which accounts for all the necessary sign changes needed to compute probabilities for y's equal to zero and one. Thus,[70]

$$\ln L = \sum_{i=1}^{n} \ln \Phi_2(w_{i1}, w_{i2}, \rho_{i*}).$$

The derivatives of the log likelihood then reduce to

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n} \left(\frac{q_{ij} g_{ij}}{\Phi_2} \right) \mathbf{x}_{ij}, \quad j = 1, 2, \\ \frac{\partial \ln L}{\partial \rho} &= \sum_{i=1}^{n} \frac{q_{i1} q_{i2} \phi_2}{\Phi_2}, \end{aligned} \tag{17-51}$$

where

$$g_{i1} = \phi(w_{i1}) \Phi\left[\frac{w_{i2} - \rho_{i*} w_{i1}}{\sqrt{1 - \rho_{i*}^2}} \right] \tag{17-52}$$

and the subscripts 1 and 2 in g_{i1} are reversed to obtain g_{i2}. Before considering the Hessian, it is useful to note what becomes of the preceding if $\rho = 0$. For $\partial \ln L / \partial \boldsymbol{\beta}_1$, if $\rho = \rho_{i*} = 0$, then g_{i1} reduces to $\phi(w_{i1})\Phi(w_{i2})$, ϕ_2 is $\phi(w_{i1})\phi(w_{i2})$, and Φ_2 is $\Phi(w_{i1})\Phi(w_{i2})$. Inserting these results in (17-51) with q_{i1} and q_{i2} produces (17-20). Because both functions in $\partial \ln L / \partial \rho$ factor into the product of the univariate functions, $\partial \ln L / \partial \rho$ reduces to $\sum_{i=1}^{n} \lambda_{i1} \lambda_{i2}$, where λ_{ij}, $j = 1, 2$, is defined in (17-20). (This result will reappear in the LM statistic shown later.)

The maximum likelihood estimates are obtained by simultaneously setting the three derivatives to zero. The second derivatives are relatively straightforward but tedious. Some simplifications are useful. Let

$$\begin{aligned} \delta_i &= \frac{1}{\sqrt{1 - \rho_{i*}^2}}, \\ v_{i1} &= \delta_i (w_{i2} - \rho_{i*} w_{i1}), \quad \text{so } g_{i1} = \phi(w_{i1}) \Phi(v_{i1}), \\ v_{i2} &= \delta_i (w_{i1} - \rho_{i*} w_{i2}), \quad \text{so } g_{i2} = \phi(w_{i2}) \Phi(v_{i2}). \end{aligned}$$

By multiplying it out, you can show that

$$\delta_i \phi(w_{i1}) \phi(v_{i1}) = \delta_i \phi(w_{i2}) \phi(v_{i2}) = \phi_2.$$

[70] To avoid further ambiguity, and for convenience, the observation subscript will be omitted from $\Phi_2 = \Phi_2(w_{i1}, w_{i2}, \rho_{i*})$ and from $\phi_2 = \phi_2(w_{i1}, w_{i2}, \rho_{i*})$.

Then

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1'} = \sum_{i=1}^{n} \mathbf{x}_{i1}\mathbf{x}_{i1}'\left[\frac{-w_{i1}g_{i1}}{\Phi_2} - \frac{\rho_{i*}\phi_2}{\Phi_2} - \frac{g_{i1}^2}{\Phi_2^2}\right],$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2'} = \sum_{i=1}^{n} q_{i1}q_{i2}\mathbf{x}_{i1}\mathbf{x}_{i2}'\left[\frac{\phi_2}{\Phi_2} - \frac{g_{i1}g_{i2}}{\Phi_2^2}\right],$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}_1 \partial \rho} = \sum_{i=1}^{n} q_{i2}\mathbf{x}_{i1}\frac{\phi_2}{\Phi_2}\left[\rho_{i*}\delta_i v_{i1} - w_{i1} - \frac{g_{i1}}{\Phi_2}\right],$$

$$\frac{\partial^2 \ln L}{\partial \rho^2} = \sum_{i=1}^{n} \frac{\phi_2}{\Phi_2}\left[\delta_i^2\rho_{i*}(1 - \mathbf{w}_i'\mathbf{R}_i^{-1}\mathbf{w}_i) + \delta_i^2 w_{i1}w_{i2} - \frac{\phi_2}{\Phi_2}\right], \qquad \textbf{(17-53)}$$

where $\mathbf{w}_i'\mathbf{R}_i^{-1}\mathbf{w}_i = \delta_i^2(w_{i1}^2 + w_{i2}^2 - 2\rho_{i*}w_{i1}w_{i2})$. (For $\boldsymbol{\beta}_2$, change the subscripts in $\partial^2 \ln L/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1'$ and $\partial^2 \ln L/\partial\boldsymbol{\beta}_1\partial\rho$ accordingly.) The complexity of the second derivatives for this model makes it an excellent candidate for the Berndt et al. estimator of the variance matrix of the maximum likelihood estimator.

Example 17.31 Tetrachoric Correlation

Returning once again to the health care application of Example 17.6 and several others, we now consider a second binary variable,

$$Hospital_{it} = \mathbf{1}(HospVis_{it} > 0).$$

Our previous analyses have focused on

$$Doctor_{it} = \mathbf{1}(DocVis_{it} > 0).$$

A simple bivariate frequency count for these two variables is:

	Hospital		
Doctor	***0***	***1***	***Total***
0	9,715	420	10,135
1	15,216	1,975	17,191
Total	24,931	2,395	27,326

Looking at the very large value in the lower-left cell, one might surmise that these two binary variables (and the underlying phenomena that they represent) are negatively correlated. The usual Pearson product moment correlation would be inappropriate as a measure of this correlation because it is used for continuous variables. Consider, instead, a bivariate probit model,

$$H_{it}^* = \mu_1 + \varepsilon_{1,it}, \quad Hospital_{it} = \mathbf{1}(H_{it}^* > 0),$$
$$D_{it}^* = \mu_2 + \varepsilon_{2,it}, \quad Doctor_{it} = \mathbf{1}(D_{it}^* > 0),$$

where $(\varepsilon_1, \varepsilon_2)$ have a bivariate normal distribution with means (0, 0), variances (1, 1), and correlation ρ. This is the model in (17-48) without independent variables. In this representation, the **tetrachoric correlation**, which is a correlation measure for a pair of binary variables, is precisely the ρ in this model—it is the correlation that would be measured between the underlying continuous variables if they could be observed. This suggests an interpretation of the correlation coefficient in a bivariate probit model—as the conditional tetrachoric correlation.

It also suggests a method of easily estimating the tetrachoric correlation coefficient using a program that is built into nearly all commercial software packages.

Applied to the hospital/doctor data defined earlier, we obtained an estimate of ρ of 0.31106, with an estimated asymptotic standard error of 0.01357. Apparently, our earlier intuition was incorrect.

17.9.2 TESTING FOR ZERO CORRELATION

The Lagrange multiplier statistic is a convenient device for testing for the absence of correlation in this model. Under the null hypothesis that ρ equals zero, the model consists of independent probit equations, which can be estimated separately. Moreover, in the multivariate model, all the bivariate (or multivariate) densities and probabilities factor into the products of the marginals if the correlations are zero, which makes construction of the test statistic a simple matter of manipulating the results of the independent probits. The Lagrange multiplier statistic for testing H_0: $\rho = 0$ in a bivariate probit model is[71]

$$\text{LM} = \frac{\left[\sum_{i=1}^{n} q_{i1}q_{i2}\dfrac{\phi(w_{i1})\phi(w_{i2})}{\Phi(w_{i1})\Phi(w_{i2})}\right]^2}{\sum_{i=1}^{n}\dfrac{[\phi(w_{i1})\phi(w_{i2})]^2}{\Phi(w_{i1})\Phi(-w_{i1})\Phi(w_{i2})\Phi(-w_{i2})}}.$$

As usual, the advantage of the LM statistic is that it obviates computing the bivariate probit model. But the full unrestricted model is now fairly common in commercial software, so that advantage is minor. The likelihood ratio or Wald test can be used with equal ease. To carry out the likelihood ratio test, we note first that if ρ equals zero, then the bivariate probit model becomes two independent univariate probits models. The log likelihood in that case would simply be the sum of the two separate log likelihoods. The test statistic would be

$$\lambda_{\text{LR}} = 2[\ln L_{\text{BIVARIATE}} - (\ln L_1 + \ln L_2)].$$

This would converge to a chi-squared variable with one degree of freedom. The Wald test is carried out by referring

$$\lambda_{\text{WALD}} = \left[\hat{\rho}_{MLE}/\sqrt{\text{Est.Asy.Var}[\hat{\rho}_{MLE}}\right]^2$$

to the chi-squared distribution with one degree of freedom. For 95% significance, the critical value is 3.84 (or one can refer the positive square root to the standard normal critical value of 1.96). Example 17.32 demonstrates.

17.9.3 PARTIAL EFFECTS

There are several partial effects one might want to evaluate in a bivariate probit model.[72] A natural first step would be the derivatives of $\text{Prob}[y_1 = 1, y_2 = 1|\mathbf{x}_1, \mathbf{x}_2]$. These can be deduced from (17-51) by multiplying by Φ_2, removing the sign carrier, q_{ij}, and differentiating with respect to $\mathbf{x}_j$ rather than $\boldsymbol{\beta}_j$. The result is

[71] This is derived in Kiefer (1982).

[72] See Greene (1996b) and Christofides et al. (1997, 2000).

$$\frac{\partial \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho)}{\partial \mathbf{x}_1} = \phi(\mathbf{x}_1'\boldsymbol{\beta}_1)\Phi\left(\frac{\mathbf{x}_2'\boldsymbol{\beta}_2 - \rho\mathbf{x}_1'\boldsymbol{\beta}_1}{\sqrt{1-\rho^2}}\right)\boldsymbol{\beta}_1.$$

Note, however, the bivariate probability, albeit possibly of interest in its own right, is not a conditional mean function. As such, the preceding does not correspond to a regression coefficient or a slope of a conditional expectation.

For convenience in evaluating the conditional mean and its partial effects, we will define a vector $\mathbf{x} = \mathbf{x}_1 \bigcup \mathbf{x}_2$ and let $\mathbf{x}_1'\boldsymbol{\beta}_1 = \mathbf{x}'\boldsymbol{\gamma}_1$. Thus, $\boldsymbol{\gamma}_1$ contains all the nonzero elements of $\boldsymbol{\beta}_1$ and possibly some zeros in the positions of variables in $\mathbf{x}$ that appear only in the other equation; $\boldsymbol{\gamma}_2$ is defined likewise. The bivariate probability is

$$\text{Prob}[y_1 = 1, y_2 = 1|\mathbf{x}] = \Phi_2[\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho].$$

Signs are changed appropriately if the probability of the zero outcome is desired in either case. (See 17-48.) The partial effects of changes in $\mathbf{x}$ on this probability are given by

$$\frac{\partial \Phi_2}{\partial \mathbf{x}} = g_1\boldsymbol{\gamma}_1 + g_2\boldsymbol{\gamma}_2,$$

where g_1 and g_2 are defined in (17-52). The familiar univariate cases will arise if $\rho = 0$, and effects specific to one equation or the other will be produced by zeros in the corresponding position in one or the other parameter vector. There are also some probabilities to consider. The marginal probabilities are given by the univariate probabilities,

$$\text{Prob}[y_j = 1\,|\mathbf{x}] = \Phi(\mathbf{x}'\boldsymbol{\gamma}_j), \quad j = 1, 2,$$

so the analysis of (17-11) and (17-12) applies. One pair of probabilities that might be of interest are

$$\begin{aligned}\text{Prob}[y_1 = 1|y_2 = 1, \mathbf{x}] &= \frac{\text{Prob}[y_1 = 1, y_2 = 1|\mathbf{x}]}{\text{Prob}[y_2 = 1|\mathbf{x}]} \\ &= \frac{\Phi_2(\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho)}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\end{aligned}$$

and similarly for $\text{Prob}[y_2 = 1|y_1 = 1, \mathbf{x}]$. The partial effects for this function are given by

$$\frac{\partial \text{Prob}[y_1 = 1\,|y_2 = 1, \mathbf{x}]}{\partial \mathbf{x}} = \left(\frac{1}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\right)\left[g_1\boldsymbol{\gamma}_1 + \left(g_2 - \Phi_2\frac{\phi(\mathbf{x}'\boldsymbol{\gamma}_2)}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\right)\boldsymbol{\gamma}_2\right].$$

Finally, one might construct the probability function,

$$\text{Prob}(y_1 = 1|y_2, \mathbf{x}) = \frac{\Phi_2[\mathbf{x}'\boldsymbol{\gamma}_1, (2y_2 - 1)\mathbf{x}'\boldsymbol{\gamma}_2, (2y_2 - 1)\rho]}{\Phi[(2y_2 - 1)\mathbf{x}'\boldsymbol{\gamma}_2]}.$$

The derivatives of this function are the same as those presented earlier, with sign changes in several places if $y_2 = 0$ is the argument.

Example 17.32 Bivariate Probit Model for Health Care Utilization

We have extended the bivariate probit model of the previous example by specifying a set of independent variables,

$$\mathbf{x}_i = \textit{Constant}, \textit{Female}_i, \textit{Age}_{it}, \textit{Income}_{it}, \textit{Kids}_{it}, \textit{Education}_{it}, \textit{Married}_{it}.$$

We have specified that the same exogenous variables appear in both equations. (There is no requirement that different variables appear in the equations, nor that a variable be excluded from each equation.) The correct analogy here is to the seemingly unrelated regressions model, not to the linear simultaneous-equations model. Unlike the SUR model of Chapter 10, it is not the case here that having the same variables in the two equations implies that the model can be fit equation by equation, one equation at a time. That result only applies to the estimation of sets of linear regression equations.

Table 17.24 contains the estimates of the parameters of the univariate and bivariate probit models. The tests of the null hypothesis of zero correlation strongly reject the hypothesis that ρ equals zero. The t statistic for ρ based on the full model is 0.2981/0.0139 = 21.446, which is much larger than the critical value of 1.96. For the likelihood ratio test, we compute

$$\lambda_{LR} = 2\{-25{,}285.07 - [-17{,}422.72 + (-8{,}073.604)]\} = 422.508.$$

Once again, the hypothesis is rejected. (The Wald statistic is $21.446^2 = 459.957$.) The LM statistic is 383.953. The coefficient estimates agree with expectations. The income coefficient is statistically significant in the doctor equation, but not in the hospital equation, suggesting, perhaps, that physican visits are at least to some extent discretionary while hospital visits occur on an emergency basis that would be much less tied to income. The table also contains the decomposition of the partial effects for Prob[$y_1 = 1 | y_2 = 1$]. The direct effect is $[g_1/\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)]\boldsymbol{\gamma}_1$ in the definition given earlier. The mean estimate of Prob[$y_1 = 1 | y_2 = 1$] is 0.821285. In the table in Example 17.31, this would correspond to the raw proportion $P(D = 1, H = 1)/P(H = 1) = (1{,}975/27{,}326)/(2{,}395/27{,}326) = 0.8246$.

TABLE 17.24 Estimated Bivariate Probit Model[a]

	Doctor					*Hospital*	
	Model Estimates		*Partial Effects*			*Model Estimates*	
Variable	*Univariate*	*Bivariate*	*Direct*	*Indirect*	*Total*	*Univariate*	*Bivariate*
Constant	−0.1243	−0.1243				−1.3328	−1.3385
	(0.05815)	(0.05814)				(0.08320)	(0.07957)
Female	0.3559	0.3551	0.09650	−0.00724	0.08926	0.1023	0.1050
	(0.01602)	(0.01604)	(0.00500)	(0.00152)	(0.00513)	(0.02195)	(0.02174)
Age	0.01189	0.01188	0.00323	0.00032	0.00291	0.00461	0.00461
	(0.00080)	(0.00080)	(0.00023)	(0.00007)	(0.00024)	(0.00108)	(0.00106)
Income	−0.1324	−0.1337	−0.03632	−0.00306	−0.03939	0.03739	0.04441
	(0.04655)	(0.04628)	(0.01260)	(0.00411)	0.01254)	(0.06329)	(0.05946)
Kids	−0.1521	−0.1523	−0.04140	0.00105	−0.04036	−0.01714	−0.01517
	(0.01833)	(0.01825)	(0.00505)	(0.00177)	(0.00517)	(0.02562)	(0.02570)
Education	−0.01497	−0.01484	−0.00403	0.00151	−0.00252	−0.02196	−0.02191
	(0.00358)	(0.00358)	(0.00010)	(0.00035)	(0.00100)	(0.00522)	(0.00511)
Married	0.07352	0.07351	0.01998	0.00330	0.02328	−0.04824	−0.04789
	(0.02064)	(0.02063)	0.00563)	(0.00192)	(0.00574)	(0.02788)	(0.02777)
ln L	−17422.72	−25285.07				−8073.604	−25285.07

[a]Estimated correlation coefficient = 0.2981 (0.0139).

17.9.4 A PANEL DATA MODEL FOR BIVARIATE BINARY RESPONSE

Extending multiple equation models to accommodate unobserved common effects in panel data settings is straightforward in theory, but complicated in practice. For the bivariate probit case, for example, the natural extension of (17-48) would be

$$y^*_{1,it} = \mathbf{x}'_{1,it}\boldsymbol{\beta}_1 + \varepsilon_{1,it} + \alpha_{1,i} \quad y_{1,it} = 1 \text{ if } y^*_{1,it} > 0, 0 \text{ otherwise},$$
$$y^*_{2,it} = \mathbf{x}'_{2,it}\boldsymbol{\beta}_2 + \varepsilon_{2,it} + \alpha_{2,i} \quad y_{2,it} = 1 \text{ if } y^*_{2,it} > 0, 0 \text{ otherwise},$$
$$\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix} \Big| \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right].$$

The complication will be in how to treat (α_1, α_2). A fixed effects treatment will require estimation of two full sets of dummy variable coefficients, will likely encounter the incidental parameters problem in double measure, and will be complicated in practical terms. As in all earlier cases, the fixed effects case also preempts any specification involving time-invariant variables. It is also unclear in a fixed effects model how any correlation between α_1 and α_2 would be handled. It should be noted that strictly from a consistency standpoint, these considerations are moot. The two equations can be estimated separately, only with some loss of efficiency. The analogous situation would be the seemingly unrelated regressions model in Chapter 10. A random effects treatment (perhaps accommodated with Mundlak's approach of adding the group means to the equations as in Section 17.7.3.b) offers greater promise. If $(\alpha_1, \alpha_2) = (u_1, u_2)$ are normally distributed random effects, with

$$\begin{pmatrix}u_{1,i} \\ u_{2,i}\end{pmatrix} \Big| \mathbf{X}_{1,i}, \mathbf{X}_{2,i} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right],$$

then the unconditional log likelihood for the bivariate probit model,

$$\ln L = \sum_{i=1}^{n} \ln \int_{u_1, u_2} \prod_{t=1}^{T_i} \Phi_2[(w_{1,it}|u_{1,i}), (w_{2,it}|u_{2,i}), \rho^*_{it}]\, f(u_{1,i}, u_{2,i})\, du_{1,i}\, du_{2,i},$$

can be maximized using simulation or quadrature as we have done in previous applications. A possible variation on this specification would specify that the same common effect enter both equations. In that instance, the integration would only be over a single dimension. In this case, there would only be a single new parameter to estimate, σ^2, the variance of the common random effect while ρ would equal one. A refinement on this form of the model would allow the scaling to be different in the two equations by placing u_i in the first equation and θu_i in the second. This would introduce the additional scaling parameter, but ρ would still equal one. This is the formulation of a common random effect used in Heckman's formulation of the dynamic panel probit model in Section 17.7.4.

Example 17.33 Bivariate Random Effects Model for Doctor and Hospital Visits

We will extend the pooled bivariate probit model presented in Example 17.32 by allowing a general random effects formulation, with free correlation between the time-varying components, $(\varepsilon_1, \varepsilon_2)$, and between the time-invariant effects, (u_1, u_2). We used simulation to fit the model. Table 17.25 presents the pooled and random effects estimates. The log-likelihood functions for the pooled and random effects models are −25,285.07 and

TABLE 17.25 Estimated Random Effects Bivariate Probit Model

	Doctor		*Hospital*	
	Pooled	***Random Effects***	***Pooled***	***Random Effects***
Constant	−0.1243	−0.2976	−1.3385	−1.5855
	(0.0581)	(0.0965)	(0.0796)	(0.1085)
Female	0.3551	0.4548	0.1050	0.1280
	(0.0160)	(0.0286)	(0.0217)	(0.0295)
Age	0.0119	0.0199	0.0046	0.0050
	(0.0008)	(0.0013)	(0.0011)	(0.0014)
Income	−0.1337	−0.0106	0.0444	0.1336
	(0.0463)	(0.0640)	(0.0595)	(0.0773)
Kids	−0.1523	−0.1544	−0.0152	0.0216
	(0.0183)	(0.0269)	(0.0257)	(0.0321)
Education	−0.0148	−0.0257	−0.0219	−0.0244
	(0.0036)	(0.0061)	(0.0051)	(0.0068)
Married	0.0735	0.0288	−0.0479	−0.1050
	(0.0206)	(0.0317)	(0.0278)	(0.0355)
$Corr(\varepsilon_1, \varepsilon_2)$	0.2981	0.1501	0.2981	0.1501
$Corr(u_1, u_2)$	0.0000	0.5382	0.0000	0.5382
Std. Dev. u	0.0000	0.2233	0.0000	0.6338
Std. Dev. ε	1.0000	1.0000	1.0000	1.0000

−23,769.67, respectively. Two times the difference is 3,030.76. This would be a chi squared with three degrees of freedom (for the three free elements in the covariance matrix of u_1 and u_2). The 95% critical value is 7.81, so the pooling hypothesis would be rejected. The change in the correlation coefficient from 0.2981 to 0.1501 suggests that we have decomposed the disturbance in the model into a time-varying part and a time-invariant part. The latter seems to be the smaller of the two. Although the time-invariant elements are more highly correlated, their variances are only $0.2233^2 = 0.0499$ and $0.6338^2 = 0.4017$ compared to 1.0 for both ε_1 and ε_2.

17.9.5 A RECURSIVE BIVARIATE PROBIT MODEL

Section 17.6.2 examines a case in which there is an endogenous continuous variable in a binary choice (probit) model. The model is

$$
\begin{aligned}
T &= \mathbf{x}_T'\boldsymbol{\beta}_T + \varepsilon_T, \\
y^* &= \mathbf{x}_y'\boldsymbol{\beta}_y + \gamma T + \varepsilon_y, \quad y = \mathbf{1}(y^* > 0), \\
\begin{pmatrix} \varepsilon_T \\ \varepsilon_y \end{pmatrix} \Big| \mathbf{x}_T, \mathbf{x}_y &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right].
\end{aligned}
$$

The application examined there involved a labor force participation model that was conditioned on an endogenous variable, the non-wife part of family income. In many cases, the endogenous variable in the equation is also binary. In the application we will examine below, the presence of a gender economics course in the economics curriculum

at liberal arts colleges is conditioned on whether or not there is a women's studies program on the campus. The model in this case becomes

$$T^* = \mathbf{x}_T'\boldsymbol{\beta}_T + \varepsilon_T, \qquad T = \mathbf{1}(T^* > 0),$$
$$y^* = \mathbf{x}_y'\boldsymbol{\beta}_y + \gamma T + \varepsilon_y, \qquad y = \mathbf{1}(y^* > 0),$$
$$\begin{pmatrix} \varepsilon_T \\ \varepsilon_y \end{pmatrix} \Big| \mathbf{x}_T, \mathbf{x}_y \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

This model is qualitatively different from the bivariate probit model in (17-48); the first dependent variable, T, appears on the right-hand side of the second equation.[73] This model is a **recursive**, simultaneous-equations model. Surprisingly, the endogenous nature of one of the variables on the right-hand side of the second equation does not need special consideration in formulating the log likelihood.[74] We can establish this fact with the following (admittedly trivial) argument: The term that enters the log likelihood is $P(y = 1, T = 1) = P(y = 1 \,|\, T = 1)P(T = 1)$. Given the model as stated, the marginal probability for $T = 1$ is just $\Phi(\mathbf{x}_T'\boldsymbol{\beta}_T)$, whereas the conditional probability is $\Phi_2(\cdots)/\Phi(\mathbf{x}_T'\boldsymbol{\beta}_T)$. The product returns the bivariate normal probability we had earlier. The other three terms in the log likelihood are derived similarly, which produces:

$$P(y = 1, T = 1) = \Phi(\mathbf{x}_y'\boldsymbol{\beta}_y + \gamma, \mathbf{x}_T'\boldsymbol{\beta}_T, \rho),$$
$$P(y = 1, T = 0) = \Phi(\mathbf{x}_y'\boldsymbol{\beta}_y, -\mathbf{x}_T'\boldsymbol{\beta}_T, -\rho),$$
$$P(y = 0, T = 1) = \Phi[-(\mathbf{x}_y'\boldsymbol{\beta}_y + \gamma), \mathbf{x}_T'\boldsymbol{\beta}_T, -\rho],$$
$$P(y = 0, T = 0) = \Phi(-\mathbf{x}_y'\boldsymbol{\beta}_y, -\mathbf{x}_T'\boldsymbol{\beta}_T, \rho).$$

These terms are exactly those of (17-48) that we obtain just by carrying T in the second equation with no special attention to its endogenous nature. We can ignore the simultaneity in this model and we cannot in the linear regression model. In this instance, we are maximizing the full log likelihood, whereas in the linear regression case, we are manipulating certain sample moments that do not converge to the necessary population parameters in the presence of simultaneity. The log likelihood for this model is

$$\ln L = \sum_{i=1}^{n} \ln \Phi[q_{y,i}(\mathbf{x}_{yi}'\boldsymbol{\beta}_y + \gamma T_i), q_{T,i}(\mathbf{x}_{T,i}'\boldsymbol{\beta}_T), q_{y,i}q_{\mathrm{T},i}\rho],$$

where $q_{y,i} = (2y_i - 1)$ and $q_{T,i} = 2(T_i - 1)$.[75]

---

[73] Eisenberg and Rowe (2006) is another application of this model. In their study, they analyzed the joint (recursive) effect of T = veteran status on y, smoking behavior. The estimator they used was two-stage least squares and GMM. Evans and Schwab (1995), examined below, fit their model by MLE and by 2SLS for comparison.

[74] The model appears in Maddala (1983, p. 123).

[75] If one were armed with only a univariate probit estimator, it might be tempting to mimic 2SLS to estimate this model using a two-step procedure: (1) estimate $\boldsymbol{\beta}_T$ by a probit regression of T on $\mathbf{x}_T$, then (2) estimate $(\boldsymbol{\beta}_y, \gamma)$ by probit regression of y on $[\mathbf{x}_y, \Phi(\mathbf{x}_T'\hat{\boldsymbol{\beta}}_T)]$. This would be an example of a forbidden regression. [See Wooldridge (2010, pp. 267, 594).] The first step works, but the second does not produce consistent estimators of the parameters of interest. The estimating equation at the second is improper—the conditional probability is conditioned on T, not on the probability that T equals one. The temptation should be easy to resist; the recursive bivariate probit model is a built-in procedure in contemporary software.

Example 17.34 The Impact of Catholic School Attendance on High School Performance

Evans and Schwab (1995) considered the effect of Catholic school attendance on two success measures, graduation from high school and entrance to college. Their model is

$$C^* = \mathbf{x}'\boldsymbol{\beta}_C + \varepsilon_C, \qquad C = \mathbf{1}(C^* > 0),$$
$$G^* = \mathbf{x}'\boldsymbol{\beta}_G + \delta R + \gamma C + \varepsilon_G, \quad G = \mathbf{1}(G^* > 0),$$
$$\begin{pmatrix}\varepsilon_C \\ \varepsilon_G\end{pmatrix} \Big| \mathbf{x}_C, \mathbf{x}_G \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right].$$

The binary variables are $C = \mathbf{1}$(Attended Catholic School) and $G = \mathbf{1}$(Graduated from high school). In a second specification of the model, $G = \mathbf{1}$(Entered a four-year college after graduation). Covariates included race, gender, family income, parents' education, family structure, religiosity, and a tenth-grade test score. The parameters of the model are all identified (estimable) whether or not there are variables in the G equation that are not in the C equation (i.e., whether or not there are exclusion restrictions) by dint of the nonlinearity of the structure. However, mindful of the dubiousness of a model that is identified *only* by the nonlinearity, the authors included $\mathrm{R} = \mathbf{1}$ (Student is Catholic) in the equation, to aid identification. That would seem important here, as of more than 30 variables in the equations, only two, the test score and a "% Catholic in County of Residence," were not also dummy variables. (Income was categorized.)

Example 17.35 Gender Economics Courses at Liberal Arts Colleges

Burnett (1997) proposed the following bivariate probit model for the presence of a gender economics course in the curriculum of a liberal arts college:

$$\text{Prob}[G = 1, W = 1 \,|\, \mathbf{x}_G, \mathbf{x}_W] = \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G + \gamma W, \mathbf{x}_W'\beta_W, \rho).$$

The dependent variables in the model are

G = presence of a gender economics course
W = presence of a women's studies program on the campus.

The independent variables in the model are

z_1 = constant term,
z_2 = academic reputation of the college, coded 1(best), 2, . . . to 141,
z_3 = size of the full-time economics faculty, a count,
z_4 = percentage of the economics faculty that are women, proportion (0 to 1),
z_5 = religious affiliation of the college, 0 = no, 1 = yes,
z_6 = percentage of the college faculty that are women, proportion (0 to 1),
$z_7 - z_{10}$ = regional dummy variables, South, Midwest, Northeast, West.

The regressor vectors are

$$\mathbf{x}_G = z_1, z_2, z_3, z_4, z_5 \quad \text{(gender economics course equation)},$$
$$\mathbf{x}_W = z_2, z_5, z_6, z_7 - z_{10} \text{ (women's studies program equation)}.$$

Maximum likelihood estimates of the parameters of Burnett's model were computed by Greene (1998) using her sample of 132 liberal arts colleges; 31 of the schools offer gender economics, 58 have women's studies programs, and 29 have both. (See Appendix Table F17.1.) The estimated parameters are given in Table 17.26. Both bivariate probit and single-equation estimates are given. The estimate of ρ is only 0.1359, with a standard error of 1.2359. The Wald statistic for the test of the hypothesis that ρ equals zero is $(0.1359/1.2539)^2 = 0.011753$. For a single restriction, the critical value from the chi-squared

TABLE 17.26 Estimates of a Recursive Simultaneous Bivariate Probit Model (estimated standard errors in parentheses)

	Single Equation		*Bivariate Probit*	
Variable	***Coefficient***	***Std. Err.***	***Coefficient***	***Std. Err.***
Gender Economics Equation				
Constant	−1.4176	(0.8768)	−1.1911	(2.2155)
AcRep	−0.0114	(0.0036)	−0.0123	(0.0079)
WomStud	1.1095	(0.4699)	0.8835	(2.2603)
EconFac	0.0673	(0.0569)	0.0677	(0.0695)
PctWEcon	2.5391	(0.8997)	2.5636	(1.0144)
Relig	−0.3482	(0.4212)	−0.3741	(0.5264)
Women's Studies Equation				
AcRep	−0.0196	(0.0042)	−0.0194	(0.0057)
PctWFac	1.9429	(0.9001)	1.8914	(0.8714)
Relig	−0.4494	(0.3072)	−0.4584	(0.3403)
South	1.3597	(0.5948)	1.3471	(0.6897)
West	2.3386	(0.6449)	2.3376	(0.8611)
North	1.8867	(0.5927)	1.9009	(0.8495)
Midwest	1.8248	(0.6595)	1.8070	(0.8952)
ρ	0.0000	(0.0000)	0.1359	(1.2539)
ln L	−85.6458		−85.6317	

table is 3.84, so the hypothesis cannot be rejected. The likelihood ratio statistic for the same hypothesis is $2[-85.6317 - (-85.6458)] = 0.0282$, which leads to the same conclusion. The Lagrange multiplier statistic is 0.003807, which is consistent. This result might seem counterintuitive, given the setting. Surely gender economics and women's studies are highly correlated, but this finding does not contradict that proposition. The correlation coefficient measures the correlation between the disturbances in the equations, the omitted factors. That is, ρ measures (roughly) the correlation between the outcomes after the influence of the included factors is accounted for. Thus, the value 0.1359 measures the effect after the influence of women's studies is already accounted for. As discussed in the next paragraph, the proposition turns out to be right. The single most important determinant (at least within this model) of whether a gender economics course will be offered is indeed whether the college offers a women's studies program.

The partial effects in this model are fairly involved, and as before, we can consider several different types. Consider, for example, z_2, academic reputation. There is a direct effect produced by its presence in the gender economics course equation. But there is also an indirect effect. Academic reputation enters the women's studies equation and, therefore, influences the probability that W equals one. Because W appears in the gender economics course equation, this effect is transmitted back to G. The total effect of academic reputation and, likewise, religious affiliation is the sum of these two parts. Consider first the gender economics variable, G. The conditional probability is

$$\begin{aligned}\text{Prob}[G = 1 \,|\, \mathbf{x}_G, \mathbf{x}_W] &= \text{Prob}[G = 1 \,|\, W = 1, \mathbf{x}_G, \mathbf{x}_W]\text{Prob}[W = 1] \\ &\quad + \text{Prob}[G = 1 \,|\, W = 0, \mathbf{x}_G, \mathbf{x}_W]\,\text{Prob}[W = 0] \\ &= \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G + \gamma, \mathbf{x}_w\boldsymbol{\beta}_w, \rho) + \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G, -\mathbf{x}_W'\boldsymbol{\beta}_W, -\rho).\end{aligned}$$

TABLE 17.27 Partial Effects in Gender Economics Model

	Direct	*Indirect*	*Total*	*(Type of Variable, Mean)*
AcRep	−0.0017	−0.0005	−0.0022	(Continuous, 119.242)
PctWEcon	0.3602		0.3602	(Continuous, 0.24787)
EconFac	0.0095		0.0095	(Continuous, 6.74242)
Relig			−0.0716[a]	(Binary, 0.57576)
PctWFac		0.0508	0.0508	(Continuous, 0.35772)

[a]Direct and indirect effects for binary variables are the same.

Derivatives can be computed using our earlier results. We are also interested in the effect of religious affiliation. Because this variable is binary, simply differentiating the probability function may not produce an accurate result. Instead, we would compute the probability with this variable set to one and then zero, and take the difference. Finally, what is the effect of the presence of a women's studies program on the probability that the college will offer a gender economics course? To compute this effect, we would first compute the average treatment effect (see Section 17.6.1) by averaging

$$\text{TE} = \Phi(\mathbf{x}_G'\boldsymbol{\beta}_G + \gamma) - \Phi(\mathbf{x}_G'\boldsymbol{\beta}_G)$$

over the full sample of schools. The average treatment effect for the schools that actually do have a women's studies program would be

$$\text{TET} = \Phi\left[\frac{(\mathbf{x}_G'\boldsymbol{\beta}_G + \gamma) - \rho(\mathbf{x}_W'\boldsymbol{\beta}_W)}{\sqrt{1-\rho^2}}\right] - \Phi\left[\frac{(\mathbf{x}_G'\boldsymbol{\beta}_G) - \rho(\mathbf{x}_W'\boldsymbol{\beta}_W)}{\sqrt{1-\rho^2}}\right]$$

and averaging over the schools that have a women's studies program ($W = 1$).

Table 17.27 presents the estimates of the partial effects and some descriptive statistics for the data. Numerically, the strongest effect appears to be exerted by the representation of women on the faculty; its coefficient of 0.3602 is by far the largest. However, this variable cannot change by a full unit because it is a proportion. An increase of 1% in the presence of women on the economics faculty raises the probability by only 0.0036, which is comparable in scale to the effect of academic reputation. The effect of women on the faculty is likewise fairly small, only 0.000508 per 1% change. As might have been expected, the single most important influence is the presence of a women's studies program. The estimated average treatment effect is 0.1452 (0.3891). The average treatment effect on the schools that have women's studies programs (ATET) is 0.2293 (0.5165). Of course, the raw data would have anticipated this result. Of the 31 schools that offer a gender economics course, 29 also have a women's studies program and only two do not. Note finally that the effect of religious affiliation (whatever it is) is mostly direct.

17.10 A MULTIVARIATE PROBIT MODEL

In principle, a multivariate probit model would simply extend (17-48) to more than two outcome variables just by adding equations. The resulting equation system, again analogous to the seemingly unrelated regressions model, would be

$$
\begin{aligned}
&y_m^* = \mathbf{x}_m'\boldsymbol{\beta}_m + \varepsilon_m, y_m = \mathbf{1}(y_m^* > 0), m = 1, \ldots, M,\\
&E[\varepsilon_m | \mathbf{x}_1, \ldots, \mathbf{x}_M] = 0,\\
&\text{Var}[\varepsilon_m | \mathbf{x}_1, \ldots, \mathbf{x}_M] = 1,\\
&\text{Cov}[\varepsilon_j, \varepsilon_m | \mathbf{x}_1, \ldots, \mathbf{x}_M] = \rho_{jm},\\
&(\varepsilon_1, \ldots, \varepsilon_M) \sim \text{N}_M[\mathbf{0}, \mathbf{R}].
\end{aligned}
$$

The joint probabilities of the observed events, $[y_{i1}, y_{i2} \ldots, y_{iM} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iM}]$, $i = 1, \ldots, n$ that form the basis for the log-likelihood function are the M-variate normal probabilities,

$$L_i = \Phi_M(q_{i1}\mathbf{x}_{i1}'\beta_1, \ldots, q_{iM}\mathbf{x}_{iM}'\beta_M, \mathbf{R}*),$$

where

$$
\begin{aligned}
q_{im} &= 2y_{im} - 1,\\
R_{jm}^* &= q_{ij}q_{im}\rho_{jm}.
\end{aligned}
$$

The practical obstacle to this extension is the evaluation of the M-variate normal integrals and their derivatives. Simulation-based integration using the GHK simulator or simulated likelihood methods (see Chapter 15) allow for estimation of relatively large models. We consider an application in Example 17.36.[76]

The **multivariate probit model** in another form presents a useful extension of the random effects probit model for panel data (Section 17.7.2). If the parameter vectors in all equations are constrained to be equal, we obtain what Bertschek and Lechner (1998) call the "panel probit model",

$$
\begin{aligned}
&y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, y_{it} = \mathbf{1}(y_{it}^* > 0), i = 1, \ldots, n, t = 1, \ldots, T,\\
&(\varepsilon_{i1}, \ldots, \varepsilon_{iT}) \sim \text{N}[\mathbf{0}, \mathbf{R}].
\end{aligned}
$$

The Butler and Moffitt (1982) approach for this model (see Section 17.4.2) has proved useful in many applications. But the underlying assumption that $\text{Cov}[\varepsilon_{it}, \varepsilon_{is}] = \rho$ is a substantive restriction. By treating this structure as a multivariate probit model with the restriction that the coefficient vector be the same in every period, one can obtain a model with free correlations across periods.[77] Hyslop (1999), Bertschek and Lechner (1998), Greene (2004 and Example 17.26), and Cappellari and Jenkins (2006) are applications.

Example 17.36 A Multivariate Probit Model for Product Innovations

Bertschek and Lechner applied the panel probit model to an analysis of the innovation activity of 1,270 German firms observed in five years, 1984–1988, in response to imports and foreign direct investment.[78] The probit model to be estimated is based on the latent regression

[76]Studies that propose improved methods of simulating probabilities include Pakes and Pollard (1989) and especially Börsch-Supan and Hajivassiliou (1993), Geweke (1989), and Keane (1994). A symposium in the November 1994 issue of *Review of Economics and Statistics* presents discussion of numerous issues in specification and estimation of models based on simulation of probabilities. Applications that employ simulation techniques for evaluation of multivariate normal integrals are now fairly numerous. See, for example, Hyslop (1999) (Example 17.26), which applies the technique to a panel data application with $T = 7$. Example 17.23 develops a five-variate application.

[77]By assuming the coefficient vectors are the same in all periods, we actually obviate the normalization that the diagonal elements of R are all equal to one as well. The restriction identifies $T - 1$ relative variances $\rho_{tt} = \sigma_t^2/\sigma_T^2$. This aspect is examined in Greene (2004).

[78]See Bertschek (1995).

$$y_{it}^{*} = \beta_1 + \sum_{k=2}^{8} x_{k,it}\beta_k + \varepsilon_{it},\ y_{it} = \mathbf{1}(y_{it}^{*} > 0),\ i = 1, \ldots, 1{,}270,\ t = 1984, \ldots, 1988,$$

where

y_{it} = 1 if a product innovation was realized by firm i in year t, 0 otherwise,
$x_{2,it}$ = Log of industry sales in DM,
$x_{3,it}$ = Import share = ratio of industry imports to (industry sales plus imports),
$x_{4,it}$ = Relative firm size = ratio of employment in business unit to employment in the industry (times 30),
$x_{5,it}$ = FDI share = ratio of industry foreign direct investment to, (industry sales plus imports),
$x_{6,it}$ = Productivity = ratio of industry value added to industry employment,
$x_{7,it}$ = Raw materials sector = 1 if the firm is in this sector,
$x_{8,it}$ = Investment goods sector = 1 if the firm is in this sector.

The coefficients on import share (β_3) and FDI share (β_5) were of particular interest. The objectives of the study were the empirical investigation of innovation and the methodological development of an estimator that could obviate computing the five-variate normal probabilities necessary for a full maximum likelihood estimation of the model.

Table 17.28 presents the single-equation, pooled probit model estimates.[79] Given the structure of the model, the parameter vector could be estimated consistently with any single period's data. Hence, pooling the observations, which produces a mixture of the estimators, will also be consistent. Given the panel data nature of the data set, however, the conventional standard errors from the pooled estimator are dubious. Because the marginal distribution will produce a consistent estimator of the parameter vector, this is a case in which the cluster estimator (see Section 14.8.2) provides an appropriate asymptotic covariance matrix. Note

TABLE 17.28 Estimated Pooled Probit Model

		Estimated Standard Errors				*Partial Effects*		
Variable	***Estimate***[a]	***SE(1)***[b]	***SE(2)***[c]	***SE(3)***[d]	***SE(4)***[e]	***Partial***	***Std. Err.***	***t ratio***
Constant	−1.960	0.239	0.377	0.230	0.373	—	—	—
ln Sales	0.177	0.0250	0.0375	0.0222	0.0358	0.0683[f]	0.0138	4.96
Rel Size	1.072	0.206	0.306	0.142	0.269	0.413[f]	0.103	4.01
Imports	1.134	0.153	0.246	0.151	0.243	0.437[f]	0.0938	4.66
FDI	2.853	0.467	0.679	0.402	0.642	1.099[f]	0.247	4.44
Prod.	−2.341	1.114	1.300	0.715	1.115	−0.902[f]	0.429	−2.10
Raw Mtl	−0.279	0.0966	0.133	0.0807	0.126	−0.110[g]	0.0503	−2.18
Inv Good	0.188	0.0404	0.0630	0.0392	0.0628	0.0723[g]	0.0241	3.00

[a]Recomputed. Only two digits were reported in the earlier paper.
[b]Obtained from results in Bertschek and Lechner, Table 9.
[c]Based on the Avery et al. (1983) GMM estimator.
[d]Square roots of the diagonals of the negative inverse of the Hessian.
[e]Based on the cluster estimator.
[f]Coefficient scaled by the density evaluated at the sample means.
[g]Computed as the difference in the fitted probability with the dummy variable equal to one, then zero.

[79]We are grateful to the authors of this study who have generously loaned us their data for our continued analysis. The data are proprietary and cannot be made publicly available, unlike the other data sets used in our examples.

TABLE 17.29 Estimated Constrained Multivariate Probit Model (Estimated standard errors in parentheses)

Coefficients	***Full Maximum Likelihood Using GHK Simulator***	***Random Effects*** $\rho = 0.578$ **(0.0189)**
Constant	−1.797** (0.341)	−2.839 (0.534)
ln Sales	0.154** (0.0334)	0.245 (0.052)
Relative size	0.953** (0.160)	1.522 (0.259)
Imports	1.155** (0.228)	1.779 (0.360)
FDI	2.426** (0.573)	3.652 (0.870)
Productivity	−1.578 (1.216)	−2.307 (1.911)
Raw material	−0.292** (0.130)	−0.477 (0.202)
Investment goods	0.224** (0.0605	0.331 (0.095)
log likelihood	−3,522.85	−3,535.55
	Estimated Correlations	
1984, 1985	0.460** (0.0301)	
1984, 1986	0.599** (0.0323)	
1985, 1986	0.643** (0.0308)	
1984, 1987	0.540** (0.0308)	
1985, 1987	0.546** (0.0348)	
1986, 1987	0.610** (0.0322)	
1984, 1988	0.483** (0.0364)	
1985, 1988	0.446** (0.0380)	
1986, 1988	0.524** (0.0355)	
1987, 1988	0.605** (0.0325)	

*Indicates significant at 95% level,
**Indicates significant at 99% level based on a two-tailed test.

that the standard errors in column SE(4) of the table are considerably higher than the uncorrected ones in columns 1 and 3.

The pooled estimator is consistent, so the further development of the estimator is a matter of (1) obtaining a more efficient estimator of $\boldsymbol{\beta}$ and (2) computing estimates of the cross-period correlation coefficients. The FIML estimates of the model can be computed using the GHK simulator. The FIML estimates and the random effects model using the Butler and Moffitt (1982) quadrature method are reported in Table 17.29. The correlations reported are based on the FIML estimates. Also noteworthy in Table 17.30 is the divergence of the random effects estimates from the FIML estimates. The log-likelihood function is −3,535.55 for the random effects model and −3,522.85 for the unrestricted model. The chi-squared statistic for the nine restrictions of the equicorrelation model is 25.4. The critical value from the chi-squared table for nine degrees of freedom is 16.9 for 95% and 21.7 for 99% significance, so the hypothesis of the random effects model would be rejected in favor of the more general panel probit model.

17.11 SUMMARY AND CONCLUSIONS

This chapter has surveyed a large range of techniques for modeling binary choice variables. The model for choice between two alternatives provides the framework for a large proportion of the analysis of microeconomic data. Thus, we have given a very large amount

of space to this model in its own right. In addition, many issues in model specification and estimation that appear in more elaborate settings, such as those we will examine in the next chapter, can be formulated as extensions of the binary choice model of this chapter. Binary choice modeling provides a convenient point to study endogeneity in a nonlinear model, issues of nonresponse in panel data sets, and general problems of estimation and inference with longitudinal data. The binary probit model in particular has provided the laboratory case for theoretical econometricians such as those who have developed methods of bias reduction for the fixed effects estimator in dynamic nonlinear models.

We began the analysis with the fundamental parametric probit and logit models for binary choice. Estimation and inference issues such as the computation of appropriate covariance matrices for estimators and partial effects are considered here. We then examined familiar issues in modeling, including goodness of fit and specification issues such as the distributional assumption, heteroscedasticity, and missing variables. As in other modeling settings, endogeneity of some right-hand variables presents a substantial complication in the estimation and use of nonlinear models such as the probit model. We examined models with endogenous right-hand-side variables, and in two applications, problems of endogenous sampling. The analysis of binary choice with panel data provides a setting to examine a large range of issues that reappear in other applications. We reconsidered the familiar pooled, fixed, and random effects estimator estimators, and found that much of the wisdom obtained in the linear case does not carry over to the nonlinear case. The incidental parameters problem, in particular, motivates a considerable amount of effort to reconstruct the estimators of binary choice models. Finally, we considered some multivariate extensions of the probit model. As before, the models are useful in their own right. Once again, they also provide a convenient setting in which to examine broader issues, such as more detailed models of endogeneity nonrandom sampling, and computation requiring simulation.

Chapter 18 will continue the analysis of discrete choice models with three frameworks: unordered multinomial choice, ordered choice, and models for count data. Most of the estimation and specification issues we have examined in this chapter will reappear in these settings.

Key Terms and Concepts

- Attributes
- Average partial effect
- Binary choice model
- Bivariate probit
- Butler and Moffitt method
- Characteristics
- Choice-based sampling
- Complementary log log model
- Conditional likelihood function
- Control function
- Event count
- Fixed effects model
- Generalized residual
- Gumbel model
- Incidental parameters problem
- Index function model
- Initial conditions
- Interaction effect
- Inverse probability weighted (IPW)
- Latent regression
- Linear probability model (LPM)
- Logit
- Marginal effects
- Maximum simulated likelihood (MSL)
- Method of scoring
- Microeconometrics
- Minimal sufficient statistic
- Multinomial choice
- Multivariate probit model
- Nonresponse bias
- Ordered choice model
- Persistence
- Quadrature
- Qualitative response (QR)
- Random effects model
- Recursive model
- Selection on unobservables
- State dependence
- Tetrachoric correlation
- Unbalanced sample

Exercises

1. A binomial probability model is to be based on the following index function model:

$$y^* = \alpha + \boldsymbol{\beta} d + \varepsilon,$$
$$y = 1, \text{ if } y^* > 0,$$
$$y = 0 \text{ otherwise.}$$

The only regressor, d, is a dummy variable. The data consist of 100 observations that have the following:

		y	
		0	1
d	0	24	28
	1	32	16

Obtain the maximum likelihood estimators of α and β, and estimate the asymptotic standard errors of your estimates. Test the hypothesis that β equals zero by using a Wald test (asymptotic t test) and a likelihood ratio test. Use the probit model and then repeat, using the logit model. Do your results change? (*Hint:* Formulate the log likelihood in terms of α and $\delta = \alpha + \beta$.)

2. Suppose that a linear probability model is to be fit to a set of observations on a dependent variable y that takes values zero and one, and a single regressor x that varies continuously across observations. Obtain the exact expressions for the least squares slope in the regression in terms of the mean(s) and variance of x, and interpret the result.
3. Given the data set

y	1	0	0	1	1	0	0	1	1	1
x	9	2	5	4	6	7	3	5	2	6

estimate a probit model and test the hypothesis that x is not influential in determining the probability that y equals one.
4. Construct the Lagrange multiplier statistic for testing the hypothesis that all the slopes (but not the constant term) equal zero in the binomial logit model. Prove that the Lagrange multiplier statistic is nR^2 in the regression of $(y_i - p)$ on the xs, where p is the sample proportion of 1s.
5. The following hypothetical data give the participation rates in a particular type of recycling program and the number of trucks purchased for collection by 10 towns in a small mid-Atlantic state:

Town	1	2	3	4	5	6	7	8	9	10
Trucks	160	250	170	365	210	206	203	305	270	340
Participation%	11	74	8	87	62	83	48	84	71	79

The town of Eleven is contemplating initiating a recycling program but wishes to achieve a 95% rate of participation. Using a probit model for your analysis,

a. How many trucks would the town expect to have to purchase to achieve its goal? (*Hint:* You can form the log likelihood by replacing y_i with the participation rate (for example, 0.11 for observation 1) and $(1 - y_i)$ with (1 − the rate), in (17-16).
b. If trucks cost $20,000 each, then is a goal of 90% reachable within a budget of $6.5 million? (That is, should they *expect* to reach the goal?)
c. According to your model, what is the marginal value of the 301st truck in terms of the increase in the percentage participation?

6. A data set consists of $n = n_1 + n_2 + n_3$ observations on y and x. For the first n_1 observations, $y = 1$ and $x = 1$. For the next n_2 observations, $y = 0$ and $x = 1$. For the last n_3 observations, $y = 0$ and $x = 0$. Prove that neither (17-18) nor (17-20) has a solution.
7. Prove (17-26).
8. In the panel data models estimated in Section 17.7, neither the logit nor the probit model provides a framework for applying a Hausman test to determine whether fixed or random effects is preferred. Explain. (*Hint:* Unlike our application in the linear model, the incidental parameters problem persists here.)

Application

1. Appendix Table F17.2 provides Fair's (1978) *Redbook* survey on extramarital affairs. The data are described in Application 1 at the end of Chapter 18 and in Appendix F. The variables in the data set are as follows:

id = an identification number,
C = constant, value = 1,
yrb = a constructed measure of time spent in extramarital affairs,
$v1$ = a rating of the marriage, coded 1 to 4,
$v2$ = age, in years, aggregated,
$v3$ = number of years married,
$v4$ = number of children, top coded at 5,
$v5$ = religiosity, 1 to 4, 1 = not, 4 = very,
$v6$ = education, coded 9, 12, 14, 16, 17, 20,
$v7$ = occupation,
$v8$ = husband's occupation,

and three other variables that are not used. The sample contains a survey of 6,366 married women, conducted by *Redbook* magazine. For this exercise, we will analyze, first, the binary variable,

$$A = 1 \text{ if } yrb > 0, 0 \text{ otherwise.}$$

The regressors of interest are v_1 to v_8; however, not all of them necessarily belong in your model. Use these data to build a binary choice model for A. Report all computed results for the model. Compute the partial effects for the variables you choose. Compare the results you obtain for a probit model to those for a logit model. Are there any substantial differences in the results for the two models?

18
MULTINOMIAL CHOICES AND EVENT COUNTS

18.1 INTRODUCTION

Chapter 17 presented most of the econometric issues that arise in analyzing discrete dependent variables, including specification, estimation, inference, and a variety of variations on the basic model. All of these were developed in the context of a model of binary choice, the choice between two alternatives. This chapter will use those results in extending the choice model to three specific settings:

Multinomial Choice: The individual chooses from more than two choices, once again, making the choice that provides the greatest utility. Applications include the choices of political candidates, how to commute to work, which energy supplier to use, what health care plan to choose, where to live, or what brand of car, appliance, or food product to buy.

Ordered Choice: The individual reveals the strength of his or her preferences with respect to a single outcome. Familiar cases involve survey questions about strength of feelings regarding a particular commodity such as a movie, a book, or a consumer product, or self-assessments of social outcomes such as health in general or self-assessed well-being. Although preferences will probably vary continuously in the space of individual utility, the expression of those preferences for purposes of analyses is given in a discrete outcome on a scale with a limited number of choices, such as the typical five-point scale used in marketing surveys.

Event Counts: The observed outcome is a count of the number of occurrences. In many cases, this is similar to the preceding settings in that the "dependent variable" measures an individual choice, such as the number of visits to the physician or the hospital, the number of derogatory reports in one's credit history, or the number of visits to a particular recreation site. In other cases, the event count might be the outcome of some less focused natural process, such as prevalence of a disease in a population or the number of defects per unit of time in a production process, the number of traffic accidents that occur at a particular location per month, the number of customers that arrive at a service point per unit of time, or the number of messages that arrive at a switch per unit of time over the course of a day. In this setting, we will be doing a more familiar sort of regression modeling.

Most of the methodological underpinnings needed to analyze these cases were presented in Chapter 17. In this chapter, we will be able to develop variations on these basic model types that accommodate different choice situations. As in Chapter 17, we are focused on discrete outcomes, so the analysis is framed in terms of models of the probabilities attached to those outcomes.

18.2 MODELS FOR UNORDERED MULTIPLE CHOICES

Some studies of multiple-choice settings include the following:

1. Hensher (1986, 1991), McFadden (1974), and many others have analyzed the travel mode of urban commuters. Hensher and Greene (2007b) analyze commuting between Sydney and Melbourne by a sample of individuals who choose from air, train, bus, and car as the mode of travel.
2. Schmidt and Strauss (1975a, b) and Boskin (1974) have analyzed occupational choice among multiple alternatives.
3. Rossi and Allenby (1999, 2003) studied consumer brand choices in a repeated choice (panel data) model.
4. Train (2009) studied the choice of electricity supplier by a sample of California electricity customers.
5. Michelsen and Madlener (2012) studied homeowners' choice of type of heating appliance to install in a new home.
6. Hensher, Rose, and Greene (2015) analyzed choices of automobile models by a sample of consumers offered a hypothetical menu of features.
7. Lagarde (2013) examined the choice of different sets of guidelines for preventing malaria by a sample of individuals in Ghana.

In each of these cases, there is a single decision based on two or more alternatives. In this and the next section, we will encounter two broad types of multinomial choice sets, **unordered choices** and **ordered choices**. All of the choice sets listed above are unordered. In contrast, a bond rating or a preference scale is, by design, a ranking; that is its purpose. Quite different techniques are used for the two types of models. We will examine models for ordered choices in Section 18.3. This section will examine models for unordered choice sets. General references on the topics discussed here include Hensher, Louviere, and Swait (2000), Train (2009), and Hensher, Rose, and Greene (2015).

18.2.1 RANDOM UTILITY BASIS OF THE MULTINOMIAL LOGIT MODEL

Unordered choice models can be motivated by a random utility model. For the ith consumer faced with J choices, suppose that the utility of choice j is

$$U_{ij} = \mathbf{z}_{ij}'\boldsymbol{\theta} + \varepsilon_{ij}.$$

If the consumer makes choice j in particular, then we assume that U_{ij} is the maximum among the J utilities. Hence, the statistical model is driven by the probability that choice j is made, which is

$$\text{Prob}(U_{ij} > U_{ik}) \qquad \text{for all other } k \neq j.$$

The model is made operational by a particular choice of distribution for the disturbances. As in the binary choice case, two models are usually considered: logit and probit. Because of the need to evaluate multiple integrals of the normal distribution, the probit model has found rather limited use in this setting. The logit model, in contrast, has been widely used in many fields, including economics, market research, politics, finance, and transportation engineering. Let Y_i be a random variable that indicates the choice made.

McFadden (1974a) has shown that if (and only if) the J disturbances are independent and identically distributed with Gumbel (type 1 extreme value) distributions,

$$F(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij})), \tag{18-1}$$

then

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{z}_{ij}'\boldsymbol{\theta})}{\sum_{j=1}^{J}\exp(\mathbf{z}_{ij}'\boldsymbol{\theta})}, \tag{18-2}$$

which leads to what is called the **conditional logit model**. (lt is often labeled the **multinomial logit model**, but this wording conflicts with the usual name for the model discussed in the next section, which differs slightly. Although the distinction turns out to be purely artificial, we will maintain it for the present.)

Utility depends on $\mathbf{z}_{ij}$, which includes aspects specific to the individual as well as to the choices. It is useful to distinguish them. Let $\mathbf{z}_{ij} = [\mathbf{x}_{ij}, \mathbf{w}_i]$ and partition $\boldsymbol{\theta}$ conformably into $[\boldsymbol{\beta}', \boldsymbol{\alpha}']'$. Then $\mathbf{x}_{ij}$ varies across the choices and possibly across the individuals as well. The components of $\mathbf{x}_{ij}$ are called the attributes of the choices. But $\mathbf{w}_i$ contains the **characteristics** of the individual and is, therefore, the same for all choices. If we incorporate this fact in the model, then (18-2) becomes

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\alpha})}{\sum_{j=1}^{J}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\alpha})} = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})\exp(\mathbf{w}_i'\boldsymbol{\alpha})}{\left[\sum_{j=1}^{J}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})\right]\exp(\mathbf{w}_i'\boldsymbol{\alpha})}. \tag{18-3}$$

Terms that do not vary across alternatives—that is, those specific to the individual—fall out of the probability. This is as expected in a model that compares the utilities of the alternatives.

Consider a model of shopping center choice by individuals in various cities that depends on the number of stores at the mall, S_{ij}, the distance from the central business district, D_{ij}, and the shoppers' incomes, I_i, the utilities for three choices would be

$$\begin{aligned}
U_{i1} &= D_{i1}\beta_1 + S_{i1}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i1};\\
U_{i2} &= D_{i2}\beta_1 + S_{i2}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i2};\\
U_{i3} &= D_{i3}\beta_1 + S_{i3}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i3}.
\end{aligned}$$

The choice of alternative 1, for example, reveals that

$$\begin{aligned}
U_{i1} - U_{i2} &= (D_{i1} - D_{i2})\beta_1 + (S_{i1} - S_{i2})\beta_2 + (\varepsilon_{i1} - \varepsilon_{i2}) > 0 \text{ and}\\
U_{i1} - U_{i3} &= (D_{i1} - D_{i3})\beta_1 + (S_{i1} - S_{i3})\beta_2 + (\varepsilon_{i1} - \varepsilon_{i3}) > 0.
\end{aligned}$$

The constant term and *Income* have fallen out of the comparison. The result follows from the fact that the random utility model is ultimately based on comparisons of pairs of alternatives, not the alternatives themselves. Evidently, if the model is to allow individual specific effects, then it must be modified. One method is to create a set of dummy variables (alternative specific constants), A_j, for the choices and multiply each of them by the common $\mathbf{w}$. We then allow the coefficients on these choice invariant

characteristics to vary across the choices instead of the characteristics. Analogously to the linear model, a complete set of interaction terms creates a singularity, so one of them must be dropped. For this example, the matrix of attributes and characteristics would be

$$\mathbf{Z}_i = \begin{bmatrix} S_{i1} & D_{i1} & 1 & 0 & I_i & 0 \\ S_{i2} & D_{i2} & 0 & 1 & 0 & I_i \\ S_{i3} & D_{i3} & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The probabilities for this model would be

$$\text{Prob}(Y_i = j|\mathbf{Z}_i) = \frac{\exp\begin{pmatrix} Stores_{ij}\,\beta_1 + Distance_{ij}\,\beta_2 + \\ A_j\alpha_j + A_j\,Income_i\,\gamma_j \end{pmatrix}}{\sum_{j=1}^{3}\exp\begin{pmatrix} Stores_{ij}\,\beta_1 + Distance_{ij}\,\beta_2 + \\ A_j\alpha_j + A_j\,Income_i\gamma_j \end{pmatrix}}, \alpha_3 = \gamma_3 = 0.$$

18.2.2 THE MULTINOMIAL LOGIT MODEL

To set up the model that applies when data are individual specific, it will help to consider an example. Schmidt and Strauss (1975a, b) estimated a model of occupational choice based on a sample of 1,000 observations drawn from the Public Use Sample for three years: 1960, 1967, and 1970. For each sample, the data for each individual in the sample consist of the following:

1. *Occupation:* 0 = menial, 1 = blue collar, 2 = craft, 3 = white collar, 4 = professional. (Note the slightly different numbering convention, starting at zero, which is standard.)
2. *Characteristics:* constant, education, experience, race, sex.

The multinomial logit model[1] for occupational choice is

$$\text{Prob}(Y_i = j|\mathbf{w}_i) = \frac{\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}{\sum_{j=0}^{4}\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}, \quad j = 0, 1, \ldots, 4. \tag{18-4}$$

(The binomial logit model in Section 17.3 is conveniently produced as the special case of $J = 1$.) The estimated equations provide a set of probabilities for the $J + 1$ choices for a decision maker with characteristics $\mathbf{w}_i$. Before proceeding, we must remove an indeterminacy in the model. If we define $\boldsymbol{\alpha}_j^* = \boldsymbol{\alpha}_j + \mathbf{q}$ for any nonzero vector $\mathbf{q}$, then recomputing the probabilities in (18-4) using $\boldsymbol{\alpha}_j^*$ instead of $\boldsymbol{\alpha}_j$ produces the identical set of probabilities because all the terms involving $\mathbf{q}$ drop out. A convenient normalization that solves the problem is $\boldsymbol{\alpha}_0 = \mathbf{0}$. (This arises because the probabilities sum to one, so only J parameter vectors are needed to determine the $J + 1$ probabilities.) Therefore, the probabilities are

$$\text{Prob}(Y_i = j|\mathbf{w}_i) = P_{ij} = \frac{\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}{1 + \sum_{k=1}^{J}\exp(\mathbf{w}_i'\boldsymbol{\alpha}_k)}, \quad j = 0, 1, \ldots, J. \tag{18-5}$$

[1]Nerlove and Press (1973) is a pioneering study in this literature, also about labor market choices.

The form of the binary choice model examined in Section 17.2 results if $J = 1$. The model implies that we can compute J **log-odds**,

$$\ln\left[\frac{P_{ij}}{P_{ik}}\right] = \mathbf{w}_i'(\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_k) = \mathbf{w}_i'\boldsymbol{\alpha}_j \quad \text{if } k = 0.$$

From the point of view of estimation, it is useful that the odds ratio, P_{ij}/P_{ik}, does not depend on the other choices, which follows from the independence and identical distributions of the random terms in the original model. From a behavioral viewpoint, this fact turns out not to be very attractive. We shall return to this problem in Section 18.2.4.

The log likelihood can be derived by defining, for each individual, $d_{ij} = 1$ if alternative j is chosen by individual i, and 0 if not, for the $J + 1$ possible outcomes. Then, for each i, one and only one of the d_{ij}'s is 1. The log likelihood is a generalization of that for the binomial probit or logit model,

$$\ln L = \sum_{i=1}^{n}\sum_{j=0}^{J} d_{ij} \ln \text{Prob}(Y_i = j|\mathbf{w}_i).$$

The derivatives have the characteristically simple form

$$\frac{\partial \ln L}{\partial \boldsymbol{\alpha}_j} = \sum_{i=}^{n}(d_{ij} - P_{ij})\mathbf{w}_i \quad \text{for } j = 1, \ldots, J.$$

The exact second derivatives matrix has J^2 $K \times K$ blocks,[2]

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\alpha}_j\, \partial \boldsymbol{\alpha}_l'} = -\sum_{i=1}^{n} P_{ij}[\mathbf{1}(j = l) - P_{il}]\mathbf{w}_i\mathbf{w}_i',$$

where $\mathbf{1}(j = l)$ equals 1 if j equals l and 0 if not. Because the Hessian does not involve d_{ij}, these are the expected values, and Newton's method is equivalent to the method of scoring. It is worth noting that the number of parameters in this model proliferates with the number of choices, which is inconvenient because the typical cross section sometimes involves a fairly large number of characteristics.

The coefficients in this model are difficult to interpret. It is tempting to associate $\boldsymbol{\alpha}_j$ with the jth outcome, but that would be misleading. Note that all of the $\boldsymbol{\alpha}_j$'s appear in the denominator of P_{ij}. By differentiating (18-5), we find that the partial effects of the characteristics on the probabilities are

$$\boldsymbol{\delta}_{ij} = \frac{\partial P_{ij}}{\partial \mathbf{w}_i} = P_{ij}\left[\boldsymbol{\alpha}_j - \sum_{k=0}^{J} P_{ik}\,\boldsymbol{\alpha}_k\right] = P_{ij}[\boldsymbol{\alpha}_j - \overline{\boldsymbol{\alpha}}]. \tag{18-6}$$

Therefore, every subvector of $\boldsymbol{\alpha}$ enters every partial effect, both through the probabilities and through the weighted average that appears in $\boldsymbol{\delta}_{ij}$. These values can be computed from the parameter estimates. Although the usual focus is on the coefficient estimates, equation (18-6) suggests that there is at least some potential for confusion. Note, for example, that for any particular w_{ik}, $\partial P_{ij}/\partial w_{ik}$ need not have the same sign as α_{jk}.

[2]If the data were in the form of proportions, such as market shares, then the appropriate log likelihood and derivatives are $\Sigma_i\Sigma_j n_i \ln p_{ij}$ and $\Sigma_i\Sigma_j n_i(p_{ij} - P_{ij})\mathbf{w}_i$, respectively. The terms in the Hessian are multiplied by n_i.

Standard errors can be estimated using the delta method. (See Section 4.6.) For purposes of the computation, let $\boldsymbol{\alpha} = [\mathbf{0}, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \ldots, \boldsymbol{\alpha}_J']'$. We include the fixed $\mathbf{0}$ vector for outcome 0 because although $\boldsymbol{\alpha}_0 = \mathbf{0}$, $\boldsymbol{\delta}_{i0} = -P_{i0}\overline{\boldsymbol{\alpha}}$, which is not $\mathbf{0}$. Note as well that $\text{Asy.Cov}[\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_j] = \mathbf{0}$ for $j = 1, \ldots, J$. Then

$$\text{Asy.Var}[\hat{\boldsymbol{\delta}}_{ij}] = \sum_{l=0}^{J}\sum_{m=0}^{J}\left(\frac{\partial \boldsymbol{\delta}_{ij}}{\partial \boldsymbol{\alpha}_l'}\right) \quad \text{Asy.Cov}[\hat{\boldsymbol{\alpha}}_l', \hat{\boldsymbol{\alpha}}_m']\left(\frac{\partial \boldsymbol{\delta}_{ij}'}{\partial \boldsymbol{\alpha}_m}\right),$$

$$\frac{\partial \boldsymbol{\delta}_{ij}}{\partial \boldsymbol{\alpha}_l'} = [\mathbf{1}(j = l) - P_{il}][P_{ij}\mathbf{I} + \boldsymbol{\delta}_{ij}\mathbf{w}_i'] - P_{ij}[\boldsymbol{\delta}_{il}\mathbf{w}_i'].$$

Finding adequate fit measures in this setting presents the same difficulties as in the binomial models. As before, it is useful to report the log likelihood. If the model contains no covariates and no constant terms, then the log likelihood will be

$$\ln L_c = \sum_{j=0}^{J} n_j \ln\left(\frac{1}{J+1}\right),$$

where n_j is the number of individuals who choose outcome j. If the characteristic vector includes only a constant term, then the restricted log likelihood is

$$\ln L_0 = \sum_{j=0}^{J} n_j \ln\left(\frac{n_j}{n}\right) = \sum_{j=0}^{J} n_j \ln p_j,$$

where p_j is the sample proportion of observations that make choice j. A useful table will give a listing of hits and misses of the prediction rule "predict $Y_i = j$ if $\hat{P}_{ij}$ is the maximum of the predicted probabilities."[3]

Example 18.1 Hollingshead Scale of Occupations

Fair's (1977) study of extramarital affairs is based on a cross section of 601 responses to a survey by *Psychology Today*. One of the covariates is a category of occupations on a seven-point scale, the Hollingshead (1975) scale.[4] The Hollingshead scale is intended to be a measure on a prestige scale, a fact which we'll ignore (or disagree with) for the present. The seven levels on the scale are, broadly,

1. Higher executives,
2. Managers and proprietors of medium-sized businesses,
3. Administrative personnel and owners of small businesses,
4. Clerical and sales workers and technicians,
5. Skilled manual employees,
6. Machine operators and semiskilled employees,
7. Unskilled employees.

Among the other variables in the data set are *Age*, *Sex*, and *Education*. The data are given in Appendix Table F18.1. Table 18.1 lists estimates of a multinomial logit model. (We emphasize that the data are a self-selected sample of *Psychology Today* readers in 1976, so it is unclear what contemporary population would be represented. The following serves as an uncluttered numerical example that readers could reproduce. Note, as well, that at least

[3]It is common for this rule to predict all observations with the same value in an unbalanced sample or a model with little explanatory power. This is not a contradiction of an estimated model with many significant coefficients because the coefficients are not estimated so as to maximize the number of correct predictions.

[4]See, also Bornstein and Bradley (2003).

TABLE 18.1 Estimated Multinomial Logit Model for Occupation (*t* ratios in parentheses)

	α_0	α_1	α_2	α_3	α_4	α_5	α_6
				Parameters			
Constant	0.0	3.1506	2.0156	−1.9849	−6.6539	−15.0779	−12.8919
		(1.14)	(1.28)	(−1.38)	(−5.49)	(−9.18)	(−4.61)
Age	0.0	−0.0244	−0.0361	−0.0123	0.0038	0.0225	0.0588
		(−0.73)	(−1.64)	(−0.63)	(0.25)	(1.22)	(1.92)
Sex	0.0	6.2361	4.6294	4.9976	4.0586	5.2086	5.8457
		(5.08)	(4.39)	(4.82)	(3.98)	(5.02)	(4.57)
Education	0.0	−0.4391	−0.1661	0.0684	0.4288	0.8149	0.4506
		(−2.62)	(−1.75)	(0.79)	(5.92)	(8.56)	(2.92)
				Partial Effects			
Age	−0.0001	−0.0002	−0.0028	−0.0022	0.0006	0.0036	0.0011
	(−.19)	(−0.92)	(−2.23)	(−1.15)	(0.23)	(1.89)	(1.90)
Sex	−0.2149	0.0164	0.0233	0.1041	−0.1264	0.1667	0.0308
	(−4.24)	(1.98)	(1.00)	(2.87)	(−2.15)	(4.20)	(2.35)
Education	−0.0187	−0.0069	−0.0387	−0.0460	0.0278	0.0810	0.0015
	(−2.22)	(−2.31)	(−6.29)	(−5.1)	(2.12)	(8.61)	(0.56)

by some viewpoint, the outcome for this experiment is ordered so the model in Section 18.3 might be more appropriate.) The log likelihood for the model is −770.28141 while that for the model with only the constant terms is −982.20533. The likelihood ratio statistic for the hypothesis that all 18 coefficients of the model are zero is 423.85, which is far larger than the critical value of 28.87. In the estimated parameters, it appears that only gender is consistently statistically significant. However, it is unclear how to interpret the fact that *Education* is significant in some of the parameter vectors and not others. The partial effects give a similarly unclear picture, though in this case, the effect can be associated with a particular outcome. However, we note that the implication of a test of significance of a partial effect in this model is itself ambiguous. For example, *Education* is not significant in the partial effect for outcome 6, though the coefficient on *Education* in α_6 is. This is an aspect of modeling with multinomial choice models that calls for careful interpretation by the model builder. Note that the rows of partial effects sum to zero. The interpretation of this result is that when a characteristic such as age changes, the probabilities change in turn. But they sum to one before and after the change.

Example 18.2 Home Heating Systems

Michelsen and Madlener (2012) studied the preferences of homeowners for adoption of innovative residential heating systems. The analysis was based on a survey of 2,240 German homeowners who installed one of four types of new heating systems*: GAS-ST* = gas-fired condensing boiler with solar thermal support, *OIL-ST* = oil-fired condensing boiler with solar thermal support, *HEAT-P* = heat pump, and *PELLET* = wood pellet-fired boiler. Variables in the model included sociodemographics such as age, income and gender; home characteristics such as size, age, and previous type of heating system; location and some specific characteristics, including preference for energy savings (on a five-point scale), preference for more independence from fossil fuels and, also on a five-point scale, preference for environmental protection. The authors reported only the average partial effects for the many variables (not the estimated coefficients). Two, in particular, were the survey data on

environmental protection and energy independence. They reported the following average partial effects for these two variables:

	GAS-ST	*OIL-ST*	*HEAT-P*	*PELLET*
Environment	0.002	−0.003	−0.022	0.024
Independence	−0.150	−0.043	0.100	0.093

The precise meaning of the changes in the two variables are unclear, as they are five-point scales treated as if they were continuous. Nonetheless, the substitution of technologies away from fossil fuels is suggested in the results. The desire to reduce CO_2 emissions is less obvious in the environmental protection results.[5]

18.2.3 THE CONDITIONAL LOGIT MODEL

When the data consist of choice-specific attributes instead of individual-specific characteristics, the natural model formulation would be

$$\text{Prob}(Y_i = j|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iJ}) = \text{Prob}(Y_i = j|\mathbf{X}_i) = P_{ij} = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{j=1}^{J}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}. \quad \textbf{(18-7)}$$

Here, in accordance with the convention in the literature, we let $j = 1, 2, \ldots, J$ for a total of J alternatives. The model is otherwise essentially the same as the multinomial logit. Even more care will be required in interpreting the parameters, however. Once again, an example will help focus ideas.

In this model, the coefficients are not directly tied to the marginal effects. The marginal effects for continuous variables can be obtained by differentiating (18-7) with respect to a particular $\mathbf{x}_m$ to obtain

$$\frac{\partial P_{ij}}{\partial \mathbf{x}_{im}} = [P_{ij}(\mathbf{1}(j = m) - P_{im})]\boldsymbol{\beta}, \quad m = 1, \ldots, J.$$

It is clear that through its presence in P_{ij} and P_{im}, every attribute set $\mathbf{x}_m$ affects all the probabilities. Hensher (1991) suggests that one might prefer to report elasticities of the probabilities. The effect of attribute k of choice m on P_{ij} would be

$$\frac{\partial \ln P_{ij}}{\partial \ln x_{mk}} = \frac{x_{mk}}{P_{ij}}\frac{\partial P_{ij}}{\partial x_{mk}} = x_{mk}[\mathbf{1}(j = m) - P_{im}]\beta_k.$$

Because there is no ambiguity about the scale of the probability itself, whether one should report the derivatives or the elasticities is largely a matter of taste. There is a striking result in the elasticity; $\partial \ln P_{ij}/\partial \ln x_{mk}$ is not a function of P_{ij}. This is a strong implication of the particular functional form assumed at the outset. It implies the rather peculiar substitution pattern that can be seen in the top panel of Table 18.8, below. We will explore this result in Section 18.2.4. Much of the research on multinomial choice modeling over the past several decades has focused on more general forms (including several that we will examine here) that provide more realistic behavioral results. Some applications are developed in Example 18.3.

[5]The results were extracted from their Table 6, p. 1279.

Estimation of the conditional logit model is simplest by Newton's method or the method of scoring. The log likelihood is the same as for the multinomial logit model. Once again, we define $d_{ij} = 1$ if $Y_i = j$ and 0 otherwise. Then

$$\ln L = \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} \ln \text{Prob}(Y_i = j).$$

Market share and frequency data are common in this setting. If the data are in this form, then the only change needed is, once again, to define d_{ij} as the proportion or frequency.

Because of the simple form of $\ln L$, the gradient and Hessian also have particularly convenient forms: Let $\bar{\mathbf{x}}_i = \sum_{j=1}^{J} P_{ij}\mathbf{x}_{ij}$. Then,

$$\begin{aligned}\frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i),\\ \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= -\sum_{i=1}^{n}\sum_{j=1}^{J} P_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'.\end{aligned} \tag{18-8}$$

The usual problems of fit measures appear here. The log-likelihood ratio and tabulation of actual versus predicted choices will be useful. There are two possible constrained log likelihoods. The model cannot contain a constant term, so the constraint $\boldsymbol{\beta} = \mathbf{0}$ renders all probabilities equal to $1/J$. The constrained log likelihood for this constraint is then $L_c = -n \ln J$. Of course, it is unlikely that this hypothesis would fail to be rejected. Alternatively, we could fit the model with only the $J - 1$ choice-specific constants, which makes the constrained log likelihood the same as in the multinomial logit model, $\ln L_0^* = \sum_j n_j \ln p_j$, where, as before, n_j is the number of individuals who choose alternative j.

We have maintained a distinction between the multinomial logit model based on characteristics of the individual and the conditional logit model based on the attributes of the choices). The distinction is completely artificial. Applications of multinomial choice modeling usually mix the two forms—our example below related to travel mode choice includes attributes of the modes as well as household income. The general form of the multinomial logit model that appears in applications, based on (18-3), would be

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\alpha}_j)}{\sum_{m=1}^{J}\exp(\mathbf{x}_{im}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\alpha}_m)}.$$

18.2.4 THE INDEPENDENCE FROM IRRELEVANT ALTERNATIVES ASSUMPTION

We noted earlier that the odds ratios in the multinomial logit or conditional logit models are independent of the other alternatives. This property is convenient for estimation, but it is not a particularly appealing restriction to place on consumer behavior. An additional consequence, also unattractive, is the peculiar pattern of substitution elasticities that is implied by the multinomial logit form. The property of the logit model whereby P_{ij}/P_{im} is independent of the remaining probabilities, and $\partial \ln P_{ij}/\partial \ln x_{im}$ is not a function of P_{ij}, is called the **independence from irrelevant alternatives (IIA)**.

The independence assumption follows from the initial assumption that the random components of the utility functions are independent and homoscedastic. Later we will discuss several models that have been developed to relax this assumption. Before doing so, we consider a test that has been developed for testing the validity of the assumption. The unconditional probability of choice j in the MNL model is

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{m=1}^{J}\exp(\mathbf{x}_{im}'\boldsymbol{\beta}_m)}.$$

Consider the probability of choice j in a reduced choice set, say in alternatives 1 to $J-1$. This would be

$$\frac{\text{Prob}[Y = j \text{ and } j \in (1, \ldots, J-1)]}{\text{Prob}(j \in (1, \ldots, J-1))} = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{m=1}^{J}\exp(\mathbf{x}_{im}'\boldsymbol{\beta}_m)} \Big/ \frac{\sum_{j=1}^{J-1}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{m=1}^{J}\exp(\mathbf{x}_{im}'\boldsymbol{\beta}_m)}$$
$$= \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{m=1}^{J-1}\exp(\mathbf{x}_{im}'\boldsymbol{\beta}_m)}.$$

This is the same model, with the denominator summed from 1 to $J-1$, instead. The MNL model survives the restriction of the choice set—that is, the parameters of the model would be the same. Hausman and McFadden (1984) suggest that if a subset of the choice set truly is irrelevant, then omitting it from the model altogether will not change parameter estimates systematically. Exclusion of these choices (and the observations that choose them) will be inefficient but will not lead to inconsistency. But if the remaining odds ratios are not truly independent from these alternatives, then the parameter estimators obtained when these choices are excluded will be inconsistent. This observation is the usual basis for Hausman's specification test. The statistic is

$$\chi^2 = (\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_f)'[\hat{\mathbf{V}}_s - \hat{\mathbf{V}}_f]^{-1}(\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_f),$$

where s indicates the estimators based on the restricted subset, f indicates the estimator based on the full set of choices, and $\hat{\mathbf{V}}_s$ and $\hat{\mathbf{V}}_f$ are the respective estimates of the asymptotic covariance matrices. The statistic has a limiting chi-squared distribution with K degrees of freedom. We will examine an application in Example 18.3.

18.2.5 ALTERNATIVE CHOICE MODELS

The multinomial logit form imposes some unattractive restrictions on the pattern of behavior in the choice process. A large variety of alternative models in a long thread of research have been developed that relax the restrictions of the MNL model.[6] Two specific restrictions are the homoscedasticity across choices and individuals of the utility functions and the lack of correlation across the choices. We consider three alternatives to the MNL model. Note it is not simply the distribution at work. Changing the model to a *multinomial probit* model based on the normal distribution, but still independent and homoscedastic, does not solve the problem.

[6]One of the earliest contributions to this literature is Gaudry and Dagenais's (1979) "DOGIT" model that "[D]odges the researcher's dilemma of choosing a priori between a format which commits to IIA restrictions and one which excludes them ...." (p. 105.) The DOGIT functional form is $P_j = (V_j + \lambda_j\Sigma_m V_m)/[(1 + \Sigma_m\lambda_m)\Sigma_m V_m]$, where $V_j = \exp(\mathbf{x}_{ij}'\boldsymbol{\beta})$ and $\lambda_j \geq 0$.

18.2.5.a Heteroscedastic Extreme Value Model

The variance of ε_{ij} in (18-1) is equal to $\pi^2/6$. The heteroscedastic extreme value (HEV) specification developed by Bhat (1995) allows a separate variance,

$$\sigma_j^2 = \pi^2/(6\theta_j^2), \tag{18-9}$$

for each ε_{ij} in (18-1). One of the θ's must be normalized to 1.0 because we can only compare ratios of variances. We can allow heterogeneity across individuals as well as across choices by specifying

$$\theta_{ij} = \theta_j \times \exp(\boldsymbol{\phi}'\mathbf{h}_i). \tag{18-10}$$

[See Salisbury and Feinberg (2010) and Louviere and Swait (2010) for applications of this type of HEV model.] The heteroscedasticity alone interrupts the IIA assumption.

18.2.5.b Multinomial Probit Model

A natural alternative model that relaxes the independence restrictions built into the multinomial logit (MNL) model is the **multinomial probit model (MNP)**. The structural equations of the MNP model are

$$U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}, j = 1, \ldots, J, [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iJ}] \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

The term in the log likelihood that corresponds to the choice of alternative q is

$$\text{Prob}[\text{choice}_{iq}] = \text{Prob}[U_{iq} > U_{ij}, j = 1, \ldots, J, j \neq q].$$

The probability for this occurrence is

$$\text{Prob}[\text{choice}_{iq}] = \text{Prob}[\varepsilon_{i1} - \varepsilon_{iq} < (\mathbf{x}_{iq} - \mathbf{x}_{i1})'\boldsymbol{\beta}, \ldots, \varepsilon_{iJ} - \varepsilon_{iq} < (\mathbf{x}_{iq} - \mathbf{x}_{iJ})'\boldsymbol{\beta}]$$

for the $J - 1$ other choices, which is a cumulative probability from a $(J - 1)$-variate normal distribution. Because we are only making comparisons, one of the variances in this $J - 1$ variate structure—that is, one of the diagonal elements in the reduced $\boldsymbol{\Sigma}$—must be normalized to 1.0. Because only comparisons are ever observable in this model, for identification, $J - 1$ of the covariances must also be normalized, to zero. The MNP model allows an unrestricted $(J - 1) \times (J - 1)$ correlation structure and $J - 2$ free standard deviations for the disturbances in the model. (Thus, a two-choice model returns to the univariate probit model of Section 17.2.3.) For more than two choices, this specification is far more general than the MNL model, which assumes that $\boldsymbol{\Sigma} = (\pi^2/6)\mathbf{I}$. (The scaling is absorbed in the coefficient vector in the MNL model.) It adds the unrestricted correlations to the heteroscedastic model of the previous section.

The greater generality of the multinomial probit is produced by the correlations across the alternatives (and, to a lesser extent, by the possible heteroscedasticity). The distribution itself is a lesser extension. An MNP model that simply substitutes a normal distribution with $\Sigma = \mathbf{I}$ will produce virtually the same results (probabilities and elasticities) as the multinomial logit model. An obstacle to implementation of the MNP model has been the difficulty in computing the multivariate normal probabilities for models with many alternatives.[7] Results on accurate simulation of multinormal integrals

[7]Hausman and Wise (1978) point out that the probit model may not be as impractical as it might seem. First, for J choices, the comparisons implicit in $U_{ij} > U_{im}$ for $m \neq j$ involve the $J - 1$ differences, $\varepsilon_j - \varepsilon_m$. Thus, starting with a J-dimensional problem, we need only consider derivatives of $(J - 1)$-order probabilities. Therefore, for example, a model with four choices requires only the evaluation of trivariate normal integrals, bivariate if only the derivatives of the log likelihood are needed.

using the GHK simulator have made estimation of the MNP model feasible. (See Section 15.6.2.b and a symposium in the November 1994 issue of the *Review of Economics and Statistics*.) Computation is exceedingly time consuming. It is also necessary to ensure that $\boldsymbol{\Sigma}$ remain a positive definite matrix. One way often suggested is to construct the Cholesky decomposition of $\boldsymbol{\Sigma}$, $\mathbf{LL}'$, where $\mathbf{L}$ is a lower triangular matrix, and estimate the elements of $\mathbf{L}$. The normalizations and zero restrictions can be imposed by making the last row of the $J \times J$ matrix $\boldsymbol{\Sigma}$ equal $(0, 0, \ldots, 1)$ and using $\mathbf{LL}'$ to create the upper $(J-1) \times (J-1)$ matrix. The additional normalization restriction is obtained by imposing $\mathbf{L}_{11} = 1$.

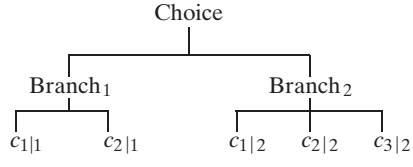
The identification restrictions in Σ needed to identify the model can appear in different places. For example, it is arbitrary which alternative provides the numeraire, and any other row of $\boldsymbol{\Sigma}$ can be normalized. One consequence is that it is not possible to compare directly the estimated coefficient vectors, $\boldsymbol{\beta}$, in the MNP and MNL models. The substantive differences between estimated models are revealed by the predicted probabilities and the estimated elasticities.

18.2.5.c The Nested Logit Model

One way to relax the homoscedasticity assumption in the conditional logit model that also provides an intuitively appealing structure is to group the alternatives into subgroups that allow the variance to differ across the groups while maintaining the IIA assumption within the groups. This specification defines a **nested logit model**. To fix ideas, it is useful to think of this specification as a two- (or more) level choice problem (although, once again, the model arises as a modification of the stochastic specification in the original conditional logit model, not necessarily as a model of behavior). Suppose, then, that the J alternatives can be divided into B subgroups (branches) such that the choice set can be written

$$[c_1, \ldots, c_J] = [(c_{1|1}, \ldots, c_{J_1|1}), (c_{1|2}, \ldots, c_{J_2|2}) \ldots, (c_{1|B}, \ldots, c_{J_B|B})].$$

Logically, we may think of the choice process as that of choosing among the B choice sets and then making the specific choice within the chosen set. This method produces a tree structure, which for two branches and, say, five choices (twigs) might look as follows:



Suppose as well that the data consist of observations on the attributes of the choices $\mathbf{x}_{ij|b}$ and attributes of the choice sets $\mathbf{z}_{ib}$.

To derive the mathematical form of the model, we begin with the unconditional probability

$$\text{Prob}[twig_j, branch_b] = P_{ijb} = \frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma})}{\sum_{b=1}^{B}\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma})}.$$

Now write this probability as

$$P_{ijb} = P_{ij|b}P_b$$
$$= \left(\frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}{\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}\right)\left(\frac{\exp(\mathbf{z}'_{ib}\boldsymbol{\gamma})}{\sum_{l=1}^{L}\exp(\mathbf{z}'_{ib}\boldsymbol{\gamma})}\right)\frac{\left(\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})\right)\left(\sum_{l=1}^{L}\exp(\mathbf{z}'_{ib}\boldsymbol{\gamma})\right)}{\left(\sum_{l=1}^{L}\sum_{j=1}^{J_l}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma})\right)}.$$

Define the **inclusive value** for the lth branch as

$$IV_{ib} = \ln\left(\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})\right).$$

Then, after canceling terms and using this result, we find

$$P_{ij|b} = \frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}{\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})} \quad \text{and} \quad P_b = \frac{\exp[\tau_b(\mathbf{z}'_{ib}\boldsymbol{\gamma} + IV_{ib})]}{\sum_{b=1}^{B}\exp[\tau_b(\mathbf{z}'_{ib}\boldsymbol{\gamma} + IV_{ib})]}, \qquad \textbf{(18-11)}$$

where the new parameters τ_l must equal 1 to produce the original MNL model. Therefore, we use the restriction $\tau_l = 1$ to recover the conditional logit model, and the preceding equation just writes this model in another form. The nested logit model arises if this restriction is relaxed. The inclusive value coefficients, unrestricted in this fashion, allow the model to incorporate some degree of heteroscedasticity and cross alternative correlation. Within each branch, the IIA restriction continues to hold. The equal variance of the disturbances within the jth branch are now[8]

$$\sigma_b^2 = \frac{\pi^2}{6\tau_b}. \qquad \textbf{(18-12)}$$

With $\tau_j = 1$, this reverts to the basic result for the multinomial logit model. The nested logit model is equivalent to a random utility model with block diagonal covariance matrix. For example, for the four-choice model examined in Example 18.3, the model is equivalent to a RUM with

$$\Sigma = \begin{bmatrix} \sigma_F^2 & 0 & 0 & 0 \\ 0 & \sigma_G^2 & \sigma_G^2\rho & \sigma_G^2\rho \\ 0 & \sigma_G^2\rho & \sigma_G^2 & \sigma_G^2\rho \\ 0 & \sigma_G^2\rho & \sigma_G^2\rho & \sigma_G^2 \end{bmatrix}.$$

As usual, the coefficients in the model are not directly interpretable. The derivatives that describe covariation of the attributes and probabilities are

$$\frac{\partial \ln \text{Prob}[\textit{choice} = m, \textit{branch} = b]}{\partial x_k \text{ in } \textit{choice } M \text{ and } \textit{branch } B}$$
$$= \{\mathbf{1}(b = B)[\mathbf{1}(m = M) - P_{M|B}] + \tau_B[\mathbf{1}(b = B) - P_B]P_{M|B}\}\beta_k.$$

[8]See Hensher, Louviere, and Swait (2000). See Greene and Hensher (2002) for alternative formulations of the nested logit model.

The nested logit model has been extended to three and higher levels. The complexity of the model increases rapidly with the number of levels. But the model has been found to be extremely flexible and is widely used for modeling consumer choice in the marketing and transportation literatures, to name a few.

There are two ways to estimate the parameters of the nested logit model. A **limited information**, two-step maximum likelihood approach can be done as follows:

1. Estimate $\boldsymbol{\beta}$ by treating the choice within branches as a simple conditional logit model.
2. Compute the inclusive values for all the branches in the model. Estimate $\boldsymbol{\gamma}$ and the $\boldsymbol{\tau}$ parameters by treating the choice among branches as a conditional logit model with attributes $\mathbf{z}_{ib}$ and I_{ib}.

Because this approach is a two-step estimator, the estimate of the asymptotic covariance matrix of the estimates at the second step must be corrected.[9] For full information maximum likelihood (FIML) estimation of the model, the log likelihood is[10]

$$\ln L = \sum_{i=1}^{n} \ln[\text{Prob}(twig|branch)_i \times \text{Prob}(branch)_i].$$

The information matrix is not block diagonal in $\boldsymbol{\beta}$ and $(\boldsymbol{\gamma}, \boldsymbol{\tau})$, so FIML estimation will be more efficient than two-step estimation. The FIML estimator is now available in several commercial computer packages. (It also solves the problem if efficiently mixing the B different estimators of $\boldsymbol{\beta}$ that are produced by reestimation with each branch.)

To specify the nested logit model, it is necessary to partition the choice set into branches. Sometimes there will be a natural partition, such as in the example given by Maddala (1983) when the choice of residence is made first by community, then by dwelling type within the community. In other instances, however, the partitioning of the choice set is ad hoc and leads to the troubling possibility that the results might be dependent on the branches so defined. (Many studies in this literature present several sets of results based on different specifications of the tree structure.) There is no well-defined testing procedure for discriminating among tree structures, which is a problematic aspect of the model.

Example 18.3 Multinomial Choice Model for Travel Mode

Hensher and Greene[11] report estimates of a model of travel mode choice for travel between Sydney and Melbourne, Australia. The data set contains 210 observations on choice among four travel modes, *air*, *train*, *bus*, and *car*. (See Appendix Table F18.2.) The attributes used for their example were: choice-specific constants; two choice-specific continuous measures; *GC*, a measure of the generalized cost of the travel that is equal to the sum of in-vehicle cost, *INVC*, and a wage-like measure times *INVT*, the amount of time spent traveling; and *TTME*, the terminal time (zero for car); and for the choice between air and the other modes, *HINC*, the household income. A summary of the sample data is given in Table 18.2. The sample is choice based so as to balance it among the four choices—the true population allocation, as shown in the last column of Table 18.2, is dominated by drivers.

The model specified is

$$U_{ij} = \alpha_{air} d_{i,\,air} + \alpha_{train} d_{i,\,train} + \alpha_{bus} d_{i,\,bus} + \beta_G GC_{ij} + \beta_T TTME_{ij} + \gamma_H d_{i,\,air} HINC_i + \varepsilon_{ij},$$

[9]See McFadden (1984).

[10]See Hensher (1986, 1991) and Greene (2007b).

[11]See Greene (2016).

TABLE 18.2 Summary Statistics for Travel Mode Choice Data

	GC	*TTME*	*INVC*	*INVT*	*HINC*	*Number Choosing*	*p*	*True Prop.*
Air	102.648	61.010	85.522	133.710	34.548	58	0.28	0.14
	113.522	46.534	97.569	124.828	41.274			
Train	130.200	35.690	51.338	608.286	34.548	63	0.30	0.13
	106.619	28.524	37.460	532.667	23.063			
Bus	115.257	41.657	33.457	629.462	34.548	30	0.14	0.09
	108.133	25.200	33.733	618.833	29.700			
Car	94.414	0	20.995	573.205	34.548	59	0.28	0.64
	89.095	0	15.694	527.373	42.22			

Note: The upper figure in each cell is the average for all 210 observations. The lower figure is the mean for the observations that made that choice.

where for each j, ε_{ij} has the same independent, type 1 extreme value distribution,

$$F_\varepsilon(\varepsilon_{ij}) = \exp(-exp(-\varepsilon_{ij})),$$

which has variance $\pi^2/6$. The mean of -0.5772 is absorbed in the constants. Estimates of the conditional logit model are shown in Table 18.3. The model was fit with and without the corrections for choice-based sampling. (See Section 17.5.4.) Because the sample shares do not differ radically from the population proportions, the effect on the estimated parameters is fairly modest. Nonetheless, it is apparent that the choice-based sampling is not completely innocent. A cross tabulation of the predicted versus actual outcomes is given in Table 18.4. The predictions are generated by tabulating the integer parts of $m_{jk} = \sum_{i=1}^{210} \hat{p}_{ij} d_{ik}$, $j, k = air, train, bus, car$, where $\hat{p}_{ij}$ is the predicted probability of outcome j for observation i and d_{ik} is the binary variable that indicates if individual i made choice k.

Are the odds ratios *train*/*bus* and *car*/*bus* really independent from the presence of the *air* alternative? To use the Hausman test, we would eliminate choice *air* from the choice set and estimate a three-choice model. Because 58 respondents chose this mode, we would lose 58 observations. In addition, for every data vector left in the sample, the air-specific constant

TABLE 18.3 Parameter Estimates for Multinomial Logit Model

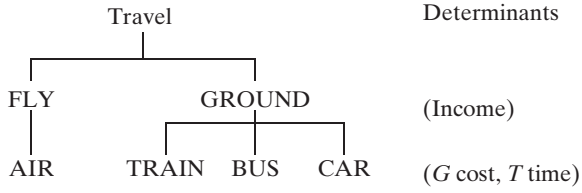
	Unweighted Sample		*Choice-Based Sample Weighting*	
	Estimate	*t Ratio*	*Estimate*	*t Ratio*
β_G	−0.01550	−3.517	−0.01333	−2.711
β_T	−0.09612	−9.207	−0.13405	−5.216
γ_H	0.01329	1.295	−0.00108	−0.097
α_{air}	5.2074	6.684	6.5940	4.075
α_{train}	3.8690	8.731	3.6190	4.317
α_{bus}	3.1632	7.025	3.3218	3.822
Log likelihood at $\boldsymbol{\beta} = 0$	−291.1218		−291.1218	
Log likelihood (sample shares)	−283.7588		−218.9929	
Log likelihood at convergence	−199.1284		−147.5896	

TABLE 18.4 Predicted Choices Based on MNL Model Probabilities (predictions based on choice-based sampling in parentheses)

	Air	***Train***	***Bus***	***Car***	***Total (Actual)***
Air	32 (30)	8 (3)	5 (3)	13 (23)	58
Train	7 (3)	37 (30)	5 (3)	14 (27)	63
Bus	3 (1)	5 (2)	15 (14)	6 (12)	30
Car	16 (5)	13 (5)	6 (3)	25 (45)	59
Total (Predicted)	58 (39)	63 (40)	30 (23)	59 (108)	210

and the interaction, $d_{i,\,air} \times HINC_i$ would be zero for every remaining individual. Thus, these parameters could not be estimated in the restricted model. We would drop these variables. The test would be based on the two estimators of the remaining four coefficients in the model, $[\beta_G, \beta_T, \alpha_{train}, \alpha_{bus}]$. The results for the test are as shown in Table 18.5. The hypothesis that the odds ratios for the other three choices are independent from *air* would be rejected based on these results, as the chi-squared statistic exceeds the critical value.

After IIA was rejected, the authors estimated a nested logit model of the following type:



Note that one of the branches has only a single choice (this is called a "degenerate" branch), so the conditional probability, $P_{j|fly}$, $= P_{air|fly} = 1$. The estimates in Table 18.6 are the simple conditional (multinomial) logit (MNL) model for choice among the four alternatives that was reported earlier. Both inclusive value parameters are constrained (by construction) to equal 1.0000. The FIML estimates are obtained by maximizing the full log likelihood for the nested logit model. In this model,

$$\begin{aligned}
\text{Prob}(choice|branch) &= \text{P}(\alpha_{air}d_{air} + \alpha_{train}d_{train} + \alpha_{bus}d_{bus} + \beta_G GC + \beta_T TTME),\\
\text{Prob}(branch) &= \text{P}(\gamma \text{d}_{air} HINC + \tau_{fly} IV_{fly} + \tau_{ground} IV_{ground}),\\
\text{Prob}(choice, branch) &= \text{Prob}(choice|branch) \times \text{Prob}(branch).
\end{aligned}$$

TABLE 18.5 Results for IIA Test

	Full-Choice Set				***Restricted-Choice Set***			
	β_G	β_T	α_{train}	α_{bus}	β_G	β_T	α_{train}	α_{bus}
Estimate	−0.0155	−0.0961	3.869	3.163	−0.0639	−0.0699	4.464	3.105
	Estimated Asymptotic Covariance Matrix				***Estimated Asymptotic Covariance Matrix***			
β_G	0.0000194				0.000101			
β_T	−0.0000005	0.000109			−0.000013	0.000221		
α_{train}	−0.00060	−0.0038	0.196		−0.00244	−0.00759	0.410	
α_{bus}	−0.00026	−0.0038	0.161	0.203	−0.00113	−0.00753	0.336	0.371

$H = 33.3367$. Critical chi-squared[4] $= 9.488$.

TABLE 18.6 Estimates of a Nested Logit Model (standard errors in parentheses)

Parameter	*Nested Logit*		*Multinomial Logit*	
α_{air}	6.0423	(1.1989)	5.2074	(0.7791)
α_{bus}	4.0963	(0.6152)	3.1632	(0.4503)
α_{train}	5.0646	(0.6620)	3.8690	(0.4431)
β_{GC}	−0.0316	(0.0082)	−0.1550	(0.0044)
β_{TTME}	−0.1126	(0.0141)	−0.0961	(0.0104)
γ_H	0.0153	(0.0094)	0.0133	(0.0103)
τ_{fly}	0.5860	(0.1406)	1.0000	(0.0000)
τ_{ground}	0.3890	(0.1237)	1.0000	(0.0000)
σ_{fly}	2.1886	(0.5255)	1.2825	(0.0000)
σ_{ground}	3.2974	(1.0487)	1.2825	(0.0000)
ln L	−193.6561		−199.1284	

The likelihood ratio statistic for the nesting against the null hypothesis of homoscedasticity is $-2[-199.1284 - (-193.6561)] = 10.945$. The 95% critical value from the chi-squared distribution with two degrees of freedom is 5.99, so the hypothesis is rejected. We can also carry out a Wald test. The asymptotic covariance matrix for the two inclusive value parameters is [0.01977 / 0.009621, 0.01529]. The Wald statistic for the joint test of the hypothesis that $\tau_{fly} = \tau_{ground} = 1$ is

$$W = (0.586 - 1.0 \quad 0.389 - 1.0)\begin{bmatrix} 0.1977 & 0.009621 \\ 0.009621 & 0.01529 \end{bmatrix}^{-1}\begin{pmatrix} 0.586 - 1.0 \\ 0.389 - 1.0 \end{pmatrix} = 24.475.$$

The hypothesis is rejected, once again.

The choice model was reestimated under the assumptions of a heteroscedastic extreme value (HEV) specification. The simplest form allows a separate variance, $\sigma_j^2 = \pi^2/(6\theta_j^2)$, for each ε_{ij} in (18-1). (One of the θs must be normalized to 1.0 because we can only compare ratios of variances.) The results for this model are shown in Table 18.7. This model is less restrictive than the nested logit model. To make them comparable, we note that we found that $\sigma_{air} = \pi/(\tau_{fly}\sqrt{6}) = 2.1886$ and $\sigma_{train} = \sigma_{bus} = \sigma_{car} = \pi/(\tau_{ground}\sqrt{6}) = 3.2974$. The HEV model thus relaxes an additional restriction because it has three free variances whereas the nested logit model has two. But the important degree of freedom is that the HEV model does not impose the IIA assumptions anywhere in the choices, whereas the nested logit does, within each branch. Table 18.7 contains additional results for HEV specifications. In the "Restricted HEV Model," the variance of $\varepsilon_{i,Air}$ is allowed to differ from the others.

A primary virtue of the HEV model, the nested logit model, and other alternative models is that they relax the IIA assumption. This assumption has implications for the cross elasticities between attributes in the different probabilities. Table 18.8 lists the estimated elasticities of the estimated probabilities with respect to changes in the generalized cost variable. Elasticities are computed by averaging the individual sample values rather than computing them once at the sample means. The implication of the IIA assumption can be seen in the table entries. Thus, in the estimates for the multinomial logit (MNL) model, the cross elasticities for each attribute are all equal. In the nested logit model, the IIA property only holds within the branch. Thus, in the first column, the effect of GC of air affects all ground modes equally, whereas the effect of GC for train is the same for bus and car, but different from these two for air. All these elasticities vary freely in the HEV model.

Table 18.9 lists the estimates of the parameters of the multinomial probit and random parameters logit models. The multinomial probit model produces free correlations among

TABLE 18.7 Estimates of a Heteroscedastic Extreme Value Model (standard errors in parentheses)

Parameter	*HEV Model*		*Restricted HEV Model*	
α_{air}	2.228	(1.047)	1.622	(1.247)
α_{train}	3.412	(0.895)	3.942	(0.489)
α_{bus}	3.286	(0.836)	2.866	(0.418)
β_{GC}	−0.026	(0.009)	−0.033	(0.006)
β_{TTME}	−0.071	(0.024)	−0.075	(0.005)
γ	0.028	(0.019)	0.039	(0.021)
θ_{air}	0.472	(0.199)	0.380	(0.095)
θ_{train}	0.886	(0.460)	1.000	(0.000)
θ_{bus}	3.143	(3.551)	1.000	(0.000)
θ_{car}	1.000	(0.000)	1.000	(0.000)
	Implied Standard Deviations			
σ_{air}	2.720	(1.149)		
σ_{train}	1.448	(0.752)		
σ_{bus}	0.408	(0.461)		
σ_{car}	1.283	(0.000)		
ln L	−199.0306		−203.2679	

TABLE 18.8 Estimated Elasticities with Respect to Generalized Cost

	Cost Is That of Alternative			
Effect on	*Air*	*Train*	*Bus*	*Car*
Multinomial Logit				
Air	−1.136	0.498	0.238	0.418
Train	0.456	−1.520	0.238	0.418
Bus	0.456	0.498	−1.549	0.418
Car	0.456	0.498	0.238	−1.061
Nested Logit				
Air	−1.377	0.523	0.523	0.523
Train	0.377	−2.955	1.168	1.168
Bus	0.196	0.604	−3.037	0.604
Car	0.337	1.142	1.142	−1.872
Heteroscedastic Extreme Value				
Air	−1.019	0.410	0.954	0.429
Train	0.395	−3.026	3.184	0.898
Bus	0.282	0.999	−8.161	1.326
Car	0.314	0.708	2.733	−2.589
Multinomial Probit				
Air	−1.092	0.606	0.530	0.290
Train	0.591	−4.078	3.187	1.043
Bus	0.245	1.294	−7.694	1.218
Car	0.255	1.009	2.942	−2.364

TABLE 18.9 Parameter Estimates for Normal-Based Multinomial Choice Models

Parameter	*Multinomial Probit*	*Random Parameters*	
α_{air}	1.799 (1.705)	4.393 (1.698)	
σ_{air}	4.638 (2.251)	4.267 (2.224)	[4.455][a]
α_{train}	4.347 (1.789)	5.649 (1.383	
σ_{train}	1.877 (1.222)	1.097 (1.388)	[1.688][a]
α_{bus}	3.652 (1.421)	4.587 (1.260)	
σ_{bus}	1.000[b]	0.677 (0.958)	[1.450][a]
α_{car}	0.000[b]	0.000[b]	
σ_{car}	1.000[b]	0.000[b]	[1.283][a]
β_G	−0.035 (0.134)	−0.036 (0.014)	
β_T	−0.081 (0.039)	−0.118 (0.022)	
γ_H	0.056 (0.038)	0.047 (0.035)	
ρ_{AT}	0.507 (0.491)	−0.707 (1.268)[c]	
ρ_{AB}	0.457 (0.853)	−0.696 (1.619)[c]	
ρ_{BT}	0.653 (0.346)	−0.014 (2.923)[c]	
ρ_{AC}	0.000[b]	0.000[b]	
ρ_{BC}	0.000[b]	0.000[b]	
ρ_{TC}	0.000[b]	0.000[b]	
$\ln L$	−196.927	−195.646	

[a] Computed as the square root of $(\pi^2/6 + \sigma_j^2)$.
[b] Restricted to this fixed value.
[c] Computed using the delta method.

the choices, which implies an unrestricted 3 × 3 correlation matrix and two free standard deviations.

Table 18.9 reports a variant of the random parameters logit model in which the alternative specific constants are random and freely correlated. The variance for each utility function is $\sigma_j^2 + \theta_j^2$ where σ_j^2 is the contribution of the logit model, which is $\pi^2/6 = 1.645$, and θ_j^2 is the estimated constant specific variance estimated in the random parameters model. The estimates of the specific parameters, θ_j, are given in the table. The estimated model allows unrestricted variation and correlation among the three intercept parameters—this parallels the general specification of the multinomial probit model. The standard deviations and correlations shown for the multinomial probit model are parameters of the distribution of ε_{ij}, the overall randomness in the model. The counterparts in the random parameters model apply to the distributions of the parameters. Thus, the full disturbance in the model in which only the constants are random is $\varepsilon_{iair} + u_{air}$ for air, and likewise for train and bus. It should be noted that in the random parameters model, the disturbances have a distribution that is that of a sum of an extreme value and a normal variable, while in the probit model, the disturbances are normally distributed. With these considerations, the models in each case are comparable and are, in fact, fairly similar.

None of this discussion suggests a preference for one model or the other. The likelihood values are not comparable, so a direct test is precluded. Both relax the IIA assumption, which is a crucial consideration. The random parameters model enjoys a significant practical advantage, as discussed earlier, and also allows a much richer specification of the utility function itself. But, the question still warrants additional study. Both models are making their way into the applied literature.

18.2.6 MODELING HETEROGENEITY

Much of the recent development of choice models has been directed toward accommodating individual heterogeneity. We will consider a few of these, including the mixed logit, which has attracted most of the focus of recent research. The mixed logit model is the extension of the random parameters framework of Sections 15.6–15.10 to multinomial choice models. We will also examine the latent class MNL model.

18.2.6.a The Mixed Logit Model

The **random parameters logit model (RPL)** is also called the **mixed logit model**. [See Revelt and Train (1996); Bhat (1996); Berry, Levinsohn, and Pakes (1995); Jain, Vilcassim, and Chintagunta (1994); Hensher and Greene (2010a); and Hensher, Rose and Greene (2015).] Train's (2009) formulation of the RPL model (which encompasses the others) is a modification of the MNL model. The model is a **random coefficients** formulation. The change to the basic MNL model is the parameter specification in the distribution of the parameters across individuals, i,

$$\beta_{ik} = \beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k + \sigma_k u_{ik}, \tag{18-13}$$

where u_{ik}, $k = 1, \ldots, K$, is multivariate normally distributed with correlation matrix $\mathbf{R}$, σ_k is the standard deviation of the kth distribution, $\beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k$ is the mean of the distribution, and $\mathbf{z}_i$ is a vector of person-specific characteristics (such as age and income) that do not vary across choices. This formulation contains all the earlier models. For example, if $\boldsymbol{\theta}_k = \mathbf{0}$ for all the coefficients and $\sigma_k = 0$ for all the coefficients except for choice-specific constants, then the original MNL model with a normal-logistic mixture for the random part of the MNL model arises (hence the name). (Most of the received applications have $\boldsymbol{\theta}_k = \mathbf{0}$ – that is, homogeneous means of the random parameters.

The model is estimated by simulating the log-likelihood function rather than direct integration to compute the probabilities, which would be infeasible because the mixture distribution composed of the original ε_{ij} and the random part of the coefficient is unknown. For any individual,

$$\text{Prob}[\text{choice } j \,|\, \mathbf{u}_i] = \text{MNL probability} \,|\, \beta_i(\mathbf{u}_i),$$

with all restrictions imposed on the coefficients. The appropriate probability is

$$E_{\boldsymbol{u}}[\text{Prob}(\text{choice } j \,|\, \mathbf{u})] = \int_{u_1, \ldots, u_k} \text{Prob}[\text{choice } j \,|\, \mathbf{u}] f(\mathbf{u}) d\mathbf{u},$$

which can be estimated by simulation, using

$$\text{Est. } E_{\boldsymbol{u}}[\text{Prob}(\text{choice } j \,|\, \mathbf{u})] = \frac{1}{R}\sum_{r=1}^{R} \text{Prob}[\text{choice } j \,|\, \boldsymbol{\beta}_i(\mathbf{u}_{ir})],$$

where $\mathbf{u}_{ir}$ is the rth of R draws for observation i. (There are nkR draws in total. The draws for observation i must be the same from one computation to the next, which can be accomplished by assigning to each individual his or her own seed for the random number generator and restarting it each time the probability is to be computed.) By this method, the log likelihood and its derivatives with respect to $(\beta_k, \boldsymbol{\theta}_k, \sigma_k)$, $k = 1, \ldots, K$ and $\mathbf{R}$ are simulated to find the values that maximize the simulated log likelihood.

The mixed model enjoys two considerable advantages not available in any of the other forms suggested. In a panel data or repeated-choices setting (see Section 18.2.8),

one can formulate a random effects model simply by making the variation in the coefficients time invariant. Thus, the model is changed to

$$U_{ijt} = \mathbf{x}_{ijt}'\boldsymbol{\beta}_i + \varepsilon_{ijt}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, J, \quad t = 1, \ldots, T,$$
$$\beta_{i,k} = \beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k + \sigma_k u_{i,k}.$$

Habit persistence is carried by the time-invariant random effect, u_{ik}. If only the constant terms vary and they are assumed to be uncorrelated, then this is logically equivalent to the familiar random effects model. But much greater generality can be achieved by allowing the other coefficients to vary randomly across individuals and by allowing correlation of these effects.[12] A second degree of flexibility is in (18-13). The random components, u_i, are not restricted to normality. Other distributions that can be simulated will be appropriate when the range of parameter variation consistent with consumer behavior must be restricted, for example to narrow ranges or to positive values (such as based on the lognormal distribution). We will make use of both of these features in the application in Example 18.8.

18.2.6.b A Generalized Mixed Logit Model

The development of functional forms for multinomial choice models begins with the conditional (now usually called the multinomial) logit model that we considered in Section 18.2.3. Subsequent proposals including the multinomial probit and nested logit models (and a wide range of variations on these themes) were motivated by a desire to extend the model beyond the IIA assumptions. These were achieved by allowing correlation across the utility functions or heteroscedasticity such as that in the heteroscedastic extreme value model in (18-10). That issue has been settled in the current generation of multinomial choice models, culminating with the mixed logit model that appears to provide all the flexibility needed to depart from the IIA assumptions. [See McFadden and Train (2000) for a strong endorsement of this idea.]

Recent research in choice modeling has focused on enriching the models to accommodate individual heterogeneity in the choice specification. To a degree, including observable characteristics, such as household income, serves this purpose. In this case, the observed heterogeneity enters the deterministic part of the utility functions. The heteroscedastic HEV model shown in (18-10) moves the observable heterogeneity to the scaling of the utility function instead of the mean. The mixed logit model in (18-13) accommodates both observed and unobserved heterogeneity in the preference parameters. A recent thread of research including Keane (2006), Feibig, et al. (2009), and Greene and Hensher (2010a) has considered functional forms that accommodate individual heterogeneity in both taste parameters (marginal utilities) and overall scaling of the preference structure. Feibig et al.'s **generalized mixed logit model** is

$$U_{i,j} = \mathbf{x}_{ij}'\boldsymbol{\beta}_i + \varepsilon_{ij},$$
$$\boldsymbol{\beta}_i = \sigma_i\boldsymbol{\beta} + [\gamma + \sigma_i(1 - \gamma)]\mathbf{u}_i$$
$$\sigma_i = \exp[\overline{\sigma} + \tau w_i],$$

where $0 \leq \gamma \leq 1$ and w_i is an additional source of unobserved random variation in preferences along with $\mathbf{u}_i$. In this formulation, the weighting parameter, γ, distributes the

[12] A stated choice experiment in which consumers make several choices in sequence about automobile features appears in Hensher, Rose, and Greene (2015).

individual heterogeneity in the preference weights, $\mathbf{u}_i$, and the overall scaling parameter, σ_i. Heterogeneity across individuals in the overall scaling of preference structures is introduced by a nonzero τ while $\overline{\sigma}$ is chosen so that $E_w[\sigma_i] = 1$. Greene and Hensher (2010a) proposed including the observable heterogeneity already in the mixed logit model, and adding it to the scaling parameter as well. Also allowing the random parameters to be correlated (via the nonzero elements in $\mathbf{\Gamma}$) produces a multilayered form of the generalized mixed logit model,

$$\boldsymbol{\beta}_i = \sigma_i[\boldsymbol{\beta} + \mathbf{\Delta}\mathbf{z}_i] + [\gamma + \sigma_i(1 - \gamma)]\mathbf{\Gamma}\mathbf{u}_i$$
$$\sigma_i = \exp[\overline{\sigma} + \boldsymbol{\delta}'\mathbf{h}_i + \tau w_i].$$

Ongoing research has continued to produce refinements that can accommodate realistic forms of individual heterogeneity in the basic multinomial logit framework.

Example 18.4 Using Mixed Logit to Evaluate a Rebate Program

In 2005, Australia led OECD countries and most of the world in per capita greenhouse gas emissions. Among the many federal and state programs aimed at promoting energy efficiency was a water heater rebate program for the New South Wales residential sector. Wasi and Carson (2013) sought to evaluate the impact of the program on Sydney area homeowners' demand for efficient water heaters. The study assessed the effect of the rebate program in shifting existing stocks of electric (primarily coal generated) heaters toward more climate-friendly technologies. Two studies were undertaken: a "revealed preference" (RP) analysis of choices made by recent purchasers of new water heaters and a "stated preference" (SP) study of households that had not replaced their water heaters in the past ten years (and were likely to be in the market in the near future). Broad conclusions drawn from the study included:

> *Our results suggest that households who do not have access to natural gas are more responsive to the rebate program.Without incentive, these households are more likely to replace their electric heater with another electric heater. For those with access to natural gas, many of them would have chosen to replace their electric heater with a gas heater even if the rebate programs had not been in place. These findings are consistent in both ex-post and ex-ante evaluation. From actual purchase data, we also find that the rebate programs appear to work largely on households that deliberately set out to replace their water heater rather than on households that replaced their water heater on an emergency/ urgent basis.* (p. 646.)

Data for the study were obtained through a web-based panel by a major survey research firm. A total of 3,322 respondents out of 9,400 invitees were interested in participating. Access to natural gas is a key determinant of the technology choices that households make. The RP (ex-post) sample included 408 with gas access and 504 without; the SP (ex-ante) sample included 547 with access and 354 without.

Modeling the RP respondents was complicated by the fact that many did not remember the available choice set or could not accurately provide data for the installation cost and running cost. The authors opted for a difference in differences approach based on a simple logit model, as shown in Table 18.10 (which is extracted from their Table 3).[13] (Results are based on a binary logit model for households with no gas access and trinomial logit for those with gas access.)

The SP choice model was based on a mixed logit framework: Attributes of the choices included setup cost net of the rebate, running cost, and a dummy variable for a mail-in rebate.

[13]Wasi and Carson (2013).

TABLE 18.10 Results from Table 6

Estimated Policy Effects on Probability of Switching from Electric for Households with Gas Access

Probability of Switching to	Before Policy	After Policy	Change in Shares
Electric	0.28**	0.19**	−0.09
Gas	0.69**	0.55**	−0.14**
Solar/Heat Pump	0.03**	0.26**	0.23**
Probability of Switching to	Before Policy 2004-2005	2006-Sep 2007	Change in Shares
Electric	0.39**	0.22**	−0.17*
Gas	0.61**	0.74**	0.13
Solar/Heat Pump	0.00	0.04*	0.04*
Effects of Policy on Probability of Switching to			Difference of Changes in Shares
Electric			0.08
Gas			−0.27**
Solar/Heat Pump			0.19**

**,* = Statistically significant at 1%, 5%, respectively.

TABLE 18.11 Results from Table 14

Estimated SP Choice Models

	MNL	***GMNL***		***MM-MNL***			
				Class 1		***Class 2***	
		Mean	StdDev	Mean	StdDev	Mean	StdDev
Cost after rebate/10000	−8.62**	−27.13**	12.53**	−27.3**	14.66**	−16.93**	12.9**
1 if mail-in rebate	0.002	0.01	0.61**	0.01	0.07	−0.28	1.33**
Annual running cost/1000	−3.99**	−17.66**	9.21**	−22.02**	15.42**	−9.35**	6.94**
Class probability				0.66**		0.34**	
τ		0.75**					
γ		−0.81					

**,* = Statistically significant at 1%, 5%, respectively.

The choice experiment included 16 repetitions. The choice set for new installations included electric, gas storage, gas instantaneous, solar, and heat pump. A variety of models were considered: multinomial logit (MNL), mixed logit (MXL), generalized mixed logit (GMXL), latent class logit (LCM), and a mixture of two normals (MM), which is a latent class model in which each class is defined by a mixed logit model. Based on the BIC values, it was determined that the GMXL and MM models were preferred. Some of the results are shown in Table 18.11, which is extracted from their Table 6.

Column 1 of Table 18.11 reports the estimates from the MNL model for the gas access sample.[14] The two cost variables have negative coefficients as expected. The coefficient of

[14]Ibid.

the rebate dummy is positive but not statistically different from zero. The coefficient is large and negative in one of the two classes, suggesting that in this segment, there is substantial disutility attached to filing for the rebate. The average WTP for $1 saved annually is $-3.99 \times 10/-8.62 = 4.62$. Assuming the durability of 15 years, this implies a discount rate of 20%. Column 2 presents the result from the GMNL (generalized mixed logit) model using the full covariance matrix version. The average WTP for $1 saved annually from this model is $6.55, implying a discount rate of 12.8%. Policy evaluations were carried out by simulating the market shares of the different water heater technologies and evaluating the implied impacts on emissions. For households with gas access, the share of electric and gas heaters would reduce by 8% and 11%, respectively. The share of solar/heat pump would increase by 19%. Households with no access to natural gas, while still possessing more electric heaters, are more responsive to the rebate policy (38% reduction in the share of electric heaters). The final step is the evaluation of the cost of the rebate for emission reduction. It was determined that the average costs of carbon reduction from the SP data are $254/ton using a gas access sample and $105/ton from a sample with no access to natural gas. These values were significantly higher than U.S results ($47/ton) but similar to other results from Mexico. Notably, they are much larger than provided for by the NSW climate change fund ($26/ton).

18.2.6.c Latent Classes

We examined the latent class model in Sections 14.15 and 17.7.6. The framework has been used in a number of choice experiments to model heterogeneity semiparametrically. The base framework is

$$\begin{aligned}\text{Prob}(\text{choice}_{it} = j \,|\, \mathbf{X}_{it}, \text{class} = c) &= \frac{\exp(\mathbf{x}_{ijt}'\boldsymbol{\beta}_c)}{\sum_{m=1}^{J}\exp(\mathbf{x}_{imt}'\boldsymbol{\beta}_c)},\\ \text{Prob}(\text{class} = c) &= \pi_c, c = 1, \ldots, C.\end{aligned}$$

The latent class model can usefully be cast as a random parameters specification in which the support of the parameter space is a finite set of points. By this hierarchical structure, the parameter vector, $\boldsymbol{\beta}$, has a discrete distribution, such that

$$\text{Prob}(\boldsymbol{\beta}_i = \boldsymbol{\beta}_c) = \pi_c, 0 \le \pi_c \le 1, \Sigma_c \pi_c = 1.$$

The unconditional choice probability is

$$\text{Prob}(\text{choice}_{it} = j \,|\, \mathbf{X}_{it}) = \sum_{c=1}^{C} \pi_c \frac{\exp(\mathbf{x}_{ijt}'\beta_c)}{\sum_{m=1}^{J}\exp(\mathbf{x}_{imt}'\beta_c)}.$$

Wasi and Carson (2013), in Example 18.4, settled on a latent class specification in which each class defined a mixed logit model. (In Wasi and Carson's specification, $\boldsymbol{\beta}_{i|c} \sim \text{N}[\boldsymbol{\beta}_c, \boldsymbol{\Sigma}_c]$.)

Example 18.5 Latent Class Analysis of the Demand for Green Energy

Ndebele and Marsh (2014) examined preferences for Green Energy among electricity consumers in New Zealand. The study was motivated by a New Zealand study by the Electricity Commission (2008) that reported that nearly 50% of respondents indicated that they would consider the environment when choosing an electricity retailer whilst 17% indicated they would "very seriously" consider switching to a retailer which promotes itself for using renewable resources.

Ndebele and Marsh used a latent class choice modeling framework in which the integration of Environmental Attitude (EA) with stated choices is either direct via the utility function as interactions with the attribute levels of alternatives or as a variable in the class membership

probability model. They identified three latent classes with different preferences for the attributes of electricity suppliers. A typical respondent with a high New Ecological Paradigm (NEP) scale score is willing to pay on average $12.80 more per month on his or her power bill to secure a 10% increase in electricity generated from renewable energy sources compared to respondents with low NEP scores.

An online survey questionnaire was developed to collect the data required for this research. The first part of the survey questionnaire elicited socio-demographic and EA. EA was measured using the 15 items of the NEP scale. The NEP scale is a measure of environmental attitude.[15] The NEP scale is a five-point Likert-type scale consisting of 15 items or statements about the human-environment relationship. The design for the SP experiment is shown in Table 18.12, which is extracted from their Table 2.[16]

An online survey was administered by a market research company in January 2014 to a sample of 224 New Zealand residential electricity bill payers. Stratification was based on age group, gender, and income group. The NEP scores were obtained through online interview. As part of the debriefing, respondents were asked to state the attributes they ignored in choosing their preferred supplier. Attitudinal questions also included questions measuring *awareness of the consequences (AC)* of switching to a supplier that generates most of its electricity from renewables and how far they felt personally responsible—that is, ascription of responsibility (AR)—for reducing CO_2 emissions by switching to a supplier that generates electricity from renewable energy sources. The authors report that "[t]o account for attribute non-attendance in model estimation we coded our data to reflect stated serial non-attendance to specific attributes." Attribute nonattendance is examined in Section 18.2.6d and Example 18.6.

Estimated models are shown in Table 18.13, which is extracted from their Table 13. Based on the MNL model, consumers with moderate NEP scale scores are willing to pay ($10 × 0.0066/0.0255) ≈ $2.60 more per month to secure a 10% increase in electricity generated from renewable sources compared to consumers with a low NEP scale score or low EA. Consumers with strong EA (high NEP scale score) are willing to pay ($10 × 0.0105/0.0255) ≈ $4.10 more per month to secure a 10% increase in electricity generated from renewables compared with customers with low EA. A supplier that is offering a 10% higher prompt payment discount may charge $3.80 more per month than other suppliers *ceteris paribus* and still retain its customers.

TABLE 18.12 Experimental Design: Attributes in Stated Choice Experiment

Attribute	*Description*
Time	= Average wait time for customer service calls (minutes)
Fixed	= Amount of time prices are guaranteed (months)
Discount	= Percent discount for paying bills on time
Rewards	= Presence of a loyalty program (yes/no)
Renewable	= Proportion of electricity generated by green technologies
Ownership	= Proportion of supplier New Zealand owned
Supplier Type	= New or well known company (yes/no)
Bill	= Average monthly bill

[15]See (Dunlap (2008) and Hawcroft and Milfont (2010).

[16]From Ndebele and Marsh (2014).

TABLE 18.13 Estimated Models

Selected Estimates of MNL and Latent Class Model Parameters

Variables	*MNL*	*Latent Class*		
		Class 1	Class 2	Class 3
ASC_{QC}	0.5766***	0.5213***	0.0953	3.2544***
Time (Minutes)	−0.0430***	−0.0378***	−0.0340***	−0.0420
Fixed Term (Months)	0.0046**	0.0057	0.0103**	−0.0033
Discount	0.0096***	0.0054	0.0157***	0.0516***
Loyalty Rewards	0.3691***	0.2698*	0.3607***	0.4891
%Renewable	0.0031	0.0019	0.0079	−0.0042
MNEP × Renewable	0.0066**	0.0075	0.0056	0.0230*
SNEP × Renewable	0.0105***	0.0145*	0.0099**	−0.0003
%NZ Ownership	0.0082***	0.0135***	0.0122***	0.0057
Monthly Power Bill	−0.0255***	−0.0572***	−0.0139***	−0.0147***
Class Probability		0.5374***	0.3479***	0.1147***
Log Likelihood	−2153.4	−1748.41		

*,**, *** Significant at 0.10, 0.05, 0.01, respectively.

18.2.6.d Attribute Nonattendance

In the choice model,

$$U_{ijt} = \alpha_j + \beta_1 x_{ijt,1} + \beta_2 x_{ijt,2} + \ \ldots \ + \varepsilon_{ijt},$$

and the familiar multinomial logit probability, the presence of a nonzero part worth (β) on attribute k suggests a nonzero marginal utility (or disutility) of that attribute for individual i. One possible misspecification of the model would be an assumption of homogeneous attendance. In a given population, one form of heterogeneity might be attribute nonattendance for some (or all) of the attributes.[17] **Attribute nonattendance** (ANA) can represent a rational result of zero marginal utility or it can result from a deliberate strategy to simplify the choice process. These outcomes might be directly observable in a choice experiment in which respondents are specifically queried about them. In Example 18.5, we noted that Ndebele and Marsh solicited this information in the debriefing interview. Nonattendance might only be indirectly observable by behavior that seems to suggest its presence. Consider, for example, a stated choice experiment in which large variation in an attribute such as price appears not to induce switching behavior.

Attribute nonattendance represents a form of individual heterogeneity. Consider the utility function suggested above, which suggests full attendance of both attributes. In a heterogeneous population, there could be (at least) four types of individuals

$$\begin{aligned}
&\text{(Type 1, 2)} \quad U_{ijt} = \alpha_j + \beta_1 x_{ijt,1} + \beta_2 x_{ijt,2} + \ \ldots \ + \varepsilon_{ijt},\\
&\text{(Type 0, 2)} \quad U_{ijt} = \alpha_j + 0 \qquad + \beta_2 x_{ijt,2} + \ \ldots \ + \varepsilon_{ijt},\\
&\text{(Type 1, 0)} \quad U_{ijt} = \alpha_j + \beta_1 x_{ijt,1} + \quad 0 \quad + \ \ldots \ + \varepsilon_{ijt},\\
&\text{(Type 0, 0)} \quad U_{ijt} = \alpha_j + 0 \qquad + \quad 0 \quad + \ \ldots \ + \varepsilon_{ijt.}
\end{aligned}$$

[17]See, for example, Alemu et al. (2013), Hensher, Rose, and Greene (2005, 2012), Hensher and Greene (2010), Hess and Hensher (2012), Hole (2011), and Scarpa, Thiene, and Hensher (2010). The first of these is an extensive survey of the subject.

If the partitioning of the population is observed—Ndebele and Marsh note "we coded our data to reflect stated serial non-attendance to specific attributes"—then the appropriate estimation strategy is to impose the implied zero constraints on $\boldsymbol{\beta}$ selectively, observation by observation. The indicator of which attributes are nonttended by each individual, d_{Type}, becomes part of the "coding" of the data. The log likelihood to be maximized would be

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Bigg[d_{i,Type1,2} \ln L_i\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + d_{i,Type0,2} \ln L_i\begin{pmatrix}0\\ \beta_2\end{pmatrix} + d_{i,Type1,0} \ln L_i\begin{pmatrix}\beta_1\\ 0\end{pmatrix} + d_{i,Type0,0} \ln L_i\begin{pmatrix}0\\ 0\end{pmatrix}\Bigg].$$

(Only one of the indicators, $d_{i,Type}$, equals one.)

One framework for analyzing attribute nonattendance when it is only indirectly observed is a form of latent class model. If the analyst has not directly observed the types, then this suggests a latent class approach to modeling attribute nonattendance. In the model above, this case is simply a missing data application. Since d_{Type} is unobserved, it is replaced in the log likelihood with the probabilities, π_{Type} (which are to be estimated as well) and the model becomes a familiar latent class model,

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{n}\Bigg[\pi_{Type1,2} \ln L_i\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \pi_{Type0,2} \ln L_i\begin{pmatrix}0\\ \beta_2\end{pmatrix} + \pi_{Type1,0} \ln L_i\begin{pmatrix}\beta_1\\ 0\end{pmatrix} + \pi_{Type0,0} \ln L_i\begin{pmatrix}0\\ 0\end{pmatrix}\Bigg].$$

For the example above, the latent class structure would have four classes. For reasons apparent in the listing above, Hensher and Greene (2010) label this the "2^K model." Note that the implied latent class model has two types of restrictions. There is only a single parameter vector in the model — there are cross-class restrictions on the parameters — and there are fixed zeros at different positions in the parameter vector.[18] We will examine an application in Example 18.6.

Example 18.6 Malaria Control During Pregnancy

Lagarde (2013) used the 2^K approach to model attribute nonattendance in a choice experiment about adoption of guidelines for malaria control during pregnancy. The discrete choice experiment was administered to health care providers in Ghana to evaluate their potential resistance to changes in clinical guidelines. The choice task involved whether or not to accept a new set of clinical guidelines. Results showed that less than 3% of the respondents considered all six attributes when choosing between the two hypothetical scenarios proposed, with a majority looking at only one or two attributes. Accounting for ANA strategies affected the magnitude of some of the coefficients and willingness-to-pay estimates.

Guidelines involved six attributes, hence 64 combinations of attendance: The attributes were

1. **Approach:** preventive or curative,
2. **Antimalarial drugs:** SP (Fansidar) or SS-AQ Artesunate-amodiaquine,
3. **Prevalence of anemia for mothers treated with protocol:** 1% or 15%,

[18] A natural extension would be to relax the restriction of equal coefficients across the classes. This is testable.

4. **Prevalence of low birth weight among infants of mothers treated:** 10% or 15%,
5. **Staffing level for the SN clinic:** Under-staffed or adequately staffed,
6. **Salary supplement included in the protocol:** GH. C10, GH. C20.

The author devised a stepwise simplification in the estimation strategy to allow analysis of the excessively large number of classes (64) in the base case model. Accounting for ANA produced fairly large changes in model estimates and estimates of WTP. For examples the estimated coefficients on Anemia Risk and Treatment changed from −0.127 (0.086) to −0.214 (0.016) and from −0.096 (0.077) to −1.840 (0.540). The main results suggested that WTP measures were very sensitive to the presence of ANA. The estimated WTP for the SP drug rose from 8.75 to 24.59 when ANA was considered.[19]

18.2.7 ESTIMATING WILLINGNESS TO PAY

One of the standard applications of choice models is to estimate how much consumers value the attributes of the choices. Recall that we are not able to observe the scale of the utilities in the choice model. However, we can use the marginal utility of income, also scaled in the same unobservable way, to effect the valuation. In principle, we could estimate

$$\begin{aligned}\text{WTP} &= (\text{Marginal Utility of Attribute}/\sigma)/(\text{Marginal Utility of Income}/\sigma)\\ &= \beta_{attribute}/\gamma_{Income},\end{aligned}$$

where σ is the unknown scaling of the utility functions. Note that σ cancels out of the ratio. In our application, for example, we might assess how much consumers would be willing to pay to have shorter waits at the terminal for the public modes of transportation by using

$$\widehat{\text{WTP}}_{time} = -\hat{\beta}_{TIME}/\hat{\gamma}_{Income}.$$

(We use the negative because additional time spent waiting at the terminal provides disutility, as evidenced by its coefficient's negative sign.) In settings in which income is not observed, researchers often use the negative of the coefficient on a cost variable as a proxy for the marginal utility of income. Standard errors for estimates of WTP can be computed using the delta method or the method of Krinsky and Robb. (See Sections 4.6 and 15.3.)

In the basic multinomial logit model, the estimator of WTP is a simple ratio of parameters. In our estimated model in Table 18.3, for example, using the household income coefficient as the numeraire, the estimate of WTP for a shorter wait at the terminal is $-(-0.09612)/0.01329 = 7.23$. The units of measurement must be resolved in this computation, since terminal time is measured in minutes while income is in \$1,000/year. Multiplying this result by 60 minutes/hour and dividing by the equivalent hourly income times 8,760/1,000 gives \$49.52 per hour of waiting time. To compute the estimated asymptotic standard error, for convenience, we first rescaled the terminal time to hours by dividing it by 60 and the income variable to \$/hour by multiplying it by 1,000/8,760. The resulting estimated asymptotic distribution for the estimators is

$$\begin{pmatrix}\hat{\beta}_{TTME}\\ \hat{\gamma}_{HINC}\end{pmatrix} \sim N\left[\begin{pmatrix}-5.76749\\ 0.11639\end{pmatrix}, \begin{pmatrix}0.392365 & 0.00193095\\ 0.00193095 & 0.00808177\end{pmatrix}\right].$$

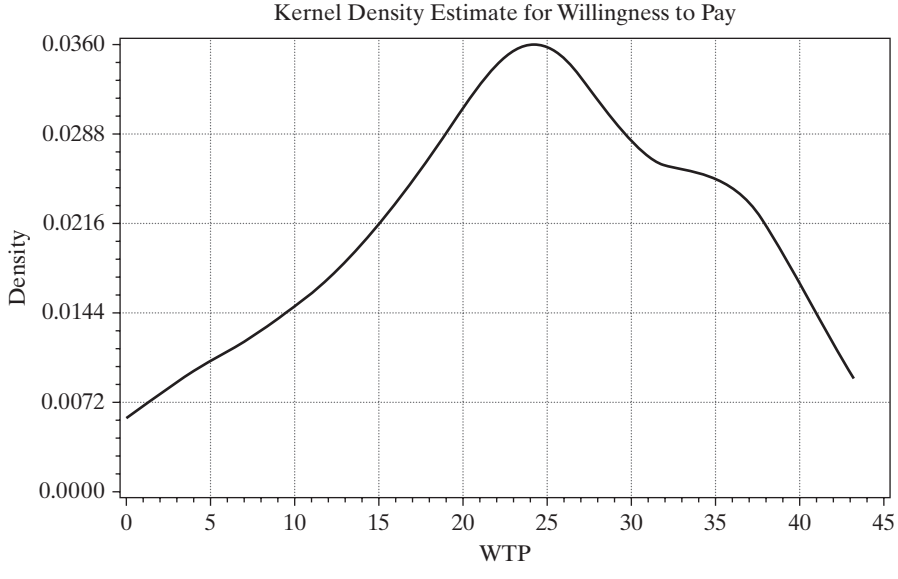
[19]Figures from Lagarde (2013) Tables IV and V.

The derivatives of $\widehat{\text{WTP}}_{TIME} = -\hat{\beta}_{TIME}/\hat{\gamma}_{HINC}$ are $-1/\hat{\gamma}_{HINC}$ for $\hat{\beta}_{TTME}$ and $-\widehat{\text{WTP}}/\hat{\gamma}_{HINC}$ for $\hat{\gamma}_{HINC}$. This provides an estimator of 38.8304 for the standard error. The confidence interval for this parameter would be −26.55 to +125.66. This seems extremely wide. We will return to this issue later.

In the mixed logit model, if either of the coefficients in the computation is random, then the preceding simple computation above will not reveal the heterogeneity in the result. In many studies of WTP using mixed logit models, it is common to allow the utility parameter on the attribute (numerator) to be random and treat the numeraire (income or cost coefficient) as nonrandom. (See Example 18.8.) Using our mode choice application, we refit the model with $\hat{\beta}_{TTME,i} = \hat{\beta}_{TTME} + \hat{\sigma}_{TTME}v_i$ and all other coefficients nonrandom. We then used the method described in Section 15.10 to estimate the mixed logit model and $\text{E}[\hat{\beta}_{TTME,i}|\mathbf{X}_i, choice_i]/\hat{\gamma}_H$ to estimate the expected WTP for each individual in the sample. Income and terminal time were scaled as before. Figure 18.1 displays a kernel estimator of the estimates of WTP_i by this method. The density estimator reveals the heterogeneity in the population of this parameter.

Willingness to pay measures computed as suggested above are ultimately based on a ratio of two asymptotically normally distributed parameter estimators. In general, ratios of normally distributed random variables do not have a finite variance. This often becomes apparent when using the delta method, as it seems previously. A number of writers, notably, Daly, Hess, and Train (2009), have documented the problem of extreme results of WTP computations and why they should be expected. One solution suggested, for example, by Train and Weeks (2005), Sonnier, Ainsle, and Otter (2007), and Scarpa, Thiene, and Train (2008), is to recast the original model in **willingness to pay space**. In

FIGURE 18.1 Estimated Willingness to Pay for Decreased Terminal Time.



the multinomial logit case, this amounts to a trivial reparameterization of the model. Using our application as an example, we would write

$$\begin{aligned} U_{ij} &= \alpha_j + \beta_{GC}GC_i + \gamma_{HINC}[(\beta_{TTME}/\gamma_{HINC})TTME_i + (A_{AIR}HINC_i)] + \varepsilon_{ij} \\ &= \alpha_j + \beta_{GC}GC_i + \gamma_{HINC}[\lambda_{TTME}TTME_i + (A_{AIR}HINC_i)] + \varepsilon_{ij}. \end{aligned}$$

This obviously returns the original model, though in the process, it transforms a linear estimation problem into a nonlinear one. But, in principle, with the model reparameterized in WTP space, we have sidestepped the problem noted earlier; $-\hat{\lambda}_{TTME}$ is the estimator of WTP with no further transformation of the parameters needed. As noted, this will return the numerically identical results for a multinomial logit model. It will not return the identical results for a mixed logit model, in which we write $\hat{\lambda}_{TTME,i} = \hat{\lambda}_{TTME} + \hat{\theta}_{TTME}v_{TTME,i}$. Greene and Hensher (2010b) apply this method to the generalized mixed logit model in Section 18.2.8.

Example 18.7 Willingness to Pay for Renewable Energy

Scarpa and Willis (2010) examined the willingness to pay for renewable energy in the UK with a stated choice experiment. A sample of 1,279 UK households were interviewed about their preferences for heating systems. One analysis in the study considered answers to the following question:

> *"Please imagine that your current heating system needs replacement. I would like you to think about some alternative heating systems for your home. All of the following systems would fully replace your current system. For example, if you had a gas boiler, it would be taken out and replaced by the new system. The rest of your heating system, such as the radiators, would not need to be changed."*

This *primary* experiment included alternative systems such as biomass boilers and supplementary heat pumps with their associated attributes (with space requirements for fuel storage and hot water storage tanks), compared to combi-gas boilers, which deliver central heating and hot water on-demand without the need for hot water storage or fuel storage or the inconvenience associated with tending solid fuel boilers. Notably, in this experiment, the authors did not suggest an opt-out choice. The experiment assumed that the heating system had failed and needed to be replaced. A second experiment, the one discussed below, was based on the *discretionary* case, *"Now I would like you to imagine that your current heating system is functioning completely normally, and to think about supplementing your existing system with an additional system."*

Respondents were asked to choose the type of heating system they would prefer between two alternatives, in four different scenarios. Results for multinomial logit models estimated in preference space and WTP space are shown in Table 18.14 in the results extracted from their Table 5.[20] In addition to the MNL models, they estimated a nested logit model (not shown) and a mixed logit model in WTP space. (We will examine a stated choice experiment based on a mixed logit model in the next application.) Note the two MNL models produce the same log likelihood and related statistics. This is a result of the fact that the WTP space model is a 1:1 transformation of the preference space model. (This is an application of the invariance principle in Section 14.4.5.d.) We can deduce the second model from the first. For example, the numeraire coefficient is the capital cost, equal to −0.3288. Thus, in the WTP space model, the coefficient on solar energy is 0.9312/0.3288 = 2.8316. The coefficient on energy savings is 0.0973/0.3288 = 0.2957 (plus some rounding error) and likewise for the other coefficients in the WTP space model. (This leaves a loose end. The coefficient on capital costs should

[20]Scarpa and Willis (2009).

TABLE 18.14 Estimated Models

Estimated Multiomial Logit Models (1,241 Individuals, 7,280 observations)

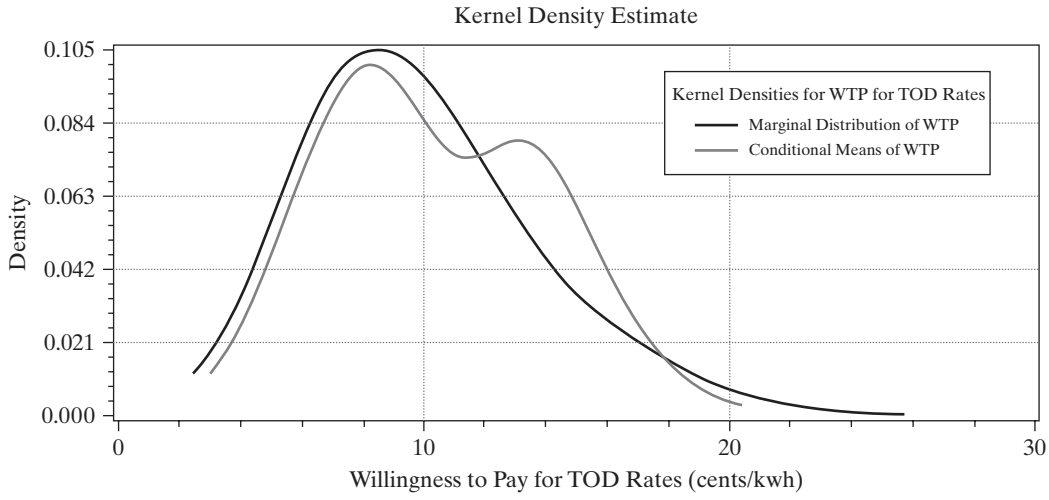
	MNL Preference Space		*MNL WTP-Space*	
	Coefficient	***|t|***	***Coefficient***	***Std. Error***
Solar electricity	0.9312	11.01	2.8316	0.2441
Solar hot water	0.9547	10.84	2.90322	0.2555
Wind turbine	0.4236	5.15	1.2882	0.2408
Capital cost/mean ln(λ)	−0.3288	24.13	−1.1122	0.0415
Friend	−0.0698	1.31	−0.2120	0.1627
Heating engineer	0.0864	1.43	0.2626	0.1834
Both	0.1820	3.52	0.5534	0.1575
Maintenance cost	−0.0303	5.08	−0.0922	0.0184
Energy savings	0.0973	5.20	0.2957	0.0590
Log likelihood	−7328.88		−7328.88	
Rho-square	0.08091		0.08091	

be 1.0000. The authors do not make clear where the 1.1122 comes from.) By adjusting for the units of measurement, the 2.3816 for solar energy translates to a value of 2381.6 GBP. The average installation costs for a 2 kWh solar PV unit in 2008 was 10,638 GBP, 3,904 GBP for a 2 kWh solar hot water unit, and 4,998 GBP for a 1 kWh micro-wind unit. The implied WTP values from the model in Table 5 are 2,381 GBP, 2,903 GBP and 1,288 GBP, respectively. The estimates from the CE data also permitted the evaluation of the relative importance consumers attached to capital in relation to ongoing energy savings. Consumers were WTP 2.91 ± 0.30 GBP in capital costs to reduce annual fuel bills by 1 GBP. The authors conclude that "whilst renewable energy adoption is significantly valued by households, this value is not sufficiently large, for the vast majority of households, to cover the higher capital costs of micro-generation energy technologies, and in relation of annual savings in energy running costs." (p. 135)

18.2.8 PANEL DATA AND STATED CHOICE EXPERIMENTS

The counterpart to panel data in the multinomial choice context is usually the "stated choice experiment," such as the study discussed in Example 18.7. In a stated choice experiment, the analyst (typically) hypothesizes several variations on a general scenario and requests the respondent's preferences among several alternatives each time. In Example 18.8, the sampled individuals are offered a choice of four different electricity suppliers. Each alternative supplier is a specific bundle of rate structure types, contract length, familiarity, and other attributes. The respondent is presented with from 8 to 12 such scenarios, and makes a choice each time. The panel data aspect of this setup is that the same individual makes the choice each time. Any chooser-specific feature, including the underlying preference, is repeated and carried across from scenario to scenario. The MNL model (whether analyzed in preference or WTP space) does not explicitly account for the common underlying characteristics of the individual. The analogous case in the regression and binary choice cases we have already examined would be the pooled model. Several modeling approaches have been used to accommodate the underlying individual heterogeneity in the choice model. The mixed logit model is the most common. Note the third set of results in Figure 18.2 is based on a mixed logit model,

FIGURE 18.2 WTP for Time of Day Rates.



$$\text{Prob}(choice_{it} = j|\mathbf{X}_{it}) = \frac{\exp(\mathbf{x}_{ijt}'\boldsymbol{\beta}_i)}{\Sigma_{m=1}^{J}\exp(\mathbf{x}_{imt}'\boldsymbol{\beta}_i)}, \boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i; i = 1, \ldots, n; t = 1, \ldots, T_i.$$

The random elements in the coefficients are analogous to random effects in the settings we have already examined.

18.2.8.a The Mixed Logit Model

Panel data in the unordered discrete choice setting typically come in the form of sequential choices. Train (2009, Chapter 6) reports an analysis of the site choices of 258 anglers who chose among 59 possible fishing sites for a total of 962 visits. Rossi and Allenby (1999) modeled brand choice for a sample of shoppers who made multiple store trips. The mixed logit model is a framework that allows the counterpart to a random effects model. The random utility model would appear

$$U_{ij,,t} = \mathbf{x}_{ij,t}'\boldsymbol{\beta}_i + \varepsilon_{ij,t},$$

where conditioned on $\boldsymbol{\beta}_i$, a multinomial logit model applies. The random coefficients carry the common effects across choice situations. For example, if the random coefficients include choice-specific constant terms, then the random utility model becomes essentially a random effects model. A modification of the model that resembles Mundlak's correction for the random effects model is

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}^0 + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{u}_i,$$

where, typically, $\mathbf{z}_i$ would contain demographic and socioeconomic information. The scaling matrix, $\boldsymbol{\Gamma}$, allows the random elements of $\boldsymbol{\beta}$ to be correlated; a diagonal $\boldsymbol{\Gamma}$ returns the more familiar case.

The **stated choice experiment** is similar to the repeated choice situation, with a crucial difference. In a stated choice survey, the respondent is asked about his or her preferences over a series of hypothetical choices, often including one or more that are actually available

and others that might not be available (yet). Hensher, Rose, and Greene (2015) describe a survey of Australian commuters who were asked about hypothetical commutation modes in a choice set that included the one they currently took and a variety of proposed alternatives. Revelt and Train (2000) analyzed a stated choice experiment in which California electricity consumers were asked to choose among alternative hypothetical energy suppliers. The advantage of the stated choice experiment is that it allows the analyst to study choice situations over a range of variation of the attributes or a range of choices that might not exist within the observed, actual outcomes. Thus, the original work on the MNL by McFadden et al. concerned survey data on whether commuters would ride a (then-hypothetical) underground train system to work in the San Francisco Bay area. The disadvantage of **stated choice data** is that they are hypothetical. Particularly when they are mixed with **revealed preference data**, the researcher must assume that the same preference patterns govern both types of outcomes. This is likely to be a dubious assumption. One method of accommodating the mixture of underlying preferences is to build different scaling parameters into the model for the stated and revealed preference components of the model. Greene and Hensher (2007) suggested a nested logit model that groups the hypothetical choices in one branch of a tree and the observed choices in another.

18.2.8.b Random Effects and the Nested Logit Model

The mixed logit model in a stated choice experiment setting can be restricted to produce a random effects model. Consider the four-choice example below. The corresponding formulation would be

$$\begin{aligned}
U_{i1,t} &= (\alpha_1 + u_{i1}) + \mathbf{x}'_{i1,t}\boldsymbol{\beta} + \varepsilon_{i1,t},\\
U_{i2,t} &= (\alpha_2 + u_{i2}) + \mathbf{x}'_{i2,t}\boldsymbol{\beta} + \varepsilon_{i2,t},\\
U_{i3,t} &= (\alpha_3 + u_{i3}) + \mathbf{x}'_{i3,t}\boldsymbol{\beta} + \varepsilon_{i3,t},\\
U_{i4,t} &= \qquad\qquad\quad \mathbf{x}'_{i4,t}\boldsymbol{\beta} + \varepsilon_{i4,t}.
\end{aligned}$$

This is simply a restricted version of the random parameters model in which the constant terms are the random parameters. This formulation also provides a way to specify the nested logit model by imposing a further restriction. For example, the nested logit model in the mode choice in Example 18.3 results from an error components model,

$$\begin{aligned}
U_{i,air} &= \qquad u_{i,fly} \qquad\quad + \mathbf{x}'_{i,air}\boldsymbol{\beta} + \varepsilon_{i,air},\\
U_{i,train} &= (\alpha_{train} + u_{i,ground}) + \mathbf{x}'_{i,train}\boldsymbol{\beta} + \varepsilon_{i,train},\\
U_{i,bus} &= (\alpha_{bus} + u_{i,ground}) \;+ \mathbf{x}'_{i,bus}\boldsymbol{\beta} + \varepsilon_{i,bus},\\
U_{i,car} &= (\alpha_{car} + u_{i,ground}) \;+ \mathbf{x}'_{i,car}\boldsymbol{\beta} + \varepsilon_{i,car}.
\end{aligned}$$

This is the model suggested after (18-12). The implied covariance matrix for the four utility functions would be

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_F^2 & 0 & 0 & 0 \\ 0 & \sigma_G^2 & \sigma_G^2\rho & \sigma_G^2\rho \\ 0 & \sigma_G^2\rho & \sigma_G^2 & \sigma_G^2\rho \\ 0 & \sigma_G^2\rho & \sigma_G^2\rho & \sigma_G^2 \end{bmatrix}.$$

FIML estimates of the nested logit model from Table 18.6 in Example 18.3 are reported in Table 18.15 below. We have refit the model as an error components model with the two components shown above. This is a model with random constant terms. The estimated

parameters in Table 18.15 are similar as would be expected. The estimated standard deviations for the FIML estimated model are 2.1886 and 3.2974 for *Fly* and *Ground*, respectively. For the random parameters model, we would calculate these using $v = (\pi^2/6 + \sigma_b^2)^{1/2} = 3.48$ for *Fly* and 1.3899 for *Ground*. The similarity of the results carries over to the estimated elasticities, some of which are shown in Table 18.16.

18.2.8.c A Fixed Effects Multinomial Logit Model

A fixed effects multinomial logit model can be formulated as

$$\text{Prob}(y_{it} = j) = \frac{\exp(\alpha_{ij} + \mathbf{x}_{it,j}'\boldsymbol{\beta})}{\Sigma_{m=1}^{J} \exp(\alpha_{im} + \mathbf{x}_{it,m}'\boldsymbol{\beta})}.$$

Because the probabilities are based on comparisons, one of the utility functions must be normalized at zero. We take that to be the last (Jth) alternative, so the normalized model is

$$\text{Prob}(y_{it} = j) = \frac{\exp(\alpha_{ij} + \mathbf{x}_{it,j}'\boldsymbol{\beta})}{1 + \Sigma_{m=1}^{J-1} \exp(\alpha_{im} + \mathbf{x}_{it,m}'\boldsymbol{\beta})}, j = 1, \ldots, J-1.$$

We examined the binary logit model with fixed effects in Section 17.7.3. The model here is a direct extension. The Rasch/Chamberlain method for the fixed effects logit model can be used, in principle, for this multinomial logit case. [Chamberlain (1980) mentions this possibility briefly.] However, the amount of computation involved in doing so increases vastly with J. Part of the complexity stems from the difficulty of constructing

TABLE 18.15 Estimated Nested Logit Models

	FIML Nested Logit		*Mixed Logit*	
	Estimate	***Std. Error***	***Estimate***	***Std. Error***
Air	6.04234	(1.19888)	4.65134	(1.26475)
Train	5.06460	(0.66202)	5.13427	(0.67043)
Bus	4.09632	(0.61516)	4.15790	(0.62631)
GC	−0.03159	(0.00816)	−0.03228	(0.00689)
TTME	−0.11262	(0.01413)	−0.11423	(0.01183)
HINC	0.02616	(0.01761)	0.03571	(0.02468)
Fly	0.58601	(0.14062)	3.24032	(1.71679)
Ground	0.38896	(0.12367)	0.53580	(10.65887)
ln *L*	−193.65615		−195.72711	

TABLE 18.16 Elasticities with Respect to Generalized Cost

	AIR		*TRAIN*		*BUS*		*CAR*	
	NL	***MXL***	***NL***	***MXL***	***NL***	***MXL***	***NL***	***MXL***
AIR	−1.3772	−1.1551	0.5228	0.4358	0.5228	0.4358	0.5228	0.4358
TRAIN	0.3775	0.4906	−2.9452	−3.0467	1.1675	1.1562	1.1675	1.1562
BUS	0.1958	0.2502	0.6039	0.5982	−3.0368	−3.1223	0.6039	0.5982
CAR	0.3372	0.3879	1.1424	1.1236	1.1424	1.1236	−1.8715	−1.9564

the denominator of the conditional probability. The terms in the sum are the different ways that the sequence of $J \times T$ outcomes can sum to T *including the constraint that within each block of J, the outcomes sum to one*. The amount of computation is potentially prohibitive. For our example below, with $J = 4$ and $T = 12$, the number of terms is roughly 6×10^{10}. The Krailo and Pike algorithm is less useful here due to the need to impose the constraint that only one choice be made in each period. However, there is a much simpler approach available based on the minimum distance principle that uses the same information.[21] (See Section 13.3.) For each of outcomes 1 to $J - 1$, the choice between observation j and the numeraire, alternative J, produces a fixed effects binary logit. For each of the $J - 1$ outcomes, then, the $\Sigma_{i=1}^{n} T_i$ observations that chose either outcome j or outcome J can be used to fit a binary logit model to estimate $\boldsymbol{\beta}$. This produces $J - 1$ estimates, $\hat{\boldsymbol{\beta}}_j$, each with estimated asymptotic covariance matrix $\mathbf{V}_j$. The minimum distance estimator of the single $\boldsymbol{\beta}$ would then be

$$\hat{\boldsymbol{\beta}} = \left[\sum_{j=1}^{J-1} \mathbf{V}_j^{-1}\right]^{-1} \sum_{j=1}^{J-1} (\mathbf{V}_j^{-1} \hat{\boldsymbol{\beta}}_j).$$

The estimated asymptotic covariance matrix would be the first term. Each of the binary logit estimates and the averaging at the last step require an insignificant amount of computation.

It does remain true that, like the binary choice estimator, the post-estimation analysis is severely limited because the fixed effects are not actually estimated. It is not possible to compute probabilities and partial effects, etc.

Example 18.8 Stated Choice Experiment: Preference for Electricity Supplier

Revelt and Train (2000) studied the preferences for different prices of a sample of California electricity customers.[22] The authors were particularly interested in individual heterogeneity and used a mixed logit approach. The choice experiment examines the choices among electricity suppliers in which a supplier is defined by a set of attributes. The choice model is based on

$$U_{ijt} = \beta_1 PRICE_{ijt} + \beta_2 TOD_{ijt} + \beta_3 SEAS_{ijt} + \beta_4 CNTL_{ijt} + \beta_5 LOCAL_{ijt} + \beta_6 KNOWN_{ijt} + \varepsilon_{ijt},$$

where

PRICE	= Fixed rates, cents/kwh = 7 or 9, or 0 if seasonal or time of day rates,
TOD	= Dummy for time of day rates, 11 cents 8AM-8PM, 5 cents 8PM – 8AM,
SEAS	= Dummy for seasonal rates, 10 summer, 8 winter, 6 spring and fall,
CNTL	= Fixed term contract with exit penalty, length 0, 1 year, 5 years,
LOCAL, KNOWN	= Dummies for familiarity: local utility, known but not local, unknown.

Data were collected in 1997 by the Research Triangle Institute for the Electric Power Research Institute.[23] The sample contains 361 individuals, each asked to make 12 choices from a set of 4 candidate firms.[24] There were a total of 4,308 choice situations analyzed.

---

[21] Pforr (2011) reports results for a moderate-sized problem with 4,344 individuals, about six periods and only two outcomes with four attributes. Using the brute force method takes over 100 seconds. The minimum distance estimator for the same problem takes 0.2 seconds to produce the identical results. The time advantage would be far greater for the four-choice model analyzed in Example 18.8.

[22] See also Train (2009, Chapter 11).

[23] Professor Train has generously provided the data for this experiment for us (and readers) to replicate, analyze, and extend the models in this example.

[24] A handful of the 361 individuals answered fewer than 12 choice tasks: two each answered 8 or 9; one answered 10 and eight answered 11.

This is an **unlabeled choice** experiment. There is no inherent distinction between the firms in the choice set other than the attributes. Firm 1 in the choice set is only labeled Firm 1 because it is first in the list. The choice situations we have examined in this chapter have varied in this dimension:

Example 18.2 Heating system types	labeled,
Example 18.3 Travel mode	labeled,
Example 18.4 Water heating type	labeled,
Example 18.5 Green energy	unlabeled,
Example 18.6 Malaria control guidelines	unlabeled,
Example 18.7 Heating systems	labeled,
Example 18.8 Electricity pricing	unlabeled.

One of the main uses of choice models is to analyze substitution patterns. In Example 18.3, we estimated elasticities of substitution among travel modes. Unlabeled choice experiments generally do not provide information about substitution between alternatives. They do provide information about willingness to pay. That will be the focus of the study in this example. When the utility function is based on price, rather than income, the marginal disutility of an increase in price is treated as a surrogate for the marginal utility of an increase in income for purposes of measuring willingness to pay. In general, the interpretation of the sign of the WTP is context specific. In the example below, we are interested in the perceived value of time of day rates, measured by the *TOD*/*PRICE* coefficients. Both coefficients are negative in the MNL model. But the negative of the price change is the surrogate for income. We interpret the WTP of approximately 10 cents/kwh as the amount the customer would accept as a fixed rate if he or she could avoid the *TOD* rates. But, the *LOCAL* brand value of the utility is positive, so the positive WTP is interpreted as the extra amount the customer would be willing to pay to be supplied by the local utility as opposed to an unknown supplier.

Table 18.17 reports estimates of the choice models for rate structures and utility companies. The MNL model shows marginal valuations of contract length, time, and seasonal rates relative to the fixed rates and the brand value of the utility. The WTP results are shown in Table 18.18. The negative coefficient on *Contract Length* implies that the average customer is willing to pay a premium of (0.17 cents/kwh)/year to avoid a fixed length contract. The offered contracts are one and five years, so customers appear to be willing to pay up to 0.85 cents/kwh to avoid a long-term contract. The brand value of the local utility compared to a new and unknown supplier is 2.3 cents/kwh. Since the average rate across the different scenarios is about 9 cents, this is quite a large premium. The value is somewhat less for a known, but not the local, utility. The coefficients on time of day and seasonal rates suggest the equivalent valuations of the rates compared to the fixed rate schedule. Based on the MNL model, the average customer would value the time of day rates as equivalent to a fixed rate schedule of 8.74 cents. The fixed rate offer was 7 or 9 cents/kwh, so this is on the high end.

The mixed logit model allows heterogeneity in the valuations. A normal distribution is used for the contract length and brand value coefficients. These allow the distributions to extend on both sides of zero so that, for example, some customers prefer the local utility while others do not. With an estimated mean of 2.16117 and standard deviation of 1.50097, these results suggest that $(1 - \Phi(2.16117/1.50097)) = 7.5\%$ of customers actually prefer an unknown outside supplier to their local utility. The coefficients on *TOD* and seasonal rates have been specified to have lognormal distributions. Because they are assumed to be negative, the specified coefficient is $-\exp(\beta + \sigma v)$. (The negative sign is attached to the variable and the coefficient on $-TOD$ is then specified with a positive lognormal distribution.) The mean value of this coefficient in the population distribution is then $E[\beta_{TOD}] = -\exp(2.11304 + 0.38651^2/2) = 8.915$., so the average customer is roughly indifferent between the *TOD* rates and the fixed rate schedule. Figure 18.2 shows a kernel

TABLE 18.17 Estimated Choice Models for Electricity Supplier (Standard errors in parentheses)

		Mixed Logit[b]				
Variable	***MNL***[a]	***Mean β***	***Std. Dev. σ***	***FEM***	***REM***[c]	***ANA***[d]
Price	−0.62523	−0.86814	0.00000	−0.38841	−0.63762	−0.54713
	(0.03349)	(0.02273)	(0.00000)	(0.02039)	(0.07432)	(0.03962)
Contract	−0.10830	−0.21831	0.36379	−0.05586	−0.10940	−0.10937
	(0.01402)	(0.01659)	(0.01736)	(0.00682)	(0.00964)	(0.00862)
Time of Day[e]	−5.46276	2.11304[e]	0.38651	−3.46145	−5.57917	−5.11061
	(0.27815)	(0.02693)	(0.01847)	(0.16622)	(0.59680)	(0.30446)
Seasonal[e]	−5.84003	2.13564[e]	0.27607	−3.59727	−5.95563	−5.34035
	(0.27272)	(0.02571)	(0.01589)	(0.16596)	(0.61004)	(0.30811)
Local	1.44224	2.16117	1.50097	0.83266	1.47522	1.44016
	(0.07887)	(0.08915)	(0.08985)	(0.04106)	(0.09103)	(0.05510)
Known	0.99550	1.46173	0.97705	0.47649	1.02153	0.97419
	(0.06387)	(0.06538)	(0.07272)	(0.03319)	(0.07962)	(0.04944)
ln L	−4958.65	−3959.73		−4586.93	−4945.98	−4882.34

[a] Robust standard errors are clustered over individuals. Conventional standard errors for MNL are 0.02322, 0.00824, 0.18371, 0.18668, 0.05056, 0.04478, respectively.
[b] Train (2009) reports point estimates (b,s) = (−0.8827,0), (−0.2125, 0.3865), (2.1328, 0.4113), (2.1577, 0.2812), (2.2297, 1.7514), (1.5906, 0.9621) for Price, Cntl, TOD, Seas, Local, Known, respectively.
[c] Estimated Standard Deviations in RE Model are 0.00655 (0.02245), 0.47463 (0.06049), 0.016062 (0.04259).
[d] Class probabilities are 0.93739, 0.06261.
[e] Lognormal coefficient in mixed logit model is $\exp(\beta + \sigma v)$.

TABLE 18.18 Estimated WTP Based on Different Models

	Contract	***Local***	***Known***	***TOD***	***Seasonal***
Multinomial Logit Fixed Parameter					
Estimate	0.17322	2.30675	1.59223	8.73723	9.34065
Standard Error	0.02364	0.18894	0.13870	0.15126	0.15222
Lower Confidence Limit	0.12689	1.93643	1.32038	8.44076	9.04230
Upper Confidence Limit	0.21955	2.67707	1.86407	9.03370	9.63899
Mixed Logit WTP for Rates					
Lognormal					
Estimated Mean = $\exp(\beta + \sigma^2/2)$				8.91500	8.79116
Estimated Std. Dev. = Mean × $[\exp(\sigma^2) - 1]^{1/2}$				3.57852	2.47396
5% Lower Limit				1.90110	3.94220
95% Upper Limit				15.92900	13.64012
Triangular					
Estimated Mean = β				7.83937	8.19676
Estimated Spread = $\beta \pm \sigma$				5.90744	4.15295
Estimated Std. Dev. = $[\sigma^2/6]^{1/2}$				2.41170	1.69543
5% Lower Limit				3.11244	4.87370
95% Upper Limit				12.56630	11.51981

density estimator of the estimated population distribution of marginal valuations of the *TOD* rates. The bimodal distribution shows the sample of estimated values of E[$-\beta_{TOD}$| choices made]. Train notes, if the model is properly specified and the estimates appropriate, the means of these two distributions should be the same. The sample mean of the estimated conditional means is 10.4 cents/kwh while the estimated population mean is 9.9. The estimated standard deviation of the population distribution is $8.915 \times [\exp(0.38651^2) - 1]^{1/2} = 3.578$. Thus, about 95% of the population is estimated to value the *TOD* rates in the interval 9.9 +/− 7.156. Note that a very high valuation of the *TOD* rates suggests a strong aversion to *TOD* rates. The lognormal distribution tends to produce implausibly large values such as those here in the thick tail of the distribution. We refit the model using triangular distributions that have fixed widths $\beta \pm \sigma$. The estimated distributions have range 7.839 ± 5.907 for *TOD* and 8.197 ± 4.152 for *Seasonal*. Computation of 95% probability intervals (based on a normal approximation, m $\pm$ 1.96s) are shown in Table 18.18.

Results are also shown for simple fixed and random effects estimates. The random effects results are essentially identical to the MNL results while the fixed effects results depart substantially from both the MNL and mixed logit results. The ANA model relates to whether, in spite of the earlier findings, there are customers who do not consider the brand value of the local utility in choosing their suppliers. The ANA model specifies two classes, one with full attendance and one in which coefficients on *LOCAL* and *KNOWN* are both equal to zero. The results suggest that 6.26% of the population ignores the brand value of the supplier in making their choices.

18.2.9 AGGREGATE MARKET SHARE DATA—THE BLP RANDOM PARAMETERS MODEL

The structural demand model of Berry, Levinsohn, and Pakes (BLP) (1995) is an important application of the mixed logit model. Demand models for differentiated products such as automobiles [BLP (1995), Goldberg (1995)], ready-to-eat cereals [Nevo (2001)], and consumer electronics [Das, Olley, and Pakes (1996)], have been constructed using the mixed logit model with market share data.[25] A basic structure is defined for

$$\begin{aligned}&\text{Markets, denoted } t = 1, \ldots, T,\\ &\text{Consumers in the markets, denoted } i = 1, \ldots, n_t,\\ &\text{Products, denoted } j = 1, \ldots, J.\end{aligned}$$

The definition of a market varies by application; BLP analyzed the U.S. national automobile market for 20 years; Nevo examined a cross section of cities over 20 quarters so the city-quarter is a market; Das et al. defined a market as the annual sales to consumers in particular income levels.

For market t, we base the analysis on average prices, p_{jt}; aggregate quantities, q_{jt}; consumer incomes, y_i; observed product attributes, $\mathbf{x}_{jt}$; and unobserved (by the analyst) product attributes, Δ_{jt}. The indirect utility function for consumer i, for product j in market t is

$$u_{ijt} = \alpha_i(y_i - p_{jt}) + \mathbf{x}_{jt}'\boldsymbol{\beta}_i + \Delta_{jt} + \varepsilon_{ijt}, \qquad \textbf{(18-14)}$$

where α_i is the marginal utility of income and $\boldsymbol{\beta}_i$ are marginal utilities attached to specific observable attributes of the products. The fact that some unobservable product attributes, Δ_{jt}, will be reflected in the prices implies that prices will be endogenous in a demand

[25] We draw heavily on Nevo (2000) for this discussion.

model that is based on only the observable attributes. Heterogeneity in preferences is reflected (as we did earlier) in the formulation of the random parameters,

$$\begin{pmatrix}\alpha_i\\ \boldsymbol{\beta}_i\end{pmatrix} = \begin{pmatrix}\alpha\\ \boldsymbol{\beta}\end{pmatrix} + \begin{pmatrix}\boldsymbol{\pi}'\\ \boldsymbol{\Pi}\end{pmatrix}\mathbf{d}_i + \begin{pmatrix}\gamma w_i\\ \boldsymbol{\Gamma}\mathbf{v}_i\end{pmatrix}, \tag{18-15}$$

where $\mathbf{d}_i$ is a vector of demographics such as gender and age while α, $\boldsymbol{\beta}$, $\boldsymbol{\pi}$, $\boldsymbol{\Pi}$, γ, and $\boldsymbol{\Gamma}$ are structural parameters to be estimated (assuming they are identified). A utility function is also defined for an "outside good" that is (presumably) chosen if the consumer chooses none of the brands, $1, \ldots, J$,

$$u_{i0t} = \alpha_i y_i + \Delta_{0t} + \boldsymbol{\pi}_0'\mathbf{d}_i + \varepsilon_{i0t}.$$

Since there is no variation in income across the choices, $\alpha_i y_i$ will fall out of the logit probabilities, as we saw earlier. A normalization is used instead, $u_{i0t} = \varepsilon_{i0t}$, so that comparisons of utilities are against the outside good. The resulting model can be reconstructed by inserting (18-15) into (18-14),

$$\begin{aligned}
u_{ijt} &= \alpha_i y_i + \delta_{jt}(\mathbf{x}_{jt}, p_{jt}, \Delta_{jt}: \alpha, \boldsymbol{\beta}) + \tau_{ijt}(\mathbf{x}_{jt}, p_{jt}, \mathbf{v}_i, w_i: \boldsymbol{\pi}, \boldsymbol{\Pi}, \gamma, \boldsymbol{\Gamma}) + \varepsilon_{ijt},\\
\delta_{jt} &= \mathbf{x}_{jt}'\boldsymbol{\beta} - \alpha p_{jt} + \Delta_{jt},\\
\tau_{jt} &= [-p_{jt}, \mathbf{x}_{jt}']\left[\begin{pmatrix}\boldsymbol{\pi}'\\ \boldsymbol{\Pi}\end{pmatrix}d_i + \begin{pmatrix}\gamma w_i\\ \boldsymbol{\Gamma}\mathbf{v}_i\end{pmatrix}\right].
\end{aligned}$$

The preceding model defines the random utility model for consumer i in market t. Each consumer is assumed to purchase the one good that maximizes utility. The market share of the jth product in this market is obtained by summing over the choices made by those consumers. With the assumption of homogeneous tastes ($\boldsymbol{\Gamma} = \mathbf{0}$ and $\gamma = 0$) and i.i.d., type I extreme value distributions for ε_{ijt}, it follows that the market share of product j is

$$s_{jt} = \frac{\exp(\mathbf{x}_{jt}'\boldsymbol{\beta} - \alpha p_{jt} + \Delta_{jt})}{1 + \sum_{k=1}^{J}\exp(\mathbf{x}_{kt}'\boldsymbol{\beta} - \alpha p_{kt} + \Delta_{kt})}.$$

The IIA assumptions produce the familiar problems of peculiar and unrealistic substitution patterns among the goods. Alternatives considered include a nested logit, a "generalized extreme value" model and, finally, the mixed logit model, now applied to the aggregate data.

Estimation cannot proceed along the lines of Section 18.2.7 because Δ_{jt} is unobserved and p_{jt} is, therefore, endogenous. BLP propose, instead, to use a GMM estimator, based on the moment equations,

$$E\{[S_{jt} - s_{jt}(\mathbf{x}_{jt}, p_{jt}|\alpha, \boldsymbol{\beta})]\mathbf{z}_{jt}\} = \mathbf{0},$$

for a suitable set of instruments. Layering in the random parameters specification, we obtain an estimation based on **method of simulated moments**, rather than a maximum simulated log likelihood. The simulated moments would be based on

$$E_{w,\mathbf{v}}[\mathbf{s_{jt}}(\mathbf{x}_{jt}, p_{jt}|\alpha_i, \boldsymbol{\beta}_i)] = \int_{w,\mathbf{v}}\{s_{jt}[\mathbf{x}_{jt}, p_{jt}|\alpha_i(w), \boldsymbol{\beta}_i(\mathbf{v})]\}dF(w)dF(\mathbf{v}).$$

These would be simulated using the method of Section 18.2.7. The algorithm developed by BLP for estimation of the model is famously intricate and complicated. Several authors have proposed faster, less complicated methods of estimation. Lee and Seo (2011) proposed a useful device that is straightforward to implement.

Example 18.9 Health Insurance Market

Tamm, Tauchmann, Wasem, and Greβ (2007) analyzed the German health insurance market in this framework. The study was motivated by the introduction of competition into the German social health insurance system in 1996. The authors looked for evidence of competition in estimates of the price elasticities of the market shares of the firms using an extensive panel data set spanning 2001–2004. The starting point is a model for the market shares,

$$s_{it} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \gamma_i + \varepsilon_{it})}{\sum_{i=1}^{N}\exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \gamma_i + \varepsilon_{it})}, i = 1, \ldots, N.$$

Taking logs produces

$$\ln(s_{it}) = \boldsymbol{\beta}'\mathbf{x}_{it} + \delta_t + \gamma_i + \varepsilon_{it},$$

where δ_t is the log of the denominator, which is the same for all firms, and γ_i is an endogenous firm effect. Since consumers do not change their insurer every period, the model is augmented to account for persistence,

$$\ln(s_{it}) = \alpha \ln(s_{i,t-1}) + \boldsymbol{\beta}'\mathbf{x}_{it} + \delta_t + \gamma_i + \varepsilon_{it}.$$

The limiting cases of $\alpha = 0$ (the static case) and $\alpha = 1$ (random walk) are examined in the study, as well as the intermediate cases. GMM estimators are formulated for the three cases. The preferred estimate of the premium elasticity (from their Table VII) is -1.09, with a confidence interval of (-1.43 to -0.75), which suggests the influence of price competition in this market.

18.3 RANDOM UTILITY MODELS FOR ORDERED CHOICES

The analysts at bond rating agencies such as Moody's and Standard & Poor's provide an evaluation of the quality of a bond that is, in practice, a discrete listing of the continuously varying underlying features of the security. The rating scales are as follows:

Rating	*S&P Rating*	*Moody's Rating*
Highest quality	AAA	Aaa
High quality	AA	Aa
Upper medium quality	A	A
Medium grade	BBB	Baa
Somewhat speculative	BB	Ba
Low grade, speculative	B	B
Low grade, default possible	CCC	Caa
Low grade, partial recovery possible	CC	Ca
Default, recovery unlikely	C	C

For another example, Netflix (www.netflix.com) is an Internet company that, among other activities, streams movies to subscribers. After a subscriber streams a movie, the next time

he or she logs onto the Web site, he or she is invited to rate that movie on a five-point scale, where five is the highest, most favorable rating. The ratings of the many thousands of subscribers who streamed that movie are averaged to provide a recommendation to prospective viewers. As of April 5, 2009, the average rating of the 2007 movie *National Treasure: Book of Secrets* given by approximately 12,900 visitors to the site was 3.8. Many other Internet sellers of products and services, such as Barnes & Noble, Amazon, Hewlett Packard, and Best Buy, employ rating schemes such as this. Many recently developed national survey data sets, such as the British Household Panel Data Set (BHPS) (www.iser.essex.ac.uk/bhps), the Australian HILDA data (www.melbourneinstitute.com/hilda/), and the German Socioeconomic Panel (GSOEP) (www.diw.de/en/soep), all contain questions that elicit self-assessed ratings of health, health satisfaction, or overall well-being. Like the other examples listed, these survey questions are answered on a discrete scale, such as the 0 to 10 scale of the question about health satisfaction in the GSOEP.[26] Ratings such as these provide applications of the models and methods that interest us in this section.[27]

For an individual respondent, we hypothesize that there is a continuously varying strength of preferences that underlies the rating he or she submits. For convenience and consistency with what follows, we will label that strength of preference "utility," U^*. Continuing the Netflix example, we describe utility as ranging over the entire real line,

$$-\infty < U_{im}^* < +\infty,$$

where i indicates the individual and m indicates the movie. Individuals are invited to rate the movie on an integer scale from 1 to 5. Logically, then, the translation from underlying utility to a rating could be viewed as a *censoring* of the underlying utility,

$$\begin{aligned}
R_{im} &= 1 \text{ if } -\infty < U_{im}^* \leq \mu_1,\\
R_{im} &= 2 \text{ if } \mu_1 < U_{im}^* \leq \mu_2,\\
R_{im} &= 3 \text{ if } \mu_2 < U_{im}^* \leq \mu_3,\\
R_{im} &= 4 \text{ if } \mu_3 < U_{im}^* \leq \mu_4,\\
R_{im} &= 5 \text{ if } \mu_4 < U_{im}^* < \infty.
\end{aligned}$$

The same mapping would characterize the bond ratings, since the qualities of bonds that produce the ratings will vary continuously, and the self-assessed health and well-being questions in the panel survey data sets are based on an underlying utility or preference structure. The crucial feature of the description thus far is that underlying the discrete response is a continuous range of preferences. Therefore, the observed rating represents a censored version of the true underlying preferences. Providing a rating of five could be an outcome ranging from general enjoyment to wild enthusiasm. Note that for thresholds, μ_j, number $(J - 1)$, where J is the number of possible ratings (here, five)—$J - 1$ values are needed to divide the range of utility into J cells. The thresholds are an important element of the model; they divide the range of utility into cells that are then identified with the observed outcomes. Importantly, the difference between

[26]The original survey used a 0–10 scale for self-assessed health. It is currently based on a five-point scale.

[27]Greene and Hensher (2010a) provide a survey of ordered choice modeling. Other textbook and monograph treatments include DeMaris (2004), Long (1997), Johnson and Albert (1999), and Long and Freese (2006). Introductions to the model also appear in journal articles such as Winship and Mare (1984), Becker and Kennedy (1992), Daykin and Moffatt (2002), and Boes and Winkelmann (2006).

two levels of a rating scale (for example, one compared to two, two compared to three) is not the same as on a utility scale. Hence, we have a strictly nonlinear transformation captured by the thresholds, which are estimable parameters in an ordered choice model.

The model as suggested thus far provides a crude description of the mechanism underlying an observed rating. Any individual brings his or her own set of characteristics to the utility function, such as age, income, education, gender, where he or she lives, family situation, and so on, which we denote $x_{i1}, x_{i2}, \ldots, x_{iK}$. They also bring their own aggregates of unmeasured and unmeasurable (by the statistician) idiosyncrasies, denoted ε_{im}. How these features enter the utility function is uncertain, but it is conventional to use a linear function, which produces a familiar random utility function,

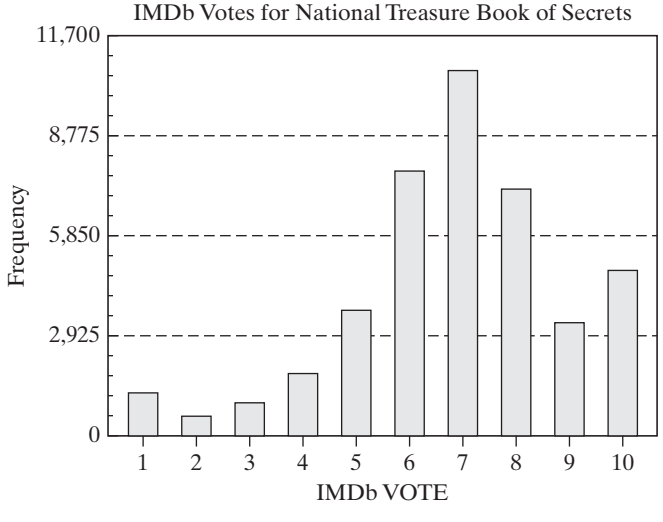
$$U_{im}^{*} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_{im}.$$

Example 18.10 Movie Ratings

The Web site www.IMDb.com invites visitors to rate movies that they have seen. This site uses a 10-point scale. It reported the results in Figure 18.3 for the movie *National Treasure: Book of Secrets* for 41,771 users of the site.[28] The figure at the left shows the overall ratings. The panel at the right shows how the average rating varies across age, gender, and whether the rater is a U.S. viewer or not. The rating mechanism we have constructed is

$$\begin{aligned}
R_{im} &= 1 \text{ if } -\infty < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_1,\\
R_{im} &= 2 \text{ if } \mu_1 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_2,\\
&\cdots\\
R_{im} &= 9 \text{ if } \mu_8 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_9,\\
R_{im} &= 10 \text{ if } \mu_9 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} < \infty.
\end{aligned}$$

FIGURE 18.3 IMDb.com Ratings.



	Votes	Mean
Males	33,644	6.5
Females	5,464	6.8
Aged under 18	2,492	6.6
Males under 18	1,795	6.5
Females under 18	695	7.2
Aged 18–29	26,045	6.6
Males Aged 18–29	22,603	6.6
Females Aged 18–29	3,372	6.8
Aged 30–44	8,210	6.4
Males Aged 30–44	7,216	6.4
Females Aged 30–44	936	6.7
Aged 45+	2,258	6.5
Males Aged 45+	1,814	6.4
Females Aged 45+	420	6.9
US Users	14,792	6.6
Non-US Users	24,283	6.5

[28]The data are as of December 1, 2008. A rating for the same movie as of August 1, 2016 at www.imdb.com/title/tt0465234/ratings?ref_=tt_ov_rt shows essentially the same pattern for 182,780 viewers.

Relying on a central limit theorem to aggregate the innumerable small influences that add up to the individual idiosyncrasies and movie attraction, we assume that the random component, ε_{im}, is normally distributed with zero mean and (for now) constant variance. The assumption of normality will allow us to attach probabilities to the ratings. In particular, arguably the most interesting one is

$$\text{Prob}(R_{im} = 10 \,|\, \mathbf{x}_i) = \text{Prob}[\varepsilon_{im} > \mu_9 - \mathbf{x}_i'\boldsymbol{\beta}].$$

The structure provides the framework for an econometric model of how individuals rate movies (that they stream from Netflix). The resemblance of this model to familiar models of binary choice is more than superficial. For example, one might translate this econometric model directly into a simple probit model by focusing on the variable

$$E_{im} = 1 \text{ if } R_{im} = 10$$
$$E_{im} = 0 \text{ if } R_{im} < 10.$$

Thus, the model is an extension of a binary choice model to a setting of more than two choices. But the crucial feature of the model is the ordered nature of the observed outcomes and the correspondingly ordered nature of the underlying preference scale.

The model described here is an *ordered choice model*. (The use of the normal distribution for the random term makes it an *ordered probit model*.) Ordered choice models are appropriate for a wide variety of settings in the social and biological sciences. The essential ingredient is the mapping from an underlying, naturally ordered preference scale to a discrete ordered observed outcome, such as the rating scheme just described. The model of ordered choice pioneered by Aitcheson and Silvey (1957), Snell (1964), and Walker and Duncan (1967) and articulated in its modern form by Zavoina and McElvey (1975) has become a widely used tool in many fields. The number of applications in the current literature is large and increasing rapidly, including:

- Bond ratings [Terza (1985a)],
- Congressional voting on a Medicare bill [McElvey and Zavoina (1975)],
- Credit ratings [Cheung (1996), Metz, and Cantor (2006)],
- Driver injury severity in car accidents [Eluru, Bhat, and Hensher (2008)],
- Drug reactions [Fu, Gordon, Liu, Dale, and Christensen (2004)],
- Education [Machin and Vignoles (2005), Carneiro, Hansen, and Heckman (2003), Cunha, Heckman, and Navarro (2007)],
- Financial failure of firms [Hensher and Jones (2007)],
- Happiness [Winkelmann (2005), Zigante (2007)],
- Health status [Jones, Koolman, and Rice (2003)],
- Job skill rating [Marcus and Greene (1985)],
- Life satisfaction [Clark, Georgellis, and Sanfey (2001), Groot and ven den Brink (2003), Winkelmann (2002)],
- Monetary policy [Eichengreen, Watson, and Grossman (1985)],
- Nursing labor supply [Brewer, Kovner, Greene, and Cheng (2008)],
- Obesity [Greene, Harris, Hollingsworth, and Maitra (2008)],
- Political efficacy [King, Murray, Salomon, and Tandon (2004)],
- Pollution [Wang and Kockelman (2009)],
- Promotion and rank in nursing [Pudney and Shields (2000)],

- Stock price movements [Tsay (2005)],
- Tobacco use [Harris and Zhao (2007), Kasteridis, Munkin, and Yen (2008)], and
- Work disability [Kapteyn et al. (2007)].

18.3.1 THE ORDERED PROBIT MODEL

The ordered probit model is built around a latent regression in the same manner as the binomial probit model. We begin with

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

As usual, y^* is unobserved. What we do observe is

$$\begin{aligned} y &= 0 \quad \text{if } y^* \le 0 \\ &= 1 \quad \text{if } 0 < y^* \le \mu_1 \\ &= 2 \quad \text{if } \mu_1 < y^* \le \mu_2 \\ &\vdots \\ &= J \quad \text{if } \mu_{J-1} \le y^*, \end{aligned}$$

which is a form of censoring. The μ's are unknown parameters to be estimated with $\boldsymbol{\beta}$.

We assume that ε is normally distributed across observations.[29] For the same reasons as in the binomial probit model (which is the special case with $J = 1$), we normalize the mean and variance of ε to zero and one. We then have the following probabilities:

$$\begin{aligned} \text{Prob}(y = 0|\mathbf{x}) &= \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 1|\mathbf{x}) &= \Phi(\mu_1 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 2|\mathbf{x}) &= \Phi(\mu_2 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\mu_1 - \mathbf{x}'\boldsymbol{\beta}), \\ &\vdots \\ \text{Prob}(y = J|\mathbf{x}) &= 1 - \Phi(\mu_{J-1} - \mathbf{x}'\boldsymbol{\beta}). \end{aligned}$$

For all the probabilities to be positive, we must have

$$0 < \mu_1 < \mu_2 < \cdots < \mu_{J-1}.$$

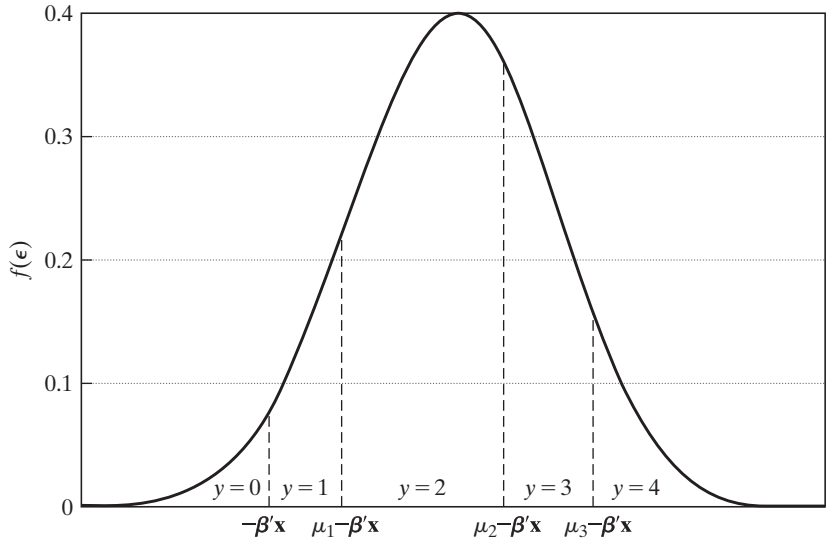
Figure 18.4 shows the implications of the structure. This is an extension of the univariate probit model we examined in Chapter 17. The log-likelihood function and its derivatives can be obtained readily, and optimization can be done by the usual means.

As usual, the partial effects of the regressors $\mathbf{x}$ on the probabilities are not equal to the coefficients. It is helpful to consider a simple example. Suppose there are three categories. The model thus has only one unknown threshold parameter. The three probabilities are

$$\begin{aligned} \text{Prob}(y = 0|\mathbf{x}) &= 1 - \Phi(\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 1|\mathbf{x}) &= \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 2|\mathbf{x}) &= 1 - \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}). \end{aligned}$$

[29] Other distributions, particularly the logistic, could be used just as easily. We assume the normal purely for convenience. The logistic and normal distributions generally give similar results in practice.

FIGURE 18.4 Probabilities in the Ordered Probit Model.



For the three probabilities, the partial effects of changes in the regressors are

$$\frac{\partial \text{Prob}(y = 0|\mathbf{x})}{\partial \mathbf{x}} = -\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta},$$

$$\frac{\partial \text{Prob}(y = 1|\mathbf{x})}{\partial \mathbf{x}} = [\phi(-\mathbf{x}'\boldsymbol{\beta}) - \phi(\mu - \mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta},$$

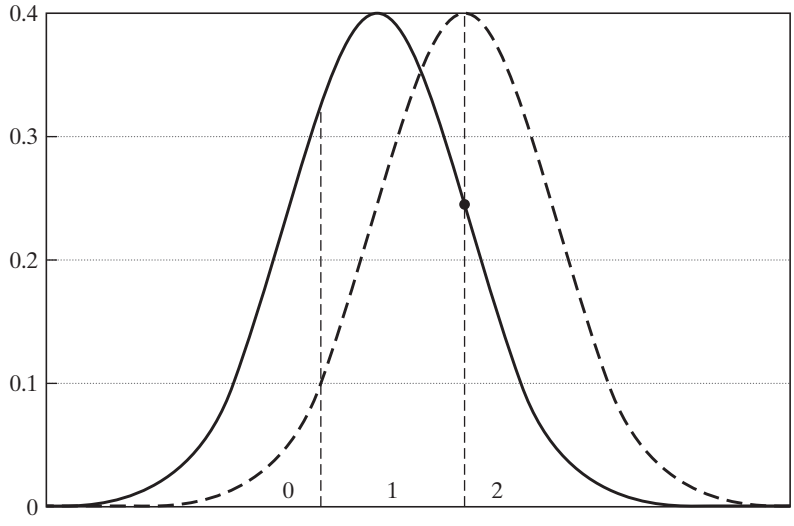
$$\frac{\partial \text{Prob}(y = 2|\mathbf{x})}{\partial \mathbf{x}} = \phi(\mu - \mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}.$$

Figure 18.5 illustrates the effect. The probability distributions of y and y^* are shown in the solid curve. Increasing one of the x's while holding $\boldsymbol{\beta}$ and μ constant is equivalent to shifting the distribution slightly to the right, which is shown as the dashed curve. The effect of the shift is unambiguously to shift some mass out of the leftmost cell. Assuming that $\boldsymbol{\beta}$ is positive (for this x), $\text{Prob}(y = 0|\mathbf{x})$ must decline. Alternatively, from the previous expression, it is obvious that the derivative of $\text{Prob}(y = 0|\mathbf{x})$ has the opposite sign from $\boldsymbol{\beta}$. By a similar logic, the change in $\text{Prob}(y = 2|\mathbf{x})$ [or $\text{Prob}(y = J|\mathbf{x})$ in the general case] must have the same sign as $\boldsymbol{\beta}$. Assuming that the particular $\boldsymbol{\beta}$ is positive, we are shifting some probability into the rightmost cell. But what happens to the middle cell is ambiguous. It depends on the two densities. In the general case, relative to the signs of the coefficients, only the signs of the changes in $\text{Prob}(y = 0|\mathbf{x})$ and $\text{Prob}(y = J|\mathbf{x})$ are unambiguous! The upshot is that we must be very careful in interpreting the coefficients in this model. Indeed, without a fair amount of extra calculation, it is quite unclear how the coefficients in the ordered probit model should be interpreted.

Example 18.11 Rating Assignments

Marcus and Greene (1985) estimated an ordered probit model for the job assignments of new Navy recruits. The Navy attempts to direct recruits into job classifications in which they will be

FIGURE 18.5 Effects of Change in *x* on Predicted Probabilities.



most productive. The broad classifications the authors analyzed were technical jobs with three clearly ranked skill ratings: "medium skilled," "highly skilled," and "nuclear qualified/highly skilled." Because the assignment is partly based on the Navy's own assessment and needs and partly on factors specific to the individual, an ordered probit model was used with the following determinants: (1) ENSPE = a dummy variable indicating that the individual entered the Navy with an "A school" (technical training) guarantee; (2) EDMA = educational level of the entrant's mother; (3) AFQT = score on the Armed Forces Qualifying Test; (4) EDYR = years of education completed by the trainee; (5) MARR = a dummy variable indicating that the individual was married at the time of enlistment; and (6) AGEAT = trainee's age at the time of enlistment. (The data used in this study are not available for distribution.) The sample size was 5,641. The results are reported in Table 18.19. The extremely large *t* ratio on the AFQT score is to be expected, as it is a primary sorting device used to assign job classifications.

To obtain the marginal effects of the continuous variables, we require the standard normal density evaluated at $-\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}} = -0.8479$ and $\hat{\mu} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}} = 0.9421$. The predicted probabilities are $\Phi(-0.8479) = 0.198$, $\Phi(0.9421) - \Phi(-0.8479) = 0.628$, and $1 - \Phi(0.9421) = 0.174$. (The

TABLE 18.19 Estimated Rating Assignment Equation

Variable	*Estimate*	*t Ratio*	*Mean of Variable*
Constant	−4.34	—	—
ENSPA	0.057	1.7	0.66
EDMA	0.007	0.8	12.1
AFQT	0.039	39.9	71.2
EDYRS	0.190	8.7	12.1
MARR	−0.48	−9.0	0.08
AGEAT	0.0015	0.1	18.8
μ	1.79	80.8	—

TABLE 18.20 Partial Effect of a Binary Variable

	$-\hat{\boldsymbol{\beta}}'\mathbf{x}$	$\hat{\mu} - \hat{\boldsymbol{\beta}}'\mathbf{x}$	$Prob[y = 0]$	$Prob[y = 1]$	$Prob[y = 2]$
$MARR = 0$	−0.8863	0.9037	0.187	0.629	0.184
$MARR = 1$	−0.4063	1.3837	0.342	0.574	0.084
Change			0.155	−0.055	−0.100

actual frequencies were 0.25, 0.52, and 0.23.) The two densities are $\phi(-0.8479) = 0.278$ and $\phi(0.9421) = 0.255$. Therefore, the derivatives of the three probabilities with respect to AFQT, for example, are

$$\frac{\partial P_0}{\partial \text{AFQT}} = (-0.278)0.039 = -0.01084,$$

$$\frac{\partial P_1}{\partial \text{AFQT}} = (0.278 - 0.255)0.039 = 0.0009,$$

$$\frac{\partial P_2}{\partial \text{AFQT}} = 0.255(0.039) = 0.00995.$$

Note that the marginal effects sum to zero, which follows from the requirement that the probabilities add to one. This approach is not appropriate for evaluating the effect of a dummy variable. We can analyze a dummy variable by comparing the probabilities that result when the variable takes its two different values with those that occur with the other variables held at their sample means. For example, for the MARR variable, we have the results given in Table 18.20.

18.3.2.a SPECIFICATION TEST FOR THE ORDERED CHOICE MODEL

The basic formulation of the ordered choice model implies that for constructed binary variables,

$$w_{ij} = 1 \text{ if } y_i \leq j, 0 \text{ otherwise}, j = 1, 2, \ldots, J-1, \qquad \textbf{(18-16)}$$

$$\text{Prob}(w_{ij} = 1 \,|\, \mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta} - \mu_j).$$

The first of these, when $j = 1$, is the binary choice model of Section 17.2. One implication is that we could estimate the slopes, but not the threshold parameters, in the ordered choice model just by using w_{i1} and $\mathbf{x}_i$ in a binary probit or logit model. (Note that this result also implies the validity of combining adjacent cells in the ordered choice model.) But (18-16) also defines a set of $J - 1$ binary choice models with different constants but common slope vector, $\boldsymbol{\beta}$. This equality of the parameter vectors in (18-16) has been labeled the **parallel regression assumption**. Although it is merely an implication of the model specification, this has been viewed as an implicit restriction on the model.[30] Brant (1990) suggests a test of the parallel regressions assumption based on (18-16). One can, in principle, fit $J - 1$ such binary choice models separately. Each will produce its own constant term and a consistent estimator of the common $\boldsymbol{\beta}$. Brant's Wald test examines the linear restrictions $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_{J-1}$, or H_0: $\boldsymbol{\beta}_q - \boldsymbol{\beta}_1 = \mathbf{0}$, $q = 2, \ldots, J-1$. The Wald statistic will be

[30] See, for example, Long (1997, p. 141).

$$\chi^2[(J-2)K] = (\mathbf{R}\hat{\boldsymbol{\beta}}^*)'[\mathbf{R} \times \text{Asy.Var}[\hat{\boldsymbol{\beta}}^*] \times \mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}^*),$$

where $\hat{\boldsymbol{\beta}}^*$ is obtained by stacking the individual binary logit or probit estimates of $\boldsymbol{\beta}$ (without the constant terms).[31]

Rejection of the null hypothesis calls the model specification into question. An alternative model in which there is a different $\boldsymbol{\beta}$ for each value of y has two problems: it does not force the probabilities to be positive and it is internally inconsistent. On the latter point, consider the suggested latent regression, $y^* = \mathbf{x}'\boldsymbol{\beta}_j + \varepsilon$. If the $\boldsymbol{\beta}$ is different for each j, then it is not possible to construct a data-generating mechanism for y^* (or, for example, simulate it); the realized value of y^* cannot be defined without knowing y (that is, the realized j), since the applicable $\boldsymbol{\beta}$ depends on j, but y is supposed to be determined from y^* through, for example, (18-16). There is no parametric restriction other than the one we seek to avoid that will preserve the ordering of the probabilities for all values of the data and maintain the coherency of the model. This still leaves the question of what specification failure would logically explain the finding. Some suggestions in Brant (1990) include: (1) misspecification of the latent regression, $\mathbf{x}'\boldsymbol{\beta}$; (2) heteroscedasticity of ε; and (3) misspecification of the distributional form for the latent variable, that is, "nonlogistic link function."

Example 18.12 Brant Test for an Ordered Probit Model of Health Satisfaction

In Examples 17.6 through 17.10 and several others, we studied the health care usage of a sample of households in the German Socioeconomic Panel (GSOEP). The data include a self-reported measure of health satisfaction (HSAT) that is coded 0 to 10. This variable provides a natural application of the ordered choice models in this chapter. The data are an unbalanced panel. For purposes of this exercise, we have used the first (1984) wave of the data set, which is a cross section of 4,483 observations. We then collapsed the 11 cells into 5 [(0–2), (3–5), (6–8), (9), (10)] for this example. The utility function is

$$HSAT_i^* = \beta_1 + \beta_2\, AGE_i + \beta_3\, INCOME_i + \beta_4\, KIDS_i + \beta_5\, EDUC_i + \beta_6\, MARRIED_i\, \beta_7\, WORKING_i + \varepsilon_i.$$

Variables *KIDS*, M*ARRIED*, and *WORKING* are binary indicators of whether there are children in the household, marital status, and whether the individual was working at the time of the survey. (These data are examined further in Example 18.14.) The model contains six variables, and there are four binary choice models fit, so there are $(J-2)(K) = (3)(6) = 18$ restrictions. The chi squared for the probit model is 87.836. The critical value for 95% is 28.87, so the homogeneity restriction is rejected. The corresponding value for the logit model is 77.84, which leads to the same conclusion.

18.3.3 BIVARIATE ORDERED PROBIT MODELS

There are several extensions of the ordered probit model that follow the logic of the bivariate probit model we examined in Section 17.9. A direct analog to the base case two-equation model is used in the study in Example 18.13.

Example 18.13 Calculus and Intermediate Economics Courses

Butler et al. (1994) analyzed the relationship between the level of calculus attained and grades in intermediate economics courses for a sample of Vanderbilt University students. The two-step estimation approach involved the following strategy. (We are stylizing the precise formulation a bit to compress the description.) Step 1 involved a direct application of the

[31]See Brant (1990), Long (1997), or Greene and Hensher (2010a, p. 187) for details on computing the statistic.

ordered probit model of Section 18.3.1 to the level of calculus achievement, which is coded 0,1, . . . , 6:

$$\begin{aligned} m_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i|\mathbf{x}_i \sim \text{N}[0, 1], \\ m_i &= 0 \text{ if } -\infty < m_i^* \leq 0 \\ &= 1 \text{ if } 0 < m_i^* \leq \mu_1 \\ &\cdots \\ &= 6 \text{ if } \mu_5 < m_i^* < +\infty. \end{aligned}$$

The authors argued that although the various calculus courses can be ordered discretely by the material covered, the differences between the levels cannot be measured directly. Thus, this is an application of the ordered probit model. The independent variables in this first-step model included SAT scores, foreign language proficiency, indicators of intended major, and several other variables related to areas of study.

The second step of the estimator involves regression analysis of the grade in the intermediate microeconomics or macroeconomics course. Grades in these courses were translated to a granular continuous scale (A = 4.0, A− = 3.7, etc.). A linear regression is specified,

$$\textit{Grade}_i = \mathbf{z}_i'\delta + u_i, \quad \text{where} \quad u_i|\mathbf{z}_i \sim \text{N}[0, \sigma_u^2].$$

Independent variables in this regression include, among others: (1) dummy variables for which outcome in the ordered probit model applies to the student (with the zero reference case omitted), (2) grade in the last calculus course, (3) several other variables related to prior courses, (4) class size, (5) freshman GPA, and so on. The unobservables in the *Grade* equation and the math attainment are clearly correlated, a feature captured by the additional assumption that $(\varepsilon_i, u_i|\mathbf{x}_i, \mathbf{z}_i) \sim \text{N}_2[(0, 0), (1, \sigma_u^2), \rho\sigma_u]$. A nonzero ρ captures this "selection" effect. With this in place, the dummy variables in (1) have now become endogenous. The solution is a *selection* correction that we will examine in detail in Chapter 19. The modified equation becomes

$$\begin{aligned} \textit{Grade}_i|m_i &= \mathbf{z}_i'\delta + E[u_i|m_i] + v_i \\ &= \mathbf{z}_i'\delta + (\rho\sigma_u)[\lambda(\mathbf{x}_i'\boldsymbol{\beta}, \mu_1, \ \ldots, \mu_5)] + v_i. \end{aligned}$$

They thus adopt a "control function" approach to accommodate the endogeneity of the math attainment dummy variables. [See Sections 17.6.2d and 17.6.2e) for another application of this method.] The term $\lambda(\mathbf{x}_i'\boldsymbol{\beta}, \mu_1, \ \ldots, \mu_5)$ is a generalized residual that is constructed using the estimates from the first-stage ordered probit model.[32] Linear regression of the course grade on $\mathbf{z}_i$ and this constructed regressor is computed at the second step. The standard errors at the second step must be corrected for the use of the estimated regressor using what amounts to a Murphy and Topel (2002) correction. (See Section 14.7.)

Li and Tobias (2006) in a replication of and comment on Butler et al. (1994), after roughly replicating the classical estimation results with a Bayesian estimator, observe that the preceding *Grade* equation above could also be treated as an ordered probit model. The resulting **bivariate ordered probit** model would be

$$\begin{aligned} m_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \text{and} \qquad & g_i^* &= \mathbf{z}_i'\boldsymbol{\delta} + u_i, \\ m_i &= 0 \text{ if } -\infty < m_i^* \leq 0 & g_i &= 0 \text{ if } -\infty < g_i^* \leq 0 \\ &= 1 \text{ if } 0 < m_i^* \leq \mu_1 & &= 1 \text{ if } 0 < g_i^* \leq \alpha_1 \\ &\cdots & &\cdots \\ &= 6 \text{ if } \mu_5 < m_i^* < +\infty & &= 11 \text{ if } \mu_9 < g_i^* < +\infty, \end{aligned}$$

[32] A precise statement of the form of this variable is given in Li and Tobias (2006).

where

$$(\varepsilon_i, u_i | \mathbf{x}_i, \mathbf{z}_i) \sim \mathbf{N}_2[(0, 0),(1, \sigma_u^2), \rho\sigma_u].$$

Li and Tobias extended their analysis to this case simply by transforming the dependent variable in Butler et al.'s second equation. Computing the log likelihood using sets of bivariate normal probabilities is fairly straightforward for the bivariate ordered probit model.[33] However, the classical study of these data using the bivariate ordered approach remains to be done, so a side-by-side comparison to Li and Tobias's Bayesian alternative estimator is not possible. The endogeneity of the calculus dummy variables in (1) remains a feature of the model, so both the MLE and the Bayesian posterior are less straightforward than they might appear. Whether the results in Section 17.9.5 on the recursive bivariate probit model extend to this case also remains to be determined.

The bivariate ordered probit model has been applied in a number of settings in the recent empirical literature, including husband and wife's education levels [Magee et al. (2000)], family size [(Calhoun (1995)], and many others. In two early contributions to the field of pet econometrics, Butler and Chatterjee analyze ownership of cats and dogs (1995), and dogs and televisions (1997).

18.3.4 PANEL DATA APPLICATIONS

The ordered probit model is used to model discrete scales that represent indicators of a continuous underlying variable such as strength of preference, performance, or level of attainment. Many of the recently assembled national panel data sets contain survey questions that ask about subjective assessments of health satisfaction, or well-being, all of which are applications of this interpretation. Examples include the following:

- The European Community Household Panel (ECHP) includes questions about job satisfaction.[34]
- The British Household Panel Survey (BHPS) and the Australian HILDA data include questions about health status.[35]
- The German Socioeconomic Household Panel (GSOEP) includes questions about subjective well-being[36] and subjective assessment of health satisfaction.[37]

Ostensibly, the applications would fit well into the ordered probit frameworks already described. However, given the panel nature of the data, it will be desirable to augment the model with some accommodation of the individual heterogeneity that is likely to be present. The two standard models, fixed and random effects, have both been applied to the analyses of these survey data.

18.3.4.a Ordered Probit Models with Fixed Effects

D'Addio et al. (2003), using methodology developed by Frijters et al. (2004) and Ferrer-i-Carbonell et al. (2004), analyzed survey data on job satisfaction using the Danish

[33]See Greene (2007b).

[34]See D'Addio (2004).

[35]See Contoyannis et al. (2004).

[36]See Winkelmann (2005).

[37]See Riphahn et al. (2003) and Example 18.4.

component of the European Community Household Panel (ECHP). Their estimator for an ordered logit model is built around the logic of Chamberlain's estimator for the binary logit model. [See Section 17.7.3.] Because the approach is robust to individual specific threshold parameters and allows time-invariant variables, it differs sharply from the fixed effects models we have considered thus far as well as from the ordered probit model of Section 18.3.1.[38] Unlike Chamberlain's estimator for the binary logit model, however, their conditional estimator is not a function of minimal sufficient statistics. As such, the incidental parameters problem remains an issue.

Das and van Soest (2000) proposed a somewhat simpler approach.[39] Consider the base case ordered logit model with fixed effects,

$$y_{it}^{*} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_{it}|\mathbf{X}_i \sim \text{logistic}[0, \pi^2/3],$$

$$y_{it} = j \quad \text{if} \quad \mu_{j-1} < y_{it}^{*} < \mu_j, j = 0, 1, \ldots, J \quad \text{and} \quad \mu_{-1} = -\infty, \mu_0 = 0, \mu_J = +\infty.$$

The model assumptions imply that

$$\text{Prob}(y_{it} = \text{j}|\mathbf{X}_i) = \Lambda(\mu_j - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) - \Lambda(\mu_{j-1} - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where $\Lambda(t)$ is the cdf of the logistic distribution. Now, define a binary variable

$$w_{it,j} = 1 \text{ if } y_{it} > j, \quad j = 0, \ldots, J - 1.$$

It follows that

$$\begin{aligned}\text{Prob}[w_{it,j} = 1|\mathbf{X}_i] &= \Lambda(\alpha_i - \mu_j + \mathbf{x}_{it}'\boldsymbol{\beta}) \\ &= \Lambda(\theta_i + \mathbf{x}_{it}'\boldsymbol{\beta}).\end{aligned}$$

The j specific constant, which is the same for all individuals, is absorbed in θ_i. Thus, a fixed effects binary logit model applies to each of the $J - 1$ binary random variables, $w_{it,j}$. The method in Section 17.7.3 can now be applied to each of the $J - 1$ random samples. This provides $J - 1$ estimators of the parameter vector $\boldsymbol{\beta}$ (but no estimator of the threshold parameters). The authors propose to reconcile these different estimators by using a minimum distance estimator of the common true $\boldsymbol{\beta}$. (See Section 13.3 and 18.2.8c.) The minimum distance estimator at the second step is chosen to minimize

$$q = \sum_{j=0}^{J-1}\sum_{m=0}^{J-1}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})'[\mathbf{V}_{jm}^{-1}](\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}),$$

where $[\mathbf{V}_{jm}^{-1}]$ is the j, m block of the inverse of the $(J - 1)K \times (J - 1)K$ partitioned matrix $\mathbf{V}$ that contains Asy.Cov$[\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_m]$. The appropriate form of this matrix for a set of cross-section estimators is given in Brant (1990). Das and van Soest (2000) used the counterpart for Chamberlain's fixed effects estimator but do not provide the specifics for computing the off-diagonal blocks in $\mathbf{V}$.

[38]Cross-section versions of the ordered probit model with individual specific thresholds appear in Terza (1985a), Pudney and Shields (2000), and Greene (2009a).

[39]See Long's (1997) discussion of the "parallel regressions assumption," which employs this device in a cross-section framework.

The full ordered probit model with fixed effects, including the individual specific constants, can be estimated by unconditional maximum likelihood using the results in Section 14.9.6.d. The likelihood function is concave, so despite its superficial complexity, the estimation is straightforward.[40] (In the following application, with more than 27,000 observations and 7,293 individual effects, estimation of the full model required roughly five seconds of computation.) No theoretical counterpart to the Hsiao (1986, 2003) and Abrevaya (1997) results on the small T bias (incidental parameters problem) of the MLE in the presence of fixed effects has been derived for the ordered probit model. The Monte Carlo results in Greene (2004a) (see, as well, Section 15.5.2), suggest that biases comparable to those in the binary choice models persist in the ordered probit model as well. (See, also, Bester and Hansen (2009) and Carro (2007).) As in the binary choice case, the complication of the fixed effects model is the small sample bias, not the computation. The Das and van Soest approach finesses this problem—their estimator is consistent—but at the cost of losing the information needed to compute partial effects or predicted probabilities.

18.3.4.b Ordered Probit Models with Random Effects

The random effects ordered probit model has been much more widely used than the fixed effects model. Applications include Groot and van den Brink (2003), who studied training levels of employees, with firm effects; Winkelmann (2005), who examined subjective measures of well-being with individual and family effects; Contoyannis et al. (2004), who analyzed self-reported measures of health status; and numerous others. In the simplest case, the Butler and Moffitt (1982) quadrature method (Section 14.9.6.c) can be extended to this model.

Winkelmann (2005) used the random effects approach to analyze the **subjective well-being (SWB)** question (also coded 0 to 10) in the German Socioeconomic Panel (GSOEP) data set. The ordered probit model in this study is based on the latent regression,

$$y^*_{imt} = \mathbf{x}'_{imt}\boldsymbol{\beta} + \varepsilon_{imt} + u_{im} + v_i.$$

The independent variables include age, gender, employment status, income, family size, and an indicator for good health. An unusual feature of the model is the nested random effects (see Section 14.14.2), which include a family effect, v_i, as well as the individual family member (i in family m) effect, u_{im}. The GLS/MLE approach we applied to the linear regression model in Section 14.9.6.b is unavailable in this nonlinear setting. Winkelmann instead employed a Hermite quadrature procedure to maximize the log-likelihood function.

18.14 Example Health Satisfaction

The GSOEP German Health Care data that we have used in Examples 11.16, 17.4, and others includes a self-reported measure of health satisfaction, *HSAT*, that takes values 0, 1, . . . ,10.[41] This is a typical application of a scale variable that reflects an underlying continuous variable, "health." The frequencies and sample proportions for the reported values are as follows:

[40] See Pratt (1981).

[41] In the original data set, 40 (of 27,326) observations on this variable were coded with noninteger values between 6 and 7. For purposes of our example, we have recoded all 40 observations to 7.

HSAT	*Frequency*	*Proportion (%)*
0	447	1.6
1	255	0.9
2	642	2.3
3	1,173	4.2
4	1,390	5.0
5	4,233	15.4
6	2,530	9.2
7	4,231	15.4
8	6,172	22.5
9	3,061	11.2
10	3,192	11.6

We have fit pooled and panel data versions of the ordered probit model to these data. The model is

$$y_{it}^{*} = \beta_1 + \beta_2\, Age_{it} + \beta_3\, Income_{it} + \beta_4\, Kids_{it} + \beta_5\, Education_{it} + \beta_6\, Married_{it} + \beta_7\, Working_{it} + \varepsilon_{it} + c_i,$$

where c_i will be the common fixed or random effect. (We are interested in comparing the fixed and random effects estimators, so we have not included any time-invariant variables such as gender in the equation.) Table 18.21 lists five estimated models. (Standard errors for the estimated threshold parameters are omitted.) The first is the pooled ordered probit model. The second and third are fixed effects. Column 2 shows the unconditional fixed effects estimates using the results of Section 14.9.6.d. Column 3 shows the Das and van Soest estimator. For the minimum distance estimator, we used an inefficient weighting matrix, the block-diagonal matrix in which the *j*th block is the inverse of the *j*th asymptotic covariance matrix for the individual logit estimators. With this weighting matrix, the estimator is

$$\hat{\boldsymbol{\beta}}_{MDE} = \left[\sum_{j=0}^{9} \mathbf{V}_j^{-1}\right]^{-1} \sum_{j=0}^{9} \mathbf{V}_j^{-1}\, \hat{\boldsymbol{\beta}}_j$$

and the estimator of the asymptotic covariance matrix is approximately equal to the bracketed inverse matrix. The fourth set of results is the random effects estimator computed using the maximum simulated likelihood method. This model can be estimated using Butler and Moffitt's quadrature method; however, we found that even with a large number of nodes, the quadrature estimator converged to a point where the log likelihood was far lower than the MSL estimator, and at parameter values that were implausibly different from the other estimates. Using different starting values and different numbers of quadrature points did not change this outcome. The MSL estimator for a random constant term (see Section 15.6.3) is considerably lower but produces more reasonable results. The fifth set of results is the Mundlak form of the random effects model, which includes the group means in the models as controls to accommodate possible correlation between the latent heterogeneity and the included variables. As noted in Example 18.3, the components of the ordered choice model must be interpreted with some care. By construction, the partial effects of the variables on the probabilities of the outcomes must change sign, so the simple coefficients do not show the complete picture implied by the estimated model. Table 18.22 shows the partial effects for the pooled model to illustrate the computations.

Example 18.15 A Dynamic Ordered Choice Model:

Contoyannis, Jones, and Rice (2004) analyzed a self-assessed health (SAH) scale that ranged from 1 (very poor) to 5 (excellent) in the British Household Panel Survey. The data set examined consisted of the first eight waves of the data set, from 1991 to 1999, roughly 5,000

TABLE 18.21 Estimated Ordered Probit Models for Health Satisfaction

	(1)	(2) Fixed Effects	(3) Fixed Effects	(4) Random	(5) Random Effects Mundlak	
Variable	*Pooled*	*Uncond.*	*Conditional*	*Effects*	*Variables*	*Means*
Constant	2.4739			3.8577	3.2603	
	(0.04669)			(0.05072)	(0.05323)	
Age	−0.01913	−0.07162	−0.1011	−0.03319	−0.06282	0.03940
	(0.00064)	(0.002743)	(0.002878)	(0.00065)	(0.00234)	(0.00244)
Income	0.1811	0.2992	0.4353	0.09436	0.2618	0.1461
	(0.03774)	(0.07058)	(0.07462)	(0.03632)	(0.06156)	(0.07695)
Kids	0.06081	−0.06385	−0.1170	0.01410	−0.05458	0.1854
	(0.01459)	(0.02837)	(0.03041)	(0.01421)	(0.02566)	(0.03129)
Education	0.03421	0.02590	0.06013	0.04728	0.02296	0.02257
	(0.002828)	(0.02677)	(0.02819)	(0.002863)	(0.02793)	(0.02807)
Married	0.02574	0.05157	0.08505	0.07327	0.04605	−0.04829
	(0.01623)	(0.04030)	(0.04181)	(0.01575)	(0.03506)	(0.03963)
Working	0.1292	−0.02659	−0.00797	0.07108	−0.02383	0.2702
	(0.01403)	(0.02758)	(0.02830)	(0.01338)	(0.02311)	(0.02856)
μ_1	0.1949	0.3249		0.2726	0.2752	
μ_2	0.5029	0.8449		0.7060	0.7119	
μ_3	0.8411	1.3940		1.1778	1.1867	
μ_4	1.111	1.8230		1.5512	1.5623	
μ_5	1.6700	2.6992		2.3244	2.3379	
μ_6	1.9350	3.1272		2.6957	2.7097	
μ_7	2.3468	3.7923		3.2757	3.2911	
μ_8	3.0023	4.8436		4.1967	4.2168	
μ_9	3.4615	5.5727		4.8308	4.8569	
σ_u	0.0000	0.0000		1.0078	0.9936	
ln L	−56,813.52	−41,875.63		−53,215.54	−53,070.43	

TABLE 18.22 Estimated Partial Effects: Pooled Model

HSAT	*Age*	*Income*	*Kids*	*Education*	*Married*	*Working*
0	0.0006	−0.0061	−0.0020	−0.0012	−0.0009	−0.0046
1	0.0003	−0.0031	−0.0010	−0.0006	−0.0004	−0.0023
2	0.0008	−0.0072	−0.0024	−0.0014	−0.0010	−0.0053
3	0.0012	−0.0113	−0.0038	−0.0021	−0.0016	−0.0083
4	0.0012	−0.0111	−0.0037	−0.0021	−0.0016	−0.0080
5	0.0024	−0.0231	−0.0078	−0.0044	−0.0033	−0.0163
6	0.0008	−0.0073	−0.0025	−0.0014	−0.0010	−0.0050
7	0.0003	−0.0024	−0.0009	−0.0005	−0.0003	−0.0012
8	−0.0019	0.0184	0.0061	0.0035	0.0026	0.0136
9	−0.0021	0.0198	0.0066	0.0037	0.0028	0.0141
10	−0.0035	0.0336	0.0114	0.0063	0.0047	0.0233

households. Their model accommodated a variety of complications in survey data. The latent regression underlying their ordered probit model is

$$h_{it}^{*} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{H}_{i,t-1}'\boldsymbol{\gamma} + \alpha_i + \varepsilon_{it},$$

where $\mathbf{x}_{it}$ includes marital status, race, education, household size, age, income, and number of children in the household. The lagged value, $\mathbf{H}_{i,\,t-1}$, is a set of binary variables for the observed health status in the previous period.[42] In this case, the lagged values capture state dependence—the assumption that the health outcome is redrawn randomly in each period is inconsistent with evident runs in the data. The initial formulation of the regression is a fixed effects model. To control for the possible correlation between the effects, α_i, and the regressors, and the initial conditions problem that helps to explain the state dependence, they use a hybrid of Mundlak's (1978) correction and a suggestion by Wooldridge (2010) for modeling the initial conditions,

$$\alpha_i = \alpha_0 + \bar{\mathbf{x}}'\boldsymbol{\alpha}_1 + \mathbf{H}_{i,1}'\boldsymbol{\delta} + u_i,$$

where u_i is exogenous. Inserting the second equation into the first produces a random effects model that can be fit using the quadrature method we considered earlier.

The authors were interested in transitions in the reported health status, especially to and from the highest level. Based on the balanced panel for women, the authors estimated the unconditional probabilities of transition to Excellent Health from (Excellent, Good, Fair, Poor, and Very Poor) to be (0.572, 0.150, 0.040, 0.021, 0.014).[43]

The presence of attrition complicates the analysis. The authors examined the issue in a set of tests, and found evidence of nonrandom attrition for men in the sample, but not women. (See Example 11.2 in Section 11.2.5, where we have examined their study.) Table 18.23, extracted from their Table XII, displays a few of the partial effects of most interest, the implications for the probability of reporting the highest value of SAH.[44] Several specifications were considered. Model (4) in the results includes the IPW treatment for possible attrition (see Section 17.7.7). Model (6) is the most general specification considered. Surprisingly, the income effect is extremely small. However, given the considerable inertia suggested by the transition probabilities, one might expect that it would require a large change in the covariates to induce switching out of the top cell. The mean log income in the data is about 0.5 and the proportion of responders who report *EX* is roughly 4884/23,408 = 0.2086. If log income rises by 0.1, or 20%, the average probability for *EX* would rise by only 0.1 × 0.008 = 0.0008, which is trivial. Having reported *EX* in the previous period is expected to raise the probability by 0.074 compared to the value if SAH were *GOOD* (the omitted cell is the second one), which is substantial.

TABLE 18.23 Average Partial Effects on Probability of Reporting Excellent Health

	Pooled Model (4)	***Random Effects Model (6)***
ln Income	0.004 (0.002)	0.008 (0.004)
SAH EX(*t*-1)	0.208 (0.092)	0.074 (0.035)
SAH FAIR(*t*-1)	−0.127 (0.074)	−0.061 (0.033)

[42]This is the same device that was used by Butler et al. (1994) in Example 18.13. Van Ooijen, Alessie, and Knoef (2015) also analyzed self-assessed health in the context of a dynamic ordered choice model, using the Dutch Longitudinal Internet Study in the Social Sciences.

[43]Figures from Contoyannis, Jones, and Rice (2004), Table II.

[44]Contoyannis et al. (2004).

18.3.5 EXTENSIONS OF THE ORDERED PROBIT MODEL

The basic specification of the ordered probit model can be extended in the same directions as we considered in constructing models for binary choice in Chapter 17. These include heteroscedasticity in the random utility function[45] and heterogeneity in the preferences (i.e., random parameters and latent classes).[46] Two specification issues that are specific to the ordered choice model are accommodating heterogeneity in the threshold parameters and reconciling differences in the meaning of the preference scale across different groups. We will sketch the model extensions in this section. Further details are given in Chapters 6 and 7 of Greene and Hensher (2010a).

18.3.5.a Threshold Models—Generalized Ordered Choice Models

The model analyzed thus far assumes that the thresholds μ_j are the same for every individual in the sample. Terza (1985a), Pudney and Shields (2000), King, Murray, Salomon, and Tandon (KMST, 2004), Boes and Winkelmann (2006a), Greene, Harris, Hollingsworth and Maitra (2008), and Greene and Hensher (2010a) all present applications that include individual variation in the thresholds of the ordered choice model.

In his analysis of bond ratings, Terza (1985a) suggested the generalization,

$$\mu_{ij} = \mu_j + \mathbf{x}_i'\boldsymbol{\delta}.$$

With three outcomes, the probabilities are formed from

$$y_i^* = \boldsymbol{\alpha} + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

and

$$\begin{aligned} y_i = 0 \text{ if } & y_i^* \leq 0, \\ 1 \text{ if } & 0 < y_i^* \leq \mu + \mathbf{x}_i'\boldsymbol{\delta}, \\ 2 \text{ if } & y_i^* > \mu + \mathbf{x}_i'\boldsymbol{\delta}. \end{aligned}$$

For three outcomes, the model has two thresholds, $\mu_0 = 0$ and $\mu_1 = \mu + \mathbf{x}_i'\boldsymbol{\delta}$. The three probabilities can be written

$$\begin{aligned} P_0 &= \text{Prob}(y_i = 0 \,|\, \mathbf{x}_i) = \Phi[-(\alpha + \mathbf{x}_i'\boldsymbol{\beta})], \\ P_1 &= \text{Prob}(y_i = 1 \,|\, \mathbf{x}_i) = \Phi[(\mu + \mathbf{x}_i'\boldsymbol{\delta}) - (\alpha + \mathbf{x}_i'\boldsymbol{\beta})] - \Phi[-(\alpha + \mathbf{x}_i'\boldsymbol{\beta})], \\ P_2 &= \text{Prob}(y_i = 2 \,|\, \mathbf{x}_i) = 1 - \Phi[(\mu + \mathbf{x}_i'\boldsymbol{\delta}) - (\alpha + \mathbf{x}_i'\boldsymbol{\beta})]. \end{aligned}$$

For applications of this approach, see, for example, Kerkhofs and Lindeboom (1995), Groot and van den Brink (2003), and Lindeboom and van Doorslayer (2003). Note that if $\boldsymbol{\delta}$ is unrestricted, then $\text{Prob}(y_i = 1 \,|\, \mathbf{x}_i)$ can be negative. This is a shortcoming of the model when specified in this form. Subsequent development of the generalized model involves specifications that avoid this internal inconsistency. Note, as well, that if the model is recast in terms of μ and $\boldsymbol{\gamma} = [\alpha, (\boldsymbol{\beta} - \boldsymbol{\delta})]$, then the model is not distinguished from the original ordered probit model with a constant threshold parameter. This identification issue emerges prominently in Pudney and Shield's (2000) continued development of this model.

---

[45] See Section 17.5.2, Keele and Park (2005), and Wang and Kockelman (2005), for an application.

[46] An extensive study of heterogeneity in health satisfaction based on 22 waves of the GSOEP is Jones and Schurer (2010).

Pudney and Shields's (2000) "generalized ordered probit model" was also formulated to accommodate *observable* individual heterogeneity in the threshold parameters. Their application was in the context of job promotion for UK nurses in which the steps on the promotion ladder are individual specific. In their setting, in contrast to Terza's, some of the variables in the threshold equations are explicitly different from those in the regression. The authors constructed a generalized model and a test of threshold constancy by defining $\mathbf{q}_i$ to include a constant term and those variables that are unique to the threshold model. Variables that are common to both the thresholds and the regression are placed in $\mathbf{x}_i$ and the model is reparameterized as

$$\Pr(y_i = g|\mathbf{x}_i, \mathbf{q}_i) = \Phi[\mathbf{q}_i'\boldsymbol{\delta}_g - \mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\delta}_g)] - \Phi[\mathbf{q}_i'\boldsymbol{\delta}_{g-1} - \mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\delta}_{g-1})].$$

An important point noted by the authors is that the same model results if these common variables are placed in the thresholds instead. This is a minor algebraic result, but it exposes an ambiguity in the interpretation of the model—whether a particular variable affects the regression or the thresholds is one of the issues that was developed in the original model specification.

As will be evident in the application in the next section, the specification of the threshold parameters is a crucial feature of the ordered choice model. KMST (2004), Greene (2007b), Eluru, Bhat, and Hensher (2008), and Greene and Hensher (2010a) employ a hierarchical ordered probit, or HOPIT model,

$$\begin{aligned}
y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \\
y_i &= j \text{ if } \mu_{i,j-1} \leq y_i^* < \mu_{ij}, \\
\mu_0 &= 0, \\
\mu_{i,j} &= \exp(\lambda_j + \mathbf{z}_i'\boldsymbol{\gamma}) \quad \text{(case 1)}, \\
\text{or } \mu_{i,j} &= \exp(\lambda_j + \mathbf{z}_i'\boldsymbol{\gamma}_j) \quad \text{(case 2)}.
\end{aligned}$$

Case 2 is the Terza (1985a) and Pudney and Shields's (2000) model with an exponential rather than linear function for the thresholds. This formulation addresses two problems: (1) the thresholds are mathematically distinct from the regression; (2) by this construction, the threshold parameters must be positive. With a slight modification, the ordering of the thresholds can also be imposed. In case 1,

$$\mu_{i,j} = [\exp(\lambda_1) + \exp(\lambda_2) + \cdots + \exp(\lambda_j)] \times \exp(\mathbf{z}_i'\boldsymbol{\gamma}),$$

and in case 2,

$$\mu_{i,j} = \mu_{i,j-1} + \exp(\lambda_j + \mathbf{z}_i'\boldsymbol{\gamma}_j).$$

In practical terms, the model can now be fit with the constraint that all predicted probabilities are greater than zero. This is a numerical solution to the problem of ordering the thresholds for all data vectors.

This extension of the ordered choice model shows a case of **identification through functional form**. As we saw in the previous two models, the parameters $(\lambda_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta})$ would not be separately identified if all the functions were linear. The contemporary literature views models that are unidentified without a change in functional form with some skepticism. However, the underlying theory of this model does not insist on linearity of

the thresholds (or the utility function, for that matter), but it *does* insist on the ordering of the thresholds, and one might equally criticize the original model for being unidentified because the model builder insists on a linear form. That is, there is no obvious reason that the threshold parameters must be linear functions of the variables, or that linearity enjoys some claim to first precedence in the utility function. This is a methodological issue that cannot be resolved here. The nonlinearity of the preceding specification, or others that resemble it, does provide the benefit of a simple way to achieve other fundamental results, for example, coherency of the model (all positive probabilities).

18.3.5.b Thresholds and Heterogeneity—Anchoring Vignettes

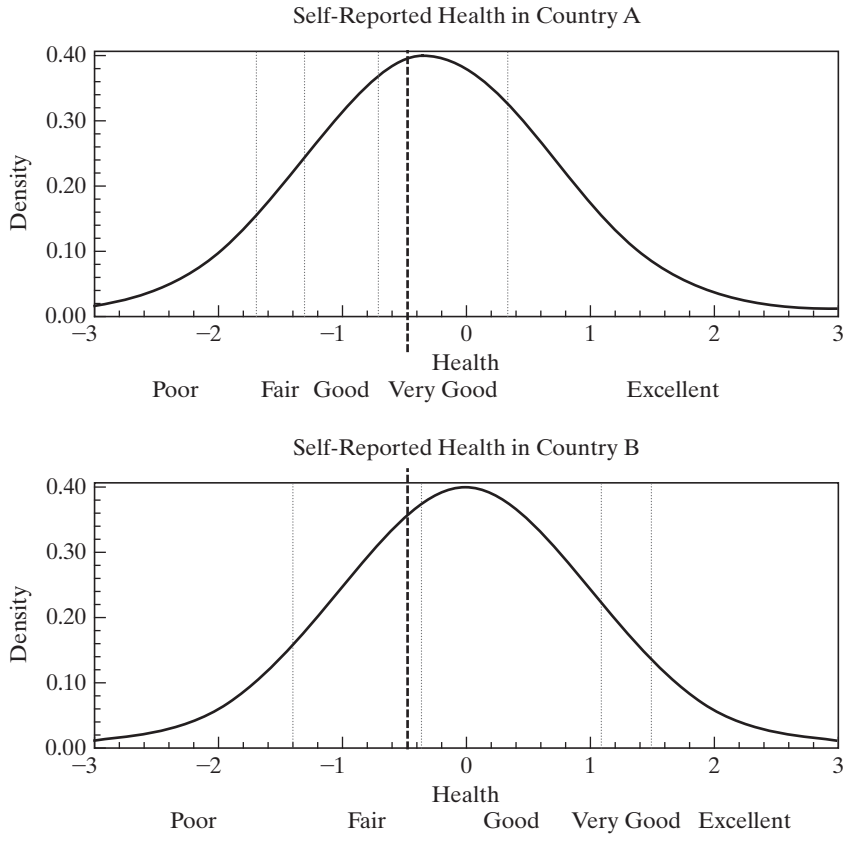
The introduction of observed heterogeneity into the threshold parameters attempts to deal with a fundamentally restrictive assumption of the ordered choice model. Survey respondents rarely view the survey questions exactly the same way. This is certainly true in surveys of health satisfaction or subjective well-being.[47] KMST (2004) identify two very basic features of survey data that will make this problematic. First, they often measure concepts that are definable only with reference to examples, such as freedom, health, satisfaction, and so on. Second, individuals do, in fact, often understand survey questions very differently, particularly with respect to answers at the extremes. A widely used term for this interpersonal incomparability is **differential item functioning (DIF)**. Kapteyn, Smith, and Van Soest (KSV, 2007) and Van Soest et al. (2007) suggest the results in Figure 18.6 to describe the implications of DIF. The figure shows the distribution of Health (or drinking behavior in the latter study) in two hypothetical countries. The density for country A (the upper figure) is to the left of that for country B, implying that, on average, people in country A are less healthy than those in country B. But the people in the two countries culturally offer very different response scales if asked to report their health on a five-point scale, as shown. In the figure, those in country A have a much more positive view of a given, objective health status than those in country B. A person in country A with health status indicated by the dotted line would report that he or she is in "Very Good" health while a person in country B with the same health status would report only "Fair." A simple frequency of the distribution of self-assessments of health status in the two countries would suggest that people in country A are much healthier than those in country B when, in fact, the opposite is true. Correcting for the influences of DIF in such a situation would be essential to obtaining a meaningful comparison of the two countries. The impact of DIF is an accepted feature of the model within a population but could be strongly distortionary when comparing very disparate groups, such as across countries, as in KMST (political groups), Murray, Tandon, Mathers, and Sudana (2002) (health outcomes), Tandon et al. (2004), and KSV (work disability), Sirven, Santos-Egglmann, and Spagnoli (2008), and Gupta, Kristensens, and Possoli (2008) (health), Angelini et al. (2008) (life satisfaction), Kristensen and Johansson (2008), and Bago d'Uva et al. (2008), all of whom used the ordered probit model to make cross-group comparisons.

KMST proposed the use of *anchoring vignettes* to resolve this difference in perceptions across groups.[48] The essential approach is to use a series of examples that, it is believed, all respondents will agree on to estimate each respondent's DIF and correct for it. The idea of using vignettes to anchor perceptions in survey questions is not itself

[47]See Boes and Winkelmann (2006b) and Ferrer-i-Carbonell and Frijters (2004).

[48]See also Kristensen and Johansson (2008).

FIGURE 18.6 Differential Item Functioning in Ordered Choices.



new; KMST cite a number of earlier uses. The innovation is their method for incorporating the approach in a formal model for ordered choices. The bivariate and multivariate probit models that they develop combine the elements described in Sections 18.3.1 through 18.3.3 and the HOPIT model in Section 18.3.5.

18.4 MODELS FOR COUNTS OF EVENTS

We have encountered behavioral variables that involve counts of events at several points in this text. In Examples 14.13 and 17.33, we examined the number of times an individual visited the physician using the GSOEP data. The credit default data that we used in Example 17.21 also includes another behavioral variable, the number of derogatory reports in an individual's credit history. Finally, in Example 17.36, we analyzed data on firm innovation. Innovation is often analyzed in terms of the number of patents that the firm obtains (or applies for).[49] In each of these cases, the variable of interest is a count

[49]For example, by Hausman, Hall, and Griliches (1984) and many others.

of events. This obviously differs from the discrete dependent variables we have analyzed in the previous two sections. A count is a quantitative measure that is, at least in principle, amenable to analysis using multiple linear regression. However, the typical preponderance of zeros and small values and the discrete nature of the outcome variable suggest that the regression approach can be improved by a method that explicitly accounts for these aspects.

Like the basic multinomial logit model for unordered data in Section 18.2 and the simple probit and logit models for binary and ordered data in Sections 17.2 and 18.3, the Poisson regression model is the fundamental starting point for the analysis of count data. We will develop the elements of modeling for count data in this framework in Sections 18.4.1 through 18.4.3, and then turn to more elaborate, flexible specifications in subsequent sections. Sections 18.4.4 and 18.4.5 will present the negative binomial and other alternatives to the Poisson functional form. Section 18.4.6 will describe the implications for the model specification of some complicating features of observed data, truncation, and censoring. Truncation arises when certain values, such as zero, are absent from the observed data because of the sampling mechanism, not as a function of the data-generating process. Data on recreation site visitation that are gathered at the site, for example, will, by construction, not contain any zeros. Censoring arises when certain ranges of outcomes are all coded with the same value. In the example analyzed the response variable is censored at 12, though values larger than 12 are possible in the field. As we have done in the several earlier treatments, in Section 18.4.7, we will examine extensions of the count data models that are made possible when the analysis is based on panel data. Finally, Section 18.4.8 discusses some behavioral models that involve more than one equation. For an example, based on the large number of zeros in the observed data, it appears that our count of doctor visits might be generated by a two-part process, a first step in which the individual decides whether or not to visit the physician at all, and a second decision, given the first, how many times to do so. The hurdle model that applies here and some related variants are discussed in Sections 18.4.8 and 18.4.9.

18.4.1 THE POISSON REGRESSION MODEL

The Poisson regression model specifies that each y_i is drawn from a Poisson population with parameter λ_i, which is related to the regressors $\mathbf{x}_i$. The primary equation of the model is

$$\text{Prob}(Y = y_i|\mathbf{x}_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots. \tag{18-17}$$

The most common formulation for λ_i is the loglinear model,

$$\ln \lambda_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

It is easily shown that the expected number of events per period or per unit of space is given by

$$E[y_i|\mathbf{x}_i] = \text{Var}[y_i|\mathbf{x}_i] = \lambda_i = e^{\mathbf{x}_i'\boldsymbol{\beta}},$$

so

$$\frac{\partial E[y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i\boldsymbol{\beta}.$$

With the parameter estimates in hand, this vector can be computed using any data vector desired or averaged across the sample to estimate the average partial effects. Because the model to this point is a straightforward regression, computation of treatment effects (at this point) is simple as well. For *exogenous* treatment indicator, T,

$$E[y|\mathbf{x}, T] = \exp(\mathbf{x}'\boldsymbol{\beta} + \gamma T).$$

So, average treatment effects can be estimated with

$$\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}[\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{\gamma}) - \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})].$$

ATET is computed by averaging over only those observations with T = 1. The case of endogenous treatment is more complicated, as usual, and is examined in Section 18.4.9.

In principle, the Poisson model is simply a nonlinear regression. But it is easier to estimate the parameters with maximum likelihood techniques. The log-likelihood function is

$$\ln L = \sum_{i=1}^{n}[-\lambda_i + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln y_i!].$$

The likelihood equations are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}(y_i - \lambda_i)\mathbf{x}_i = \mathbf{0}.$$

The Hessian is

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n}\lambda_i\mathbf{x}_i\mathbf{x}_i'.$$

The Hessian is negative definite for all $\mathbf{x}$ and $\boldsymbol{\beta}$. Newton's method is a simple algorithm for this model and will usually converge rapidly. At convergence, $\left[\sum_{i=1}^{n}\hat{\lambda}_i\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$ provides an estimator of the asymptotic covariance matrix for the parameter estimator.

There are a variety of extensions of the Poisson model—some considered later in Section 18.4.5—that introduce heterogeneity or relax the assumption of equidispersion. In general, the implication of these extensions is upon the (heteroscedastic) variance of the random variable. The conditional mean function remains the same; $E[y|\mathbf{x}] = \lambda(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})$. A consequence is that the Poisson log likelihood will provide a consistent ML estimator of $\boldsymbol{\beta}$ even in the presence of a wide variety of failures of the Poisson model assumptions. Thus, the Poisson MLE is one of the fundamental examples of a QMLE. In these settings, it is generally appropriate to adjust the estimated asymptotic covariance matrix of the estimator. For this case, a robust covariance matrix is computed using

$$[-\mathbf{H}]^{-1}(\mathbf{G}'\mathbf{G})[-\mathbf{H}]^{-1} = \left[\sum_{i=1}^{n}\hat{\lambda}_i\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{n}(y_i - \hat{\lambda}_i)^2\mathbf{x}_i\mathbf{x}_i'\right]\left[\sum_{i=1}^{n}\hat{\lambda}_i\mathbf{x}_i\mathbf{x}_i'\right]^{-1}.$$

Given the estimates, the prediction for observation i is $\hat{\lambda}_i = \exp(\mathbf{x}_i\,\hat{\boldsymbol{\beta}})$. A standard error for the prediction interval can be formed by using the delta method (see Section 4.6).

The estimated variance of the prediction will be $\hat{\lambda}_i^2 \mathbf{x}_i' \mathbf{V} \mathbf{x}_i$, where $\mathbf{V}$ is the estimated asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$.

For testing hypotheses, the three standard tests are very convenient in this model. The Wald statistic is computed as usual. As in any discrete choice model, the likelihood ratio test has the intuitive form

$$\text{LR} = 2\sum_{i=1}^{n} \ln\left(\frac{\hat{P}_i}{\hat{P}_{\text{restricted},i}}\right),$$

where the probabilities in the denominator are computed with using the restricted model. Using the BHHH estimator for the asymptotic covariance matrix, the LM statistic is simply

$$\text{LM} = \left[\sum_{i=1}^{n} \mathbf{x}_i (y_i - \hat{\lambda}_i)\right]' \left[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' (y_i - \hat{\lambda}_i)^2\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i (y_i - \hat{\lambda}_i)\right] = \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i}, \tag{18-18}$$

where each row of $\mathbf{G}$ is simply the corresponding row of $\mathbf{X}$ multiplied by $e_i = (y_i - \hat{\lambda}_i)$, $\hat{\lambda}_i$ is computed using the restricted coefficient vector, and $\mathbf{i}$ is a column of ones. Characteristically, the LM statistic can be computed as nR^2 in the regression of a column of ones on $\mathbf{g}_i = e_i \mathbf{x}_i$.

18.4.2 MEASURING GOODNESS OF FIT

The Poisson model produces no natural counterpart to the R^2 in a linear regression model, as usual, because the conditional mean function is nonlinear and, moreover, because the regression is heteroscedastic. But many alternatives have been suggested.[50] A measure based on the standardized residuals is

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n} \left[\dfrac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}\right]^2}{\sum_{i=1}^{n} \left[\dfrac{y_i - \bar{y}}{\sqrt{\bar{y}}}\right]^2}.$$

This measure has the virtue that it compares the fit of the model with that provided by a model with only a constant term. But it can be negative, and it can rise when a variable is dropped from the model. For an individual observation, the **deviance** is

$$d_i = 2[y_i \ln(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i)] = 2[y_i \ln(y_i/\hat{\lambda}_i) - e_i],$$

where, by convention, $0 \ln(0) = 0$. If the model contains a constant term, then $\sum_{i=1}^{n} e_i = 0$. The sum of the deviances,

$$G^2 = \sum_{i=1}^{n} d_i = 2\sum_{i=1}^{n} y_i \ln(y_i/\hat{\lambda}_i),$$

is reported as an alternative fit measure by some computer programs. This statistic will equal 0.0 for a model that produces a perfect fit. (*Note:* because y_i is an integer while the

[50] See the surveys by Cameron and Windmeijer (1993), Gurmu and Trivedi (1994), and Greene (2005).

prediction is continuous, it could not happen.) Cameron and Windmeijer (1993) suggest that the fit measure based on the deviances,

$$R_d^2 = 1 - \frac{\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i) \right]}{\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\bar{y}}\right) \right]},$$

has a number of desirable properties. First, denote the log-likelihood function for the model in which ψ_i is used as the prediction (e.g., the mean) of y_i as $\ell(\psi_i, y_i)$. The Poisson model fit by MLE is, then, $\ell(\hat{\lambda}_i, y_i)$, the model with only a constant term is $\ell(\bar{y}, y_i)$, and a model that achieves a perfect fit (by predicting y_i with itself) is $\ell(y_i, y_i)$. Then,

$$R_d^2 = \frac{\ell(\hat{\lambda}, y_i) - \ell(\bar{y}, y_i)}{\ell(y_i, y_i) - \ell(\bar{y}, y_i)}.$$

Both numerator and denominator measure the improvement of the model over one with only a constant term. The denominator measures the maximum improvement, since one cannot improve on a perfect fit. Hence, the measure is bounded by zero and one and increases as regressors are added to the model.[51] We note, finally, the passing resemblance of R_d^2 to the "*pseudo-*R^2," or "likelihood ratio index" reported by some statistical packages (for example, *Stata*),

$$R_{\text{LRI}}^2 = 1 - \frac{\ell(\hat{\lambda}_i, y_i)}{\ell(\bar{y}, y_i)}.$$

Many modifications of the Poisson model have been analyzed by economists. In this and the next few sections, we briefly examine a few of them.

18.4.3 TESTING FOR OVERDISPERSION

The Poisson model has been criticized because of its implicit assumption that the variance of y_i equals its mean. Many extensions of the Poisson model that relax this assumption have been proposed by Hausman, Hall, and Griliches (1984), McCullagh and Nelder (1983), and Cameron and Trivedi (1986), to name but a few.

The first step in this extended analysis is usually a test for overdispersion in the context of the simple model. A number of authors have devised tests for "overdispersion" within the context of the Poisson model. [See Cameron and Trivedi (1990), Gurmu (1991), and Lee (1986).] We will consider three of the common tests, one based on a regression approach, one a conditional moment test, and a third, a Lagrange multiplier test, based on an alternative model.

Cameron and Trivedi (1990) offer several different tests for overdispersion. A simple regression-based procedure used for testing the hypothesis

$$H_0\text{: Var}[y_i] = E[y_i],$$
$$H_1\text{: Var}[y_i] = E[y_i] + \alpha g(E[y_i]),$$

[51] Note that multiplying both numerator and denominator by 2 produces the ratio of two likelihood ratio statistics, each of which is distributed as chi squared.

is carried out by regressing

$$z_i = \frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i\sqrt{2}},$$

where $\hat{\lambda}_i$ is the predicted value from the regression, on either a constant term or $\hat{\lambda}_i$ without a constant term. A simple t test of whether the coefficient is significantly different from zero tests H_0 versus H_1.

The next section presents the **negative binomial model**. This model relaxes the Poisson assumption that the mean equals the variance. The Poisson model is obtained as a parametric restriction on the negative binomial model, so a Lagrange multiplier test can be computed. In general, if an alternative distribution for which the Poisson model is obtained as a parametric restriction, such as the negative binomial model, can be specified, then a Lagrange multiplier statistic can be computed.[52] The LM statistic is

$$\text{LM} = \left[\frac{\sum_{i=1}^{n}\hat{w}_i[(y_i - \hat{\lambda}_i)^2 - y_i]}{\sqrt{2\sum_{i=1}^{n}\hat{w}_i\hat{\lambda}_i^2}}\right]^2. \quad \textbf{(18-19)}$$

The weight, $\hat{w}_i$, depends on the assumed alternative distribution. For the negative binomial model discussed later, $\hat{w}_i$ equals 1.0. Thus, under this alternative, the statistic is particularly simple to compute:

$$\text{LM} = \frac{(\mathbf{e}'\mathbf{e} - n\bar{y})^2}{2 - \hat{\boldsymbol{\lambda}}'\hat{\boldsymbol{\lambda}}}. \quad \textbf{(18-20)}$$

The main advantage of this test statistic is that one need only estimate the Poisson model to compute it. Under the hypothesis of the Poisson model, the limiting distribution of the LM statistic is chi squared with one degree of freedom.

18.4.4 HETEROGENEITY AND THE NEGATIVE BINOMIAL REGRESSION MODEL

The assumed equality of the conditional mean and variance functions is typically taken to be the major shortcoming of the Poisson regression model. Many alternatives have been suggested.[53] The most common is the negative binomial model, which arises from a natural formulation of cross-section heterogeneity. [See Hilbe (2007).] We generalize the Poisson model by introducing an individual, unobserved effect into the conditional mean,

$$\ln \mu_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i = \ln \lambda_i + \ln u_i,$$

where the disturbance ε_i reflects either specification error, as in the classical regression model, or the kind of cross-sectional heterogeneity that normally characterizes microeconomic data. Then, the distribution of y_i conditioned on $\mathbf{x}_i$ *and* u_i (i.e., ε_i) remains Poisson with conditional mean and variance μ_i:

$$f(y_i|\,\mathbf{x}_i, u_i) = \frac{e^{-(\lambda_i u_i)}(\lambda_i u_i)^{y_i}}{y_i!}.$$

[52]See Cameron and Trivedi (1986, p. 41).

[53]See Hausman, Hall, and Griliches (1984), Cameron and Trivedi (1986, 1998), Gurmu and Trivedi (1994), Johnson and Kotz (1993), and Winkelmann (2005) for discussion.

The unconditional distribution $f(y_i|\mathbf{x}_i)$ is the expected value (over u_i) of $f(y_i|\mathbf{x}_i, u_i)$,

$$f(y_i|\mathbf{x}_i) = \int_0^\infty \frac{e^{-(\lambda_i u_i)}(\lambda_i u_i)^{y_i}}{y_i!} g(u_i)\, du_i.$$

The choice of a density for u_i defines the unconditional distribution. For mathematical convenience, a gamma distribution is usually assumed for $u_i = \exp(\varepsilon_i)$.[54] As in other models of heterogeneity, the mean of the distribution is unidentified if the model contains a constant term (because the disturbance enters multiplicatively) so $E[\exp(\varepsilon_i)]$ is assumed to be 1.0. With this normalization,

$$g(u_i) = \frac{\theta^\theta}{\Gamma(\theta)} e^{-\theta u_i} u_i^{\theta-1}.$$

The density for y_i is then

$$\begin{aligned} f(y_i|\mathbf{x}_i) &= \int_0^\infty \frac{e^{-(\lambda_i u_i)}(\lambda_i u_i)^{y_i}}{y_i!} \frac{\theta^\theta u_i^{\theta-1} e^{-\theta u_i}}{\Gamma(\theta)} du_i \\ &= \frac{\theta^\theta \lambda_i^{y_i}}{\Gamma(y_i + 1)\Gamma(\theta)} \int_0^\infty e^{-(\lambda_i+\theta)u_i} u_i^{\theta+y_i-1} du_i \\ &= \frac{\theta^\theta \lambda_i^{y_i} \Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)(\lambda_i + \theta)^{\theta+y_i}} \\ &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} r_i^{y_i}(1 - r_i)^\theta, \qquad \text{where } r_i = \frac{\lambda_i}{\lambda_i + \theta}, \end{aligned}$$

which is one form of the **negative binomial distribution**. The distribution has conditional mean λ_i and conditional variance $\lambda_i(1 + (1/\theta)\lambda_i)$.[55] The negative binomial model can be estimated by maximum likelihood without much difficulty. A test of the Poisson distribution is often carried out by testing the hypothesis $\alpha = 1/\theta = 0$ using the Wald or likelihood ratio test.

18.4.5 FUNCTIONAL FORMS FOR COUNT DATA MODELS

The equidispersion assumption of the Poisson regression model, $E[y_i|\mathbf{x}_i] = \text{Var}[y_i|\mathbf{x}_i]$, is a major shortcoming. Observed data rarely, if ever, display this feature. The very large amount of research activity on functional forms for count models is often focused on testing for equidispersion and building functional forms that relax this assumption. In practice, the Poisson model is typically only the departure point for an extended specification search.

One easily remedied minor issue concerns the units of measurement of the data. In the Poisson and negative binomial models, the parameter λ_i is the expected number of events *per unit of time or space*. Thus, there is a presumption in the model formulation, for example, the Poisson, that the same amount of time is observed for each i. In a spatial

[54] An alternative approach based on the normal distribution is suggested in Terza (1998), Greene (1995b, 1997, 2005), Winkelmann (2003), and Riphahn, Wambach, and Million (2003). The normal-Poisson mixture is also easily extended to the random effects model discussed in the next section. There is no closed form for the normal-Poisson mixture model, but it can be easily approximated by using Hermite quadrature or simulation. See Sections 14.14.4 and 17.7.2.

[55] This model is Negbin 2 in Cameron and Trivedi's (1986) presentation.

context, such as measurements of the prevalence of a disease per group of N_i persons, or the number of bomb craters per square mile (London, 1940), the assumption would be that the same physical area or the same size of population applies to each observation. Where this differs by individual, it will introduce a type of heteroscedasticity in the model. The simple remedy is to modify the model to account for the **exposure**, T_i, of the observation as follows:

$$\text{Prob}(y_i = j|\mathbf{x}_i, T_i) = \frac{\exp(-T_i\phi_i)(T_i\phi_i)^j}{j!}, \quad \phi_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), j = 0, 1, \ldots.$$

The original model is returned if we write $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \ln T_i)$. Thus, when the exposure differs by observation, the appropriate accommodation is to include the log of exposure in the regression part of the model with a coefficient of 1.0. (For less than obvious reasons, the term *offset variable* is commonly associated with the exposure variable T_i.) Note that if T_i is the same for all i, $\ln T_i$ will simply vanish into the constant term of the model (assuming one is included in $\mathbf{x}_i$).

The recent literature, mostly associating the result with Cameron and Trivedi's (1986, 1998) work, defines two familiar forms of the negative binomial model. The **Negbin 2 (NB2) form** of the probability is

$$\begin{aligned}\text{Prob}(Y = y_i|\mathbf{x}_i) &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} r_i^{y_i}(1 - r_i)^\theta, \\ \lambda_i &= \exp(\mathbf{x}_i'\boldsymbol{\beta}), \\ r_i &= \lambda_i/(\theta + \lambda_i).\end{aligned} \tag{18-21}$$

This is the default form of the model in the standard econometrics packages that provide an estimator for this model. The **Negbin 1 (NB1) form** of the model results if θ in the preceding is replaced with $\theta_i = \theta\lambda_i$. Then, r_i reduces to $r = 1/(1 + \theta)$, and the density becomes

$$\text{Prob}(Y = y_i|\mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i)} r^{y_i}(1 - r)^{\theta\lambda_i}. \tag{18-22}$$

This is not a simple reparameterization of the model. The results in Example 18.15 demonstrate that the log-likelihood functions are not equal at the maxima, and the parameters are not simple transformations in one model versus the other. We are not aware of a theory that justifies using one form or the other for the negative binomial model. Neither is a restricted version of the other, so we cannot carry out a likelihood ratio test of one versus the other. The more general **Negbin *P* (NBP)** family does nest both of them, so this may provide a more general, encompassing approach to finding the right specification. [See Greene (2005, 2008b).] The Negbin P model is obtained by replacing θ in the Negbin 2 form with $\theta\lambda_i^{2-P}$. We have examined the cases of $P = 1$ and $P = 2$ in (18-21) and (18-22). The full model is

$$\text{Prob}(Y = y_i|\mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i^Q + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i^Q)} \left(\frac{\lambda_i}{\theta\lambda_i^Q + \lambda_i}\right)^{y_i} \left(\frac{\theta\lambda_i^Q}{\theta\lambda_i^Q + \lambda_i}\right)^{\theta\lambda_i^Q}, Q = 2 - P.$$

The conditional mean function for the three cases considered is

$$E[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}) = \lambda_i.$$

The parameter P is picking up the scaling. A general result is that for all three variants of the model,

$$\text{Var}[y_i|\mathbf{x}_i] = \lambda_i(1 + \alpha\lambda_i^{P-1}), \quad \text{where } \alpha = 1/\theta.$$

Thus, the NB2 form has a variance function that is quadratic in the mean while the NB1 form's variance is a simple multiple of the mean. There have been many other functional forms proposed for count data models, including the generalized Poisson, gamma, and Polya-Aeppli forms described in Winkelmann (2003) and Greene (2016).

The heteroscedasticity in the count models is induced by the relationship between the variance and the mean. The single parameter θ picks up an implicit overall scaling, so it does not contribute to this aspect of the model. As in the linear model, microeconomic data are likely to induce heterogeneity in both the mean and variance of the response variable. A specification that allows independent variation of both will be of some virtue. The result,

$$\text{Var}[y_i|\mathbf{x}_i] = \lambda_i(1 + (1/\theta)\lambda_i^{P-1}),$$

suggests that a convenient platform for separately modeling heteroscedasticity will be the dispersion parameter, θ, which we now parameterize as

$$\theta_i = \theta \exp(\mathbf{z}_i'\boldsymbol{\delta}).$$

Operationally, this is a relatively minor extension of the model. But it is likely to introduce quite a substantial increase in the flexibility of the specification. Indeed, a heterogeneous Negbin P model is likely to be sufficiently parameterized to accommodate the behavior of most data sets. (Of course, the specialized models discussed in Section 18.4.8, for example, the zero-inflation models, may yet be more appropriate for a given situation.)

Example 18.16 Count Data Models for Doctor Visits

The study by Riphahn et al. (2003) that provided the data we have used in numerous earlier examples analyzed the two count variables *DocVis* (visits to the doctor) and *HospVis* (visits to the hospital). The authors were interested in the joint determination of these two count variables. One of the issues considered in the study was whether the data contained evidence of moral hazard, that is, whether health care utilization as measured by these two outcomes was influenced by the subscription to health insurance.[56] The data contain indicators of two levels of insurance coverage, *PUBLIC*, which is the main source of insurance, and *ADDON*, which is a secondary optional insurance. In the sample of 27,326 observations (family/years), 24,203 individuals held the public insurance. (There is quite a lot of within group variation in this. Individuals did not routinely obtain the insurance for all periods.) Of these 24,203, 23,689 had only public insurance and 514 had both types. (One could not have only the ADDON insurance.) To explore the issue, we have analyzed the *DocVis* variable with the count data models described in this section. The exogenous variables in our model are

$$\mathbf{x}_{it} = (1, \textit{Age}, \textit{Education}, \textit{Income}, \textit{Kids}, \textit{AddOn}).$$

(Variables are described in Appendix Table F7.1.)

Table 18.24 presents the estimates of the several count models. In all specifications, the coefficient on *ADDON* is positive but not statistically significant, which is consistent with the results in the authors' study. They found evidence of moral hazard in a simple model,

[56]Munkin and Trivedi (2007) is a similar application to dental insurance.

TABLE 18.24 Estimated Models for *DocVis* (standard errors in parentheses)

Variable	*Poisson*	*Negbin 2*	*Negbin 2 Heterogeneous*	*Negbin 1*	*Negbin P*	*Poisson Normal*
Constant	1.05266	1.10083	1.14129	0.93184	0.97164	0.09302
	(0.11395)	(0.05970)	(0.06175)	(0.05630)	(0.06389)	(0.04364)
Age	0.01838	0.01789	0.01689	0.01571	0.01888	0.02267
	(0.00134)	(0.00079)	(0.00081)	(0.00070)	(0.00081)	(0.00051)
Education	−0.04355	−0.04797	−0.04450	−0.03127	−0.04282	−0.04595
	(0.00699)	(0.00378)	(0.00386)	(0.00355)	(0.00414)	(0.00276)
Income	−0.52502	−0.46285	−0.45443	−0.23198	−0.37774	−0.45804
	(0.08240)	(0.04600)	(0.04654)	(0.04451)	(0.05122)	(0.03235)
Kids	−0.16109	−0.15656	−0.16266	−0.13658	−0.16521	−0.18450
	(0.03118)	(0.01735)	(0.01769)	(0.01648)	(0.01855)	(0.01217)
AddOn	0.07282	0.07134	0.06839	0.17879	0.16107	0.27067
	(0.07801)	(0.07205)	(0.07142)	(0.05493)	(0.06969)	(0.04068)
	[0.06548]					
	{0.02534}					
P	0.0000	2.0000	2.0000	1.0000	1.52377	
	—	—	—	—	(0.03485)	
α	0.0000	1.92971	2.61217	6.19585	3.34512	
σ	—	(0.02009)	(0.05965)	(0.06867)	(0.13995)	1.31484
	—	—	—	—	—	(0.00425)
	—	—	—	—	—	
δ *(Female)*	—	—	−0.38157	—	—	
	—	—	(0.02040)	—	—	
δ *(Married)*	—	—	−0.13661	—	—	
	—	—	(0.02305)	—	—	
ATE	0.24018	0.23491	0.22070	0.62105	0.55460	0.42961
	(0.26637)	(0.24561)	(0.23850)	(0.20782)	(0.25929)	(0.07399)
ATET	0.21945	0.21482	0.21781	0.59304	0.51528	0.39914
	(0.24317)	(0.22454)	(0.25055)	(0.19813)	(0.24066)	(0.06810)
ln L	−104,603.0	−60,291.50	−60,149.00	−60,274.94	−60,219.19	−60,619.11

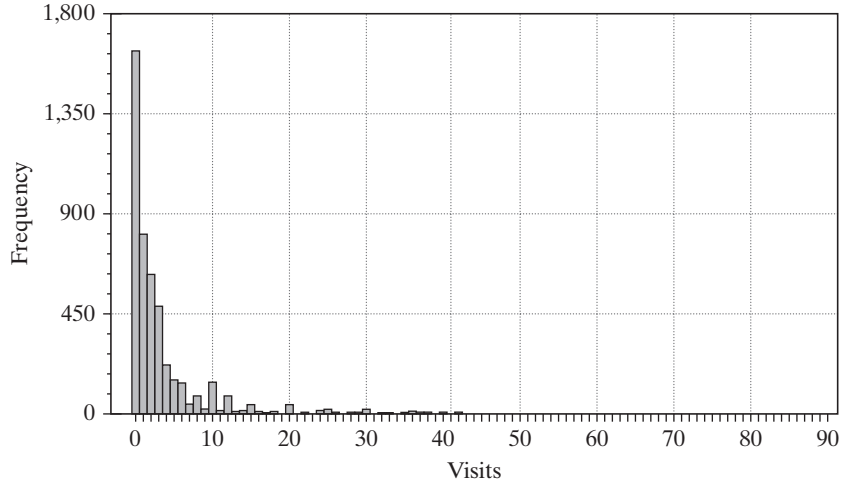
but none when their model was expanded. The various test statistics strongly reject the hypothesis of equidispersion. Cameron and Trivedi's (1990) semiparametric tests from the Poisson model (see Section 18.4.3) have t statistics of 22.151 for $g_i = \mu_i$ and 22.440 for $g_i = \mu_i^2$. Both of these are far larger than the critical value of 1.96. The LR statistic comparing to the NB model is over 80,000, which is also larger than the (any) critical value. On these bases, we would reject the hypothesis of equidispersion. The Wald and likelihood ratio tests based on the negative binomial models produce the same conclusion. For comparing the different negative binomial models, note that Negbin 2 is the worst of the four by the likelihood function, although NB1 and NB2 are not directly comparable. On the other hand, note that in the NBP model, the estimate of P is more than 10 standard errors from 1.0000 or 2.0000, so both NB1 and NB2 are rejected in favor of the unrestricted NBP form of the model. The NBP and the heterogeneous NB2 model are not nested either, but comparing the log likelihoods, it does appear that the heterogeneous model is substantially superior. We computed the Vuong statistic based on the individual contributions to the log likelihoods, with

$v_i = \ln L_i(\text{NBP}) - \ln L_i(\text{NB2-H})$. (See Section 14.6.6). The value of the statistic is −3.27. On this basis, we would reject NBP in favor of NB2-H. Finally, with regard to the original question, the ATE and ATET computed for *ADDON* are generally quite small with the Poisson and NB models—the mean of *DocVis* is about 3.2 and the effect is about 0.2 and insignificant. The effect is larger in the less restrictive NBP and normal mixture models. The evidence here, as in RWM, is mixed.

18.4.6 TRUNCATION AND CENSORING IN MODELS FOR COUNTS

Truncation and censoring are relatively common in applications of models for counts. Truncation arises as a consequence of discarding what appear to be unusable data, such as the zero values in survey data on the number of uses of recreation facilities.[57] In this setting, a more common case which also gives rise to truncation is on-site sampling. When one is interested in visitation by the entire population, which will naturally include zero visits, but one draws their sample on site, the distribution of visits is truncated at zero by construction. Every visitor has visited at least once. Shaw (1988), Englin and Shonkwiler (1995), Grogger and Carson (1991), Creel and Loomis (1990), Egan and Herriges (2006), and Martínez-Espinera and Amoako-Tuffour (2008) are studies that have treated truncation due to on-site sampling in environmental and recreation applications. Truncation will also arise when data are trimmed to remove what appear to be unusual values. Figure 18.7 displays a histogram for the number of doctor visits in the 1988 wave of the GSOEP data that we have used in several examples. There is a suspiciously large spike at zero and an extremely long right tail of what might seem to be atypical observations. For modeling purposes, it might be tempting to remove these non-Poisson appearing observations in the tails. (Other models might be a better solution.) The distribution that characterizes what remains in the sample is a truncated distribution. Truncation is not innocent. If the entire population is of interest, then

FIGURE 18.7 Number of Doctor Visits, 1988 Wave of GSOEP Data.



[57]Shaw (1988) and Bockstael et al. (1990).

conventional statistical inference (such as estimation) on the truncated sample produces a systematic bias known as (of course) truncation bias. This would arise, for example, if an ordinary Poisson model intended to characterize the full population is fit to the sample from a truncated population.

Censoring, in contrast, is generally a feature of the sampling design. In the application in Example 18.18, the dependent variable is the self-reported number of extramarital affairs in a survey taken by the magazine *Psychology Today*. The possible answers are 0, 1, 2, 3, 4 to 10 (coded as 7), and "monthly, weekly or daily" coded as 12. The two upper categories are censored. Similarly, in the doctor visits data in the previous paragraph, recognizing the possibility of truncation bias due to data trimming, we might, instead, simply censor the distribution of values at 15. The resulting variable would take values $0, \ldots, 14$, "15 or more." In both cases, applying conventional estimation methods leads to predictable biases. However, it is also possible to reconstruct the estimators specifically to account for the truncation or censoring in the data.

Truncation and censoring produce similar effects on the distribution of the random variable and on the features of the population such as the mean. For the truncation case, suppose that the original random variable has a Poisson distribution—all these results can be directly extended to the negative binomial or any of the other models considered earlier—with

$$P(y_i = j|\mathbf{x}_i) = [\exp(-\lambda_i)\lambda_i^j/j!] = P_{i,j}.$$

If the distribution is truncated at value C—that is, only values $C + 1, \ldots$ are observed—then the resulting random variable has probability distribution

$$P(y_i = j|\mathbf{x}_i, y_i > C) = \frac{P(y_i = j|\mathbf{x}_i)}{P(y_i > C|\mathbf{x}_i)} = \frac{P(y_i = j|\mathbf{x}_i)}{1 - P(y_i \leq C|\mathbf{x}_i)}.$$

The original distribution must be scaled up so that it sums to one for the cells that remain in the truncated distribution. The leading case is truncation at zero, that is, "left truncation," which, for the Poisson model produces[58]

$$P(y_i = j|\mathbf{x}_i, y_i > 0) = \frac{\exp(-\lambda_i)\lambda_i^j}{j![1 - \exp(-\lambda_i)]} = \frac{P_{i,j}}{1 - P_{i,0}}, \; j = 1, \ldots.$$

The conditional mean function is

$$E(y_i|\mathbf{x}_i, y_i > 0) = \frac{1}{[1 - \exp(-\lambda_i)]}\sum_{j=1}^{\infty}\frac{j\exp(-\lambda_i)\lambda_i^j}{j!} = \frac{\lambda_i}{[1 - \exp(-\lambda_i)]} > \lambda_i.$$

The second equality results because the sum can be started at zero—the first term is zero—and this produces the expected value of the original variable. As might be expected, truncation "from below" has the effect of increasing the expected value. It can be shown that it decreases the conditional variance, however. The partial effects are

$$\delta_i = \frac{\partial E[y_i|\mathbf{x}_i, y_i > 0]}{\partial \mathbf{x}_i} = \left[\frac{1 - P_{i,0} - \lambda_i P_{i,0}}{(1 - P_{i,0})^2}\right]\lambda_i\boldsymbol{\beta}. \qquad \textbf{(18-23)}$$

[58]See, for example, Mullahy (1986), Shaw (1988), Grogger and Carson (1991), Greene (1995a,b), and Winkelmann (2003).

The term outside the brackets is the partial effects in the absence of the truncation while the bracketed term rises from slighter greater than 0.5 to 1.0 as λ_i increases from just above zero.

Example 18.17 Major Derogatory Reports

In Examples 17.17 and 17.21, we examined a binary choice model for the accept/reject decision for a sample of applicants for a major credit card. Among the variables in that model is Major Derogatory Reports (MDRs). This is an interesting behavioral variable in its own right that can be appropriately modeled using the count data specifications in this chapter. In the sample of 13,444 individuals, 10,833 had zero MDRs while the values for the remaining 2,561 ranged from 1 to 22. This preponderance of zeros exceeds by far what one would anticipate in a Poisson model that was dispersed enough to produce the distribution of remaining individuals. As we will pursue in Example 18.18, a natural approach for these data is to treat the extremely large block of zeros explicitly in an extended model. For present purposes, we will consider the nonzero observations apart from the zeros and examine the effect of accounting for left truncation at zero on the estimated models. Estimation results are shown in Table 18.25. The first column of results compared to the second shows the suspected impact of incorrectly including the zero observations. The coefficients change only slightly, but the partial effects are far smaller when the zeros are included in the estimation. It was not possible to fit a truncated negative binomial with these data.

Censoring is handled similarly. The usual case is *right censoring*, in which realized values greater than or equal to C are all given the value C. In this case, we have a two-part distribution.[59] The observed random variable, y_i, is constructed from an underlying random variable, y_i^*, by $y_i = \text{Min}(y_i^*, C)$. Wang and Zhou (2015) applied this specification with a negative binomial count model to a study of the number of deliveries to online shoppers. The dependent variable, deliveries, ranging from 0 to 200, was censored at 10 for the analysis.

TABLE 18.25 Estimated Truncated Poison Regression Model (*t* ratios in parentheses)

	Poisson Full Sample		***Poisson***		***Truncated Poisson***	
Constant	0.8756	(17.10)	0.8698	(16.78)	0.7400	(11.99)
Age	0.0036	(2.38)	0.0035	(2.32)	0.0049	(2.75)
Income	−0.0039	(−4.78)	−0.0036	(−3.83)	−0.0051	(−4.51)
OwnRent	−0.1005	(−3.52)	−0.1020	(−3.56)	−0.1415	(−4.18)
Self-Employed	−0.0325	(−0.62)	−0.0345	(−0.66)	−0.0515	(−0.82)
Dependents	0.0445	(4.69)	0.0440	(4.62)	0.0606	(5.48)
MthsCurAdr	0.00004	(0.23)	0.0001	(0.25)	0.0001	(0.30)
ln *L*	−5,379.30		−5,378.79		−5,097.08	
			Average Partial Effects			
Age	0.0017		0.0085		0.0084	
Income	−0.0018		−0.0087		−0.0089	
OwnRent	−0.0465		−0.2477		−0.2460	
Self-Employed	−0.0150		−0.0837		−0.0895	
Dependents	0.0206		0.1068		0.1054	
MthsCurAdr	0.00002		0.0001		0.0001	
Cond'l. Mean	0.4628		2.4295		2.4295	
Scale factor	0.4628		2.4295		1.7381	

[59]See Terza (1985b).

Probabilities in the presence of censoring are constructed using the axioms of probability. This produces

$$\text{Prob}(y_i = j|\mathbf{x}_i) = P_{i,j}, j = 0, 1, \ldots, C-1,$$
$$\text{Prob}(y_i = C|\mathbf{x}_i) = \sum_{j=C}^{\infty} P_{i,j} = 1 - \sum_{j=0}^{C-1} P_{i,j}.$$

In this case, the conditional mean function is

$$\begin{aligned} E[y_i|\mathbf{x}_i] &= \sum_{j=0}^{C-1} j\, P_{i,j} + \sum_{j=C}^{\infty} C P_{i,j} \\ &= \sum_{j=0}^{\infty} j P_{i,j} - \sum_{j=C}^{\infty} (j - C) P_{i,j} \\ &= \lambda_i - \sum_{j=C}^{\infty} (j - C) P_{i,j} < \lambda_i. \end{aligned}$$

The infinite sum can be computed by using the complement. Thus,

$$\begin{aligned} E[y_i|\mathbf{x}_i] &= \lambda_i - \left[\sum_{j=0}^{\infty} (j - C) P_{i,j} - \sum_{j=0}^{C-1} (j - C) P_{i,j} \right] \\ &= \lambda_i - (\lambda_i - C) + \sum_{j=0}^{C-1} (j - C) P_{i,j} \\ &= C - \sum_{j=0}^{C-1} (C - j) P_{i,j}. \end{aligned}$$

Example 18.18 Extramarital Affairs

In 1969, the popular magazine *Psychology Today* published a 101-question survey on sex and asked its readers to mail in their answers. The results of the survey were discussed in the July 1970 issue. From the approximately 2,000 replies that were collected in electronic form (of about 20,000 received), Professor Ray Fair (1978) extracted a sample of 601 observations on men and women then currently married for the first time and analyzed their responses to a question about extramarital affairs. Fair's analysis in this frequently cited study suggests several interesting econometric questions.[60]

Fair used the tobit model that we discuss in Chapter 19 as a platform. The nonexperimental nature of the data (which can be downloaded from the Internet at http://fairmodel.econ.yale.edu/rayfair/work.ss.htm and are given in Appendix Table F18.1) provides a laboratory case that we can use to examine the relationships among the tobit, truncated regression, and probit models. Although the tobit model seems to be a natural choice for the model for these data, given the cluster of zeros, the fact that the behavioral outcome variable is a count that typically takes a small value suggests that the models for counts that we have examined in this chapter might be yet a better choice. Finally, the preponderance of zeros in the data that initially motivated the tobit model suggests that even the standard Poisson model, although an improvement, might still be inadequate. We will pursue that aspect of the data later. In this example, we will focus on just the censoring issue. Other features of the models and data are reconsidered in the exercises.

[60]In addition, his 1977 companion paper in *Econometrica* on estimation of the tobit model proposed a variant of the EM algorithm, developed by Dempster, Laird, and Rubin (1977).

The study was based on 601 observations on the following variables (full details on data coding are given in the data file and Appendix Table F18.1):

y = number of affairs in the past year, 0, 1, 2, 3, (4–10) = 7, (monthly, weekly, or daily) = 12. Sample mean = 1.46; Frequencies = (451, 34, 17, 19, 42, 38),
z_1 = sex = 0 for female, 1 for male. Sample mean = 0.476,
z_2 = age. Sample mean = 32.5,
z_3 = number of years married. Sample mean = 8.18,
z_4 = children, 0 = no, 1 = yes. Sample mean = 0.715,
z_5 = religiousness, 1 = anti, . . . , 5 = very. Sample mean = 3.12,
z_6 = education, years, 9 = grade school, 12 = high school, . . . , 20 = Ph.D or other. Sample mean = 16.2,
z_7 = occupation, "Hollingshead scale," 1–7. Sample mean = 4.19,
z_8 = self-rating of marriage, 1 = very unhappy, . . . , 5 = very happy. Sample mean = 3.93.

A tobit model was fit to y using a constant term and all eight variables. A restricted model was fit by excluding z_1, z_4, and z_6, none of which was individually statistically significant in the model. We are able to match exactly Fair's results for both equations. The tobit model should only be viewed as an approximation for these data. The dependent variable is a count, not a continuous measurement. The Poisson regression model, or perhaps one of the many variants of it, should be a preferable modeling framework. Table 18.26 presents estimates of the Poisson and negative binomial regression models. There is ample evidence of overdispersion in these data; the t ratio on the estimated overdispersion parameter is 7.015/0.945 = 7.42, which is strongly suggestive. The large absolute value of the coefficient is likewise suggestive.

Responses of 7 and 12 do not represent the actual counts. It is unclear what the effect of the first recoding would be, because it might well be the mean of the observations in this group. But the second is clearly a censored observation. To remove both of these effects, we have recoded both the values 7 and 12 as 4 and treated this observation (appropriately) as a censored observation, with 4 denoting "4 or more." As shown in the lower panel of results in Table 18.26, the effect of this treatment of the data is greatly to reduce the measured effects. Although this step does remove a deficiency in the data, it does not remove the overdispersion; at this point, the negative binomial model is still the preferred specification.

18.4.7 PANEL DATA MODELS

The familiar approaches to accommodating heterogeneity in panel data have fairly straightforward extensions in the count data setting.[61] We will examine them for the Poisson model. Hausman, Hall and Griliches (1984) and Allison (2000) also give results for the negative binomial model.

18.4.7.a Robust Covariance Matrices for Pooled Estimators

The standard asymptotic covariance matrix estimator for the Poisson model is

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \left[-\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}\, \partial \hat{\boldsymbol{\beta}}'} \right]^{-1} = \left[\sum_{i=1}^{n} \hat{\lambda}_i \mathbf{x}_i \mathbf{x}_i' \right]^{-1} = [\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}]^{-1},$$

where $\hat{\boldsymbol{\Lambda}}$ is a diagonal matrix of predicted values. The BHHH estimator is

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \left[\sum_{i=1}^{n} \left(\frac{\partial \ln P_i}{\partial \hat{\boldsymbol{\beta}}} \right) \left(\frac{\partial \ln P_i}{\partial \hat{\boldsymbol{\beta}}} \right)' \right]^{-1} = \left[\sum_{i=1_i}^{n} (y_i - \hat{\lambda}_i)^2 \mathbf{x}_i \mathbf{x}_i' \right]^{-1} = [\mathbf{X}'\hat{\mathbf{E}}^2\mathbf{X}]^{-1},$$

[61] Hausman, Hall, and Griliches (1984) give full details for these models.

TABLE 18.26 Censored Poisson and Negative Binomial Distributions

	Poisson Regression			*Negative Binomial Regression*		
Variable	*Estimate*	*Std. Error*	*Partial Effect*	*Estimate*	*Std. Error*	*Partial Effect*
			Based on Uncensored Poisson Distribution			
Constant	2.53	0.197	—	2.19	0.859	—
z_2	−0.0322	0.0059	−0.047	−0.0262	0.0180	−0.0039
z_3	0.116	0.0099	0.168	0.0848	0.0401	0.127
z_5	−0.354	0.0309	−0.515	−0.422	0.171	−0.632
z_7	0.0798	0.0194	0.116	0.0604	0.0909	0.0906
z_8	−0.409	0.0274	−0.596	−0.431	0.167	−0.646
α				7.015	0.945	
ln L	−1,427.037			−728.2441		
			Based on Poisson Distribution Right Censored at y = 4			
Constant	1.90	0.283	—	4.79	1.16	—
z_2	−0.0328	0.0084	−0.0235	−0.0166	0.0250	−0.0043
z_3	0.105	0.0140	0.0755	0.174	0.0568	0.045
z_5	−0.323	0.0437	−0.232	−0.723	0.198	−0.186
z_7	0.0798	0.0275	0.0572	0.0900	0.116	0.0232
z_8	−0.390	0.0391	−0.279	−0.854	0.216	−0.220
α				9.40	1.35	
ln L	−747.7541			−482.0505		

where $\hat{\mathbf{E}}$ is a diagonal matrix of residuals. The Poisson model is one in which the MLE is robust to certain misspecifications of the model, such as the failure to incorporate latent heterogeneity in the mean (that is, one fits the Poisson model when the negative binomial is appropriate). In this case, a robust covariance matrix is the "sandwich" estimator,

$$\text{Robust Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = [\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\mathbf{E}}^2\mathbf{X}][\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}]^{-1},$$

which is appropriate to accommodate this failure of the model. It has become common to employ this estimator with all specifications, including the negative binomial. One might question the virtue of this. Because the negative binomial model already accounts for the latent heterogeneity, it is unclear what *additional* failure of the assumptions of the model this estimator would be robust to. The questions raised in Section 14.8 about robust covariance matrices would be relevant here. However, if the model is, indeed, complete, then the robust estimator does no harm.

A related calculation is used when observations occur in groups that may be correlated. This would include a random effects setting in a panel in which observations have a common latent heterogeneity as well as more general, stratified, and clustered data sets. The parameter estimator is unchanged in this case (and an assumption is made that the estimator is still consistent), but an adjustment is made to the estimated asymptotic covariance matrix. The calculation is done as follows: Suppose the n observations are assembled in G clusters of observations, in which the number of observations in the ith cluster is n_i. Thus, $\sum_{i=1}^{G} n_i = n$. Denote by $\boldsymbol{\beta}$ the full set of model parameters in whatever variant of the model is being estimated. Let the observation-specific gradients

and Hessians be $\mathbf{g}_{ij} = \partial \ln L_{ij}/\partial\boldsymbol{\beta} = (y_{ij} - \lambda_{ij})\mathbf{x}_{ij}$ and $\mathbf{H}_{ij} = \partial^2 \ln L_{ij}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = -\lambda_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}'$. The uncorrected estimator of the asymptotic covariance matrix based on the Hessian is

$$\mathbf{V}_H = -\mathbf{H}^{-1} = \left(-\sum_{i=1}^{G}\sum_{j=1}^{n_i}\mathbf{H}_{ij}\right)^{-1}.$$

The corrected asymptotic covariance matrix is

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \mathbf{V}_H\left(\frac{G}{G-1}\right)\left[\sum_{i=1}^{G}\left(\sum_{j=1}^{n_i}\mathbf{g}_{ij}\right)\left(\sum_{j=1}^{n_i}\mathbf{g}_{ij}\right)'\right]\mathbf{V}_H.$$

Note that if there is exactly one observation per cluster, then this is $G/(G-1)$ times the sandwich (robust) estimator.

18.4.7.b Fixed Effects

With fixed effects, the Poisson distribution will have conditional mean

$$\log \lambda_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i, \tag{18-24}$$

where now $\mathbf{x}_{it}$ has been redefined to exclude the constant term. The approach used in the linear model of transforming y_{it} to group mean deviations does not remove the heterogeneity, nor does it leave a Poisson distribution for the transformed variable. However, the Poisson model with fixed effects can be fit using the methods described for the probit model in Section 17.7.3. The extension to the Poisson model requires only the minor modifications, $g_{it} = (y_{it} - \lambda_{it})$ and $h_{it} = -\lambda_{it}$. Everything else in that derivation applies with only a simple change in the notation. The first-order conditions for maximizing the log-likelihood function for the Poisson model will include

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T_i}(y_{it} - e^{\alpha_i}\mu_{it}) = 0 \quad \text{where } \mu_{it} = e^{\mathbf{x}_{it}'\boldsymbol{\beta}}.$$

This implies an explicit solution for α_i in terms of $\boldsymbol{\beta}$ in this model,

$$\hat{\alpha}_i = \ln\left(\frac{(1/T_i)\sum_{t=1}^{T_i}y_{it}}{(1/T_i)\sum_{t=1}^{T_i}\hat{\mu}_{it}}\right) = \ln\left(\frac{\bar{y}_i}{\hat{\mu}_i}\right). \tag{18-25}$$

Unlike the regression or the probit model, this estimator does not require that there be within-group variation in y_{it}—all the values can be the same. It does require that at least one observation for individual i be nonzero, however. The rest of the solution for the fixed effects estimator follows the same lines as that for the probit model. An alternative approach, albeit with little practical gain, would be to concentrate the log-likelihood function by inserting this solution for α_i back into the original log likelihood, and then maximizing the resulting function of $\boldsymbol{\beta}$. While logically this makes sense, the approach suggested earlier for the probit model is simpler to implement.

An estimator that is not a function of the fixed effects is found by obtaining the joint distribution of $(y_{i1}, \ldots, y_{iT_i})$ conditional on their sum. For the Poisson model, a close cousin to the multinomial logit model discussed earlier is produced:

$$p\left(y_{i1}, y_{i2}, \ldots, y_{iT_i}\,\middle|\,\sum_{i=1}^{T_i}y_{it}\right) = \frac{\left(\sum_{t=1}^{T_i}y_{it}\right)!}{\left(\prod_{t=1}^{T_i}y_{it}!\right)}\prod_{t=1}^{T_i}p_{it}^{y_{it}}, \tag{18-26}$$

where

$$p_{it} = \frac{e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i}}{\sum_{t=1}^{T_i} e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i}} = \frac{e^{\mathbf{x}_{it}'\boldsymbol{\beta}}}{\sum_{t=1}^{T_i} e^{\mathbf{x}'_{it}\boldsymbol{\beta}}}. \tag{18-27}$$

The contribution of group i to the conditional log likelihood is

$$\ln L_i = \sum_{t=1}^{T_i} y_{it} \ln p_{it}.$$

Note, once again, that the contribution to $\ln L$ of a group in which $y_{it} = 0$ in every period is zero. Cameron and Trivedi (1998) have shown that these two approaches give identical results.

Hausman, Hall, and Griliches (1984) (HHG) report the following conditional density for the fixed effects negative binomial (FENB) model:

$$p\left(y_{i1}, y_{i2}, \ldots, y_{iT_i} \middle| \sum_{t=1}^{T_i} y_{it}\right) = \frac{\Gamma\left(1 + \sum_{t=1}^{T_i} y_{it}\right)\Gamma\left(\sum_{t=1}^{T_i} \lambda_{it}\right)}{\Gamma\left(\sum_{t=1}^{T_i} y_{it} + \sum_{t=1}^{T_i} \lambda_{it}\right)} \prod_{t=1}^{T_i} \frac{\Gamma(y_{it} + \lambda_{it})}{\Gamma(1 + y_{it})\Gamma(\lambda_{it})},$$

which is also free of the fixed effects. This is the default FENB formulation used in popular software packages such as *SAS* and *Stata*. Researchers accustomed to the admonishments that fixed effects models cannot contain overall constants or time-invariant covariates are sometimes surprised to find (perhaps accidentally) that this fixed effects model allows both.[62] The resolution of this apparent contradiction is that the HHG FENB model is not obtained by shifting the conditional mean function by the fixed effect, $\ln \lambda_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i$, as it is in the Poisson model. Rather, the HHG model is obtained by building the fixed effect into the model as an individual-specific θ_i in the Negbin 1 form in (18-22). The conditional mean functions in the models are as follows (we have changed the notation slightly to conform to our earlier formulation):

$$\begin{aligned}
&\text{NB1(HHG):} \quad E[y_{it}|\mathbf{x}_{it}] = \theta_i \phi_{it} = \theta_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}),\\
&\text{NB2:} \qquad\quad E[y_{it}|\mathbf{x}_{it}] = \exp(\alpha_i)\phi_{it} = \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i).
\end{aligned}$$

The conditional variances are

$$\begin{aligned}
&\text{NB1(HHG):} \quad \text{Var}[y_{it}|\mathbf{x}_{it}] = \theta_i \phi_{it}[1 + \theta_i],\\
&\text{NB2:} \qquad\quad \text{Var}[y_{it}|\mathbf{x}_{it}] = \lambda_{it}[1 + \theta\lambda_{it}].
\end{aligned}$$

Letting $\mu_i = \ln \theta_i$, it appears that the HHG formulation does provide a fixed effect in the mean, as now, $E[y_{it}|\mathbf{x}_{it}] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \mu_i)$. Indeed, by this construction, it appears (as the authors suggest) that there are separate effects in both the mean and the variance. They make this explicit by writing $\theta_i = \exp(\mu_i)\gamma_i$ so that in their model,

$$\begin{aligned}
E[y_{it}|\mathbf{x}_{it}] &= \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \mu_i),\\
\text{Var}[y_{it}|\mathbf{x}_{it}] &= \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \mu_i)/[1 + \gamma_i \exp(\mu_i)].
\end{aligned}$$

The contradiction arises because the authors assert that μ_i and γ_i are separate parameters. In fact, they cannot vary separately; only θ_i can vary autonomously. The firm-specific

[62] This issue is explored at length in Allison (2000) and Allison and Waterman (2002).

effect in the HHG model is still isolated in the scaling parameter, which falls out of the conditional density. The mean is homogeneous, which explains why a separate constant, or a time-invariant regressor (or another set of firm-specific effects) can reside there.[63]

18.4.7.c Random Effects

The fixed effects approach has the same flaws and virtues in this setting as in the probit case. It is not necessary to assume that the heterogeneity is uncorrelated with the included exogenous variables. If the uncorrelatedness of the regressors and the heterogeneity can be maintained, then the random effects model is an attractive alternative model. Once again, the approach used in the linear regression model, partial deviations from the group means followed by generalized least squares (see Section 11.5), is not usable here. The approach used is to formulate the joint probability conditioned upon the heterogeneity, then integrate it out of the joint distribution. Thus, we form

$$p(y_{i1}, \ldots, y_{iT_i} | u_i) = \prod_{t=1}^{T_i} p(y_{it} | u_i).$$

Then the random effect is swept out by obtaining

$$\begin{aligned} p(y_{i1}, \ldots, y_{iT_i}) &= \int_{u_i} p(y_{i1}, \ldots, y_{i,T_i}, u_i) du_i \\ &= \int_{u_i} p(y_{i1}, \ldots, y_{iT_i} | u_i) g(u_i) du_i \\ &= E_{u_i}[p(y_{i1}, \ldots, y_{iT_i} | u_i)]. \end{aligned}$$

This is exactly the approach used earlier to condition the heterogeneity out of the Poisson model to produce the negative binomial model. If, as before, we take $p(y_{it} | u_i)$ to be Poisson with mean $\lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)$ in which $\exp(u_i)$ is distributed as gamma with mean 1.0 and variance $1/\alpha$, then the preceding steps produce a negative binomial distribution,

$$p(y_{i1}, \ldots, y_{iT_i}) = \frac{\left[\prod_{t=1}^{T_i} \lambda_{it}^{y_{it}}\right] \Gamma\left(\theta + \sum_{t=}^{T_i} y_{it}\right)}{\left[\Gamma(\theta) \prod_{t=1}^{T_i} y_{it}!\right] \left[\left(\sum_{t=1}^{T_i} \lambda_{it}\right)^{\sum_{t=1}^{T_i} y_{it}}\right]} Q_i^{\theta} (1 - Q_i)^{\sum_{t=1}^{T_i} y_{it}}, \quad \textbf{(18-28)}$$

where

$$Q_i = \frac{\theta}{\theta + \sum_{t=1}^{T_i} \lambda_{it}}.$$

For estimation purposes, we have a negative binomial distribution for $Y_i = \Sigma_t y_{it}$ with mean $\Lambda_i = \Sigma_t \lambda_{it}$.

Like the fixed effects model, introducing random effects into the negative binomial model adds some additional complexity. We do note, because the negative binomial model derives from the Poisson model by adding latent heterogeneity to the conditional mean,

[63] See Greene (2005) and Allison and Waterman (2002) for further discussion.

adding a random effect to the negative binomial model might well amount to introducing the heterogeneity a second time—the random effects NB model is a Poisson regression with $E[y_{it}|\mathbf{x}_{it}, \varepsilon_i, w_{it}] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + w_{it} + \varepsilon_i)$. However, one might prefer to interpret the negative binomial as the density for y_{it} in its own right and treat the common effects in the familiar fashion. Hausman et al.'s (1984) random effects negative binomial (RENB) model is a hierarchical model that is constructed as follows. The heterogeneity is assumed to enter λ_{it} additively with a gamma distribution with mean 1, i.e., $G(\theta_i, \theta_i)$. Then, $\theta_i/(1 + \theta_i)$ is assumed to have a beta distribution with parameters a and b (see Appendix B.4.6). The resulting unconditional density after the heterogeneity is integrated out is

$$p(y_{i1}, y_{i2}, \ldots, y_{iT_i}) = \frac{\Gamma(a + b)\Gamma\left(a + \sum_{t=1}^{T_i}\lambda_{it}\right)\Gamma\left(b + \sum_{t=1}^{T_i}y_{it}\right)}{\Gamma(a)\Gamma(b)\Gamma\left(a + \sum_{t=1}^{T_i}\lambda_{it} + b + \sum_{t=1}^{T_i}y_{it}\right)}.$$

As before, the relationship between the heterogeneity and the conditional mean function is unclear, because the random effect impacts the parameter of the scedastic function. An alternative approach that maintains the essential flavor of the Poisson model (and other random effects models) is to augment the NB2 form with the random effect,

$$\text{Prob}(Y = y_{it}|\mathbf{x}_{it}, \varepsilon_i) = \frac{\Gamma(\theta + y_{it})}{\Gamma(y_{it} + 1)\Gamma(\theta)} r_{it}^{y_{it}}(1 - r_{it})^{\theta},$$
$$\lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_i),$$
$$r_{it} = \lambda_{it}/(\theta + \lambda_{it}).$$

We then estimate the parameters by forming the conditional (on ε_i) log likelihood and integrating ε_i out either by quadrature or simulation. The parameters are simpler to interpret by this construction. Estimates of the two forms of the random effects model are presented in Example 18.19 for a comparison.

There is a preference in the received literature for the fixed effects estimators over the random effects estimators. The virtue of dispensing with the assumption of uncorrelatedness of the regressors and the group-specific effects is substantial. On the other hand, the assumption does come at a cost. To compute the probabilities or the marginal effects, it is necessary to estimate the constants, α_i. The unscaled coefficients in these models are of limited usefulness because of the nonlinearity of the conditional mean functions.

Other approaches to the random effects model have been proposed. Greene (1994, 1995a, 1995b, 1997), Riphahn et al. (2003), and Terza (1995) specify a normally distributed heterogeneity, on the assumption that this is a more natural distribution for the aggregate of small independent effects. Brannas and Johanssen (1994) have suggested a semiparametric approach based on the GMM estimator by superimposing a very general form of heterogeneity on the Poisson model. They assume that conditioned on a random effect ε_{it}, y_{it} is distributed as Poisson with mean $\varepsilon_{it}\lambda_{it}$. The covariance structure of ε_{it} is allowed to be fully general. For $t, s = 1, \ldots, T$, $\text{Var}[\varepsilon_{it}] = \sigma_i^2$, $\text{Cov}[\varepsilon_{it}, \varepsilon_{js}] = \gamma_{ij}(|t - s|)$. For a long time series, this model is likely to have far too many parameters to be identified without some restrictions, such as first-order homogeneity ($\boldsymbol{\beta}_i = \boldsymbol{\beta}\ \forall\, i$),

uncorrelatedness across groups, $[\gamma_{ij}(.) = 0 \text{ for } i \neq j]$, groupwise homoscedasticity $(\sigma_i^2 = \sigma^2 \,\forall\, i)$, and nonautocorrelatedness $[\gamma(r) = 0 \,\forall\, r \neq 0]$. With these assumptions, the estimation procedure they propose is similar to the procedures suggested earlier. If the model imposes enough restrictions, then the parameters can be estimated by the method of moments. The authors discuss estimation of the model in its full generality. Finally, the latent class model discussed in Section 14.15.4 and the random parameters model in Section 15.9 extend naturally to the Poisson model. Indeed, most of the received applications of the latent class structure have been in the Poisson or negative binomial regression framework.[64]

Example 18.19 Panel Data Models for Doctor Visits

The German health care panel data set contains 7,293 individuals with group sizes ranging from 1 to 7. Table 18.27 presents the fixed and random effects estimates of the equation. The pooled estimates are also shown for comparison. Overall, the panel data treatments bring large changes in the estimates compared to the pooled estimates. There is also a

TABLE 18.27 Estimated Panel Data Models for Doctor Visits (standard errors in parentheses)

	Poisson				***Negative Binomial***			
					Fixed Effects		***Random Effects***	
Variable	***Pooled Robust Std. Error***	***Fixed Effects***	***Random Effects***	***Pooled NB2***	***FE NB1***	***FE NB2***	***HHG Gamma***	***Normal***
Constant	1.05266	—	0.69553	1.10083	−1.14543	—	−0.41087	0.37764
	(0.11395)		(0.05266)	(0.05970)	(0.09392)		(0.06062)	(0.05499)
Age	0.01838	0.03127	0.02331	0.01789	0.02383	0.04476	0.01886	0.02230
	(0.00134)	(0.00144)	(0.00045)	(0.00079)	(0.00119)	(0.00277)	(0.00078)	0.00070)
Educ	−0.04355	−0.03934	−0.03938	−0.04797	0.01338	−0.04788	−0.02469	−0.04536
	(0.00699)	(0.01734)	(0.00434)	(0.00378)	(0.00630)	(0.02963)	(0.00386)	(0.00345)
Income	−0.52502	−0.30674	−0.27282	−0.46285	0.01635	−0.20085	−0.10785	−0.18650
	(.08240)	(0.04103)	(0.01519)	(0.04600)	(0.05541)	(0.07321)	(0.04577)	(0.04267)
Kids	−0.16109	0.00153	−0.03974	−0.15656	−0.03336	−0.00131	−0.11181	−0.12013
	(0.03118)	(0.01534)	(0.00526)	(0.01735)	(0.02117)	(0.02921)	(0.01677)	(0.01583)
AddOn	0.07282	−0.07946	−0.05654	0.07134	0.11224	−0.02158	0.15086	0.05637
	(0.07801)	(0.03568)	(0.01605)	(0.07205)	(0.06622)	(0.06739)	(0.05836)	(0.05699)
α	—	—	1.16959	1.92971	—	1.91953	—	1.08433
			(0.01949)	(0.02009)		(0.02993)		(0.01210)
a	—	—	—	—	—	—	2.13948	—
							(0.05928)	
b	—	—	—	—	—	—	3.78252	—
							(0.11377)	
σ	—	—	—	—	—	—	—	0.96860
								(0.00828)
ln L	−104,603.0	−60,327.8	−71,779.6	−60,291.5	34,015.4	−49,478.0	−58,189.5	−58,170.5

[64]See Greene (2001) for a survey.

considerable amount of variation across the specifications. With respect to the parameter of interest, *AddOn,* we find that the size of the coefficient falls substantially with all panel data treatments and it becomes negative in the Poisson models. Whether using the pooled, fixed, or random effects specifications, the test statistics (Wald, LR) all reject the Poisson model in favor of the negative binomial. Similarly, either common effects specification is preferred to the pooled estimator. There is no simple basis for choosing between the fixed and random effects models, and we have further blurred the distinction by suggesting two formulations of each of them. We do note that the two random effects estimators are producing similar results, which one might hope for. But the two fixed effects estimators are producing very different estimates. The NB1 estimates include two coefficients, *Income* and *Education,* which are positive, but negative in every other case. Moreover, the coefficient on *AddOn*, varies in sign, and is insignificant in nearly all cases. As before, the data do not suggest the presence of moral hazard, at least as measured here.

We also fit a three-class latent class model for these data. (See Section 14.10.) The three class probabilities were modeled as functions of *Married* and *Female,* which appear from the results to be significant determinants of the class sorting. The average prior probabilities for the three classes are 0.09027, 0.49332, and 0.41651. The coefficients on *AddOn* in the three classes, with associated *t* ratios, are −0.02191 (0.45), 0.36825 (5.60), and 0.01117 (0.26). The qualitative result concerning evidence of moral hazard suggested here is that there might be a segment of the population for which we have some evidence, but more generally, we find relatively little.

18.4.8 TWO-PART MODELS: ZERO-INFLATION AND HURDLE MODELS

Mullahy (1986), Heilbron (1989), Lambert (1992), Johnson and Kotz (1993), and Greene (1994) have analyzed an extension of the hurdle model in which the zero outcome can arise from one of two regimes.[65] In one regime, the outcome is always zero. In the other, the usual Poisson process is at work, which can produce the zero outcome or some other. In Lambert's application, she analyzes the number of defective items produced by a manufacturing process in a given time interval. If the process is under control, then the outcome is always zero (by definition). If it is not under control, then the number of defective items is distributed as Poisson and may be zero or positive in any period. The model at work is therefore

$$\text{Prob}(y_i = 0|\mathbf{x}_i) = \text{Prob(regime 1)} + \text{Prob}(y_i = 0|\mathbf{x}_i, \text{regime 2})\,\text{Prob(regime 2)},$$
$$\text{Prob}(y_i = j|\mathbf{x}_i) = \text{Prob}(y_i = j|\mathbf{x}_i, \text{regime 2})\,\text{Prob(regime 2)}, j = 1, 2, \ldots.$$

Let z denote a binary indicator of regime 1 ($z = 0$) or regime 2 ($z = 1$), and let y^* denote the outcome of the Poisson process in regime 2. Then the observed y is $z \times y^*$. A natural extension of the splitting model is to allow z to be determined by a set of covariates. These covariates need not be the same as those that determine the conditional probabilities in the Poisson process. Thus, the model is:

$$\text{Prob}(z_i = 0|\mathbf{w}_i) = F(\mathbf{w}_i, \boldsymbol{\gamma}), \text{(Regime 1: } y \text{ will equal zero)};$$
$$\text{Prob}(y_i = j|\mathbf{x}_i, z_i = 1) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}, \text{(Regime 2: } y \text{ will be a count outcome)}.$$

The zero-inflation model can also be viewed as a type of latent class model. The two class probabilities are $F(\mathbf{w}_i, \boldsymbol{\gamma})$ and $1 - F(\mathbf{w}_i, \boldsymbol{\gamma})$, and the two regimes are $y = 0$ and the

[65]The model is variously labeled the "with zeros," or WZ, model [Mullahy (1986)], the zero-inflated Poisson, or ZIP, model [Lambert (1992)], and "zero-altered Poisson," or ZAP, model [Greene (1994)].

Poisson or negative binomial data-generating process.[66] The extension of the ZIP formulation to the negative binomial model is widely labeled the ZINB model.[67] [See Zaninotti and Falischetti (2010) for an application.]

The mean of this random variable in the Poisson case is

$$E[y_i|\mathbf{x}_i, \mathbf{w}_i] = F_i \times 0 + (1 - F_i) \times E[y_i^*|\mathbf{x}_i, z_i = 1] = (1 - F_i)\lambda_i.$$

Lambert (1992) and Greene (1994) consider a number of alternative formulations, including logit and probit models discussed in Sections 17.2 and 17.3, for the probability of the two regimes. It might be of interest to test simply whether there is a regime splitting mechanism at work or not. Unfortunately, the basic model and the zero-inflated model are not nested. Setting the parameters of the splitting model to zero, for example, does not produce Prob$[z = 0] = 0$. In the probit case, this probability becomes 0.5, which maintains the regime split. The preceding tests for over- or underdispersion would be rather indirect. What is desired is a test of non-Poissonness. An alternative distribution may (but need not) produce a systematically different proportion of zeros than the Poisson. Testing for a different distribution, as opposed to a different set of parameters, is a difficult procedure. Because the hypotheses are necessarily nonnested, the power of any test is a function of the alternative hypothesis and may, under some, be small. Vuong (1989) has proposed a test statistic for nonnested models that is well suited for this setting when the alternative distribution can be specified. (See Section 14.6.6.) Let $f_j(y_i|\mathbf{x}_i)$ denote the predicted probability that the random variable Y equals y_i under the assumption that the distribution is $f_j(y_i|\mathbf{x}_i)$, for $j = 1, 2$, and let

$$m_i = \ln\left(\frac{f_1(y_i|\mathbf{x}_i)}{f_2(y_i|\mathbf{x}_i)}\right).$$

Then Vuong's statistic for testing the nonnested hypothesis of model 1 versus model 2 is

$$v = \frac{\sqrt{n}[\frac{1}{n}\Sigma_{i=1}^n m_i]}{\sqrt{\frac{1}{n}\Sigma_{i=1}^n (m_i - \overline{m})^2}} = \frac{\sqrt{n}\overline{m}}{s_m}.$$

This is the standard statistic for testing the hypothesis that $E[m_i]$ equals zero. Vuong shows that v has a limiting standard normal distribution. As he notes, the statistic is bidirectional. If $|v|$ is less than 2, then the test does not favor one model or the other. Otherwise, large values favor model 1 whereas small (negative) values favor model 2. Carrying out the test requires estimation of both models and computation of both sets of predicted probabilities. In Greene (1994), it is shown that the Vuong test has some power to discern the zero-inflation phenomenon. The logic of the testing procedure is to allow for overdispersion by specifying a negative binomial count data process and then examine whether, even allowing for the overdispersion, there still appear to be excess zeros. In his application, that appears to be the case.

Example 18.20 Zero-Inflation Models for Major Derogatory Reports

In Example 18.17, we examined the counts of major derogatory reports for a sample of 13,444 credit card applicants. It was noted that there are over 10,800 zeros in the counts. One might guess that among credit card users, there is a certain (probably large) proportion

[66] Harris and Zhao (2007) applied this approach to a survey of teenage smokers and nonsmokers in Australia, using an ordered probit model. (See Section 18.3.)

[67] Greene (2005) presents a survey of two-part models, including the zero-inflation models.

TABLE 18.28 Estimated Zero Inflated Count Models

	Poisson			*Negative Binomial*		
		Zero Inflation			*Zero Inflation*	
	Poisson Regression	*Regression*	*Zero Regime*	*Negative Binomial*	*Regression*	*Zero Regime*
Constant	−1.33276	0.75483	2.06919	−1.54536	−0.39628	4.18910
Age	0.01286	0.00358	−0.01741	0.01807	−0.00280	−0.14339
Income	−0.02577	−0.05127	−0.03023	−0.02482	−0.05502	−0.33903
OwnRent	−0.17801	−0.15593	−0.01738	−0.18985	−0.28591	−0.50026
Self Employment	0.04691	−0.01257		0.07920	0.06817	
Dependents	0.13760	0.06038	−0.09098	0.14054	0.08599	−0.32897
Cur. Add.	0.00195	0.00046		0.00245	0.00257	
α				6.41435	4.85653	
ln L	−15,467.71	−11,569.74		−10,582.88	−10,516.46	
Vuong		20.6981			4.5943	

of individuals who would never generate an MDR, and some other proportion who might or might not, depending on circumstances. We propose to extend the count models in Example 18.17 to accommodate the zeros. The extensions to the ZIP and ZINB models are shown in Table 18.28. Only the coefficients are shown for purpose of the comparisons. Vuong's diagnostic statistic appears to confirm intuition that the Poisson model does not adequately describe the data; the value is 20.6981. Using the model parameters to compute a prediction of the number of zeros, it is clear that the splitting model does perform better than the basic Poisson regression. For the simple Poisson model, the average probability of zero times the sample size gives a prediction of 8,609. For the ZIP model, the value is 10,914.8, which is a dramatic improvement. By the likelihood ratio test, the negative binomial is clearly preferred; comparing the two zero-inflation models, the difference in the log likelihood functions is over 1,000. As might be expected, the Vuong statistic falls considerably, to 4.5943. However, the simple model with no zero inflation is still rejected by the test.

In some settings, the zero outcome of the data generating process is qualitatively different from the positive ones. The zero or nonzero value of the outcome is the result of a separate decision whether or not to participate in the activity. On deciding to participate, the individual decides separately how much, that is, how intensively. Mullahy (1986) argues that this fact constitutes a shortcoming of the Poisson (or negative binomial) model and suggests a **hurdle model** as an alternative.[68] In his formulation, a binary probability model determines whether a zero or a nonzero outcome occurs and then, in the latter case, a (truncated) Poisson distribution describes the positive outcomes. The model is

$$\text{Prob}(y_i = 0|\mathbf{x}_i) = e^{-\theta},$$

$$\text{Prob}(y_i = j|\mathbf{x}_i) = (1 - e^{-\theta})\frac{\exp(-\lambda_i)\lambda_i^j}{j![1 - \exp(-\lambda_i)]}, \quad j = 1, 2, \ldots.$$

This formulation changes the probability of the zero outcome and scales the remaining probabilities so that they sum to one. Mullahy suggests some formulations and applies

[68]For a similar treatment in a continuous data application, see Cragg (1971).

the model to a sample of observations on daily beverage consumption. Mullahy's formulation adds a new restriction that $\text{Prob}(y_i = 0|\mathbf{x}_i)$ no longer depends on the covariates, however. The natural next step is to parameterize this probability. This extension of the hurdle model would combine a binary choice model like those in Section 17.2 and 17.3 with a truncated count model as shown in Section 18.4.6. This would produce, for example, for a logit participation equation and a Poisson intensity equation,

$$\text{Prob}(y_i = 0|\mathbf{w}_i) = \Lambda(\mathbf{w}_i'\boldsymbol{\gamma})$$
$$\text{Prob}(y_i = j|\mathbf{x}_i, \mathbf{w}_i, y_i > 0) = \frac{[1 - \Lambda(\mathbf{w}_i'\boldsymbol{\gamma})]\exp(-\lambda_i)\lambda_i^j}{j![1 - \exp(-\lambda_i)]}.$$

The conditional mean function in the hurdle model is

$$E[y_i|\mathbf{x}_i, \mathbf{w}_i] = \frac{[1 - F(\mathbf{w}_i'\boldsymbol{\gamma})]\lambda_i}{[1 - \exp(-\lambda_i)]}, \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}),$$

where $F(.)$ is the probability model used for the participation equation (probit or logit). The partial effects are obtained by differentiating with respect to the two sets of variables separately,

$$\frac{\partial E[y_i|\mathbf{x}_i, \mathbf{w}_i]}{\partial \mathbf{x}_i} = [1 - F(\mathbf{w}_i'\boldsymbol{\gamma})]\boldsymbol{\delta}_i,$$
$$\frac{\partial E[y_i|\mathbf{x}_i, \mathbf{w}_i]}{\partial \mathbf{w}_i} = \left\{\frac{-f(\mathbf{w}_i'\boldsymbol{\gamma})\lambda_i}{[1 - \exp(-\lambda_i)]}\right\}\boldsymbol{\gamma},$$

where $\boldsymbol{\delta}_i$ is defined in (18-23) and $f(.)$ is the density corresponding to $F(.)$. For variables that appear in both $\mathbf{x}_i$ and $\mathbf{w}_i$, the effects are added. For dummy variables, the preceding would be an approximation; the appropriate result would be obtained by taking the difference of the conditional means with the variable fixed at one and zero.

It might be of interest to test for hurdle effects. The hurdle model is similar to the zero-inflation model in that a model without hurdle effects is not nested within the hurdle model; setting $\boldsymbol{\gamma} = \mathbf{0}$ produces either $F = \alpha$, a constant, or $F = {}^1/_2$ if the constant term is also set to zero. Neither serves the purpose. Nor does forcing $\boldsymbol{\gamma} = \boldsymbol{\beta}$ in a model with $\mathbf{w}_i = \mathbf{x}_i$ and $F = \Lambda$ with a Poisson intensity equation, which might be intuitively appealing. A complementary log log model with

$$\text{Prob}(y_i = 0|\mathbf{w}_i) = \exp[-\exp(\mathbf{w}_i'\boldsymbol{\gamma})]$$

does produce the desired result if $\mathbf{w}_i = \mathbf{x}_i$. In this case, "hurdle effects" are absent if $\boldsymbol{\gamma} = \boldsymbol{\beta}$. The strategy in this case, then, would be a test of this restriction. But, this formulation is otherwise restrictive, first in the choice of variables and second in its unconventional functional form. The more general approach to this test would be the Vuong test used earlier to test the zero-inflation model against the simpler Poisson or negative binomial model.

The hurdle model bears some similarity to the zero-inflation model. However, the behavioral implications are different. The zero-inflation model can usefully be viewed as a latent class model. The splitting probability defines a regime determination. In the hurdle model, the splitting equation represents a behavioral outcome on the same level

as the intensity (count) equation.[69] Both of these modifications substantially alter the Poisson formulation. First, note that the equality of the mean and variance of the distribution no longer follow; both modifications induce overdispersion. On the other hand, the overdispersion does not arise from heterogeneity; it arises from the nature of the process generating the zeros. As such, an interesting identification problem arises in this model. If the data do appear to be characterized by overdispersion, then it seems less than obvious whether it should be attributed to heterogeneity or to the regime splitting mechanism. Mullahy (1986) argues the point more strongly. He demonstrates that overdispersion will always induce excess zeros. As such, in a splitting model, we may misinterpret the excess zeros as due to the splitting process instead of the heterogeneity.

Example 18.21 Hurdle Models for Doctor Visits

Jones and Schurer (2009) used the hurdle framework to study physician visits in several countries using the ECHP panel data set. The base model was a negative binomial regression, with a logit hurdle equation. The main interest was the cross-country variation in the income elasticity of health care utilization. A few of their results for general practitioners are shown in Table 18.29, which is extracted from their Table 8.[70] (Corresponding results are computed for specialists.) Note that individuals are classified as high or low users. The latent classes have been identified as a group of heavy users of the system and light users, which would seem to suggest that the classes are not latent. The class assignments are done using the method described in Section 14.15.4. The posterior (conditional) class probabilities, $\hat{\pi}_{i1}$ and $\hat{\pi}_{i2}$, are computed for each person in the sample. An individual is classified as coming from class 1 if $\hat{\pi}_{i1} \geq 0.5$ and class 2 if $\hat{\pi}_{i1} < 0.5$. With this classification, the average within group utilization is computed. The group with the higher group mean is labeled the "High users."

In Examples 18.16 and 18.21, we fit Poisson regressions with means

$$E[DocVis|\mathbf{x}] = \exp(\beta_1 + \beta_2 Age + \beta_3 Education + \beta_4 Income + \beta_5 Kids + \beta_6 AddOn).$$

TABLE 18.29 Income Elasticities

Estimated Income Coefficients and Elasticities for GP and Specialist Visits—Country-Specific LC Hurdle Models (Asymptotic t ratios in parentheses)

		GPs			
		Low Users		***High Users***	
Country		***Estimated Coefficient***	***Estimated Elasticity***	***Estimated Coefficient***	***Estimated Elasticity***
Austria	$P(Y > 0)$	−0.051 (−1.467)	−0.012	−0.109 (−0.872)	−0.005
	$E(Y\|Y > 0)$	0.012(0.693)	0.009	0.039(2.167)	0.035
Belgium	$P(Y > 0)$	0.035(1.002)	0.008	0.292(4.004)	0.010
	$E(Y\|Y > 0)$	−0.052(−3.125)	−0.037	−0.055(−4.030)	−0.050
Denmark	$P(Y > 0)$	0.083(1.746)	0.033	0.261 (2.302)	0.023
	$E(Y\|Y > 0)$	0.042 (0.992)	0.021	−0.030 (−1.009)	−0.024
Finland	$P(Y > 0)$	0.054(1.358)	0.024	−0.030 (−0.263)	−0.003
	$E(Y\|Y > 0)$	0.007(0.237)	0.004	−0.048 (−1.706)	−0.037

[69]See, for example, Jones (1989), who applied the model to cigarette consumption.

[70]From Jones and Schurer (2009).

Table 18.30 reports results for a two-class latent class model based on this specification using the 3,377 observations in the 1994 wave of the panel. The estimated prior class probabilities are 0.23298 and 0.76702. For each observation in the sample, the posterior probabilities are computed using

$$\hat{\pi}_{i1} = \frac{\hat{\pi}_1 \hat{L}_{i1}}{\hat{\pi}_1 \hat{L}_{i1} + \hat{\pi}_2 \hat{L}_{i2}}, \hat{L}_{ic} = \frac{\exp(-\hat{\lambda}_{ic})(\hat{\lambda}_{ic})^{DocVis_i}}{DocVis_i!}, \hat{\lambda}_{ic} = \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_c),\ c = 1, 2,$$

then $\hat{\pi}_{i2} = 1 - \hat{\pi}_{i1}$. The mean values of these posterior probabilities are 0.228309 and 0.771691, which, save for some minor error, match the prior probabilities. (In theory, they match perfectly.) We then define the class assignment to be class 1 if $\hat{\pi}_{i1} \geq 0.5$ and class 2 if $\hat{\pi}_{i1} < 0.5$. By this calculation, there are 771 and 2,606 observations in the two classes, respectively. The sample averages of *DocVis* for the two groups are 11.380 and 1.535, which confirms the idea of a group of high users and low users. Figure 18.8 displays histograms for the two groups. (The sample has been trimmed by dropping a handful of observations larger than 30 in group 1.)

18.4.9 ENDOGENOUS VARIABLES AND ENDOGENOUS PARTICIPATION

As in other situations, one would expect to find endogenous variables in models for counts. For example, in the study on which we have relied for our examples of health care utilization, Riphahn, Wambach, and Million (RWM, 2003), were interested in the role of the *AddOn* insurance in the usage variable. One might expect the choice to buy insurance to be at least partly influenced by some of the same factors that motivate usage of the health care system. Insurance purchase might well be endogenous in a model such as the hurdle model in Example 18.21.

The Poisson model presents a complication for modeling endogeneity that arises in some other cases as well. For simplicity, consider a continuous variable, such as *Income*, to continue our ongoing example. A model of income determination and doctor visits might appear

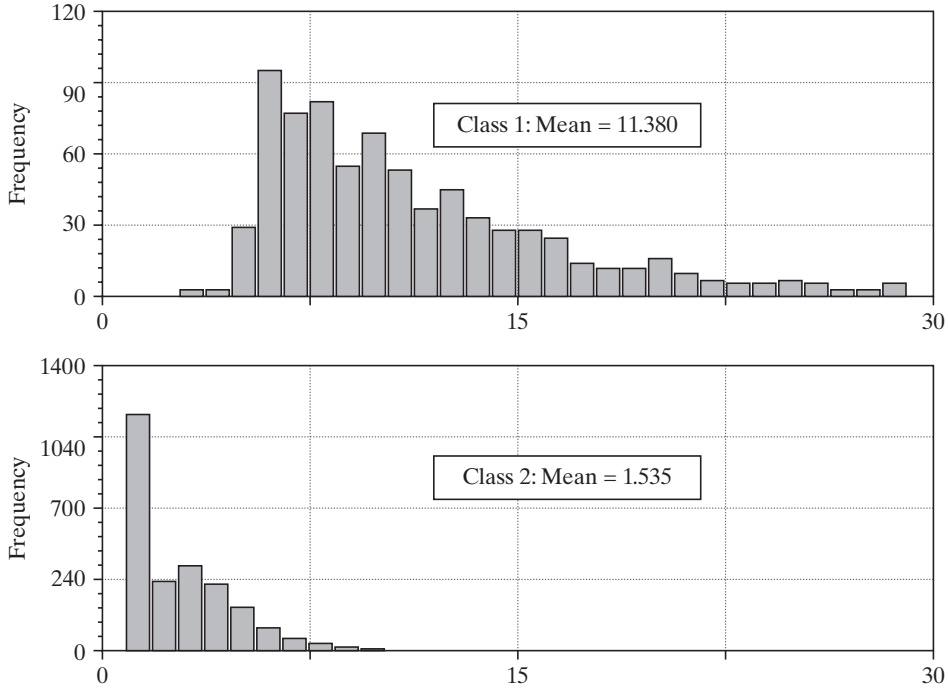
$$Income = \mathbf{z}_i'\boldsymbol{\gamma} + u_i,$$
$$\text{Prob}(DocVis_i = j \,|\, \mathbf{x}_i, Income_i) = \exp(-\lambda_i), \lambda_i^j/j!, \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \delta\, Income_i).$$

Endogeneity as we have analyzed it, for example, in Chapter 8 and Sections 17.3.5 and 17.5.5, arises through correlation between the endogenous variable and the unobserved

TABLE 18.30 Estimated Latent Class Model for Doctor Visits

	Latent Class Model				*Poisson Regression*	
	Class 1		*Class 2*			
Variable	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Constant	2.67381	0.11876	0.66690	0.17591	1.23358	0.06706
Age	0.01394	0.00149	0.01867	0.00213	0.01866	0.00082
Income	−0.39859	0.08096	−0.51861	0.12012	−0.40231	0.04632
Education	−0.05760	0.00699	−0.06516	0.01140	−0.04457	0.00435
Kids	−0.13259	0.03539	−0.32098	0.05270	−0.14477	0.02065
AddOn	0.00786	0.08795	0.06883	0.15084	0.12270	0.06129
Class Prob.	0.23298	0.00959	0.76702	0.00959	1.00000	0.00000
ln *L*	−9263.76				−13653.41	

FIGURE 18.8 Distributions of Doctor Visits by Class.



omitted factors in the main equation. But the Poisson model does not contain any unobservables. This is a major shortcoming of the specification as a regression model; all of the regression variation of the dependent variable arises through variation of the observables. There is no accommodation for unobserved heterogeneity or omitted factors. This is the compelling motivation for the negative binomial model or, in RWM's case, the Poisson-normal mixture model.[71] If the model is reformulated to accommodate heterogeneity, as in

$$\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \delta\, Income_i + \varepsilon_i),$$

then $Income_i$ will be endogenous if u_i and ε_i are correlated.

A bivariate normal model for (u_i, ε_i) with zero means, variances σ_u^2 and σ_ε^2, and correlation ρ provides a convenient (and the usual) platform to operationalize this idea. By projecting ε_i on u_i, we have

$$\varepsilon_i = (\rho\sigma_\varepsilon/\sigma_u)u_i + v_i,$$

where v_i is normally distributed with mean zero and variance $\sigma_\varepsilon^2(1 - \rho^2)$. It will prove convenient to parameterize these based on the regression and the specific parameters as follows:

$$\begin{aligned}\varepsilon_i &= \rho\sigma_\varepsilon(Income_i - \mathbf{z}_i'\boldsymbol{\gamma})/\sigma_u + v_i,\\ &= \tau[(Income_i - \mathbf{z}_i'\boldsymbol{\gamma})/\sigma_u] + \theta w_i,\end{aligned}$$

[71] See Terza (2009, pp. 555–556) for discussion of this issue.

where w_i will be normally distributed with mean zero and variance one while $\tau = \rho\sigma_\varepsilon$ and $\theta^2 = \sigma_\varepsilon^2(1 - \rho^2)$. Then, combining terms,

$$\varepsilon_i = \tau u_i^* + \theta w_i.$$

With this parameterization, the conditional mean function in the Poisson regression model is

$$\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \delta\, Income_i + \tau u_i^* + \theta w_i).$$

The parameters to be estimated are $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, δ, σ_ε, σ_u, and ρ. There are two ways to proceed. A two-step method can be based on the fact that $\boldsymbol{\gamma}$ and σ_u can consistently be estimated by linear regression of *Income* on $\mathbf{z}$. After this first step, we can compute values of u_i^* and formulate the Poisson regression model in terms of

$$\hat{\lambda}_i(w_i) = \exp[\mathbf{x}_i'\boldsymbol{\beta} + \delta\, Income_i + \tau\hat{u}_i + \theta w_i].$$

The log likelihood to be maximized at the second step is

$$\ln L(\boldsymbol{\beta}, \delta, \tau, \theta \,|\, \mathbf{w}) = \sum_{i=1}^{n} -\hat{\lambda}_i(w_i) + y_i \ln \hat{\lambda}_i(w_i) - \ln y_i!.$$

A remaining complication is that the unobserved heterogeneity, w_i, remains in the equation so it must be integrated out of the log-likelihood function. The unconditional log-likelihood function is obtained by integrating the standard normally distributed w_i out of the conditional densities,

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau, \theta) = \sum_{i=1}^{n} \ln\left\{ \int_{-\infty}^{\infty} \left[\frac{\exp(-\hat{\lambda}_i(w_i))(\hat{\lambda}_i(w_i))^{y_i}}{y_i!} \right] \phi(w_i) dw_i \right\}.$$

The method of Butler and Moffitt or maximum simulated likelihood that we used to fit a probit model in Section 17.4.2 can be used to estimate $\boldsymbol{\beta}$, δ, τ, and θ. Estimates of ρ and σ_ε can be deduced from the last two of these; $\sigma_\varepsilon^2 = \theta^2 + \tau^2$ and $\rho = \tau/\sigma_\varepsilon$. This is the control function method discussed in Section 17.6.2 and is also the "residual inclusion" method discussed by Terza, Basu, and Rathouz (2008).

The full set of parameters can be estimated in a single step using full information maximum likelihood. To estimate all parameters simultaneously and efficiently, we would form the log likelihood from the joint density of *DocVis* and *Income* as $P(DocVis \,|\, Income) f(Income)$. Thus,

$$f(DocVis, Income) = \frac{\exp[-\lambda_i(w_i)][\lambda_i(w_i)]^{y_i}}{y_i!} \frac{1}{\sigma_u} \phi\left(\frac{Income - \mathbf{z}_i'\boldsymbol{\gamma}}{\sigma_u} \right),$$
$$\lambda_i(w_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \delta\, Income_i + \tau(Income_i - \mathbf{z}_i'\boldsymbol{\gamma})/\sigma_u + \theta w_i).$$

As before, the unobserved w_i must be integrated out of the log-likelihood function. Either quadrature or simulation can be used. The parameters to be estimated by maximizing the full log likelihood are $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \sigma_u, \sigma_\varepsilon, \rho)$. The invariance principle can be used to simplify the estimation a bit by parameterizing the log-likelihood function in terms of τ and θ. Some additional simplification can also be obtained by using the Olsen (1978) [and Tobin (1958)] transformations, $\eta = 1/\sigma_u$ and $\boldsymbol{\alpha} = (1/\sigma_u)\boldsymbol{\gamma}$.

An endogenous binary variable, such as *Public* or *AddOn* in our *DocVis* example is handled similarly but is a bit simpler. The structural equations of the model are

$$
\begin{aligned}
T^* &= \mathbf{z}'\boldsymbol{\gamma} + u, \qquad && u \sim N[0,1],\\
T &= \mathbf{1}(T^* > 0),\\
\lambda &= \exp(\mathbf{x}'\boldsymbol{\beta} + \delta T + \varepsilon) \qquad && \varepsilon \sim N[0, \sigma_\varepsilon^2],
\end{aligned}
$$

with $\text{Cov}(u, \varepsilon) = \rho\sigma_\varepsilon$. The endogeneity of T is implied by a nonzero ρ. We use the bivariate normal result,

$$u = (\rho/\sigma_\varepsilon)\varepsilon + v,$$

where v is normally distributed with mean zero and variance $1 - \rho^2$. Then, using our earlier results for the probit model (Section 17.3),

$$P(T|\varepsilon) = \Phi\left[(2T-1)\left(\frac{\mathbf{z}'\boldsymbol{\gamma} + (\rho/\sigma_\varepsilon)\varepsilon}{\sqrt{1-\rho^2}}\right)\right], \quad T = 0, 1.$$

It will be convenient once again to write $\varepsilon = \sigma_\varepsilon w$ where $w \sim \text{N}[0, 1]$. Making the substitution, we have

$$P(T|w) = \Phi\left[(2T-1)\left(\frac{\mathbf{z}'\boldsymbol{\gamma} + \rho w}{\sqrt{1-\rho^2}}\right)\right], \quad T = 0, 1.$$

The probability density function for $y|T,w$ is Poisson with $\lambda(w) = \exp(\mathbf{x}'\boldsymbol{\beta} + \delta T + \sigma_\varepsilon w)$. Combining terms,

$$P(y, T|w) = \frac{\exp[-\lambda(w)][\lambda(w)]^y}{y!}\Phi\left[(2T-1)\left(\frac{\mathbf{z}'\boldsymbol{\gamma} + \rho w}{\sqrt{1-\rho^2}}\right)\right].$$

This last result provides the terms that enter the log likelihood for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \rho, \sigma_\varepsilon)$. As before, the unobserved heterogeneity, w, must be integrated out of the log likelihood, so either the quadrature or simulation method discussed in Chapter 17 is used to obtain the parameter estimates. Note that this model may also be estimated in two steps, with $\boldsymbol{\gamma}$ obtained in the first-step probit. The two-step method will not be appreciably simpler, since the second term in the density must remain to identify ρ. The residual inclusion method is not feasible here since T^* is not observed.

This same set of methods is used to allow for endogeneity of the participation equation in the hurdle model in Section 18.4.8. Mechanically, the hurdle model with endogenous participation is essentially the same as the endogenous binary variable.[72]

Example 18.22 Endogenous Treatment in Health Care Utilization

Table 18.31 reports estimates of the treatment effects model for our health care utilization data. The main result is the causal parameter on *Addon,* which is shown in the boxes in the table. We have fit the model with the full panel (pooled) and with the final (1994) wave of the panel. The results are nearly identical. The large negative value is, of course, inconsistent with any suggestion of moral hazard, and seems extreme enough to cast some suspicion on the model specification. We, like Riphahn et al. (2003) and others they discuss, did not find evidence of moral hazard in the demand for physician visits. (The authors did find more suggestive results for hospital visits.)

[72]See Greene (2005, 2007d).

TABLE 18.31 Estimated Treatment Effects Model (Standard errors in parentheses)

	Full Panel		*1994 Wave*	
	Treatment	***Outcome***	***Treatment***	***Outcome***
Variable	***(Probit: Addon)***	***(Poisson: DocVis)***	***(Probit: Addon)***	***(Poisson: DocVis)***
Health Sat.	0.10824		0.13202	
	(0.00677)		(0.00903)	
Married	0.12325		0.14827	
	(0.03564)		(0.07314)	
Income	0.61812		0.31412	
	(0.05873)		(0.14664)	
Working	−0.05864		0.19407	
	(0.03297)		(0.12375)	
Education	0.05233		0.04755	
	(0.00588)		(0.01020)	
Kids	−0.10872	−0.17063	−0.00065	−0.23349
	(0.03306)	(0.01879)	(0.07519)	(04933)
Constant	−3.56368	−0.74006	−3.70407	−0.20658
	(0.08364)	(0.04094)	(0.16509)	(0.10440)
Age		0.02099		0.01431
		(0.00079)		(0.00214)
Female		0.42599		0.50918
		(0.01619)		(0.04400)
AddOn		−2.73847		−2.86428
		(0.04978)		(0.09289)
Sigma		1.43070		1.42112
		(0.00653)		(0.01866)
Rho		0.93299		0.99644
		(0.00754)		(0.00376)
ln *L*	−62366.61		−8313.88	
N	27,326,		3,377	

18.5 SUMMARY AND CONCLUSIONS

The analysis of individual decisions in microeconometrics is largely about discrete decisions such as whether to participate in an activity or not, whether to make a purchase or not, or what brand of product to buy. This chapter and Chapter 17 have developed the four essential models used in that type of analysis. Random utility, the binary choice model, and regression-style modeling of probabilities developed in Chapter 17 are the three fundamental building blocks of discrete choice modeling. This chapter extended those tools into the three primary areas of choice modeling: unordered choice models, ordered choice models, and models for counts. In each case, we developed a core modeling framework that provides the broad platform and then developed a variety of extensions.

In the analysis of unordered choice models, such as brand or location, the multinomial logit (MNL) model has provided the essential starting point. The MNL works well to provide a basic framework, but as a behavioral model in its own right, it has some important shortcomings. Much of the recent research in this area has focused on relaxing these behavioral assumptions. The most recent research in this area, on the mixed logit model, has produced broadly flexible functional forms that can match behavioral modeling to empirical specification and estimation.

The ordered choice model is a natural extension of the binary choice setting and also a convenient bridge between models of choice between two alternatives and more complex models of choice among multiple alternatives. We began this analysis with the ordered probit and logit model pioneered by Zavoina and McKelvey (1975). Recent developments of this model have produced the same sorts of extensions to panel data and modeling heterogeneity that we considered in Chapter 17 for binary choice. We also examined some multiple-equation specifications. For all its versatility, the familiar ordered choice models have an important shortcoming in the assumed constancy underlying preference behind the rating scale. The current work on differential item functioning, such as King et al. (2004), has produced significant progress on filling this gap in the theory.

Finally, we examined probability models for counts of events. Here, the Poisson regression model provides the broad framework for the analysis. The Poisson model has two shortcomings that have motivated the current stream of research. First, the functional form binds the mean of the random variable to its variance, producing an unrealistic regression specification. Second, the basic model has no component that accommodates unmeasured heterogeneity. (This second feature is what produces the first.) Current research has produced a rich variety of models for counts, such as two-part behavioral models that account for many different aspects of the decision-making process and the mechanisms that generate the observed data.

Key Terms and Concepts

- Attribute nonattendance
- Bivariate ordered probit
- Censoring
- Characteristics
- Choice-based sample
- Conditional logit model
- Count data
- Deviance
- Differential item functioning (DIF)
- Exposure
- Generalized mixed logit model
- Hurdle model
- Identification through functional form
- Inclusive value
- Independence from irrelevant alternatives (IIA)
- Limited information
- Log-odds
- Method of simulated moments
- Mixed logit model
- Multinomial choice
- Multinomial logit model
- Multinomial probit model (MNP)
- Negative binomial distribution
- Negative binomial model
- Negbin 1 (NB1) form
- Negbin 2 (NB2) form
- Negbin P (NBP) model
- Nested logit model
- Ordered choice
- Overdispersion
- Parallel regression assumption
- Random coefficients
- Random parameters logit model (RPL)
- Revealed preference data
- Specification error
- Stated choice data
- Stated choice experiment
- Subjective well-being (SWB)
- Unlabeled choices
- Unordered choice model
- Willingness to pay space

Exercises

1. We are interested in the ordered probit model. Our data consist of 250 observations, of which the responses are

y	0	1	2	3	4
n	50	40	45	80	35

   Using the preceding data, obtain maximum likelihood estimates of the unknown parameters of the model. (*Hint:* Consider the probabilities as the unknown parameters.)
2. For the zero-inflated Poisson (ZIP) model in Section 18.4.8, we derived the conditional mean function, $E[y_i|\mathbf{x}_i, \mathbf{w}_i] = (1 - F_i)\lambda_i$.
   a. For the same model, now obtain $[Var[y_i|\mathbf{x}_i, \mathbf{w}_i]]$. Then, obtain $\tau_i = \text{Var}[y_i|\mathbf{x}_i, \mathbf{w}_i]/E[y_i|\mathbf{x}_i, \mathbf{w}_i]$. Does the zero inflation produce overdispersion? (That is, is the ratio greater than one?)
   b. Obtain the partial effect for a variable z_i that appears in both $\mathbf{w}_i$ and $\mathbf{x}_i$.
3. Consider estimation of a Poisson regression model for $y_i|x_i$. The data are truncated on the left—these are on-site observations at a recreation site, so zeros do not appear in the data set. The data are censored on the right—any response greater than 5 is recorded as a 5. Construct the log likelihood for a data set drawn under this sampling scheme.

Applications

1. Appendix Table F17.2 provides Fair's (1978) *Redbook Magazine* survey on extramarital affairs. The variables in the data set are as follows:

   id = an identification number,
   C = constant, value = 1,
   yrb = a constructed measure of time spent in extramarital affairs,
   v_1 = a rating of the marriage, coded 1 to 5,
   v_2 = age, in years, aggregated,
   v_3 = number of years married,
   v_4 = number of children, top coded at 5,
   v_5 = religiosity, 1 to 4, 1 = not, 4 = very,
   v_6 = education, coded 9, 12, 14, 16, 17, 20,
   v_7 = occupation,
   v_8 = husband's occupation,

   and three other variables that are not used. The sample contains a survey of 6,366 married women. For this exercise, we will analyze, first, the binary variable $A = 1$ if $yrb > 0$, 0 otherwise. The regressors of interest are v_1 to v_8. However, not necessarily all of them belong in your model. Use these data to build a binary choice model for A. Report all computed results for the model. Compute the partial effects for the variables you choose. Compare the results you obtain for a probit model to those for a logit model. Are there any substantial differences in the results for the two models?

2. Continuing the analysis of the first application, we now consider the self-reported rating, v_1. This is a natural candidate for an ordered choice model, because the simple five-item coding is a censored version of what would be a continuous scale on some subjective satisfaction variable. Analyze this variable using an ordered probit model. What variables appear to explain the response to this survey question? (*Note:* The variable is coded 1, 2, 3, 4, 5. Some programs accept data for ordered choice modeling in this form, for example, *Stata*, while others require the variable to be coded 0, 1, 2, 3, 4, for example, *NLOGIT*. Be sure to determine which is appropriate for the program you are using and transform the data if necessary.) Can you obtain the partial effects for your model? Report them as well. What do they suggest about the impact of the different independent variables on the reported ratings?
3. Several applications in the preceding chapters using the German health care data have examined the variable *DocVis*, the reported number of visits to the doctor. The data are described in Appendix Table F7.1. A second count variable in that data set that we have not examined is *HospVis*, the number of visits to hospital. For this application, we will examine this variable. To begin, we treat the full sample (27,326) observations as a cross section.
   a. Begin by fitting a Poisson regression model to this variable. The exogenous variables are listed in Appendix Table F7.1. Determine an appropriate specification for the right-hand side of your model. Report the regression results and the partial effects.
   b. Estimate the model using ordinary least squares and compare your least squares results to the partial effects you computed in part a. What do you find?
   c. Is there evidence of overdispersion in the data? Test for overdispersion. Now, reestimate the model using a negative binomial specification. What is the result? Do your results change? Use a likelihood ratio test to test the hypothesis of the negative binomial model against the Poisson.
4. The GSOEP data are an unbalanced panel, with 7,293 groups. Continue your analysis in Application 3 by fitting the Poisson model with fixed and with random effects and compare your results. (Recall, like the linear model, the Poisson fixed effects model may not contain any time-invariant variables.) How do the panel data results compare to the pooled results?
5. Appendix Table F18.3 contains data on ship accidents reported in McCullagh and Nelder (1983). The data set contains 40 observations on the number of incidents of wave damage for oceangoing ships. Regressors include aggregate months of service, and three sets of dummy variables, Type (1, . . . ,5), operation period (1960–1974 or 1975–1979), and construction period (1960–1964, 1965–1969, or 1970–1974). There are six missing values on the dependent variable, leaving 34 usable observations.
   a. Fit a Poisson model for these data, using the log of service months, four type dummy variables, two construction period variables, and one operation period dummy variable. Report your results.
   b. The authors note that the rate of accidents is supposed to be per period, but the exposure (aggregate months) differs by ship. Reestimate your model constraining the coefficient on log of service months to equal one.
   c. The authors take overdispersion as a given in these data. Do you find evidence of overdispersion? Show your results.

19

LIMITED DEPENDENT VARIABLES—TRUNCATION, CENSORING, AND SAMPLE SELECTION

19.1 INTRODUCTION

This chapter is concerned with **truncation** and censoring. As we saw in Section 18.4.6, these features complicate the analysis of data that might otherwise be amenable to conventional estimation methods such as regression. Truncation effects arise when one attempts to make inferences about a larger population from a sample that is drawn from a distinct subpopulation. For example, studies of income based on incomes above or below some poverty line may be of limited usefulness for inference about the whole population. Truncation is essentially a characteristic of the distribution from which the sample data are drawn. Censoring is a more common feature of recent studies. To continue the example, suppose that instead of being unobserved, all incomes below the poverty line are reported as if they were *at* the poverty line. The censoring of a range of values of the variable of interest introduces a distortion into conventional statistical results that is similar to that of truncation. Unlike truncation, however, censoring is a feature of the sample data. Presumably, if the data were not censored, they would be a representative sample from the population of interest. We will also examine a form of truncation called the **sample selection** problem. Although most empirical work in this area involves censoring rather than truncation, we will study the simpler model of truncation first. It provides most of the theoretical tools we need to analyze models of censoring and sample selection.

The discussion will examine the general characteristics of truncation, censoring, and sample selection, and then, in each case, develop a major area of application of the principles. The stochastic frontier model is a leading application of results for truncated distributions in empirical models.[1] Censoring appears prominently in the analysis of labor supply and in modeling of duration data. Finally, the sample selection model has appeared in all areas of the social sciences and plays a significant role in the evaluation of treatment effects and program evaluation.

19.2 TRUNCATION

In this section, we are concerned with inferring the characteristics of a full population from a sample drawn from a restricted part of that population.

[1]See Aigner, Lovell, and Schmidt (1977) and Fried, Lovell, and Schmidt (2008).

19.2.1 TRUNCATED DISTRIBUTIONS

A **truncated distribution** is the part of an untruncated distribution that is above or below some specified value. For instance, in Example 19.2, we are given a characteristic of the distribution of incomes above $100,000. This subset is a part of the full distribution of incomes which range from zero to (essentially) infinity.

THEOREM 19.1 Density of a Truncated Random Variable
If a continuous random variable x has pdf f(x) and a is a constant, then[2]

$$f(x|x > a) = \frac{f(x)}{\text{Prob}(x > a)}.$$

The proof follows from the definition of conditional probability and amounts merely to scaling the density so that it integrates to one over the range above a. Note that the truncated distribution is a conditional distribution.

Most recent applications based on continuous random variables use the **truncated normal distribution**. If x has a normal distribution with mean μ and standard deviation σ, then

$$\text{Prob}(x > a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi(\alpha),$$

where $\alpha = (a - \mu)/\sigma$ and $\Phi(.)$ is the standard normal cdf. The density of the truncated normal distribution is then

$$f(x|x > a) = \frac{f(x)}{1 - \Phi(\alpha)} = \frac{(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}}{1 - \Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi\left(\frac{x - \mu}{\sigma}\right)}{1 - \Phi(\alpha)},$$

where $\phi(.)$ is the standard normal pdf. The **truncated standard normal distribution**, with $\mu = 0$ and $\sigma = 1$, is illustrated for $a = -0.5, 0$, and 0.5 in Figure 19.1. Another truncated distribution that has appeared in the recent literature, this one for a discrete random variable, is the truncated at zero Poisson distribution,

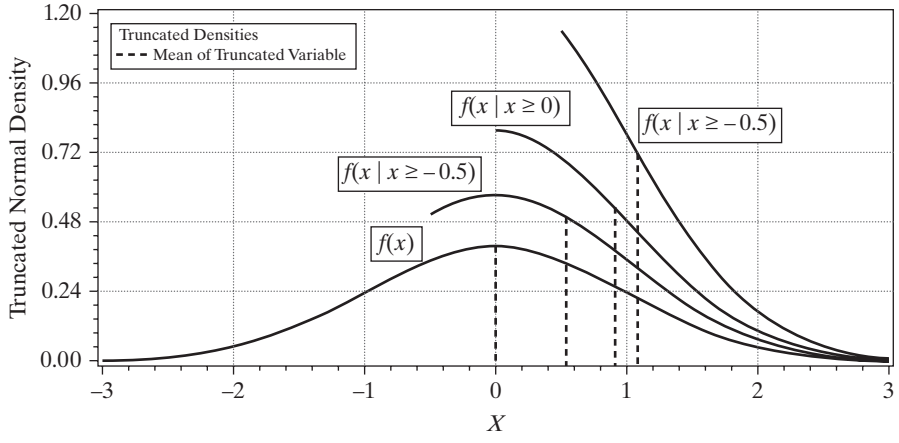
$$\text{Prob}[Y = y|y > 0] = \frac{(e^{-\lambda}\lambda^y)/y!}{\text{Prob}[Y > 0]} = \frac{(e^{-\lambda}\lambda^y)/y!}{1 - \text{Prob}[Y = 0]}$$
$$= \frac{(e^{-\lambda}\lambda^y)/y!}{1 - e^{-\lambda}}, \qquad \lambda > 0, y = 1, \ldots.$$

This distribution is used in models of uses of recreation and other kinds of facilities where observations of zero uses are discarded.[3]

[2]The case of truncation from above instead of below is handled in an analogous fashion and does not require any new results.

[3]See Shaw (1988) and Smith (1988). An application of this model appears in Section 18.4.6 and Example 18.18.

FIGURE 19.1 Truncated Normal Distributions.



For convenience in what follows, we shall call a random variable whose distribution is truncated a **truncated random variable**.

19.2.2 MOMENTS OF TRUNCATED DISTRIBUTIONS

We are usually interested in the mean and variance of the truncated random variable. They would be obtained by the general formula,

$$E[x|x > a] = \int_a^{\infty} xf(x|x > a)\, dx,$$

for the mean and likewise for the variance.

Example 19.1 Truncated Uniform Distribution

If x has a *standard* uniform distribution, denoted $U(0, 1)$, then

$$f(x) = 1, \qquad 0 \le x \le 1.$$

The truncated at $x = \frac{1}{3}$ distribution is also uniform:

$$f\left(x|x > \frac{1}{3}\right) = \frac{f(x)}{\text{Prob}(x > \frac{1}{3})} = \frac{1}{\left(\frac{2}{3}\right)} = \frac{3}{2}, \quad \frac{1}{3} \le x \le 1.$$

The expected value is

$$E\left[x|x > \frac{1}{3}\right] = \int_{1/3}^{1} x\left(\frac{3}{2}\right)dx = \frac{2}{3}.$$

For a variable distributed uniformly between L and U, the variance is $(U - L)^2/12$. Thus,

$$\text{Var}\left[x|x > \frac{1}{3}\right] = \frac{1}{27}.$$

The mean and variance of the untruncated distribution are $\frac{1}{2}$ and $\frac{1}{12}$, respectively.

Example 19.1 illustrates two results.

1. If the truncation is from below, then the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, then the mean of the truncated variable is smaller than the mean of the original one.
2. Truncation reduces the variance compared with the variance in the untruncated distribution.

Henceforth, we shall use the terms **truncated mean** and **truncated variance** to refer to the mean and variance of the random variable with a truncated distribution.

For the truncated normal distribution, we have the following theorem:[4]

THEOREM 19.2 Moments of the Truncated Normal Distribution
If $x \sim N[\mu, \sigma^2]$ *and* a *is a constant, then*

$$E[x \mid \text{truncation}] = \mu + \sigma\lambda(\alpha), \tag{19-1}$$

$$\text{Var}[x \mid \text{truncation}] = \sigma^2[1 - \delta(\alpha)], \tag{19-2}$$

where $\alpha = (a - \mu)/\sigma$, $\phi(\alpha)$ *is the standard normal density and*

$$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)] \quad \text{if truncation is } x > a, \tag{19-3a}$$

$$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha) \quad \text{if truncation is } x < a, \tag{19-3b}$$

and

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]. \tag{19-4}$$

An important result is

$$0 < \delta(\alpha) < 1 \quad \text{for all values of } \alpha,$$

which implies point 2 after Example 19.1. A result that we will use at several points below is $d\phi(\alpha)/d\alpha = -\alpha\phi(\alpha)$. The function $\lambda(\alpha)$ is called the **inverse Mills ratio**. The function in (19-3a) is also called the **hazard function** for the standard normal distribution.

Example 19.2 A Truncated Lognormal Income Distribution

An article that appeared in the *New York Post* in 1987 claimed that "The typical 'upper affluent American' . . . makes $142,000 per year . . . The people surveyed had household income of at least $100,000." Would this statistic tell us anything about the typical American? As it stands, it probably does not (popular impressions notwithstanding). The 1987 article where this appeared went on to state, "If you're in that category, pat yourself on the back—only 2% of American households make the grade, according to the survey." Because the **degree of truncation** in the sample is 98%, the $142,000 was probably quite far from the mean in the full population.

Suppose that incomes, x, in the population were lognormally distributed—see Section B.4.4. Then the log of income, y, had a normal distribution with, say, mean μ and standard deviation, σ.

[4]Details may be found in Johnson, Kotz, and Balakrishnan (1994, pp. 156–158). Proofs appear in Cameron and Trivedi (2005).

Suppose that the survey was large enough for us to treat the sample average as the true mean. Assuming so, we'll deduce μ and σ and then determine the population mean income.

Two useful numbers for this example are ln 100 = 4.605 and ln 142 = 4.956. The article states that Prob[$x \geq 100$] = Prob[exp(y) ≥ 100] = 0.02, or Prob($y < 4.605$) = 0.98. This implies that

$$\text{Prob}[(y - \mu)/\sigma < (4.605 - \mu)/\sigma] = 0.98.$$

Because $\Phi[(4.605 - \mu)/\sigma] = 0.98$, we know that $\Phi^{-1}(0.98) = 2.054 = (4.605 - \mu)/\sigma$, or

$$4.605 = \mu + 2.054\sigma.$$

The article also states that

$$E[x|x > 100] = E[\exp(y)|\exp(y) > 100] = 142,$$

or

$$E[\exp(y)|y > 4.645] = 142.$$

To proceed, we need another result for the lognormal distribution:[5]

$$\text{If } y \sim N[\mu, \sigma^2], \text{ then } E[\exp(y)|y > a] = \exp(\mu + \sigma^2/2) \times \frac{\Phi(\sigma - (a - \mu)/\sigma)}{1 - \Phi((a - \mu)/\sigma)}.$$

For our application, we would equate this expression to 142, and a to ln 100 = 4.605. This provides a second equation. To estimate the two parameters, we used the method of moments. We solved the minimization problem,

$$\text{Minimize}_{\mu,\sigma}[4.605 - (\mu + 2.054\sigma)]^2 + [142\Phi((\mu - 4.605)/\sigma) - \exp(\mu + \sigma^2/2)\Phi(\sigma - (4.605 - \mu)/\sigma)]^2.$$

The two solutions are 2.89372 and 0.83314 for μ and σ, respectively. To obtain the mean income, we now use the result that if $y \sim N[\mu, \sigma^2]$ and $x = \exp(y)$, then $E[x] = \exp(\mu + \sigma^2/2)$. Inserting our values for μ and σ gives $E[x] = \$25{,}554$. The *1987 Statistical Abstract of the United States* gave the mean of household incomes across all groups for the United States was about $25,000. So, the estimate based on surprisingly little information would have been relatively good. These meager data did, indeed, tell us something about the average American. To recap, we were able to deduce the overall mean from estimates of the truncated mean and variance and the theoretical relationships between the truncated and untruncated mean and variance.

19.2.3 THE TRUNCATED REGRESSION MODEL

In the model of the earlier examples, we now assume that

$$\mu = \mathbf{x}'\boldsymbol{\beta}$$

is the deterministic part of the classical regression model. Then

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon,$$

where

$$\varepsilon|\mathbf{x} \sim N[0, \sigma^2],$$

[5]See Johnson, Kotz, and Balakrishnan (1995, p. 241).

so that

$$y|\mathbf{x} \sim N[\mathbf{x}'\boldsymbol{\beta}, \sigma^2]. \tag{19-5}$$

We are interested in the distribution of y given that y is greater than the truncation point a. This is the result described in Theorem 19.2. It follows that

$$E[y|y > a] = \mathbf{x}'\boldsymbol{\beta} + \sigma\frac{\phi[(a - \mathbf{x}'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(a - \mathbf{x}'\boldsymbol{\beta})/\sigma]}. \tag{19-6}$$

The conditional mean is therefore a nonlinear function of a, σ, $\mathbf{x}$, and $\boldsymbol{\beta}$.

The partial effects in this model *in the subpopulation* can be obtained by writing

$$E[y|y > a] = \mathbf{x}'\boldsymbol{\beta} + \sigma\lambda(\alpha), \tag{19-7}$$

where now $\alpha = (a - \mathbf{x}'\boldsymbol{\beta})/\sigma$. For convenience, let $\lambda = \lambda(\alpha)$ and $\delta = \delta(\alpha)$. Then

$$\begin{aligned}\frac{\partial E[y|y > a]}{\partial \mathbf{x}} &= \boldsymbol{\beta} + \sigma(d\lambda/d\alpha)\frac{\partial \boldsymbol{\alpha}}{\partial \mathbf{x}} \\ &= \boldsymbol{\beta} + \sigma(\lambda^2 - \alpha\lambda)(-\boldsymbol{\beta}/\sigma) \\ &= \boldsymbol{\beta}(1 - \lambda^2 + \alpha\lambda) \\ &= \boldsymbol{\beta}(1 - \delta).\end{aligned} \tag{19-8}$$

Note the appearance of the scale factor $1 - \delta$ from the truncated variance. Because $(1 - \delta)$ is between zero and one, we conclude that for every element of $\mathbf{x}$, the partial effect is less than the corresponding coefficient. There is a similar **attenuation** of the variance. In the subpopulation $y > a$, the regression variance is not σ^2 but

$$\text{Var}[y|y > a] = \sigma^2(1 - \delta). \tag{19-9}$$

Whether the partial effect in (19-7) or the coefficient $\boldsymbol{\beta}$ itself is of interest depends on the intended inferences of the study. If the analysis is to be confined to the subpopulation, then (19-7) is of interest. If the study is intended to extend to the entire population, however, then it is the coefficients $\boldsymbol{\beta}$ that are actually of interest.

One's first inclination might be to use ordinary least squares (OLS) to estimate the parameters of this regression model. For the subpopulation from which the data are drawn, we could write (19-6) in the form

$$y|y > a = E[y|y > a] + u = \mathbf{x}'\boldsymbol{\beta} + \sigma\lambda + u, \tag{19-10}$$

where u is y minus its conditional expectation. By construction, u has a zero mean, but it is heteroscedastic,

$$\text{Var}[u] = \sigma^2(1 - \lambda^2 + \lambda\alpha) = \sigma^2(1 - \delta),$$

which is a function of $\mathbf{x}$. If we estimate (19-10) by least squares regression of $\mathbf{y}$ on $\mathbf{X}$, then we have omitted a variable, the nonlinear term λ. All the biases that arise because of an omitted variable can be expected.[6]

Without some knowledge of the distribution of $\mathbf{x}$, it is not possible to determine how serious the bias is likely to be. A result obtained by Chung and Goldberger (1984) is

[6]See Heckman (1979) who formulates this as a "specification error."

broadly suggestive. If $E[\mathbf{x}|y]$ in the full population is a linear function of y, then plim $\mathbf{b} = \boldsymbol{\beta}\tau$ for some proportionality constant τ. This result is consistent with the widely observed (albeit rather rough) proportionality relationship between least squares estimates of this model and maximum likelihood estimates.[7] The proportionality result appears to be quite general. In applications, it is usually found that, compared with consistent maximum likelihood estimates, the OLS estimates are biased toward zero. (See Example 19.5.)

19.2.4 THE STOCHASTIC FRONTIER MODEL

A lengthy literature commencing with theoretical work by Knight (1933), Debreu (1951), and Farrell (1957) and the pioneering empirical study by Aigner, Lovell, and Schmidt (1977) has been directed at models of production that specifically account for the textbook proposition that a production function is a theoretical ideal.[8] If $y = f(\mathbf{x})$ defines a production relationship between inputs, $\mathbf{x}$, and an output, y, then for any given $\mathbf{x}$, the observed value of y must be less than or equal to $f(\mathbf{x})$. The implication for an empirical regression model is that in a formulation such as $y = h(\mathbf{x}, \boldsymbol{\beta}) + u$, u must be negative. Because the theoretical production function is an ideal—the frontier of efficient production—any nonzero disturbance must be interpreted as the result of inefficiency. A strictly orthodox interpretation embedded in a Cobb–Douglas production model might produce an empirical frontier production model such as

$$\ln y = \beta_1 + \sum_k \beta_k \ln x_k - u, \quad u \geq 0.$$[9]

One-sided disturbances such as this one present a particularly difficult estimation problem. The primary theoretical problem is that any measurement error in $\ln y$ must be embedded in the disturbance. The practical problem is that the entire estimated function becomes a slave to any single errantly measured data point.

Aigner, Lovell, and Schmidt proposed instead a formulation within which observed deviations from the production function could arise from two sources: (1) productive inefficiency, as we have defined it earlier and that would necessarily be negative, and (2) idiosyncratic effects that are specific to the firm and that could enter the model with either sign. The end result was what they labeled the **stochastic frontier**:

$$\begin{aligned}\ln y &= \beta_1 + \sum_k \beta_k \ln x_k - u + v, \quad u \geq 0, \quad v \sim N\left[0, \sigma_v^2\right] \\ &= \beta_1 + \sum_k \beta_k \ln x_k + \varepsilon.\end{aligned}$$

The frontier for any particular firm is $h(\mathbf{x}, \boldsymbol{\beta}) + v$, hence the name *stochastic frontier*. The inefficiency term is u, a random variable of particular interest in this setting. Because the data are in log terms, u is a measure of the percentage by which the particular observation fails to achieve the frontier, ideal production rate.

---

[7] See the appendix in Hausman and Wise (1977), Greene (1983), Stoker (1986, 1992), and Powell (1994).

[8] A survey by Greene (2007a) appears in Fried, Lovell, and Schmidt (2008). Kumbhakar and Lovell (2000) and Kumbhakar and Parmeter (2014) are comprehensive references on the subject.

[9] For example, Greene (1990).

To complete the specification, they suggested two possible distributions for the inefficiency term: the absolute value of a normally distributed variable, which has the truncated at zero distribution shown in Figure 19.1, and an exponentially distributed variable. The density functions for these two compound variables are given by Aigner, Lovell, and Schmidt; let $\varepsilon = v - u, \lambda = \sigma_u/\sigma_v, \sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$, and $\Phi(z) =$ the probability to the left of z in the standard normal distribution (see Section B.4.1). The random variable that is obtained as $v - u$ where v and u are normal and half normal has a skew normal density,

$$f(\varepsilon) = \frac{2}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right)\Phi\left(\frac{-\lambda\varepsilon}{\sigma}\right).$$

This implies the log likelihood for the "half-normal" model,

$$\sum_{i=1}^{n} \ln h(\varepsilon_i|\boldsymbol{\beta}, \lambda, \sigma) = \sum_{i=1}^{n}\left[-\ln\sigma + \left(\frac{1}{2}\right)\ln\frac{2}{\pi} - \frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2 + \ln\Phi\left(\frac{-\varepsilon_i\lambda}{\sigma}\right)\right].$$

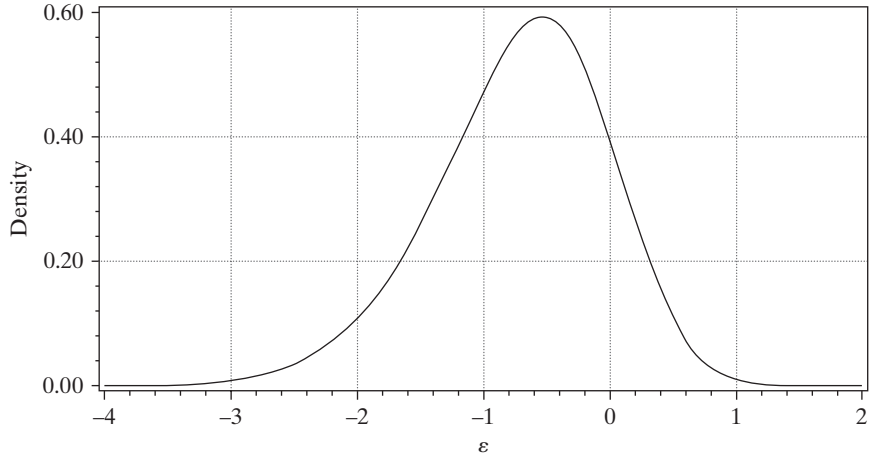
For the normal-exponential model with parameter θ, the log likelihood is

$$\sum_{i=1}^{n} \ln h(\varepsilon_i|\boldsymbol{\beta}, \theta, \sigma_v) = \sum_{i=1}^{n}\left[\ln\theta + \frac{1}{2}\theta^2\sigma_v^2 + \theta\varepsilon_i + \ln\Phi\left(-\frac{\varepsilon_i}{\sigma_v} - \theta\,\sigma_v\right)\right].$$

Both distributions are asymmetric. We thus have a regression model with a nonnormal distribution specified for the disturbance. The disturbance, ε, has a nonzero mean as well; $E[\varepsilon] = -\sigma_u(2/\pi)^{1/2}$ for the half-normal model and $-1/\theta$ for the exponential model. Figure 19.2 illustrates the density for the half-normal model with $\sigma = 1$ and $\lambda = 2$. By writing $\beta_0 = \beta_1 + E[\varepsilon]$ and $\varepsilon^* = \varepsilon - E[\varepsilon]$, we obtain a more conventional formulation,

$$\ln y = \beta_0 + \sum_k \beta_k \ln x_k + \varepsilon^*,$$

FIGURE 19.2 Skew Normal Density for the Disturbance in the Stochastic Frontier Model.



which does have a disturbance with a zero mean but an asymmetric, nonnormal distribution. The asymmetry of the distribution of ε^* does not negate our basic results for least squares in this classical regression model. This model satisfies the assumptions of the Gauss–Markov theorem, so least squares is unbiased and consistent (save for the constant term) and efficient among *linear unbiased* estimators. In this model, however, the maximum likelihood estimator is not linear, and it is more efficient than least squares.

The log-likelihood function for the half-normal model is given in Aigner, Lovell, and Schmidt (1977),

$$\ln L = -n \ln \sigma + \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\varepsilon_i}{\sigma} \right)^2 + \sum_{i=1}^{n} \ln \Phi \left(\frac{-\varepsilon_i \lambda}{\sigma} \right). \qquad \textbf{(19-11)}$$

Maximization programs for this model are built into modern software packages such as *Stata*, *SAS* and *NLOGIT*. The log likelihood is simple enough that it can also be readily adapted to the generic optimization routines in, for example, *Gauss*. Some treatments in the literature use the parameterization employed by Battese and Coelli (1992) and Coelli (1996), $\gamma = \sigma_u^2/\sigma^2$. This is a one-to-one transformation of λ; $\lambda = (\gamma/(1 - \gamma))^{1/2}$, so which parameterization is employed is a matter of convenience; the empirical results will be the same. The log-likelihood function for the exponential model can be built up from the density given earlier. For the half-normal model, we would also rely on the invariance of maximum likelihood estimators to recover estimates of the structural variance parameters, $\sigma_v^2 = \sigma^2/(1 + \lambda^2)$ and $\sigma_u^2 = \sigma^2\lambda^2/(1 + \lambda^2)$.[10] (Note, the variance of the truncated variable, u, is not σ_u^2; using (19-2), it reduces to $[(1 - 2/\pi)\sigma_u^2]$.) In addition, a structural parameter of interest is the proportion of the total variance of ε that is due to the inefficiency term. For the half-normal model, $\text{Var}[\varepsilon] = \text{Var}[u] + \text{Var}[v] = (1 - 2/\pi)\sigma_u^2 + \sigma_v^2$, whereas for the exponential model, the counterpart is $1/\theta^2 + \sigma_v^2$.

Modeling in the stochastic frontier setting is rather unlike what we are accustomed to up to this point, in that the inefficiency part of the disturbance, specifically u, not the model parameters, is the central focus of the analysis. The reason is that in this context, the disturbance, u, rather than being the catchall for the unknown and unknowable factors omitted from the equation, has a particular interpretation—it is the firm-specific inefficiency. Ideally, we would like to estimate u_i for each firm in the sample to compare them on the basis of their productive efficiency. Unfortunately, the data do not permit a direct estimate, because with estimates of $\boldsymbol{\beta}$ in hand, we are only able to compute a direct estimate of $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$. Jondrow et al. (1982), however, have derived a useful approximation that is now the standard measure in these settings,

$$E[u_i | \varepsilon_i] = \frac{\sigma\lambda}{1 + \lambda^2} \left[\frac{\phi(z_i)}{1 - \Phi(z_i)} - z_i \right], z_i = \frac{\varepsilon_i \lambda}{\sigma}$$

for the half-normal model, and

$$E[u_i | \varepsilon_i] = z_i + \sigma_v \frac{\phi(z_i/\sigma_v)}{\Phi(z_i/\sigma_v)}, z_i = -(\varepsilon_i + \theta\sigma_v^2)$$

[10] A vexing problem for estimation of the model is that if the OLS residuals are skewed in the positive (wrong) direction (see Figure 19.2), OLS with $\hat{\lambda} = 0$ will be the MLE. OLS residuals with a positive skew are apparently inconsistent with a model in which, in theory, they should have a negative skew. [See Waldman (1982) for theoretical development of this result.] There is a substantial literature on this issue, including, for example, Hafner, Manner, and Simar (2013).

for the exponential model. These values can be computed using the maximum likelihood estimates of the structural parameters in the model. In some cases in which researchers are interested in discovering best practice, the estimated values are sorted and the ranks of the individuals in the sample become of interest.[11]

Research in this area since the methodological developments beginning in the 1930s and the building of the empirical foundations in 1977 and 1982 has proceeded in several directions. Most theoretical treatments of inefficiency as envisioned here attribute it to aspects of management of the firm. It remains to establish a firm theoretical connection between the theory of firm behavior and the stochastic frontier model as a device for measurement of inefficiency.

In the context of the model, many studies have developed alternative, more flexible functional forms that (it is hoped) can provide a more realistic model for inefficiency. Two that are relevant in this chapter are Stevenson's (1980) truncated normal model and the normal-gamma frontier. One intuitively appealing form of the truncated normal model is

$$U_i \sim N[\mu + \mathbf{z}_i'\boldsymbol{\alpha}, \sigma_u^2], u_i = |U_i|.$$

The original normal–half-normal model results if μ equals zero and $\boldsymbol{\alpha}$ equals zero. This is a device by which the environmental variables noted in the next paragraph can enter the model of inefficiency. A truncated normal model is presented in Example 19.3. The half-normal, truncated normal, and exponential models all take the form of distribution shown in Figure 19.1. The gamma model,

$$f(u) = [\theta^P/\Gamma(P)]\exp(-\theta u)u^{P-1},$$

is a flexible model that presents the advantage that the distribution of inefficiency can move away from zero. If P is greater than one, then the density at $u = 0$ equals zero and the entire distribution moves away from the origin. The implication is that the distribution of inefficiency among firms can move away from zero. The gamma model is estimated by simulation methods—either Bayesian MCMC[12] or maximum simulated likelihood.[13] Many other functional forms have been proposed.[14]

There are usually elements in the environment in which the firm operates that impact the firm's output and/or costs but are not, themselves, outputs, inputs, or input prices. In Example 19.3, the costs of the Swiss railroads are affected by three variables: track width, long tunnels, and curvature. It is not yet specified how such factors should be incorporated into the model; four candidates are in the mean and variance of u_i, directly in the function, or in the variance of v_i.[15] All of these can be found in the received studies. This aspect of the model was prominent in the discussion of the famous World Health Organization (WHO) efficiency study of world health systems.[16] In Example 19.3, we have placed the environmental factors in the mean of the inefficiency distribution. This produces a rather extreme set of results for the JLMS estimates of

[11]For example, the World Health Organization (2000) and Tandon et al. (2000).

[12]Huang (2003) and Tsionas (2002).

[13]Greene (2003).

[14]See Greene (2007a) for a survey.

[15]See Hadri, Guermat, and Whittaker (2003) and Kumbhakar (1997a).

[16]WHO (2000), Tandon et al. (2000), and Greene (2004b).

inefficiency—many railroads are estimated to be extremely inefficient. An alternative formulation would be a heteroscedastic model in which $\sigma_{u,i} = \sigma_u \exp(\mathbf{z}_i'\boldsymbol{\delta})$ or $\sigma_{v,i} = \sigma_v \exp(\mathbf{z}_i'\boldsymbol{\eta})$, or both. We can see from the JLMS formula that the term heteroscedastic is actually a bit misleading, because both standard deviations enter (now) λ_i, which is, in turn, a crucial parameter in the mean of inefficiency.

How should inefficiency be modeled in panel data, such as in our example? It might be tempting to treat it as a time-invariant effect.[17] A string of studies, including Battese and Coelli (1992, 1995), Cuesta (2000), Kumbhakar (1997a), Kumbhakar and Orea (2004), and many others have proposed hybrid forms that treat the core random part of inefficiency as a time-invariant firm-specific effect that is modified over time by a deterministic, possibly firm-specific, function. The Battese–Coelli form,

$$u_{it} = \exp[-\eta(t - T)]\,|U_i| \text{ where } U_i N[0, \sigma_u^2],$$

has been used in a number of applications. Cuesta (2000) suggests allowing η to vary across firms, producing a model that bears some relationship to a fixed-effects specification. Greene (2004b) argued that a preferable approach would be to allow inefficiency to vary freely over time in a panel, and to the extent that there is a common time-invariant effect in the model, that should be treated as unobserved heterogeneity, not inefficiency. This produces the "true random effects model,"

$$\ln y_{it} = \alpha + w_i + \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} - u_{it}.$$

This is simply a random effects stochastic frontier model, as opposed to a random effects linear regression analyzed in Chapter 10. At first glance, it appears to be an extremely optimistic specification with three disturbances. The key to estimation and inference is to note that it is only a familiar two-part disturbance, the sum of a time-varying skew normal variable, $\varepsilon_{it} = v_{it} - u_{it}$, and a time-invariant normal variable w_i. Maximum simulated likelihood estimation of the model is developed in Greene (2004b).[18]

Is it reasonable to use a possibly restrictive parametric approach to modeling inefficiency? Sickles (2005) and Kumbhakar et al. (2007) are among numerous studies that have explored less parametric approaches to efficiency analysis. Proponents of **data envelopment analysis** have developed methods that impose absolutely no parametric structure on the production function.[19] Among the costs of this high degree of flexibility is a difficulty to include environmental effects anywhere in the analysis, and the uncomfortable implication that any unmeasured heterogeneity of any sort is necessarily included in the measure of inefficiency. That is, data envelopment analysis returns to the deterministic frontier approach where this section began.

Example 19.3 Stochastic Cost Frontier for Swiss Railroads

Farsi, Filippini, and Greene (2005) analyzed the cost efficiency of Swiss railroads. In order to use the stochastic frontier approach to analyze costs of production, rather than production, they rely on the fundamental duality of production and cost.[20] An appropriate cost frontier

[17] As in Schmidt and Sickles (1984) and Pitt and Lee (1984) in two pioneering papers.

[18] Colombi et al. (2014) and Filippini and Greene (2016) have extended this approach to a "generalized true random effects model," $y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + w_i - h_i + v_{it} - u_{it}$. In this case, both components of the random effects model have skew normal, rather than normal distributions. Estimation is carried out using maximum simulated likelihood.

[19] See, for example, Simar and Wilson (2000, 2007).

[20] See Samuelson (1938), Shephard (1953), and Kumbhakar and Lovell (2000).

model for a firm that produces more than one output—the Swiss railroads carry both freight and passengers—will appear as the following:

$$\ln(C/P_K) = \alpha + \sum_{k=1}^{K-1} \beta_k \ln (P_k/P_K) + \sum_{m=1}^{M} \gamma_m \ln Q_m + v + u.$$

The requirement that the cost function be homogeneous of degree one in the input prices has been imposed by normalizing total cost, C, and the first $K - 1$ prices by the Kth input price. In this application, the three factors are labor, capital, and electricity—the third is used as the numeraire in the cost function. Notice that the inefficiency term, u, enters the cost function positively; actual cost is above the frontier cost. [The MLE is modified simply by replacing ε_i with $-\varepsilon_i$ in (19-11).] In analyzing costs of production, we recognize that there is an additional source of inefficiency that is absent when we analyze production. On the production side, inefficiency measures the difference between output and frontier output, which arises because of technical inefficiency. By construction, if output fails to reach the efficient level for the given input usage, then costs must be higher than frontier costs. However, costs can be excessive even if the firm is technically efficient if it is *allocatively inefficient*. That is, the firm can be technically efficient while not using inputs in the cost minimizing mix (equating the ratio of marginal products to the input price ratios). It follows that on the cost side, "u" can contain both elements of inefficiency while on the production side, we would expect to measure only technical inefficiency.[21]

The data for this study are an unbalanced panel of 50 railroads with T_i ranging from 1 to 13. (Thirty-seven of the firms are observed 13 times, 8 are observed 12 times, and the remaining 5 are observed 10, 7, 7, 3, and 1 times, respectively.) The variables we will use here are:

CT: Total costs adjusted for inflation (1,000 Swiss francs),
QP: Total passenger-output in passenger-kilometers,
QF: Total goods-output in ton-kilometers,
PL: Labor price adjusted for inflation (in Swiss francs per person per year),
PK: Capital price with capital stock proxied by total number of seats,
PE: Price of electricity (Swiss francs per kWh).

Logs of costs and prices (ln CT, ln *PK*, ln *PL*) are normalized by *PE*. We will also use these environmental variables:

NARROW_T: Dummy for the networks with narrow track (1 meter wide) The usual width is 1.435 meters;
TUNNEL: Dummy for networks that have tunnels with an average length of more than 300 meters;
VIRAGE: Dummy for the networks whose minimum radius of curvature is 100 meters or less.

The full data set is given in Appendix Table F19.1. Several other variables not used here are presented in the appendix table. In what follows, we will ignore the panel data aspect of the data set. This would be a focal point of a more extensive study.

There have been dozens of models proposed for the inefficiency component of the stochastic frontier model. Table 19.1 presents several different forms. The basic half-normal model is given in the first column. The estimated cost function parameters across the different forms are broadly similar, as might be expected as (α, $\boldsymbol{\beta}$) are consistently estimated in all cases. There are fairly pronounced differences in the implications for the components of ε, however. There is an ambiguity in the model as to whether modifications to the distribution of u_i will affect the mean of the distribution, the variance, or both. The following results

[21]See Kumbhakar (1997b).

TABLE 19.1 Estimated Stochastic Frontier Cost Functions[a]

	Model					
Variable	*Half Normal*	*Truncated Normal*	*Exponential*	*Gamma*	*Heterosced.*	*Heterogen.*
Constant	−10.0799	−9.80624	−10.1838	−10.1944	−9.82189	−10.2891
ln QP	0.64220	0.62573	0.64403	0.64401	0.61976	0.63576
ln QF	0.06904	0.07708	0.06803	0.06810	0.07970	0.07526
ln PK	0.26005	0.26625	0.25883	0.25886	0.25464	0.25893
ln PL	0.53845	0.50474	0.56138	0.56047	0.53953	0.56036
Constant		0.44116			−2.48218[b]	
Narrow		0.29881			2.16264[b]	0.14355
Virage		−0.20738			−1.52964[b]	−0.10483
Tunnel		0.01118			0.35748[b]	−0.01914
σ	0.44240	0.38547	(0.34325)	(0.34288)	0.45392[c]	0.40597
λ	1.27944	2.35055				0.91763
P			1.0000	1.22920		
θ			13.2922	12.6915		
σ_u	(0.34857)	(0.35471)	(0.07523)	(0.09685)	0.37480[c]	0.27448
σ_v	(0.27244)	(0.15090)	0.33490	0.33197	0.25606	0.29912
Mean $E[u\|\varepsilon]$	0.27908	0.52858	0.075232	0.096616	0.29499	0.21926
ln L	−210.495	−200.67	−211.42	−211.091	−201.731	−208.349

[a]Estimates in parentheses are derived from other MLEs.
[b]Estimates used in computation of σ_u.
[c]Obtained by averaging $\lambda = \sigma_{u,i}/\sigma_v$ over observations.

suggest that it is both for these data. The gamma and exponential models appear to remove most of the inefficiency from the data. Note that the estimates of σ_u are considerably smaller under these specifications, and σ_v is correspondingly larger. The second-to-last row shows the sample averages of the Jondrow estimators—this estimates $E_\varepsilon E[u|\varepsilon] = E[u]$. There is substantial difference across the specifications.

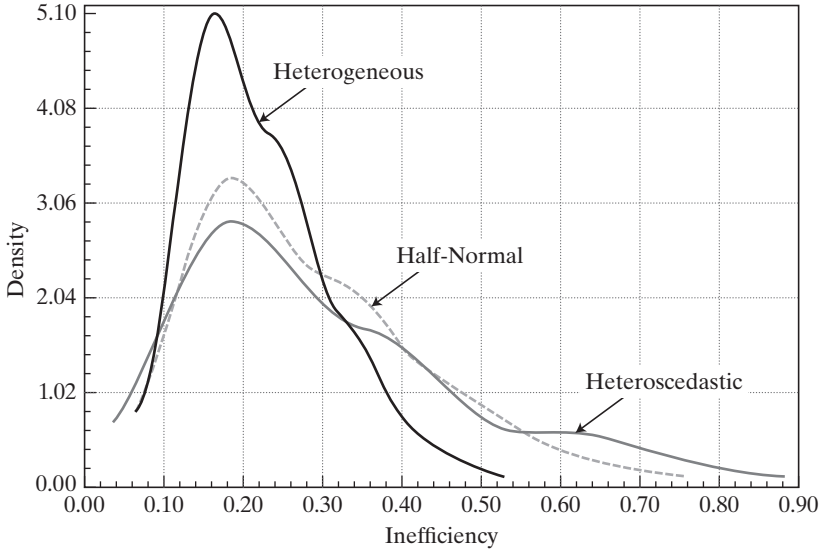
The estimates in the rightmost two columns illustrate two different placements of the measured heterogeneity: in the variance of u_i and directly in the cost function. The log-likelihood function appears to favor the first of these. However, the models are not nested and involve the same number of parameters. We used the Vuong test (see Section 14.6.6), instead and obtained a value of −2.65 in favor of the heteroscedasticity model. Figure 19.3 describes the values of $E[u_i|\varepsilon_i]$ estimated for the sample observations for the half-normal, heteroscedastic, and heterogeneous models. The smaller estimate of σ_u for the third of these is evident in the figure, which suggests a somewhat tighter concentration of values than the other two.

19.3 CENSORED DATA

A very common problem in microeconomic data is censoring of the dependent variable. When the dependent variable is censored, values in a certain range are all transformed to (or reported as) a single value. Some examples that have appeared in the empirical literature are as follows:[22]

[22]More extensive listings may be found in Amemiya (1984) and Maddala (1983).

FIGURE 19.3 Kernel Density Estimator for JLMS Estimates.



1. Household purchases of durable goods [Tobin (1958)],
2. The number of extramarital affairs [Fair (1977, 1978)],
3. The number of hours worked by a woman in the labor force [Quester and Greene (1982)],
4. The number of arrests after release from prison [Witte (1980)],
5. Household expenditures on various commodity groups [Jarque (1987)],
6. Vacation expenditures [Melenberg and van Soest (1996)],
7. Charitable donations [Brown, Harris, and Taylor (2009)].

Each of these studies analyzes a dependent variable that is zero for a significant fraction of the observations. Conventional regression methods fail to account for the qualitative difference between *limit* (zero) observations and *nonlimit* (continuous) observations.

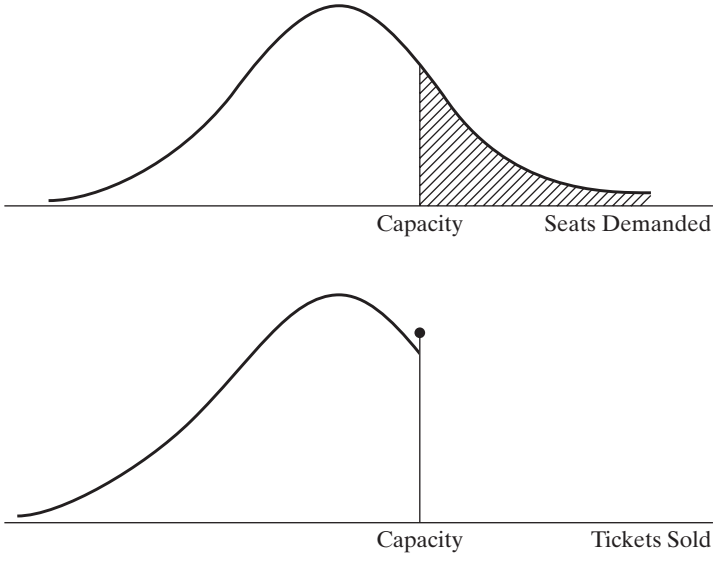
19.3.1 THE CENSORED NORMAL DISTRIBUTION

The relevant distribution theory for a **censored variable** is similar to that for a truncated one. Once again, we begin with the normal distribution, as much of the received work has been based on an assumption of normality. We also assume that the censoring point is zero, although this is only a convenient normalization. In a truncated distribution, only the part of distribution above $y = 0$ is relevant to our computations. To make the distribution integrate to one, we scale it up by the probability that an observation in the untruncated population falls in the range that interests us. When data are censored, the distribution that applies to the sample data is a mixture of discrete and continuous distributions. Figure 19.4 illustrates the effects.

To analyze this distribution, we define a new random variable y transformed from the original one, y^*, by

$$y = 0 \quad \text{if } y^* \leq 0,$$
$$y = y^* \quad \text{if } y^* > 0.$$

FIGURE 19.4 Partially Censored Distribution.



The **two-part distribution** that applies if $y^* \sim N[\mu, \sigma^2]$ is

$$\text{Prob}(y = 0) = \text{Prob}(y^* \leq 0) = \Phi(-\mu/\sigma) = 1 - \Phi(\mu/\sigma),$$

and if $y^* > 0$, *then* y has the density of y^*.

This distribution is a mixture of discrete and continuous parts. The total probability is one, as required, but instead of scaling the second part, we simply assign the full probability in the censored region to the censoring point, in this case, zero.

For the special case of $a = 0$, the mean simplifies to

$$E[y \mid a = 0] = \Phi(\mu/\sigma)(\mu + \sigma\lambda), \text{ where } \lambda = \frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}.$$

For censoring of the upper part of the distribution instead of the lower, it is only necessary to reverse the role of Φ and $1 - \Phi$ and redefine λ as in Theorem 19.2.

THEOREM 19.3 Moments of the Censored Normal Variable
If $y^* \sim N[\mu, \sigma^2]$ *and* $y = a$ *if* $y^* \leq a$ *or else* $y = y^*$, *then*

$$E[y] = \Phi a + (1 - \Phi)(\mu + \sigma\lambda),$$

and

$$\text{Var}[y] = \sigma^2(1 - \Phi)[(1 - \delta) + (\alpha - \lambda)^2\Phi],$$

where

$$\Phi[(a - \mu)/\sigma] = \Phi(\alpha) = \text{Prob}(y^* \leq a) = \Phi, \quad \lambda = \phi/(1 - \Phi),$$

and

$$\delta = \lambda^2 - \lambda\alpha.$$

Proof: *For the mean,*

$$\begin{aligned} E[y] &= \text{Prob}(y = a) \times E[y|y = a] + \text{Prob}(y > a) \times E[y|y > a] \\ &= \text{Prob}(y^* \leq a) \times a + \text{Prob}(y^* > a) \times E[y^*|y^* > a] \\ &= \Phi a + (1 - \Phi)(\mu + \sigma\lambda) \end{aligned}$$

using Theorem 19.2. For the variance, we use a counterpart to the decomposition in (B-69); that is, Var[y] = E[conditional variance] + Var[conditional mean], *and Theorem 19.2.*

Example 19.4 Censored Random Variable

We are interested in the number of tickets *demanded* for events at a certain arena. Our only measure is the number actually *sold*. Whenever an event sells out, however, we know that the actual number demanded is larger than the number sold. The number of tickets demanded is censored when it is transformed to obtain the number sold. Suppose that the arena in question has 20,000 seats and, in a recent season, sold out 25% of the time. If the average attendance, including sellouts, was 18,000, then what are the mean and standard deviation of the demand for seats? According to Theorem 19.3, the 18,000 is an estimate of

$$E[\textit{sales}] = 20{,}000(1 - \Phi) + (\mu + \sigma\lambda)\Phi.$$

Because this is censoring from above, rather than below, $\lambda = -\phi(\alpha)/\Phi(\alpha)$. The argument of Φ, ϕ, and λ is $\alpha = (20{,}000 - \mu)/\sigma$. If 25% of the events are sellouts, then $\Phi = 0.75$. Inverting the standard normal at 0.75 gives $\alpha = 0.675$. In addition, if $\alpha = 0.675$, then $-\phi(0.675)/0.75 = \lambda = -0.424$. This result provides two equations in μ and σ, (a) $18{,}000 = 0.25(20{,}000) + 0.75(\mu - 0.424\sigma)$ and (b) $0.675\sigma = 20{,}000 - \mu$. The solutions are $\sigma = 2426$ and $\mu = 18{,}362$.

For comparison, suppose that we were told that the mean of 18,000 applies only to the events that were *not* sold out and that, on average, the arena sells out 25% of the time. Now our estimates would be obtained from the equations (a) $18{,}000 = \mu - 0.424\sigma$ and (b) $0.675\sigma = 20{,}000 - \mu$. The solutions are $\sigma = 1820$ and $\mu = 18{,}772$.

19.3.2 THE CENSORED REGRESSION (TOBIT) MODEL

The regression model based on the preceding discussion is referred to as the **censored regression model** or the tobit model [in reference to Tobin (1958), where the model was first proposed]. The regression is obtained by making the mean in the preceding correspond to a classical regression model. The general formulation is usually given in terms of an index function,

$$\begin{aligned} y^* &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \\ y &= 0 \quad \text{if } y^* \leq 0, \\ y_i &= y_i^* \text{ if } y_i^* > 0. \end{aligned}$$

There are potentially three conditional mean functions to consider, depending on the purpose of the study. For the index variable, sometimes called the *latent variable,* $E[y^*|\mathbf{x}]$

is $\mathbf{x}'\boldsymbol{\beta}$. If the data are always censored, however, then this result will usually not be useful. Consistent with Theorem 19.3, for an observation randomly drawn from the population, which may or may not be censored,

$$E[y|\mathbf{x}] = \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)(\mathbf{x}'\boldsymbol{\beta} + \sigma\lambda),$$

where

$$\lambda = \frac{\phi[(0 - \mathbf{x}'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(0 - \mathbf{x}'\boldsymbol{\beta})/\sigma]} = \frac{\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)}{\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)}. \tag{19-12}$$

Finally, if we intend to confine our attention to uncensored observations, then the results for the truncated regression model apply. The limit observations should not be discarded, however, because the truncated regression model is no more amenable to least squares than the censored data model. It is an unresolved question which of these functions should be used for computing predicted values from this model. Intuition suggests that $E[y|\mathbf{x}]$ is correct, but authors differ on this point. For the setting in Example 19.4, for predicting the number of tickets sold, say, to plan for an upcoming event, the censored mean is obviously the relevant quantity. On the other hand, if the objective is to study the need for a new facility, then the mean of the latent variable y^* would be more interesting.

There are differences in the partial effects as well. For the index variable, $\dfrac{\partial E[y^*|\mathbf{x}]}{\partial \mathbf{x}} = \boldsymbol{\beta}$. But this result is not what will usually be of interest, because y_i^* is unobserved. For the observed data, y_i, the following general result will be useful:[23]

THEOREM 19.4 Partial Effects in the Censored Regression Model
In the censored regression model with latent regression $y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ and observed dependent variable, $y = a$ if $y^* \le a$, $y = b$ if $y^* \ge b$, and $y = y^*$ otherwise, where a and b are constants with $b > a$, let $f(\varepsilon)$ and $F(\varepsilon)$ denote the density and cdf of the standardized variable ε/σ where ε is a continuous random variable with mean 0 and variance σ^2, and $f(\varepsilon|\mathbf{x}) = f(\varepsilon)$. Then

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \boldsymbol{\beta} \times \text{Prob}[a < y^* < b].$$

Proof: *By definition,*

$$E[y|\mathbf{x}] = a\,\text{Prob}[y^* \le a|\mathbf{x}] + b\,\text{Prob}[y^* \ge b|\mathbf{x}] + \text{Prob}[a < y^* < b|\mathbf{x}]E[y^*|a < y^* < b|\mathbf{x}].$$

Let $\alpha_j = (j - \mathbf{x}'\boldsymbol{\beta})/\sigma$, $F_j = F(\alpha_j)$, $f_j = f(\alpha_j)$, and $j = a, b$. Then

$$E[y|\mathbf{x}] = aF_a + b(1 - F_b) + (F_b - F_a)E[y^*|a < y^* < b, \mathbf{x}].$$

[23] See Greene (1999) for the general result and Rosett and Nelson (1975) and Nakamura and Nakamura (1983) for applications based on the normal distribution.

Because $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma[(y^* - \mathbf{x}'\boldsymbol{\beta})/\sigma]$, *the conditional mean may be written*

$$E[y^* | a < y^* < b, \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} + \sigma E\left[\frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \middle| \frac{a - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{b - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right]$$
$$= \mathbf{x}'\boldsymbol{\beta} + \sigma \int_{\alpha_a}^{\alpha_b} \frac{(\varepsilon/\sigma) f(\varepsilon/\sigma)}{F_b - F_a} d\left(\frac{\varepsilon}{\sigma}\right).$$

Collecting terms, we have

$$E[y|\mathbf{x}] = a\, F_a + b(1 - F_b) + (F_b - F_a)\mathbf{x}'\boldsymbol{\beta} + \sigma \int_{\alpha_a}^{\alpha_b} \left(\frac{\varepsilon}{\sigma}\right) f\left(\frac{\varepsilon}{\sigma}\right) d\left(\frac{\varepsilon}{\sigma}\right).$$

Now, differentiate with respect to x. The only complication is the last term, for which the differentiation is with respect to the limits of integration. We use Leibnitz's theorem and use the assumption that $f(\varepsilon)$ *does not involve x. Thus,*

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \left(\frac{-\boldsymbol{\beta}}{\sigma}\right) a\, f_a - \left(\frac{-\boldsymbol{\beta}}{\sigma}\right) b f_b + (F_b - F_a)\boldsymbol{\beta} + (\mathbf{x}'\boldsymbol{\beta})(f_b - f_a)\left(\frac{-\boldsymbol{\beta}}{\sigma}\right)$$
$$+ \sigma[\alpha_b f_b - \alpha_a f_a]\left(\frac{-\boldsymbol{\beta}}{\sigma}\right).$$

After inserting the definitions of α_a *and* α_b, *and collecting terms, we find all terms sum to zero save for the desired result,*

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = (F_b - F_a)\boldsymbol{\beta} = \boldsymbol{\beta} \times \text{Prob}[a < y_i^* < b].$$

Note that this general result includes censoring in either or both tails of the distribution, and it does not assume that ε is normally distributed. For the standard case with censoring at zero and normally distributed disturbances, the result specializes to

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \boldsymbol{\beta}\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right).$$

Although not a formal result, this does suggest a reason why, in general, least squares estimates of the coefficients in a tobit model usually resemble the MLEs times the proportion of nonlimit observations in the sample.

McDonald and Moffitt (1980) suggested a useful decomposition of $\partial E[y_i|\mathbf{x}_i]/\partial \mathbf{x}_i$,

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \boldsymbol{\beta} \times \{\Phi[1 - \lambda(\alpha + \lambda)] + \phi(\alpha + \lambda)\},$$

where $\alpha = \mathbf{x}'\boldsymbol{\beta}/\sigma$, $\Phi = \Phi(\alpha)$, and $\lambda = \phi/\Phi$. Taking the two parts separately, this result decomposes the slope vector into

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \text{Prob}[y > 0]\frac{\partial E[y|\mathbf{x}, y > 0]}{\partial \mathbf{x}} + E[y|\mathbf{x}, y > 0]\frac{\partial\, \text{Prob}[y > 0]}{\partial \mathbf{x}}.$$

Thus, a change in $\mathbf{x}$ has two effects: It affects the conditional mean of y^* in the positive part of the distribution, and it affects the probability that the observation will fall in that part of the distribution.

19.3.3 ESTIMATION

The tobit model has become so routine and been incorporated in so many computer packages that despite formidable obstacles in years past, estimation is now essentially on the level of ordinary linear regression. The log likelihood for the censored regression model is

$$\ln L = \sum_{y_i>0} -\frac{1}{2}\left[\ln(2\pi) + \ln\sigma^2 + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right] + \sum_{y_i=0} \ln\left[1 - \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]. \quad \textbf{(19-13)}$$

The two parts correspond to the linear regression for the nonlimit observations and the relevant probabilities for the limit observations, respectively. This likelihood is a nonstandard type, because it is a mixture of discrete and continuous distributions. In a seminal paper, Amemiya (1973) showed that despite the complications, proceeding in the usual fashion to maximize $\ln L$ would produce an estimator with all the familiar desirable properties attained by MLEs.

The log-likelihood function is fairly involved, but **Olsen's reparameterization** (1978) simplifies things considerably. With $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$ and $\theta = 1/\sigma$, the log likelihood is

$$\ln L = \sum_{y_i>0} -\frac{1}{2}[\ln(2\pi) - \ln\theta^2 + (\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})^2] + \sum_{y_i=0} \ln[1 - \Phi(\mathbf{x}_i'\boldsymbol{\gamma})]. \quad \textbf{(19-14)}$$

The results in this setting are now very similar to those for the truncated regression. The Hessian is always negative definite, so Newton's method is simple to use and usually converges quickly. After convergence, the original parameters can be recovered using $\sigma = 1/\theta$ and $\beta = \boldsymbol{\gamma}/\theta$. The asymptotic covariance matrix for these estimates can be obtained from that for the estimates of $[\boldsymbol{\gamma}, \theta]$ using the delta method:

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}, \hat{\sigma}] = \hat{\mathbf{J}}\,\text{Asy.Var}[\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}]\hat{\mathbf{J}}',$$

where

$$\mathbf{J} = \begin{bmatrix} \partial\boldsymbol{\beta}/\partial\boldsymbol{\gamma}' & \partial\beta/\partial\theta \\ \partial\sigma/\partial\boldsymbol{\gamma}' & \partial\sigma/\partial\theta \end{bmatrix} = \begin{bmatrix} (1/\theta)\mathbf{I} & (-1/\theta^2)\boldsymbol{\gamma} \\ \mathbf{0}' & (-1/\theta^2) \end{bmatrix}.$$

Researchers often compute OLS estimates despite their inconsistency. Almost without exception, it is found that the OLS estimates are smaller in absolute value than the MLEs. A striking empirical regularity is that the maximum likelihood estimates can often be approximated by dividing the OLS estimates by the proportion of nonlimit observations in the sample.[24] The effect is illustrated in the last two columns of Table 19.2. Another strategy is to discard the limit observations, but we now see that just trades the censoring problem for the truncation problem.

[24]This concept is explored further in Greene (1980b), Goldberger (1981), and Chung and Goldberger (1984).

Example 19.5 Estimated Tobit Equations for Hours Worked

In their study of the number of hours worked in a survey year by a large sample of wives, Quester and Greene (1982) were interested in whether wives whose marriages were statistically more likely to dissolve hedged against that possibility by spending, on average, more time working. They reported the tobit estimates given in Table 19.2. The last figure in the table implies that a very large proportion of the women reported zero hours, so least squares regression would be inappropriate.

The figures in parentheses are the ratio of the coefficient estimate to the estimated asymptotic standard error. The dependent variable is hours worked in the survey year. *Small kids* is a dummy variable indicating whether there were children in the household. The *education difference* and *relative wage* variables compare husband and wife on these two dimensions. The wage rate used for wives was predicted using a previously estimated regression model and is thus available for all individuals, whether working or not. *Second marriage* is a dummy variable. Divorce probabilities were produced by a large microsimulation model presented in another study.[25] The variables used here were dummy variables indicating *mean* if the predicted probability was between 0.01 and 0.03 and *high* if it was greater than 0.03. The slopes are the partial effects described earlier.

Note the partial effects compared with the tobit coefficients. Likewise, the estimate of σ is quite misleading as an estimate of the standard deviation of hours worked. The effects of the divorce probability variables were as expected and were quite large. One of the questions raised in connection with this study was whether the divorce probabilities could reasonably be treated as independent variables. It might be that for these individuals, the number of hours worked was a significant determinant of the probability.

TABLE 19.2 Tobit Estimates of an Hours Worked Equation

	White Wives		*Black Wives*		*Least Squares*	*Scaled OLS*
	Coefficient	*Slope*	*Coefficient*	*Slope*		
Constant	−1,803.13		−2,753.87			
	(−8.64)		(−9.68)			
Small kids	−1324.84	−385.89	−824.19	−376.53	−352.63	−766.56
	(−19.78)		(−10.14)			
Education difference	−48.08	−14.00	22.59	10.32	11.47	24.93
	(−4.77)		(1.96)			
Relative wage	312.07	90.90	286.39	130.93	123.95	269.46
	(5.71)		(3.32)			
Second marriage	175.85	51.51	25.33	11.57	13.14	28.57
	(3.47)		(0.41)			
Mean divorce robability	417.39	121.58	481.02	219.75	219.22	476.57
	(6.52)		(5.28)			
High divorce probability	670.22	195.22	578.66	264.36	244.17	530.80
	(8.40)		(5.33)			
σ	1,559	618	1,511	826		
Sample size	7459		2798			
Proportion working	0.29		0.46			

[25]Orcutt, Caldwell, and Wertheimer (1976).

19.3.4 TWO-PART MODELS AND CORNER SOLUTIONS

The tobit model contains a restriction that might be unreasonable in an economic setting. Consider a behavioral outcome, $y =$ charitable donation. Two implications of the tobit model are that

$$\text{Prob}(y > 0|\mathbf{x}) = \text{Prob}(\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0|\mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)$$

and [from (19-7)]

$$E[y|y > 0, \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} + \sigma\,\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)/\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma).$$

Differentiating both of these, we find from (17-11) and (19-8),

$$\partial\,\text{Prob}(y > 0|\mathbf{x})/\partial\mathbf{x} = [\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)/\sigma]\boldsymbol{\beta} = \text{a positive multiple of } \boldsymbol{\beta},$$

$$\partial E[y|y > 0, \mathbf{x}]/\partial\mathbf{x} = \{[1 - \delta(\mathbf{x}'\boldsymbol{\beta}/\sigma)]/\sigma\}\boldsymbol{\beta} = \text{a positive multiple of } \boldsymbol{\beta}.$$

Thus, any variable that appears in the model affects the participation probability and the intensity equation with the same sign. In the case suggested, for example, it is conceivable that age might affect participation and intensity in different directions. Fin and Schmidt (1984) suggest another application, loss due to fire in buildings; older buildings might be more likely to have fires but, because of the greater value of newer buildings, the actual damage might be greater in newer buildings. This fact would require the coefficient on age to have different signs in the two functions, which is impossible in the tobit model because they are the same coefficient.

In an early study in this literature, Cragg (1971) proposed a somewhat more general model in which the probability of a limit observation is independent of the regression model for the nonlimit data. One can imagine, for instance, the decision of whether or not to purchase a car as being different from the decision of how much to spend on the car, having decided to buy one.

A more general, two-part model that accommodates these objections is as follows:

1. Participation equation:

$$\begin{aligned}&\text{Prob}[y^* > 0] = \Phi(\mathbf{x}'\boldsymbol{\gamma}),\ d = 1 \text{ if } y^* > 0,\\ &\text{Prob}[y^* \le 0] = 1 - \Phi(\mathbf{x}'\boldsymbol{\gamma}),\ d = 0 \text{ if } y^* \le 0.\end{aligned} \tag{19-15}$$

2. Intensity equation *for nonlimit observations*:

$$E[y|d = 1] = \mathbf{x}'\boldsymbol{\beta} + \sigma\lambda,$$

according to Theorem 19.2. This two-part model is a combination of the truncated regression model of Section 19.2 and the univariate probit model of Section 17.3, which suggests a method of analyzing it. Note that it is precisely the same approach we considered in Section 18.4.8 and Example 18.21 where we used a hurdle model to model doctor visits. The tobit model returns if $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$. The parameters of the regression (intensity) equation can be estimated independently using the truncated regression model of Section 19.2. An application is Melenberg and van Soest (1996).

Based only on the tobit model, Fin and Schmidt (1984) devised a Lagrange multiplier test of the restriction of the tobit model that, although a bit cumbersome algebraically, can be computed without great difficulty. If one is able to estimate the truncated regression model, the tobit model, and the probit model separately, then

there is a simpler way to test the hypothesis. The tobit log likelihood is the sum of the log likelihoods for the truncated regression and probit models. To show this result, add and subtract $\sum_{y_i=1} \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta})$ in (19-13). This produces the log likelihood for the truncated regression model (considered in the exercises) plus (17-19) for the probit model. Therefore, a likelihood ratio statistic can be computed using

$$\lambda = -2[\ln LT - (\ln LP + \ln LTR)],$$

where

$$\begin{aligned} LT &= \text{likelihood for the tobit model in (19-13), with the same coefficients,} \\ LP &= \text{likelihood for the probit model in (17-16), fit separately,} \\ LTR &= \text{likelihood for the truncated regression model, fit separately.} \end{aligned}$$

The two-part model just considered extends the tobit model, but it stops a bit short of the generality we might achieve. In the preceding hurdle model, we have assumed that the same regressors appear in both equations. Although this produces a convenient way to retreat to the tobit model as a parametric restriction, it couples the two decisions perhaps unreasonably. In our example to follow, where we model extramarital affairs, the decision whether or not to spend any time in an affair may well be an entirely different decision from how much time to spend having once made that commitment. The obvious way to proceed is to reformulate the hurdle model as

1. Participation equation

$$\begin{aligned} \text{Prob}[d^* > 0] &= \Phi(\mathbf{z}'\boldsymbol{\gamma}), & d &= 1 \text{ if } d^* > 0, \\ \text{Prob}[d^* \le 0] &= 1 - \Phi(\mathbf{z}'\boldsymbol{\gamma}), & d &= 0 \text{ if } d^* \le 0. \end{aligned} \qquad \textbf{(19-16)}$$

2. Intensity equation for nonlimit observations

$$E[y \mid d = 1] = \mathbf{x}'\boldsymbol{\beta} + \sigma\,\lambda.$$

This extension, however, omits an important element; it seems unlikely that the two decisions would be uncorrelated; that is, the implicit disturbances in the equations should be correlated. The combination of these produces what has been labeled a **type II tobit model**. [Amemiya (1985) identified five possible permutations of the model specification and observation mechanism. The familiar tobit model is type I; this is type II.] The full model is:

1. Participation equation

$$\begin{aligned} d^* &= \mathbf{z}'\boldsymbol{\gamma} + u, & u &\sim N[0, 1] \\ d &= 1 \text{ if } d^* > 0, & &0 \text{ otherwise.} \end{aligned}$$

2. Intensity equation

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N[0, \sigma^2].$$

3. Observation mechanism

$$\begin{aligned} &\text{(a) } y^* = 0 \text{ if } d = 0 \text{ and } y = y^* \text{ if } d = 1. \\ &\text{(b) } y = y^* \text{ if } d = 1 \text{ and } y \text{ is unobserved if } d = 0. \end{aligned}$$

4. Endogeneity

$$(u, \varepsilon) \sim \text{bivariate normal with correlation } \rho.$$

Mechanism (a) produces Amemiya's type II model. Amemiya (1984) blends these two interpretations. In the statement of the model, he presents (a), but in the subsequent discussion, assumes (b). The difference is substantive if $\mathbf{x}$ is observed in case (b). Otherwise, they are the same, and "$y = 0$" is not actually meaningful. Amemiya notes, "$y^* = 0$ merely signifies the event $d^* \leq 0$." If $\mathbf{x}$ is observed when $d = 0$, then these observations will contribute to the likelihood for the full sample. If not, then they will not. We will develop this idea later when we consider Heckman's selection model [which is case (b) without observed $\mathbf{x}$ when $d = 0$].

There are two estimation strategies that can be used to fit the type II model. A two-step method can proceed as follows: The probit model for d can be estimated using maximum likelihood as shown in Section 17.3. For the second step, we make use of our theorems on truncation (and Theorem 19.5 that will appear later) to write

$$\begin{aligned} E[y \mid d = 1, \mathbf{x}, \mathbf{z}] &= \mathbf{x}'\boldsymbol{\beta} + E[\varepsilon \mid d = 1, \mathbf{x}, \mathbf{z}] \\ &= \mathbf{x}'\boldsymbol{\beta} + \rho\sigma \frac{\phi(\mathbf{z}'\boldsymbol{\gamma})}{\Phi(\mathbf{z}'\boldsymbol{\gamma})} \\ &= \mathbf{x}'\boldsymbol{\beta} + \rho\sigma\lambda. \end{aligned} \tag{19-17}$$

Since we have estimated $\boldsymbol{\gamma}$ at step 1, we can compute $\hat{\lambda} = \phi(\mathbf{z}'\hat{\boldsymbol{\gamma}})/\Phi(\mathbf{z}'\hat{\boldsymbol{\gamma}})$ using the first-step estimates, and we can estimate $\boldsymbol{\beta}$ and $\theta = (\rho\sigma)$ by least squares regression of y on $\mathbf{x}$ and $\hat{\lambda}$. It will be necessary to correct the asymptotic covariance matrix that is computed for $(\hat{\boldsymbol{\beta}}, \hat{\theta})$. This is a template application of the Murphy and Topel (2002) results that appear in Section 14.7. The second approach is full information maximum likelihood, estimating all the parameters in both equations simultaneously. We will return to the details of estimation of the type II tobit model in Section 19.4 where we examine Heckman's model of "sample selection" model (which is the type II tobit model).

Many of the applications of the tobit model in the received literature are constructed not to accommodate censoring of the underlying data, but, rather, to model the appearance of a large cluster of zeros. Cragg's application is clearly related to this phenomenon. Consider, for example, survey data on purchases of consumer durables, firm expenditure on research and development, household charitable contributions, or consumer savings. In each case, the observed data will consist of zero or some positive amount. Arguably, there are two decisions at work in these scenarios: First, whether to engage in the activity or not, and second, given that the answer to the first question is yes, how intensively to engage in it—how much to spend, for example. This is precisely the motivation behind the hurdle model. This specification has been labeled a "**corner solution model**"; see Wooldridge (2010, Chapter 17).

In practical terms, the difference between the **hurdle model** and the tobit model should be evident in the data. Often overlooked in tobit analyses is that the model predicts not only a cluster of zeros (or limit observations), but also a grouping of observations *near zero* (or the limit point). For example, the tobit model is surely misspecified for the sort of (hypothetical) spending data shown in Figure 19.5 for a sample of 1,000 observations. Neglecting for the moment the earlier point about the underlying decision process, Figure 19.6 shows the characteristic appearance of a (substantively) censored variable. The implication for the model builder is that an appropriate specification would consist of two equations, one for the "participation decision," and one for the distribution

FIGURE 19.5 Hypothetical Spending Data; Vertical Axis Is Sample Proportions.

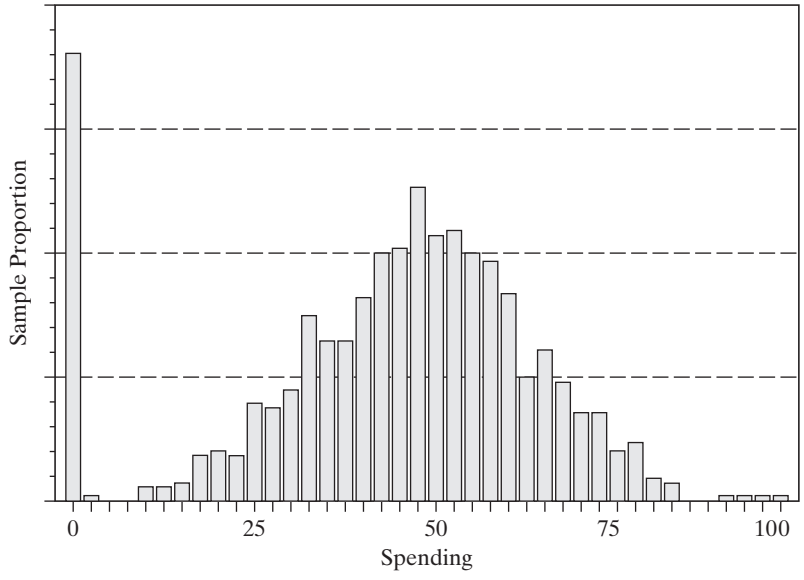
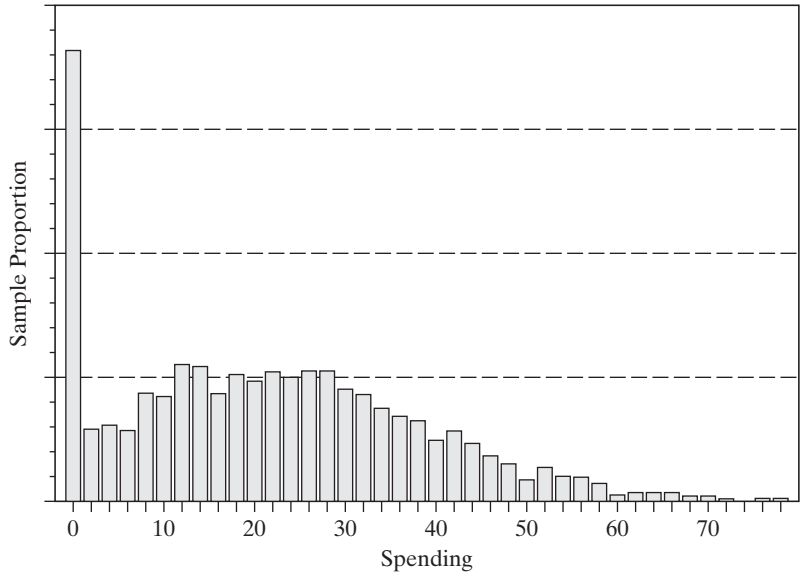


FIGURE 19.6 Hypothetical Censored Data; Vertical Axis Is Sample Proportions.



of the positive dependent variable. Formally, we might, continuing the development of Cragg's specification, model the first decision with a binary choice (e.g., probit or logit model). The second equation is a model for $y \mid y > 0$, for which the truncated regression model of Section 19.2.3 is a natural candidate. As we will see, this is essentially the model behind the sample selection treatment developed in Section 19.4.

Two practical issues frequently intervene at this point. First, one might well have a model in mind for the intensity (regression) equation, but none for the participation equation. This is the usual backdrop for the uses of the tobit model, which produces the considerations in the previous section. The second issue concerns the appropriateness of the truncation or censoring model to data such as those in Figure 19.6. If we consider only the nonlimit observations in Figure 19.5, the underlying distribution does not appear to be truncated at all. The truncated regression model in Section 19.2.3 fit to these data will not depart significantly from the linear model fit by OLS [because the underlying probability in the denominator of (19-6) will equal one and the numerator will equal zero]. But this is not the case of a tobit model forced on these same data. Forcing the model in (19-13) on data such as these will significantly distort the estimator—all else equal, it will significantly attenuate the coefficients, the more so the larger the proportion of limit observations in the sample. Once again, this stands as a caveat for the model builder. The tobit model is manifestly misspecified for data such as those in Figure 19.5.

Example 19.6 Two-Part Model for Extramarital Affairs

In Example 18.18, we examined Fair's (1977) *Psychology Today* survey data on extramarital affairs. The 601 observations in the data set are mostly zero—451 of the 601. This feature of the data motivated the author to use a tobit model to analyze these data. In our example, we reconsidered the model, because the nonzero observations were a count, not a continuous variable. Another data set in Fair's study was the *Redbook Magazine* survey of 6,366 married women. Once again, the outcome variable of interest was extramarital affairs. However, in this instance, the outcome data were transformed to a measure of time spent, which, being continuous, lends itself more naturally to the tobit model we are studying here. The variables in the data set are as follows (excluding three unidentified and not used):

id = Identification number,
C = Constant, value = 1
yrb = Constructed measure of time spent in extramarital affairs,
v_1 = Rating of the marriage, coded 1 to 5,
v_2 = Age, in years, aggregated,
v_3 = Number of years married,
v_4 = Number of children, top coded at 5,
v_5 = Religiosity, 1 to 4, 1 = not, 4 = very,
v_6 = Education, coded 9, 12, 14, 16, 17, 20,
v_7 = Wife's occupation—Hollingshead scale,
v_8 = Husband's occupation—Hollingshead scale.

This is a cross section of 6,366 observations with 4,313 zeros and 2,053 positive values.

Table 19.3 presents estimates of various models for *yrb*. The leftmost column presents the OLS estimates. The least squares estimator is inconsistent in this model. The empirical

regularity is that the OLS estimator appears to be biased toward zero, the more so the larger the proportion of limit observations. Here, the ratio, based on the tobit estimates in the second column, appears to be about 4 or 5 to 1. Likewise, the OLS estimator of σ appears to be greatly underestimated. This would be expected, as the OLS estimator is treating the limit observations, which have no variation in the dependent variable, as if they were nonlimit observations. The third set of results is the truncated regression estimator. In principle, the truncated regression estimator is also consistent. However, it will be less efficient as it is based on less information. In our example, this estimator seems to be quite erratic, again compared to the tobit estimator. Note, for example, the coefficient on years married, which, although it is "significant" in both cases, changes sign. The *t* ratio on Religiousness falls from -11.11 to -1.29 in the truncation model. The probit estimator based on $yrb > 0$ appears next. As a rough check on the corner solution aspect of our model, we would expect the normalized tobit coefficients ($\boldsymbol{\beta}/\sigma$) to approximate the probit coefficients, which they appear to. However, the likelihood ratio statistic for testing the internal consistency based on the three estimated models is $2[7{,}804.38 - 3{,}463.71 - 3{,}469.58] = 1{,}742.18$ with nine degrees of freedom. The hypothesis of parameter constancy implied by the tobit model is rejected. The last two sets of results are for a hurdle model in which the intensity equation is fit by the two-step method.

TABLE 19.3 Estimated Censored Regression Models (*t* ratios in parentheses)

	Model						
	Linear OLS	***Tobit***	***Truncated Regression***	***Probit***	***Tobit/σ***	***Hurdle Participation***	***Hurdle Intensity***
Constant	3.62346	7.83653	8.89449	2.21010	1.74189	1.56419	4.84602
	(13.63)	(10.98)	(2.90)	(12.60)		(17.75)	(5.87)
RateMarr	−0.42053	−1.53071	−0.44303	−0.42874	−0.34024	−0.42582	−0.24603
	(−14.79)	(−20.85)	(−1.45)	(−23.40)		(−23.61)	(−.46)
Age	−0.01457	−0.10514	−0.22394	−0.03542	−0.02337		−0.01903
	(−1.59)	(−4.24)	(−1.83)	(−5.87)			(−.77)
YrsMarr	−0.01599	0.12829	−0.94437	0.06563	0.02852		−0.16822
	(−1.62)	(4.86)	(−7.27)	(10.18)			(−6.52)
NumKids	−0.01705	−0.02777	−0.02280	−0.00394	−0.00617	0.14024	−0.28365
	(−.57)	(−0.36)	(−0.06)	(−0.21)		(11.55)	(−1.49)
Religious	−0.24374	−0.94350	−0.50490	−0.22281	−0.20972	−0.21466	−0.05452
	(−7.83)	(−11.11)	(−1.29)	(−10.88)		(−10.64)	(−0.19)
Education	−0.01743	−0.08598	−0.06406	−0.02373	−0.01911		0.00338
	(−1.24)	(−2.28)	(−0.38)	(−2.60)			(0.09)
Wife Occ.	0.06577	0.31284	0.00805	0.09539	0.06954		0.01505
	(2.10)	(3.82)	(0.02)	(4.75)			(0.19)
Hus. Occ.	0.00405	0.01421	−0.09946	0.00659	0.00316		−0.02911
	(0.19)	(0.26)	(−0.41)	(0.49)			(−0.53)
σ	2.14351	4.49887	5.46846				3.43748
$\ln L$	$R^2 = 0.05479$	−7,804.38	−3,463.71	−3,469.58			

19.3.5 SPECIFICATION ISSUES

Three issues that commonly arise in microeconomic data, endogeneity, heteroscedasticity, and nonnormality, have been analyzed at length in the tobit setting.[26]

19.3.5a Endogenous Right-Hand-Side Variables

We consider the case of an endogenous variable on the right-hand side of the tobit model. The structure is

$$\text{(latent variable) } y^* = \mathbf{x}'\boldsymbol{\beta} + \gamma T + \varepsilon, \text{ (observed variable) } y = \text{Max}(0, y^*)$$

$$\text{(latent variable) } T^* = \mathbf{z}'\boldsymbol{\alpha} + u, \text{ (observed variable) } T = h(T^*)$$

$$(u,\varepsilon) \sim \text{Bivariate normal } [(0, 1), (\sigma_u^2, \sigma_{u\varepsilon}, \sigma_\varepsilon^2)], \rho = \sigma_{u\varepsilon}/(\sigma_u\sigma_\varepsilon).$$

As usual, there are two cases to consider, the continuous variable, $h(T^*) = T^*$, and the endogenous dummy variable, $h(T^*) = \mathbf{1}(T^* > 0)$. (A probit model governs the observation of T.)

For the continuous case, again as usual, there are two-step and FIML estimators. We can use the reduced form to set up the two-step estimator. If (u,ε) are bivariate normally distributed, then $\varepsilon = \delta u + w$ where w is independent of u, $\text{Var}[w] = \sigma_\varepsilon^2(1 - \rho^2)$ and $\delta = \sigma_{u\varepsilon}/\sigma_u^2 = \rho\sigma_\varepsilon/\sigma_u$. Insert this in the tobit equation,

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \gamma T + \delta u + w, y = \text{Max}(0, y^*),$$

and recall $u = T - \mathbf{z}'\boldsymbol{\alpha}$, so

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \gamma T + \delta(T - \mathbf{z}'\boldsymbol{\alpha}) + w, y = \text{Max}(0, y^*).$$

The model can now be estimated in two steps: (1) Least squares regression of T on $\mathbf{z}$ consistently estimates $\boldsymbol{\alpha}$. Use these OLS estimates to construct residuals (control functions), $\hat{u}_i$ for each observation. (2) The second step consists of ML estimation of $\boldsymbol{\beta}$, γ and δ by ML tobit estimation of y on $(\mathbf{x}, T, \hat{u})$. The original parameters can be deduced from the second-step estimators, using $\delta = \rho\sigma_\varepsilon/\sigma_\text{u}$, the ML estimator of σ_w^2, which estimates $\sigma_\varepsilon^2(1 - \rho^2)$, and $\hat{\sigma}_u^2 = (1/n)\Sigma_{i=1}^n \hat{u}_i^2$. [27] For inference purposes, the covariance matrix at the second step must be corrected for the presence of $\hat{\boldsymbol{\alpha}}$. The Murphy and Topel correction promises to be quite complicated; bootstrapping would seem to be an appealing alternative.[28]

Blundell and Smith (1986) have devised an FIML estimator by reparameterizing the bivariate normal $f(u, \varepsilon) = f(\varepsilon|u)f(u)$ and concentrating the log-likelihood function

[26] Two symposia that contain numerous results on these subjects are Blundell (1987) and Duncan (1986b). An application that explores these two issues in detail is Melenberg and van Soest (1996). Developing specification tests for the tobit model has been a popular enterprise. A sampling of the received literature includes Nelson (1981), Bera, Jarque, and Lee (1982), Chesher and Irish (1987), Chesher, Lancaster, and Irish (1985), Gourieroux et al. (1984, 1987), Newey (1985a,b, 1986), Rivers and Vuong (1988), Horowitz and Neumann (1989), and Pagan and Vella (1989) are useful references on the general subject of conditional moment testing. More general treatments of specification testing are Godfrey (1988) and Ruud (1982).

[27] For example, $\hat{\tau}^2 = \hat{\delta}^2\hat{\sigma}_u^2/\hat{\sigma}_w^2$ then $\hat{\rho} = \text{sgn}(\hat{\delta})\hat{\tau}/\sqrt{1 + \hat{\tau}^2}$.

[28] A test for endogeneity can be carried out in the tobit model without correcting the covariance matrix with a simple t test of the hypothesis that δ equals zero.

over $\hat{\sigma}_u^2 = (1/n)\Sigma_{i=1}^n (T_i - \mathbf{z}_i'\boldsymbol{\alpha})^2 = (1/n)\Sigma_{i=1}^n u_i^2$. However $\boldsymbol{\alpha}$ is estimated, this will be the estimator of σ_u^2. This is inserted into the log likelihood to concentrate σ_u out of this estimation step. Then, define $e_i = (y_i - \mathbf{x}_i'\boldsymbol{\beta} - \gamma T_i - \psi u_i)$, where $\psi = \sigma_{u\varepsilon}/{\sigma_u}^2$ and $\omega = [\sigma_\varepsilon^2(1 - \rho^2)]^{1/2}$. Assembling the parts,

$$\ln L_c(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}, \psi, \omega) = \frac{-n}{2}\ln\frac{1}{n}\sum_{i=1}^n (T_i - \mathbf{z}_i'\boldsymbol{\alpha})^2 + \sum_{y_i>0}\ln\left[\frac{1}{\psi}\phi\left(\frac{e_i}{\psi}\right)\right] + \sum_{y_i=0}\ln\Phi\left[\frac{e_i - y_i}{\omega}\right].$$

The function is maximized over $\boldsymbol{\beta}$, γ, $\boldsymbol{\alpha}$, ψ, and ω; σ_u is estimated residually with the mean squared residual from T_i. Estimates of ρ and σ_ε can be recovered by the method of moments.

There is no simple two-step approach when $h(T^*) = \mathbf{1}(T^* > 0)$, the endogenous treatment case. However, there is a straightforward FIML estimator. There are four terms in the log likelihood for (y_i, T_i). For the two "limit" cases when $y = 0$, the terms are exactly those in the log likelihood for the bivariate probit with endogenous treatment effect in Section 17.6.1. Thus, for these two cases, $(y = 0, T = 0)$ and $(y = 0, T = 1)$,

$$\ln L_i = \ln \Phi_2\left[\frac{-\mathbf{x}_i'\boldsymbol{\beta} - \gamma T_i}{\sigma_\varepsilon}, (2T_i - 1)\mathbf{z}_i'\boldsymbol{\alpha}, -(2T_i - 1)\rho\right].$$

For the cases when y^* is observed, the terms in the log likelihood are exactly those in the FIML estimator for the sample selection model in (19-25) (Section 19.4.3). For these cases with $(y,T = 0)$ and $(y,T = 1)$,

$$\ln L_i = \ln\left[\frac{\exp(-(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \gamma T_i)^2/(2\sigma_\varepsilon^2))}{\sigma_\varepsilon\sqrt{2\pi}}\Phi\left(\frac{\rho(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \gamma T_i)/\sigma_\varepsilon + (2T - 1)\mathbf{z}_i'\boldsymbol{\alpha}}{\sqrt{1 - \rho^2}}\right)\right].$$

With any of these consistent estimators in hand, estimates of the average partial effects can be estimated based on $\partial E[y|\mathbf{x},T]/\partial T = \gamma\Phi[(\mathbf{x}'\boldsymbol{\beta} + \gamma T)/\sigma_\varepsilon]$ and likewise for the variables in $\mathbf{x}$. For the treatment effect case, we would use

$$\Delta E[y|\mathbf{x},T] =$$
$$\left[(\mathbf{x}'\boldsymbol{\beta} + \gamma)\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} + \gamma}{\sigma_\varepsilon}\right) + \sigma_\varepsilon\phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} + \gamma}{\sigma_\varepsilon}\right)\right] - \left[\left(\mathbf{x}'\boldsymbol{\beta}\right)\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_\varepsilon}\right) + \sigma_\varepsilon\phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_\varepsilon}\right)\right].$$

19.3.5.b Heteroscedasticity

Maddala and Nelson (1975), Hurd (1979), Arabmazar and Schmidt (1982a,b), and Brown and Moffitt (1982) all suggest varying degrees of pessimism regarding how inconsistent the maximum likelihood estimator will be when **heteroscedasticity** occurs. Not surprisingly, the degree of censoring is the primary determinant. Unfortunately, all the analyses have been carried out in the setting of very specific models—for example, involving only a single dummy variable or one with groupwise heteroscedasticity—so the primary lesson is the very general conclusion that heteroscedasticity emerges as a potentially serious problem.

One can approach the heteroscedasticity problem directly. Petersen and Waldman (1981) present the computations needed to estimate a tobit model with heteroscedasticity of several types. Replacing σ with σ_i in the log-likelihood function and including σ_i^2 in the summations produces the needed generality. Specification of a particular model for σ_i provides the empirical model for estimation.

Example 19.7 Multiplicative Heteroscedasticity in the Tobit Model

Petersen and Waldman (1981) analyzed the volume of short interest in a cross section of common stocks. The regressors included a measure of the market component of heterogeneous expectations as measured by the firm's *BETA* coefficient; a company-specific measure of heterogeneous expectations, *NONMARKET*; the *NUMBER* of analysts making earnings forecasts for the company; the number of common shares to be issued for the acquisition of another firm, *MERGER*; and a dummy variable for the existence of *OPTION*s. They report the results listed in Table 19.4 for a model in which the variance is assumed to be of the form $\sigma_i^2 = \exp(\mathbf{x}_i'\boldsymbol{\alpha})$. The values in parentheses are the ratio of the coefficient to the estimated asymptotic standard error.

The effect of heteroscedasticity on the estimates is extremely large. We do note, however, a common misconception in the literature. The change in the coefficients is often misleading. The average partial effects in the heteroscedasticity model will generally be very similar to those computed from the model which assumes homoscedasticity. (The calculation is pursued in the exercises.)

A test of the hypothesis that $\boldsymbol{\alpha} = \mathbf{0}$ (except for the constant term) can be based on the likelihood ratio statistic. For these results, the statistic is $-2[-547.30 - (-466.27)] = 162.06$. This statistic has a limiting chi-squared distribution with five degrees of freedom. The sample value exceeds the critical value in the table of 11.07, so the hypothesis can be rejected.

In the preceding example, we carried out a likelihood ratio test against the hypothesis of homoscedasticity. It would be desirable to be able to carry out the test without having to estimate the unrestricted model. A **Lagrange multiplier test** can be used for that purpose. Consider the heteroscedastic tobit model in which we specify that

$$\sigma_i^2 = \sigma^2[\exp(\mathbf{w}_i'\boldsymbol{\alpha})]^2. \tag{19-18}$$

This model is a fairly general specification that includes many familiar ones as special cases. The null hypothesis of homoscedasticity is $\boldsymbol{\alpha} = \mathbf{0}$. (We used this specification in the probit model in Section 17.5.5 and in the linear regression model in Section 9.7.1) Using the BHHH estimator of the Hessian as usual, we can produce a Lagrange multiplier statistic as follows: Let $z_i = 1$ if y_i is positive and 0 otherwise and

TABLE 19.4 Estimates of a Tobit Model (Standard errors in parentheses)

	Homoscedastic	*Heteroscedastic*	
	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$
Constant	−18.28 (5.10)	−4.11 (3.28)	−0.47 (0.60)
Beta	10.97 (3.61)	2.22 (2.00)	1.20 (1.81)
Nonmarket	0.65 (7.41)	0.12 (1.90)	0.08 (7.55)
Number	0.75 (5.74)	0.33 (4.50)	0.15 (4.58)
Merger	0.50 (5.90)	0.24 (3.00)	0.06 (4.17)
Option	2.56 (1.51)	2.96 (2.99)	0.83 (1.70)
ln L	−547.30	−466.27	
Sample size	200	200	

$$a_i = z_i\left(\frac{\varepsilon_i}{\sigma^2}\right) + (1 - z_i)\left(\frac{(-1)\lambda_i}{\sigma}\right),$$

$$b_i = z_i\left(\frac{(\varepsilon_i^2/\sigma^2 - 1)}{2\sigma^2}\right) + (1 - z_i)\left(\frac{(\mathbf{x}_i'\boldsymbol{\beta})\lambda_i}{2\sigma^3}\right),$$

$$\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}, \qquad \lambda_i = \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}. \tag{19-19}$$

The data vector is $\mathbf{g}_i = [a_i\mathbf{x}_i', b_i, b_i\mathbf{w}_i']'$. The sums are taken over all observations, and all functions involving unknown parameters (ε_i, ϕ_i, Φ_i, $\mathbf{x}_i'\boldsymbol{\beta}$, σ, λ_i) are evaluated at the restricted (homoscedastic) maximum likelihood estimates. Then,

$$\text{LM} = \mathbf{i}'\mathbf{G}[\mathbf{G}'\mathbf{G}]^{-1}\mathbf{G}'\mathbf{i} = nR^2 \tag{19-20}$$

in the regression of a column of ones on the $K + 1 + P$ derivatives of the log-likelihood function for the model with multiplicative heteroscedasticity, evaluated at the estimates from the restricted model. (If there were no limit observations, then it would reduce to the Breusch–Pagan statistic discussed in Section 9.6.2.) Given the maximum likelihood estimates of the tobit model coefficients, it is quite simple to compute. The statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of variables in $\mathbf{w}_i$.

19.3.5.c Nonnormality

Nonnormality is an especially difficult problem in this setting. It has been shown that if the underlying disturbances are not normally distributed, then the estimator based on (19-13) may be inconsistent. Of course, the pertinent question is how the misspecification affects estimation of the quantities of interest, usually the partial effects. Here, the issue is less clear, as we saw for the binary choice models in Section 17.2.4. Research is ongoing both on alternative estimators and on methods for testing for this type of misspecification.[29]

One approach to the estimation is to use an alternative distribution. Kalbfleisch and Prentice (2002) present a unifying treatment that includes several distributions such as the exponential, lognormal, and Weibull. (Their primary focus is on survival analysis in a medical statistics setting, which is an interesting convergence of the techniques in very different disciplines.) Of course, assuming some other specific distribution does not necessarily solve the problem and may make it worse. A preferable alternative would be to devise an estimator that is robust to changes in the distribution. Powell's (1981, 1984) least absolute deviations (LAD) estimator offers some promise.[30] The main drawback to its use is its computational complexity. An extensive application of the LAD estimator is Melenberg and van Soest (1996). Although estimation in the nonnormal case is relatively difficult, testing for this failure of the model is worthwhile to assess the estimates obtained by the conventional methods. Among the tests that

[29]See Duncan (1986a,b), Goldberger (1983), Pagan and Vella (1989), Lee (1996), and Fernandez (1986).

[30]See Duncan (1986a,b) for a symposium on the subject and Amemiya (1984). Additional references are Newey, Powell, and Walker (1990), Lee (1996), and Robinson (1988).

have been developed are Hausman tests, Lagrange multiplier tests [Bera and Jarque (1981, 1982) and Bera, Jarque, and Lee (1982)], and **conditional moment tests** [Nelson (1981)].

19.3.6 PANEL DATA APPLICATIONS

Extension of the familiar panel data results to the tobit model parallel the probit model, with the attendant problems. The random effects or random parameters models discussed in Chapter 17 can be adapted to the censored regression model using simulation or quadrature. The same reservations with respect to the orthogonality of the effects and the regressors will apply here, as will the applicability of the Mundlak (1978) correction to accommodate it.

Much of the attention in the theoretical literature on panel data methods for the tobit model has been focused on fixed effects. The departure point would be the maximum likelihood estimator for the static fixed effects model,

$$y_{it}^* = \alpha_i + x_{it}'\beta + \varepsilon_{it}, \varepsilon_{it} \sim \mathrm{N}[0, \sigma^2],$$
$$y_{it} = \mathrm{Max}(0, y_{it}).$$

However, there are no firm theoretical results on the behavior of the MLE in this model. Intuition might suggest, based on the findings for the binary probit model, that the MLE would be biased in the same fashion, away from zero. Perhaps surprisingly, the results in Greene (2004a) persistently found that not to be the case in a variety of model specifications. Rather, the incidental parameters, such as it is, manifests in a downward bias in the estimator of σ, not an upward (or downward) bias in the MLE of $\boldsymbol{\beta}$. However, this is less surprising when the tobit estimator is juxtaposed with the MLE in the linear regression model with fixed effects. In that model, the MLE is the within-groups (LSDV) estimator, which is unbiased and consistent. But, the ML estimator of the disturbance variance in the linear regression model is $\mathbf{e}_{\mathrm{LSDV}}'\mathbf{e}_{\mathrm{LSDV}}/(nT)$, which is biased downward by a factor of $(T-1)/T$. [This is the result found in the original source on the incidental parameters problem, Neyman and Scott (1948).] So, what evidence there is suggests that unconditional estimation of the tobit model behaves essentially like that for the linear regression model. That does not settle the problem, however; if the evidence is correct, then it implies that although consistent estimation of $\boldsymbol{\beta}$ is possible, appropriate statistical inference is not. The bias in the estimation of σ shows up in the computation of partial effects.

There is no conditional estimator of $\boldsymbol{\beta}$ for the tobit (or truncated regression) model. First differencing or taking group mean deviations does not preserve the model. Because the latent variable is censored before observation, these transformations are not meaningful. Some progress has been made on theoretical, **semiparametric estimators** for this model. See, for example, Honoré and Kyriazidou (2000) for a survey. Much of the theoretical development has also been directed at dynamic models where the benign result of the previous paragraph (such as it is) is lost once again. Arellano (2001) contains some general results. Hahn and Kuersteiner (2004) have characterized the bias of the MLE and suggested methods of reducing the bias of the estimators in dynamic binary choice and censored regression models.

19.4 SAMPLE SELECTION AND INCIDENTAL TRUNCATION

The topic of sample selection, or **incidental truncation**, has been the subject of an enormous literature, both theoretical and applied.[31] This analysis combines both of the previous topics.

Example 19.8 Incidental Truncation

In the high-income survey discussed in Example 19.2, respondents were also included in the survey if their net worth, not including their homes, was at least $500,000. Suppose that the survey of incomes was based *only* on people whose net worth was at least $500,000. This selection is a form of truncation, but not quite the same as in Section 19.2. This selection criterion does not necessarily exclude individuals whose incomes might be quite low. Still, one would expect that individuals with higher than average high net worth would have higher than average incomes as well. Thus, the average income in this subpopulation would in all likelihood also be misleading as an indication of the income of the typical American. The data in such a survey would be nonrandomly selected or incidentally truncated.

Econometric studies of nonrandom sampling have analyzed the deleterious effects of sample selection on the properties of conventional estimators such as least squares; have produced a variety of alternative estimation techniques; and, in the process, have yielded a rich crop of empirical models. In some cases, the analysis has led to a reinterpretation of earlier results.

19.4.1 INCIDENTAL TRUNCATION IN A BIVARIATE DISTRIBUTION

Suppose that y and z have a bivariate distribution with correlation ρ. We are interested in the distribution, in particular, the mean of y given that z exceeds a particular value. Intuition suggests that if y and z are positively correlated, then the truncation of z should push the distribution and the mean of y to the right. As before, we are interested in (1) the form of the incidentally truncated distribution and (2) the mean and variance of the incidentally truncated random variable. We will develop some generalities, and then, because it has dominated the empirical literature, focus on the bivariate normal distribution.

The truncated *joint* density of y and z is

$$f(y, z \mid z > a) = \frac{f(y, z)}{\text{Prob}(z > a)}.$$

To obtain the incidentally truncated marginal density for y, we would then integrate z out of this expression. The moments of the incidentally truncated normal distribution are given in Theorem 19.5.[32]

[31] A large proportion of the analysis in this framework has been in the area of labor economics. See, for example, Vella (1998), which is an extensive survey for practitioners. The results, however, have been applied in many other fields, including, for example, long series of stock market returns by financial economists ("survivorship bias") and medical treatment and response in long-term studies by clinical researchers ("attrition bias"). Some studies that comment on methodological issues are Barnow, Cain, and Goldberger (1981), Heckman (1990), Manski (1989, 1990, 1992), Newey, Powell, and Walker (1990), and Wooldridge (1995).

[32] More general forms of the result that apply to multivariate distributions are given in Kotz, Balakrishnan, and Johnson (2000).

THEOREM 19.5 Moments of the Incidentally Truncated Bivariate Normal Distribution

If y and z have a bivariate normal distribution with means μ_y *and* μ_z, standard deviations σ_y *and* σ_z, *and correlation* ρ, *then*

$$E[y \mid z > a] = \mu_y + \rho\sigma_y\lambda(\alpha_z),$$
$$\text{Var}[y \mid z > a] = \sigma_y^2[1 - \rho^2\delta(\alpha_z)],$$

where

$$\alpha_z = (a - \mu_z)/\sigma_z,\ \lambda(\alpha_z) = \phi(\alpha_z)/[1 - \Phi(\alpha_z)],\ \text{and}\ \delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z].$$

Note that the expressions involving z are analogous to the moments of the truncated distribution of x given in Theorem 19.2. If the truncation is $z < a$, then we make the replacement $\lambda(\alpha_z) = -\phi(\alpha_z)/\Phi(\alpha_z)$. It is clear that if ρ is positive, then $E[y \mid z > a] > E[y]$ as σ_y and $\lambda(\alpha_z)$ are both positive. As expected, the truncated mean is pushed in the direction of the correlation if the truncation is from below and in the opposite direction if it is from above. In addition, the incidental truncation reduces the variance, because both $\delta(\alpha)$ and ρ^2 are between zero and one. This second result is less obvious, but essentially it follows from the general principle that conditioning reduces variance.

19.4.2 REGRESSION IN A MODEL OF SELECTION

To motivate a regression model that corresponds to the results in Theorem 19.5, we consider the following example.

Example 19.9 A Model of Labor Supply

A simple model of female labor supply consists of two equations:[33]

1. Wage equation. The difference between a person's *market wage,* what she could command in the labor market, and her *reservation wage,* the wage rate necessary to make her choose to participate in the labor market, is a function of characteristics such as age and education as well as, for example, number of children and where a person lives.
2. Hours equation. The desired number of labor hours supplied depends on the wage, home characteristics such as whether there are small children present, marital status, and so on.

The problem of truncation surfaces when we consider that the second equation describes desired hours, but an actual figure is observed only if the individual is working. (In most such studies, only a *participation equation,* that is, whether hours are positive or zero, is observable.) We infer from this that the market wage exceeds the reservation wage. Thus, the hours variable in the second equation is incidentally truncated based on (offered wage – reservation wage).

To put the preceding examples in a general framework, let the equation that determines the sample selection be

$$z^* = \mathbf{w}'\boldsymbol{\gamma} + u,$$

[33]See, for example, Heckman (1976). This strand of literature begins with an exchange by Gronau (1974) and Lewis (1974).

where $\mathbf{w}$ is exogenous (mean independent of u and ε), and let the equation of primary interest be

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

The sampling rule is that y is observed only when z^* is greater than zero. The conditional mean that applies to the "selected" observations is

$$\begin{aligned} E[y|\mathbf{x},z^* > 0] &= \mathbf{x}'\boldsymbol{\beta} + E[\varepsilon|\mathbf{w}'\boldsymbol{\gamma} + u > 0] \\ &= \mathbf{x}'\boldsymbol{\beta} + E[\varepsilon|u > -\mathbf{w}'\boldsymbol{\gamma}]. \end{aligned}$$

If ε and u are uncorrelated, then $E[\varepsilon|u > a] = E[\varepsilon] = 0$. But, if they are correlated, then we would expect $E[\varepsilon|u > a]$ to be a function of a, as in Theorem 9.5 for the bivariate normal distribution. For this case, if $E[\varepsilon|u > -\mathbf{w}'\boldsymbol{\gamma}] = h(\mathbf{w}'\boldsymbol{\gamma},\sigma_u)$, then the relevant regression is

$$E[y|\mathbf{x},z^* > 0] = \mathbf{x}'\boldsymbol{\beta} + h(\mathbf{w}'\boldsymbol{\gamma}, \rho, \sigma_u, \sigma_\varepsilon) + v,$$

where $E[v|\mathbf{x}, h(\mathbf{w}'\boldsymbol{\gamma}, \rho, \sigma_u, \sigma_\varepsilon)] = 0$. The immediate implication is that simple linear regression of y on $\mathbf{x}$ will not consistently estimate $\boldsymbol{\beta}$ because of the omitted variable(s) contained in $h(\mathbf{w}'\boldsymbol{\gamma}, \rho, \sigma_u, \sigma_\varepsilon)$. In our example, the suggested wage equation contains age and education. But, conditioned on positive hours, which are being determined by children and marital status, the simple regression conditioned on positive hours is missing (a function of) these two variables. This omitted variables bias of least squares in this setting is "selection bias."[34]

In order to progress from this point, it will be necessary to be more specific about the omitted term in the tainted regression. If ε and u have a bivariate normal distribution with zero means and correlation ρ, then we may insert these in Theorem 19.5 to obtain the specific model *that applies to the observations in the selected sample:*

$$\begin{aligned} E[y|y \text{ is observed}] &= E[y|z^* > 0] \\ &= E[y|u > -\mathbf{w}'\boldsymbol{\gamma}] \\ &= \mathbf{x}'\boldsymbol{\beta} + E[\varepsilon|u > -\mathbf{w}'\boldsymbol{\gamma}] \\ &= \mathbf{x}'\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda(\alpha_u) \\ &= \mathbf{x}'\boldsymbol{\beta} + \beta_\lambda\lambda(\alpha_u), \end{aligned}$$

where $\alpha_u = (-\mathbf{w}'\boldsymbol{\gamma} - 0)/\sigma_u$ and

$$\lambda(\alpha_u) = \phi(-\mathbf{w}'\boldsymbol{\gamma}/\sigma_u)/[1 - \Phi(-\mathbf{w}'\boldsymbol{\gamma}/\sigma_u)] = \phi(\mathbf{w}'\boldsymbol{\gamma}/\sigma_u)/\Phi(\mathbf{w}'\boldsymbol{\gamma}/\sigma_u). \quad \textbf{(19-21)}$$

[We have used the symmetry of the normal distribution, $\phi(-b) = \phi(b)$ and $(1 - \Phi(-b)) = \Phi(b)$.] So,

$$\begin{aligned} y|z^* > 0 &= E[y|z^* > 0] + v_i \\ &= \mathbf{x}'\boldsymbol{\beta} + \beta_\lambda\lambda(\mathbf{w}'\boldsymbol{\gamma}/\sigma_u) + v. \end{aligned}$$

Least squares regression using the observed data—for instance, OLS regression of hours on its determinants, using only data for women who are working—produces inconsistent

[34] Any number of commentators have suggested in a given context that "the data are subject to selection bias." As we can see in the preceding, it is the OLS estimator, not the data, that is biased. We will find shortly that under suitable assumptions, there is a different estimator, based on the same data, that is not biased. (Or at least is consistent.)

estimates of $\boldsymbol{\beta}$. Once again, we can view the problem as an omitted variable. Least squares regression of y on $\mathbf{x}$ *and* λ would be a consistent estimator, but if λ is omitted, then the **specification error** of an omitted variable is committed. Finally, note that the second part of Theorem 19.5 implies that even if $\lambda(\mathbf{w}'\boldsymbol{\gamma}/\sigma_u)$ were observed and included in the regression, then least squares would be inefficient. The disturbance, v, is heteroscedastic.

The partial effect of the regressors on $E[y|\mathbf{x}, \mathbf{w}]$ *in the observed sample* consists of two components. There is the direct effect on the mean of y, which is $\boldsymbol{\beta}$. In addition, for a particular independent variable, if it appears in the probability that z^* is positive, then it will influence $E[y|\mathbf{x}, \mathbf{w}]$ through its presence in $\lambda(\mathbf{w}'\boldsymbol{\gamma}/\sigma_u)$. The full effect of changes in a regressor that appears in both $\mathbf{x}_i$ and $\mathbf{w}_i$ on y is

$$\frac{\partial E[y|z^* > 0]}{\partial x_k} = \beta_k - \gamma_k \left\{ \left(\frac{\rho \sigma_\varepsilon}{\sigma_u} \right) [\delta(\mathbf{w}'\boldsymbol{\gamma}/\sigma_u)] \right\},$$
$$\delta(\mathbf{w}'\boldsymbol{\gamma}/\sigma_u) = \lambda^2 - \alpha\, \lambda,$$

where λ is defined in (19-21) and $\alpha = -\mathbf{w}'\boldsymbol{\gamma}/\sigma_u$. Suppose that ρ is positive and $E[y|\mathbf{x}, \mathbf{w}]$ is greater when z^* is positive than when it is negative. Because $0 < \delta < 1$, the additional term serves to reduce the partial effect. The change in the probability affects the mean of y in that the mean in the group $z^* > 0$ is higher. The second term in the derivative compensates for this effect, leaving only the partial effect of a change *given that* $z_i^* > 0$ *to begin with*. Consider Example 19.11, and suppose that education affects both the probability of migration and the income in either state. If we suppose that the income of migrants is higher than that of otherwise identical people who do not migrate, then the partial effect of education has two parts, one due to its influence in increasing the probability of the individual's entering a higher-income group and one due to its influence on income within the group. As such, the coefficient on education in the regression overstates the partial effect of the education of migrants and understates it for nonmigrants. The sizes of the various parts depend on the setting. It is quite possible that the magnitude, sign, and statistical significance of the effect might all be different from those of the estimate of $\boldsymbol{\beta}$, a point that appears frequently to be overlooked in empirical studies.

In most cases, the selection variable z^* is not observed. Rather, we observe only its sign. To consider our two examples, the observation might be only whether someone is working outside the home or whether an individual migrated or not. We can infer the sign of z^*, but not its magnitude, from such information. Because there is no information on the scale of z^*, the disturbance variance in the selection equation cannot be estimated. (We encountered this problem in Chapter 17 in connection with the binary choice models.) Thus (retaining the joint normality assumption), we reformulate the model as follows:

$$\begin{aligned} \text{Selection mechanism: } z^* &= \mathbf{w}'\boldsymbol{\gamma} + u,\ z = \mathbf{1}(z^* > 0), \\ \text{Prob}(z = 1|\mathbf{w}) &= \Phi(\mathbf{w}'\boldsymbol{\gamma}), \\ \text{Prob}(z = 0|\mathbf{w}) &= 1 - \Phi(\mathbf{w}'\boldsymbol{\gamma}). \end{aligned} \qquad \textbf{(19-22)}$$

$$\text{Regression model} \qquad y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon \text{ observed only if } z = 1,$$
$$(\varepsilon, u) \sim \text{bivariate normal}[0, 0, \sigma_\varepsilon, 1, \rho].^{35}$$

[35] In some treatments of this model, it is noted (invoking Occam's razor) that this specification actually relies only on the marginal normality of u and $E[\varepsilon|u > -\mathbf{w}'\boldsymbol{\gamma}] = \theta\, \lambda(\mathbf{w}'\boldsymbol{\gamma})$. Neither joint normality nor even marginal normality of ε is essential. It is debatable how narrow the bivariate normality assumption is given the very specific assumption made about $E[\varepsilon|u > -\mathbf{w}'\boldsymbol{\gamma}]$.

Suppose that z_i and $\mathbf{w}_i$ are observed for a random sample of individuals but y_i is observed only when $z_i = 1$. This model is precisely the one we examined earlier, with

$$E[y_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda(\mathbf{w}_i'\boldsymbol{\gamma}).$$

19.4.3 TWO-STEP AND MAXIMUM LIKELIHOOD ESTIMATION

The parameters of the sample selection model in (19-22) can be estimated by maximum likelihood.[36] However, Heckman's (1979) **two-step estimation** procedure is usually used instead. Heckman's method is as follows:[37]

1. Estimate the probit equation by maximum likelihood to obtain estimates of $\boldsymbol{\gamma}$. For each observation in the selected sample, compute $\hat{\lambda}_i = \phi(\mathbf{w}_i'\hat{\boldsymbol{\gamma}})/\Phi(\mathbf{w}_i'\hat{\boldsymbol{\gamma}})$ and $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \mathbf{w}_i'\hat{\boldsymbol{\gamma}})$.
2. Estimate $\boldsymbol{\beta}$ and $\beta_\lambda = \rho\sigma_\varepsilon$ by least squares regression of y on $\mathbf{x}$ and $\hat{\lambda}$.[38]

It is possible also to construct consistent estimators of the individual parameters ρ and σ_ε (again, assuming bivariate normality). At each observation, the true conditional variance of the disturbance would be

$$\sigma_i^2 = \sigma_\varepsilon^2(1 - \rho^2\delta_i).$$

The average conditional variance for the sample would converge to

$$\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 = \sigma_\varepsilon^2(1 - \rho^2\bar{\delta}),$$

which is what is estimated by the least squares residual variance $\mathbf{e}'\mathbf{e}/n$. For the square of the coefficient on $\hat{\lambda}$, we have

$$\text{plim}\, b_\lambda^2 = \rho^2\sigma_\varepsilon^2,$$

whereas based on the probit results we have

$$\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_i = \bar{\delta}.$$

We can then obtain a consistent estimator of σ_ε^2 using

$$\hat{\sigma}_\varepsilon^2 = \frac{\mathbf{e}'\mathbf{e}}{n} + \hat{\bar{\delta}}b_\lambda^2.$$

Finally, an estimator of ρ^2 is

$$\hat{\rho} = \text{sgn}(b_\lambda)\sqrt{\frac{b_\lambda^2}{\hat{\sigma}_\varepsilon^2}}, \qquad \textbf{(19-23)}$$

[36] See Greene (1995a). Note in view of footnote 39, the MLE is only appropriate under the specific assumption of bivariate normality. Absent that assumption, we would use two-step least squares or instrumental variables in all cases. See, also, Newey (1991) for details.

[37] Perhaps in a mimicry of the "tobit" estimator described earlier, this procedure has come to be known as the "Heckit" estimator.

[38] As a modest first step, it is possible to "test for selectivity," that is, the null hypothesis that β_λ equals zero, simply by regressing y on $(\mathbf{x}, \hat{\lambda})$ using the selected sample and using a conventional t test based on the usual estimator of the asymptotic covariance matrix. Under the null hypothesis, this amounts to an ordinary test of the specification of the linear regression. Upon rejection of the null hypothesis, one would proceed with the additional calculations about to be suggested.

which provides a complete set of estimators of the model's parameters.[39]

To test hypotheses, an estimator of the asymptotic covariance matrix of $[\mathbf{b}', b_\lambda]'$ is needed. We have two problems to contend with. First, we can see in Theorem 19.5 that the disturbance term in

$$(y_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda_i + v_i \qquad \textbf{(19-24)}$$

is heteroscedastic;

$$\text{Var}[v_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \sigma_\varepsilon^2(1 - \rho^2\delta_i).$$

Second, there are estimated parameters in $\hat{\lambda}_i$. Suppose that we assume for the moment that λ_i and δ_i are known (i.e., we do not have to estimate $\boldsymbol{\gamma}$). For convenience, let $\mathbf{x}_i^* = [\mathbf{x}_i, \lambda_i]$, and let $\mathbf{b}^*$ be the least squares coefficient vector in the regression of y on $\mathbf{x}^*$ in the selected data. Then, using the appropriate form of the variance of OLS in a heteroscedastic model from Chapter 9, we would have to estimate

$$\begin{aligned}\text{Var}[\mathbf{b}^* | \mathbf{X}_*] &= \sigma_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}\left[\sum_{i=1}^{n}(1 - \rho^2\delta_i)\mathbf{x}_i^*\mathbf{x}_i^{*\prime}\right][\mathbf{X}_*'\mathbf{X}_*]^{-1} \\ &= \sigma_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I} - \rho^2\boldsymbol{\Delta})\mathbf{X}_*][\mathbf{X}_*'\mathbf{X}_*]^{-1},\end{aligned}$$

where $\mathbf{I} - \rho^2\Delta$ is a diagonal matrix with $(1 - \rho^2\delta_i)$ on the diagonal. (Note that this collapses to the simple least squares result if ρ equals zero, which motivates the test in Footnote 38.) Without any other complications, this result could be computed fairly easily using $\mathbf{X}$, the sample estimates of σ_ε^2 and ρ^2, and the assumed known values of λ_i and δ_i.

The parameters in $\boldsymbol{\gamma}$ do have to be estimated using the probit equation. Rewrite (19-24) as

$$(y_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \beta_\lambda\hat{\lambda}_i + v_i - \beta_\lambda(\hat{\lambda}_i - \lambda_i).$$

In this form, we see that in the preceding expression we have ignored both an additional source of variation in the compound disturbance and correlation across observations; the same estimate of $\boldsymbol{\gamma}$ is used to compute $\hat{\lambda}_i$ for every observation. Heckman has shown that the earlier covariance matrix can be appropriately corrected by adding a term inside the brackets,

$$\mathbf{Q} = \hat{\rho}^2(\mathbf{X}_*'\hat{\boldsymbol{\Delta}}\mathbf{W})\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}](\mathbf{W}'\hat{\boldsymbol{\Delta}}\mathbf{X}_*) = \hat{\rho}^2\hat{\mathbf{F}}\hat{\mathbf{V}}\hat{\mathbf{F}}',$$

where $\hat{\mathbf{V}} = \text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}]$, the estimator of the asymptotic covariance of the probit coefficients. Any of the estimators in (17-22) to (17-24) may be used to compute $\hat{\mathbf{V}}$. The complete expression is

$$\text{Est.Asy.Var}[\mathbf{b}, b_\lambda] = \hat{\sigma}_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I} - \hat{\rho}^2\hat{\Delta})\mathbf{X}_* + \mathbf{Q}][\mathbf{X}_*'\mathbf{X}_*]^{-1}.$$

This is the estimator that is embedded in contemporary software such as *Stata*. We note three useful further aspects of the two-step estimator:

1. This is an application of the two-step procedures we developed in Section 8.4.1 and 14.7 and that were formalized by Murphy and Topel (1985).[40]

[39]The proposed estimator is suggested in Heckman (1979). Note that $\hat{\rho}$ is not a sample correlation and, as such, is not limited to [0, 1]. See Greene (1981) for discussion.

[40]This matrix formulation is derived in Greene (1981). Note that the Murphy and Topel (2002) results for two-step estimators given in Theorem 14.8 would apply here as well. Asymptotically, this method would give the same answer. The Heckman formulation has become standard.

2. The two-step estimator is a control function estimator. See Section 14.7. The control function is a generalized residual [see Chesher, Lancaster, and Irish (1985)] so this estimator is an example for Terza, Basu, and Rathouz's (2008) residual inclusion method.

3. Absent the bivariate normality assumption (or with it), the appropriate asymptotic covariance matrix could be estimated by bootstrapping the two-step least squares estimator.

The sample selection model can also be estimated by maximum likelihood. The full log-likelihood function for the data is built up from the two components for the observed data:

$$\text{Prob(selection)} \times \text{density}|\text{selection for observations with } z_i = 1,$$

and

$$\text{Prob(nonselection) for observations with } z_i = 0.$$

Combining the parts produces the full log-likelihood function,

$$\ln L = \sum_{z=1} \ln\left[\frac{\exp(-(1/2)\varepsilon_i^2/\sigma_\varepsilon^2)}{\sigma_\varepsilon\sqrt{2\pi}}\Phi\left(\frac{\rho\varepsilon_i/\sigma_\varepsilon + \mathbf{w}_i'\boldsymbol{\gamma}}{\sqrt{1-\rho^2}}\right)\right] + \sum_{z=0} \ln\left[1 - \Phi\left(\mathbf{w}_i'\boldsymbol{\gamma}\right)\right], \quad \textbf{(19-25)}$$

where $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$.[41]

Two virtues of the FIML estimator will be the greater efficiency brought by using the full likelihood function rather than the method of moments and, second, the estimation of ρ subject to the constraint $-1 < \rho < 1$. (This is typically done by reparameterizing the model in terms of the monotonic inverse hyperbolic tangent, $\tau = (1/2)\ln[(1+\rho)/(1-\rho)] = \text{atanh}(\rho)$. The transformed parameter, τ, is unrestricted. The inverse transformation is $\rho = [\exp(2\tau) - 1]/[\exp(2\tau) + 1]$ which is bounded between zero and one.) One possible drawback (it might be argued) could be the complexity of the likelihood function that would make estimation more difficult than the two-step estimator. However, the MLE for the selection model appears as a built-in procedure in modern software such as *Stata* and *NLOGIT*, and it is straightforward to implement in *Gauss*, so this might be a moot point. Surprisingly, the MLE is by far less common than the two-step estimator in the received applications. The estimation of ρ is the difficult part of the estimation process (this is often the case). It is quite common for the method of moments estimator and the FIML estimator to be very different—our application in Example 19.10 is such a case. Perhaps surprisingly so, the moment-based estimator of ρ in (19-23) is not bounded by zero and one.[42] This would seem to recommend the MLE.

The fully parametric bivariate normality assumption of the model has been viewed as a potential drawback. However, relatively little progress has been made on devising informative semi- and nonparametric estimators—see, for one example, Gallant and

[41]Note the FIML estimator does suggest an alternative two-step estimator. Because $\boldsymbol{\gamma}$ can be estimated by the probit estimator at a first step, the index $\mathbf{w}_i'\hat{\boldsymbol{\gamma}}$ can be inserted into the log likelihood, then $\boldsymbol{\beta}$, ρ, and σ_ε can be estimated by maximizing only the first sum (over the selected observations). The usefulness of this third approach remains to be investigated.

[42]See Greene (1981).

Nychka (1987). The obstacle here is that, ultimately, the model hangs on a parameterization of the correlation of the unobservables in the two equations. So, method of moment estimators, IV or kernel-based estimators must still incorporate some feature of a bivariate distribution. Some results have been obtained using the method of copula functions.[43] Martins (2001) considers other semiparametric approaches.

Example 19.10 Female Labor Supply

Example 17.15 proposed a labor force participation model for a sample of 753 married women in a sample analyzed by Mroz (1987). The data set contains wage and hours information for the 428 women who participated in the formal market (LFP = 1). Following Mroz, we suppose that for these 428 individuals, the offered wage exceeded the reservation wage and, moreover, the unobserved effects in the two wage equations are correlated. As such, a wage equation based on the market data should account for the sample selection problem. We specify a simple wage model:

$$\ln Wage = \beta_1 + \beta_2 Exper + \beta_3 Exper^2 + \beta_4 Education + \varepsilon^{44},$$

where *Exper* is labor market experience. Maximum likelihood, Heckman two-step, and OLS estimates of the wage equation are shown in Table 19.5. The maximum likelihood estimates are FIML estimates—the labor force participation equation is reestimated at the same time. Only the parameters of the wage equation are shown next. Note as well that the two-step estimator estimates the single coefficient on λ_i and the structural parameters σ and ρ are deduced by the method of moments. The maximum likelihood estimator computes estimates of these parameters directly.

The two-step and ML estimators both provide a direct test of "selectivity," that is, $\rho = 0$. In both cases, the estimate is small and the standard error is large. Both tests fail to reject the hypothesis. The correction to the standard errors in the two-step estimator is also minor. This is to be expected. Both terms in the adjustment involve ρ^2, which is small here—the unadjusted OLS standard error for the two-step estimator is essentially the same, 0.13439. (It does not follow algebraically that the adjustment will increase the estimated standard errors.) Because there is scant impact of the sample selection correction in this model, the OLS estimates will provide reasonable values for the average partial effects. The OLS education coefficient, in particular, is 0.107. The average partial effects from the two-step and ML results are 0.10668 and 0.10821, respectively.

TABLE 19.5 Estimated Selection Corrected Wage Equation

	Two-Step		*Maximum Likelihood*		*Least Squares*	
	Estimate	***Std. Error***	***Estimate***	***Std. Error***	***Estimate***	***Std. Error***
Constant	−0.57810	(0.30502)	−0.55270	(0.32495)	−0.52204	(0.19863)
Experience	0.04389	(0.01626)	0.04284	(0.01729)	0.04157	(0.01318)
$Experience^2$	−0.00086	(0.00044)	−0.00084	(0.00048)	−0.00081	(0.00039)
Education	0.10907	(0.01552)	0.10835	(0.01674)	0.10749	(0.01415)
$(\rho\sigma)$	0.03226	(0.13362)				
ρ	0.04861		0.02661	(0.18227)	0.00000	
σ	0.66363		0.66340	(0.01445)	0.66642	

[43]See Smith (2003, 2005) and Trivedi and Zimmer (2007). Extensive commentary appears in Wooldridge (2010).

[44]As in Example 17.15, for comparison purposes, we have replicated the specification in Wooldridge (2010, p. 807).

Example 19.11 A Mover-Stayer Model for Migration

The model of migration analyzed by Nakosteen and Zimmer (1980) fits into the framework described in this section. The equations of the model are

$$\begin{aligned} &\text{Net benefit of moving:} && M^* = \mathbf{w}'\boldsymbol{\gamma} + u, \\ &\text{Income if moves:} && I_1 = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \\ &\text{Income if stays:} && I_0 = \mathbf{x}_0'\boldsymbol{\beta}_0 + \varepsilon_0. \end{aligned}$$

One component of the net benefit is the market wage individuals could achieve if they move, compared with what they could obtain if they stay. Therefore, among the determinants of the net benefit are factors that also affect the income received in either place. An analysis of income in a sample of migrants must account for the incidental truncation of the mover's income on a positive net benefit. Likewise, the income of the stayer is incidentally truncated on a nonpositive net benefit. The model implies an income after moving for all observations, but we observe it only for those who actually do move. Nakosteen and Zimmer (1980) applied the selectivity model to a sample of 9,223 individuals with data for two years (1971 and 1973) sampled from the Social Security Administration's Continuous Work History Sample. Over the period, 1,078 individuals migrated and the remaining 8,145 did not. The independent variables in the migration equation were as follows:

$$\begin{aligned} SE &= \text{self-employment dummy variable; 1 if yes,} \\ \Delta EMP &= \text{rate of growth of state employment,} \\ \Delta PCI &= \text{growth of per capita income} \\ \mathbf{x} &= \text{age, race (nonwhite = 1), sex (female = 1),} \\ \Delta SIC &= \text{1 if individual changes industry.} \end{aligned}$$

The earnings equations included ΔSIC and SE. The authors reported the results in Table 19.6.

19.4.4 SAMPLE SELECTION IN NONLINEAR MODELS

The preceding analysis has focused on an extension of the linear regression (or the estimation of simple averages of the data). The method of analysis changes in nonlinear models. To begin, it is not necessarily obvious what the impact of the sample selection is on the response variable, or how it can be accommodated in a model. Consider the model analyzed by Boyes, Hoffman, and Low (1989):

$$\begin{aligned} y_{i1} &= 1 \text{ if individual } i \text{ defaults on a loan, 0 otherwise,} \\ y_{i2} &= 1 \text{ if the individual is granted a loan, 0 otherwise.} \end{aligned}$$

TABLE 19.6 Estimated Earnings Equations (Asymptotic *t* ratios in parentheses)

	Migration	*Migrant Earnings*	*Nonmigrant Earnings*
Constant	−1.509	9.041	8.593
SE	−0.708 (−5.72)	−4.104 (−9.54)	−4.161 (−57.71)
ΔEMP	−1.488 (−2.60)	—	—
ΔPCI	1.455 (3.14)	—	—
Age	−0.008 (−5.29)	—	—
Race	−0.065 (−1.17)	—	—
Sex	−0.082 (−2.14)	—	—
ΔSIC	0.948 (24.15)	−0.790 (−2.24)	−0.927 (−9.35)
λ	—	0.212 (0.50)	0.863 (2.84)

Wynand and van Praag (1981) also used this framework to analyze consumer insurance purchases in the first application of the selection methodology in a nonlinear model. Greene (1992) applied the same model to y_1 = default on credit card loans, in which y_{i2} denotes whether an application for the card was accepted or not. [Mohanty (2002) also used this model to analyze teen employment in California.] For a given individual, y_1 is not observed unless $y_2 = 1$. Following the lead of the linear regression case in Section 19.4.3, a natural approach might seem to be to fit the second (selection) equation using a univariate probit model, compute the inverse Mills ratio, λ_i, and add it to the first equation as an additional "control" variable to accommodate the selection effect. [This is the approach used by Wynand and van Praag (1981) and Greene (1994).] The problems with this control function approach are, first, it is unclear what in the model is being "controlled" and, second, assuming the first model is correct, the appropriate model conditioned on the sample selection is unlikely to contain an inverse Mills ratio anywhere in it. [See Terza (2009) for discussion.] That result is specific to the linear model, where it arises as $E[\varepsilon_i|\text{selection}]$. What would seem to be the apparent counterpart for this probit model,

$$\text{Prob}(y_{i1} = 1|\text{selection on } y_{i2} = 1) = \Phi(\mathbf{x}_{i1}'\boldsymbol{\beta}_1 + \theta\,\lambda_i),$$

is not, in fact, the appropriate probability.[45] For this particular application, the appropriate conditional probability (extending the bivariate probit model of Section 17.9) would be

$$\text{Prob}[y_{i1} = 1|y_{i2} = 1] = \frac{\Phi_2(\mathbf{x}_{i1}'\boldsymbol{\beta}_1, \mathbf{x}_{i2}'\boldsymbol{\beta}_2, \rho)}{\Phi(\mathbf{x}_{i2}'\boldsymbol{\beta}_2)}.$$

We would use this result to build up the likelihood function for the three observed outcomes, as follows: The three types of observations in the sample, with their unconditional probabilities, are

$$\begin{aligned}
y_{i2} = 0\text{: Prob}(y_{i2} = 0 \qquad\qquad |\mathbf{x}_{i1}, \mathbf{x}_{i2}) &= 1 - \Phi(\mathbf{x}_{i2}'\boldsymbol{\beta}_2),\\
y_{i1} = 0, y_{i2} = 1\text{: Prob}(y_{i1} = 0, y_{i2} = 1)|\mathbf{x}_{i1}, \mathbf{x}_{i2}) &= \Phi_2(-\mathbf{x}_{i1}'\boldsymbol{\beta}_1, \mathbf{x}_{i2}'\boldsymbol{\beta}_2, -\rho), \qquad \textbf{(19-26)}\\
y_{i1} = 1, y_{i2} = 1\text{: Prob}(y_{i1} = 1, y_{i2} = 1|\mathbf{x}_{i1}, \mathbf{x}_{i2}) &= \Phi_2(\mathbf{x}_{i1}'\boldsymbol{\beta}_1, \mathbf{x}_{i2}'\beta_2, \rho).
\end{aligned}$$

The log-likelihood function is based on these probabilities.[46] An application appears in Section 17.6.

Example 19.12 Doctor Visits and Insurance

Continuing our analysis of the utilization of the German health care system, we observe that the data set contains an indicator of whether the individual subscribes to the "Public" health insurance or not. Roughly 87% of the observations in the sample do. We might ask whether the selection on public insurance reveals any substantive difference in visits to the physician. We estimated a logit specification for this model in Example 17.19. Using (19-26) as the framework, we define y_{i2} to be presence of insurance and y_{i1} to be the binary variable defined to equal 1 if the individual makes at least one visit to the doctor in the survey year.

[45] As in the linear case, the augmented single-equation model does provide a valid framework for a simple t test of the null hypothesis that θ equals zero, but if the test rejects the null hypothesis, an altogether different approach is called for.

[46] Extensions of the bivariate probit model to other types of censoring are discussed in Poirier (1980) and Abowd and Farber (1982).

The estimation results are given in Table 19.7. Based on these results, there does appear to be a very strong relationship. The coefficients do change somewhat in the conditional model. A Wald test for the presence of the selection effect against the null hypothesis that ρ equals zero produces a test statistic of $(-7.188)^2 = 51.667$, which is larger than the critical value of 3.84. Thus, the hypothesis is rejected. A likelihood ratio statistic is computed as the difference between the log likelihood for the full model and the sum of the two separate log likelihoods for the independent probit models when ρ equals zero. The result is

$$\lambda_{LR} = 2[-23969.58 - (-15536.39 + (-8471.508)) = 77.796.$$

The hypothesis is rejected once again. Partial effects were computed using the results in Section 17.6.

The large correlation coefficient can be misleading. The estimated -0.9299 does not state that the presence of insurance makes it much less likely to go to the doctor. This is the correlation among the unobserved factors in each equation. The factors that make it more likely to purchase public insurance make it less likely to use a physician. (In this system, everyone has insurance. We are actully examining the difference between those who obtain the public insurance and those who obtain private insurance.) To obtain a simple correlation between the two variables, we might use the tetrachoric correlation defined in Example 17.31. This would be computed by fitting a bivariate probit model for the two binary variables without any other variables. The estimated value is 0.120.

More general cases are typically much less straightforward. Greene (2005, 2007d) and Terza (1998, 2010) present sample selection models for nonlinear specifications based on the underlying logic of the Heckman model in Section 19.4.3, that the influence of the incidental truncation acts on the unobservable variables in the model. (That is the source of the "selection bias" in conventional estimators.) The modeling extension introduces the unobservables into the model in a natural fashion that parallels the regression model. Terza (1985b, 2009) presents a survey of the general results.

TABLE 19.7 Estimated Probit Equations for Doctor Visits

	Independent: No Selection			*Sample Selection Model*		
Variable	*Estimate*	*Std. Error*	*Partial Effect*	*Estimate*	*Std. Error*	*Partial Effect*
Constant	0.05588	0.0656		−9.4366	0.0676	
Age	0.01331	0.0008	0.0050	0.0128	0.0008	0.0050
Income	−0.1034	0.0509	−0.0386	−0.1030	0.0458	−0.0406
Kids	−0.1349	0.0195	−0.0506	−0.1264	0.0179	−0.0498
Education	−0.0192	0.0043	−0.0072	0.0366	0.0047	0.0027
Married	0.03586	0.0217	0.0134	0.0356	0.0202	0.0140
ln L	−15,536.39					
Constant	3.3585	0.0700		3.2699	0.0692	
Age	0.0002	0.0010		−0.0003	0.0010	
Education	−0.1854	0.0039		−0.1807	0.0039	
Female	0.1150	0.0219	0.0000[a]	0.2230	0.0210	0.0145[a]
ln L	−8,471.51					
ρ	0.0000	0.0000		−0.9299	0.1294	
ln L	−24,007.91			−23,969.58		

[a]Indirect effect from second equation.

The generic model will take the form

1. Probit selection equation:

$$z^* = \mathbf{w}'\boldsymbol{\alpha} + u \text{ in which } u \sim \text{N}[0, 1],$$
$$z = \mathbf{1}(z^* > 0),$$

2. Nonlinear index function model with unobserved heterogeneity and sample selection:

$$\begin{aligned} &\mu|\varepsilon = \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon, \varepsilon \sim \text{N}[0, 1], \\ &y|\mathbf{x}, \varepsilon \sim \text{density } g(y|\mathbf{x}, \varepsilon) = f(y|\mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon), \\ &y, \mathbf{x} \text{ are observed only when } z = 1, \\ &[u, \varepsilon] \sim \text{N}[(0, 1), (1, \rho, 1)]. \end{aligned} \tag{19-27}$$

For example, in a Poisson regression model, the conditional mean function becomes $E(y|\mathbf{x}) = \lambda = \exp(\mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon) = \exp(\mu)$. (We used this specification of the model in Chapter 18 to introduce random effects in the Poisson regression model for panel data.)

The log-likelihood function for the full model is the joint density for the observed data. When z equals one, $(y, \mathbf{x}, z, \mathbf{w})$ are all observed. To obtain the joint density $p(y, z = 1|\mathbf{x}, \mathbf{w})$, we proceed as follows:

$$p(y, z = 1|\mathbf{x}, \mathbf{w}) = \int_{-\infty}^{\infty} p(y, z = 1|\mathbf{x}, \mathbf{w}, \varepsilon) f(\varepsilon) d\varepsilon.$$

Conditioned on ε, z and y are independent. Therefore, the joint density is the product,

$$p(y, z = 1|\mathbf{x}, \mathbf{w}, \varepsilon) = f(y|\mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon)\text{Prob}(z = 1|\mathbf{w}, \varepsilon).$$

The first part, $f(y|\mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon)$ is the conditional index function model in (19-27). By joint normality, $f(u|\varepsilon) = N[\rho\varepsilon, (1 - \rho^2)]$, so $u|\varepsilon = \rho\varepsilon + (u - \rho\varepsilon) = \rho\varepsilon + v_i$ where $E[v] = 0$ and $\text{Var}[v] = (1 - \rho^2)$. Therefore,

$$\text{Prob}(z = 1|\mathbf{w}, \varepsilon) = \Phi\left(\frac{\mathbf{w}'\boldsymbol{\alpha} + \rho\varepsilon}{\sqrt{1 - \rho^2}}\right).$$

Combining terms and using the earlier approach, the unconditional joint density is

$$p(y, z = 1|\mathbf{x}, \mathbf{w}) = \int_{-\infty}^{\infty} f(y|\mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon)\Phi\left(\frac{\mathbf{w}'\boldsymbol{\alpha} + \rho\varepsilon}{\sqrt{1 - \rho^2}}\right)\frac{\exp(-\varepsilon^2/2)}{\sqrt{2\pi}} d\varepsilon. \tag{19-28}$$

The other part of the likelihood function for the observations with $z_i = 0$ will be

$$\begin{aligned} \text{Prob}(z = 0|\mathbf{w}) &= \int_{-\infty}^{\infty} \text{Prob}(z = 0|\mathbf{w}, \varepsilon) f(\varepsilon) d\varepsilon. \\ &= \int_{-\infty}^{\infty} \left[1 - \Phi\left(\frac{\mathbf{w}'\boldsymbol{\alpha} + \rho\varepsilon}{\sqrt{1 - \rho^2}}\right)\right] f(\varepsilon) d\varepsilon \\ &= \int_{-\infty}^{\infty} \Phi\left(\frac{-(\mathbf{w}'\boldsymbol{\alpha} + \rho\varepsilon)}{\sqrt{1 - \rho^2}}\right)\frac{\exp(-\varepsilon^2/2)}{\sqrt{2\pi}} d\varepsilon. \end{aligned} \tag{19-29}$$

For convenience, we can use the invariance principle to reparameterize the log-likelihood function in terms of $\boldsymbol{\gamma} = \boldsymbol{\alpha}/\sqrt{1 - \rho^2}$ and $\tau = \rho/\sqrt{1 - \rho^2}$. Combining all the preceding terms, the log-likelihood function to be maximized is

$$\ln L = \sum_{i=1}^{n} \ln \int_{-\infty}^{\infty} [(1 - z_i) + z_i f(y_i | \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i)]\Phi[(2z_i - 1)(\mathbf{w}_i'\boldsymbol{\gamma} + \tau\varepsilon_i)]\phi(\varepsilon_i)d\varepsilon_i. \quad \textbf{(19-30)}$$

This can be maximized with respect to $(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}, \tau)$ using quadrature or simulation. When done, ρ can be recovered from $\rho = \tau/(1 + \tau^2)^{1/2}$ and $\boldsymbol{\alpha} = (1 - \rho^2)^{1/2}\boldsymbol{\gamma}$. All that differs from one model to another is the specification of $f(y_i | \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i)$. This is the specification used in Terza (1998) and Terza and Kenkel (2001). (In these two papers, the authors also analyzed $E[y | z = 1]$. This estimator was based on nonlinear least squares, but as earlier, it is necessary to integrate the unobserved heterogeneity out of the conditional mean function.) Greene (2010a) applies the method to a stochastic frontier model.

19.4.5 PANEL DATA APPLICATIONS OF SAMPLE SELECTION MODELS

The development of methods for extending sample selection models to panel data settings parallels the literature on cross-section methods. It begins with Hausman and Wise (1977, 1979) who devised a maximum likelihood estimator for a two-period model with attrition—the "selection equation" was a formal model for attrition from the sample. Subsequent research has drawn the analogy between attrition and sample selection in a variety of applications, such as Keane et al. (1988) and Verbeek and Nijman (1992), and produced theoretical developments including Wooldridge (1995, 2010, Section 19.9). We have noted some of these issues in Section 11.2.5.

The direct extension of panel data methods to sample selection brings several new issues for the modeler. An immediate question arises concerning the nature of the selection itself. Although much of the theoretical literature [For example, Kyriazidou (1997, 2001) and Honoré and Kyrizadou (1997, 2000)] treat the panel as if the selection mechanism is run anew in every period, in practice, the selection process often comes in two very different forms. First, selection may take the form of selection of the entire group of observations into the panel data set. Thus, the selection mechanism operates once, perhaps even before the observation window opens. Consider the entry (or not) of eligible candidates for a job training program. In this case, it is not appropriate to build the model to allow entry, exit, and then reentry. Second, for most applications, selection comes in the form of attrition or retention. Once an observation is "deselected," it does not return. Leading examples would include "survivorship" in time-series–cross-section models of firm performance and attrition in medical trials and in panel data applications involving large national survey data bases, such as Contoyannis et al. (2004). Each of these cases suggests the utility of a more structured approach to the selection mechanism.

19.4.5.a Common Effects in Sample Selection Models

A formal "effects" treatment for sample selection was first suggested in complete form by Verbeek (1990), who formulated a random effects model for the probit equation and a fixed effects approach for the main regression. Zabel (1992) criticized the specification for its asymmetry in the treatment of the effects in the two equations. He also argued that the likelihood function that neglected correlation between the effects and regressors in

the probit model would render the FIML estimator inconsistent. His proposal involved fixed effects in both equations. Recognizing the difficulty of fitting such a model, he then proposed using the Mundlak correction (Section 11.5.7). The full model is

$$\begin{aligned} y_{it}^{*} &= \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \qquad \eta_i = \bar{\mathbf{x}}_i'\boldsymbol{\pi} + \tau w_i, w_i \sim N[0,1], \\ d_{it}^{*} &= \theta_i + \mathbf{z}_{it}'\alpha + u_{it}, \qquad \theta_i = \mathbf{z}_i'\boldsymbol{\delta} + \omega v_i, v_i \sim N[0,1], \\ (\varepsilon_{it}, u_{it}) &\sim N_2[(0,0),(\sigma^2, 1, \rho\sigma)]. \end{aligned} \tag{19-31}$$

The selectivity in the model is carried through the correlation between ε_{it} and u_{it}. The resulting log likelihood is built up from the contribution of individual i,

$$\begin{aligned} L_i = &\int_{-\infty}^{\infty} \prod_{d_{it}=0} \Phi[-\mathbf{z}_{it}'\boldsymbol{\alpha} - \bar{\mathbf{z}}_i'\boldsymbol{\delta} - \omega v_i]\phi(v_i)dv_i \\ &\times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \prod_{d_{it}=1} \Phi\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha} + \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_i + (\rho/\sigma)\varepsilon_{it}}{\sqrt{1-\rho^2}}\right] \\ &\times \frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it}}{\sigma}\right)\phi_2(v_i, w_i)\, dv_i dw_i, \\ \varepsilon_{it} = &\, y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \bar{\mathbf{x}}_i'\boldsymbol{\pi} - \tau w_i. \end{aligned} \tag{19-32}$$

The log likelihood is then $\ln L = \sum_i \ln L_i$.

The log likelihood requires integration in two dimensions for any selected observations. Vella (1998) suggested two-step procedures to avoid the integration. However, the bivariate normal integration is actually the product of two univariate normals, because in the preceding specification, v_i and w_i are assumed to be uncorrelated. As such, the likelihood function in (19-32) can be readily evaluated using familiar simulation or quadrature techniques.[47] To show this, note that the first line in the log likelihood is of the form $E_v\left[\prod_{d=0}\Phi(\dots)\right]$ and the second line is of the form $E_w[E_v[\Phi(\dots)\phi(\dots)/\sigma]]$. Either of these expectations can be satisfactorily approximated with the average of a sufficient number of draws from the standard normal populations that generate w_i and v_i. The term in the simulated likelihood that follows this prescription is

$$\begin{aligned} L_i^S = &\frac{1}{R}\sum_{r=1}^{R}\prod_{d_{it}=0}[\Phi - \mathbf{z}_{it}'\boldsymbol{\alpha} - \mathbf{z}_i'\boldsymbol{\delta} - \omega v_{i,r}] \\ &\times \frac{1}{R}\sum_{r=1}^{R}\prod_{d_{it}=1}\Phi\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha} + \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_{i,r} + (\rho/\sigma)\varepsilon_{it,r}}{\sqrt{1-\rho^2}}\right]\frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it,r}}{\sigma}\right) \\ \varepsilon_{it,r} = &\, y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \bar{\mathbf{x}}_i'\boldsymbol{\pi} - \boldsymbol{\tau} w_{i,r}. \end{aligned} \tag{19-33}$$

Maximization of this log likelihood with respect to $(\boldsymbol{\beta}, \sigma, \rho, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\pi}, \tau, \omega)$ by conventional gradient methods is quite feasible. Indeed, this formulation provides a means by which the likely correlation between v_i and w_i can be accommodated in the model. Suppose

[47] See Sections 14.14.4 and 15.6.2.b. Vella and Verbeek (1999) suggest this in a footnote, but do not pursue it.

that w_i and v_i are bivariate standard normal with correlation ρ_{vw}. We can project w_i on v_i and write

$$w_i = \rho_{vw} v_i + \left(1 - \rho_{vw}^2\right)^{1/2} h_i,$$

where h_i has a standard normal distribution. To allow the correlation, we now simply substitute this expression for w_i in the simulated (or original) log likelihood and add ρ_{vw} to the list of parameters to be estimated. The simulation is still over independent normal variates, v_i and h_i.

Notwithstanding the preceding derivation, much of the recent attention has focused on simpler two-step estimators. Building on Ridder and Wansbeek (1990) and Verbeek and Nijman (1992),[48] Vella and Verbeek (1999) propose a two-step methodology that involves a random effects framework similar to the one in (19-31). As they note, there is some loss in efficiency by not using the FIML estimator. But, with the sample sizes typical in contemporary panel data sets, that efficiency loss may not be large. As they note, their two-step template encompasses a variety of models including the tobit model examined in the preceding sections and the mover-stayer model noted earlier.

The Vella and Verbeek model requires some fairly intricate maximum likelihood procedures. Wooldridge (1995) proposes an estimator that, with a few probably—but not necessarily—innocent assumptions, can be based on straightforward applications of conventional, everyday methods. We depart from a fixed effects specification,

$$\begin{aligned} y_{it}^* &= \eta_i + \mathbf{x}_{iT}'\boldsymbol{\beta} + \varepsilon_{it}, \\ d_{it}^* &= \theta_i + \mathbf{z}_{iT}'\boldsymbol{\alpha} + u_{it}, \\ (\varepsilon_{it}, u_{it}) &\sim N_2[(0,0), (\sigma^2, 1, \rho\sigma)]. \end{aligned}$$

Under the **mean independence assumption**,

$$E[\varepsilon_{it} | \eta_i, \theta_i, \mathbf{z}_{i1}, \ \ldots, \mathbf{z}_{iT}, v_{i1}, \ \ldots, v_{it}, d_{i1}, \ \ldots, d_{it}] = \rho u_{it},$$

it will follow that

$$E[y_{it} | \mathbf{x}_{i1}, \ \ldots, \mathbf{x}_{iT}, \eta_i, \theta_i, \mathbf{z}_{i1}, \ \ldots, \mathbf{z}_{iT}, v_{i1}, \ \ldots, v_{iT}, d_{i1}, \ \ldots, d_{iT}] = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho u_{it}.$$

This suggests an approach to estimating the model parameters; however, it requires computation of u_{it}. That would require estimation of θ_i, which cannot be done, at least not consistently—and that precludes simple estimation of u_{it}. To escape the dilemma, Wooldridge (2002c) suggests Chamberlain's approach to the fixed effects model,

$$\theta_i = f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \ \cdots \ + \mathbf{z}_{iT}'\mathbf{f}_T + h_i.$$

With this substitution,

$$\begin{aligned} d_{it}^* &= \mathbf{z}_{iT}'\boldsymbol{\alpha} + f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \ \cdots \ + \mathbf{z}_{iT}'\mathbf{f}_T + h_i + u_{it} \\ &= \mathbf{z}_{iT}'\boldsymbol{\alpha} + f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \ \cdots \ + \mathbf{z}_{iT}'\mathbf{f}_T + w_{it}, \end{aligned}$$

where w_{it} is independent of $\mathbf{z}_{iT}$, $t = 1, \ \ldots, T$. This now implies that

$$\begin{aligned} E[y_{it} | \mathbf{x}_{i1}, \ \ldots, \mathbf{x}_{iT}, \eta_i, \theta_i, \mathbf{z}_{i1}, \ \ldots, \mathbf{z}_{iT}, v_{i1}, \ \ldots, v_{iT}, d_{i1}, \ \ldots, d_{iT}] &= \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho(w_{it} - h_i) \\ &= (\eta_i - \rho h_i) + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho w_{it}. \end{aligned}$$

[48] See Vella (1998) for numerous additional references.

To complete the estimation procedure, we now compute T cross-sectional probit models (reestimating $f_0, \mathbf{f}_1, \ldots$ each time) and compute $\hat{\lambda}_{it}$ from each one. The resulting equation,

$$y_{it} = a_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho\hat{\lambda}_{it} + v_{it},$$

now forms the basis for estimation of $\boldsymbol{\beta}$ and ρ by using a conventional fixed effects linear regression with the observed data.

19.4.5.b Attrition

The literature on sample selection contains numerous analyses of two-period models, such as Kyriazidou (1997, 2001). They generally focus on nonparametric and semiparametric analyses. An early parametric contribution of Hausman and Wise (1979) is also a two-period model of attrition, which would seem to characterize many of the studies suggested in the current literature. The model formulation is a two-period random effects specification,

$$y_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta} + \varepsilon_{i1} + u_i \quad \text{(first period regression)},$$
$$y_{i2} = \mathbf{x}_{i2}'\boldsymbol{\beta} + \varepsilon_{i2} + u_i \quad \text{(second period regression)}.$$

Attrition is likely in the second period (to begin the study, the individual must have been observed in the first period). The authors suggest that the probability that an observation is made in the second period varies with the value of y_{i2} as well as some other variables,

$$z_{i2}^* = \delta y_{i2} + \mathbf{x}_{i2}'\boldsymbol{\theta} + \mathbf{w}_{i2}'\boldsymbol{\alpha} + v_{i2}.$$

Attrition occurs if $z_{i2}^* \leq 0$, which produces a probit model,

$$z_{i2} = \mathbf{1}(z_{i2}^* > 0) \quad \text{(attrition indicator observed in period 2)}.$$

An observation is made in the second period if $z_{i2} = 1$, which makes this an early version of the familiar sample selection model. The reduced form of the observation equation is

$$\begin{aligned} z_{i2}^* &= \mathbf{x}_{i2}'(\delta\boldsymbol{\beta} + \boldsymbol{\theta}) + \mathbf{w}_{i2}'\boldsymbol{\alpha} + \delta\varepsilon_{i2} + v_{i2} \\ &= \mathbf{x}_{i2}'\boldsymbol{\pi} + \mathbf{w}_{i2}'\boldsymbol{\alpha} + h_{i2} \\ &= \mathbf{r}_{i2}'\boldsymbol{\gamma} + h_{i2}. \end{aligned}$$

The variables in the probit equation are all those in the second period regression plus any additional ones dictated by the application. The estimable parameters in this model are $\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 = \text{Var}[\varepsilon_{it} + u_i]$, and two correlation coefficients, and $\rho_{12} = \text{Corr}[\varepsilon_{i1} + u_i, \varepsilon_{i2} + u_i] = \text{Var}[u_i]/\sigma^2$. All disturbances are assumed to be normally distributed. (Readers are referred to the paper for motivation and details on this specification.)

The authors propose a full information maximum likelihood estimator. Estimation can be simplified somewhat by using two steps. The parameters of the probit model can be estimated first by maximum likelihood. Then the remaining parameters are estimated by maximum likelihood, conditionally on these first-step estimates. The Murphy and Topel adjustment is made after the second step. [See Greene (2007d).]

The Hausman and Wise model covers the case of two periods in which there is a formal mechanism in the model for retention in the second period. It is unclear how the procedure could be extended to a multiple-period application such as that in Contoyannis et al. (2004), which involved a panel data set with eight waves. In addition, in that study,

the variables in the main equations were counts of hospital visits and physican visits, which complicates the use of linear regression. A workable solution to the problem of attrition in a multiperiod panel is the **inverse probability weighted estimator** [Wooldridge (2010, 2006b) and Rotnitzky and Robins (2005)]. In the Contoyannis application, there are eight waves in the panel. Attrition is taken to be "ignorable" so that the unobservables in the attrition equation and in the main equation(s) of interest are uncorrelated. (Note that Hausman and Wise do not make this assumption.) This enables Contoyannis et al. to fit a "retention" probit equation for each observation present at wave 1, for waves 2–8, using characteristics observed at the entry to the panel. [This defines, then, "selection (retention) on observables." See Moffitt, Fitzgerald, and Gottschalk (1999) for discussion.] Defining d_{it} to be the indicator for presence ($d_{it} = 1$) or absence ($d_{it} = 0$) of observation i in wave t, it will follow that the sequence of observations will begin at 1 and either stay at 1 or change to 0 for the remaining waves. Let $\hat{p}_{it}$ denote the predicted probability from the probit estimator at wave t. Then, their full log likelihood is constructed as

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{d_{it}}{\hat{p}_{it}} \ln L_{it}.$$

Wooldridge (2010) presents the underlying theory for the properties of this weighted maximum likelihood estimator. [Further details on the use of the inverse probability weighted estimator in the Contoyannis et al. (2004) study appear in Jones, Koolman, and Rice (2006) and in Section 17.7.7.]

19.5 MODELS FOR DURATION

The leading application of the censoring models we examined in Section 19.3 is models for durations and events. We consider the time until some kind of transition as the duration, and the transition, itself, as the event. The length of a spell of unemployment (until rehire or exit from the market), the duration of a strike, the amount of time until a patient ends a health-related spell in connection with a disease or operation, and the length of time between origination and termination (via prepayment, default, or some other mechanism) of a mortgage are all examples of durations and transitions. The role that censoring plays in these scenarios is that, in almost all cases in which we, as analysts, study duration data, some or even many of the spells we observe do not end in transitions. For example, in studying the lengths of unemployment spells, many of the individuals in the sample may still be unemployed at the time the study ends—the analyst observes (or believes) that the spell will end some time after the observation window closes. These data on spell lengths are, by construction, censored. Models of duration will generally account explicitly for censoring of the duration data.

This section is concerned with models of duration. In some aspects, the regression-like models we have studied, such as the discrete choice models, are the appropriate tools. As in the previous two chapters, however, the models are nonlinear, and the familiar regression methods are not appropriate. Most of this analysis focuses on maximum likelihood estimators. In modeling duration, although an underlying regression model is, in fact, at work, it is generally not the conditional mean function that is of interest. More likely, as we will explore next, the objects of estimation are certain probabilities of events, for example in the conditional probability of a transition in a given interval

given that the spell has lasted up to the point of interest. These are known as *hazard models*—the probability is labeled the hazard function—and are a central focus of this type of analysis.

19.5.1 MODELS FOR DURATION DATA[49]

Intuition might suggest that the longer a strike persists, the more likely it is that it will end within, say, the next week. Or is it? It seems equally plausible to suggest that the longer a strike has lasted, the more difficult must be the problems that led to it in the first place, and hence the *less* likely it is that it will end in the next short time interval. A similar kind of reasoning could be applied to spells of unemployment or the interval between conceptions. In each of these cases, it is not only the duration of the event, per se, that is interesting, but also the likelihood that the event will end in the *next period* given that it has lasted as long as it has.

Analysis of the length of *time until failure* has interested engineers for decades. For example, the models discussed in this section were applied to the durability of electric and electronic components long before economists discovered their usefulness. Likewise, the analysis of *survival times*—for example, the length of survival after the onset of a disease or after an operation such as a heart transplant—has long been a staple of biomedical research. Social scientists have recently applied the same body of techniques to strike duration, length of unemployment spells, intervals between conception, time until business failure, length of time between arrests, length of time from purchase until a warranty claim is made, intervals between purchases, and so on.

This section will give a brief introduction to the econometric analysis of duration data. As usual, we will restrict our attention to a few straightforward, relatively uncomplicated techniques and applications, primarily to introduce terms and concepts. The reader can then wade into the literature to find the extensions and variations. We will concentrate primarily on what are known as parametric models. These apply familiar inference techniques and provide a convenient departure point. Alternative approaches are considered at the end of the discussion.

19.5.2 DURATION DATA

The variable of interest in the analysis of duration is the length of time that elapses from the beginning of some event either until its end or until the measurement is taken, which may precede termination. Observations will typically consist of a cross section of durations, $t_1, t_2, \ldots, t_n$. The process being observed may have begun at different points in calendar time for the different individuals in the sample. For example, the strike duration data examined in Example 19.14 are drawn from nine different years.

Censoring is a pervasive and usually unavoidable problem in the analysis of duration data. The common cause is that the measurement is made while the process is ongoing. An obvious example can be drawn from medical research. Consider analyzing the survival times of heart transplant patients. Although the beginning times may be known with precision, at the time of the measurement, observations on any individuals who

[49]There are a large number of highly technical articles on this topic, but relatively few accessible sources for the uninitiated. A particularly useful introductory survey is Kiefer (1985, 1988), upon which we have drawn heavily for this section. Other useful sources are Kalbfleisch and Prentice (2002), Heckman and Singer (1984a), Lancaster (1990), Florens, Fougere, and Mouchart (1996), and Cameron and Trivedi (2005, Chapters 17–19).

are still alive are necessarily censored. Likewise, samples of spells of unemployment drawn from surveys will probably include some individuals who are still unemployed at the time the survey is taken. For these individuals, duration or survival is at least the observed t_i, but not equal to it. Estimation must account for the censored nature of the data for the same reasons as considered in Section 19.3. The consequences of ignoring censoring in duration data are similar to those that arise in regression analysis.

In a conventional regression model that characterizes the conditional mean and variance of a distribution, the regressors can be taken as fixed characteristics at the point in time or for the individual for which the measurement is taken. When measuring duration, the observation is implicitly on a process that has been under way for an interval of time from zero to t. If the analysis is conditioned on a set of covariates (the counterparts to regressors) $\mathbf{x}_t$, then the duration is implicitly a function of the time path of the variable $\mathbf{x}(t), t = (0, t)$, which may have changed during the interval. For example, the observed duration of employment in a job may be a function of the individual's rank in the firm. But that rank may have changed several times between the time of hire and when the observation was made. As such, observed rank at the end of the job tenure is not necessarily a complete description of the individual's rank *while he or she was employed*. Likewise, marital status, family size, and amount of education are all variables that can change during the duration of unemployment and that one would like to account for in the duration model. The treatment of **time-varying covariates** is a considerable complication.[50]

19.5.3 A REGRESSION-LIKE APPROACH: PARAMETRIC MODELS OF DURATION

We will use the term *spell* as a catchall for the different duration variables we might measure. Spell length is represented by the random variable T. A simple approach to duration analysis would be to apply regression analysis to the sample of observed spells. By this device, we could characterize the expected duration, perhaps conditioned on a set of covariates whose values were measured at the end of the period. We could also assume that conditioned on an $\mathbf{x}$ that has remained fixed from $T = 0$ to $T = t, t$ has a normal distribution, as we commonly do in regression. We could then characterize the probability distribution of observed duration times. But normality turns out not to be particularly attractive in this setting for a number of reasons, not least of which is that duration is positive by construction, while a normally distributed variable can take negative values. (*Log* normality turns out to be a palatable alternative, but it is only one in a long list of candidates.)

19.5.3.a Theoretical Background

Suppose that the random variable T has a continuous probability distribution $f(t)$, where t is a realization of T. The cumulative probability is

$$F(t) = \int_0^t f(s)\, ds = \text{Prob}(T \leq t).$$

We will usually be more interested in the probability that the spell is of length *at least* t, which is given by the **survival function**,

$$S(t) = 1 - F(t) = \text{Prob}(T \geq t).$$

[50] See Petersen (1986) for one approach to this problem.

Consider the question raised in the introduction: Given that the spell has lasted until time t, what is the probability that it will end in the next short interval of time, say, Δt? It is

$$l(t, \Delta t) = \text{Prob}(t \leq T \leq t + \Delta t | T \geq t).$$

A useful function for characterizing this aspect of the distribution is the **hazard rate**,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t\, S(t)} = \frac{f(t)}{S(t)}.$$

Roughly, the hazard rate is the rate at which spells are completed after duration t, given that they last at least until t. As such, the hazard function gives an answer to our original question.

The hazard function, the density, the CDF, and the survival function are all related. The hazard function is

$$\lambda(t) = \frac{-d \ln S(t)}{dt},$$

so

$$f(t) = S(t)\lambda(t).$$

Another useful function is the **integrated hazard function**,

$$\Lambda(t) = \int_0^t \lambda(s)\, ds,$$

for which

$$S(t) = e^{-\Lambda(t)},$$

so

$$\Lambda(t) = -\ln S(t).$$

The integrated hazard function is a **generalized residual** in this setting.[51]

19.5.3.b Models of the Hazard Function

For present purposes, the hazard function is more interesting than the survival rate or the density. Based on the previous results, one might consider modeling the hazard function itself, rather than, say, modeling the survival function and then obtaining the density and the hazard. For example, the base case for many analyses is a hazard rate that does not vary over time. That is, $\lambda(t)$ is a constant λ. This is characteristic of a process that has no memory; the *conditional* probability of failure in a given short interval is the same regardless of when the observation is made. Thus,

$$\lambda(t) = \lambda.$$

From the earlier definition, we obtain the simple differential equation,

$$\frac{-d \ln S(t)}{dt} = \lambda.$$

The solution is

$$\ln S(t) = k - \lambda t,$$

[51] See Chesher and Irish (1987) and Example 19.13.

or

$$S(t) = Ke^{-\lambda t},$$

where K is the constant of integration. The terminal condition that $S(0) = 1$ implies that $K = 1$, and the solution is

$$S(t) = e^{-\lambda t}.$$

This solution is the **exponential** distribution, which has been used to model the time until failure of electronic components. Estimation of λ is simple, because with an exponential distribution, $E[t] = 1/\lambda$. The maximum likelihood estimator of λ would be the reciprocal of the sample mean.

A natural extension might be to model the hazard rate as a linear function, $\lambda(t) = \alpha + \beta t$. Then $\Lambda(t) = \alpha t + \frac{1}{2}\beta t^2$ and $f(t) = \lambda(t)S(t) = \lambda(t)\exp[-\Lambda(t)]$. To avoid a negative hazard function, one might depart from $\lambda(t) = \exp[g(t, \boldsymbol{\theta})]$, where $\boldsymbol{\theta}$ is a vector of parameters to be estimated. With an observed sample of durations, estimation of α and β is, at least in principle, a straightforward problem in maximum likelihood.[52]

A distribution whose hazard function slopes upward is said to have **positive duration dependence**. For such distributions, the likelihood of failure at time t, conditional upon duration up to time t, is increasing in t. The opposite case is that of decreasing hazard or **negative duration dependence**. Our question in the introduction about whether the strike is more or less likely to end at time t given that it has lasted until time t can be framed in terms of positive or negative duration dependence. The assumed distribution has a considerable bearing on the answer. If one is unsure at the outset of the analysis whether the data can be characterized by positive or negative duration dependence, then it is counterproductive to assume a distribution that displays one characteristic or the other over the entire range of t. Thus, the exponential distribution and our suggested extension could be problematic. The literature contains a variety of choices for duration models: normal, inverse normal,[53] lognormal, F, gamma, Weibull (which is a popular choice), and many others.[54] To illustrate the differences, we will examine a few of the simpler ones. Table 19.8 lists the hazard functions and survival functions for four

TABLE 19.8 Survival Distributions

Distribution	*Hazard Function,* $\lambda(t)$	*Survival Function, S(t)*
Exponential	$\lambda(t) = \lambda$	$S(t) = e^{-\lambda t}$
Weibull	$\lambda(t) = \lambda p(\lambda t)^{p-1}$	$S(t) = e^{-(\lambda t)^p}$
Lognormal	$f(t) = (p/t)\phi[p \ln(\lambda t)]$	$S(t) = \Phi[-p \ln(\lambda t)]$
	[ln t is normally distributed with mean $-\ln \lambda$ and standard deviation $1/p$.]	
Loglogistic	$\lambda(t) = \lambda p(\lambda t)^{p-1}/[1 + (\lambda t)^p]$	$S(t) = 1/[1 + (\lambda t)^p]$
	[ln t has a logistic distribution with mean $-\ln \lambda$ and variance $\pi^2/(3p^2)$.]	

[52]Kennan (1985) used a similar approach.

[53]Inverse Gaussian; see Lancaster (1990).

[54]Three sources that contain numerous specifications are Kalbfleisch and Prentice (2002), Cox and Oakes (1985), and Lancaster (1990).

commonly used distributions. Each involves two parameters, a location parameter, λ, and a scale parameter, p.[55]

All these are distributions for a nonnegative random variable. Their hazard functions display very different behaviors, as can be seen in Figure 19.7. The hazard function for the exponential distribution is constant, that for the Weibull is monotonically increasing or decreasing depending on p, and the hazards for lognormal and loglogistic distributions first increase and then decrease. Which among these or the many alternatives is likely to be best in any application is uncertain.

19.5.3.c Maximum Likelihood Estimation

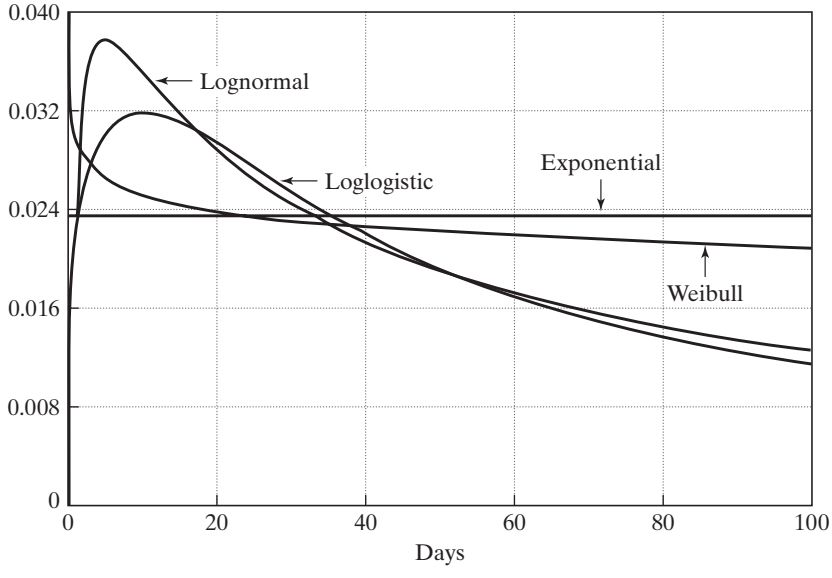
The parameters λ and p of the models in Table 19.8 can be estimated by maximum likelihood. For observed duration data, $t_1, t_2, \ldots, t_n$, the log-likelihood function can be formulated and maximized in the ways we have become familiar with in earlier chapters. Censored observations can be incorporated as in Section 19.3 for the tobit model. [See (19-13).] As such,

$$\ln L(\boldsymbol{\theta}) = \sum_{\substack{\text{uncensored}\\ \text{observations}}} \ln f(t|\boldsymbol{\theta}) + \sum_{\substack{\text{censored}\\ \text{observations}}} \ln S(t|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\lambda, p)$. For some distributions, it is convenient to formulate the log-likelihood function in terms of $f(t) = \lambda(t)S(t)$ so that

$$\ln L(\boldsymbol{\theta}) = \sum_{\substack{\text{uncensored}\\ \text{observations}}} \ln \lambda(t|\boldsymbol{\theta}) + \sum_{\substack{\text{all}\\ \text{observations}}} \ln S(t|\boldsymbol{\theta}).$$

FIGURE 19.7 Parametric Hazard Functions.



[55] *Note:* In the benchmark case of the exponential distribution, λ is the hazard function. In all other cases, the hazard function is a function of λ, p, and, where there is duration dependence, t as well. Different authors, for example, Kiefer (1988), use different parameterizations of these models. We follow the convention of Kalbfleisch and Prentice (2002).

Inference about the parameters can be done in the usual way. Either the BHHH estimator or actual second derivatives can be used to estimate asymptotic standard errors for the estimates.[56] The transformation to $w = p(\ln t + \ln \lambda)$ for these distributions greatly facilitates maximum likelihood estimation. For example, for the Weibull model, by defining $w = p(\ln t + \ln \lambda)$, we obtain the very simple density $f(w) = \exp[w - \exp(w)]$ and survival function $S(w) = \exp(-\exp(w))$.[57] Therefore, by using $\ln t$ instead of t, we greatly simplify the log-likelihood function. Details for these and several other distributions may be found in Kalbfleisch and Prentice (2002, pp. 68–70). The Weibull distribution is examined in detail in the next section.

19.5.3.d Exogenous Variables

One limitation of the models given earlier is that external factors are not given a role in the survival distribution. The addition of covariates to duration models is fairly straightforward, although the interpretation of the coefficients in the model is less so. Consider, for example, the Weibull model. (The extension to other distributions will be similar.) Let

$$\lambda_i = e^{-\mathbf{x}_i'\boldsymbol{\beta}},$$

where $\mathbf{x}_i$ is a constant term and a set of variables that are assumed not to change from time $T = 0$ until the failure time, $T = t_i$. Making λ_i a function of a set of covariates is equivalent to changing the units of measurement on the time axis. For this reason, these models are sometimes called **accelerated failure time models**. Note as well that in all the models listed (and generally), the regressors do not bear on the question of duration dependence, which, when it varies, is a function of p.

Let $\sigma = 1/p$ and let $\delta_i = 1$ if the spell is completed and $\delta_i = 0$ if it is censored. As before, let

$$w_i = p \ln(\lambda_i t_i) = \frac{(\ln t_i - \mathbf{x}_i'\boldsymbol{\beta})}{\sigma},$$

and denote the density and survival functions $f(w_i)$ and $S(w_i)$. The observed random variable is $\ln t_i = \sigma w_i + \mathbf{x}_i'\boldsymbol{\beta}$. The Jacobian of the transformation from w_i to $\ln t_i$ is $dw_i/d \ln t_i = 1/\sigma$, so the density and survival functions for $\ln t_i$ are

$$f(\ln t_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma} f\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right), \quad \text{and} \quad S(\ln t_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma) = S\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right).$$

The log likelihood for the observed data is

$$\ln L(\boldsymbol{\beta}, \sigma | \text{data}) = \sum_{i=1}^{n} [\delta_i \ln f(\ln t_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma) + (1 - \delta_i) \ln S(\ln t_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma)].$$

For the **Weibull model**, for example (see Footnote 57),

$$f(w_i) = \exp(w_i - e^{w_i}),$$

[56] One can compute a robust covariance matrix as considered in Chapter 14. It is unclear in this context what *failures* of the model assumptions would be considered.

[57] The transformation is $\exp(w) = (\lambda t)^p$ so $t = (1/\lambda)[\exp(w)]^{1/p}$. The Jacobian of the transformation is $dt/dw = [\exp(w)]^{1/p}/(\lambda p)$. The density in Table 19.8 is $\lambda p[\exp(w)]^{-(1/p)-1}[\exp(-\exp(w))]$. Multiplying by the Jacobian produces the result, $f(w) = \exp[w - \exp(w)]$. The survival function is the antiderivative, $[\exp(-\exp(w))]$.

and

$$S(w_i) = \exp(-e^{w_i}).$$

Making the transformation to $\ln t_i$ and collecting terms reduces the log likelihood to

$$\ln L(\boldsymbol{\beta}, \sigma \,|\, \text{data}) = \sum_i \left[\delta_i \left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma} - \ln \sigma \right) - \exp\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma} \right) \right].$$

(Many other distributions, including the others in Table 19.8, simplify in the same way. The exponential model is obtained by setting σ to one.) The derivatives can be equated to zero using the methods described in Section E.3. The individual terms can also be used to form the BHHH estimator of the asymptotic covariance matrix for the estimator.[58] The Hessian is also simple to derive, so Newton's method could be used instead.[59]

Note that the hazard function generally depends on t, p, and $\mathbf{x}$. The sign of an estimated coefficient suggests the direction of the effect of the variable on the hazard function when the hazard is monotonic. But in those cases, such as the loglogistic, in which the hazard is nonmonotonic, even this may be ambiguous. The magnitudes of the effects may also be difficult to interpret in terms of the hazard function. In a few cases, we do get a regression-like interpretation. In the Weibull and exponential models, $E[t \,|\, \mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})\Gamma[(1/p) + 1]$, whereas for the lognormal and loglogistic models, $E[\ln t \,|\, \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$. In these cases, β_k is the derivative (or a multiple of the derivative) of this conditional mean. For some other distributions, the conditional median of t is easily obtained. Numerous cases are discussed by Kiefer (1985, 1988), Kalbfleisch and Prentice (2002), and Lancaster (1990).

19.5.3.e Heterogeneity

The problem of unobserved heterogeneity in duration models can be viewed essentially as the result of an incomplete specification. Individual specific covariates are intended to incorporate observation specific effects. But if the model specification is incomplete and if systematic individual differences in the distribution remain after the observed effects are accounted for, then inference based on the improperly specified model is likely to be problematic. We have already encountered several settings in which the possibility of heterogeneity mandated a change in the model specification; the fixed and random effects regression, logit, and probit models all incorporate observation-specific effects. Indeed, all the failures of the linear regression model discussed in the preceding chapters can be interpreted as a consequence of heterogeneity arising from an incomplete specification.

There are a number of ways of extending duration models to account for heterogeneity. The strictly nonparametric approach of the Kaplan–Meier (1958) estimator (see Section 19.5.4) is largely immune to the problem, but it is also rather limited in how much information can be obtained from it. One direct approach is to model heterogeneity in the parametric model. Suppose that we posit a survival function conditioned on the individual specific effect v_i. We treat the survival function as $S(t_i \,|\, v_i)$. Then add to that a model for the unobserved heterogeneity $f(v_i)$. (Note that this is a

[58]Note that the log-likelihood function has the same form as that for the tobit model in Section 19.3.2. By just reinterpreting the nonlimit observations in a tobit setting, we can, therefore, use this framework to apply a wide range of distributions to the tobit model.

[59]See Kalbfleisch and Prentice (2002) for numerous other examples.

counterpart to the incorporation of a disturbance in a regression model and follows the same procedures that we used in the Poisson model with random effects.) Then

$$S(t) = E_v[S(t|v)] = \int_v S(t|v)f(v)\,dv.$$

The gamma distribution is frequently used for this purpose.[60] Consider, for example, using this device to incorporate heterogeneity into the Weibull model we used earlier. As is typical, we assume that v has a gamma distribution with mean 1 and variance $\theta = 1/k$. Then

$$f(v) = \frac{k^k}{\Gamma(k)} e^{-kv} v^{k-1},$$

and

$$S(t|v) = e^{-(v\lambda t)^p}.$$

After a bit of manipulation, we obtain the unconditional distribution,

$$S(t) = \int_0^\infty S(t|v)\, f(v)\, dv = [1 + \theta(\lambda t)^p]^{-1/\theta}.$$

The limiting value, with $\theta = 0$, is the **Weibull survival model**, so $\theta = 0$ corresponds to $\text{Var}[v] = 0$, or no heterogeneity.[61] The hazard function for this model is

$$\lambda(t) = \lambda p(\lambda t)^{p-1}[S(t)]^\theta,$$

which shows the relationship to the Weibull model.

This approach is common in parametric modeling of heterogeneity. In an important paper on this subject, Heckman and Singer (1984b) argued that this approach tends to overparameterize the survival distribution and can lead to rather serious errors in inference. They gave some dramatic examples to make the point. They also expressed some concern that researchers tend to choose the distribution of heterogeneity more on the basis of mathematical convenience than on any sensible economic basis. (We examined Heckman and Singer's approach in a probit model in Example 17.27.)

19.5.4 NONPARAMETRIC AND SEMIPARAMETRIC APPROACHES

The parametric models are attractive for their simplicity. But by imposing as much structure on the data as they do, the models may distort the estimated hazard rates. It may be that a more accurate representation can be obtained by imposing fewer restrictions.

The Kaplan–Meier (1958) **product limit estimator** is a strictly empirical, nonparametric approach to survival and hazard function estimation. Assume that the observations on duration are sorted in ascending order so that $t_1 \leq t_2$ and so on and, for now, that no observations are censored. Suppose as well that there are K distinct survival times in the data, denoted T_k; K will equal n unless there are ties. Let n_k denote

[60] See, for example, Hausman, Hall, and Griliches (1984), who use it to incorporate heterogeneity in the Poisson regression model. The application is developed in Section 18.4.7.c.

[61] For the strike data analyzed in Figure 19.7, the maximum likelihood estimate of θ is 0.0004, which suggests that at least in the context of the Weibull model, latent heterogeneity does not appear to be a feature of these data.

the number of individuals whose observed duration is at least T_k. The set of individuals whose duration is at least T_k is called the **risk set** at this duration. (We borrow, once again, from biostatistics, where the risk set is those individuals still "at risk" at time T_k.) Thus, n_k is the size of the risk set at time T_k. Let h_k denote the number of observed spells completed at time T_k. A strictly empirical estimate of the survivor function would be

$$\hat{S}(T_k) = \prod_{i=1}^{k} \frac{n_i - h_i}{n_i} = \frac{n_i - h_i}{n_1}.$$

The estimator of the hazard rate is

$$\hat{\lambda}(T_k) = \frac{h_k}{n_k}. \tag{19-34}$$

Corrections are necessary for observations that are censored. Lawless (1982), Kalbfleisch and Prentice (2002), Greene (1995a) and Kiefer (1988) and give details. Susin (2001) points out a fundamental ambiguity in this calculation (one which he argues appears in the 1958 source). The estimator in (19-34) is not a *rate* as such, as the width of the time window is undefined, and could be very different at different points in the chain of calculations. Because many intervals, particularly those late in the observation period, might have zeros, the failure to acknowledge these intervals should impart an upward bias to the estimator. His proposed alternative computes the counterpart to (19-34) over a mesh of defined intervals as follows,

$$\hat{\lambda}(I_a^b) = \frac{\sum_{j=a}^{b} h_j}{\sum_{j=a}^{b} n_j b_j},$$

where the interval is from $t = a$ to $t = b$, h_j is the number of failures in each period in this interval, n_j is the number of individuals at risk in that period, and b_j is the width of the period. Thus, an interval (a, b) is likely to include several "periods."

Cox's (1972) approach to the **proportional hazard** model is another popular, semiparametric method of analyzing the effect of covariates on the hazard rate. The model specifies that

$$\lambda(t_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta})\lambda_0(t_i).$$

The function λ_0 is the "baseline" hazard, which is the individual heterogeneity. In principle, this hazard is a parameter for each observation that must be estimated. Cox's **partial likelihood** estimator provides a method of estimating $\boldsymbol{\beta}$ without requiring estimation of λ_0. The estimator is somewhat similar to Chamberlain's estimator for the logit model with panel data in that a conditioning operation is used to remove the heterogeneity. (See Section 17.7.3.a) Suppose that the sample contains K distinct exit times, $T_1, \ldots, T_K$. For any time T_k, the risk set, denoted R_k, is all individuals whose exit time is at least T_k. The risk set is defined with respect to any moment in time T as the set of individuals who have not yet exited just prior to that time. For every individual i in risk set R_k, $t_i \geq T_k$. The probability that an individual exits at time T_k *given that exactly one individual exits at this time* (which is the counterpart to the conditioning in the binary logit model in Chapter 17) is

$$\text{Prob}[t_i = T_k \,|\, \text{risk set}_k] = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\sum_{j \in R_k} e^{\mathbf{x}_j'\boldsymbol{\beta}}}.$$

Thus, the conditioning sweeps out the baseline hazard functions. For the simplest case in which exactly one individual exits at each distinct exit time and there are no censored observations, the partial log likelihood is

$$\ln L = \sum_{k=1}^{K}\left[\mathbf{x}_k'\boldsymbol{\beta} - \ln \sum_{j\in R_k} e^{\mathbf{x}_j'\boldsymbol{\beta}}\right].$$

If m_k individuals exit at time T_k, then the contribution to the log likelihood is the sum of the terms for each of these individuals.

The proportional hazard model is a common choice for modeling durations because it is a reasonable compromise between the Kaplan–Meier estimator and the possibly excessively structured parametric models. Hausman and Han (1990) and Meyer (1988), among others, have devised other, semiparametric specifications for hazard models.

Example 19.13 Survival Models for Strike Duration

The strike duration data given in Kennan (1985, pp. 14–16) have become a familiar standard for the demonstration of hazard models. Appendix Table F19.2 lists the durations, in days, of 62 strikes that commenced in June of the years 1968 to 1976. Each involved at least 1,000 workers and began at the expiration or reopening of a contract. Kennan reported the actual duration. In his survey, Kiefer (1985), using the same observations, censored the data at 80 days to demonstrate the effects of censoring. We have kept the data in their original form; the interested reader is referred to Kiefer for further analysis of the censoring problem.[62]

Parameter estimates for the four duration models are given in Table 19.8. The estimate of the median of the survival distribution is obtained by solving the equation $S(t) = 0.5$. For example, for the Weibull model, $S(M) = 0.5 = \exp[-(\lambda M)^P]$, or $M = [(\ln 2)^{1/P}]/\lambda$. For the exponential model, $p = 1$. For the lognormal and loglogistic models, $M = 1/\lambda$. The delta method is then used to estimate the standard error of this function of the parameter estimates. (See Section 4.6.) All these distributions are skewed to the right. As such, $E[t]$ is greater than the median. For the exponential and Weibull models, $E[t] = [1/\lambda]\Gamma[(1/p) + 1]$; for the normal, $E[t] = (1/\lambda)[\exp(1/p^2)]^{1/2}$. The implied hazard functions are shown in Figure 19.7.

The variable x reported with the strike duration data is a measure of unanticipated aggregate industrial production net of seasonal and trend components. It is computed as the residual in a regression of the log of industrial production in manufacturing on time, time squared, and monthly dummy variables. With the industrial production variable included as a covariate, the estimated Weibull model is

$$\begin{aligned} -\ln \lambda &= 3.7772 - 9.3515\,x, & p &= 1.00288,\\ &\quad (0.1394)\quad (2.973) & &\quad (0.1217),\\ \text{median strike length} &= 27.35(3.667)\ \text{days}, & E[t] &= 39.83\ \text{days}. \end{aligned}$$

TABLE 19.9 Estimated Duration Models (Estimated standard errors in parentheses)

	λ	p	*Median Duration*
Exponential	0.02344 (0.00298)	1.00000 (0.00000)	29.571 (3.522)
Weibull	0.02439 (0.00354)	0.92083 (0.11086)	27.543 (3.997)
Loglogistic	0.04153 (0.00707)	1.33148 (0.17201)	24.079 (4.102)
Lognormal	0.04514 (0.00806)	0.77206 (0.08865)	22.152 (3.954)

[62]Our statistical results are nearly the same as Kiefer's despite the censoring.

Note that the Weibull model is now almost identical to the exponential model ($p = 1$). Because the hazard conditioned on x is approximately equal to λ_i, it follows that the hazard function is increasing in *unexpected* industrial production. A 1% increase in x leads to a 9.35% increase in λ, which because $p \approx 1$ translates into a 9.35% decrease in the median strike length or about 2.6 days. (Note that $M = \ln 2/\lambda$.)

The proportional hazard model does not have a constant term. (The baseline hazard is an individual specific constant.) The estimate of β is -9.0726, with an estimated standard error of 3.225. This is very similar to the estimate obtained for the Weibull model.

Example 19.14 Time Until Retirement

Christensen and Kallestrup-Lamb (2012) studied the duration of labor market attachment until retirement in a large sample of Danish workers. The observations were a sample of 9,329 individuals who were 50 years old in 1985. They were followed until 2001. Duration is defined as the number of years since 1985 until retirement, with right censoring occurring at $T = 17$. This is a *stock sample*—all individuals enter the initial state at the same point in time (1985), and are observed with reference to the same absolute time interval, 1985, where $T = 0$, to 2001, where $T = 17$. In a *flow sample*, individuals would enter at different points in the observation window, and time $T = 0$ would vary with each person as he or she entered. Data on labor market experience were augmented with matched data on health measures and health shocks, as well as socioeconomic and financial information including wealth and own and household income. The authors were interested in controlling for a sample selection effect suggested by initial participation in the program, and in other forms of unobserved heterogeneity. For the latter, they considered discrete, semiparametric approaches based on Heckman and Singer (1984a,b) (see Section 17.7.6) and the continuous forms in Section 19.5.3e. The primary approach involved a single "risk," identified broadly as exit from the labor force for any reason. However, the authors were also interested in a competing risks framework. They noted that exit could be motivated by five states: early retirement, disability, unemployment then early retirement, unemployment then some other form of exit, some other form. A variety of models were investigated. The Kaplan-Meier approach suggested the overall pattern of retirements, but did not allow for the influence of the time-varying covariates, especially the health status. Formal models based on the exponential (no duration dependence) and the Weibill model (with variable duration dependence) were considered. The two noted forms of latent heterogeneity were added to the specifications. This paper provides a lengthy application of most of the methods discussed in this section.

19.6 SUMMARY AND CONCLUSIONS

This chapter has examined settings in which, in principle, the linear regression model of Chapter 2 would apply, but the data-generating mechanism produces a nonlinear form: truncation, censoring, and sample selection or endogenous sampling. For each case, we develop the basic theory of the effect and then use the results in a major area of research in econometrics.

In the truncated regression model, the range of the dependent variable is restricted substantively. Certainly all economic data are restricted in this way—aggregate income data cannot be negative, for example. But when data are truncated so that plausible values of the dependent variable are precluded, for example, when zero values for expenditure are discarded, the data that remain are analyzed with models that explicitly account for the truncation. The stochastic frontier model is based on

a composite disturbance in which one part follows the assumptions of the familiar regression model while the second component is built on a platform of the truncated regression.

When data are censored, values of the dependent variable that could in principle be observed are masked. Ranges of values of the true variable being studied are observed as a single value. The basic problem this presents for model building is that in such a case, we observe variation of the independent variables without the corresponding variation in the dependent variable that might be expected. Consistent estimation and useful interpretation of estimation results are based on maximum likelihood or some other technique that explicitly accounts for the censoring mechanism. The most common case of censoring in observed data arises in the context of duration analysis, or survival functions (which borrows a term from medical statistics where this style of model building originated). It is useful to think of duration, or survival data, as the measurement of time between transitions or changes of state. We examined three modeling approaches that correspond to the description in Chapter 12, nonparametric (survival tables), semiparametric (the proportional hazard models), and parametric (various forms such as the Weibull model).

Finally, the issue of sample selection arises when the observed data are not drawn randomly from the population of interest. Failure to account for this nonrandom sampling produces a model that describes only the nonrandom subsample, not the larger population. In each case, we examined the model specification and estimation techniques which are appropriate for these variations of the regression model. Maximum likelihood is usually the method of choice, but for the third case, a two-step estimator has become more common. The leading contemporary application of selection methods and endogenous sampling is in the measure of treatment effects that are examined in Chapter 6.

Key Terms and Concepts

- Accelerated failure time model
- Attenuation
- Censored regression model
- Censored variable
- Conditional moment test
- Corner solution model
- Data envelopment analysis
- Degree of truncation
- Exponential
- Generalized residual
- Hazard function
- Hazard rate
- Incidental truncation
- Integrated hazard function
- Inverse Mills ratio
- Mean independence assumption
- Negative duration dependence
- Olsen's reparameterization
- Partial likelihood
- Positive duration dependence
- Product limit estimator
- Proportional hazard
- Risk set
- Sample selection
- Semiparametric estimator
- Stochastic frontier model
- Survival function
- Time-varying covariate
- Tobit model
- Truncated distribution
- Truncated mean
- Truncated normal distribution
- Truncated random variable
- Truncated standard normal distribution
- Truncated variance
- Truncation
- Two-part distribution
- Two-step estimation
- Weibull model
- Weibull survival model

Exercises

1. The following 20 observations are drawn from a censored normal distribution:

3.8396	7.2040	0.00000	0.00000	4.4132	8.0230
5.7971	7.0828	0.00000	0.80260	13.0670	4.3211
0.00000	8.6801	5.4571	0.00000	8.1021	0.00000
1.2526	5.6016				

   The applicable model is

$$y_i^* = \mu + \varepsilon_i,$$
$$y_i = y_i^* \quad \text{if } \mu + \varepsilon_i > 0, 0 \text{ otherwise},$$
$$\varepsilon_i \sim N[0, \sigma^2].$$

   Exercises 1 through 4 in this section are based on the preceding information. The OLS estimator of μ in the context of this tobit model is simply the sample mean. Compute the mean of all 20 observations. Would you expect this estimator to over- or underestimate μ? If we consider only the nonzero observations, then the truncated regression model applies. The sample mean of the nonlimit observations is the least squares estimator in this context. Compute it and then comment on whether this sample mean should be an overestimate or an underestimate of the true mean.
2. We now consider the tobit model that applies to the full data set.
   a. Formulate the log likelihood for this very simple tobit model.
   b. Reformulate the log likelihood in terms of $\theta = 1/\sigma$ and $\gamma = \mu/\sigma$. Then derive the necessary conditions for maximizing the log likelihood with respect to θ and γ.
   c. Discuss how you would obtain the values of θ and γ to solve the problem in part b.
   d. Compute the maximum likelihood estimates of μ and σ.
3. Using only the nonlimit observations, repeat Exercise 2 in the context of the truncated regression model. Estimate μ and σ by using the method of moments estimator outlined in Example 19.2. Compare your results with those in the previous exercises.
4. Continuing to use the data in Exercise 1, consider once again only the nonzero observations. Suppose that the sampling mechanism is as follows: y^* and another normally distributed random variable z have population correlation 0.7. The two variables, y^* and z, are sampled jointly. When z is greater than zero, y is reported. When z is less than zero, both z and y are discarded. Exactly 35 draws were required to obtain the preceding sample. Estimate μ and σ. (*Hint:* Use Theorem 19.5.)
5. Derive the partial effects for the tobit model with heteroscedasticity that is described in Section 19.3.5.b.
6. Prove that the Hessian for the tobit model in (19-14) is negative definite after Olsen's transformation is applied to the parameters.

Applications

1. We examined Ray Fair's famous analysis (*Journal of Political Economy,* 1978) of a *Psychology Today* survey on extramarital affairs in Example 18.18 using a Poisson

regression model. Although the dependent variable used in that study was a count, Fair (1978) used the tobit model as the platform for his study. You can reproduce the tobit estimates in Fair's paper easily with any software package that contains a tobit estimator—most do. The data appear in Appendix Table F18.1. Reproduce Fair's least squares and tobit estimates. Compute the partial effects for the model and interpret all results.

2. The Mroz (1975) data used in Example 19.10 (see Appendix Table F5.1) also describe a setting in which the tobit model has been frequently applied. The sample contains 753 observations on labor market outcomes for married women, including the following variables:

$$
\begin{aligned}
lfp &= \text{indicator (0/1) for whether in the formal labor market (lfp = 1) or not (lfp = 0)},\\
whrs &= \text{wife's hours worked},\\
kl6 &= \text{number of children under 6 years old in the household},\\
k618 &= \text{number of children from 6 to 18 years old in the household},\\
we &= \text{wife's age},\\
we &= \text{wife's education},\\
ww &= \text{wife's hourly wage},\\
hhrs &= \text{husband's hours worked},\\
ha &= \text{husband's age},\\
hw &= \text{husband's wage},\\
faminc &= \text{family income from other sources},\\
wmed &= \text{wife's mother's education}\\
wfed &= \text{wife's father's education}\\
cit &= \text{dummy variable for living in an urban area},\\
ax &= \text{labor market experience} = \text{wa} - \text{we} - 5,
\end{aligned}
$$

   and several variables that will not be useful here. Using these data, estimate a tobit model for the wife's hours worked. Report all results including partial effects and relevant diagnostic statistics. Repeat the analysis for the wife's labor earnings, $ww \times whrs$. Which is a more plausible model?

3. Continuing the analysis of the previous application, note that these data conform precisely to the description of corner solutions in Section 19.3.4. The dependent variable is not censored in the fashion usually assumed for a tobit model. To investigate whether the dependent variable is determined by a two-part decision process (yes/no and, if yes, how much), specify and estimate a two-equation model in which the first equation analyzes the binary decision $lfp = 1$ if $whrs > 0$ and 0 otherwise and the second equation analyzes $whrs \,|\, whrs > 0$. What is the appropriate model? What do you find? Report all results.

4. Stochastic Frontier Model. Section 10.3.1 presents estimates of a Cobb–Douglas cost function using Nerlove's 1955 data on the U.S. electric power industry. Christensen and Greene's 1976 update of this study used 1970 data for this industry. The Christensen and Greene data are given in Appendix Table F4.4. These data have

provided a standard test data set for estimating different forms of production and cost functions, including the stochastic frontier model discussed in Section 19.2.4. It has been suggested that one explanation for the apparent finding of economies of scale in these data is that the smaller firms were inefficient for other reasons. The stochastic frontier might allow one to disentangle these effects. Use these data to fit a frontier cost function which includes a quadratic term in log output in addition to the linear term and the factor prices. Then examine the estimated Jondrow et al. residuals to see if they do indeed vary negatively with output, as suggested. (This will require either some programming on your part or specialized software. The stochastic frontier model is provided as an option in *Stata* and *NLOGIT*. Or, the likelihood function can be programmed easily for *R*, or *GAUSS*.) (*Note:* For a cost frontier as opposed to a production frontier, it is necessary to reverse the sign on the argument in the Φ function that appears in the log likelihood.)

20
SERIAL CORRELATION

20.1 INTRODUCTION

Time-series data often display autocorrelation or serial correlation of the disturbances across periods. Consider, for example, the plot of the least squares residuals in the following example.

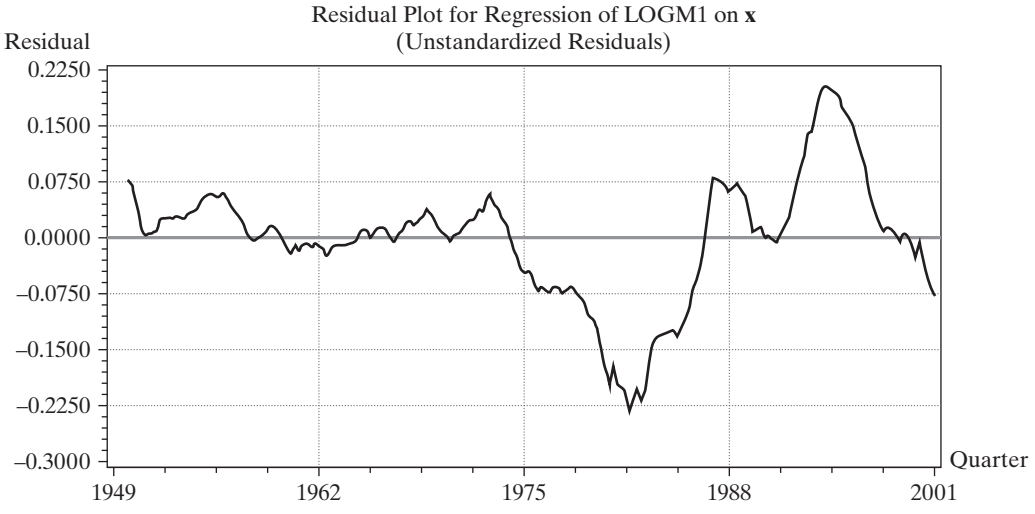
Example 20.1 Money Demand Equation

Appendix Table F5.2 contains quarterly data from 1950I to 2000IV on the U.S. money stock (M1), output (real GDP), and the price level (CPI_U). Consider a simple (extremely) model of money demand,[1]

$$\ln M1_t = \beta_1 + \beta_2 \ln GDP_t + \beta_3 \ln CPI_t + \varepsilon_t.$$

A plot of the least squares residuals is shown in Figure 20.1. The pattern in the residuals suggests that knowledge of the sign of a residual in one period is a good indicator of the sign of the residual in the next period. This knowledge suggests that the effect of a

FIGURE 20.1 Autocorrelated Least Squares Residuals.



[1]Because this chapter deals exclusively with time-series data, we shall use the index t for observations and T for the sample size throughout.

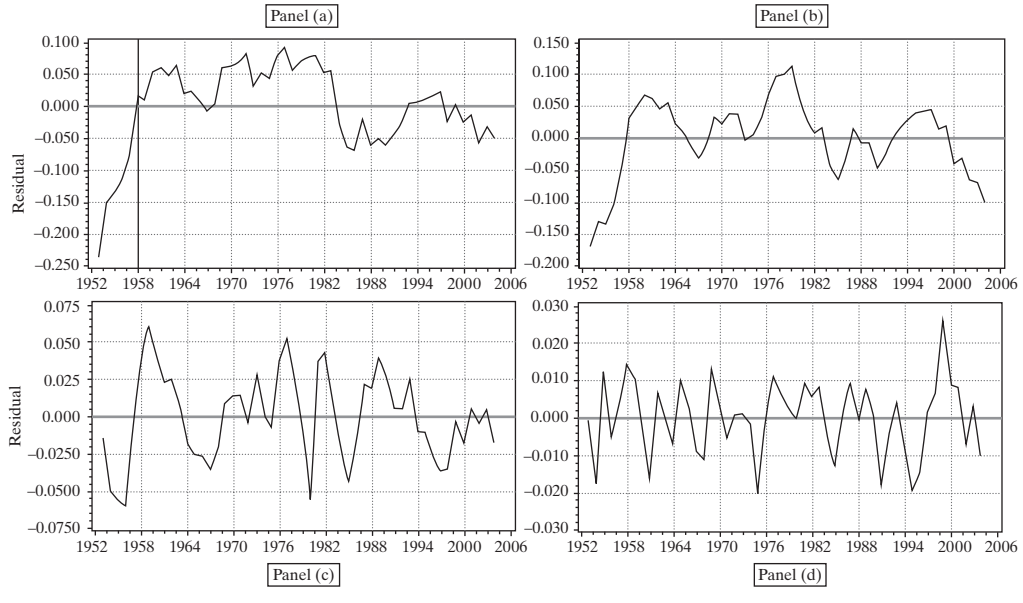
given disturbance is carried, at least in part, across periods. This sort of memory in the disturbances creates the long, slow swings from positive values to negative ones that are evident in Figure 20.1. One might argue that this pattern is the result of an obviously naïve model, but that is one of the important points in this discussion. Patterns such as this usually do not arise spontaneously; to a large extent, they are, indeed, a result of an incomplete or flawed model specification.

One explanation for autocorrelation is that relevant factors omitted from the time-series regression, like those included, are correlated across periods. This fact may be due to serial correlation in factors that should be in the regression model. It is easy to see why this situation would arise. Example 20.2 shows an obvious case.

Example 20.2 Autocorrelation Induced by Misspecification of the Model

In Examples 2.3, 4.2, 4.7, and 4.8, we examined yearly time-series data on the U.S. gasoline market from 1953 to 2004. The evidence in the examples was convincing that a regression model of variation in ln *G/Pop* should include, at a minimum, a constant, ln P_G and ln *Income/Pop* price variables and a time trend also provide significant explanatory power, but these two are a bare minimum. Moreover, we also found on the basis of a Chow test of structural change that apparently this market changed structurally after 1974. Figure 20.2 displays plots of four sets of least squares residuals. Parts (a) through (c) show clearly that as the specification of the regression is expanded, the autocorrelation in the "residuals" diminishes. Part (c) shows the effect of forcing the coefficients in the equation to be the same both before and after the structural shift. In part (d), the residuals in the two subperiods 1953 to 1974 and 1975 to 2004 are produced by separate unrestricted regressions. This latter set of residuals is almost nonautocorrelated. (*Note:* The range of variation of the residuals falls as the model is improved, i.e., as its fit improves.) The full equation is

FIGURE 20.2 Regression Residuals.



$$\ln\frac{G_t}{Pop_t} = \beta_1 + \beta_2 \ln P_{Gt} + \beta_3 \ln\frac{I_t}{Pop_t} + \beta_4 \ln P_{NCt} + \beta_5 \ln P_{UCt} + \beta_6 \ln P_{PTt} + \beta_7 \ln P_{Nt} + \beta_8 \ln P_{Dt} + \beta_9 \ln P_{St} + \beta_{10} t + \varepsilon_t.$$

Finally, we consider an example in which serial correlation is an anticipated part of the model.

Example 20.3 Negative Autocorrelation in the Phillips Curve

The Phillips curve [Phillips (1957)] has been one of the most intensively studied relationships in the macroeconomics literature. As originally proposed, the model specifies a negative relationship between wage inflation and unemployment in the United Kingdom over a period of 100 years. Recent research has documented a similar relationship between unemployment and price inflation. It is difficult to justify the model when cast in simple levels; labor market theories of the relationship rely on an uncomfortable proposition that markets persistently fall victim to money illusion, even when the inflation can be anticipated. Recent research[2] has reformulated a short-run (disequilibrium) "expectations augmented Phillips curve" in terms of unexpected inflation and unemployment that deviates from a long-run equilibrium or "natural rate." The **expectations-augmented Phillips curve** can be written as

$$\Delta p_t - E[\Delta p_t | \Psi_{t-1}] = \beta[u_t - u^*] + \varepsilon_t,$$

where Δp_t is the rate of inflation in year t, $E[\Delta p_t | \Psi_{t-1}]$ is the forecast of Δp_t made in period $t - 1$ based on information available at time $t - 1$, Ψ_{t-1}, u_t is the unemployment rate, and u^* is the natural, or equilibrium rate. (Whether u^* can be treated as an unchanging parameter, as we are about to do, is controversial.) By construction, $[u_t - u^*]$ is disequilibrium, or cyclical unemployment. In this formulation, ε_t would be the supply shock (i.e., the stimulus that produces the disequilibrium situation). To complete the model, we require a model for the expected inflation. For the present, we'll assume that economic agents are rank empiricists. The forecast of next year's inflation is simply this year's value. This produces the estimating equation,

$$\Delta p_t - \Delta p_{t-1} = \beta_1 + \beta_2 u_t + \varepsilon_t,$$

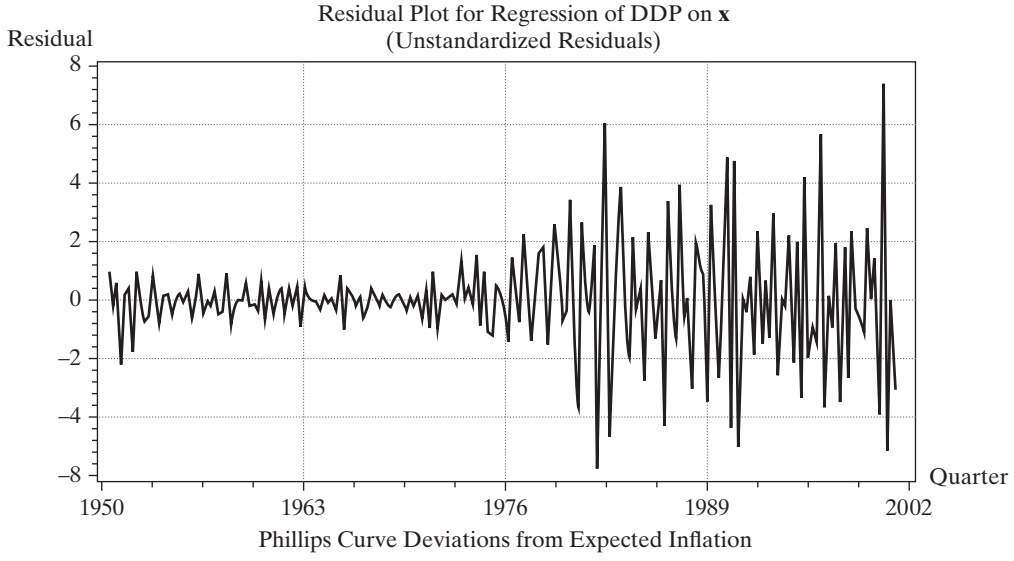
where $\beta_2 = \beta$ and $\beta_1 = -\beta u^*$. Note that there is an implied estimate of the natural rate of unemployment embedded in the equation. After estimation, u^* can be estimated by $-b_1/b_2$. The equation was estimated with the 1950.1 to 2000.4 data in Appendix Table F5.2 that were used in Example 20.1 (minus two quarters for the change in the rate of inflation). Least squares estimates (with standard errors in parentheses) are as follows:

$$\Delta p_t - \Delta p_{t-1} = 2.23567\ (0.49213) - 0.04155\ (0.08360)\ u_t + e_t,\ R^2 = 0.00123,\ T = 202.$$

The implied estimate of the natural rate of unemployment is 5.67 percent, which is in line with other estimates. The estimated asymptotic covariance of b_1 and b_2 is -0.03964. Using the delta method, we obtain a standard error of 3.17524 for this estimate, so a confidence interval for the natural rate is $5.67\% \pm 1.96(3.17\%) = (-0.55\%, 11.89\%)$. (This seems fairly wide, but, again, whether it is reasonable to treat this as a parameter is at least questionable). The regression of the least squares residuals on their past values gives a slope of -0.51843 with a highly significant t ratio of -8.48. We thus conclude that the residuals (and, apparently, the disturbances) in this model are highly negatively autocorrelated. This is consistent with the striking pattern in Figure 20.3.

[2]For example, Staiger et al. (1996).

FIGURE 20.3 Negatively Autocorrelated Residuals.



The problems for estimation and inference caused by autocorrelation are similar to (although, unfortunately, more involved than) those caused by heteroscedasticity. As before, least squares is inefficient, and inference based on the least squares estimates is adversely affected. Depending on the underlying process, however, GLS and FGLS estimators can be devised that circumvent these problems. There is one qualitative difference to be noted. In Section 20.10, we will examine models in which the generalized regression model can be viewed as an extension of the regression model to the conditional second moment of the dependent variable. In the case of autocorrelation, the phenomenon arises in almost all cases from a misspecification of the model. Views differ on how one should react to this failure of the classical assumptions, from a pragmatic one that treats it as another problem in the data to an orthodox methodological view that it represents a major specification issue.[3]

We should emphasize that the models we shall examine here are quite far removed from the classical regression. The exact or small-sample properties of the estimators are rarely known, and only their asymptotic properties have been derived.

20.2 THE ANALYSIS OF TIME-SERIES DATA

The treatment in this chapter will be the first structured analysis of time-series data in the text. Time-series analysis requires some revision of the interpretation of both data generation and sampling that we have maintained thus far.

[3]See, for example, "A Simple Message to Autocorrelation Correctors: Don't" [Mizon (1995)].

A time-series model will typically describe the path of a variable y_t in terms of contemporaneous (and perhaps lagged) factors $\mathbf{x}_t$, disturbances (**innovations**), ε_t, and its own past, $y_{t-1}, \ldots$. For example,

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \varepsilon_t.$$

The time series is a single occurrence of a random event. For example, the quarterly series on real output in the United States from 1950 to 2000 that we examined in Example 20.1 is a single realization of a process, GDP_t. The entire history over this period constitutes a realization of the process. At least in economics, the process could not be repeated. There is no counterpart to repeated sampling in a cross section or replication of an experiment involving a time-series process in physics or engineering. Nonetheless, were circumstances different at the end of World War II, the observed history *could* have been different. In principle, a completely different realization of the entire series might have occurred. The sequence of observations, $\{y_t\}_{t=-\infty}^{t=\infty}$, is a **time-series process**, which is characterized by its time ordering and its systematic correlation between observations in the sequence. The signature characteristic of a time-series process is that empirically, the data-generating mechanism produces exactly one realization of the sequence. Statistical results based on sampling characteristics concern not random sampling from a population, but from distributions of statistics constructed from sets of observations taken from this realization in a **time window**, $t = 1, \ldots, T$. Asymptotic distribution theory in this context concerns behavior of statistics constructed from an increasingly long window in this sequence.

The properties of y_t as a random variable in a cross section are straightforward and are conveniently summarized in a statement about its mean and variance or the probability distribution generating y_t. The statement is less obvious here. It is common to assume that innovations are generated independently from one period to the next, with the familiar assumptions

$$E[\varepsilon_t] = 0,$$
$$\text{Var}[\varepsilon_t] = \sigma_\varepsilon^2,$$

and

$$\text{Cov}[\varepsilon_t, \varepsilon_s] = 0 \quad \text{for } t \neq s.$$

In the current context, this distribution of ε_t is said to be **covariance stationary** or **weakly stationary**. Thus, although the substantive notion of random sampling must be extended for the time series ε_t, the mathematical results based on that notion apply here. It can be said, for example, that ε_t is generated by a time-series process whose mean and variance are not changing over time. As such, by the method we will discuss in this chapter, we could, at least in principle, obtain sample information and use it to characterize the distribution of ε_t. Could the same be said of y_t? There is an obvious difference between the series ε_t and y_t; observations on y_t at different points in time are necessarily correlated. Suppose that the y_t series *is* weakly stationary and that, for the moment, $\beta_2 = 0$. Then we could say that

$$E[y_t] = \beta_1 + \beta_3 E[y_{t-1}] + E[\varepsilon_t] = \beta_1/(1 - \beta_3)$$

and

$$\text{Var}[y_t] = \beta_3^2 \text{Var}[y_{t-1}] + \text{Var}[\varepsilon_t],$$

or

$$\gamma_0 = \beta_3^2\gamma_0 + \sigma_\varepsilon^2,$$

so that

$$\gamma_0 = \frac{\sigma_\varepsilon^2}{1 - \beta_3^2}.$$

Thus, γ_0, the variance of y_t, is a fixed characteristic of the process generating y_t. Note how the stationarity assumption, which apparently includes $|\beta_3| < 1$, has been used. The assumption that $|\beta_3| < 1$ is needed to ensure a finite and positive variance.[4] Finally, the same results can be obtained for nonzero β_2 if it is further assumed that x_t is a weakly stationary series.[5]

Alternatively, consider simply repeated substitution of lagged values into the expression for y_t,

$$y_t = \beta_1 + \beta_3(\beta_1 + \beta_3 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t, \tag{20-1}$$

and so on. We see that, in fact, the current y_t is an accumulation of the entire history of the innovations, ε_t. So if we wish to characterize the distribution of y_t, then we might do so in terms of sums of random variables. By continuing to substitute for y_{t-2}, then $y_{t-3}, \ldots$ in (20-1), we obtain an explicit representation of this idea,

$$y_t = \sum_{i=0}^{\infty} \beta_3^i(\beta_1 + \varepsilon_{t-i}).$$

Do sums that reach back into infinite past make any sense? We might view the process as having begun generating data at some remote, effectively infinite past. As long as distant observations become progressively less important, the extension to an infinite past is merely a mathematical convenience. The diminishing importance of past observations is implied by $|\beta_3| < 1$. Notice that, not coincidentally, this requirement is the same as that needed to solve for γ_0 in the preceding paragraphs. A second possibility is to assume that the *observation* of *this* time series begins at some time 0 [with (x_0, ε_0) called the *initial conditions*], by which time the underlying process has reached a state such that the mean and variance of y_t are not (or are no longer) changing over time. The mathematics is slightly different, but we are led to the same characterization of the random process generating y_t. In fact, the same weak stationarity assumption ensures both of them.

Except in very special cases, we would expect all the elements in the T component random vector $(y_1, \ldots, y_T)$ to be correlated. In this instance, said correlation is called *autocorrelation*. As such, the results pertaining to estimation with independent or uncorrelated observations that we used in the previous chapters are no longer usable. In point of fact, we have a sample of but one observation on the multivariate random variable $[y_t, t = 1, \ldots, T]$. There is a counterpart to the cross-sectional notion of parameter estimation, but only under assumptions (e.g., weak stationarity) that establish that parameters in the familiar sense even exist. Even with stationarity, it will emerge that for

[4]The current literature in macroeconometrics and time series analysis is dominated by analysis of cases in which $\beta_3 = 1$ (or counterparts in different models). We will return to this subject in Chapter 21.

[5]See Section 20.4.1 on the stationarity assumption.

estimation and inference, none of our earlier finite-sample results are usable. Consistency and asymptotic normality of estimators are somewhat more difficult to establish in time-series settings because results that require independent observations, such as the central limit theorems, are no longer usable. Nonetheless, counterparts to our earlier results have been established for most of the estimation problems we consider here.

20.3 DISTURBANCE PROCESSES

The preceding section has introduced a bit of the vocabulary and aspects of time-series specification. To obtain the theoretical results, we need to draw some conclusions about autocorrelation and add some details to that discussion.

20.3.1 CHARACTERISTICS OF DISTURBANCE PROCESSES

In the usual time-series setting, the disturbances are assumed to be homoscedastic but correlated across observations, so that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\boldsymbol{\Omega},$$

where $\sigma^2\boldsymbol{\Omega}$ is a full, positive definite matrix with a constant $\sigma^2 = \text{Var}[\varepsilon_t|\mathbf{X}]$ on the diagonal. As will be clear in the following discussion, we shall also assume that $\boldsymbol{\Omega}_{ts}$ is a function of $|t - s|$, but not of t or s alone, which is a **stationarity** assumption. (See the preceding section.) It implies that the covariance between observations t and s is a function only of $|t - s|$, the distance apart in time of the observations. Because σ^2 is not restricted, we normalize $\boldsymbol{\Omega}_{tt} = 1$. We define the **autocovariances,**

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s}|\mathbf{X}] = \text{Cov}[\varepsilon_{t+s}, \varepsilon_t|\mathbf{X}] = \sigma^2\boldsymbol{\Omega}_{t,t-s} = \gamma_s = \gamma_{-s}.$$

Note that $\sigma^2\boldsymbol{\Omega}_{tt} = \gamma_0$. The correlation between ε_t and ε_{t-s} is their autocorrelation,

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s}|\mathbf{X}] = \frac{\text{Cov}[\varepsilon_t, \varepsilon_{t-s}|\mathbf{X}]}{\sqrt{\text{Var}[\varepsilon_t|\mathbf{X}]\text{Var}[\varepsilon_{t-s}|\mathbf{X}]}} = \frac{\gamma_s}{\gamma_0} = \rho_s = \rho_{-s}.$$

We can then write

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \boldsymbol{\Gamma} = \gamma_0\mathbf{R},$$

where $\boldsymbol{\Gamma}$ is an **autocovariance matrix** and $\mathbf{R}$ is an **autocorrelation matrix**—the ts element is an **autocorrelation coefficient**,

$$\rho_s = \frac{\gamma_{|t-s|}}{\gamma_0}.$$

(*Note:* The matrix $\boldsymbol{\Gamma} = \gamma_0\mathbf{R}$ is the same as $\sigma^2\,\boldsymbol{\Omega}$.) We will usually use the abbreviation ρ_s to denote the autocorrelation between observations s periods apart.

Different types of processes imply different patterns in $\mathbf{R}$. For example, the most frequently analyzed process is a **first-order autoregression** or **AR(1)** process,

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

where u_t is a stationary, nonautocorrelated (**white noise**) process and ρ is a parameter. We will verify later that for this process, $\rho_s = \rho^s$. Higher-order **autoregressive processes** of the form

$$\varepsilon_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_p\varepsilon_{t-p} + u_t$$

imply more involved patterns, including, for some values of the parameters, cyclical behavior of the autocorrelations.[6] Stationary autoregressions are structured so that the influence of a given disturbance fades as it recedes into the more distant past but vanishes only asymptotically. For example, for the AR(1), $\text{Cov}[\varepsilon_t, \varepsilon_{t-s}]$ is never zero, but it does become negligible if $|\rho|$ is less than 1. **Moving-average processes**, conversely, have a short memory. For the MA(1) process,

$$\varepsilon_t = u_t - \lambda u_{t-1},$$

the memory in the process is only one period: $\gamma_0 = \sigma_u^2(1 + \lambda^2)$, $\gamma_1 = -\lambda\sigma_u^2$, but $\gamma_s = 0$ if $s > 1$.

Example 20.4 Autocorrelation Function for the Rate of Inflation

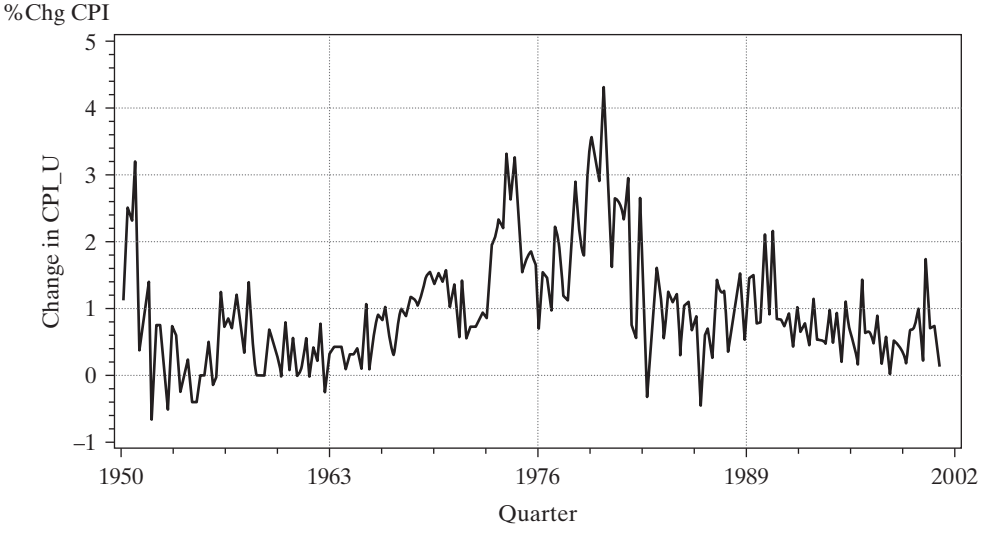
The autocorrelation function for a time series is a useful statistic for describing the nature of the underlying process. The function is computed as

$$ACF(s) = r_s = \frac{c_s}{c_0} = \frac{(1/(T - S))\Sigma_{t=s+1}^{T}(z_t - \bar{z})(z_{t-s} - \bar{z})}{(1/T)\Sigma_{t=s+1}^{T}(z_t - \bar{z})^2}, s = 1, \ldots.$$

The pattern of values of the *ACF* will help reveal the form of the time-series process. For an AR(1) process, the autocorrelations r_s will tend to appear like a geometric series, r^s. For a moving average series such as the MA(1), r_s will show one or a few significant values, then fall sharply to (approximately) zero. The characteristic pattern of an MA(1) process is $r_s = r$ for $s = 1$ and $r_s = 0$ for $s > 1$.

Figure 20.4 shows the quarterly percentage change in the U.S. Consumer Price Index from 1950 to 2000. (We will examine these data in some detail in Chapter 21.) The first 10 autocorrelations for this series are as follows:

FIGURE 20.4 Rate of Inflation in the Consumer Price Index.



[6] This model is considered in more detail in Section 20.9.2.

Lag	1	2	3	4	5	6	7	8	9	10
ACF	0.657	0.602	0.624	0.599	0.469	0.418	0.390	0.360	0.302	0.260

The persistence of the autocorrelations indicates a strongly autoregressive process.

20.3.2 AR(1) DISTURBANCES

Time-series processes such as the ones listed here can be characterized by their order, the values of their parameters, and the behavior of their autocorrelations.[7] We shall consider various forms at different points. The received empirical literature is overwhelmingly dominated by the AR(1) model, which is partly a matter of convenience. Processes more involved than this model are usually extremely difficult to analyze. There is, however, a more practical reason. It is very optimistic to expect to know precisely the correct form of the appropriate model for the disturbance in any given situation. The first-order autoregression has withstood the test of time and experimentation as a reasonable model for underlying processes that probably, in truth, are impenetrably complex. AR(1) works as a first pass—higher-order models are often constructed as a refinement.

The first-order autoregressive disturbance, or AR(1) process, is represented in the **autoregressive form** as

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \tag{20-2}$$

where

$$E[u_t|\mathbf{X}] = 0,$$
$$E[u_t^2|\mathbf{X}] = \sigma_u^2,$$

and

$$\text{Cov}[u_t, u_s|\mathbf{X}] = 0 \quad \text{if } t \neq s.$$

Because u_t is white noise, the conditional moments equal the unconditional moments. Thus $E[\varepsilon_t|\mathbf{X}] = E[\varepsilon_t]$ and so on.

By repeated substitution, we have

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots. \tag{20-3}$$

From the preceding **moving-average form**, it is evident that each disturbance ε_t embodies the entire past history of the u's, with the most recent observations receiving greater weight than those in the distant past. Depending on the sign of ρ, the series will exhibit clusters of positive and then negative observations or, if ρ is negative, regular oscillations of sign (as in Example 20.3).

Because the successive values of u_t are uncorrelated, the variance of ε_t is the variance of the right-hand side of (20-3):

$$\text{Var}[\varepsilon_t] = \sigma_u^2 + \rho^2\sigma_u^2 + \rho^4\sigma_u^2 + \cdots. \tag{20-4}$$

To proceed, a restriction must be placed on ρ,

$$|\rho| < 1, \tag{20-5}$$

[7] See Box and Jenkins (1984) for an authoritative study.

because otherwise, the right-hand side of (20-4) will become infinite. This result is the stationarity assumption discussed earlier. With (20-5), which implies that $\lim_{s\to\infty} \rho^s = 0$, $E[\varepsilon_t] = 0$ and

$$\text{Var}[\varepsilon_t] = \frac{\sigma_u^2}{1-\rho^2} = \sigma_\varepsilon^2. \tag{20-6}$$

With the stationarity assumption, there is an easier way to obtain the variance

$$\text{Var}[\varepsilon_t] = \rho^2 \,\text{Var}[\varepsilon_{t-1}] + \sigma_u^2$$

because $\text{Cov}[u_t, \varepsilon_s] = 0$ if $t > s$. With stationarity, $\text{Var}[\varepsilon_{t-1}] = \text{Var}[\varepsilon_t]$, which implies (20-6). Proceeding in the same fashion,

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-1}] = E[\varepsilon_t\varepsilon_{t-1}] = E[\varepsilon_{t-1}(\rho\varepsilon_{t-1} + u_t)] = \rho \,\text{Var}[\varepsilon_{t-1}] = \frac{\rho\sigma_u^2}{1-\rho^2}. \tag{20-7}$$

By repeated substitution in (20-2), we see that for any s,

$$\varepsilon_t = \rho^s\varepsilon_{t-s} + \sum_{i=0}^{s-1}\rho^i u_{t-i}$$

(e.g., $\varepsilon_t = \rho^3\varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t$). Therefore, because ε_s is not correlated with any u_t for which $t > s$ (i.e., any subsequent u_t), it follows that

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s}] = E[\varepsilon_t\varepsilon_{t-s}] = \frac{\rho^s\sigma_u^2}{1-\rho^2}. \tag{20-8}$$

Dividing by $\gamma_0 = \sigma_u^2/(1-\rho^2)$ provides the autocorrelations,

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s}] = \rho_s = \rho^s. \tag{20-9}$$

With the stationarity assumption, the autocorrelations fade over time. Depending on the sign of ρ, they will either be declining in geometric progression or alternating in sign if ρ is negative. Collecting terms, we have

$$\sigma^2\boldsymbol{\Omega} = \frac{\sigma_u^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \rho \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & \rho & 1 \end{bmatrix}. \tag{20-10}$$

20.4 SOME ASYMPTOTIC RESULTS FOR ANALYZING TIME-SERIES DATA

Because $\boldsymbol{\Omega}$ is not equal to $\mathbf{I}$, the now-familiar complications will arise in establishing the properties of estimators of $\boldsymbol{\beta}$, in particular of the least squares estimator. The finite sample properties of the OLS and GLS estimators remain intact. Least squares will continue to be unbiased. The earlier general proof allows for autocorrelated disturbances. The Aitken theorem (Theorem 9.4) and the distributional results for normally distributed disturbances can still be established conditionally on $\mathbf{X}$. (However, even these will be complicated when $\mathbf{X}$ contains lagged values of the dependent variable.) But finite

sample properties are of very limited usefulness in time-series contexts. Nearly all that can be said about estimators involving time-series data is based on their asymptotic properties.

As we saw in our analysis of heteroscedasticity, whether least squares is consistent or not depends on the matrices

$$\mathbf{Q}_T = (1/T)\mathbf{X}'\mathbf{X}$$

and

$$\mathbf{Q}_T^* = (1/T)\mathbf{X}'\mathbf{\Omega}\mathbf{X}.$$

In our earlier analyses, we were able to argue for convergence of $\mathbf{Q}_T$ to a positive definite matrix of constants, $\mathbf{Q}$, by invoking laws of large numbers. But these theorems assume that the observations in the sums are independent, which as suggested in Section 20.2, is surely not the case here. Thus, we require a different tool for this result. We can expand the matrix $\mathbf{Q}_T^*$ as

$$\mathbf{Q}_T^* = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\rho_{ts}\mathbf{x}_t\mathbf{x}_s', \tag{20-11}$$

where $\mathbf{x}_t'$ and $\mathbf{x}_s'$ are rows of $\mathbf{X}$ and ρ_{ts} is the autocorrelation between ε_t and ε_s. Sufficient conditions for this matrix to converge are that $\mathbf{Q}_T$ converge and that the correlations between disturbances diminish reasonably rapidly as the observations become further apart in time. For example, if the disturbances follow the AR(1) process described earlier, then $\rho_{ts} = \rho^{|t-s|}$ and if $\mathbf{x}_t$ is sufficiently well behaved, $\mathbf{Q}_T^*$ will converge to a positive definite matrix $\mathbf{Q}^*$ as $T \to \infty$. **Asymptotic normality** of the least squares and GLS estimators will depend on the behavior of sums such as

$$\sqrt{T}\overline{\mathbf{w}}_T = \sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right) = \sqrt{T}\left(\frac{1}{T}\mathbf{X}'\boldsymbol{\varepsilon}\right).$$

Asymptotic normality of least squares is difficult to establish for this general model. The central limit theorems we have relied on thus far do not extend to sums of *dependent* observations. The results of Amemiya (1985), Mann and Wald (1943), and Anderson (1971) do carry over to most of the familiar types of autocorrelated disturbances, including those that interest us here, so we shall ultimately conclude that ordinary least squares, GLS, and instrumental variables continue to be consistent and asymptotically normally distributed, and, in the case of OLS, inefficient. This section will provide a brief introduction to some of the underlying principles that are used to reach these conclusions.

20.4.1 CONVERGENCE OF MOMENTS—THE ERGODIC THEOREM

The discussion thus far has suggested (appropriately) that stationarity (or its absence) is an important characteristic of a process. The points at which we have encountered this notion concerned requirements that certain sums converge to finite values. In particular, for the AR(1) model, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, for the variance of the process to be finite, we require $|\rho| < 1$, which is a sufficient condition. However, this result is only a byproduct. Stationarity (at least, the weak stationarity we have examined) is only a characteristic of the sequence of moments of a distribution.

DEFINITION 20.1 Strong Stationarity
A time-series process, $\{z_t\}_{t=-\infty}^{t=\infty}$, is strongly stationary, or "stationary," if the joint probability distribution of any adjacent set of k observations in the sequence $[z_t, z_{t+1}, \ldots, z_{t+k-1}]$ is the same regardless of the origin, t , in the time scale.

For example, in (20-2), if we add $u_t \sim N[0, \sigma_u^2]$, then the resulting process, $\{\varepsilon_t\}_{t=-\infty}^{t=\infty}$, can easily be shown to be strongly stationary.

DEFINITION 20.2 Weak Stationarity
A time-series process, $\{z_t\}_{t=-\infty}^{t=\infty}$, is weakly stationary (or covariance stationary) if $E[z_t]$ is finite and is the same for all t and if the covariances between any two observations (labeled their autocovariance), $\text{Cov}[z_t, z_{t-k}]$, is a finite function only of model parameters and their distance apart in time, k, but not of the absolute location of either observation on the time scale.

Weak stationary is obviously implied by strong stationary, although it requires less because the distribution can, at least in principle, be changing on the time axis. The distinction is rarely necessary in applied work. In general, save for narrow theoretical examples, it will be difficult to come up with a process that is weakly but not strongly stationary. The reason for the distinction is that in much of our work, only weak stationary is required, and, as always, when possible, econometricians will dispense with unnecessary assumptions.

As we will discover shortly, stationarity is a crucial characteristic at this point in the analysis. If we are going to proceed to parameter estimation in this context, we will also require another characteristic of a time series, **ergodicity**. There are various ways to delineate this characteristic, none of them particularly intuitive. We borrow one definition from Davidson and MacKinnon (1993, p. 132) which comes close:

DEFINITION 20.3 Ergodicity
A strongly stationary time-series process, $\{z_t\}_{t=-\infty}^{t=\infty}$, is ergodic if for any two bounded functions that map vectors in the a and b dimensional real vector spaces to real scalars, f: $\boldsymbol{R}^a \rightarrow \boldsymbol{R}^1$ and g: $\boldsymbol{R}^b \rightarrow \boldsymbol{R}^1$,

$$\lim_{k \rightarrow \infty} |E[f(z_t, z_{t+1}, \ldots, z_{t+a-1})g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b-1})]|$$
$$= |E[f(z_t, z_{t+1}, \ldots, z_{t+a-1})]| \, |E[g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b-1})]|.$$

The definition states essentially that if events are separated far enough in time, then they are *asymptotically independent*. An implication is that in a time series, every observation will contain at least some unique information. Ergodicity is a crucial element of our

theory of estimation. When a time series has this property (with stationarity), then we can consider estimation of parameters in a meaningful sense.[8] The analysis relies heavily on the following theorem:

THEOREM 20.1 The Ergodic Theorem
If $\{z_t\}_{t=-\infty}^{t=\infty}$ is a time-series process that is strongly stationary and ergodic and $E[|z_t|]$ is a finite constant, and if $\bar{z}_T = (1/T)\sum_{t=1}^{T} z_t$, then $\bar{z}_T \xrightarrow{a.s.} \mu$, where $\mu = E[z_t]$. Note that the convergence is almost surely not in probability (which is implied) or in mean square (which is also implied). [See White (2001, p. 44) and Davidson and MacKinnon (1993, p. 133).]

What we have in the ergodic theorem is, for sums of dependent observations, a counterpart to the laws of large numbers that we have used at many points in the preceding chapters. Note, once again, the need for this extension is that to this point, our laws of large numbers have required sums of independent observations. But, in this context, by design, observations are distinctly not independent.

For this result to be useful, we will require an extension.

THEOREM 20.2 Ergodicity of Functions
If $\{z_t\}_{t=-\infty}^{t=\infty}$ is a time-series process that is strongly stationary and ergodic and if $y_t = f\{z_t\}$ is a measurable function in the probability space that defines z_t, then y_t is also stationary and ergodic. Let $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$ define a $K \times 1$ vector valued stochastic process—each element of the vector is an ergodic and stationary series, and the characteristics of ergodicity and stationarity apply to the joint distribution of the elements of $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$. Then, the ergodic theorem applies to functions of $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$.[9]

Theorem 20.2 produces the results we need to characterize the least squares (and other) estimators. In particular, by applying the assumptions of Theorem 20.2 to the data series, $[\mathbf{x}_t, \varepsilon_t]_{t=-\infty}^{t=\infty}$ we obtain that $y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$ is a stationary and ergodic process.

[8]Much of the analysis to follow will involve nonstationary series, which are the focus of most of the current literature—tests for nonstationarity largely dominate the recent study in time-series analysis. Ergodicity is a much more subtle and difficult concept. For any process that we will consider, ergodicity will have to be a given, at least at this level. A classic reference on the subject is Doob (1953). Another authoritative treatise is Billingsley (1995). White (2001) provides a concise analysis of many of these concepts as used in econometrics, and some useful commentary.

[9]See White (2001, pp. 44–45) for discussion.

By analyzing terms element by element we can use these results directly to assert that averages of $\mathbf{w}_t = \mathbf{x}_t\varepsilon_t$, $\mathbf{Q}_{tt} = \mathbf{x}_t\mathbf{x}_t'$, and $\mathbf{Q}_{tt}^* = \varepsilon_t^2\mathbf{x}_t\mathbf{x}_t'$ will converge to their population counterparts, $\mathbf{0}$, $\mathbf{Q}$ and $\mathbf{Q}^*$.

20.4.2 CONVERGENCE TO NORMALITY—A CENTRAL LIMIT THEOREM

To form a distribution theory for least squares, GLS, ML, and GMM, we will need a counterpart to the central limit theorem. In particular, we need to establish a large sample distribution theory for quantities of the form

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right) = \sqrt{T}\overline{\mathbf{w}}.$$

As noted earlier, we cannot invoke the familiar central limit theorems (Lindeberg–Levy, Lindeberg–Feller, Liapounov) because the observations in the sum are not independent. But, with the assumptions already made, we do have an alternative result. Some needed preliminaries are as follows:

DEFINITION 20.4 Martingale Sequence
A vector sequence $\mathbf{z}_t$ is a martingale sequence if $E[\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots] = \mathbf{z}_{t-1}$.

An important example of a martingale sequence is the **random walk**,

$$z_t = z_{t-1} + u_t,$$

where $\text{Cov}[u_t, u_s] = 0$ for all $t \neq s$. Then

$$E[z_t|z_{t-1}, z_{t-2}, \ldots] = E[z_{t-1}|z_{t-1}, z_{t-2}, \ldots] + E[u_t|z_{t-1}, z_{t-2}, \ldots] = z_{t-1} + 0 = z_{t-1}.$$

DEFINITION 20.5 Martingale Difference Sequence
A vector sequence $\mathbf{z}_t$ is a martingale difference sequence if $E[\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots] = \mathbf{0}$.

With Definition 20.5, we have the following broadly encompassing result:

THEOREM 20.3 Martingale Difference Central Limit Theorem
If $\mathbf{z}_t$ is a vector valued stationary and ergodic martingale difference sequence, with $E[\mathbf{z}_t\mathbf{z}_t'] = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a finite positive definite matrix, and if $\overline{\mathbf{z}}_T = (1/T)\sum_{t=1}^{T}\mathbf{z}_t$, then $\sqrt{T}\,\overline{\mathbf{z}}_T \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Sigma}]$. [For discussion, see Davidson and MacKinnon (1993, Sections. 4.7 and 4.8).][10]

[10]For convenience, we are bypassing a step in this discussion: establishing multivariate normality requires that the result first be established for the marginal normal distribution of each component, then that every linear combination of the variables also be normally distributed. (See Theorems D.17 and D.18A.) Our interest at this point is merely to collect the useful end results. Interested users may find the detailed discussions of the many subtleties and narrower points in White (2001) and Davidson and MacKinnon (1993, Chapter 4).

Theorem 20.3 is a generalization of the Lindeberg–Levy central limit theorem. It is not yet broad enough to cover cases of autocorrelation, but it does go beyond Lindeberg–Levy, for example, in extending to the GARCH model of Section 20.13.3.[11] But, looking ahead, this result encompasses what will be a very important application. Suppose in the classical linear regression model, $\{\mathbf{x}_t\}_{t=-\infty}^{t=\infty}$ is a stationary and ergodic multivariate stochastic process and $\{\varepsilon_t\}_{t=-\infty}^{t=\infty}$ is an i.i.d. process—that is, not autocorrelated and not heteroscedastic. Then, this is the most general case of the classical model that still maintains the assumptions about ε_t that we made in Chapter 2. In this case, the process $\{\mathbf{w}_t\}_{t=-\infty}^{t=\infty} = \{\mathbf{x}_t\varepsilon_t\}_{t=-\infty}^{t=\infty}$ is a martingale difference sequence, so that with sufficient assumptions on the moments of $\mathbf{x}_t$ we could use this result to establish consistency and asymptotic normality of the least squares estimator.[12]

We now consider a central limit theorem that is broad enough to include the case that interested us at the outset, stochastically dependent observations on $\mathbf{x}_t$ and autocorrelation in ε_t.[13] Suppose as before that $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$ is a stationary and ergodic stochastic process. We consider $\sqrt{T}\bar{\mathbf{z}}_T$. The following conditions are assumed:[14]

1. **Asymptotic uncorrelatedness:** $E[\mathbf{z}_t|\,\mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \ldots]$ converges in mean square to zero as $k \to \infty$. Note that is similar to the condition for ergodicity. White (2001) demonstrates that a (nonobvious) implication of this assumption is $E[\mathbf{z}_t] = \mathbf{0}$.
2. **Summability of autocovariances:** With dependent observations,

$$\lim_{T\to\infty} \text{Var}[\sqrt{T}\,\bar{\mathbf{z}}_T] = \sum_{t=1}^{\infty}\sum_{s=1}^{\infty} \text{Cov}[\mathbf{z}_t, \mathbf{z}_s'] = \sum_{k=-\infty}^{\infty} \boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}^*.$$

   To begin, we will need to assume that this matrix is finite, a condition called **summability**. Note this is the condition needed for convergence of $\mathbf{Q}_T^*$ in (20-11). If the sum is to be finite, then the $k = 0$ term must be finite, which gives us a necessary condition,

$$E[\mathbf{z}_t\mathbf{z}_t'] = \Gamma_0, \quad \text{a finite matrix.}$$

3. **Asymptotic negligibility of innovations:** Let

$$\mathbf{r}_{tk} = E[\mathbf{z}_t|\,\mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \ldots] - E[\mathbf{z}_t|\,\mathbf{z}_{t-k-1}, \mathbf{z}_{t-k-2}, \ldots].$$

An observation $\mathbf{z}_t$ may be viewed as the accumulated information that has entered the process since it began up to time t. Thus, it can be shown that

$$\mathbf{z}_t = \sum_{s=0}^{\infty} \mathbf{r}_{ts}.$$

The vector $\mathbf{r}_{tk}$ can be viewed as the information in this accumulated sum that entered the process at time $t - k$. The condition imposed on the process is that $\sum_{s=0}^{\infty}\sqrt{E[\mathbf{r}_{ts}'\mathbf{r}_{ts}]}$ be

[11]Forms of the theorem that surpass Lindeberg–Feller (D.19) and Liapounov (Theorem D.20) by allowing for different variances at each time, t, appear in Ruud (2000, p. 479) and White (2001, p. 133). These variants extend beyond our requirements in this treatment.

[12]See, for example, Hamilton (1994, pp. 208–212).

[13]Detailed analysis of this case is quite intricate and well beyond the scope of this book. Some fairly terse analysis may be found in White (2001, pp. 122–133) and Hayashi (2000).

[14]See Hayashi (2000, p. 405) who attributes the results to Gordin (1969).

finite. In words, condition 3 states that information eventually becomes negligible as it fades far back in time from the current observation. The AR(1) model (as usual) helps illustrate this point. If $z_t = \rho z_{t-1} + u_t$, then

$$\begin{aligned} r_{t0} &= E[z_t|\, z_t, z_{t-1}, \ldots] - E[z_t|\, z_{t-1}, z_{t-2}, \ldots] = z_t - \rho z_{t-1} = u_t, \\ r_{t1} &= E[z_t|\, z_{t-1}, z_{t-2} \ldots] - E[z_t|\, z_{t-2}, z_{t-3} \ldots] \\ &= E[\rho z_{t-1} + u_t|\, z_{t-1}, z_{t-2} \ldots] - E[\rho(\rho z_{t-2} + u_{t-1}) + u_t|\, z_{t-2}, z_{t-3}, \ldots] \\ &= \rho(z_{t-1} - \rho z_{t-2}) \\ &= \rho u_{t-1}. \end{aligned}$$

By a similar construction, $r_{tk} = \rho^k u_{t-k}$ from which it follows that $z_t = \sum_{s=0}^{\infty} \rho^s u_{t-s}$, which we saw earlier in (20-3). You can verify that if $|\rho| < 1$, the negligibility condition will be met.

THEOREM 20.4 Gordin's Central Limit Theorem
If $\underline{\mathbf{z}}_t$ is strongly stationary and ergodic and if conditions **1**–**3** *are met, then* $\sqrt{T}\,\bar{\mathbf{z}}_T \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Gamma}^*]$.

With all this machinery in place, we now have the theorem we will need. We will be able to employ these tools when we consider the least squares, IV, and GLS estimators in the discussion to follow.

20.5 LEAST SQUARES ESTIMATION

The least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T}\right).$$

Unbiasedness follows from the results in Chapter 4—no modification is needed. We know from Chapter 9 that the Gauss–Markov theorem has been lost—assuming it exists (that remains to be established), the GLS estimator is efficient, and OLS is not. How much information is lost by using least squares instead of GLS depends on the data. Broadly, least squares fares better in data that have long periods and little cyclical variation, such as aggregate output series. As might be expected, the greater the autocorrelation in $\boldsymbol{\varepsilon}$, the greater will be the benefit to using generalized least squares (when this is possible). Even if the disturbances are normally distributed, the usual F and t statistics do not have those distributions. So, not much remains of the finite sample properties we obtained in Chapter 4. The asymptotic properties remain to be established.

20.5.1 ASYMPTOTIC PROPERTIES OF LEAST SQUARES

The asymptotic properties of $\mathbf{b}$ are straightforward to establish given our earlier results. If we assume that the process generating $\mathbf{x}_t$ is stationary and ergodic, then by Theorems 20.1 and 20.2, $(1/T)(\mathbf{X}'\mathbf{X})$ converges to $\mathbf{Q}$ and we can apply the Slutsky theorem to the

inverse. If ε_t is not serially correlated, then $\mathbf{w}_t = \mathbf{x}_t\varepsilon_t$ is a martingale difference sequence, so $(1/T)(\mathbf{X}'\boldsymbol{\varepsilon})$ converges to zero. This establishes consistency for the simple case. On the other hand, if $[\mathbf{x}_t, \varepsilon_t]$ are jointly stationary and ergodic, then we can invoke the ergodic theorems 20.1 and 20.2 for both moment matrices and establish consistency. Asymptotic normality is a bit more subtle. For the case without serial correlation in ε_t, we can employ Theorem 20.3 for $\sqrt{T}\,\overline{\mathbf{w}}$. The involved case is the one that interested us at the outset of this discussion, that is, where there is autocorrelation in ε_t and dependence in $\mathbf{x}_t$. Theorem 20.4 is in place for this case. Once again, the conditions described in the preceding section must apply and, moreover, the assumptions needed will have to be established both for $\mathbf{x}_t$ and ε_t. Commentary on these cases may be found in Davidson and MacKinnon (1993), Hamilton (1994), White (2001), and Hayashi (2000). Formal presentation extends beyond the scope of this text, so at this point, we will proceed, and assume that the conditions underlying Theorem 20.4 are met. The results suggested here are quite general, albeit only sketched for the general case. For the remainder of our examination, at least in this chapter, we will confine attention to fairly simple processes in which the necessary conditions for the asymptotic distribution theory will be fairly evident.

There is an important exception to the results in the preceding paragraph. If the regression contains any lagged values of the dependent variable, then in most cases, least squares will no longer be unbiased or consistent. (We will examine the exceptions in Section 20.9.3.) To take the simplest case, suppose that

$$\begin{aligned} y_t &= \beta y_{t-1} + \varepsilon_t, \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t, \end{aligned} \tag{20-12}$$

and assume $|\beta| < 1$, $|\rho| < 1$. In this model, the regressor and the disturbance are correlated. There are various ways to approach the analysis. One useful way is to rearrange (20-12) by subtracting ρy_{t-1} from y_t. Then,

$$y_t = (\beta + \rho)y_{t-1} - \beta\rho y_{t-2} + u_t, \tag{20-13}$$

which is a classical regression with stochastic regressors. Because u_t is an innovation in period t, it is uncorrelated with both regressors, and least squares regression of y_t on (y_{t-1}, y_{t-2}) estimates $\rho_1 = (\beta + \rho)$ and $\rho_2 = -\beta\rho$. What is estimated by regression of y_t on y_{t-1} alone? Let $\gamma_k = \text{Cov}[y_t, y_{t-k}] = \text{Cov}[y_t, y_{t+k}]$. By stationarity, $\text{Var}[y_t] = \text{Var}[y_{t-1}]$, and $\text{Cov}[y_t, y_{t-1}] = \text{Cov}[y_{t-1}, y_{t-2}]$, and so on. These and (20-13) imply the following relationships:

$$\begin{aligned} \gamma_0 &= \rho_1\gamma_1 + \rho_2\gamma_2 + \sigma_u^2, \\ \gamma_1 &= \rho_1\gamma_0 + \rho_2\gamma_1, \\ \gamma_2 &= \rho_1\gamma_1 + \rho_2\gamma_0. \end{aligned} \tag{20-14}$$

(These are the **Yule–Walker equations** for this model.) The slope in the simple regression estimates γ_1/γ_0, which can be found in the solutions to these three equations. (An alternative approach is to use the left-out variable formula, which is a useful way to interpret this estimator.) In this case, we see that the slope in the short regression is an estimator of $(\beta + \rho) - \beta\rho(\gamma_1/\gamma_0)$. In either case, solving the three equations in (20-14) for γ_0, γ_1, and γ_2 in terms of ρ_1, ρ_2, and σ_u^2 produces

$$\text{plim } b = \frac{\beta + \rho}{1 + \beta\rho}. \tag{20-15}$$

This result is between β (when $\rho = 0$) and 1 (when both β and $\rho = 1$). Therefore, least squares is inconsistent unless ρ equals zero. The more general case that includes regressors, $\mathbf{x}_t$, involves more complicated algebra but gives essentially the same result. This is a general result; when the equation contains a lagged dependent variable in the presence of autocorrelation, OLS and GLS are inconsistent. The problem can be viewed as one of an omitted variable.

20.5.2 ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

As usual, $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is an inappropriate estimator of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$, both because s^2 is a biased estimator of σ^2 and because the matrix is incorrect. Generalities are scarce, but in general, for economic time series that are positively related to their past values, the standard errors conventionally *estimated* by least squares are likely to be too small. For slowly changing, trending aggregates such as output and consumption, this is probably the norm. For highly variable data such as inflation, exchange rates, and market returns, the situation is less clear. Nonetheless, as a general proposition, one would normally not want to rely on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ as an estimator of the asymptotic covariance matrix of the least squares estimator.

In view of this situation, if one is going to use least squares, then it is desirable to have an appropriate estimator of the covariance matrix of the least squares estimator. There are two approaches. If the form of the autocorrelation is known, then one can estimate the parameters of $\mathbf{\Omega}$ directly and compute a consistent estimator. Of course, if so, then it would be more sensible to use feasible generalized least squares instead and not waste the sample information on an inefficient estimator. The second approach parallels the use of the White estimator for heteroscedasticity.

The extension of White's result to the more general case of autocorrelation is much more difficult than in the heteroscedasticity case. The natural counterpart for estimating

$$\mathbf{Q}_* = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{x}_i\mathbf{x}_j' \qquad \textbf{(20-16)}$$

in (9-3) would be

$$\mathbf{Q}_* = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}e_t e_s \mathbf{x}_t\mathbf{x}_s'.$$

But there are two problems with this estimator, one theoretical and one practical.

Unlike the heteroscedasticity case, the matrix in (20-16) is $1/T$ times a sum of T^2 terms, so it is difficult to conclude yet that it will converge to anything at all. This application is most likely to arise in a time-series setting. To obtain convergence, it is necessary to assume that the terms involving unequal subscripts in (20-16) diminish in importance as T grows. A sufficient condition is that terms with subscript pairs $|t - s|$ grow smaller as the distance between them grows larger. In practical terms, observation pairs are progressively less correlated as their separation in time grows. Intuitively, if one can think of weights with the diagonal elements getting a weight of 1.0, then in the sum, the weights in the sum grow smaller as we move away from the diagonal. If we think of the sum of the weights rather than just the number of terms, then this sum falls off sufficiently rapidly that as n grows large, the sum is of order T rather than T^2. Thus, we achieve convergence of $\mathbf{Q}^*$ by assuming that the rows of $\mathbf{X}$ are well behaved and that the correlations diminish with increasing separation in time. (See Section 9.2. for a more formal statement of this condition.)

TABLE 20.1 Robust Covariance Estimation

Variable	*OLS Estimate*	*OLS SE*	*Corrected SE*
Constant	−1.6331	0.2286	0.3335
ln *Output*	0.2871	0.04738	0.07806
ln *CPI*	0.9718	0.03377	0.06585
$R^2 = 0.98952, r = 0.98762$			

The practical problem is that $\hat{\mathbf{Q}}_*$ need not be positive definite. Newey and West (1987a) have devised an estimator that overcomes this difficulty,

$$\hat{\mathbf{Q}}_* = \mathbf{S}_0 + \frac{1}{T}\sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-l}(\mathbf{x}_t\mathbf{x}_{t-l}' + \mathbf{x}_{t-l}\mathbf{x}_t'),$$

$$w_l = 1 - \frac{l}{(L+1)}. \tag{20-17}$$

[See (9-5).] [The weight in (20-17) is the Bartlett weight.] The **Newey–West autocorrelation consistent covariance estimator** is surprisingly simple and relatively easy to implement.[15] There is a final problem to be solved. It must be determined in advance how large L is to be. In general, there is little theoretical guidance. Current practice specifies $L \approx T^{1/4}$. Unfortunately, the result is not quite as crisp as that for the heteroscedasticity consistent estimator.

We have the result that $\mathbf{b}$ and $\mathbf{b}_{\text{IV}}$ are asymptotically normally distributed, and we have an appropriate estimator for the asymptotic covariance matrix. We have not specified the distribution of the disturbances, however. Thus, for inference purposes, the F statistic is approximate at best. Moreover, for more involved hypotheses, the likelihood ratio and Lagrange multiplier tests are unavailable. That leaves the Wald statistic, including asymptotic t ratios, as the main tool for statistical inference. We will examine a number of applications in the chapters to follow.

The White and Newey–West estimators are standard in the econometrics literature. We will encounter them at many points in the discussion to follow.

Example 20.5 Autocorrelation Consistent Covariance Estimation

For the model shown in Example 20.1, the regression results with the uncorrected standard errors and the Newey–West autocorrelation robust covariance matrix for lags of five quarters are shown in Table 20.1. The effect of the very high degree of autocorrelation is evident.

20.6 GMM ESTIMATION

The GMM estimator in the regression model with autocorrelated disturbances is produced by the empirical moment equations,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_{GMM}) = \frac{1}{T}\mathbf{X}'\hat{\boldsymbol{\varepsilon}}(\hat{\boldsymbol{\beta}}_{GMM}) = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}) = \mathbf{0}. \tag{20-18}$$

[15]Both estimators are now standard features in modern econometrics computer programs. Further results on different weighting schemes may be found in Hayashi (2000, pp. 406–410).

The estimator is obtained by minimizing

$$q = \bar{\mathbf{m}}'(\hat{\boldsymbol{\beta}}_{GMM})\mathbf{W}\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM})$$

where $\mathbf{W}$ is a positive definite weighting matrix. The optimal weighting matrix would be

$$\mathbf{W} = \{\text{Asy. Var}[\sqrt{T}\,\bar{\mathbf{m}}(\boldsymbol{\beta})]\}^{-1},$$

which is the inverse of

$$\text{Asy. Var}\,[\sqrt{T}\,\bar{\mathbf{m}}(\boldsymbol{\beta})] = \text{Asy. Var}\left[\frac{1}{\sqrt{T}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right] = \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\sigma^2\rho_{ts}\mathbf{x}_t\mathbf{x}_s' = \sigma^2\mathbf{Q}^*.$$

The optimal weighting matrix would be $[\sigma^2\mathbf{Q}^*]^{-1}$. As in the heteroscedasticity case, this minimization problem is an exactly identified case, so, the weighting matrix is actually irrelevant to the solution. *The GMM estimator for the regression model with autocorrelated disturbances is ordinary least squares*. We can use the results in Section 20.5.2 to construct the asymptotic covariance matrix. We will require the assumptions in Section 20.4 to obtain convergence of the moments and asymptotic normality. We will wish to extend this simple result in one instance. In the common case in which $\mathbf{x}_t$ contains lagged values of y_t, we will want to use an instrumental variable estimator. We will return to that estimation problem in Section 20.9.3.

20.7 TESTING FOR AUTOCORRELATION

The available tests for autocorrelation are based on the principle that if the true disturbances are autocorrelated, then this fact can be detected through the autocorrelations of the least squares residuals. The simplest indicator is the slope in the artificial regression

$$\begin{aligned} e_t &= re_{t-1} + v_t, \\ e_t &= y_t - \mathbf{x}_t'\mathbf{b}, \\ r &= \left(\sum_{t=2}^{T} e_t e_{t-1}\right) \Big/ \left(\sum_{t=1}^{T-1} e_t^2\right). \end{aligned} \qquad \textbf{(20-19)}$$

If there is autocorrelation, then the slope in this regression will be an estimator of $\rho = \text{Corr}[\varepsilon_t, \varepsilon_{t-1}]$. The complication in the analysis lies in determining a formal means of evaluating when the estimator is *large*, that is, on what statistical basis to reject the null hypothesis that ρ equals zero. As a first approximation, treating (20-19) as a classical linear model and using a t or F (squared t) test to test the hypothesis is a valid way to proceed based on the Lagrange multiplier principle. We used this device in Example 20.3. The tests we consider here are refinements of this approach.

20.7.1 LAGRANGE MULTIPLIER TEST

The Breusch (1978)–Godfrey (1978) test is a Lagrange multiplier test of H_0: no autocorrelation versus H_1: $\varepsilon_t = \text{AR}(P)$ or $\varepsilon_t = \text{MA}(P)$. The same test is used for either structure. The test statistic is

$$LM = T\left(\frac{\mathbf{e}'\mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{e}}{\mathbf{e}'\mathbf{e}}\right) = TR_0^2, \qquad \textbf{(20-20)}$$

where $\mathbf{X}_0$ is the original $\mathbf{X}$ matrix augmented by P additional columns containing the lagged OLS residuals, $e_{t-1}, \ldots, e_{t-P}$. The test can be carried out simply by regressing the ordinary least squares residuals e_t on $\mathbf{x}_{t0}$ (filling in missing values for lagged residuals with zeros) and referring TR_0^2 to the tabled critical value for the chi-squared distribution with P degrees of freedom.[16] Because $\mathbf{X}'\mathbf{e} = \mathbf{0}$, the test is equivalent to regressing e_t on the part of the lagged residuals that is unexplained by $\mathbf{X}$. There is therefore a compelling logic to it; if any fit is found, then it is due to correlation between the current and lagged residuals. The test is a joint test of the first P autocorrelations of ε_t, not just the first.

Example 20.6 Test for Autocorrelation

For the model shown in Examples 20.1 and 20.4, the regression of the least squares residuals on a constant, ln*GDP*, ln*CPI* and two lagged values of the residuals (with initial values filled with zeros) produces $R^2 = 0.97632$. With $T = 204$, the Lagrange multiplier statistic is 199.17. The critical value from the chi-squared table for 2 degrees of freedom is 5.99. The hypothesis that there is no second (or greater) degree autocorrelation is rejected.

20.7.2 BOX AND PIERCE'S TEST AND LJUNG'S REFINEMENT

An alternative test that is asymptotically equivalent to the LM test when the null hypothesis, $\rho = 0$, is true and when $\mathbf{X}$ does not contain lagged values of y is due to Box and Pierce (1970). The Q **test** is carried out by referring

$$Q = T\sum_{j=1}^{P} r_j^2, \tag{20-21}$$

where $r_j = \left(\sum_{t=j+1}^{T} e_t e_{t-j}\right) \Big/ \left(\sum_{t=1}^{T} e_t^2\right)$, to the critical values of the chi-squared table with P degrees of freedom. A refinement suggested by Ljung and Box (1979) is

$$Q' = T(T+2)\sum_{j=1}^{P} \frac{r_j^2}{T-j}. \tag{20-22}$$

The essential difference between the Godfrey–Breusch and the Box–Pierce tests is the use of partial correlations (controlling for $\mathbf{X}$ and the other variables) in the former and simple correlations in the latter. Under the null hypothesis, there is no autocorrelation in ε_t, and no correlation between $\mathbf{x}_t$ and ε_s in any event, so the two tests are asymptotically equivalent. On the other hand, because it does not condition on $\mathbf{x}_t$, the Box–Pierce test is less powerful than the LM test when the null hypothesis is false, as intuition might suggest.

20.7.3 THE DURBIN–WATSON TEST

The Durbin–Watson statistic[17] was the first formal procedure developed for testing for autocorrelation using the least squares residuals. The test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = 2(1 - r) - \frac{e_1^2 + e_T^2}{\sum_{t=1}^{T} e_t^2}, \tag{20-23}$$

[16] A warning to practitioners: Current software varies on whether the lagged residuals are filled with zeros or the first P observations are simply dropped when computing this statistic. In the interest of replicability, users should determine which is the case before reporting results.

[17] Durbin and Watson (1950, 1951, 1971).

where r is the same first-order autocorrelation that underlies the preceding two statistics. If the sample is reasonably large, then the last term will be negligible, leaving $d \approx 2(1 - r)$. The statistic takes this form because the authors were able to determine the exact distribution of this transformation of the autocorrelation and could provide tables of critical values for specific values of T and K. The one-sided test for H_0: $\rho = 0$ against H_1: $\rho > 0$ is carried out by comparing d to values $d_L(T, K)$ and $d_U(T, K)$. If $d < d_L$, the null hypothesis is rejected; if $d > d_U$, the hypothesis is not rejected. If d lies between d_L and d_U, then no conclusion is drawn.

20.7.4 TESTING IN THE PRESENCE OF A LAGGED DEPENDENT VARIABLE

The Durbin–Watson test is not likely to be valid when there is a lagged dependent variable in the equation.[18] The statistic will usually be biased toward a finding of no autocorrelation. Three alternatives have been devised. The LM and Q tests can be used whether or not the regression contains a lagged dependent variable. (In the absence of a lagged dependent variable, they are asymptotically equivalent.) As an alternative to the standard test, Durbin (1970) derived a Lagrange multiplier test that is appropriate in the presence of a lagged dependent variable. The test may be carried out by referring

$$h = r\sqrt{T/(1 - Ts_c^2)}, \qquad \textbf{(20-24)}$$

where s_c^2 is the estimated variance of the least squares regression coefficient on y_{t-1}, to the standard normal tables. Large values of h lead to rejection of H_0. The test has the virtues that it can be used even if the regression contains additional lags of y_t, and it can be computed using the standard results from the initial regression without any further regressions. If $s_c^2 > 1/T$, however, then it cannot be computed. An alternative is to regress e_t on $\mathbf{x}_t, y_{t-1}, \ldots, e_{t-1}$, and any additional lags that are appropriate for e_t and then to test the joint significance of the coefficient(s) on the lagged residual(s) with the standard F test. This method is a minor modification of the Breusch–Godfrey test. Under H_0, the coefficients on the remaining variables will be zero, so the tests are the same asymptotically.

20.7.5 SUMMARY OF TESTING PROCEDURES

The preceding has examined several testing procedures for locating autocorrelation in the disturbances. In all cases, the procedure examines the least squares residuals. We can summarize the procedures as follows:

LM test. $LM = TR^2$ in a regression of the least squares residuals on $[\mathbf{x}_t, e_{t-1}, \ldots e_{t-P}]$. Reject H_0 if $LM > \chi_*^2[P]$. This test examines the covariance of the residuals with lagged values, controlling for the intervening effect of the independent variables.

Q test. $Q = T(T + 2)\sum_{j=1}^{P} r_j^2/(T - j)$. Reject H_0 if $Q > \chi_*^2[P]$. This test examines the raw correlations between the residuals and P lagged values of the residuals.

Durbin–Watson test. $d = 2(1 - r)$. Reject H_0: $\rho = 0$ if $d < d_L^*$. This test looks directly at the first-order autocorrelation of the residuals.

Durbin's test. $F_D =$ the F statistic for the joint significance of P lags of the residuals in the regression of the least squares residuals on $[\mathbf{x}_t, y_{t-1}, \ldots y_{t-R}, e_{t-1}, \ldots e_{t-P}]$. Reject H_0 if $F_D > F_*[P, T - K - P]$. This test examines the partial correlations between the residuals and the lagged residuals, controlling for the intervening effect of the independent variables and the lagged dependent variable.

[18] This issue has been studied by Nerlove and Wallis (1966), Durbin (1970), and Dezhbaksh (1990).

The Durbin–Watson test has some major shortcomings. The inconclusive region is large if T is small or moderate. The bounding distributions, while free of the parameters $\boldsymbol{\beta}$ and σ, do depend on the data (and assume that $\mathbf{X}$ is nonstochastic). An exact version based on an algorithm developed by Imhof (1980) avoids the inconclusive region, but is rarely used. The LM and Box–Pierce statistics do not share these shortcomings—their limiting distributions are chi squared independently of the data and the parameters. For this reason, the LM test has become the standard method in applied research.

20.8 EFFICIENT ESTIMATION WHEN $\boldsymbol{\Omega}$ IS KNOWN

As a prelude to deriving feasible estimators for $\boldsymbol{\beta}$ in this model, we consider full generalized least squares estimation assuming that $\boldsymbol{\Omega}$ is known. In the next section, we will turn to the more realistic case in which $\boldsymbol{\Omega}$ must be estimated as well.

If the parameters of $\boldsymbol{\Omega}$ are known, then the GLS estimator,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}), \tag{20-25}$$

and the estimate of its sampling variance,

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}_\varepsilon^2[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}, \tag{20-26}$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{T} \tag{20-27}$$

can be computed in one step. For the AR(1) case, data for the transformed model are

$$\mathbf{y}_* = \begin{bmatrix} \sqrt{1-\rho^2}\,y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}, \qquad \mathbf{X}_* = \begin{bmatrix} \sqrt{1-\rho^2}\,\mathbf{x}_1 \\ \mathbf{x}_2 - \rho \mathbf{x}_1 \\ \mathbf{x}_3 - \rho \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_T - \rho \mathbf{x}_{T-1} \end{bmatrix}. \tag{20-28}$$

These transformations are variously labeled **partial differences, quasi differences**, or **pseudo-differences**. Note that in the transformed model, every observation except the first contains a constant term. What was the column of 1s in $\mathbf{X}$ is transformed to $[(1-\rho^2)^{1/2}, (1-\rho), (1-\rho), \ldots]$. Therefore, if the sample is relatively small, then the problems with measures of fit noted in Section 3.5 will reappear.

The variance of the transformed disturbance is

$$\text{Var}[\varepsilon_t - \rho\varepsilon_{t-1}] = \text{Var}[u_t] = \sigma_u^2.$$

The variance of the first disturbance is also σ_u^2; [see (20-6)]. This can be estimated using $(1-\rho^2)\hat{\sigma}_\varepsilon^2$.

Corresponding results have been derived for higher-order autoregressive processes. For the AR(2) model,

$$\varepsilon_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + u_t, \tag{20-29}$$

the transformed data for generalized least squares are obtained by

$$\mathbf{z}_{*1} = \left[\frac{(1+\theta_2)[(1-\theta_2)^2-\theta_1^2]}{1-\theta_2}\right]^{1/2}\mathbf{z}_1,$$

$$\mathbf{z}_{*2} = (1-\theta_2^2)^{1/2}\mathbf{z}_2 - \frac{\theta_1(1-\theta_1^2)^{1/2}}{1-\theta_2}\mathbf{z}_1,$$

$$\mathbf{z}_{*t} = \mathbf{z}_t - \theta_1\mathbf{z}_{t-1} - \theta_2\mathbf{z}_{t-2}, \quad t > 2, \tag{20-30}$$

where $\mathbf{z}_t$ is used for y_t or $\mathbf{x}_t$. The transformation becomes progressively more complex for higher-order processes.[19]

Note that in both the AR(1) and AR(2) models, the transformation to $\mathbf{y}_*$ and $\mathbf{X}_*$ involves starting values for the processes that depend only on the first one or two observations. We can view the process as having begun in the infinite past. Because the sample contains only T observations, however, it is convenient to treat the first one or two (or P) observations as shown and consider them as initial values. Whether we view the process as having begun at time $t = 1$ or in the infinite past is ultimately immaterial in regard to the asymptotic properties of the estimators.

The asymptotic properties for the GLS estimator are quite straightforward given the apparatus we assembled in Section 20.4. We begin by assuming that $\{\mathbf{x}_t, \varepsilon_t\}$ are jointly an ergodic, stationary process. Then, after the GLS transformation, $\{\mathbf{x}_{*t}, \varepsilon_{*t}\}$ is also stationary and ergodic. Moreover, ε_{*t} is nonautocorrelated by construction. In the transformed model, then, $\{\mathbf{w}_{*t}\} = \{\mathbf{x}_{*t}\varepsilon_{*t}\}$ is a stationary and ergodic martingale difference sequence. We can use the ergodic theorem to establish consistency and the central limit theorem for martingale difference sequences to establish asymptotic normality for GLS in this model. Formal arrangement of the relevant results is left as an exercise.

20.9 ESTIMATION WHEN $\boldsymbol{\Omega}$ IS UNKNOWN

For an unknown $\boldsymbol{\Omega}$, there are a variety of approaches. Any consistent estimator of $\boldsymbol{\Omega}(\rho)$ will suffice—recall from Theorem 9.5 in Section 9.4.2, all that is needed for efficient estimation of $\boldsymbol{\beta}$ is a consistent estimator of $\boldsymbol{\Omega}(\rho)$. The complication arises, as might be expected, in estimating the autocorrelation parameter(s).

20.9.1 AR(1) DISTURBANCES

The AR(1) model is the one most widely used and studied. The most common procedure is to begin FGLS with a natural estimator of ρ, the autocorrelation of the residuals. Because $\mathbf{b}$ is consistent, we can use r. Others that have been suggested include Theil's (1971) estimator, $r[(T-K)/(T-1)]$ and Durbin's (1970), the slope on y_{t-1} in a regression of y_t on y_{t-1}, $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$. The second step is FGLS based on (20-25)–(20-28). This is the **Prais and Winsten** (1954) **estimator**. The **Cochrane and Orcutt** (1949) **estimator** (based on computational ease) omits the first observation.

It is possible to iterate any of these estimators to convergence. Because the estimator is asymptotically efficient at every iteration, nothing is gained by doing so. Unlike the heteroscedastic model, iterating when there is autocorrelation does not produce the

[19]See Box and Jenkins (1984) and Fuller (1976).

maximum likelihood estimator. The iterated FGLS estimator, regardless of the estimator of ρ, does not account for the term $(1/2)\ln(1-\rho^2)$ in the log-likelihood function [see the following (20-31)].

Maximum likelihood estimators can be obtained by maximizing the log likelihood with respect to $\boldsymbol{\beta}$, σ_u^2, and ρ. The log-likelihood function may be written

$$\ln L = -\frac{\sum_{t=1}^{T} u_t^2}{2\sigma_u^2} + \frac{1}{2}\ln(1-\rho^2) - \frac{T}{2}(\ln 2\pi + \ln \sigma_u^2), \tag{20-31}$$

where, as before, the first observation is computed differently from the others using (20-28). Based on the MLE, the standard approximations to the asymptotic variances of the estimators are

$$\begin{aligned}
\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}_{ML}] &= \hat{\sigma}^2_{\varepsilon,ML}[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}_{ML}\mathbf{X}]^{-1},\\
\text{Est.Asy.Var}[\hat{\sigma}^2_{u,ML}] &= 2\hat{\sigma}^4_{u,ML}/T,\\
\text{Est.Asy.Var}[\hat{\rho}_{ML}] &= (1-\hat{\rho}^2_{ML})/T.
\end{aligned} \tag{20-32}$$

All the foregoing estimators have the same asymptotic properties. The available evidence on their small-sample properties comes from Monte Carlo studies and is, unfortunately, only suggestive. Griliches and Rao (1969) find evidence that if the sample is relatively small and ρ is not particularly large, say, less than 0.3, then least squares is as good as or better than FGLS. The problem is the additional variation introduced into the sampling variance by the variance of r. Beyond these, the results are rather mixed. Maximum likelihood seems to perform well in general, but the Prais–Winsten estimator is evidently nearly as efficient. Both estimators have been incorporated in all contemporary software. In practice, the Prais and Winsten (1954) and Beach and MacKinnon (1978a) maximum likelihood estimators are probably the most common choices.

20.9.2 APPLICATION: ESTIMATION OF A MODEL WITH AUTOCORRELATION

The model of the U.S. gasoline market that appears in Example 6.20 is

$$\ln\left(\frac{G}{Pop}\right)_t = \beta_1 + \beta_2 \ln\left(\frac{Income}{Pop}\right)_t + \beta_3 \ln P_{G,t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \beta_6 t + \varepsilon_t.$$

The results in Figure 20.2 suggest that the specification may be incomplete, and, if so, there may be autocorrelation in the disturbances in this specification. Least squares estimates of the parameters using the data in Appendix Table F2.2 appear in the first row of Table 20.2. [The dependent variable is ln (*Gas expenditure*/(*price* × *population*)). These are the OLS results reported in Example 6.20.] The first five autocorrelations of the least squares residuals are 0.667, 0.438, 0.142, −0.018, and −0.198. This produces Box–Pierce and Box–Ljung statistics of 36.217 and 38.789, respectively, both of which are larger than the critical value from the chi-squared table of 11.07. We regressed the least squares residuals on the independent variables and five lags of the residuals. (The missing values in the first five years were filled with zeros.) The coefficients on the lagged residuals and the associated t statistics are 0.741(4.635), 0.153(0.789), −0.246(−1.262), 0.0942(0.472), and −0.125(−0.658). The R^2 in this regression is 0.549086, which produces a chi-squared value of 28.55. This is larger than the critical value of 11.07, so once again,

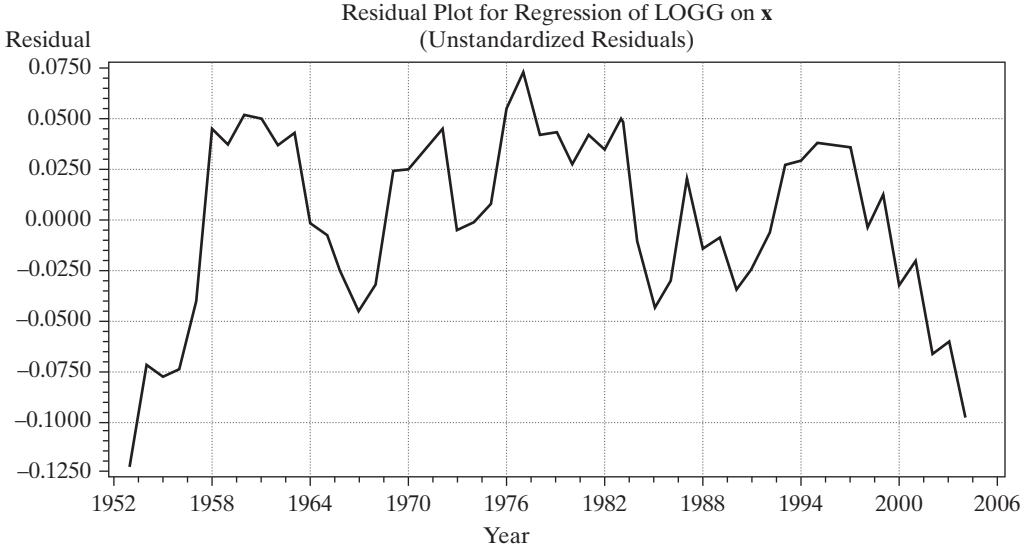
TABLE 20.2 Parameter Estimates (Standard errors in parentheses)

	β_1	β_2	β_3	β_4	β_5	β_6	ρ
OLS	−26.68	1.6250	−0.05392	−0.0834	−0.08467	−0.01393	0.0000
$R^2 = 0.96493$	(2.000)	(0.1952)	(0.04216)	(0.1765)	(0.1024)	(0.00477)	(0.0000)
Prais–	−18.58	0.7447	−0.1138	−0.1364	−0.08956	0.006689	0.9567
Winsten	(1.768)	(0.1761)	(0.03689)	(0.1528)	(0.07213)	(0.004974)	(0.04078)
Cochrane–	−18.76	0.7300	−0.1080	−0.06675	0.04190	−0.0001653	0.9695
Orcutt	(1.382)	(0.1377)	(0.02885)	(0.1201)	(0.05713)	(0.004082)	(0.03434)
Maximum	−16.25	0.4690	−0.1387	−0.09682	−0.001485	0.01280	0.9792
Likelihood	(1.391)	(0.1350)	(0.02794)	(0.1270)	(0.05198)	(0.004427)	(0.02816)
AR(2)	−19.45	0.8116	−0.09538	−0.09099	0.04091	−0.001374	0.8610
	(1.495)	(0.1502)	(0.03117)	(0.1297)	(0.06558)	(0.004227)	(0.07053)

the null hypothesis of zero autocorrelation is rejected. The plot of the residuals shown in Figure 20.5 seems consistent with this conclusion.

The Prais and Winsten FGLS estimates appear in the second row of Table 20.2, followed by the Cochrane and Orcutt results, then the maximum likelihood estimates. [The autocorrelation coefficient computed using $(1 - d/2)$ (see Section 20.7.3) is 0.78750. The MLE is computed using the Beach and MacKinnon algorithm.] Finally, we fit the AR(2) model by first regressing the least squares residuals, e_t, on e_{t-1} and e_{t-2} (without a constant term and filling the first two observations with zeros). The two estimates are 0.751941 and −0.022464, respectively. With the estimates of θ_1 and θ_2, we transformed the data using $y_t^* = y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2}$ and likewise for each regressor. Two observations are then discarded, so the AR(2) regression uses 50 observations while

FIGURE 20.5 Least Squares Residuals.



the Prais–Winsten estimator uses 52 and the Cochrane–Orcutt regression uses 51. In each case, the autocorrelation of the FGLS residuals is computed and reported in the last column of the table.

One might want to examine the residuals after estimation to ascertain whether the AR(1) model is appropriate. In the results just presented, there are two large autocorrelation coefficients listed with the residual-based tests, and in computing the LM statistic, we found that the first two coefficients were statistically significant. If the AR(1) model is appropriate, then one should find that only the coefficient on the first lagged residual is statistically significant in this auxiliary, second-step regression. Another indicator is provided by the FGLS residuals themselves. After computing the FGLS regression, the estimated residuals,

$$\hat{\varepsilon} = y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}},$$

will still be autocorrelated. In our results using the Prais–Winsten estimates, the autocorrelation of the FGLS residuals is 0.957. This is to be expected. However, if the model is correct, then the transformed residuals,

$$\hat{u}_t = \hat{\varepsilon}_t - \hat{\rho}\hat{\varepsilon}_{t-1},$$

should be at least close to nonautocorrelated. But, for our data, the autocorrelation of these adjusted residuals is only 0.292. It appears on this basis that, in fact, the AR(1) model has largely completed the specification.

20.9.3 ESTIMATION WITH A LAGGED DEPENDENT VARIABLE

In Section 20.5.1, we encountered the problem of estimation by least squares when the model contains both autocorrelation and lagged dependent variable(s). Because the OLS estimator is inconsistent, the residuals on which an estimator of ρ would be based are likewise inconsistent. Therefore, $\hat{\rho}$ will be inconsistent as well. The consequence is that the FGLS estimators described earlier are not usable in this case. There is, however, an alternative way to proceed, based on the method of instrumental variables. The method of instrumental variables was introduced in Section 8.3.2. To review, the general problem is that in the regression model, if

$$\text{plim}(1/T)\mathbf{X}'\boldsymbol{\varepsilon} \neq \mathbf{0},$$

then the least squares estimator is not consistent. A consistent estimator is

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y}),$$

where $\mathbf{Z}$ is a set of K variables chosen such that $\text{plim}(1/T)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}$ but $\text{plim}(1/T)\mathbf{Z}'\mathbf{X} \neq \mathbf{0}$. For the purpose of consistency only, any such set of instrumental variables will suffice. The relevance of that here is that the obstacle to consistent FGLS is, at least for the present, the lack of a consistent estimator of ρ. By using the technique of instrumental variables, we may estimate $\boldsymbol{\beta}$ consistently, then estimate ρ and proceed.

Hatanaka (1974, 1976) has devised an efficient two-step estimator based on this principle. To put the estimator in the current context, we consider estimation of the model

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \gamma y_{t-1} + \varepsilon_t,$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t.$$

To get to the second step of FGLS, we require a consistent estimator of the slope parameters. These estimates can be obtained using an IV estimator, where the column of $\mathbf{Z}$ corresponding to y_{t-1} is the only one that need be different from that of $\mathbf{X}$. An appropriate instrument can be obtained by using the fitted values in the regression of y_t on $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$. The residuals from the IV regression are then used to construct

$$\hat{\rho} = \frac{\sum_{t=3}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=3}^{T} \hat{\varepsilon}_t^2}, \tag{20-33}$$

where

$$\hat{\varepsilon}_t = y_t - \mathbf{b}_{\mathrm{IV}}'\mathbf{x}_t - c_{\mathrm{IV}} y_{t-1}.$$

FGLS estimates may now be computed by regressing $y_{*_t} = y_t - \hat{\rho} y_{t-1}$ on

$$\begin{aligned}
\mathbf{x}_{*_t} &= \mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1}, \\
y_{*_{t-1}} &= y_{t-1} - \hat{\rho} y_{t-2}, \\
\hat{\varepsilon}_{t-1} &= y_{t-1} - \mathbf{b}_{\mathrm{IV}}'\mathbf{x}_{t-1} - c_{\mathrm{IV}} y_{t-2}.
\end{aligned}$$

Let d be the coefficient on $\hat{\varepsilon}_{t-1}$ in this regression. The efficient estimator of ρ is

$$\hat{\hat{\rho}} = \hat{\rho} + d.$$

Appropriate asymptotic standard errors for the estimators, including $\hat{\hat{\rho}}$, are obtained from the $s^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}$ computed at the second step. These estimators are asymptotically equivalent to maximum likelihood estimators.[20]

One could argue that the concern about the bias of least squares is misdirected. Consider, again, the model in (20-12),

$$\begin{aligned}
y_t &= \beta y_{t-1} + \varepsilon_t, \\
\varepsilon_t &= \rho \varepsilon_{t-1} + u_t.
\end{aligned}$$

We established that linear regression of y_t on y_{t-1} estimates not β, but $\gamma = (\beta + \rho)/(1 - \beta\rho)$. It would follow that

$$E[y_t | y_{t-1}] = \gamma y_{t-1},$$

and this is what was of interest from the outset. If so, then the existence of autocorrelation in ε_t is a moot point. In a more completely specified model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \gamma y_{t-1} + \varepsilon_t,$$

what is likely to be of interest is $E[y_t | \mathbf{x}_t, y_{t-1}] = \mathbf{x}_t'\boldsymbol{\lambda} + \delta y_{t-1}$, and the question of autocorrelation of ε_t is a side issue. The nature of the autocorrelation in ε_t will determine whether $\boldsymbol{\beta} = \boldsymbol{\lambda}$ and $\gamma = \delta$. In the simplest case, as we saw earlier, if $\mathrm{Cov}[\varepsilon_t, \varepsilon_{t-s}] = 0$ for all s, then these equalities will hold. If ε_t is autocorrelated, then they will not. There is a fundamental ambiguity in this treatment, however. In the simple model, we also found earlier that $\mathrm{E}[y_t | y_{t-1}, y_{t-2}] = \gamma_1 y_{t-1} + \gamma_2 y_{t-2}$. There is no argument that the second-order equation is more or less correct than the first. They are two different

[20] See Hatanaka (2000).

representations of the same time series.[21] This idea calls into question the notion of "correcting" for autocorrelation in a regression. We saw in Example 20.2 another implication. The objective of the model builder would be to build residual autocorrelation out of the model. The presence of autocorrelation in the disturbance suggests that the regression part of the equation is incomplete.

Example 20.7 Dynamically Complete Regression

Figure 20.6 shows the residuals from two specifications of the gasoline demand model from Section 20.9.2: a static form,

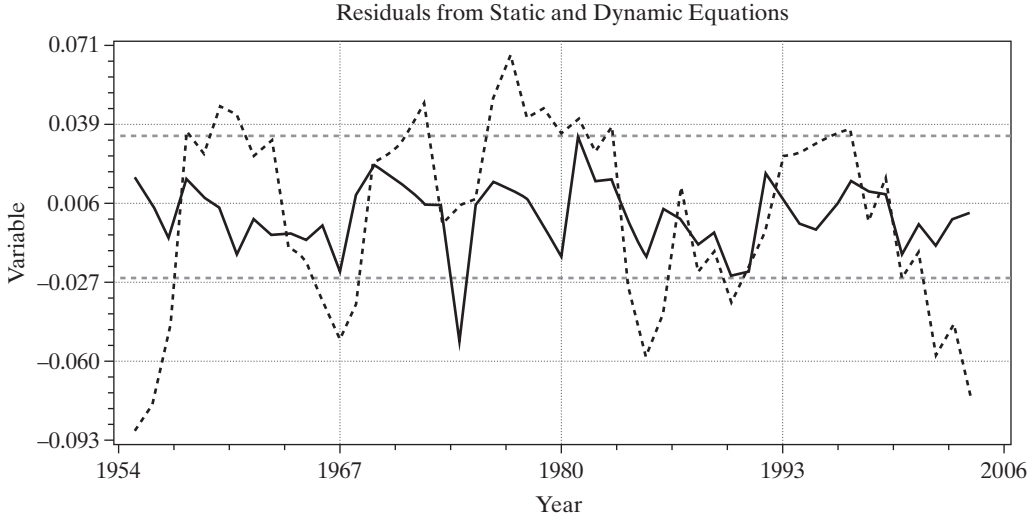
$$\ln\left(\frac{G}{pop}\right)_t = \beta_1 + \beta_2 \ln\left(\frac{Income}{Pop}\right)_t + \beta_3 \ln P_{G,t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \beta_6 t + \varepsilon_t,$$

and a dynamic form,

$$\ln\left(\frac{G}{Pop}\right)_t = \beta_1 + \beta_2 \ln\left(\frac{Income}{Pop}\right)_t + \beta_3 \ln P_{G,t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \beta_6 t + \gamma \ln\left(\frac{G}{Pop}\right)_{t-1} + \varepsilon_t.$$

The residuals from the dynamic model are shown with the solid lines. The horizontal bars contain the full range of variation of these residuals. The dashed figure shows the residuals from the static model. The much narrower range of the first set reflects the better fit of the model with the additional (highly significant) regressor. Note, as well, the more substantial amount of fluctuation which suggests less autocorrelation of the residuals from the more dynamically complete regression. To test for autocorrelation of the residuals, we computed the residuals from each regression and regressed them on the lagged residual and the other variables in the equations. For the dynamic model, the LM statistic (TR^2) equaled 1.641. This would be a

FIGURE 20.6 Regression Residuals.



[21]This is an implication of Wold's Decomposition Theorem. See Anderson (1971) or Greene (2003b, p. 619).

TABLE 20.3 Estimated Gasoline Demand Equations

	Dynamic Model				*Static Model*	
		Std.	*Elasticity*			*Std.*
Variable	*Estimate*	*Error*	*S.R.*	*L.R.*	*Estimate*	*Error*
Constant	−5.31920	1.45463	–	–	−26.4319	1.83501
ln *Income*	0.33945	0.10203	0.339	1.642	1.60170	0.17904
ln *Price*	−0.07617	0.01463	−0.076	−0.368	−0.06167	0.03872
ln *P New Cars*	−0.11713	0.06144	−0.117	−0.567	−0.14083	0.16284
ln *P Used Cars*	0.10016	0.03709	0.100	0.484	−0.01293	0.09664
Time trend	−0.00362	0.00180	–	–	−0.01518	0.00439
ln *Demand*[−1]	0.79327	0.04807	–	–	–	–
R^2	0.99552				0.96780	
LM Statistic (1)	1.641				29.787	

chi-squared variable with one degree of freedom. The critical value is 3.84, so the hypothesis of no autocorrelation is not rejected. The equation would appear to be dynamically complete. The same computation for the static model produces a chi-squared value of 29.787.

The estimates of the parameters for the two equations are given in Table 20.3. The fit of the model is high in both cases, but approaches one in the dynamic case. Long-run income and price elasticities are computed as $\eta = \beta_k/(1 - \gamma)$.

20.10 AUTOREGRESSIVE CONDITIONAL HETEROSCEDASTICITY

Heteroscedasticity is often associated with cross-sectional data, whereas time series are usually studied in the context of homoscedastic processes. In analyses of macroeconomic data, Engle (1982, 1983) and Cragg (1982) found evidence that for some kinds of data, the disturbance variances in time-series models were less stable than usually assumed. Engle's results suggested that in models of inflation, large and small forecast errors appeared to occur in clusters, suggesting a form of heteroscedasticity in which the variance of the forecast error depends on the size of the previous disturbance. He suggested the autoregressive, conditionally heteroscedastic, or ARCH, model as an alternative to the usual time-series process. More recent studies of financial markets suggest that the phenomenon is quite common. The ARCH model has proven to be useful in studying the volatility of inflation,[22] the term structure of interest rates,[23] the volatility of stock market returns,[24] and the behavior of foreign exchange markets,[25] to name but a few. This section will describe specification, estimation, and testing, in the basic formulations of the ARCH model and some extensions.[26]

[22]Coulson and Robins (1985).

[23]Engle, Hendry, and Trumble (1985).

[24]Engle, Lilien, and Robins (1987).

[25]Domowitz and Hakkio (1985) and Bollerslev and Ghysels (1996).

[26]Engle and Rothschild (1992) give a survey of this literature which describes many extensions. Mills (1993) also presents several applications. See, as well, Bollerslev (1986) and Li, Ling, and McAleer (2001). See McCullough and Renfro (1999) for discussion of estimation of this model.

Example 20.8 Stochastic Volatility

Figure 20.7 shows Bollerslev and Ghysel's 1974 data on the daily percentage nominal return for the Deutschmark/Pound exchange rate. (These data are given in Appendix Table F20.1.) The variation in the series appears to be fluctuating, with several clusters of large and small movements.

20.10.1 THE ARCH(1) MODEL

The simplest form of this model is the ARCH(1) model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t,$$
$$\varepsilon_t = u_t\sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}. \qquad \textbf{(20-34)}$$

where u_t is distributed as standard normal.[27] It follows that $E[\varepsilon_t|\mathbf{x}_t, \varepsilon_{t-1}] = 0$, so that $E[\varepsilon_t|\mathbf{x}_t] = 0$ and $E[y_t|\mathbf{x}_t] = \mathbf{x}_t'\boldsymbol{\beta}$. Therefore, this model is a classical regression model. But

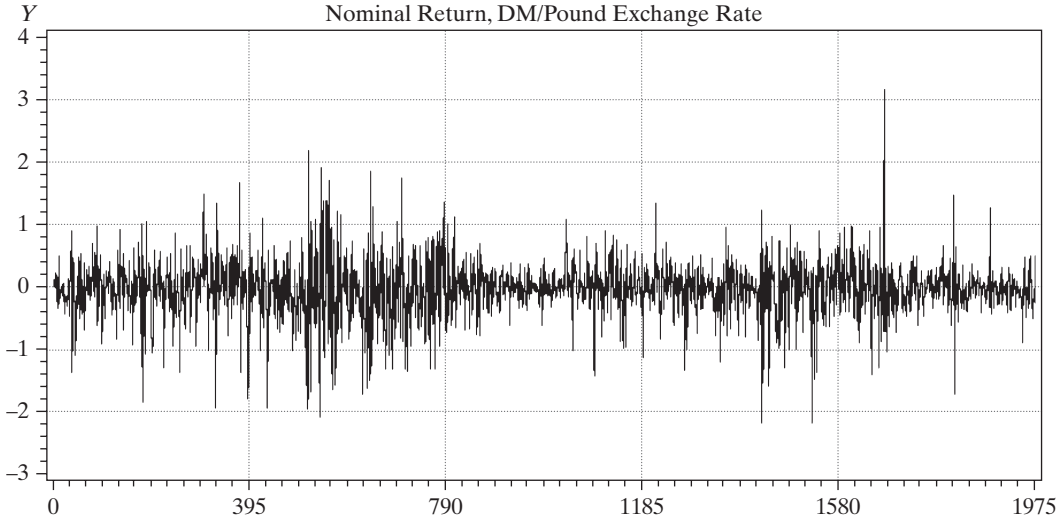
$$\text{Var}[\varepsilon_t|\varepsilon_{t-1}] = E[\varepsilon_t^2|\varepsilon_{t-1}] = E[u_t^2][\alpha_0 + \alpha_1\varepsilon_{t-1}^2] = \alpha_0 + \alpha_1\varepsilon_{t-1}^2,$$

so ε_t is *conditionally heteroscedastic*, not with respect to $\mathbf{x}_t$ as we considered in Chapter 9, but with respect to ε_{t-1}. The unconditional variance of ε_t is

$$\text{Var}[\varepsilon_t] = \text{Var}\{E[\varepsilon_t|\varepsilon_{t-1}]\} + E\{\text{Var}[\varepsilon_t|\varepsilon_{t-1}]\} = \alpha_0 + \alpha_1 E[\varepsilon_{t-1}^2] = \alpha_0 + \alpha_1\text{Var}[\varepsilon_{t-1}].$$

If the process generating the disturbances is weakly (covariance) stationary (see Definition 19.2),[28] then the unconditional variance is not changing over time so

FIGURE 20.7 Nominal Exchange Rate Returns.



[27]The assumption that u_t has unit variance is not a restriction. The scaling implied by any other variance would be absorbed by the other parameters.

[28]This discussion will draw on the results and terminology of time-series analysis in Section 20.3. The reader may wish to peruse this material at this point.

$$\text{Var}[\varepsilon_t] = \text{Var}[\varepsilon_{t-1}] = \alpha_0 + \alpha_1 \text{Var}[\varepsilon_{t-1}] = \frac{\alpha_0}{1 - \alpha_1}.$$

For this ratio to be finite and positive, $|\alpha_1|$ must be less than 1. Then, unconditionally, ε_t is distributed with mean zero and variance $\sigma^2 = \alpha_0/(1 - \alpha_1)$. Therefore, the model obeys the classical assumptions, and ordinary least squares is the most efficient *linear* unbiased estimator of $\boldsymbol{\beta}$.

But there is a more efficient *nonlinear* estimator. The log-likelihood function for this model is given by Engle (1982). Conditioned on starting values y_0 and $\mathbf{x}_0$ (and ε_0), the conditional log likelihood for observations $t = 1, \ldots, T$ is the one we examined in Section 14.9.2.a for the general heteroscedastic regression model [see (14-58)],

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln(\alpha_0 + \alpha_1\varepsilon_{t-1}^2) - \frac{1}{2}\sum_{t=1}^{T}\frac{\varepsilon_t^2}{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}, \quad \varepsilon_t = y_t - \boldsymbol{\beta}'\mathbf{x}_t. \quad \textbf{(20-35)}$$

Maximization of log L can be done with the conventional methods, as discussed in Appendix E.[29]

20.10.2 ARCH(*q*), ARCH-IN-MEAN, AND GENERALIZED ARCH MODELS

The natural extension of the ARCH(1) model presented before is a more general model with longer lags. The ARCH(q) process,

$$\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_q\varepsilon_{t-q}^2,$$

is a qth order moving average [MA(q)] process.[30] This section will generalize the ARCH(q) model, as suggested by Bollerslev (1986), in the direction of an autoregressive-moving average (ARMA) model of Section 21.2. The discussion will parallel his development, although many details are omitted for brevity. The reader is referred to that paper for background and for some of the less critical details.

Among the many variants of the capital asset pricing model (CAPM) is an intertemporal formulation by Merton (1980) that suggests an approximate linear relationship between the return and variance of the market portfolio. One of the possible flaws in this model is its assumption of a constant variance of the market portfolio. In this connection, then, the **ARCH-in-Mean**, or ARCH-M, model suggested by Engle, Lilien, and Robins (1987) is a natural extension. The model states that

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \delta\sigma_t^2 + \varepsilon_t,$$
$$\text{Var}[\varepsilon_t|\Psi_t] = \text{ARCH}(q).$$

Among the interesting implications of this modification of the standard model is that under certain assumptions, δ is the coefficient of relative risk aversion. The ARCH-M model has been applied in a wide variety of studies of volatility in asset returns, including

[29]Engle (1982) and Judge et al. (1985, pp. 441–444) suggest a four-step procedure based on the method of scoring that resembles the two-step method we used for the multiplicative heteroscedasticity model in Section 14.10.3. However, the full MLE is now incorporated in most modern software, so the simple regression-based methods, which are difficult to generalize, are less attractive in the current literature. But see McCullough and Renfro (1999) and Fiorentini, Calzolari, and Panattoni (1996) for commentary and some cautions related to maximum likelihood estimation.

[30]Once again, see Engle (1982).

the daily Standard & Poor's Index[31] and weekly New York Stock Exchange returns.[32] A lengthy list of applications is given in Bollerslev, Chou, and Kroner (1992).

The ARCH-M model has several noteworthy statistical characteristics. Unlike the standard regression model, misspecification of the variance function does affect the consistency of estimators of the parameters of the mean.[33] Recall that in the classical regression setting, weighted least squares is consistent even if the weights are misspecified as long as the weights are uncorrelated with the disturbances. That is not true here. If the ARCH part of the model is misspecified, then conventional estimators of $\boldsymbol{\beta}$ and δ will not be consistent. Bollerslev, Chou, and Kroner (1992) list a large number of studies that called into question the specification of the ARCH-M model, and they subsequently obtained quite different results after respecifying the model. A closely related practical problem is that the mean and variance parameters in this model are no longer uncorrelated. In analysis up to this point, we made quite profitable use of the block diagonality of the Hessian of the log-likelihood function for the model of heteroscedasticity. But the Hessian for the ARCH-M model is not block diagonal. In practical terms, the estimation problem cannot be segmented as we have done previously with the heteroscedastic regression model. All the parameters must be estimated simultaneously.

The generalized autoregressive conditional heteroscedasticity (GARCH) model is defined as follows.[34] The underlying regression is the usual one in (20-34). Conditioned on an information set at time t, denoted Ψ_t, the distribution of the disturbance is assumed to be

$$\varepsilon_t | \Psi_t \sim N[0, \sigma_t^2],$$

where the conditional variance is

$$\sigma_t^2 = \alpha_0 + \delta_1\sigma_{t-1}^2 + \delta_2\sigma_{t-2}^2 + \cdots + \delta_p\sigma_{t-p}^2 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_q\varepsilon_{t-q}^2. \quad \textbf{(20-36)}$$

Define

$$\mathbf{z}_t = [1, \sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \sigma_{t-p}^2, \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \ldots, \varepsilon_{t-q}^2]'$$

and

$$\boldsymbol{\gamma} = [\alpha_0, \delta_1, \delta_2, \ldots, \delta_p, \alpha_1, \ldots, \alpha_q]' = [\alpha_0, \boldsymbol{\delta}', \boldsymbol{\alpha}']'.$$

Then,

$$\sigma_t^2 = \boldsymbol{\gamma}'\mathbf{z}_t.$$

Notice that the conditional variance is defined by an autoregressive-moving average [ARMA (p, q)] process in the innovations ε_t^2. The difference here is that the *mean* of the

[31]See French, Schwert, and Stambaugh (1987).

[32]See Chou (1988).

[33]See Pagan and Ullah (1988) for a formal analysis of this point.

[34]As have most areas in time-series econometrics, the line of literature on GARCH models has progressed rapidly in recent years and will surely continue to do so. We have presented Bollerslev's model in some detail, despite many recent extensions, not only to introduce the topic as a bridge to the literature, but also because it provides a convenient and interesting setting in which to discuss several related topics such as double-length regression and pseudo-maximum likelihood estimation.

random variable of interest y_t is described completely by a heteroscedastic, but otherwise ordinary, regression model. The *conditional variance*, however, evolves over time in what might be a very complicated manner, depending on the parameter values and on p and q. The model in (20-36) is a GARCH(p,q) model, where p refers, as before, to the order of the autoregressive part.[35] As Bollerslev (1986) demonstrates with an example, the virtue of this approach is that a GARCH model with a small number of terms appears to perform as well as or better than an ARCH model with many.

The **stationarity conditions** are important in this context to ensure that the moments of the normal distribution are finite. The reason is that higher moments of the normal distribution are finite powers of the variance. A normal distribution with variance σ_t^2 has fourth moment $3\sigma_t^4$, sixth moment $15\sigma_t^6$, and so on. [The precise relationship of the even moments of the normal distribution to the variance is $\mu_{2k} = (\sigma^2)^k(2k)!/(k!2^k)$.] Simply ensuring that σ_t^2 is stable does not ensure that higher powers are as well.[36] Bollerslev presents a useful figure that shows the conditions needed to ensure stability for moments up to order 12 for a GARCH(1,1) model and gives some additional discussion. For example, for a GARCH(1,1) process, for the fourth moment to exist, $3\alpha_1^2 + 2\alpha_1\delta_1 + \delta_1^2$ must be less than 1.

It is convenient to write (20-36) in terms of polynomials in the lag operator,

$$\sigma_t^2 = \alpha_0 + D(L)\sigma_t^2 + A(L)\varepsilon_t^2.$$

The stationarity condition for such an equation is that the roots of the characteristic equation, $1 - D(z) = 0$, must lie outside the unit circle. For the present, we will assume that this case is true for the model we are considering and that $A(1) + D(1) < 1$. [This assumption is stronger than that needed to ensure stationarity in a higher-order autoregressive model, which would depend only on $D(L)$.] The implication is that the GARCH process is covariance stationary with $E[\varepsilon_t] = 0$ (unconditionally), $\text{Var}[\varepsilon_t] = \alpha_0/[1 - A(1) - D(1)]$, and $\text{Cov}[\varepsilon_t, \varepsilon_s] = 0$ for all $t \neq s$. Thus, unconditionally the model is the classical regression model that we examined in Chapters 2–6.

The usefulness of the GARCH specification is that it allows the variance to evolve over time in a way that is much more general than the simple specification of the ARCH model. For the example discussed in his paper, Bollerslev reports that although Engle and Kraft's (1983) ARCH(8) model for the rate of inflation in the GNP deflator appears to remove all ARCH effects, a closer look reveals GARCH effects at several lags. By fitting a GARCH(1,1) model to the same data, Bollerslev finds that the ARCH effects out to the same eight-period lag as fit by Engle and Kraft and his observed GARCH effects are all satisfactorily accounted for.

20.10.3 MAXIMUM LIKELIHOOD ESTIMATION OF THE GARCH MODEL

Bollerslev describes a method of estimation based on the BHHH algorithm. As he shows, the method is relatively simple, although with the line search and first derivative

[35] We have changed Bollerslev's notation slightly so as not to conflict with our previous presentation. He used $\boldsymbol{\beta}$ instead of our $\boldsymbol{\delta}$ in (20-36) and $\mathbf{b}$ instead of our $\boldsymbol{\beta}$ in (20-34).

[36] The conditions cannot be imposed a priori. In fact, there is no nonzero set of parameters that guarantees stability of *all* moments, even though the normal distribution has finite moments of all orders. As such, the normality assumption must be viewed as an approximation.

method that he suggests, it probably involves more computation and more iterations than necessary. Following the suggestions of Harvey (1976), it turns out that there is a simpler way to estimate the GARCH model that is also very illuminating. This model is actually very similar to the more conventional model of multiplicative heteroscedasticity that we examined in Section 14.10.3.

For normally distributed disturbances, the log likelihood for a sample of T observations is[37]

$$\ln L = \sum_{t=1}^{T} -\frac{1}{2}\left[\ln(2\pi) + \ln \sigma_t^2 + \frac{\varepsilon_t^2}{\sigma_t^2}\right] = \sum_{t=1}^{T} \ln f_t(\boldsymbol{\theta}) = \sum_{t=1}^{T} l_t(\boldsymbol{\theta}),$$

where $\varepsilon_t = y_t - \mathbf{x}_t'\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha_0, \boldsymbol{\alpha}', \boldsymbol{\delta}')' = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$. Derivatives of $\ln L$ are obtained by summation. Let l_t denote $\ln f_t(\boldsymbol{\theta})$. The first derivatives with respect to the variance parameters are

$$\frac{\partial l_t}{\partial \boldsymbol{\gamma}} = -\frac{1}{2}\left[\frac{1}{\sigma_t^2} - \frac{\varepsilon_t^2}{(\sigma_t^2)^2}\right]\frac{\partial \sigma_t^2}{\partial \boldsymbol{\gamma}} = \frac{1}{2}\left(\frac{1}{\sigma_t^2}\right)\frac{\partial \sigma_t^2}{\partial \boldsymbol{\gamma}}\left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1\right) = \frac{1}{2}\left(\frac{1}{\sigma_t^2}\right)\mathbf{g}_t v_t = \mathbf{b}_t v_t. \qquad \textbf{(20-37)}$$

Note that $E[v_t] = 0$. Suppose, for now, that there are no regression parameters. Newton's method for estimating the variance parameters would be

$$\hat{\boldsymbol{\gamma}}^{i+1} = \hat{\boldsymbol{\gamma}}^{i} - \mathbf{H}^{-1}\mathbf{g}, \qquad \textbf{(20-38)}$$

where $\mathbf{H}$ indicates the Hessian and $\mathbf{g}$ is the first derivatives vector. Following Harvey's suggestion (see Section 14.10.3), we will use the method of scoring instead. To do this, we make use of $E[v_t] = 0$ and $E[\varepsilon_t^2/\sigma_t^2] = 1$. After taking expectations in (20-37), the iteration reduces to a linear regression of $v_{*_t} = (1/\sqrt{2})v_t$ on regressors $\mathbf{w}_{*_t} = (1/\sqrt{2})\mathbf{g}_t/\sigma_t^2$. That is,

$$\hat{\boldsymbol{\gamma}}^{i+1} = \hat{\boldsymbol{\gamma}}^{i} + [\mathbf{W}_*'\mathbf{W}_*]^{-1}\mathbf{W}_*'\mathbf{v}_* = \hat{\boldsymbol{\gamma}}^{i} + [\mathbf{W}_*'\mathbf{W}_*]^{-1}\left(\frac{\partial \ln L}{\partial \boldsymbol{\gamma}}\right), \qquad \textbf{(20-39)}$$

where row t of $\mathbf{W}_*$ is $\mathbf{w}_{*t}'$. The iteration has converged when the slope vector is zero, which happens when the first derivative vector is zero. When the iterations are complete, the estimated asymptotic covariance matrix is simply

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}] = [\hat{\mathbf{W}}_*'\mathbf{W}_*]^{-1}$$

based on the estimated parameters.

The usefulness of the result just given is that $E[\partial^2 \ln L/\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}']$ is, in fact, zero. Because the expected Hessian is block diagonal, applying the method of scoring to the full parameter vector can proceed in two parts, exactly as it did in Section 14.10.3 for the multiplicative heteroscedasticity model. That is, the updates for the mean and variance

[37] There are three minor errors in Bollerslev's derivation that we note here to avoid the apparent inconsistencies. In his (22), $\frac{1}{2}h_t$ should be $\frac{1}{2}h_t^{-1}$. In (23), $-2h_t^{-2}$ should be $-h_t^{-2}$. In (28), $h\ \partial h/\partial\omega$ should, in each case, be $(1/h)\ \partial h/\partial\omega$. [In his (8), $\alpha_0\alpha_1$ should be $\alpha_0 + \alpha_1$, but this has no implications for our derivation.]

parameter vectors can be computed separately. Consider then the slope parameters, $\boldsymbol{\beta}$. The same type of modified scoring method as used earlier produces the iteration

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{i+1} &= \hat{\boldsymbol{\beta}}^{i} + \left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\mathbf{x}_t'}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)'\right]^{-1}\left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\varepsilon_t}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)v_t\right]\\ &= \hat{\boldsymbol{\beta}}^{i} + \left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\mathbf{x}_t'}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)'\right]^{-1}\left(\frac{\partial \ln L}{\partial \boldsymbol{\beta}}\right)\\ &= \hat{\boldsymbol{\beta}}^{i} + \mathbf{h}^{i}, \end{aligned} \tag{20-40}$$

which has been referred to as a **double-length regression**.[38] The update vector $\mathbf{h}^i$ is the vector of slopes in an augmented or double-length generalized regression,

$$\mathbf{h}^{i} = [\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}[\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{a}], \tag{20-41}$$

where $\mathbf{C}$ is a $2T \times K$ matrix whose first T rows are the $\mathbf{X}$ from the original regression model and whose next T rows are $(1/\sqrt{2})\mathbf{d}_t'/\sigma_t^2$, $t = 1, \ldots, T$; $\mathbf{a}$ is a $2T \times 1$ vector whose first T elements are ε_t and whose next T elements are $(1/\sqrt{2})v_t/\sigma_t^2$, $t = 1, \ldots, T$; and $\boldsymbol{\Omega}$ is a diagonal matrix with $1/\sigma_t^2$ in positions $1, \ldots, T$ and ones below observation T. At convergence, $[\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}$ provides the asymptotic covariance matrix for the MLE. The resemblance to the familiar result for the generalized regression model is striking, but note that this result is based on the double-length regression.

The iteration is done simply by computing the update vectors to the current parameters as defined earlier.[39] An important consideration is that to apply the scoring method, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are updated simultaneously. That is, one does not use the updated estimate of $\boldsymbol{\gamma}$ in (20-39) to update the weights for the GLS regression to compute the new $\boldsymbol{\beta}$ in (20-40). The same estimates (the results of the prior iteration) are used on the right-hand sides of both (20-39) and (20-40). The remaining problem is to obtain starting values for the iterations. One obvious choice is $\mathbf{b}$, the OLS estimator, for $\boldsymbol{\beta}$, $\mathbf{e}'\mathbf{e}/T = s^2$ for α_0, and zero for all the remaining parameters. The OLS slope vector will be consistent under all specifications. A useful alternative in this context would be to start $\boldsymbol{\alpha}$ at the vector of slopes in the least squares regression of e_t^2, the squared OLS residual, on a constant and q lagged values.[40] As discussed later, an LM test for the presence of GARCH effects is then a byproduct of the first iteration. In principle, the updated result of the first iteration is an **efficient two-step estimator** of all the parameters. But having gone to the full effort to set up the iterations, nothing is gained by not iterating to convergence. One virtue of allowing the procedure to iterate to convergence is that the resulting log-likelihood function can be used in likelihood ratio tests.

[38] See Orme (1990) and Davidson and MacKinnon (1993, Chapter 14).

[39] See Fiorentini et al. (1996) on computation of derivatives in GARCH models.

[40] A test for the presence of ARCH(q) effects against none can be carried out by carrying TR^2 from this regression into a table of critical values for the chi-squared distribution. But in the presence of GARCH effects, this procedure loses its validity.

20.10.4 TESTING FOR GARCH EFFECTS

The preceding development appears fairly complicated. In fact, it is not, because at each step, nothing more than a linear least squares regression is required. The intricate part of the computation is setting up the derivatives. On the other hand, it does take a fair amount of programming to get this far.[41] As Bollerslev suggests, it might be useful to test for GARCH effects first.

The simplest approach is to examine the squares of the least squares residuals. The autocorrelations (correlations with lagged values) of the squares of the residuals provide evidence about ARCH effects. An LM test of ARCH(q) against the hypothesis of no ARCH effects [ARCH(0), the classical model] can be carried out by computing $\chi^2 = TR^2$ in the regression of e_t^2 on a constant and q lagged values. Under the null hypothesis of no ARCH effects, the statistic has a limiting chi-squared distribution with q degrees of freedom. Values larger than the critical table value give evidence of the presence of ARCH (or GARCH) effects.

Bollerslev suggests a Lagrange multiplier statistic that is, in fact, surprisingly simple to compute. The LM test for GARCH(p,0) against GARCH(p,q) can be carried out by referring T times the R^2 in the linear regression defined in (20-42) to the chi-squared critical value with q degrees of freedom. There is, unfortunately, an indeterminacy in this test procedure. The test for ARCH(q) against GARCH(p,q) is exactly the same as that for ARCH(p) against ARCH($p + q$). For carrying out the test, one can use as starting values a set of estimates that includes $\boldsymbol{\delta} = \mathbf{0}$ and any consistent estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Then TR^2 for the regression at the initial iteration provides the test statistic.[42] A number of recent papers have questioned the use of test statistics based solely on normality. Wooldridge (1991) is a useful summary with several examples.

Example 20.9 GARCH Model for Exchange Rate Volatility

Bollerslev and Ghysels analyzed the exchange rate data in Appendix Table F20.1 using a GARCH(1,1) model,

$$\begin{aligned} y_t &= \mu + \varepsilon_t, \\ E[\varepsilon_t | \varepsilon_{t-1}] &= 0, \\ \text{Var}[\varepsilon_t | \varepsilon_{t-1}] &= \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \delta \sigma_{t-1}^2. \end{aligned}$$

The least squares residuals for this model are simply $e_t = y_t - \bar{y}$. Regression of the squares of these residuals on a constant and 10 lagged squared values using observations 11–1974 produces an $R^2 = 0.09795$. With $T = 1964$, the chi-squared statistic is 192.37, which is larger than the critical value from the table of 18.31. We conclude that there is evidence of GARCH effects in these residuals. The maximum likelihood estimates of the GARCH model are given in Table 20.4. Note the resemblance between the OLS unconditional variance (0.221128) and the estimated equilibrium variance from the GARCH model, 0.2631.

[41]Because this procedure is available as a preprogrammed procedure in many computer programs, including *EViews, Stata, RATS, NLOGIT, Shazam*, and other programs, this warning might itself be overstated.

[42]Bollerslev argues that, in view of the complexity of the computations involved in estimating the GARCH model, it is useful to have a test for GARCH effects. This case is one (as are many other maximum likelihood problems) in which the apparatus for carrying out the test is the same as that for estimating the model. Having computed the LM statistic for GARCH effects, one can proceed to estimate the model just by allowing the program to iterate to convergence. There is no additional cost beyond waiting for the answer.

TABLE 20.4 Maximum Likelihood Estimates of a GARCH(1,1) Model[43]

	μ	α_0	α_1	δ	$\alpha_0/(1-\alpha_1-\delta)$
Estimate	−0.006190	0.01076	0.1531	0.8060	0.2631
Std. Error	0.00873	0.00312	0.0273	0.0302	0.594
t ratio	−0.709	3.445	5.605	26.731	0.443

$\ln L = -1106.61, \ln L_{\text{OLS}} = -1311.09, \bar{y} = -0.01642, s^2 = 0.221128$

20.10.5 PSEUDO–MAXIMUM LIKELIHOOD ESTIMATION

We now consider an implication of nonnormality of the disturbances. Suppose that the assumption of normality is weakened to only

$$E[\varepsilon_t|\Psi_t] = 0, \qquad E\left[\frac{\varepsilon_t^2}{\sigma_t^2}\middle|\Psi_t\right] = 1, \qquad E\left[\frac{\varepsilon_t^4}{\sigma_t^4}\middle|\Psi_t\right] = \kappa < \infty,$$

where σ_t^2 is as defined earlier. Now the normal log-likelihood function is inappropriate. In this case, the nonlinear (ordinary or weighted) least squares estimator would have the properties discussed in Chapter 7. It would be more difficult to compute than the MLE discussed earlier, however. It has been shown[44] that the pseudo-MLE obtained by maximizing the same log likelihood as if it were correct produces a consistent estimator despite the misspecification.[45] The asymptotic covariance matrices for the parameter estimators must be adjusted, however.

The general result for cases such as this one[46] is that the appropriate asymptotic covariance matrix for the pseudo-MLE of a parameter vector $\boldsymbol{\theta}$ would be

$$\text{Asy. Var}[\hat{\boldsymbol{\theta}}] = \mathbf{H}^{-1}\mathbf{F}\mathbf{H}^{-1}, \tag{20-42}$$

where

$$\mathbf{H} = -E\left[\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right],$$

and

$$\mathbf{F} = E\left[\left(\frac{\partial \ln L}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \ln L}{\partial\boldsymbol{\theta}'}\right)\right],$$

(i.e., the BHHH estimator), and $\ln L$ is the used but inappropriate log-likelihood function. For current purposes, $\mathbf{H}$ and $\mathbf{F}$ are still block diagonal, so we can treat the mean and variance parameters separately. In addition, $E[v_t]$ is still zero, so the second derivative terms in both blocks are quite simple. (The parts involving $\partial^2\sigma_t^2/\partial\boldsymbol{\gamma}\,\partial\boldsymbol{\gamma}'$ and

[43]These data have become a standard data set for the evaluation of software for estimating GARCH models. The values given are the benchmark estimates. Standard errors differ substantially from one method to the next. Those given are the Bollerslev and Wooldridge (1992) results. See McCullough and Renfro (1999).

[44]See White (1982a) and Weiss (1982).

[45]White (1982a) gives some additional requirements for the true underlying density of ε_t. Gourieroux, Monfort, and Trognon (1984) also consider the issue. Under the assumptions given, the expectations of the matrices in (20-36) and (20-41) remain the same as under normality. The consistency and asymptotic normality of the pseudo-MLE can be argued under the logic of GMM estimators.

[46]See Gourieroux, Monfort, and Trognon (1984).

$\partial^2\sigma_t^2/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'$ fall out of the expectation.) Taking expectations and inserting the parts produces the corrected asymptotic covariance matrix for the variance parameters,

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}_{\text{PMLE}}] = [\mathbf{W}_*'\mathbf{W}_*]^{-1}\mathbf{B}'\mathbf{B}[\mathbf{W}_*'\mathbf{W}_*]^{-1},$$

where the rows of $\mathbf{W}^*$ are defined in (20-39) and those of $\mathbf{B}$ are in (20-37). For the slope parameters, the adjusted asymptotic covariance matrix would be

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}_{\text{PMLE}}] = [\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}\left[\sum_{t=1}^{T}\mathbf{b}_t\,\mathbf{b}_t'\right][\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1},$$

where the outer matrix is defined in (20-41) and, from the first derivatives given in (20-37) and (20-40),[47]

$$\mathbf{b}_t = \frac{\mathbf{x}_t\varepsilon_t}{\sigma_t^2} + \frac{1}{2}\left(\frac{v_t}{\sigma_t^2}\right)\mathbf{d}_t.$$

20.11 SUMMARY AND CONCLUSIONS

This chapter has examined the generalized regression model with serial correlation in the disturbances. We began with some general results on analysis of time-series data. When we consider dependent observations and serial correlation, the laws of large numbers and central limit theorems used to analyze independent observations no longer suffice. We presented some useful tools that extend these results to time-series settings. We then considered estimation and testing in the presence of autocorrelation. As usual, OLS is consistent but inefficient. The Newey–West estimator is a robust estimator for the asymptotic covariance matrix of the OLS estimator. This pair of estimators also constitute the GMM estimator for the regression model with autocorrelation. We then considered two-step feasible generalized least squares and maximum likelihood estimation for the special case usually analyzed by practitioners, the AR(1) model. The model with a correction for autocorrelation is a restriction on a more general model with lagged values of both dependent and independent variables. We considered a means of testing this specification as an alternative to fixing the problem of autocorrelation. The final section, on ARCH and GARCH effects, describes an extension of the models of autoregression to the conditional variance of ε as opposed to the conditional mean. This model embodies elements of both autocorrelation and heteroscedasticity. The set of methods plays a fundamental role in the modern analysis of volatility in financial data.

Key Terms and Concepts

- AR(1)
- ARCH
- ARCH-in-mean
- Asymptotic negligibility
- Asymptotic normality
- Autocorrelation coefficient
- Autocorrelation function
- Autocorrelation matrix
- Autocovariance
- Autocovariance matrix
- Autoregressive form
- Autoregressive processes

[47]McCullough and Renfro (1999) examined several approaches to computing an appropriate asymptotic covariance matrix for the GARCH model, including the conventional Hessian and BHHH estimators and three sandwich-style estimators, including the one suggested earlier and two based on the method of scoring suggested by Bollerslev and Wooldridge (1992). None stands out as obviously better, but the Bollerslev and QMLE estimator based on an actual Hessian appears to perform well in Monte Carlo studies.

- Cochrane–Orcutt estimator
- Covariance stationarity
- Double-length regression
- Durbin–Watson test
- Efficient two-step estimator
- Ergodicity
- Expectations-augmented Phillips curve
- First-order autoregression
- Innovation
- LM test
- Martingale sequence
- Martingale difference sequence
- Moving-average form
- Moving-average process
- Newey–West autocorrelation consistent covariance estimator
- Partial difference
- Prais–Winsten estimator
- Pseudo-differences
- Q test
- Quasi differences
- Random walk
- Stationarity
- Stationarity conditions
- Summability
- Time-series process
- Time window
- Weakly stationary
- White noise
- Yule–Walker equations

Exercises

1. Does first differencing reduce autocorrelation? Consider the models $y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t$, where $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ and $\varepsilon_t = u_t - \lambda u_{t-1}$. Compare the autocorrelation of ε_t in the original model with that of v_t in $y_t - y_{t-1} = \boldsymbol{\beta}'(\mathbf{x}_t - \mathbf{x}_{t-1}) + v_t$, where $v_t = \varepsilon_t - \varepsilon_{t-1}$.
2. Derive the disturbance covariance matrix for the model

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t,$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t - \lambda u_{t-1}.$$

   What parameter is estimated by the regression of the OLS residuals on their lagged values?
3. It is commonly asserted that the Durbin–Watson statistic is only appropriate for testing for first-order autoregressive disturbances. The Durbin–Watson statistic estimates $2(1 - \rho)$ where ρ is the first-order autocorrelation of the residuals. What combination of the coefficients of the model is estimated by the Durbin–Watson statistic in each of the following cases: AR(1), AR(2), MA(1)? In each case, assume that the regression model does not contain a lagged dependent variable. Comment on the impact on your results of relaxing this assumption.

Applications

1. The data used to fit the expectations augmented Phillips curve in Example 20.3 are given in Appendix Table F5.2. Using these data, reestimate the model given in the example. Carry out a formal test for first-order autocorrelation using the LM statistic. Then, reestimate the model using an AR(1) model for the disturbance process. Because the sample is large, the Prais–Winsten and Cochrane–Orcutt estimators should give essentially the same answer. Do they? After fitting the model, obtain the transformed residuals and examine them for first-order autocorrelation. Does the AR(1) model appear to have adequately fixed the problem?
2. Data for fitting an improved Phillips curve model can be obtained from many sources, including the Bureau of Economic Analysis's (BEA) own Web site, www.economagic.com, and so on. Obtain the necessary data and expand the model of

Example 20.3. Does adding additional explanatory variables to the model reduce the extreme pattern of the OLS residuals that appears in Figure 20.3?

3. (This exercise requires appropriate computer software. The computations required can be done with *RATS*, *EViews*, *Stata*, *LIMDEP*, and a variety of other software using only preprogrammed procedures.) Quarterly data on the consumer price index for 1950.1 to 2000.4 are given in Appendix Table F5.2. Use these data to fit the model proposed by Engle and Kraft (1983). The model is

$$\pi_t = \beta_0 + \beta_1\pi_{t-1} + \beta_2\pi_{t-2} + \beta_3\pi_{t-3} + \beta_4\pi_{t-4} + \varepsilon_t,$$

where $\pi_t = 100 \ln[p_t/p_{t-1}]$ and p_t is the price index.

a. Fit the model by ordinary least squares, then use the tests suggested in the text to see if ARCH effects appear to be present.

b. The authors fit an ARCH(8) model with declining weights,

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{8}\left(\frac{9-i}{36}\right)\varepsilon_{t-i}^2.$$

Fit this model. If the software does not allow constraints on the coefficients, you can still do this with a two-step least squares procedure, using the least squares residuals from the first step. What do you find?

c. Bollerslev (1986) recomputed this model as a GARCH(1,1). Use the GARCH(1,1) to form and refit your model.

21
NONSTATIONARY DATA

21.1 INTRODUCTION

Most economic variables that exhibit strong trends, such as GDP, consumption, or the price level, are not stationary and are thus not amenable to the analysis of the previous chapter. In many cases, stationarity can be achieved by simple differencing or some other simple transformation. But new statistical issues arise in analyzing nonstationary series that are understated by this superficial observation. This chapter will survey a few of the major issues in the analysis of nonstationary data.[1] We begin in Section 21.2 with results on analysis of a single nonstationary time series. Section 21.3 examines the implications of nonstationarity for analyzing regression relationship. Finally, Section 21.4 turns to the extension of the time-series results to panel data.

21.2 NONSTATIONARY PROCESSES AND UNIT ROOTS

This section will begin the analysis of nonstationary time series with some basic results for univariate time series. The fundamental results concern the characteristics of nonstationary series and statistical tests for identification of nonstationarity in observed data.

21.2.1 THE LAG AND DIFFERENCE OPERATORS

The lag operator, L, is a device that greatly simplifies the mathematics of time-series analysis. The operator defines the lagging operation,

$$Ly_t = y_{t-1}.$$

From the definition,

$$L^2y_t = L(Ly_t) = Ly_{t-1} = y_{t-2}.$$

It follows that

$$L^Py_t = y_{t-P},$$

$$(L^P)^Qy_t = L^{PQ}y_t = y_{t-PQ},$$

$$(L^P)(L^Q)y_t = L^Py_{t-Q} = L^{Q+P}y_t = y_{t-Q-P}.$$

[1]With panel data, this is one of the rapidly growing areas in econometrics, and the literature advances rapidly. We can only scratch the surface. Several surveys and books provide useful extensions. Three that will be very helpful are Hamilton (1994), Enders (2004), and Tsay (2005).

Finally, for the autoregressive series $y_t = \beta y_{t-1} + \varepsilon_t$, where $|\beta| < 1$, we find $(1 - \beta L)y_t = \varepsilon_t$ or

$$y_t = \left(\frac{1}{1 - \beta L}\right)\varepsilon_t = [1 + \beta L + \beta^2 L^2 + \ \ldots\ ]\varepsilon_t = \sum_{s=0}^{\infty} \beta^s \varepsilon_{t-s}.$$

The first difference operator is a useful shorthand that follows from the definition of L,

$$(1 - L)y_t = y_t - y_{t-1} = \Delta y_t.$$

So, for example,

$$\Delta^2 y_t = \Delta(\Delta y_t) = \Delta(y_t - y_{t-1}) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}).$$

21.2.2 INTEGRATED PROCESSES AND DIFFERENCING

A process that figures prominently in this setting is the **random walk with drift**,

$$y_t = \mu + y_{t-1} + \varepsilon_t.$$

By direct substitution,

$$y_t = \frac{\mu + \varepsilon_t}{1 - L} = \sum_{s=0}^{\infty} (\mu + \varepsilon_{t-s}).$$

That is, y_t is the simple sum of what will eventually be an infinite number of random variables, possibly with nonzero mean. If the innovations, ε_t, are being generated by the same zero-mean, constant-variance process, then the variance of y_t would obviously be infinite. As such, the random walk is clearly a **nonstationary process**, even if μ equals zero. On the other hand, the first difference of y_t,

$$z_t = y_t - y_{t-1} = \Delta y_t = \mu + \varepsilon_t,$$

is simply the innovation plus the mean of z_t, which we have already assumed is stationary.

The series y_t is said to be **integrated of order one**, denoted $I(1)$, because taking a first difference produces a stationary process. A nonstationary series is integrated of order d, denoted $I(d)$, if it becomes stationary after being first differenced d times. A generalization of the autoregressive moving average model, $y_t = \gamma y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}$, would be the series

$$z_t = (1 - L)^d y_t = \Delta^d y_t.$$

The resulting model is denoted an **autoregressive integrated moving-average** model, or **ARIMA** (p, d, q).[2] In full, the model would be

$$\Delta^d y_t = \mu + \gamma_1 \Delta^d y_{t-1} + \gamma_2 \Delta^d y_{t-2} + \ \cdots\ + \gamma_p \Delta^d y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ \cdots\ - \theta_q \varepsilon_{t-q},$$

[2]There are yet further refinements one might consider, such as removing seasonal effects from z_t by differencing by quarter or month. See Harvey (1990) and Davidson and MacKinnon (1993). Some recent work has relaxed the assumption that d is an integer. The fractionally integrated series or ARFIMA has been used to model series in which the very long-run multipliers decay more slowly than would be predicted otherwise.

where

$$\Delta y_t = y_t - y_{t-1} = (1 - L)y_t.$$

This result may be written compactly as

$$C(L)[(1 - L)^d y_t] = \mu + D(L)\varepsilon_t,$$

where $C(L)$ and $D(L)$ are the polynomials in the lag operator and $(1 - L)^d y_t = \Delta^d y_t$ is the dth difference of y_t.

An $I(1)$ series in its raw (undifferenced) form will typically be constantly growing or wandering about with no tendency to revert to a fixed mean. Most macroeconomic flows and stocks that relate to population size, such as output or employment, are $I(1)$. An $I(2)$ series is growing at an ever-increasing rate. The price-level data in Appendix Table F5.2 and shown later appear to be $I(2)$. Series that are $I(3)$ or greater are extremely unusual, but they do exist. Among the few manifestly $I(3)$ series that could be listed, one would find, for example, the money stocks or price levels in hyperinflationary economies such as interwar Germany or Hungary after World War II.

Example 21.1 A Nonstationary Series

The nominal GDP and consumer price index variables in Appendix Table F5.2 are strongly trended, so the mean is changing over time. Figures 21.1–21.3 plot the log of the consumer price index series in Table F5.2 and its first and second differences. The original series and first differences are obviously nonstationary, but the second differencing appears to have rendered the series stationary.

The first 10 autocorrelations of the log of the GNP deflator series are shown in Table 21.1. (See Example 20.4 for details on the ACF.) The autocorrelations of the original series show the signature of a strongly trended, nonstationary series. The first difference also exhibits nonstationarity, because the autocorrelations are still very large after a lag of 10 periods. The second difference appears to be stationary, with mild negative autocorrelation

FIGURE 21.1 Quarterly Data on Log Consumer Price Index.

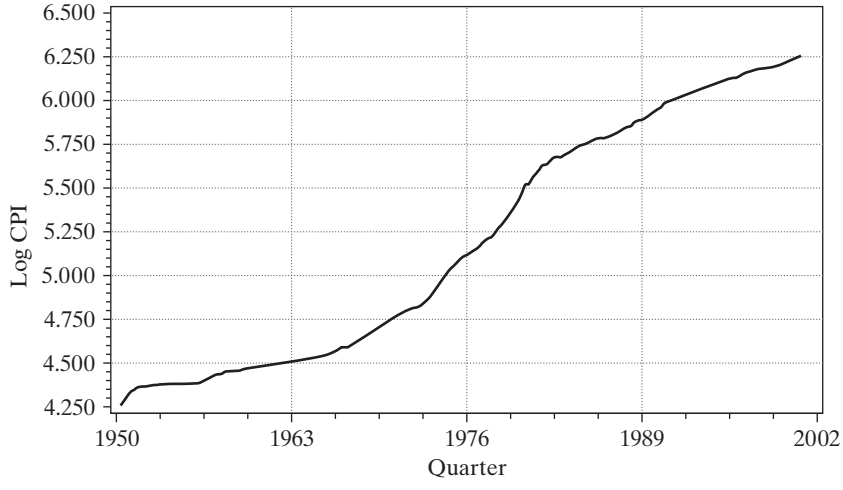
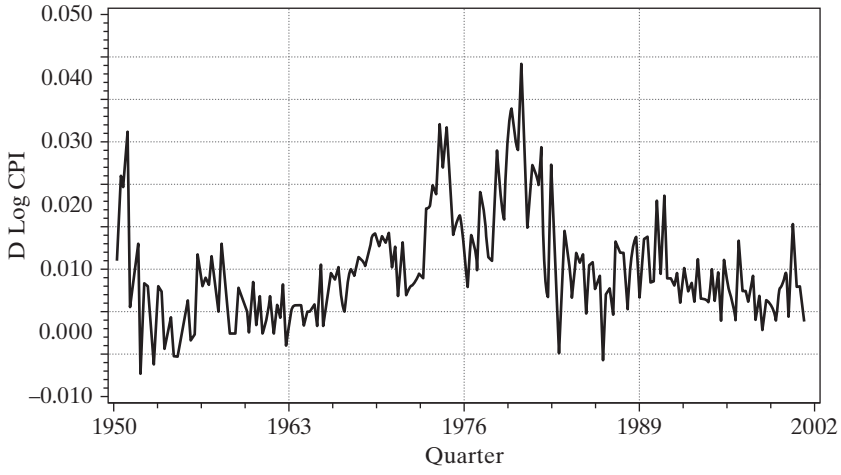


FIGURE 21.2 First Difference of Log Consumer Price Index.



at the first lag, but essentially none after that. Intuition might suggest that further differencing would reduce the autocorrelation further, but that would be incorrect. We leave as an exercise to show that, in fact, for values of γ less than about 0.5, first differencing of an AR(1) process actually increases autocorrelation.

FIGURE 21.3 Second Difference of Log Consumer Price Index.

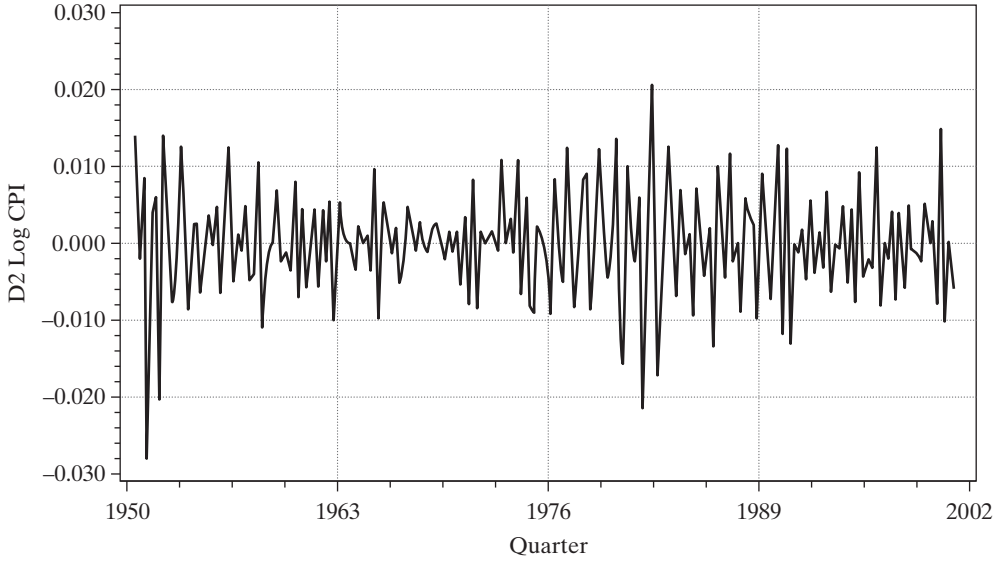


TABLE 21.1 Autocorrelations for ln Consumer Price Index

Lag	*Autocorrelation Function Original Series, log Price*		*Autocorrelation Function First Difference of log Price*		*Autocorrelation Function Second Difference of log Price*		
1	0.989	•••••••••••	0.654	•••••••	−0.422	••••••	
2	0.979	•••••••••••	0.600	•••••••	−0.111	•	
3	0.968	•••••••••••	0.621	•••••••	0.075		•
4	0.958	•••••••••••	0.600	•••••••	0.147		•
5	0.947	••••••••••	0.469	••••••	−0.112	•	
6	0.936	••••••••••	0.418	••••••	−0.037	•	
7	0.925	••••••••••	0.393	•••••	0.008		•
8	0.914	••••••••••	0.361	•••••	0.034		•
9	0.903	••••••••••	0.303	•••••	−0.023	•	
10	0.891	••••••••••	0.262	•••	−0.041	•	

21.2.3 RANDOM WALKS, TRENDS, AND SPURIOUS REGRESSIONS

In a seminal paper, Granger and Newbold (1974) argued that researchers had not paid sufficient attention to the warning of very high autocorrelation in the residuals from conventional regression models. Among their conclusions were that macroeconomic data, as a rule, were integrated and that in regressions involving the levels of such data, the standard significance tests were usually misleading. The conventional t and F tests would tend to reject the hypothesis of no relationship when, in fact, there might be none. The general result at the center of these findings is that conventional linear regression, ignoring serial correlation, of one random walk on another is virtually certain to suggest a significant relationship, even if the two are, in fact, independent. Among their extreme conclusions, Granger and Newbold suggested that researchers use a critical t value of 11.2 rather than the standard normal value of 1.96 to assess the significance of a coefficient estimate. Phillips (1986) took strong issue with this conclusion. Based on a more general model and on an analytical rather than a Monte Carlo approach, he suggested that the normalized statistic $t_\beta/\sqrt{T}$ be used for testing purposes rather than t_β itself. For the 50 observations used by Granger and Newbold, the appropriate critical value would be close to 15! If anything, Granger and Newbold were too optimistic

The random walk with drift,

$$z_t = \mu + z_{t-1} + \varepsilon_t, \tag{21-1}$$

and the **trend stationary process**,

$$z_t = \mu + \beta t + \varepsilon_t, \tag{21-2}$$

where, in both cases, ε_t is a white noise process, appear to be reasonable characterizations of many macroeconomic time series.[3] Clearly, both of these will produce strongly trended,

[3]The analysis to follow has been extended to more general disturbance processes, but that complicates matters substantially. In this case, in fact, our assumption does cost considerable generality, but the extension is beyond the scope of our work. Some references on the subject are Phillips and Perron (1988) and Davidson and MacKinnon (1993).

nonstationary series,[4] so it is not surprising that regressions involving such variables almost always produce significant relationships. The strong correlation would seem to be a consequence of the underlying trend, whether or not there really is any regression at work. But Granger and Newbold went a step further. The intuition is less clear if there is a pure **random walk** at work,

$$z_t = z_{t-1} + \varepsilon_t, \tag{21-3}$$

but even here, they found that regression "relationships" appear to persist even in unrelated series.

Each of these three series is characterized by a **unit root**. In each case, the **data-generating process (DGP)** can be written

$$(1 - L)z_t = \alpha + v_t, \tag{21-4}$$

where $\alpha = \mu$, β, and 0, respectively, and v_t is a stationary process. Thus, the characteristic equation has a single root equal to one, hence the name. The upshot of Granger and Newbold's and Phillips's findings is that the use of data characterized by unit roots has the potential to lead to serious errors in inferences.

In all three settings, differencing or detrending would seem to be a natural first step. On the other hand, it is not going to be immediately obvious which is the correct way to proceed—the data are strongly trended in all three cases—and taking the incorrect approach will not necessarily improve matters. For example, first differencing in (21-1) or (21-3) produces a white noise series, but first differencing in (21-2) trades the trend for autocorrelation in the form of an MA(1) process. On the other hand, detrending—that is, computing the residuals from a regression on time—is obviously counterproductive in (21-1) and (21-3), even though the regression of z_t on a trend will appear to be significant for the reasons we have been discussing, whereas detrending in (21-2) appears to be the right approach.[5] Because none of these approaches is likely to be obviously preferable at the outset, some means of choosing is necessary. Consider nesting all three models in a single equation,

$$z_t = \mu + \beta t + z_{t-1} + \varepsilon_t.$$

Now subtract z_{t-1} from both sides of the equation and introduce the artificial parameter γ, Then,

$$\begin{aligned} z_t - z_{t-1} &= \mu\gamma + \beta\gamma t + (\gamma - 1)z_{t-1} + \varepsilon_t \\ &= \alpha_0 + \alpha_1 t + (\gamma - 1)z_{t-1} + \varepsilon_t, \end{aligned} \tag{21-5}$$

where, by hypothesis, $\gamma = 1$. Equation (21-5) provides the basis for a variety of tests for unit roots in economic data. In principle, a test of the hypothesis that $\gamma - 1$ equals zero gives confirmation of the random walk with drift, because if γ equals 1 (and α_1 equals zero), then (21-1) results. If $\gamma - 1$ is less than zero, then the evidence favors the trend stationary (or some other) model, and detrending (or some alternative) is the preferable

[4]The constant term μ produces the deterministic trend in the random walk with drift. For convenience, suppose that the process starts at time zero. Then $z_t = \sum_{s=0}^{t}(\mu + \varepsilon_s) = \mu t + \sum_{s=0}^{t}\varepsilon_s$. Thus, z_t consists of a deterministic trend plus a stochastic trend consisting of the sum of the innovations. The result is a variable with increasing variance around a linear trend.

[5]See Nelson and Kang (1984).

approach. The practical difficulty is that standard inference procedures based on least squares and the familiar test statistics are not valid in this setting. The issue is discussed in the next section.

21.2.4 TESTS FOR UNIT ROOTS IN ECONOMIC DATA

The implications of unit roots in macroeconomic data are, at least potentially, profound. If a structural variable, such as real output, is truly $I(1)$, then shocks to it will have permanent effects. If confirmed, then this observation would mandate some rather serious reconsideration of the analysis of macroeconomic policy. For example, the argument that a change in monetary policy could have a transitory effect on real output would vanish.[6] The literature is not without its skeptics, however. This result rests on a razor's edge. Although the literature is thick with tests that have failed to reject the hypothesis that $\gamma = 1$, many have also not rejected the hypothesis that $\gamma \geq 0.95$, and at 0.95 (or even at 0.99), the entire issue becomes moot.[7]

Consider the simple AR(1) model with zero-mean, white noise innovations,

$$y_t = \gamma y_{t-1} + \varepsilon_t.$$

The downward bias of the least squares estimator when γ approaches one has been widely documented.[8] For $|\gamma| < 1$, however, the least squares estimator,

$$c = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2},$$

does have

$$\text{plim } c = \gamma$$

and

$$\sqrt{T}(c - \gamma) \xrightarrow{d} N[0, 1 - \gamma^2].$$

Does the result hold up if $\gamma = 1$? The case is called the unit root case, because in the ARMA representation $C(L)y_t = \varepsilon_t$, the characteristic equation $1 - \gamma z = 0$ has one root equal to one. That the limiting variance appears to go to zero should raise suspicions. The literature on the question dates back to Mann and Wald (1943) and Rubin (1950). But for econometric purposes, the literature has a focal point at the celebrated papers of Dickey and Fuller (1979, 1981). They showed that if γ equals one, then

$$T(c - \gamma) \xrightarrow{d} v,$$

where v is a random variable with finite, positive variance, and in finite samples, $E[c] < 1$.[9]

There are two important implications in the Dickey–Fuller results. First, the estimator of γ is biased downward if γ equals one. Second, the OLS estimator of γ converges to its

[6] The 1980s saw the appearance of literally hundreds of studies, both theoretical and applied, of unit roots in economic data. An important example is the seminal paper by Nelson and Plosser (1982). There is little question but that this observation is an early part of the radical paradigm shift that has occurred in empirical macroeconomics.

[7] A large number of issues are raised in Maddala (1992, pp. 582–588).

[8] See, for example, Evans and Savin (1981, 1984).

[9] A full derivation of this result is beyond the scope of this book. For the interested reader, a fairly comprehensive treatment at an accessible level is given in Chapter 17 of Hamilton (1994, pp. 475–542).

probability limit more rapidly than the estimators to which we are accustomed. That is, the variance of c under the null hypothesis is $O(1/T^2)$, not $O(1/T)$. (In a mean squared error sense, the OLS estimator is superconsistent.) It turns out that the implications of this finding for the regressions with trended data are considerable.

We have already observed that in some cases, differencing or detrending is required to achieve stationarity of a series. Suppose, though, that the preceding AR(1) model is fit to an $I(1)$ series, despite that fact. The upshot of the preceding discussion is that the conventional measures will tend to hide the true value of γ; the sample estimate is biased downward, and by dint of the very small *true* sampling variance, the conventional t test will tend, incorrectly, to reject the hypothesis that $\gamma = 1$. The practical solution to this problem devised by Dickey and Fuller was to derive, through Monte Carlo methods, an appropriate set of critical values for testing the hypothesis that γ equals one in an AR(1) regression when there truly is a unit root. One of their general results is that the test may be carried out using a conventional t statistic, but the critical values for the test must be revised: The standard t table is inappropriate. A number of variants of this form of testing procedure have been developed. We will consider several of them.

21.2.5 THE DICKEY–FULLER TESTS

The simplest version of the model to be analyzed is the random walk,

$$y_t = \gamma y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N[0, \sigma^2], \quad \text{and} \quad \text{Cov}[\varepsilon_t, \varepsilon_s] = 0 \,\forall\, t \neq s.$$

Under the null hypothesis that $\gamma = 1$, there are two approaches to carrying out the test. The conventional t ratio, $\text{DF}_\tau = (\hat{\gamma} - 1)/\text{Est. Std. Error}(\hat{\gamma})$, with the revised set of critical values may be used for a one-sided test. Critical values for this test are shown in the top panel of Table 21.2. Note that, in general, the critical value is considerably larger in absolute value than its counterpart from the t distribution. The second approach is based on the statistic $\text{DF}_\gamma = T(\hat{\gamma} - 1)$. Critical values for this test are shown in the top panel of Table 21.3.

The simple random walk model is inadequate for many series. Consider the rate of inflation from 1950.2 to 2000.4 (plotted in Figure 21.4) and the log of GDP over the same period (plotted in Figure 21.5). The first of these may be a random walk, but it is clearly drifting. The log GDP series, in contrast, has a strong trend. For the first of these, a random walk with drift may be specified,

$$y_t = \mu + z_t,$$
$$z_t = \gamma z_{t-1} + \varepsilon_t,$$

or

$$y_t = \mu(1 - \gamma) + \gamma y_{t-1} + \varepsilon_t.$$

For the second type of series, we may specify the trend stationary form,

$$y_t = \mu + \beta t + z_t,$$
$$z_t = \gamma z_{t-1} + \varepsilon_t$$

or

$$y_t = [\mu(1 - \gamma) + \gamma\beta] + \beta(1 - \gamma)t + \gamma y_{t-1} + \varepsilon_t.$$

TABLE 21.2 Critical Values for the Dickey–Fuller DFτ Test.

	Sample Size			
	25	***50***	***100***	***∞***
F ratio (D–F)[a]	7.24	6.73	6.49	6.25
F ratio (standard)	3.42	3.20	3.10	3.00
AR model[b] (random walk)				
0.01	−2.66	−2.62	−2.60	−2.58
0.025	−2.26	−2.25	−2.24	−2.23
0.05	−1.95	−1.95	−1.95	−1.95
0.10	−1.60	−1.61	−1.61	−1.62
0.975	1.70	1.66	1.64	1.62
AR model with constant (random walk with drift)				
0.01	−3.75	−3.59	−3.50	−3.42
0.025	−3.33	−3.23	−3.17	−3.12
0.05	−2.99	−2.93	−2.90	−2.86
0.10	−2.64	−2.60	−2.58	−2.57
0.975	0.34	0.29	0.26	0.23
AR model with constant and time trend (trend stationary)				
0.01	−4.38	−4.15	−4.04	−3.96
0.025	−3.95	−3.80	−3.69	−3.66
0.05	−3.60	−3.50	−3.45	−3.41
0.10	−3.24	−3.18	−3.15	−3.13
0.975	−0.50	−0.58	−0.62	−0.66

[a]From Dickey and Fuller (1981, p. 1063). Degrees of freedom are 2 and $T - p - 3$.
[b]From Fuller (1976, p. 373 and 1996, Table 10.A.2).

The tests for these forms may be carried out in the same fashion. For the model with drift only, the center panels of Tables 21.2 and 21.3 are used. When the trend is included, the lower panel of each table is used.

Example 21.2 Tests for Unit Roots

Cecchetti and Rich (2001) studied the effect of monetary policy on the U.S. economy. The data used in their study were the following variables:

π = one period rate of inflation = the rate of change in the CPI,
y = log of real GDP,
i = nominal interest rate = the quarterly average yield on a 90-day T-bill,
Δm = change in the log of the money stock, M1,
$i - \pi$ = ex-post real interest rate,
$\Delta m - \pi$ = real growth in the money stock,

Data used in their analysis were from the period 1959.1 to 1997.4. As part of their analysis, they checked each of these series for a unit root and suggested that the hypothesis of a unit root could only be rejected for the last two variables. We will reexamine these data for the longer interval, 1950II to 2000IV. The data are in Appendix Table F5.2. Figures 21.6–21.9 show

TABLE 21.3 Critical Values for the Dickey–Fuller DFτ Test.

	Sample Size			
	25	***50***	***100***	***∞***
AR model[a] (random walk)				
0.01	−11.8	−12.8	−13.3	−13.8
0.025	−9.3	−9.9	−10.2	−10.5
0.05	−7.3	−7.7	−7.9	−8.1
0.10	−5.3	−5.5	−5.6	−5.7
0.975	1.78	1.69	1.65	1.60
AR model with constant (random walk with drift)				
0.01	−17.2	−18.9	−19.8	−20.7
0.025	−14.6	−15.7	−16.3	−16.9
0.05	−12.5	−13.3	−13.7	−14.1
0.10	−10.2	−10.7	−11.0	−11.3
0.975	0.65	0.53	0.47	0.41
AR model with constant and time trend (trend stationary)				
0.01	−22.5	−25.8	−27.4	−29.4
0.025	−20.0	−22.4	−23.7	−24.4
0.05	−17.9	−19.7	−20.6	−21.7
0.10	−15.6	−16.8	−17.5	−18.3
0.975	−1.53	−1.667	−1.74	−1.81

[a] From Fuller (1976, p. 373 and 1996, Table 10.A.1).

FIGURE 21.4 Rate of Inflation in the Consumer Price Index.

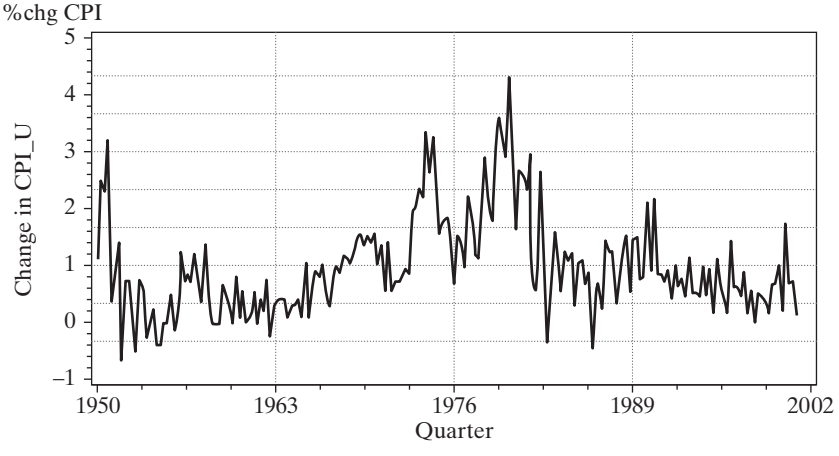
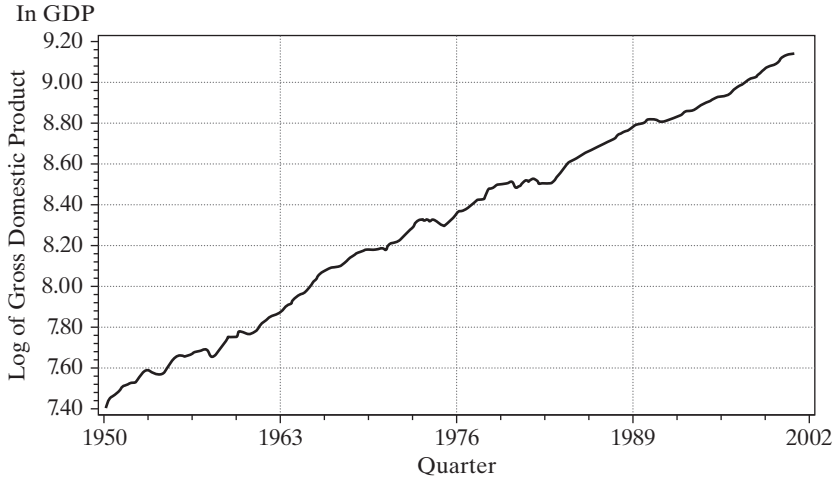


FIGURE 21.5 Log of Gross Domestic Product.



the behavior of the last four variables. The first two are shown in Figures 21.4 and 21.5. Only the real output figure shows a strong trend, so we will use the random walk with drift for all the variables except this one.

The Dickey–Fuller tests are carried out in Table 21.4. There are 203 observations used in each one. The first observation is lost when computing the rate of inflation and the change in the money stock, and one more is lost for the difference term in the regression. The critical values from interpolating to the second row, last column in each panel for 95% significance

FIGURE 21.6 T-Bill Rate.

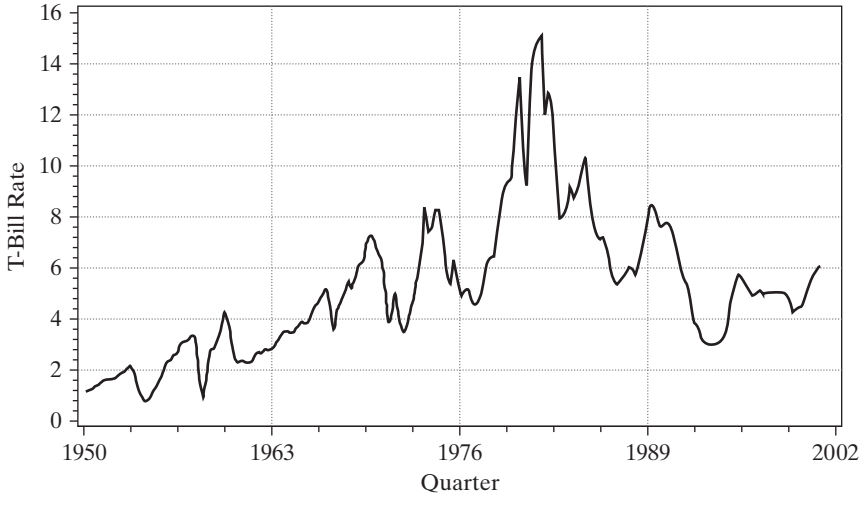
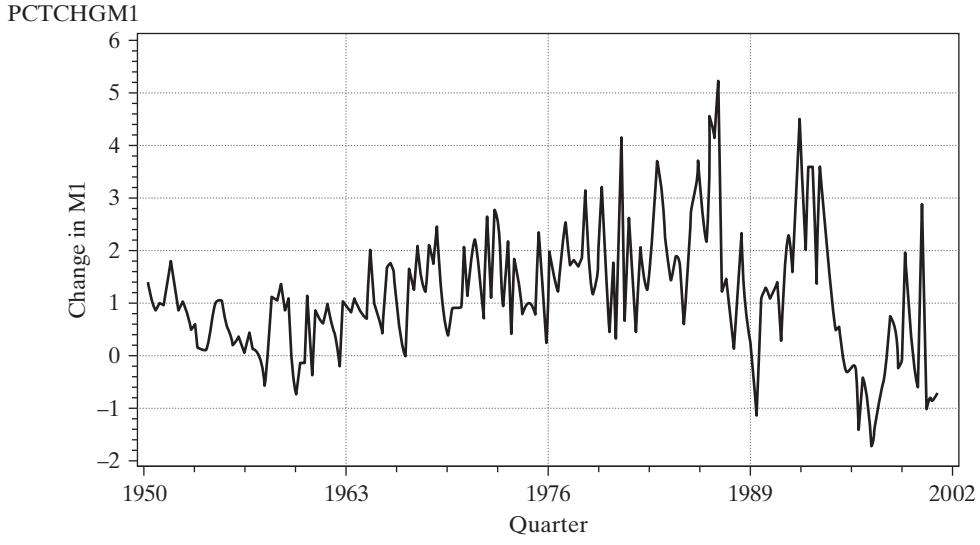


FIGURE 21.7 Percentage Change in the Money Stock.



and a one-tailed test are -3.68 and -24.2, respectively, for DF_τ and DF_γ for the output equation, which contains the time trend, and -3.14 and -16.8 for the other equations, which contain a constant but no trend. For the output equation (y), the test statistics are

$$DF_\tau = \frac{0.9584940384 - 1}{.017880922} = -2.32 > -3.44,$$

FIGURE 21.8 Ex-Post Real T-Bill Rate.

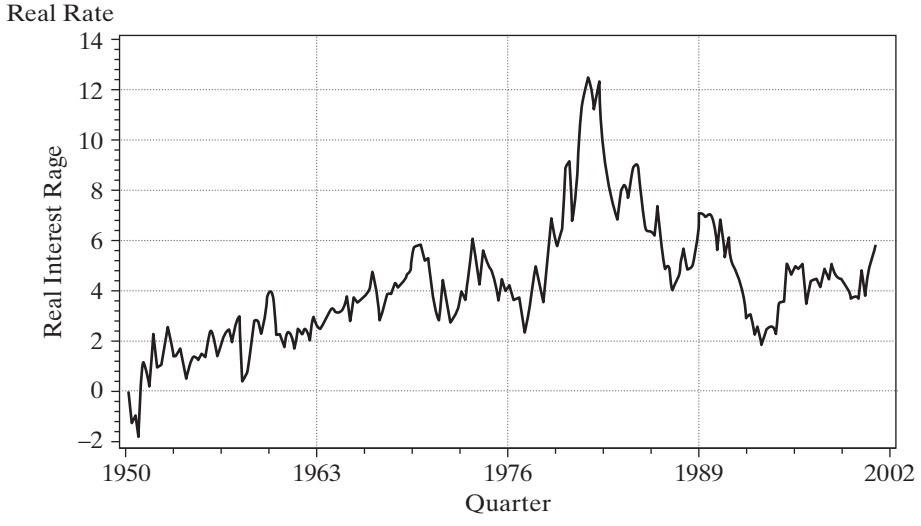
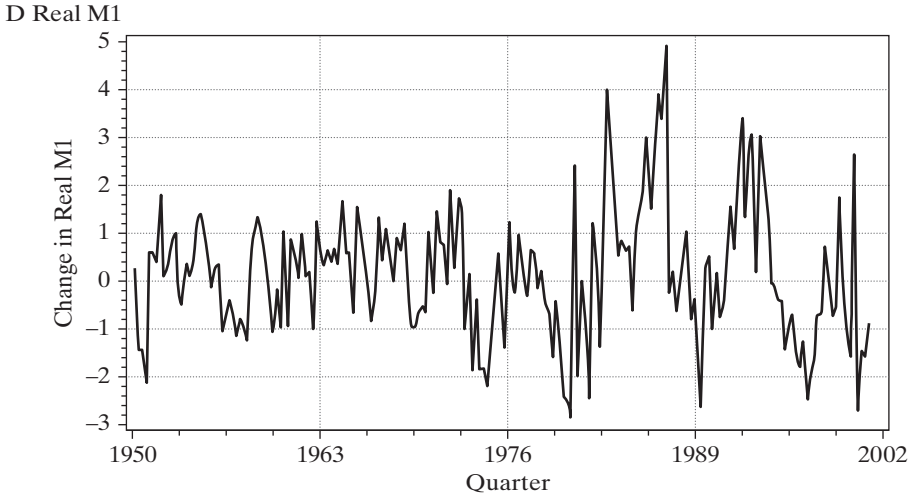


FIGURE 21.9 Change in the Real Money Stock.



and

$$\text{DF}_\gamma = 202(0.9584940384 - 1) = -8.38 > -21.2.$$

Neither is less than the critical value, so we conclude (as have others) that there is a unit root in the log GDP process. The results of the other tests are shown in Table 21.4. Surprisingly, these results do differ sharply from those obtained by Cecchetti and Rich (2001) for π and Δm. The sample period appears to matter; if we repeat the computation using Cecchetti and Rich's interval, 1959.4 to 1997.4, then DF_τ equals -3.51. This is borderline, but less contradictory. For Δm, we obtain a value of -4.204 for DF_τ when the sample is restricted to the shorter interval.

TABLE 21.4 Unit Root Tests (standard errors of estimates in parentheses).

	μ	β	γ	DF_τ	DF_γ	*Conclusion*
π	0.332		0.659	−6.40	−68.88	Reject H_0
	(0.0696)		(0.0532)	$R^2 = 0.432$	$s = 0.643$	
y	0.320	0.00033	0.958	−2.35	−8.48	Do not reject H_0
	(0.134)	(0.00015)	(0.0179)	$R^2 = 0.999$	$s = 0.001$	
i	0.228		0.961	−2.14	−7.88	Do not reject H_0
	(0.109)		(0.0182)	$R^2 = 0.933$	$s = 0.743$	
Δm	0.448		0.596	−7.05	−81.61	Reject H_0
	(0.0923)		(0.0573)	$R^2 = 0.351$	$s = 0.929$	
$i - \pi$	0.615		0.557	−7.57	−89.49	Reject H_0
	(0.185)		(0.0585)	$R^2 = 0.311$	$s = 2/395$	
$\Delta m - \pi$	0.0700		0.490	−8.25t	−103.02	Reject H_0
	(0.0833)		(0.0618)	$R^2 = 0.239$	$s = 1.176$	

The Dickey–Fuller tests described in this section assume that the disturbances in the model as stated are white noise. An extension which will accommodate some forms of serial correlation is the **augmented Dickey–Fuller test**. The augmented Dickey–Fuller test is the same one as described earlier, carried out in the context of the model

$$y_t = \mu + \beta t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t.$$

The random walk form is obtained by imposing $\mu = 0$ and $\beta = 0$; the random walk with drift has $\beta = 0$; and the trend stationary model leaves both parameters free. The two test statistics are

$$\text{DF}_\tau = \frac{\hat{\gamma} - 1}{\text{Est. Std. Error}(\hat{\gamma})},$$

exactly as constructed before, and

$$\text{DF}_\gamma = \frac{T(\hat{\gamma} - 1)}{1 - \hat{\gamma}_1 - \cdots - \hat{\gamma}_p}.$$

The advantage of this formulation is that it can accommodate higher-order autoregressive processes in ε_t.

An alternative formulation may prove convenient. By subtracting y_{t-1} from both sides of the equation, we obtain

$$\Delta y_t = \mu + \beta t + \gamma^* y_{t-1} + \sum_{j=1}^{p} \phi_j \Delta y_{t-j} + \varepsilon_t,$$

where

$$\phi_j = -\sum_{k=j+1}^{p} \gamma_k \qquad \text{and} \qquad \gamma^* = \left(\sum_{i=1}^{p} \gamma_i \right) - 1.$$

The unit root test is carried out as before by testing the null hypothesis $\gamma^* = 0$ against $\gamma^* < 0$.[10] The t test, DF_τ, may be used. If the failure to reject the unit root is taken as evidence that a unit root is present, that is, $\gamma^* = 0$, then the model specializes to the AR($p - 1$) model in the first differences, which is an ARIMA($p - 1, 1, 0$) model for y_t. For a model with a time trend,

$$\Delta y_t = \mu + \beta t + \gamma^* y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \varepsilon_t,$$

the test is carried out by testing the joint hypothesis that $\beta = \gamma^* = 0$. Dickey and Fuller (1981) present counterparts to the critical F statistics for testing the hypothesis. Some of their values are reproduced in the first row of Table 21.2. (Authors frequently focus on γ^* and ignore the time trend, maintaining it only as part of the appropriate formulation. In this case, one may use the simple test of $\gamma^* = 0$ as before, with the DF_τ critical values.)

The lag length, p, remains to be determined. As usual, we are well advised to test down to the right value instead of up. One can take the familiar approach and sequentially examine the t statistic on the last coefficient—the usual t test is appropriate.

[10] It is easily verified that one of the roots of the characteristic polynomial is $1/(\gamma_1 + \gamma_2 + \ldots + \gamma_P)$.

An alternative is to combine a measure of model fit, such as the regression s^2, with one of the information criteria. The Akaike and Schwarz (Bayesian) information criteria would produce the two information measures

$$IC(p) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{T - p_{\max} - K^*}\right) + (p + K^*)\left(\frac{A^*}{T - p_{\max} - K^*}\right),$$

$K^* =$ 1 for random walk, 2 for random walk with drift, 3 for trend stationary,

$A^* =$ 2 for Akaike criterion, $\ln(T - p_{\max} - K^*)$ for Bayesian criterion,

$p_{\max} =$ the largest lag length being considered.

The remaining detail is to decide upon $p_{\max}$. The theory provides little guidance here. On the basis of a large number of simulations, Schwert (1989) found that

$$p_{\max} = \text{integer part of } [12 \times (T/100)^{.25}]$$

gave good results.

Many alternatives to the Dickey–Fuller tests have been suggested, in some cases to improve on the finite sample properties and in others to accommodate more general modeling frameworks. The Phillips (1987) and Phillips and Perron (1988) statistic may be computed for the same three functional forms,

$$y_t = \delta_t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t, \tag{21-6}$$

where δ_t may be $0, \mu$, or $\mu + \beta t$. The procedure modifies the two Dickey–Fuller statistics we previously examined,

$$Z_\tau = \sqrt{\frac{c_0}{a}}\left(\frac{\hat{\gamma} - 1}{v}\right) - \frac{1}{2}(a - c_0)\frac{Tv}{\sqrt{as^2}},$$

$$Z_\gamma = \frac{T(\hat{\gamma} - 1)}{1 - \hat{\gamma}_1 - \cdots - \hat{\gamma}_p} - \frac{1}{2}\left(\frac{T^2 v^2}{s^2}\right)(a - c_0),$$

where

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - K},$$

$v^2 =$ estimated asymptotic variance of $\hat{\gamma}$,

$$c_j = \frac{1}{T}\sum_{s=j+1}^{T} e_t e_{t-s}, \quad j = 0, \ldots, L = j\text{th autocovariance of residuals},$$

$$c_0 = [(T - K)/T]s^2,$$

$$a = c_0 + 2\sum_{j=1}^{L}\left(1 - \frac{j}{L + 1}\right)c_j.$$

[Note the Newey–West (Bartlett) weights in the computation of a. As before, the analyst must choose L.] The test statistics are referred to the same Dickey–Fuller tables we have used before.

Elliot, Rothenberg, and Stock (1996) have proposed a method they denote the ADF-GLS procedure, which is designed to accommodate more general formulations of ε; the process generating ε_t is assumed to be an $I(0)$ stationary process, possibly an ARMA(r, s). The null hypothesis, as before, is $\gamma = 1$ in (21-6) where $\delta_t = \mu$ or $\mu + \beta t$. The method proceeds as follows:

Step 1. Linearly regress

$$\mathbf{y}^* = \begin{bmatrix} y_1 \\ y_2 - \bar{r}y_1 \\ \cdots \\ y_T - \bar{r}y_{T-1} \end{bmatrix} \quad \text{on} \quad \mathbf{X}^* = \begin{bmatrix} 1 \\ 1 - \bar{r} \\ \cdots \\ 1 - \bar{r} \end{bmatrix} \quad \text{or} \quad \mathbf{X}^* = \begin{bmatrix} 1 & 1 \\ 1 - \bar{r} & 2 - \bar{r} \\ \cdots & \\ 1 - \bar{r} & T - \bar{r}(T-1) \end{bmatrix}$$

for the random walk with drift and trend stationary cases, respectively. (Note that the second column of the matrix is simply $\bar{r} + (1 - \bar{r})t$.) Compute the residuals from this regression, $\tilde{y}_t = y_t - \hat{\delta}_t$. $\bar{r} = 1 - 7/T$ for the random walk model and $1 - 13.5/T$ for the model with a trend.

Step 2. The Dickey–Fuller DF_τ test can now be carried out using the model

$$\tilde{y}_t = \gamma\tilde{y}_{t-1} + \gamma_1\Delta\tilde{y}_{t-1} + \cdots + \gamma_p\Delta\tilde{y}_{t-p} + \eta_t.$$

If the model does not contain the time trend, then the t statistic for $(\gamma - 1)$ may be referred to the critical values in the center panel of Table 21.2. For the trend stationary model, the critical values are given in a table presented in Elliot et al. The 97.5% critical values for a one-tailed test from their table is −3.15.

As in many such cases of a new technique, as researchers develop large and small modifications of these tests, the practitioner is likely to have some difficulty deciding how to proceed. The Dickey–Fuller procedures have stood the test of time as robust tools that appear to give good results over a wide range of applications. The **Phillips–Perron tests** are very general but appear to have less than optimal small sample properties. Researchers continue to examine it and the others such as the Elliot et al. method. Other tests are catalogued in Maddala and Kim (1998).

Example 21.3 Augmented Dickey–Fuller Test for a Unit Root in GDP

Dickey and Fuller (1981) apply their methodology to a model for the log of a quarterly series on output, the Federal Reserve Board Production Index. The model used is

$$y_t = \mu + \beta t + \gamma y_{t-1} + \phi(y_{t-1} - y_{t-2}) + \varepsilon_t. \tag{21-7}$$

The test is carried out by testing the joint hypothesis that both β and γ^* are zero in the model

$$y_t - y_{t-1} = \mu^* + \beta t + \gamma^* y_{t-1} + \phi(y_{t-1} - y_{t-2}) + \varepsilon_t.$$

(If $\gamma = 0$, then μ^* will also by construction.) We will repeat the study with our data on real GDP from Appendix Table F5.2 using observations 1950.1–2000.4.

We will use the augmented Dickey–Fuller test first. Thus, the first step is to determine the appropriate lag length for the augmented regression. Using Schwert's suggestion, we find that the maximum lag length should be allowed to reach p_{max} = [integer part of $12(204/100)^{0.25}$] = 14.

The specification search uses observations 18 to 204, because as many as 17 coefficients will be estimated in the equation

$$y_t = \mu + \beta t + \gamma y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + \varepsilon_t.$$

In the sequence of 14 regressions with $j = 14, 13, \ldots$, the only statistically significant lagged difference is the first one, in the last regression, so it would appear that the model used by Dickey and Fuller would be chosen on this basis. The two information criteria produce a similar conclusion. Both of them decline monotonically from $j = 14$ all the way down to $j = 1$, so on this basis, we end the search with $j = 1$, and proceed to analyze Dickey and Fuller's model.

The linear regression results for the equation in (21-7) are

$$\begin{aligned} y_t = {} & \underset{(0.125)}{0.368} + \underset{(0.000138)}{0.000391}t + \underset{(0.0167)}{0.952}y_{t-1} + \underset{(0.0647)}{0.36025}\Delta y_{t-1} + e_t, \quad s = 0.00912 \\ & R^2 = 0.999647. \end{aligned}$$

The two test statistics are

$$\text{DF}_\tau = \frac{0.95166 - 1}{0.016716} = -2.892$$

and

$$\text{DF}_\gamma = \frac{201(0.95166 - 1)}{1 - 0.36025} = -15.263.$$

Neither statistic is less than the respective critical value, −3.70 and −24.5. On this basis, we conclude, as have many others, that there is a unit root in log GDP.

For the Phillips and Perron statistic, we need several additional intermediate statistics. Following Hamilton (1994, p. 512), we choose $L = 4$ for the long-run variance calculation. Other values we need are $T = 202$, $\hat{\gamma} = 0.9516613$, $s^2 = 0.00008311488$, $v^2 = 0.00027942647$, and the first five autocovariances, $c_0 = 0.000081469$, $c_1 = -0.00000351162$, $c_2 = 0.00000688053$, $c_3 = 0.000000597305$, and $c_4 = -0.00000128163$. Applying these to the weighted sum produces $a = 0.0000840722$, which is only a minor correction to c_0. Collecting the results, we obtain the Phillips–Perron statistics, $Z_\tau = -2.89921$ and $Z_\gamma = -15.44133$. Because these are applied to the same critical values in the Dickey–Fuller tables, we reach the same conclusion as before—we do not reject the hypothesis of a unit root in log GDP.

21.2.6 THE KPSS TEST OF STATIONARITY

Kwiatkowski et al. (1992) (KPSS) have devised an alternative to the Dickey–Fuller test for stationarity of a time series. The procedure is a test of nonstationarity against the null hypothesis of stationarity in the model

$$\begin{aligned} y_t &= \alpha + \beta t + \gamma \sum_{i=1}^{t} z_i + \varepsilon_t, \qquad t = 1, \ldots, T \\ &= \alpha + \beta t + \gamma Z_t + \varepsilon_t, \end{aligned}$$

where ε_t is a stationary series and z_t is an i.i.d. stationary series with mean zero and variance one. (These are merely convenient normalizations because a nonzero mean would move to α and a nonunit variance is absorbed in γ.) If γ equals zero, then the process is stationary if $\beta = 0$ and trend stationary if $\beta \neq 0$. Because Z_t, is $I(1)$, y_t is nonstationary if γ is nonzero.

The KPSS test of the null hypothesis, H_0: $\gamma = 0$, against the alternative that γ is nonzero reverses the strategy of the Dickey–Fuller statistic (which tests the null hypothesis $\gamma < 1$ against the alternative $\gamma = 1$). Under the null hypothesis, α and β can be estimated by OLS. Let e_t denote the tth OLS residual,

$$e_t = y_t - a - bt,$$

and let the sequence of partial sums be

$$E_t = \sum_{s=1}^{t} e_s, \quad t = 1, \ldots, T.$$

(Note $E_T = 0$.) The KPSS statistic is

$$\text{KPSS} = \frac{\sum_{t=1}^{T} E_t^2}{T^2 \hat{\sigma}^2},$$

where

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} e_t^2}{T} + 2 \sum_{j=1}^{L} \left(1 - \frac{j}{L+1}\right) r_j,$$

$$r_j = \frac{\sum_{s=j+1}^{T} e_s e_{s-j}}{T},$$

and L is chosen by the analyst. [See (20-17).] Under normality of the disturbances, ε_t, the KPSS statistic is an LM statistic. The authors derive the statistic under more general conditions. Critical values for the test statistic are estimated by simulation. The 0.05 upper-tail values reported by the authors (in their Table 1, p. 166) for $\beta = 0$ and $\beta \neq 0$ are 0.463 and 0.146, respectively.

Example 21.4 Is There a Unit Root in GDP?

Using the data used for the Dickey–Fuller tests in Example 21.3, we repeated the procedure using the KPSS test with $L = 10$. The two statistics are 1.953 without the trend and 0.312 with it. Comparing these results to the values in Table 21.4 we conclude (again) that there is, indeed, a unit root in ln GDP. Or, more precisely, we conclude that ln GDP is not a stationary series, nor even a trend stationary series.

21.3 COINTEGRATION

Studies in empirical macroeconomics almost always involve nonstationary and trending variables, such as income, consumption, money demand, the price level, trade flows, and exchange rates. Accumulated wisdom and the results of the previous sections suggest that the appropriate way to manipulate such series is to use differencing and other transformations (such as seasonal adjustment) to reduce them to stationarity and then to analyze the resulting series as VARs or with the methods of Box and Jenkins (1984). But recent research and a growing literature have shown that there are more interesting, appropriate ways to analyze trending variables.

In the *fully specified* regression model,

$$y_t = \beta x_t + \varepsilon_t,$$

there is a presumption that the disturbances ε_t are a stationary, white noise series.[11] But this presumption is unlikely to be true if y_t and x_t are integrated series. Generally, if two series are integrated to different orders, then linear combinations of them will be integrated to the higher of the two orders. Thus, if y_t and x_t are $I(1)$—that is, if both are trending variables—then we would normally expect $y_t - \beta x_t$ to be $I(1)$ regardless of the value of β, not $I(0)$ (i.e., not stationary). If y_t and x_t are each drifting upward with their own trend, then unless there is some relationship between those trends, the difference between them should also be growing, with yet another trend. There must be some kind of inconsistency in the model. On the other hand, if the two series are both $I(1)$, then there *may* be a β such that

$$\varepsilon_t = y_t - \beta x_t$$

is $I(0)$. Intuitively, if the two series are both $I(1)$, then this partial difference between them might be stable around a fixed mean. The implication would be that the series are drifting together at roughly the same rate. Two series that satisfy this requirement are said to be **cointegrated**, and the vector $[1, -\beta]$ (or any multiple of it) is a *cointegrating vector*. In such a case, we can distinguish between a long-run relationship between y_t and x_t, that is, the manner in which the two variables drift upward together, and the short-run dynamics, that is, the relationship between deviations of y_t from its long-run trend and deviations of x_t from its long-run trend. If this is the case, then differencing of the data would be counterproductive, because it would obscure the long-run relationship between y_t and x_t. Studies of cointegration and a related technique, error correction, are concerned with methods of estimation that preserve the information about both forms of covariation.[12]

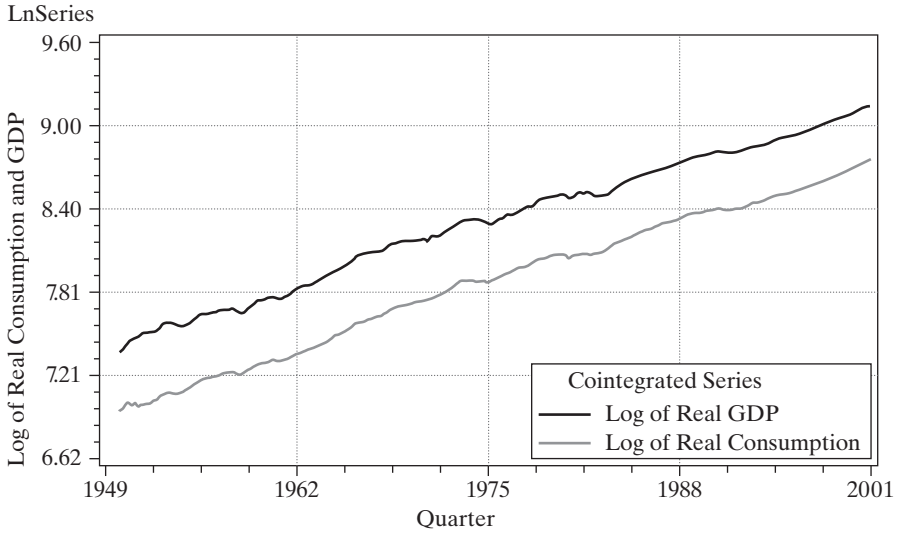
Example 21.5 Cointegration in Consumption and Output

Consumption and income provide one of the more familiar examples of the phenomenon described previously. The logs of GDP and consumption for 1950.1 to 2000.4 are plotted in Figure 21.10. Both variables are obviously nonstationary. We have already verified that there is a unit root in the income data. We leave as an exercise for the reader to verify that the consumption variable is likewise $I(1)$. Nonetheless, there is a clear relationship between consumption and output. Consider a simple regression of the log of consumption on the log of income, where both variables are manipulated in mean deviation form (so, the regression includes a constant). The slope in that regression is 1.056765. The residuals from the regression, $u_t =$ [ln Cons*, ln GDP*][1, −1.056765]′ (where the "*" indicates mean deviations) are plotted in Figure 21.11. The trend is clearly absent from the residuals. But it remains to verify whether the series of residuals is stationary. In the ADF regression of the least squares residuals on a constant (random walk with drift), the lagged value and the lagged first difference, the coefficient on u_{t-1} is 0.838488 (0.0370205) and that on $u_{t-1} - u_{t-2}$ is −0.098522. (The constant differs trivially from zero because two observations are lost in computing the ADF regression.) With 202 observations, we find $DF_\tau = -4.63$ and $DF_\gamma = -29.55$. Both are well below the critical values, which suggests that the residual series does not contain a unit root. We conclude (at least it appears so) that even after

[11] Any autocorrelation in the model has been removed through an appropriate transformation.

[12] See, for example, Engle and Granger (1987) and the lengthy literature cited in Hamilton (1994). A survey paper on VARs and cointegration is Watson (1994).

FIGURE 21.10 Cointegrated Variables: Logs of Consumption and GDP.

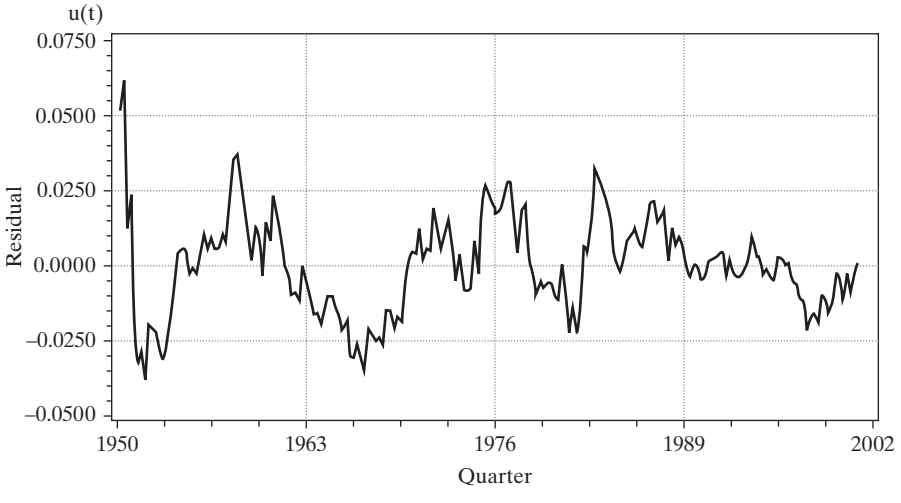


accounting for the trend, although neither of the original variables is stationary, there is a linear combination of them that is. If this conclusion holds up after a more formal treatment of the testing procedure, we will conclude that log GDP and log consumption are cointegrated.

Example 21.6 Several Cointegrated Series

The theory of purchasing power parity specifies that in long-run equilibrium, exchange rates will adjust to erase differences in purchasing power across different economies. Thus, if

FIGURE 21.11 Residuals from Consumption—Income Regression.



p_1 and p_0 are the price levels in two countries and E is the exchange rate between the two currencies, then in equilibrium,

$$v_t = E_t \frac{p_{1t}}{p_{0t}} = \mu, \qquad \text{a constant.}$$

The price levels in any two countries are likely to be strongly trended. But allowing for short-term deviations from equilibrium, the theory suggests that for a particular $\boldsymbol{\beta} = (\ln \mu, -1, 1)'$, in the model

$$\ln E_t = \beta_1 + \beta_2 \ln p_{1t} + \beta_3 \ln p_{0t} + \varepsilon_t,$$

$\varepsilon_t = \ln v_t$ would be a stationary series, which would imply that the logs of the three variables in the model are cointegrated.

We suppose that the model involves M variables, $\mathbf{y}_t = [y_{1t}, \ldots, y_{Mt}]'$, which individually may be $I(0)$ or $I(1)$, and a long-run equilibrium relationship,

$$\mathbf{y}_t'\boldsymbol{\gamma} - \mathbf{x}_t'\boldsymbol{\beta} = 0.$$

The regressors may include a constant, exogenous variables assumed to be $I(0)$, and/or a time trend. The vector of parameters $\boldsymbol{\gamma}$ is the **cointegrating vector**. In the short run, the system may deviate from its equilibrium, so the relationship is rewritten as

$$\boldsymbol{y}_t'\boldsymbol{\gamma} - \mathbf{x}_t'\boldsymbol{\beta} = \varepsilon_t,$$

where the **equilibrium error** ε_t must be a stationary series. In fact, because there are M variables in the system, at least in principle, there could be more than one cointegrating vector. In a system of M variables, there can only be up to $M - 1$ linearly independent cointegrating vectors. A proof of this proposition is very simple, but useful at this point.

Proof: Suppose that $\boldsymbol{\gamma}_i$ is a cointegrating vector and that there are M linearly independent cointegrating vectors. Then, neglecting $\mathbf{x}_t'\boldsymbol{\beta}$ for the moment, for every $\boldsymbol{\gamma}_i$, $\mathbf{y}_t'\boldsymbol{\gamma}_i$ is a stationary series ν_{ti}. Any linear combination of a set of stationary series is stationary, so it follows that every linear combination of the cointegrating vectors is also a cointegrating vector. If there are M such $M \times 1$ linearly independent vectors, then they form a basis for the M-dimensional space, so any $M \times 1$ vector can be formed from these cointegrating vectors, including the columns of an $M \times M$ identity matrix. Thus, the first column of an identity matrix would be a cointegrating vector, or y_{t1} is $I(0)$. This result is a contradiction, because we are allowing y_{t1} to be $I(1)$. It follows that there can be at most $M - 1$ cointegrating vectors.

The number of linearly independent cointegrating vectors that exist in the equilibrium system is called its **cointegrating rank**. The cointegrating rank may range from 1 to $M - 1$. If it exceeds one, then we will encounter an interesting identification problem. As a consequence of the observation in the preceding proof, we have the unfortunate result that, in general, *if the cointegrating rank of a system exceeds one,* then without out-of-sample, *exact* information, it is not possible to estimate behavioral relationships as cointegrating vectors. Enders (1995) provides a useful example.

Example 21.7 Multiple Cointegrating Vectors

We consider the logs of four variables, money demand m, the price level p, real income y, and an interest rate r. The basic relationship is

$$m = \gamma_0 + \gamma_1 p + \gamma_2 y + \gamma_3 r + \varepsilon.$$

The price level and real income are assumed to be $I(1)$. The existence of long-run equilibrium in the money market implies a cointegrating vector $\boldsymbol{\alpha}_1$. If the Fed follows a certain feedback rule, increasing the money stock when *nominal* income ($y + p$) is low and decreasing it when nominal income is high—which might make more sense in terms of rates of growth—then there is a second cointegrating vector in which $\gamma_1 = \gamma_2$ and $\gamma_3 = 0$. Suppose that we label this vector $\boldsymbol{\alpha}_2$. The parameters in the money demand equation, notably the interest elasticity, are interesting quantities, and we might seek to estimate $\boldsymbol{\alpha}_1$ to learn the value of this quantity. Because every linear combination of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ is a cointegrating vector, to this point we are only able to estimate a hash of the two cointegrating vectors.

In fact, the parameters of this model *are* identifiable from sample information (in principle). We have specified two cointegrating vectors,

$$\boldsymbol{\alpha}_1 = [1, -\gamma_{10}, -\gamma_{11}, -\gamma_{12}, -\gamma_{13}]'$$

and

$$\boldsymbol{\alpha}_2 = [1, -\gamma_{20}, \gamma_{21}, \gamma_{21}, 0]'.$$

Although it is true that every linear combination of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ is a cointegrating vector, only the original two vectors, as they are, have a 1 in the first position of both and a 0 in the last position of the second. (The equality restriction actually overidentifies the parameter matrix.) This result is, of course, exactly the sort of analysis that we used in establishing the identifiability of a simultaneous equations system in Chapter 10.

21.3.1 COMMON TRENDS

If two $I(1)$ variables are cointegrated, then some linear combination of them is $I(0)$. Intuition should suggest that the linear combination does not mysteriously create a well-behaved new variable; rather, something present in the original variables must be missing from the aggregated one. Consider an example. Suppose that two $I(1)$ variables have a linear trend,

$$y_{1t} = \alpha + \beta t + u_t,$$
$$y_{2t} = \gamma + \delta t + v_t,$$

where u_t and v_t are white noise. A linear combination of y_{1t} and y_{2t} with vector $(1, \theta)$ produces the new variable,

$$z_t = (\alpha + \theta\gamma) + (\beta + \theta\delta)t + u_t + \theta v_t,$$

which, in general, is still $I(1)$. In fact, the only way the z_t series can be made stationary is if $\theta = -\beta/\delta$. If so, then the effect of combining the two variables linearly is to remove the common linear trend, which is the basis of Stock and Watson's (1988) analysis of the problem. But their observation goes an important step beyond this one. The only way that y_{1t} and y_{2t} can be cointegrated to begin with is if they have a common trend of some sort. To continue, suppose that instead of the linear trend t, the terms on the left-hand side, y_1 and y_2, are functions of a random walk, $w_t = w_{t-1} + \eta_t$, where η_t is white noise. The analysis is identical. But now suppose that each variable y_{it} has its own random walk component

w_{it}, $i = 1, 2$. Any linear combination of y_{1t} and y_{2t} must involve *both* random walks. It is clear that they cannot be cointegrated unless, in fact, $w_{1t} = w_{2t}$. That is, once again, they must have a **common trend**. Finally, suppose that y_{1t} and y_{2t} share two common trends,

$$y_{1t} = \alpha + \beta t + \lambda w_t + u_t,$$
$$y_{2t} = \gamma + \delta t + \pi w_t + v_t.$$

We place no restriction on λ and π. Then, a bit of manipulation will show that it is not possible to find a linear combination of y_{1t} and y_{2t} that is cointegrated, even though they share common trends. The end result for this example is that if y_{1t} and y_{2t} are cointegrated, then they must share exactly one common trend.

As Stock and Watson determined, the preceding is the crux of the cointegration of economic variables. A set of M variables that are cointegrated can be written as a stationary component plus linear combinations of a smaller set of common trends. If the cointegrating rank of the system is r, then there can be up to $M - r$ linear trends and $M - r$ common random walks.[13] (The two-variable case is special. In a two-variable system, there can be only one common trend in total.) The effect of the cointegration is to purge these common trends from the resultant variables.

21.3.2 ERROR CORRECTION AND VAR REPRESENTATIONS

Suppose that the two $I(1)$ variables y_t and z_t are cointegrated and that the cointegrating vector is $[1, -\theta]$. Then all three variables, $\Delta y_t = y_t - y_{t-1}$, Δz_t, and $(y_t - \theta z_t)$ are $I(0)$. The error correction model,

$$\Delta y_t = \mathbf{x}_t'\boldsymbol{\beta} + \gamma(\Delta z_t) + \lambda(y_{t-1} - \theta z_{t-1}) + \varepsilon_t,$$

describes the variation in y_t around its long-run trend in terms of a set of $I(0)$ exogenous factors $\mathbf{x}_t$, the variation of z_t around its long-run trend, and the error correction $(y_t - \theta z_t)$, which is the equilibrium error in the model of cointegration. There is a tight connection between models of cointegration and models of error correction. The model in this form is reasonable as it stands, but in fact, it is only internally consistent if the two variables are cointegrated. If not, then the third term, and hence the right-hand side, cannot be $I(0)$, even though the left-hand side must be. The upshot is that the same assumption that we make to produce the cointegration implies (and is implied by) the existence of an error correction model.[14] As we will examine in the next section, the utility of this representation is that it suggests a way to build an elaborate model of the long-run variation in y_t as well as a test for cointegration. Looking ahead, the preceding suggests that residuals from an estimated cointegration model—that is, estimated equilibrium errors—can be included in an elaborate model of the long-run covariation of y_t and z_t. Once again, we have the foundation of Engel and Granger's approach to analyzing cointegration.

Pesaran, Shin, and Smith (2001) suggest a method of testing for a relationship in levels between a y_t and an $\mathbf{x}_t$ when there exist significant lags in the error correction form. Their **bounds test** accommodates the possibility that the regressors may be trend or difference stationary. The critical values they provide give a band that covers the polar cases in which all regressors are $I(0)$, or are $I(1)$, or are mutually cointegrated. The

[13] See Hamilton (1994, p. 578).

[14] The result in its general form is known as the Granger representation theorem. See Hamilton (1994, p. 582).

statistic is able to test for the existence of a levels equation regardless of whether the variables are $I(0)$, $I(1)$, or are cointegrated. In their application, y_t is real earnings in the UK while $\mathbf{x}_t$ includes a measure of productivity, the unemployment rate, unionization of the workforce, a *replacement ratio* that measures the difference between unemployment benefits and real wages, and a *wedge* between the real product wage and the real consumption wage. It is found that wages and productivity have unit roots. The issue then is to discern whether unionization, the wedge, and the unemployment rate, which might be $I(0)$, have level effects in the model.

Consider the vector autoregression, or VAR representation of the model

$$\begin{pmatrix} y_t \\ z_t \end{pmatrix} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{pmatrix} y_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix},$$

or

$$\mathbf{y}_t = \boldsymbol{\Gamma}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where the vector $\mathbf{y}_t$ is $[y_t, z_t]'$. Now take first differences to obtain

$$\mathbf{y}_t - \mathbf{y}_{t-1} = (\boldsymbol{\Gamma} - \mathbf{I})\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

or

$$\Delta\mathbf{y}_t = \boldsymbol{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t.$$

If all variables are $I(1)$, then all M variables on the left-hand side are $I(0)$. Whether those on the right-hand side are $I(0)$ remains to be seen. The matrix $\boldsymbol{\Pi}$ produces linear combinations of the variables in $\mathbf{y}_t$. But as we have seen, not all linear combinations can be cointegrated. The number of such independent linear combinations is $r < M$. Therefore, although there must be a VAR representation of the model, cointegration implies a restriction on the rank of $\boldsymbol{\Pi}$. It cannot have full rank; its rank is r. From another viewpoint, a different approach to discerning cointegration is suggested. Suppose that we estimate this model as an unrestricted VAR. The resultant coefficient matrix should be short-ranked. The implication is that if we fit the VAR model and impose short rank on the coefficient matrix as a restriction—how we could do that remains to be seen—then if the variables really are cointegrated, this restriction should not lead to a loss of fit. This implication is the basis of Johansen's (1988) and Stock and Watson's (1988) analysis of cointegration.

21.3.3 TESTING FOR COINTEGRATION

A natural first step in the analysis of cointegration is to establish that it is indeed a characteristic of the data. Two broad approaches for testing for cointegration have been developed. The Engle and Granger (1987) method is based on assessing whether single-equation estimates of the equilibrium errors appear to be stationary. The second approach, due to Johansen (1988, 1991) and Stock and Watson (1988), is based on the VAR approach. As noted earlier, if a set of variables is truly cointegrated, then we should be able to detect the implied restrictions in an otherwise unrestricted VAR. We will examine these two methods in turn.

Let $\mathbf{y}_t$ denote the set of M variables that are believed to be cointegrated. Step one of either analysis is to establish that the variables are indeed integrated to the same order.

The Dickey–Fuller tests discussed in Section 21.2.4 can be used for this purpose. If the evidence suggests that the variables are integrated to different orders or not at all, then the specification of the model should be reconsidered.

If the cointegration rank of the system is r, then there are r independent vectors, $\boldsymbol{\gamma}_i = [1, -\boldsymbol{\theta}_i]$, where each vector is distinguished by being normalized on a different variable. If we suppose that there are also a set of $I(0)$ exogenous variables, including a constant, in the model, then each cointegrating vector produces the equilibrium relationship,

$$\mathbf{y}_t'\boldsymbol{\gamma}_i = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_{it},$$

which we may rewrite as

$$y_{it} = \mathbf{Y}_{it}'\boldsymbol{\theta}_i + \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_{it}.$$

We can obtain estimates of $\boldsymbol{\theta}_i$ by least squares regression. If the theory is correct *and* if this OLS estimator is consistent, then residuals from this regression should estimate the equilibrium errors. There are two obstacles to consistency. First, because both sides of the equation contain $I(1)$ variables, the problem of spurious regressions appears. Second, a moment's thought should suggest that what we have done is extract an equation from an otherwise ordinary simultaneous equations model and propose to estimate its parameters by ordinary least squares. As we examined in Chapter 10, consistency is unlikely in that case. It is one of the extraordinary results of this body of theory that in this setting, neither of these considerations is a problem. In fact, as shown by a number of authors,[15] not only is $\mathbf{c}_i$, the OLS estimator of $\boldsymbol{\theta}_i$, consistent, it is **superconsistent** in that its asymptotic variance is $O(1/T^2)$ rather than $O(1/T)$ as in the usual case. Consequently, the problem of spurious regressions disappears as well. Therefore, the next step is to estimate the cointegrating vector(s), by OLS. Under all the assumptions thus far, the residuals from these regressions, e_{it}, are estimates of the equilibrium errors, ε_{it}. As such, they should be $I(0)$. The natural approach would be to apply the familiar Dickey–Fuller tests to these residuals. The logic is sound, but the Dickey–Fuller tables are inappropriate for these estimated errors. Estimates of the appropriate critical values for the tests are given by Engle and Granger (1987), Engle and Yoo (1987), Phillips and Ouliaris (1990), and Davidson and MacKinnon (1993). If autocorrelation in the equilibrium errors is suspected, then an augmented Engle and Granger test can be based on the template

$$\Delta e_{it} = \delta e_{i,t-1} + \phi_1(\Delta e_{i,t-1}) + \cdots + u_t.$$

If the null hypothesis that $\delta = 0$ cannot be rejected (against the alternative $\delta < 0$), then we conclude that the variables are not cointegrated. (Cointegration can be rejected by this method. Failing to reject does not confirm it, of course. But having failed to reject the presence of cointegration, we will proceed as if our finding had been affirmative.)

Example 21.8 Cointegration in Consumption and Output

In the example presented at the beginning of this discussion, we proposed precisely the sort of test suggested by Phillips and Ouliaris (1990) to determine if (log) consumption and (log)

[15] See, for example, Davidson and MacKinnon (1993).

GDP are cointegrated. As noted, the logic of our approach is sound, but a few considerations remain. The Dickey–Fuller critical values suggested for the test are appropriate only in a few cases, and not when several trending variables appear in the equation. For the case of only a pair of trended variables, as we have here, one may use infinite sample values in the Dickey–Fuller tables for the trend stationary form of the equation. (The drift and trend would have been removed from the residuals by the original regression, which would have these terms either embedded in the variables or explicitly in the equation.) Finally, there remains an issue of how many lagged differences to include in the ADF regression. We have specified one, although further analysis might be called for. [A lengthy discussion of this set of issues appears in Hayashi (2000, pp. 645–648).] Thus, but for the possibility of this specification issue, the ADF approach suggested in the introduction does pass muster. The sample value found earlier was −4.63. The critical values from the table are −3.45 for 5% and −3.67 for 2.5%. Thus, we conclude (as have many other analysts) that log consumption and log GDP are cointegrated.

The Johansen (1988, 1992) and Stock and Watson (1988) methods are similar, so we will describe only the first one. The theory is beyond the scope of this text, although the operational details are suggestive. To carry out the Johansen test, we first formulate the VAR,

$$\mathbf{y}_t = \mathbf{\Gamma}_1\mathbf{y}_{t-1} + \mathbf{\Gamma}_2\mathbf{y}_{t-2} + \cdots + \mathbf{\Gamma}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t.$$

The order of the model, p, must be determined in advance. Now, let $\mathbf{z}_t$ denote the vector of $M(p-1)$ variables,

$$\mathbf{z}_t = [\Delta\mathbf{y}_{t-1}, \Delta\mathbf{y}_{t-2}, \ldots, \Delta\mathbf{y}_{t-p+1}].$$

That is, $\mathbf{z}_t$ contains the lags 1 to $p-1$ of the first differences of all M variables. Now, using the T available observations, we obtain two $T \times M$ matrices of least squares residuals,

$$\mathbf{D} = \text{the residuals in the regressions of } \Delta\mathbf{y}_t \text{ on } \mathbf{z}_t,$$

$$\mathbf{E} = \text{the residuals in the regressions of } \mathbf{y}_{t-p} \text{ on } \mathbf{z}_t.$$

We now require the M^2 **canonical correlations** between the columns in $\mathbf{D}$ and those in $\mathbf{E}$. To continue, we will digress briefly to define the canonical correlations. Let $\mathbf{d}_1^*$ denote a linear combination of the columns of $\mathbf{D}$, and let $\mathbf{e}_1^*$ denote the same from $\mathbf{E}$. We wish to choose these two linear combinations so as to maximize the correlation between them. This pair of variables are the first canonical variates, and their correlation r_1^* is the first canonical correlation. In the setting of cointegration, this computation has some intuitive appeal. Now, with $\mathbf{d}_1^*$ and $\mathbf{e}_1^*$ in hand, we seek a second pair of variables $\mathbf{d}_2^*$ and $\mathbf{e}_2^*$ to maximize *their* correlation, subject to the constraint that this second variable in each pair be orthogonal to the first. This procedure continues for all M pairs of variables. It turns out that the computation of all these is quite simple. We will not need to compute the coefficient vectors for the linear combinations. The squared canonical correlations are simply the ordered characteristic roots of the matrix,

$$\mathbf{R}^* = \mathbf{R}_{DD}^{-1/2}\mathbf{R}_{DE}\mathbf{R}_{EE}^{-1}\mathbf{R}_{ED}\mathbf{R}_{DD}^{-1/2},$$

where $\mathbf{R}_{ij}$ is the (cross-) correlation matrix between variables in set i and set j, for $i, j = D, E$.

Finally, the null hypothesis that there are r or fewer cointegrating vectors is tested using the test statistic,

$$\text{TRACE TEST} = -T \sum_{i=r+1}^{M} \ln[1 - (r_i^*)^2].$$

If the correlations based on actual disturbances had been observed instead of estimated, then we would refer this statistic to the chi-squared distribution with $M - r$ degrees of freedom. Alternative sets of appropriate tables are given by Johansen and Juselius (1990) and Osterwald-Lenum (1992). Large values give evidence against the hypothesis of r or fewer cointegrating vectors.

21.3.4 ESTIMATING COINTEGRATION RELATIONSHIPS

Both of the testing procedures discussed earlier involve actually estimating the cointegrating vectors, so this additional section is actually superfluous. In the Engle and Granger framework, at a second step after the cointegration test, we can use the residuals from the static regression as an error correction term in a dynamic, first-difference regression, as shown in Section 21.3.2. One can then *test down* to find a satisfactory structure. In the Johansen test shown earlier, the characteristic vectors corresponding to the canonical correlations are the sample estimates of the cointegrating vectors. Once again, computation of an error correction model based on these first-step results is a natural next step. We will explore these in an application.

21.3.5 APPLICATION: GERMAN MONEY DEMAND

The demand for money has provided a convenient and well-targeted illustration of methods of cointegration analysis. The central equation of the model is

$$m_t - p_t = \mu + \beta y_t + \gamma i_t + \varepsilon_t, \tag{21-8}$$

where m_t, p_t, and y_t are the logs of nominal money demand, the price level, and output, and i is the nominal interest rate (not the log of). The equation involves trending variables (m_t, p_t, y_t), and one that we found earlier appears to be a random walk with drift (i_t). As such, the usual form of statistical inference for estimation of the income elasticity and interest semielasticity based on stationary data is likely to be misleading.

Beyer (1998) analyzed the demand for money in Germany over the period 1975 to 1994. A central focus of the study was whether the 1990 reunification produced a structural break in the long-run demand function. (The analysis extended an earlier study by the same author that was based on data that predated the reunification.) One of the interesting questions pursued in this literature concerns the stability of the long-term demand equation,

$$(m - p)_t - y_t = \mu + \gamma i_t + \varepsilon_t. \tag{21-9}$$

The left-hand side is the log of the inverse of the velocity of money, as suggested by Lucas (1988). An issue to be confronted in this specification is the exogeneity of the interest variable—exogeneity [in the Engle, Hendry, and Richard (1993) sense] of income is moot in the long-run equation as its coefficient is assumed (per Lucas) to equal one. Beyer explored this latter issue in the framework developed by Engle et al. (see Section 21.3.5).

TABLE 21.5 Augmented Dickey–Fuller Tests for Variables in the Beyer Model

Variable	m	Δm	$\Delta^2 m$	p	Δp	$\Delta^2 p$	$\Delta_4 p$	$\Delta\Delta_4 p$
Spec.	TS	RW	RW	TS	RW/D	RW	RW/D	RW
Lag	0	4	3	4	3	2	2	2
DF_τ	−1.82	−1.61	−6.87	−2.09	−2.14	−10.6	−2.66	−5.48
Crit.Value	−3.47	−1.95	−1.95	−3.47	−2.90	−1.95	−2.90	−1.95
Variable	y	Δy	RS	ΔRS	RL	ΔRL	$(m - p)$	$\Delta(m - p)$
Spec.	TS	RW/D	TS	RW	TS	RW	RW/D	RW/D
Lag	4	3	1	0	1	0	0	0
DF_τ	−1.83	−2.91	−2.33	−5.26	−2.40	−6.01	−1.65	−8.50
Crit.Value	−3.47	−2.90	−2.90	−1.95	−2.90	−1.95	−3.47	−2.90

The analytical platform of Beyer's study is a long-run function for the real money stock M3 (we adopt the author's notation)

$$(m - p)^* = \delta_0 + \delta_1 y + \delta_2 RS + \delta_3 RL + \delta_4 \Delta_4 p, \qquad \textbf{(21-10)}$$

where RS is a short-term interest rate, RL is a long-term interest rate, and $\Delta_4 p$ is the annual inflation rate—the data are quarterly. The first step is an examination of the data. Augmented Dickey–Fuller tests suggest that for these German data in this period, m_t and p_t are $I(2)$, while $(m_t - p_t)$, y_t, $\Delta_4 p_t$, RS_t, and RL_t are all $I(1)$. Some of Beyer's results which produced these conclusions are shown in Table 21.5. Note that although both m_t and p_t appear to be $I(2)$, their simple difference (linear combination) is $I(1)$, that is, integrated to a lower order. That produces the long-run specification given by (21-10). The Lucas specification is layered onto this to produce the model for the long-run velocity,

$$(m - p - y)^* = \delta_0^* + \delta_2^* RS + \delta_3^* RL + \delta_4^* \Delta_4 p. \qquad \textbf{(21-11)}$$

21.3.5.a Cointegration Analysis and a Long-Run Theoretical Model

For (21-10) to be a valid model, there must be at least one cointegrating vector that transforms $\mathbf{z}_t = [(m_t - p_t), y_t, RS_t, RL_t, \Delta_4 p_t]$ to stationarity. The Johansen trace test described in Section 21.3.3 was applied to the VAR consisting of these five $I(1)$ variables. A lag length of two was chosen for the analysis. The results of the trace test are a bit ambiguous; the hypothesis that $r = 0$ is rejected, albeit not strongly (sample value = 90.17 against a 95% critical value = 87.31) while the hypothesis that $r \leq 1$ is not rejected (sample value = 60.15 against a 95% critical value of 62.99). (These borderline results follow from the result that Beyer's first three eigenvalues—canonical correlations in the trace test statistic—are nearly equal. Variation in the test statistic results from variation in the correlations.) On this basis, it is concluded that the cointegrating rank equals one. The unrestricted cointegrating vector for the equation with a time trend added, is found to be

$$(m - p) = 0.936y - 1.780\Delta_4 p + 1.601RS - 3.279RL + 0.002t. \qquad \textbf{(21-12)}$$

(These are the coefficients from the first characteristic vector of the canonical correlation analysis in the Johansen computations detailed in Section 21.3.3.) An exogeneity test—we have not developed this in detail; see Beyer (1998, p. 59), Hendry and Ericsson (1991), and Engle and Hendry (1993)—confirms weak exogeneity of all four right-hand-side variables in this specification. The final specification test is for the Lucas formulation and elimination of the time trend, both of which are found to pass, producing the cointegration vector,

$$(m - p - y) = -1.832\Delta_4 p + 4.352RS - 10.89RL.$$

The conclusion drawn from the cointegration analysis is that a single-equation model for the long-run money demand is appropriate and a valid way to proceed. A last step before this analysis is a series of Granger causality tests for feedback between changes in the money stock and the four right-hand-side variables in (21-12) (not including the trend). The test results are generally favorable, with some mixed results for exogeneity of GDP.

21.3.5.b Testing for Model Instability

Let $z_t = [(m_t - p_t), y_t, \Delta_4 p_t, RS_t, RL_t]$ and let $\mathbf{z}_{t-1}^0$ denote the entire history of $\mathbf{z}_t$ up to the previous period. The joint distribution for $\mathbf{z}_t$, conditioned on $\mathbf{z}_{t-1}^0$ and a set of parameters, $\boldsymbol{\Psi}$, factors one level further into

$$f(\mathbf{z}_t | \mathbf{z}_{t-1}^0, \Psi) = f[(m - p)_t | y_t, \Delta_4 p_t, RS_t, RL_t, \mathbf{z}_{t-1}^0, \boldsymbol{\Psi}_1] \times g(y_t, \Delta_4 p_t, RS_t, RL_t | \mathbf{z}_{t-1}^0, \boldsymbol{\Psi}_2).$$

The result of the exogeneity tests carried out earlier implies that the conditional distribution may be analyzed apart from the marginal distribution—that is, the implication of the Engle, Hendry, and Richard results noted earlier. Note the partitioning of the parameter vector. Thus, the conditional model is represented by an error correction form that explains $\Delta(m - p)_t$ in terms of its own lags, the error correction term, and contemporaneous and lagged changes in the (now established) weakly exogenous variables as well as other terms such as a constant term, trend, and certain dummy variables which pick up particular events. The error correction model specified is

$$\begin{aligned}\Delta(m - p)_t = &\sum_{i=1}^{4} c_i \Delta(m - p)_{t-i} + \sum_{i=0}^{4} d_{1,i} \Delta(\Delta_{4 p_{t-i}}) + \sum_{i=0}^{4} d_{2,i} \Delta y_{t-i} \\ &+ \sum_{i=0}^{4} d_{3,i} \Delta RS_{t-i} + \sum_{i=0}^{4} d_{4,i} \Delta RL_{t-i} + \lambda(m - p - y)_{t-1} \\ &+ \gamma_1 RS_{t-1} + \gamma_2 RL_{t-1} + \mathbf{d}_t' \boldsymbol{\phi} + \omega_t, \end{aligned} \quad \textbf{(21-13)}$$

where $\mathbf{d}_t$ is the set of additional variables, including the constant and five one-period dummy variables that single out specific events such as a currency crisis in September, 1992.[16] The model is estimated by least squares, "stepwise simplified and reparameterized." (The number of parameters in the equation is reduced from 32 to 15.[17])

[16] Beyer (1998, p. 62, footnote 4).

[17] The equation ultimately used is $\Delta(m_t - p_t) = h[\Delta(m - p)_{t-4}, \Delta\Delta_4 p_t, \Delta^2 y_{t-2}, \Delta RS_{t-1} + \Delta RS_{t-3}, \Delta^2 RL_t, RS_{t-1}, RL_{t-1}, \Delta_4 p_{t-1}, (m - p - y)_{t-1}, \mathbf{d})_t]$.

The estimated form of (21-13) is an autoregressive distributed lag model. We proceed to use the model to solve for the long-run, steady-state growth path of the real money stock, (21-10). The annual growth rates $\Delta_4 m = g_m$, $\Delta_4 p = g_p$, $\Delta_4 y = g_y$ and (assumed) $\Delta_4 RS = g_{RS} = \Delta_4 RL = g_{RL} = 0$ are used for the solution[18]

$$\frac{1}{4}(g_m - g_p) = \frac{c_4}{4}(g_m - g_p) - d_{1,1}g_p + \frac{d_{2,2}}{2}g_y + \gamma_1 RS + \gamma_2 RL + \lambda(m - p - y).$$

This equation is solved for $(m - p)^*$ under the assumption that $g_m = (g_y + g_p)$,

$$(m - p)^* = \hat{\delta}_0 + \hat{\delta}_1 g_y + y + \hat{\delta}_2 \Delta_4 p + \hat{\delta}_3 RS + \hat{\delta}_4 RL.$$

Analysis then proceeds based on this estimated long-run relationship.

The primary interest of the study is the stability of the demand equation pre- and postunification. A comparison of the parameter estimates from the same set of procedures using the period 1976 to 1989 shows them to be surprisingly similar, $[(1.22 - 3.67g_y), 1, -3.67, 3.67, -6.44]$ for the earlier period and $[(1.25 - 2.09g_y), 1, -3.625, 3.5, -7.25]$ for the later one. This suggests, albeit informally, that the function has not changed (at least by much). A variety of testing procedures for structural break led to the conclusion that in spite of the dramatic changes of 1990, the long-run money demand function had not materially changed in the sample period.

21.4 NONSTATIONARY PANEL DATA

In Section 11.10, we began to examine panel data settings in which T, the number of observations in each group (e.g., country), became large as well as n. Applications include cross-country studies of growth using the Penn World Tables,[19] studies of purchasing power parity,[20] and analyses of health care expenditures.[21] In the small T cases of longitudinal, microeconomic data sets, the time-series properties of the data are a side issue that is usually of little interest. But when T is growing at essentially the same rate as n, for example, in the cross-country studies, these properties become a central focus of the analysis.

The large T, large n case presents several complications for the analyst. In the longitudinal analysis, pooling of the data is usually a given, although we developed several extensions of the models to accommodate parameter heterogeneity (see Section 11.10). In a long-term cross-country model, any type of pooling would be especially suspect. The time series are long, so this would seem to suggest that the appropriate modeling strategy would be simply to analyze each country separately. But this would neglect the hypothesized commonalities across countries such as a (proposed) common growth rate. Thus, the time-series panel data literature seeks to reconcile these opposing features of the data.

As in the single time-series cases examined earlier in this chapter, long-term aggregate series are usually nonstationary, which calls conventional methods (such as

[18]The division of the coefficients is done because the intervening lags do not appear in the estimated equation.

[19]Im, Pesaran, and Shin (2003) and Sala-i-Martin (1996).

[20]Pedroni (2001).

[21]McCoskey and Selden (1998).

those in Section 11.10) into question. A focus of the recent literature, for example, is on testing for unit roots in an analog to the platform for the augmented Dickey–Fuller tests (Section 21.2),

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{m=1}^{L_i} \gamma_{im} \Delta y_{i,t-m} + \alpha_i + \beta_i t + \varepsilon_{it}.$$

Different formulations of this model have been analyzed, for example, by Levin, Lin, and Chu (2002), who assume $\rho_i = \rho$; Im, Pesaran, and Shin (2003), who relax that restriction; and Breitung (2000), who considers various mixtures of the cases. An extension of the KPSS test in Section 21.2.5 that is particularly simple to compute is Hadri's (2000) LM statistic,

$$\text{LM} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\sum_{t=1}^{T} E_{it}^2}{T^2 \hat{\sigma}_\varepsilon^2} \right) = \frac{\sum_{i=1}^{n} KPSS_i}{n}.$$

This is the sample average of the KPSS statistics for the n countries. Note that it includes two assumptions: that the countries are independent and that there is a common σ_ε^2 for all countries. An alternative is suggested that allows σ_ε^2 to vary across countries.

As it stands, the preceding model would suggest that separate analyses for each country would be appropriate. An issue to consider, then, would be how to combine, if possible, the separate results in some optimal fashion. Maddala and Wu (1999), for example, suggested a "Fisher-type" chi-squared test based on $P = -2\Sigma_i \ln p_i$, where p_i is the p value from the individual tests. Under the null hypothesis that ρ_i equals zero, the limiting distribution is chi squared with $2n$ degrees of freedom.

Analysis of cointegration, and models of cointegrated series in the panel data setting, parallel the single time-series case, but also differ in a crucial respect.[22] Whereas in the single time-series case, the analysis of cointegration focuses on the long-run relationships between, say, x_t and z_t for two variables for the same country, in the panel data setting, say, in the analysis of exchange rates, inflation, purchasing power parity or international R & D spillovers, interest may focus on a long-run relationship between x_{it} and x_{mt} for two different countries (or n countries). This substantially complicates the analyses. It is also well beyond the scope of this text. Extensive surveys of these issues may be found in Baltagi (2005, Chapter 12) and Smith (2000).

21.5 SUMMARY AND CONCLUSIONS

This chapter has completed our survey of techniques for the analysis of time-series data. Most of the results in this chapter focus on the internal structure of the individual time series themselves. While the empirical distinction between, say, AR(p) and MA(q) series may seem ad hoc, the Wold decomposition theorem assures that with enough care, a variety of models can be used to analyze a time series. This chapter described what is arguably the fundamental tool of modern macroeconometrics: the tests for nonstationarity. Contemporary econometric analysis of macroeconomic data has added considerable structure and formality to trending variables, which are more

[22] See, for example, Kao (1999), McCoskey and Kao (1999), and Pedroni (2000, 2004).

common than not in that setting. The variants of the Dickey–Fuller and KPSS tests for unit roots are indispensable tools for the analyst of time-series data. Section 21.4 then considered the subject of cointegration. This modeling framework is a distinct extension of the regression modeling where this discussion began. Cointegrated relationships and equilibrium relationships form the basis of the time-series counterpart to regression relationships. But, in this case, it is not the conditional mean as such that is of interest. Here, both the long-run equilibrium and short-run relationships around trends are of interest and are studied in the data.

Key Terms and Concepts

- Augmented Dickey–Fuller test
- Autoregressive integrated moving-average (ARIMA) process
- Bounds test
- Canonical correlation
- Cointegrated
- Cointegration
- Cointegration rank
- Cointegrating vector
- Common trend
- Data-generating process (DGP)
- Dickey–Fuller test
- Equilibrium error
- Integrated of order one
- Nonstationary process
- Phillips–Perron test
- Random walk
- Random walk with drift
- Superconsistent
- Trend stationary process
- Unit root

Exercise

1. Find the first two autocorrelations and partial autocorrelations for the MA(2) process

$$\varepsilon_t = v_t - \theta_1 v_{t-1} - \theta_2 v_{t-2}.$$

Applications

1. Using the macroeconomic data in Appendix Table F5.2, estimate by least squares the parameters of the model $c_t = \beta_0 + \beta_1 y_t + \beta_2 c_{t-1} + \beta_3 c_{t-2} + \varepsilon_t$, where c_t is the log of real consumption and y_t is the log of real disposable income.
   a. Use the Breusch and Pagan LM test to examine the residuals for autocorrelation.
   b. Is the estimated equation stable? What is the characteristic equation for the autoregressive part of this model? What are the roots of the characteristic equation, using your estimated parameters?
   c. What is your implied estimate of the short-run (impact) multiplier for change in y_t on c_t? Compute the estimated long-run multiplier.
2. Carry out an ADF test for a unit root in the rate of inflation using the subset of the data in Appendix Table F5.2 since 1974.1. (This is the first quarter after the oil shock of 1973.)
3. Estimate the parameters of the model in Example 10.4 using two-stage least squares. Obtain the residuals from the two equations. Do these residuals appear to be white noise series? Based on your findings, what do you conclude about the specification of the model?

REFERENCES

Abowd, J., and H. Farber. "Job Queues and Union Status of Workers." *Industrial and Labor Relations Review,* 35, 1982, pp. 354–367.

Abramovitz, M., and I. Stegun. *Handbook of Mathematical Functions*. New York: Dover Press, 1971.

Abrevaya, J. "The Equivalence of Two Estimators of the Fixed Effects Logit Model." *Economics Letters,* 55, 1997, pp. 41–43.

Achen, C. "Two-Step Hierarchical Estimation: Beyond Regression Analysis." *Political Analysis,* 13, 4, 2005 pp. 447–456.

Afifi, T., and R. Elashoff. "Missing Observations in Multivariate Statistics." *Journal of the American Statistical Association,* 61, 1966, pp. 595–604.

Afifi, T., and R. Elashoff. "Missing Observations in Multivariate Statistics." *Journal of the American Statistical Association,* 62, 1967, pp. 10–29.

Agresti, A. *Categorical Data Analysis*. 2nd ed., John Wiley and Sons, New York, 2002.

Aguirregabiria, V., and P. Mira. "Dynamic Discrete Choice Structural Models: A Survey." *Journal of Econometrics*, 156, 1, 2010, pp. 38–67.

Ahn, S., and P. Schmidt. "Efficient Estimation of Models for Dynamic Panel Data." *Journal of Econometrics,* 68, 1, 1995, pp. 5–28.

Ai, C., and E. Norton. "Interaction Terms in Logit and Probit Models." *Economics Letters*, 80, 2003, pp. 123–129.

Aigner, D. "MSE Dominance of Least Squares with Errors of Observation." *Journal of Econometrics,* 2, 1974, pp. 365–372.

Aigner, D., K. Lovell, and P. Schmidt. "Formulation and Estimation of Stochastic Frontier Production Models." *Journal of Econometrics,* 6, 1977, pp. 21–37.

Aitcheson, J., and S. Silvey. "The Generalization of Probit Analysis to the Case of Multiple Responses." *Biometrika*, 44, 1957, pp. 131–140.

Aitchison, J., and J. Brown. *The Lognormal Distribution with Special Reference to Its Uses in Economics*. New York: Cambridge University Press, 1969.

Aitken, A. C. "On Least Squares and Linear Combinations of Observations." *Proceedings of the Royal Statistical Society,* 55, 1935, pp. 42–48.

Akin, J., D. Guilkey, and R. Sickles. "A Random Coefficient Probit Model with an Application to a Study of Migration." *Journal of Econometrics,* 11, 1979, pp. 233–246.

Albert, J., and S. Chib. "Bayesian Analysis of Binary and Polytomous Response Data." *Journal of the American Statistical Association,* 88, 1993a, pp. 669–679.

Aldrich, J., and F. Nelson. *Linear Probability, Logit, and Probit Models*. Beverly Hills: Sage Publications, 1984.

Alemu, H., M. Morkbak, S. Olsen, and C. L. Jensen. "Attending the Reasons for Attribute Non-attendance in Choice Experiments." *Environmental and Resource Economics*, 54, 3, 2013, pp. 333–359.

Allison, P. "Problems with Fixed-Effects Negative Binomial Models." Manuscript, Department of Sociology, University of Pennsylvania, 2000.

Allison, P., and R. Waterman. "Fixed-Effects Negative Binomial Regression Models." *Sociological Methodology,* 32, 2002, pp. 247–256.

Allison, P. *Missing Data*. Beverly Hills: Sage Publications, 2002.

Allison, P. "What's the Best R-Squared for Logistic Regression." accessed July 6, 2016, http://statisticalhorizons.com/r2logistic, 2/13/2013.

Amemiya, T. "The Estimation of Variances in a Variance-Components Model." *International Economic Review,* 12, 1971, pp. 1–13.

Amemiya, T. "Some Theorems in the Linear Probability Model." *International Economic Review,* 18, 1977, pp. 645–650.

Amemiya, T. "Qualitative Response Models: A Survey." *Journal of Economic Literature,* 19, 4, 1981, pp. 481–536.

Amemiya, T. "Tobit Models: A Survey." *Journal of Econometrics,* 24, 1984, pp. 3–63.

Amemiya, T. *Advanced Econometrics*. Cambridge: Harvard University Press, 1985.

Amemiya, T., and T. MaCurdy. "Instrumental Variable Estimation of an Error Components Model." *Econometrica,* 54, 1986, pp. 869–881.

Andersen, D. "Asymptotic Properties of Conditional Maximum Likelihood Estimators." *Journal of the Royal Statistical Society,* Series B, 32, 1970, pp. 283–301.

Anderson, T. *The Statistical Analysis of Time Series*. New York: John Wiley and Sons, 1971.

Anderson, R., and J. Thursby. "Confidence Intervals for Elasticity Estimators in Translog Models." *Review of Economics and Statistics,* 68, 1986, pp. 647–657.

Anderson, G. and R. Blundell. "Estimation and Hypothesis Testing in Dynamic Singular Equation Systems." *Econometrica*, 50, 6, 1982, pp. 1559–1571.

Anderson, T., and C. Hsiao. "Estimation of Dynamic Models with Error Components." *Journal of the American Statistical Association,* 76, 1981, pp. 598–606.

Anderson, T., and C. Hsiao. "Formulation and Estimation of Dynamic Models Using Panel Data." *Journal of Econometrics,* 18, 1982, pp. 67–82.

Anderson, T., and H. Rubin. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics,* 20, 1949, pp. 46–63.

Anderson, T., and H. Rubin. "The Asymptotic Properties of Estimators of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics,* 21, 1950, pp. 570–582.

Anderson, K., R. Burkhauser, J. Raymond, and C. Russell. "Mixed Signals in the Job Training Partnership Act." *Growth and Change*, 22, 3, 1991, pp. 32–48.

Andrews, D. "A Robust Method for Multiple Linear Regression." *Technometrics,* 16, 1974, pp. 523–531.

Andrews, D. "Hypothesis Tests with a Restricted Parameter Space." *Journal of Econometrics,* 84, 1998, pp. 155–199.

Andrews, D. "Estimation When a Parameter Is on a Boundary." *Econometrica,* 67, 1999, pp. 1341–1382.

Andrews, D. "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space." *Econometrica,* 68, 2000, pp. 399–405.

Andrews, D. "Testing When a Parameter Is on a Boundary of the Maintained Hypothesis." *Econometrica,* 69, 2001, pp. 683–734.

Andrews, D. "GMM Estimation When a Parameter Is on a Boundary." *Journal of Business and Economic Statistics,* 20, 2002, pp. 530–544.

Andrews, D. and M. Buchinsky. "A Three Step Method for Choosing the Number of Bootstrap Replication." *Econometrica,* 68, 2000, pp. 23–51.

Andrews, D., and R. Fair. "Inference in Nonlinear Econometric Models with Structural Change." *Review of Economic Studies,* 55, 1988, pp. 615–640.

Andrews, D., and W. Ploberger. "Optimal Tests When a Nuisance Parameter Is Present Only Under the Alternative." *Econometrica,* 62, 1994, pp. 1383–1414.

Andrews, D., and W. Ploberger. "Admissability of the LR Test When a Nuisance Parameter Is Present Only Under the Alternative." *Annals of Statistics,* 23, 1995, pp. 1609–1629.

Angelini, V., D. Cavapozzi, L. Corazzini, and O. Paccagnella. "Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Specific Scale Biases?" manuscript, University of Padua, 2008.

Angrist, J. "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice." *Journal of Business and Economic Statistics,* 29, 1, 2001, pp. 2–15.

Angrist, J., G. Imbens, and D. Rubin. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91, 1996, pp. 444–455.

Angrist, J., and A. Krueger. "Does Compulsory School Attendance Affect Schooling and Earnings." *Quarterly Journal of Economics*, 106, 4, 1991, pp. 979–1014.

Angrist, J., and A. Krueger. "The Effect of Age at School Entry on Educational Attainment:

An Application of Instrumental Variables with Moments Form Two Samples." *Journal of the American Statistical Association*, 87, 1992, pp 328–336.

Angrist, J., and A. Krueger. "Why Do World War II Veterans Earn More Than Nonveterans?" *Journal of Labor Economics*, 12, 1994, pp. 74–97.

Angrist, J., and A. Krueger. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives*, 15, 4, 2001, pp. 69–85.

Angrist, J., and V. Lavy. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics*. 144, 1999, pp. 533–576.

Angrist, J., and V. Lavy. "The Effect of High School Matriculation Awards; Evidence from Randomized Trials." Working paper, Department of Economics, MIT, NJ, 2002.

Angrist, J., and J. Pischke. *Mostly Harmless Econometrics*. Princeton: Princeton University Press, 2009.

Angrist, J., and J. Pischke. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives*, 24, 2, 2010, pp. 3–30.

Aneuryn-Evans, G., and A. Deaton. "Testing Linear versus Logarithmic Regression Models." *Review of Economic Studies,* 47, 1980, pp. 275–291.

Anselin, L. *Spatial Econometrics: Methods and Models,* Dordrecht: Kluwer Academic Publishers, 1988.

Anselin, L. "Spatial Econometrics." In *A Companion to Theoretical Econometrics,* edited by B. Baltagi, Oxford: Blackwell Puplishers, 2001, pp. 310–330.

Anselin, L., and S. Hudak. "Spatial Econometrics in Practice: A Review of Software Options." *Regional Science and Urban Economics,* 22, 3, 1992, pp. 509–536.

Antweiler, W. "Nested Random Effects Estimation in Unbalanced Panel Data." *Journal of Econometrics,* 101, 2001, pp. 295–312.

Arabmazar, A., and P. Schmidt. "An Investigation into the Robustness of the Tobit Estimator to Nonnormality." *Econometrica,* 50, 1982a, pp. 1055–1063.

Arabmazar, A., and P. Schmidt. "Further Evidence on the Robustness of the Tobit Estimator to Heteroscedasticity." *Journal of Econometrics,* 17, 1982b, pp. 253–258.

Arellano, M. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics,* 49, 1987, pp. 431–434.

Arellano, M. "A Note on the Anderson-Hsiao Estimator for Panel Data." *Economics Letters,* 31, 1989, pp. 337–341.

Arellano, M. "Discrete Choices with Panel Data." *Investigaciones Economica,* Lecture 25, 2000.

Arellano, M. "Panel Data: Some Recent Developments." In *Handbook of Econometrics,* Vol. 5, edited by J. Heckman and E. Leamer, North Holland, Amsterdam, 2001.

Arellano, M. *Panel Data Econometrics*. Oxford: Oxford University Press, 2003.

Arellano, M., and S. Bond. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economics Studies,* 58, 1991, pp. 277–297.

Arellano, M., and C. Borrego. "Symmetrically Normalized Instrumental Variable Estimation Using Panel Data." *Journal of Business and Economic Statistics,* 17, 1999, pp. 36–49.

Arellano, M., and O. Bover. "Another Look at the Instrumental Variables Estimation of Error Components Models." *Journal of Econometrics,* 68, 1, 1995, pp. 29–52.

Arellano, M., and J. Hahn. "A Likelihood Based Approximate Solution to the Incidental Parameters Problem in Dynamic Nonlinear Models with Multiple Effects." unpublished manuscript, CEMFI, 2006.

Arellano, M., and J. Hahn. "Understanding Bias in Nonlinear Panel Models: Some Recent Developments." In *Advances in Economics and Econometrics; Theory and Applications, Ninth World Congress*, Vol. 3, edited by R. Blundell, W. Newey, and T. Persson, Cambridge: Cambridge University Press, 2007.

Arrow, K., H. Chenery, B. Minhas, and R. Solow. "Capital-Labor Substitution and Economic Efficiency." *Review of Economics and Statistics,* 45, 1961, pp. 225–247.

Arulampalam, W., and M. Stewart. "Simplified Implementation of the Heckman Estimator

of the Dynamic Probit Model and a Comparison with Alternative Estimators." *Oxford Bulletin of Economics and Statistics*, 71, 5, 2009, pp. 659–681.

Asche, F., and R. Tveteras. "Modeling Production Risk with a Two-Step Procedure." *Journal of Agricultural and Resource Economics*, 24, 2, 1999, pp. 424–439.

Ashenfelter, O., and D. Card. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics*, 67, 4, 1985, pp. 648–660.

Ashenfelter, O., and J. Heckman. "The Estimation of Income and Substitution Effects in a Model of Family Labor Supply." *Econometrica,* 42, 1974, pp. 73–85.

Ashenfelter, O., and A. Kreuger. "Estimates of the Economic Return to Schooling from a New Sample of Twins." *American Economic Review,* 84, 1994, pp. 1157–1173.

Ashenfelter, O., and C. Rouse. "Income, Schooling and Ability: Evidence from a New Sample of Identical Twins." *Quarterly Journal of Economics*, 113, 1, 1998, pp. 253–284.

Ashenfelter, O., and D. Zimmerman. "Estimates of the Returns to Schooling from Sibling Data: Fathers, Sons, and Brothers." *The Review of Economics and Statistics*, 79, 1, 1997, pp. 1–9.

Avery, R., L. Hansen, and J. Hotz. "Multiperiod Probit Models and Orthogonality Condition Estimation." *International Economic Review,* 24, 1983, pp. 21–35.

Bago d'Uva, T., and A. Jones. "Health Care Utilization in Europe: New Evidence from the ECHP." *Journal of Health Economics*, 28, 2, 2009, pp. 265–279.

Bago d'Uva, T., E. van Doorslaer, M. Lindeboom, and O. O'Donnell. "Does Reporting Heterogeneity Bias the Measurement of Health Disparities?" *Health Economics*, 17, 2008, pp. 351–375.

Balestra, P., and M. Nerlove. "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas." *Econometrica,* 34, 1966, pp. 585–612.

Baltagi, B. "Pooling Under Misspecification: Some Monte Carlo Evidence on the Kmenta and Error Components Techniques." *Econometric Theory,* 2, 1986, pp. 429–441.

Baltagi, B. "Applications of a Necessary and Sufficient Condition for OLS to be BLUE." *Statistics and Probability Letters,* 8, 1989, pp. 457–461.

Baltagi, B. *Econometric Analysis of Panel Data.* 2nd ed., New York: John Wiley and Sons, 2001.

Baltagi, B. *Econometric Analysis of Panel Data.* 3rd ed., New York: John Wiley and Sons, 2005.

Baltagi, B. *Econometric Analysis of Panel Data.* 5th ed., New York: John Wiley and Sons, 2013.

Baltagi, B., *Oxford Handbook of Panel Data*, Badi H. Baltagi, editor, Oxford University Press, Oxford, 2015.

Baltagi, B., and Griffin, J. "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures." *European Economic Review,* 22, 1983, pp. 117–137.

Baltagi, B., and C. Kao. "Nonstationary Panels, Cointegration in Panels and Dynamic Panels: A Survey." *Advances in Econometrics,* 15, 2000, pp. 7–51.

Baltagi, B., and D. Levin. "Estimating Dynamic Demand for Cigarettes Using Panel Data: The Effects of Bootlegging, Taxation and Advertising Reconsidered." *Review of Economics and Statistics*, 68, 1, 1986, pp. 148–155.

Baltagi. B., and Q. Li. "Double Length Artificial Regressions for Testing Spatial Dependence." *Econometric Reviews,* 20, 2001, pp. 31–40.

Baltagi, B., S. Song, and B. Jung. "The Unbalanced Nested Error Component Regression Model." *Journal of Econometrics,* 101, 2001, pp. 357–381.

Bannerjee, A. "Panel Data Unit Roots and Cointegration: An Overview." *Oxford Bulletin of Economics and Statistics,* 61, 1999, pp. 607–629.

Barnow, B., G. Cain, and A. Goldberger. "Issues in the Analysis of Selectivity Bias." In *Evaluation Studies Review Annual,* Vol. 5, edited by E. Stromsdorfer and G. Farkas, Beverly Hills: Sage Publications, 1981.

Bartels, R., and D. Fiebig. "A Simple Characterization of Seemingly Unrelated Regressions Models in Which OLS is BLUE." *American Statistician,* 45, 1992, pp. 137–140.

Battese, G., and T. Coelli. "Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India." *Journal of Productivity Analysis*, 3, 1/2, 1992, pp. 153–169.

Battese, G., and T. Coelli. "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production for Panel Data." *Empirical Economics*, 20, 1995, pp. 325–332.

Bazaraa, M., and C. Shetty. *Nonlinear Programming: Theory and Algorithms.* New York: John Wiley and Sons, 1979.

Beach, C., and J. MacKinnon. "A Maximum Likelihood Procedure for Regression with Autocorrelated Errors." *Econometrica,* 46, 1978a, pp. 51–58.

Beach, C., and J. MacKinnon. "Full Maximum Likelihood Estimation of Second Order Autoregressive Error Models." *Journal of Econometrics,* 7, 1978b, pp. 187–198.

Beck, N., D. Epstein, and S. Jackman. "Estimating Dynamic Time Series Cross Section Models with a Binary Dependent Variable." Manuscript, Department of Political Science, University of California, San Diego, 2001.

Beck, N., D. Epstein, S. Jackman, and S. O' Halloran. "Alternative Models of Dynamics in Binary Time-Series Cross-Section Models: The Example of State Failure." Manuscript, Department of Political Science, University of California, San Diego, 2001.

Becker, G., and K. Murphy. "A Theory of Rational Addiction." *Journal of Political Economy*, 96, 4, 1988, pp. 675–700.

Becker, S., and A. Ichino. "Estimation of Average Treatment Effects Based on Propensity Scores." *The Stata Journal,* 2, 2002, pp. 358–377.

Becker, W., and P. Kennedy. "A Graphical Exposition of the Ordered Probit Model." *Econometric Theory*, 8, 1992, pp. 127–131.

Behrman, J., and M. Rosenzweig. "'Ability' Biases in Schooling Returns and Twins: A Test and New Estimates." *Economics of Education Review*, 18, 2, 1999, pp. 159–67.

Behrman, J., and P. Taubman. "Is Schooling 'Mostly in the Genes'? Nature-Nurture Decomposition Using Data on Relatives." *Journal of Political Economy,* 97, 6, 1989, pp. 1425–1446

Bekker, P., and T. Wansbeek. "Identification in Parametric Models." In *A Companion to Theoretical Econometrics,* edited by B. Baltagi, Oxford: Blackwell, 2001.

Bell, K., and N. Bockstael. "Applying the Generalized Method of Moments Approach to Spatial Problems Involving Micro-Level Data." *Review of Economic and Statistics* 82, 1, 2000, pp. 72–82

Belsley, D., E. Kuh, and R. Welsh. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons, 1980.

Ben-Akiva, M., and S. Lerman. *Discrete Choice Analysis*. London: MIT Press, 1985.

Bera, A., and C. Jarque. "Efficient Tests for Normality, Heteroscedasticity, and Serial Independence of Regression Residuals: Monte Carlo Evidence." *Economics Letters,* 7, 1981, pp. 313–318.

Bera, A., and C. Jarque. "Model Specification Tests: A Simultaneous Approach." *Journal of Econometrics,* 20, 1982, pp. 59–82.

Bera, A., C. Jarque, and L. Lee. "Testing for the Normality Assumption in Limited Dependent Variable Models." Mimeo, Department of Economics, University of Minnesota, 1982.

Berndt, E. *The Practice of Econometrics*. Reading, MA: Addison-Wesley, 1990.

Berndt, E., and L. Christensen. "The Translog Function and the Substitution of Equipment, Structures, and Labor in U.S. Manufacturing, 1929–1968." *Journal of Econometrics,* 1, 1973, pp. 81–114.

Berndt, E., B. Hall, R. Hall, and J. Hausman. "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurement,* 3/4, 1974, pp. 653–665.

Berndt, E., and E. Savin. "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model." *Econometrica,* 45, 1977, pp. 1263– 1277.

Berndt, E., and D. Wood. "Technology, Prices, and the Derived Demand for Energy." *Review of Economics and Statistics,* 57, 1975, pp. 376–384.

Beron, K., J. Murdoch, and M. Thayer. "Hierarchical Linear Models with Application to Air Pollution in the South Coast Air Basin." *American Journal of Agricultural Economics,* 81, 5, 1999, pp. 1123–1127.

Berry, S., J. Levinsohn, and A. Pakes. "Automobile Prices in Market Equilibrium." *Econometrica,* 63, 4, 1995, pp. 841–890.

Bertrand, M., E. Dufflo, and S. Mullainathan. "How Much Should We Trust Difference

in Differences Estimates?" Working paper, Department of Economics, MIT, 2002.

Bertschek, I. "Product and Process Innovation as a Response to Increasing Imports and Foreign Direct Investment, *Journal of Industrial Economics*. 43, 4, 1995, pp. 341–357.

Bertschek, I., and M. Lechner. "Convenient Estimators for the Panel Probit Model." *Journal of Econometrics,* 87, 2, 1998, pp. 329–372.

Berzeg, K. "The Error Components Model: Conditions for the Existence of Maximum Likelihood Estimates." *Journal of Econometrics,* 10, 1979, pp. 99–102.

Bester, C., and A. Hansen. "A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects." *Journal of Business and Economic Statistics*, 27, 2, 2009, pp. 235–250.

Beyer, A. "Modelling Money Demand in Germany." *Journal of Applied Econometrics,* 13, 1, 1998, pp. 57–76.

Bhargava, A., and J. Sargan. "Testing Residuals from Least Squares Regression for Being Generated by the Gaussian Random Walk." *Econometrica,* 51, 1, 1983, pp. 153–174.

Bhat, C. "A Heteroscedastic Extreme Value Model of Intercity Mode Choice." *Transportation Research,* 30, 1, 1995, pp. 16–29.

Bhat, C. "Accommodating Variations in Responsiveness to Level-of-Service Measures in Travel Mode Choice Modeling." Department of Civil Engineering, University of Massachusetts, Amherst, 1996.

Bhat, C. "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model." Manuscript, Department of Civil Engineering, University of Texas, Austin, 1999.

Billingsley, P. *Probability and Measure*. 3rd ed. New York: John Wiley and Sons, 1995.

Binkley, J. "The Effect of Variable Correlation on the Efficiency of Seemingly Unrelated Regression in a Two Equation Model." *Journal of the American Statistical Association,* 77, 1982, pp. 890–895.

Binkley, J., and C. Nelson. "A Note on the Efficiency of Seemingly Unrelated Regression." *American Statistician,* 42, 1988, pp. 137–139.

Birkes, D., and Y. Dodge. *Alternative Methods of Regression*. New York: John Wiley and Sons, 1993.

Blinder, A. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources*, 8, 4, 1973, pp. 436–455.

Blundell, R., ed. "Specification Testing in Limited and Discrete Dependent Variable Models." *Journal of Econometrics,* 34, 1/2, 1987, pp. 1–274.

Blundell, R., and S. Bond. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics,* 87, 1998, pp. 115–143.

Blundell, R., M. Browning, and I. Crawford. "Nonparametric Engel Curves and Revealed Preference." *Econometrica,* 71, 1, 2003, pp. 205–240.

Blundell, R., F. Laisney, and M. Lechner. "Alternative Interpretations of Hours Information in an Econometric Model of Labour Supply." *Empirical Economics,* 18, 1993, pp. 393–415.

Blundell, R., and J. Powell. "Endogeneity in Semiparametric Binary Response Models." *Review of Economic Studies*, 71, 2004, pp. 655–679.

Bockstael, N., I. Strand, K. McConnell, and F. Arsanjani. "Sample Selection Bias in the Estimation of Recreation Demand Functions: An Application to Sport Fishing." *Land Economics,* 66, 1990, pp. 40–49.

Boes, S., and R. Winkelmann. "Ordered Response Models." Working paper 0507, Socioeconomic Institute, University of Zurich, 2005.

Boes, S., and R. Winkelmann. "Ordered Response Models." *Allgemeines Statistisches Archiv*, 90, 1, 2006a, pp. 165–180.

Boes, S., and R. Winkelmann. "The Effect of Income on Positive and Negative Subjective Well-Being." University of Zurich, Socioeconomic Institute, manuscript, IZA discussion paper Number 1175, 2006b.

Bogart, W., and B. Cromwell. "How Much Is a Neighborhood School Worth?" *Journal of Urban Economics*, 47, 2000, pp. 280–305.

Bollerslev, T. "Generalized Autoregressive Conditional Heteroscedasticity." *Journal of Econometrics,* 31, 1986, pp. 307–327.

Bollerslev, T., R. Chou, and K. Kroner. "ARCH Modeling in Finance." *Journal of Econometrics,* 52, 1992, pp. 5–59.

Bollerslev, T., and E. Ghysels. "Periodic Autoregressive Conditional Heteroscedasticity." *Journal of Business and Economic Statistics,* 14, 1996, pp. 139–151.

Bollerslev, T., and J. Wooldridge. "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances." *Econometric Reviews,* 11, 1992, pp. 143–172.

Bonjour, D., L. Cherkas, J. Haskel, D. Hawkes, and T. Spector. "Returns to Education: Evidence from U.K. Twins." *The American Economic Review*, 92, 5, 2003, pp. 1719–1812.

Boot, J., and G. deWitt. "Investment Demand: An Empirical Contribution to the Aggregation Problem." *International Economic Review,* 1, 1960, pp. 3–30.

Bornstein, M., and R. Bradley. *Socioeconomic Status, Parenting, and Child Development,* Lawrence Erlbaum Associates, London, 2003.

Börsch-Supan, A., and V. Hajivassiliou. "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models." *Journal of Econometrics,* 58, 3, 1990, pp. 347–368.

Boskin, M. "A Conditional Logit Model of Occupational Choice." *Journal of Political Economy,* 82, 1974, pp. 389–398.

Bound, J., D. Jaeger, and R. Baker. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables Is Weak." *Journal of the American Statistical Association,* 90, 1995, pp. 443–450.

Bourguignon, F., F. Ferriera, and P. Leite. "Beyond Oaxaca-Blinder: Accounting for Differences in Household Income Distributions Across Countries." Discussion Paper 452, Department of Economics, Pontifica University, Catolica do Rio de Janiero, 2002. www.econ.pucrio.rb/pdf/td452.pdf.

Bover, O., and M. Arellano. "Estimating Dynamic Limited Dependent Variable Models from Panel Data." *Investigaciones Economicas, Econometrics Special Issue,* 21, 1997, pp. 141–165.

Box, G., and D. Cox. "An Analysis of Transformations." *Journal of the Royal Statistical Society,* 1964, Series B, 1964, pp. 211–264.

Box, G., and G. Jenkins. *Time Series Analysis: Forecasting and Control.* 2nd ed. San Francisco: Holden Day, 1984.

Box, G., and M. Muller. "A Note on the Generation of Random Normal Deviates." *Annals of Mathematical Statistics,* 29, 1958, pp. 610–611.

Boyes, W., D. Hoffman, and S. Low. "An Econometric Analysis of the Bank Credit Scoring Problem." *Journal of Econometrics,* 40, 1989, pp. 3–14.

Brannas, K. "Explanatory Variables in the AR(1) Count Data Model." Working paper no. 381, Department of Economics, University of Umea, Sweden, 1995.

Brant, R. "Assessing Proportionality in the Proportional Odds Model for Ordered Logistic Regression." *Biometrics,* 46, 1990, pp. 1171–1178.

Breitung, J. "The Local Power of Some Unit Root Tests for Panel Data." *Advances in Econometrics,* 15, 2000, pp. 161–177.

Breslaw, J. "Evaluation of Multivariate Normal Probabilities Using a Low Variance Simulator." *Review of Economics and Statistics,* 76, 1994, pp. 673–682.

Breusch, T., and A. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica,* 47, 1979, pp. 1287–1294.

Breusch, T., and A. Pagan. "The LM Test and Its Applications to Model Specification in Econometrics." *Review of Economic Studies,* 47, 1980, pp. 239–254.

Brewer, C., C. Kovner, W. Greene, and Y. Cheng. "Predictors of RNs' Intent to Work and Work Decisions One Year Later in a U.S. National Sample." *The International Journal of Nursing Studies*, 46, 7, 2009, pp. 940–956.

Brock, W., and S. Durlauf. "Discrete Choice with Social Interactions." Working paper #2007, Department of Economics, University of Wisconsin, Madison, 2000.

Brown, S., M. Harris, and K. Taylor. "Modeling Charitable Donations toan Unexpected Natural Disaster: Evidence from the U.S. Panel Study of Income Dynamics." Institute for the Study of Labor, IZA, Working paper 4424, 2009.

Brown, C., and R. Moffitt. "The Effect of Ignoring Heteroscedasticity on Estimates of the Tobit Model." Mimeo, University of Maryland, Department of Economics, June 1982.

Brownstone, D., and C. Kazimi. "Applying the Bootstrap." Manuscript, Department of Economics, University of California Irvine, 1998.

Brundy, J., and D. Jorgenson. "Consistent and Efficient Estimation of Systems of Simultaneous Equations by Means of Instrumental Variables." *Review of Economics and Statistics,* 53, 1971, pp. 207–224.

Buchinsky, M. "Recent Advances in Quantile Regression Models: A Practical Guide for Empirical Research." *Journal of Human Resources*, 33, 1998, pp. 88–126.

Buckles, K., and D. Hungerman. "Season of Birth and Later Outcomes: Old Questions and New Answers." NBER working paper 14573, Cambridge, MA, 2008.

Burnett, N. "Gender Economics Courses in Liberal Arts Colleges." *Journal of Economic Education,* 28, 4, 1997, pp. 369–377.

Burnside, C., and M. Eichenbaum. "Small-Sample Properties of GMM-Based Wald Tests." *Journal of Business and Economic Statistics,* 14, 3, 1996, pp. 294–308.

Buse, A. "Goodness of Fit in Generalized Least Squares Estimation." *American Statistician,* 27, 1973, pp. 106–108.

Buse, A. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note." *American Statistician,* 36, 1982, pp. 153–157.

Business Week. "Learning Labor Market Lessons from Germany." accessed April 30, 2009, http://www.bloomberg.com/news/articles/2009-04-30/learning-labor-market-lessons-from-germany.

Butler, J., and R. Moffitt. "A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model." *Econometrica*, 50, 1982, pp. 761–764.

Butler, J., and P. Chatterjee. "Pet Econometrics: Ownershop of Cats and Dogs." Working paper 95-WP1, Department of Economics, Vanderbilt University, 1995.

Butler, J., and P. Chatterjee. "Tests of the Specification of Univariate and Bivariate Ordered Probit." *Review of Economics and Statistics,* 79, 1997, pp. 343–347.

Butler, J., T. Finegan, and J. Siegfried. "Does More Calculus Improve Student Learning in Intermediate Micro and Macro Economic Theory?" *American Economic Review,* 84, 1994, pp. 206–210.

Butler, R., J. McDonald, R. Nelson, and S. White. "Robust and Partially Adaptive Estimation of Regression Models." *Review of Economics and Statistics,* 72, 1990, pp. 321–327.

Calhoun, C. "BIVOPROB: Computer Program for Maximum-Likelihood Estimation of Bivariate Ordered-Probit Models for Censored Data, Version 11.92." *Economic Journal,* 105, 1995, pp. 786–787.

Cameron, C., and D. Miller. "A Practioner's Guide to Cluster-Robust Inference." *Journal of Human Resources,* 50, 2, 2015, pp. 317–373.

Cameron, A., and P. Trivedi. "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests." *Journal of Applied Econometrics,* 1, 1986, pp. 29–54.

Cameron, A., and P. Trivedi. "Regression-Based Tests for Overdispersion in the Poisson Model." *Journal of Econometrics,* 46, 1990, pp. 347–364.

Cameron, C., and P. Trivedi. *Regression Analysis of Count Data.* New York: Cambridge University Press, 1998.

Cameron, C., and P. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press, 2005.

Cameron, C., T. Li, P. Trivedi, and D. Zimmer. "Modeling the Differences in Counted Outcomes Using Bivariate Copula Models: With Applications to Mismeasured Counts." *Econometrics Journal,* 7, 2004, pp. 566–584.

Cameron, C., and F. Windmeijer. "*R*-Squared Measures for Count Data Regression Models with Applications to Health Care Utilization." Working paper no. 93–24, Department of Economics, University of California, Davis, 1993.

Campbell, J., A. Lo, and A. MacKinlay. *The Econometrics of Financial Markets*. Princeton: Princeton University Press, 1997.

Campbell, J., and G. Mankiw. "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence." Working paper 2924, NBER, Cambridge, MA, 1989.

Cappellari, L., and S. Jenkins. "Calculation of Multivariate Normal Probabilities by Simulation, with Applications to Maximum Simulated Likelihood Estimation." Discussion Paper 2112, IZA, 2006.

Card, D. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, 43, 1990, pp. 245–257.

Card, D. "The Effect of Unions on Wage Inequality in the U.S. Labor Market." *Industrial and Labor Relations Review*, 54, 2, 2001, pp. 296–315.

Card, D., and A. Krueger. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review*, 14, 4, 1994, pp. 772–794.

Card, D., and A. Krueger. "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania: Reply." *The American Economic Review*, 90, 2000, pp. 397–420.

Card, D., D. Lee, Z. Pei, and A. Weber. "Nonlinear Policy Rules and the Identification and Estimation of Causal Effects in a Generalized Regression Kink Design." NBER, Cambridge, MA, Working paper 18564, Nov. 2012.

Carey, K. "A Panel Data Design for Estimation of Hospital Cost Functions." *Review of Economics and Statistics,* 79, 3, 1997, pp. 443–453.

Carneiro, P., K. Hansen, and J. Heckman. "Estimating Distributions of Treatment Effects with an Application to Schooling and Measurement of the Effects of Uncertainty on College Choice." *International Economic Review*, 44, 2003, pp. 361–422.

Carro, J. "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects." *Journal of Econometrics*, 140, 2, 2007, pp. 503–528.

Case, A. "Spatial Patterns in Household Demand." *Econometrica,* 59, 4, 1991, pp. 953–965.

Case, A. "Neighborhood Influence and Technological Change." *Regional Science and Urban Economics*, 22, 3, September 1992: 491–508.

Casella, G., and E. George. "Explaining the Gibbs Sampler." *American Statistician,* 46, 3, 1992, pp. 167–174.

Cecchetti, S. "The Frequency of Price Adjustment: A Study of the Newsstand Prices of Magazines." *Journal of Econometrics,* 31, 3, 1986, pp. 255–274.

Cecchetti, S., and R. Rich. "Structural Estimates of the U.S. Sacrifice Ratio." *Journal of Business and Economic Statistics,* 19, 4, 2001, pp. 416–427.

Chamberlain, G. "Omitted Variable Bias in Panel Data: Estimating the Returns to Schooling." *Annales de L'Insee,* 30/31, 1978, pp. 49–82.

Chamberlain, G. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies,* 47, 1980, pp. 225–238.

Chamberlain, G. "Multivariate Regression Models for Panel Data." *Journal of Econometrics,* 18, 1, 1982, pp. 5–46.

Chamberlain, G. "Panel Data." In *Handbook of Econometrics*, edited by Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1984.

Chamberlain, G. "Heterogeneity, Omitted Variable Bias and Duration Dependence." In *Longitudinal Analysis of Labor Market Data,* Edited by J. Heckman and B. Singer, Cambridge: Cambridge University Press, 1985.

Chamberlain, G. "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions." *Journal of Econometrics,* 34, 1987, pp. 305–334.

Chen, T. "Root N Consistent Estimation of a Panel Data Sample Selection Model." Manuscript, Hong Kong University of Science and Technology, 1998.

Cheng, T., and P. Trivedi. "Attrition Bias in Panel Data: A Sheep in Wolf's Clothing? A Case Study Based on the MABEL Survey." *Health Economics*, 24, 9, 2015, pp. 1101–1117.

Chesher, A., and M. Irish. "Residual Analysis in the Grouped Data and Censored Normal Linear Model." *Journal of Econometrics,* 34, 1987, pp. 33–62.

Chesher, A., T. Lancaster, and M. Irish. "On Detecting the Failure of Distributional Assumptions." *Annales de L'Insee,* 59/60, 1985, pp. 7–44.

Cheung, S. "Provincial Credit Rating in Canada: An Ordered Probit Analysis." Bank of Canada, working paper 96–6, http://www.bankofcanada.ca/wp-content/uploads/2010/05/wp96-6.pdf, 1996.

Chung, C., and A. Goldberger. "Proportional Projections in Limited Dependent Variable Models." *Econometrica*, 52, 1984, pp. 531–534.

Chiappori, R. "Econometric Models of Insurance Under Asymmetric Information."

Manuscript, Department of Economics, University of Chicago, 1998.

Chou, R. "Volatility Persistence and Stock Valuations: Some Empirical Evidence Using GARCH." *Journal of Applied Econometrics,* 3, 1988, pp. 279–294.

Chib, S., and E. Greenberg. "Understanding the Metropolis-Hastings Alagorithm." *The American Statistician,* 49, 4, 1995, pp. 327–335.

Chib, S., and E. Greenberg. "Markov Chain Monte Carlo Simulation Methods in Econometrics." *Econometric Theory,* 12, 1996, pp. 409–431.

Chow, G. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica,* 28, 1960, pp. 591–605.

Chow, G. "Random and Changing Coefficient Models." In *Handbook of Econometrics,* Vol. 2, edited by Griliches, Z. and M. Intriligator, Amsterdam: North Holland, 1984.

Christensen, B., and Kallestrup-Lamb, M. "The Impact of Health Changes on Labor Supply: Evidence from Merged Data on Individual Objective Medical Diagnosis Codes and Early Retirement Behavior." *Health Economics*, 21, 2012, pp. 56–100.

Christensen, L., and W. Greene. "Economies of Scale in U.S. Electric Power Generation." *Journal of Political Economy,* 84, 1976, pp. 655–676.

Christensen, L., D. Jorgenson, and L. Lau. "Transcendental Logarithmic Utility Functions." *American Economic Review,* 65, 1975, pp. 367–383.

Christofides, L., T. Stengos, and R Swidinsky. "On the Calculation of Marginal Effects in the Bivariate Probit Model." *Economics Letters,* 54, 3, 1997, pp. 203–208.

Christofides, L., T. Hardin, and R. Stengos. "On the Calculation of Marginal Effects in the Bivariate Probit Model: Corrigendum." *Economics Letters,* 68, 2000, pp. 339–340.

Chung, C., and A. Goldberger. "Proportional Projections in Limited Dependent Variable Models." *Econometrica*, 52, 1984, pp. 531–534.

CIC. "Penn World Tables." Center for International Comparisons of Production, Income and Prices, University of Pennsylvania, http://cid.econ.ucdavis.edu/pwt.html, 2010.

Clark, A., Y. Georgellis, and P. Sanfey. "Scarring: The Psychological Impact of Past Unemployment." *Economica*, 68, 2001, pp. 221–241.

Cleveland, W. "Robust Locally Weighted Regression and Smoothing Scatter Plots." *Journal of the American Statistical Association,* 74, 1979, pp. 829–836.

Coakley, J., F. Kulasi, and R. Smith. "Current Account Solvency and the Feldstein-Horioka Puzzle." *Economic Journal,* 106, 1996, pp. 620–627.

Cochrane, D., and G. Orcutt. "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms." *Journal of the American Statistical Association,* 44, 1949, pp. 32–61.

Coelli, T. "Recent Developments in Frontier Modelling and Efficiency Measurement." *Australian Journal of Agricultural and Resource Economics*, 39, 3, 1995, pp. 219–245.

Coelli, T. "Frontier 4.1." CEPA working paper, Centre for Efficiency and Productivity Analysis, University of Queensland, 1996, www.uq.edu.au/economics/cepa/frontier.htm

Colombi, C., A. Martini, and S. Vittadini. "Closed Skew Normality in Stochastic Frontiers with Individual Effects and Long/Short Run Efficiency." *Journal of Productivity Analysis*, 42, 2014, pp. 123–136.

Cohen, R., and Wallace, J. "A-Rod: Signing the Best Player in Baseball." *Harvard Business School, Case 9-203-047*, Cambridge, 2003.

Congdon, P. *Bayesian Models for Categorical Data*. New York: John Wiley and Sons, 2005.

Conway, D., and H. Roberts. "Reverse Regression, Fairness and Employment Discrimination." *Journal of Business and Economic Statistics,* 1, 1, 1983, pp. 75–85.

Contoyannis, C., A. Jones, and N. Rice. "The Dynamics of Health in the British Household Panel Survey." *Journal of Applied Econometrics,* 19, 4, 2004, pp. 473–503.

Cook, D. "Influential Observations in Linear Regression." *Journal of the American Statistical Association*, 74, 365, 1977, pp. 169–174.

Cornwell, C., and P. Rupert. "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators." *Journal of Applied Econometrics,* 3, 1988, pp. 149–155.

Cornwell, C., and P. Schmidt. "Panel Data with Cross-Sectional Variation in Slopes as Well as in Intercept." Econometrics workshop paper no. 8404, Michigan State University, Department of Economics, 1984.

Coulson, N., and R. Robins. "Aggregate Economic Activity and the Variance of Inflation: Another Look." *Economics Letters,* 17, 1985, pp. 71–75.

Council of Economic Advisors. *Economic Report of the President*. Washington, D.C.: United States Government Printing Office, 1994.

Council of Economic Advisors. *Economic Report of the President*. Washington, D.C.: United States Government Printing Office, 2016.

Cox, D. "Tests of Separate Families of Hypotheses." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1. Berkeley: University of California Press, 1961.

Cox, D. "Further Results on Tests of Separate Families of Hypotheses." *Journal of the Royal Statistical Society, Series B,* 24, 1962, pp. 406–424.

Cox, D. "Regression Models and Life Tables." *Journal of the Royal Statistical Society, Series B,* 34, 1972, pp. 187–220.

Cox, D., and D. Oakes. *Analysis of Survival Data*. New York: Chapman and Hall, 1985.

Cragg. J. "On the Relative Small Sample Properties of Several Structural Equations Estimators." *Econometrica*, 35, 1967, pp. 136–151.

Cragg, J. "Some Statistical Models for Limited Dependent Variables with Application to theDemand for Durable Goods.' *Econometrica*, 39, 5, 1971, pp. 829–844.

Cragg, J. "Estimation and Testing in Testing in Time Series Regression Models with Heteroscedastic Disturbances." *Journal of Econometrics,* 20, 1982, pp. 135–157.

Cragg., J. "Using Higher Moments to Estimate the Simple Errors in Variables Model." *Rand Journal of Economics,* 28, 0, 1997, pp. S71–S91.

Cragg, J., and R. Uhler. "The Demand for Automobiles." *Canadian Journal of Economics,* 3, 1970, pp. 386–406.

Cramèr, H. *Mathematical Methods of Statistics*. Princeton: Princeton University Press, 1948.

Cramer, J. "Predictive Performance of the Binary Logit Model in Unbalanced Samples." *Journal of the Royal Statistical Society, Series D (The Statistician),* 48, 1999, pp. 85–94.

Creel, M., and J. Loomis. "Theoretical and Empirical Advantages of Truncated Count Data Estimators for Analysis of Deer Hunting in California." *American Journal of Agricultural Economics*, 72, 1990, pp. 434–441.

Culver, S., and D. Pappell. "Is There a Unit Root in the Inflation Rate? Evidence from Sequential Break and Panel Data Model." *Journal of Applied Econometrics,* 12, 1997, pp. 435–444.

Cumby, R., J. Huizinga, and M. Obstfeld. "Two-Step, Two-Stage Least Squares Estimation in Models with Rational Expectations." *Journal of Econometrics,* 21, 1983, pp. 333–355.

Cuesta, R. "A Production Model with Firm-Specific Temporal Variation in Technical Inefficiency: With Application to Spanish Dairy Farms." *Journal of Productivity Analysis*, 13, 2, 2000, pp. 139–158.

Cunha, F., J. Heckman, and S. Navarro. "The Identification & Economic Content of Ordered Choice Models with Stochastic Thresholds." University College Dublin, Gery Institute, discussion paper WP/26/2007, 2007.

D'Addio, A., T. Eriksson, and P. Frijters. "An Analysis of the Determinants of Job Satisfaction When Individuals' Baseline Satisfaction Levels May Differ." Working paper 2003-16, Center for Applied Microeconometrics, University of Copenhagen, 2003.

Dahlberg, M., and E. Johansson. "An Examination of the Dynamic Behaviour of Local Governments Using GMM Bootstrapping Methods." *Journal of Applied Econometrics,* 15, 2000, pp. 401–416.

Dale, S., and A. Krueger. "Estimating the Return to College Selectivity of the Career Using Administrative Earnings Data." NBER, Cambridge, MA, Working paper 17159, 2011.

Dale, S., and A. Krueger. "Estimating the Payoff of Attending a More Selective College: An Application of Selection on Observables and Unobservables." *Quarterly Journal of Economics*, 107, 4, 2002, pp. 1491–1527.

Daly, A., S. Hess, and K. Train. "Assuring Finite Moments for Willingness to Pay in Random Coefficient Models." Institute for Transport Studies, University of Leeds, October, 2009.

Das, M., S. Olley, and A. Pakes. "The Evolution of the Market for Consumer Electronics." mimeo, Department of Economics, Harvard University, 1996.

Das, M., and A. van Soest. "A Panel Data Model for Subjective Information on Household Income Growth." *Journal of Economic Behavior and Organization,* 40, 2000, 409–426.

Dastoor, N. "Some Aspects of Testing Nonnested Hypotheses." *Journal of Econometrics,* 21, 1983, pp. 213–228.

Davidson, A., and D. Hinkley. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press, 1997.

Davidson, J. *Econometric Theory*. Oxford: Blackwell, 2000.

Davidson, R., and J. MacKinnon. "Several Tests for Model Specification in the Presence of Alternative Hypotheses." *Econometrica,* 49, 1981, pp. 781–793.

Davidson, R., and J. MacKinnon. "Model Specification Tests Based on Artificial Linear Regressions." *International Economic Review*, 25, 1984, pp. 485–502.

Davidson, R., and J. MacKinnon. *Estimation and Inference in Econometrics*. New York: Oxford University Press, 1993.

Davidson, R., and J. MacKinnon. *Econometric Theory and Methods.* New York: Oxford University Press, 2004.

Davidson, R., and J. MacKinnon. "Bootstrap Methods in Econometrics." In *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory,* edited by T. Mills and K. Patterson, Hampshire: Palgrave Macmillan, 2006.

Davies, R. "Evaluation of an OFT Intervention." UK Office of Fair Trading, WP 1416, http://dera.ioe.ac.uk/14610/1/oft1416.pdf, 2012.

Daykin, A., and P. Moffatt. "Analyzing Ordered Responses: A Review of the Ordered Probit Model." *Understanding Statistics*, I, 3, 2002, pp. 157–166.

Deaton, A. "Model Selection Procedures, or, Does the Consumption Function Exist?" In *Evaluating the Reliability of Macroeconomic Models*, edited by G. Chow and P. Corsi, New York: John Wiley and Sons, 1982.

Deaton A. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore: Johns Hopkins University Press, 1997.

Deaton, A. "Health, Inequality and Economic Development." *Journal of Economic Literature*, 41, 1, 2003, pp. 113–150.

Deaton, A., and J. Muellbauer. *Economics and Consumer Behavior*. New York: Cambridge University Press, 1980.

Deb, P., and P. K. Trivedi. "The Structure of Demand for Health Care: Latent Class versus Two-part Models." *Journal of Health Economics,* 21, 2002, pp. 601– 625.

Debreu, G. "The Coefficient of Resource Utilization." *Econometrica,* 19, 3, 1951, pp. 273–292.

DeFusco, A., and A. Paciorek. "The Interest Rate Elasticity of Mortgage Demand: Evidence from Bunching at the Conforming Loan Limit." *American Economic Journal: Economic Policy*, 2016, Forthcoming

DeFusco, A., and A. Paciorek. "The Interest Rate Elasticity of Mortgage Demand: Evidence from Bunching at the Conforming Loan Limit." Finance and Research Discussion Series, Federal Reserve Board, Washington DC, Working paper 2014-11, 2014.

Dehejia, R. "Practical Propensity Score Matching, A Reply to Smith and Todd." *Journal of Econometrics*, 125, 2005, pp. 355–364.

Dehejia, R., and S. Wahba. "Causal Effects in Non-experimental Studies: Evaluating the Valuation of Training Programs." *Journal of the American Statistical Association,* 94, 1999, pp. 1053–1062.

DeMaris, A. *Regression with Social Data: Modeling Continuous and Limited Response Variables*. Hoboken, NJ: Wiley, 2004.

Dempster, A., N. Laird, and D. Rubin. "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B,* 39, 1977, pp. 1–38.

DesChamps, P. "Full Maximum Likelihood Estimation of Dynamic Demand Models." *Journal of Econometrics,* 82, 1998, pp. 335–359.

De Vany, A. *Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry*. New York: Routledge, 2003.

De Vany, A., and D. Walls. "Uncertainty in the Movies: Can Star Power Reduce the Terror of the Box Office?" *Journal of Cultural Economics*, 23, 4, 1999, pp. 285–318.

De Vany, A., and D. Walls. "Does Hollywood Make Too Many R-Rated Movies? Risk, Stochastic Dominance, and the Illusion of Expectation." *The Journal of Business*, 75, 3, 2002, pp. 425–451.

De Vany, A., and D. Walls. "Movie Stars, Big Budgets, and Wide Releases: Empirical Analysis of the Blockbuster Strategy." In *Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry*, edited by Arthur De Vany, New York: Routledge, 2003.

Dezhbaksh, H. "The Inappropriate Use of Serial Correlation Tests in Dynamic Linear Models." *Review of Economics and Statistics,* 72, 1990, pp. 126–132.

Dhrymes, P. "Limited Dependent Variables." In *Handbook of Econometrics,* Vol. 2, edited by Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1984.

Dickey, D., and W. Fuller. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association,* 74, 1979, pp. 427–431.

Dickey, D., and W. Fuller. "Likelihood Ratio Tests for Autoregressive Time Series with a Unit Root." *Econometrica,* 49, 1981, pp. 1057–1072.

Diebold, F. *Elements of Forecasting*. Cincinnati: South-Western. 4th ed., 2007.

Dielman, T. *Pooled Cross-Sectional and Time Series Data Analysis*. New York: Marcel-Dekker, 1989.

Diewert, E. "Applications of Duality Theory." In *Frontiers in Quantitative Economics*, edited by M. Intriligator and D. Kendrick, Amsterdam: North Holland, 1974.

Di Maria, C., S. Ferreira, and E. Lazarova. "Shedding Light on the Light Bulb Puzzle: The Role of Attitudes and Perceptions in the Adoption of Energy Efficient Light Bulbs." *Scottish Journal of Political Economy*, 57, 1, 2010, pp. 48–68.

Domowitz, I., and C. Hakkio. "Conditional Variance and the Risk Premium in the Foreign Exchange Market." *Journal of International Economics,* 19, 1985, pp. 47–66.

Donald, S., and K. Lang. "Inference with Difference-in-Differences and Other Panel Data." *Review of Economics and Statistics*, 89, 2, 2007, pp. 221–233.

Dong, Y., and A. Lewbel. "Simple Estimators for Binary Choice Models with Endogenous Regressors." unpublished manuscript, Department of Economics, Boston College, 2010 (posted at http://www2.bc.edu/~lewbel/simplenew8.pdf).

Doob, J. *Stochastic Processes.* New York: John Wiley and Sons, 1953.

Doppelhofer, G., R. Miller, and S. Sala-i-Martin. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." NBER Working paper no. 7750, June, 2000.

Dowd, B., W. Greene, and E. Norton. "Computation of Standard Errors." *Health Services Research*, 29, 2, 2014, pp. 731–750.

Duan, N. "Smearing Estimate: A Nonparametrics Retransformation Method." *Journal of the American Statistical Association*, 78, 1983, pp. 605–612.

Duncan, G. "A Semi-parametric Censored Regression Estimator." *Journal of Econometrics,* 31, 1986a, pp. 5–34.

Duncan, G., ed. "Continuous/Discrete Econometric Models with Unspecified Error Distribution." *Journal of Econometrics,* 32, 1, 1986b, pp. 1–187.

Dunlap, R. "The New Environmental Paradigm Scale: From Marginality to Worldwide Use." *Journal of Environmental Education*, 40, 1, 2008, pp. 3–18.

Durbin, J. "Errors in Variables." *Review of the International Statistical Institute,* 22, 1954, pp. 23–32.

Durbin, J. "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors Are Lagged Dependent Variables." *Econometrica,* 38, 1970, pp. 410–421.

Durbin, J., and G. Watson. "Testing for Serial Correlation in Least Squares Regression—I." *Biometrika,* 37, 1950, pp. 409–428.

Durbin, J., and G. Watson. "Testing for Serial Correlation in Least Squares Regression—II." *Biometrika,* 38, 1951, pp. 159–178.

Durbin, J., and G. Watson. "Testing for Serial Correlation in Least Squares Regression—III." *Biometrika,* 58, 1971, pp. 1–42.

Dwivedi, T., and K. Srivastava. "Optimality of Least Squares in the Seemingly Unrelated Regressions Model." *Journal of Econometrics,* 7, 1978, pp. 391–395.

Efron, B. "Regression and ANOVA with Zero-One Data: Measures of Residual Variation." *Journal of the American Statistical Association,* 73, 1978, pp. 113–212.

Efron, B. "Bootstrapping Methods: Another Look at the Jackknife." *Annals of Statistics,* 7, 1979, pp. 1–26.

Efron, B., and R. Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1994.

Egan, K., and J. Herriges. "Multivariate Count Data Regression Models with Individual Panel Data from an On-Site Sample." *Journal of Environmental Economics and Management*, 52, 2, 2006, pp. 567–581.

Eichengreen, B., M. Watson, and R. Grossman. "Bank Rate Policy Under the Interwar Gold Standard: A Dynamic Probit Approach." *Economic Journal*, 95, 1985, pp. 725–745.

Eicker, F. "Limit Theorems for Regression with Unequal and Dependent Errors." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. LeCam and J. Neyman, Berkeley: University of California Press, 1967, pp. 59–82.

Eisenberg, D., and B. Rowe. "The Effect of Serving in the Vietnam War on Smoking Behavior Later in Life." Manuscript, School of Public Health, University of Michigan, 2006.

Elliot, G., T. Rothenberg, and J. Stock. "Efficient Tests for an Autoregressive Unit Root." *Econometrica,* 64, 1996, pp. 813–836.

Eluru, N., C. Bhat, and D. Hensher. "A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Levels in Traffic Crashes." *Accident Analysis and Prevention*, 40, 3, 2008, pp. 1033–1054.

Enders, W. *Applied Econometric Time Series.* 2nd ed., New York: John Wiley and Sons, 2004.

Engle, R. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflations." *Econometrica,* 50, 1982, pp. 987–1008.

Engle, R. "Estimates of the Variance of U.S. Inflation Based on the ARCH Model." *Journal of Money, Credit, and Banking,* 15, 1983, pp. 286–301.

Engle, R. "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics." In *Handbook of Econometrics,* Vol. 2, edited by Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1984.

Engle, R., and C. Granger. "Co-integration and Error Correction: Representation, Estimation, and Testing." *Econometrica,* 35, 1987, pp. 251–276.

Engle, R., and D. Hendry. "Testing Super Exogeneity and Invariance." *Journal of Econometrics,* 56, 1993, pp. 119–139.

Engle, R., D. Hendry, and J. Richard. "Exogeneity." *Econometrica,* 51, 1983, pp. 277–304.

Engle, R., D. Hendry, and D. Trumble. "Small Sample Properties of ARCH Estimators and Tests." *Canadian Journal of Economics,* 18, 1985, pp. 66–93.

Engle, R., and M. Rothschild. "ARCH Models in Finance." *Journal of Econometrics,* 52, 1992, pp. 1–311.

Engle, R., D. Lilen, and R. Robins. "Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model." *Econometrica,* 55, 1987, pp. 391–407.

Engle, R., and B. Yoo. "Forecasting and Testing in Cointegrated Systems." *Journal of Econometrics,* 35, 1987, pp. 143–159.

Englin, J., and J. Shonkwiler. "Estimating Social Welfare Using Count Data Models: An Application to Long-Run Recreation Demand Under Conditions of Endogenous Stratification and Truncation." *Review of Economics and Statistics*, 77, 1995, pp. 104–112.

Estes, E., and B. Honorè. "Partially Linear Regression Using One Nearest Neighbor." Manuscript, Department of Economics, Princeton University, 1995.

Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions,* 4th ed. New York: John Wiley and Sons, 2010.

Evans, D., A. Tandon, C. Murray, and J. Lauer. "The Comparative Efficiency of National Health Systems in Producing Health: An Analysis of 191 Countries." World Health Organization, GPE discussion paper, no. 29, EIP/GPE/EQC, 2000a.

Evans D., A. Tandon, C. Murray, and J. Lauer. "Measuring Overall Health System Performance for 191 Countries." World Health Organization GPE Discussion Paper, No. 30, EIP/GPE/EQC, 2000b.

Evans, G., and N. Savin. "Testing for Unit Roots: I." *Econometrica,* 49, 1981, pp. 753–779.

Evans, G., and N. Savin. "Testing for Unit Roots: II." *Econometrica,* 52, 1984, pp. 1241–1269.

Evans, W., and R. Schwab. "Finishing High and Starting College: Do Catholic Schools Make a Difference." *Quarterly Journal of Economics*, 110, 4, 1995, pp. 971–974.

Fair, R. "A Note on Computation of the Tobit Estimator." *Econometrica,* 45, 1977, pp. 1723–1727.

Fair, R. "A Theory of Extramarital Affairs." *Journal of Political Economy,* 86, 1978, pp. 45–61.

Fair, R. *Specification and Analysis of Macroeconomic Models*. Cambridge: Harvard University Press, 1984.

Farrell, M. "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society, Series A, General,* 120, part 3, 1957, pp. 253–291.

Farsi, M., M. Filippini, and W. Greene. "Efficiency Measurement in Network Industries, Application to the Swiss Railroads." *Journal of Regulatory Economics*, 28, 1, 2005, pp. 69–90.

Feldstein, M. "The Error of Forecast in Econometric Models When the Forecast-Period Exogenous Variables Are Stochastic." *Econometrica,* 39, 1971, pp. 55–60.

Fernandez, A., and J. Rodriguez-Poo. "Estimation and Testing in Female Labor Participation Models: Parametric and Semiparametric Models." *Econometric Reviews,* 16, 1997, pp. 229–248.

Fernandez, L. "Nonparametric Maximum Likelihood Estimation of Censored Regression Models." *Journal of Econometrics,* 32, 1, 1986, pp. 35–38.

Fernandez-Val, I. "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models." *Journal of Econometrics*, 150, 1, 2009, pp. 71–85.

Ferrer-i-Carbonel, A., and P. Frijters. "The Effect of Metholodogy on the Determinants of Happiness." *Economic Journal,* 114, 2004, pp. 641–659.

Fiebig, D., M. Keane, J. Louviere, and N. Wasi. "The Generalized Multinomial Logit: Accounting for Scale and Coefficient Heterogeneity." *Marketing Science*, published online before print July 23, DOI:10.1287/mksc.1090.0508, 2009.

Filippini, M., and W. Greene. "Persistent and Transient Productive Inefficiency: A Maximum Simulated Likelihood Approach." *Journal of Productivity Analysis*, 45, 2, 2016, pp. 187–196.

Fin, T., and P. Schmidt. "A Test for the Tobit Specification versus an Alternative Suggested by Cragg." *Review of Economics and Statistics,* 66, 1984, pp. 174–177.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. Newhouse, H. Allen, and K. Baicker. "The Oregon Health Insurance Experiment: Evidence from the First Year." The Oregon Health Study Group, NBER Working paper no. 17190, 2011.

Finney, D. *Probit Analysis*. Cambridge: Cambridge University Press, 1971.

Fiorentini, G., G. Calzolari, and L. Panattoni. "Analytic Derivatives and the Computation of GARCH Estimates." *Journal of Applied Econometrics,* 11, 1996, pp. 399–417.

Fisher, R. "The Theory of Statistical Estimation." *Proceedings of the Cambridge Philosophical Society,* 22, 1925, pp. 700–725.

Fisher, G., and D. Nagin. "Random versus Fixed Coefficients Coefficient Quantal Choice Models." In *Structural Analysis of Discrete Data with Econometric Applications*, C. Manski and D. McFadden, Cambridge: MIT Press, 1981.

Fitzgerald, J., P. Gottshalk, and R. Moffitt. "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study on Income Dynamics." *Journal of Human Resources,* 33, 1998, pp. 251–299.

Fleissig, A., and J. Strauss. "Unit Root Tests on Real Wage Panel Data for the G7." *Economics Letters,* 54, 1997, pp. 149–155.

Fletcher, R. *Practical Methods of Optimization*. New York: John Wiley and Sons, 1980.

Flores-Lagunes, A. and K. Schnier. "Sample Selection and Spatial Dependence." *Journal of Applied Econometrics*, 27, 2, 2012, pp. 173–204.

Florens, J., D. Fougere, and M. Mouchart. "Duration Models." In *The Econometrics of Panel Data,* 2nd ed., edited by L. Matyas and P. Sevestre, Norwell, MA: Kluwer, 1996.

Fomby, T., C. Hill, and S. Johnson. *Advanced Econometric Methods*. Needham, MA: Springer-Verlag, 1984.

Fowler, C., J. Cover, and R. Kleit. "The Geography of Fringe Banking." *Journal of Regional Science*, 54, 4, 2014, pp. 688–710.

Frankel, J., and A. Rose. "A Panel Project on Purchasing Power Parity: Mean Reversion Within and Between Countries." *Journal of International Economics,* 40, 1996, pp. 209–224.

Freedman, D. "On the So-Called 'Huber Sandwich Estimator' and Robust Standard Errors." *The American Statistician,* 60, 4, 2006, pp. 299–302.

French, K., W. Schwert, and R. Stambaugh. "Expected Stock Returns and Volatility." *Journal of Financial Economics,* 19, 1987, pp. 3–30.

Fried, H., K. Lovell, and S. Schmidt, eds. *The Measurement of Efficiency*. Oxford: Oxford University Press, 2008.

Friedman, M. *A Theory of the Consumption Function*. Princeton: Princeton University Press, 1957.

Frijters P., J. Haisken-DeNew, and M. Shields. "The Value of Reunification in Germany: An Analysis of Changes in Life Satisfaction." *Journal of Human Resources,* 39, 3, 2004, pp. 649–674

Frisch, R. "Editorial." *Econometrica,* 1, 1933, pp. 1–4.

Frisch, R., and F. Waugh. "Partial Time Regressions as Compared with Individual Trends." *Econometrica,* 1, 1933, pp. 387–401.

Frolich, M. "Nonparametric Regression for Binary Dependent Variables." *Econometrics Journal*, 9, 2006, pp. 511–540.

Fu, A., M. Gordon, G. Liu, B. Dale, and R. Christensen. "Inappropriate Medication Use and Health Outcomes in the Elderly." *Journal of the American Geriatrics Society*, 52, 11, 2004, pp. 1934–1939.

Fuller, W., and G. Battese. "Estimation of Linear Models with Crossed-Error Structure." *Journal of Econometrics,* 2, 1974, pp. 67–78.

Gallant, A. *Nonlinear Statistical Models*. New York: John Wiley and Sons, 1987.

Gallant, A., and A. Holly. "Statistical Inference in an Implicit Nonlinear Simultaneous Equation in the Context of Maximum Likelihood Estimation." *Econometrica,* 48, 1980, pp. 697–720.

Gallant, R., and T. Nychka. "Semiparametric Maximum Likelihood Estimation." *Econometrica*, 55, 1987, pp. 363–390.

Gallant, R., and H. White. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Basil Blackwell, 1988.

Gannon, B. "A Dynamic Analysis of Disability and Labour Force Participation in Ireland 1995–2000." *Health Economics*, 14, 2005, pp. 925–938.

Garber, S., and S. Klepper. "Extending the Classical Normal Errors in Variables Model." *Econometrica,* 48, 1980, pp. 1541–1546.

Garrett, T., G. Wagner, and D. Wheelock. "A Spatial Analysis of State Banking Regulation." St. Louis Federal Reserve Bank working paper 2003-044, St. Louis, 2003.

Gaudry, M., and M. Dagenais. "The Dogit Model." *Transportation Research*, 13, 1979, pp. 105–112.

Gaver, K., and M. Geisel. "Discriminating Among Alternative Models: Bayesian and Non-Bayesian Methods." In *Frontiers in Econometrics,* P. Zarembka, ed., New York, Academic Press, pp. 49–77.

Gentle, J. *Random Number Generation and Monte Carlo Methods.* 2nd ed., Springer, New York, 2003.

Gelman, A. *Bayesian Data Analysis*, Boca Raton, FL, Chapman and Hall, 2003.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis.* 2nd ed., Suffolk: Chapman and Hall, 2004.

Gentle, J. *Elements of Computational Statistics*. New York: Springer-Verlag, 2002.

Gentle, J. *Random Number Generation and Monte Carlo Methods.* 2nd ed., New York: Springer-Verlag, 2003.

Gertler, P. "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESSA's Control Randomized Experiment." *The American Economic Review*, 94, 2, 2004, pp. 336–341.

Geweke, J. "Exact Inference in the Inequality Constrained Normal Linear Regression Model." *Journal of Applied Econometrics,* 2, 1986, pp. 127–142.

Geweke, J. "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference." *Journal of Econometrics,* 38, 1988, pp. 73–90.

Geweke, J. "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica,* 57, 1989, pp. 1317–1340.

Geweke, J. *Contemporary Bayesian Econometrics and Statistics.* New York: John Wiley and Sons, 2005.

Geweke, J., M. Keane, and D. Runkle. "Alternative Computational Approaches to Inference in the Multinomial Probit Model." *Review of Economics and Statistics,* 76, 1994, pp. 609–632.

Geweke, J., M. Keane, and D. Runkle. "Statistical Inference in the Multinomial Multiperiod Probit Model." *Journal of Econometrics,* 81, 1, 1997, pp. 125– 166.

Gill, J. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Suffolk: Chapman and Hall, 2002.

Godfrey, L. *Misspecification Tests in Econometrics*. Cambridge: Cambridge University Press, 1988.

Godfrey, L. "Instrument Relevance in Multivariate Linear Models." *Review of Economics and Statistics,* 81, 1999, pp. 550–552.

Goffe, W., G. Ferrier, and J. Rodgers. "Global Optimization of Statistical Functions with Simulated Annealing." *Journal of Econometrics,* 60, 1/2, 1994, pp. 65–100.

Golan, A. "Information and Entropy Econometrics—A Review and Synthesis." *Foundations and Trends in Econometrics*, 2, 1–2, pp. 1–145, 2009.

Golan, A., G. Judge, and D. Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. New York: John Wiley and Sons, 1996.

Goldberg, P. "Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry." *Econometrica*, 63, 4, 1995, pp. 891–951.

Goldberger, A. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Discussion paper 123-72, Institute for Research on Poverty, University of Wisconsin, Madison, 1972.

Goldberger, A. "Dependency Rates and Savings Rates: Further Comment." *American Economic Review,* 63, 1, 1973, pp. 232–233.

Goldberger, A. "Linear Regression After Selection." *Journal of Econometrics,* 15, 1981, pp. 357–366.

Goldberger, A. "Abnormal Selection Bias." In *Studies in Econometrics, Time Series, and Multivariate Statistics*, edited by S. Karlin, T. Amemiya, and L. Goodman, New York: Academic Press, 1983.

Goldberger, A. *A Course in Econometrics*. Cambridge: Harvard University Press, 1991.

Goldberger, A. "Selection Bias in Evaluating Greatment Effects: Some Formal Illustrations." In *Modelling and Evaluating Treatment Effects in Econometrics, Advances in Econometrics*, 21, edited by S. Karlin, T. Amemiya, and L. Goodman, Oxford: Elsevier, 2008.

Goldfeld, S., and R. Quandt. *Nonlinear Methods in Econometrics*. Amsterdam: North Holland, 1971.

Goldfeld, S., R. Quandt, and H. Trotter. "Maximization by Quadratic Hill Climbing." *Econometrica,* 1966, pp. 541–551.

Gonzaláz, P., and W. Maloney. "Logit Analysis in a Rotating Panel Context and an Application to Self-Employment Decisions." Policy Research working paper no. 2069, Washington, D.C: World Bank, 1999.

Gordin, M. "The Central Limit Theorem for Stationary Processes." *Soviet Mathematical Dokl.*, 10, 1969, pp. 1174–1176.

Gourieroux, C., and A. Monfort. "Testing Nonnested Hypotheses." In *Handbook of Econometrics,* Vol. 4, edited by Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1994.

Gourieroux, C., and A. Monfort. "Testing, Encompassing, and Simulating Dynamic Econometric Models." *Econometric Theory,* 11, 1995, pp. 195–228.

Gourieroux, C., and A. Monfort. *Simulation-Based Methods Econometric Methods*. Oxford: Oxford University Press, 1996.

Gourieroux, C., A. Monfort, and A. Trognon. "Testing Nested or Nonnested Hypotheses." *Journal of Econometrics,* 21, 1983, pp. 83–115.

Gourieroux, C., A. Monfort, and A. Trognon. "Pseudo Maximum Likelihood Methods: Applications to Poisson Models." *Econometrica,* 52, 1984, pp. 701–720.

Gourieroux, C., A. Monfort, E. Renault, and A. Trognon. "Generalized Residuals." *Journal of Econometrics,* 34, 1987, pp. 5–32.

Granger, C., and P. Newbold. "Spurious Regressions in Econometrics." *Journal of Econometrics,* 2, 1974, pp. 111–120.

Granger, C., and M. Pesaran. "A Decision Theoretic Approach to Forecast Evaluation." In *Statistics and Finance: An Interface*, edited by W. S. Chan, W. Li, and H. Tong, London: Imperial College Press, 2000.

Gravelle H., R. Jacobs, A. Jones, and A. Street. "Comparing the Efficiency of National Health Systems: Econometric Analysis Should Be Handled with Care." Manuscript, University of York, Health Economics, UK, 2002a.

Gravelle H., R. Jacobs, A. Jones, and A. Street. "Comparing the Efficiency of National Health Systems: A Sensitivity Approach." Manuscript, University of York, *Health Economics*, UK, 2002b.

Greenberg, E., and C. Webster. *Advanced Econometrics: A Bridge to the Literature*. New York: John Wiley and Sons, 1983.

Greene, W. "Maximum Likelihood Estimation of Econometric Frontier Functions." *Journal of Econometrics,* 13, 1980a, pp. 27–56.

Greene, W. "On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model." *Econometrica,* 48, 1980b, pp. 505–514.

Greene, W. "Sample Selection Bias as a Specification Error: Comment." *Econometrica,* 49, 1981, pp. 795–798.

Greene, W. "Estimation of Limited Dependent Variable Models by Ordinary Least Squares and the Method of Moments." *Journal of Econometrics,* 21, 1983, pp. 195–212.

Greene, W. "A Gamma Distributed Stochastic Frontier Model." *Journal of Econometrics,* 46, 1990, pp. 141–163.

Greene, W. "A Statistical Model for Credit Scoring." Working paper no. EC-92-29, Department of Economics, Stern School of Business, New York University, 1992.

Greene, W. "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models." Working paper no. EC-94-10, Department of Economics, Stern School of Business, New York University, 1994.

Greene, W. "Count Data." Manuscript, Department of Economics, Stern School of Business, New York University, 1995a.

Greene, W. "Sample Selection in the Poisson Regression Model." Working paper no. EC-95-6, Department of Economics, Stern School of Business, New York University, 1995b.

Greene, W. "Marginal Effects in the Bivariate Probit Model." Working paper no. 96-11, Department of Economics, Stern School of Business, New York University, 1996.

Greene, W. "FIML Estimation of Sample Selection Models for Count Data." Working paper no. 97-02, Department of Economics, Stern School of Business, New York University, 1997.

Greene, W. "Gender Economics Courses in Liberal Arts Colleges: Further Results." *Journal of Economic Education,* 29, 4, 1998, pp. 291–300.

Greene W. "Marginal Effects in the Censored Regression Model." *Economics Letters,* 64, 1, 1999, pp. 43–50.

Greene, W. "Fixed and Random Effects in Nonlinear Models." Working paper EC-01-01, Department of Economics, Stern School of Business, New York University, 2001.

Greene, W. "Simulated Maximum Likelihood Estimation of the Normal-Gamma Stochastic Frontier Model." *Journal of Productivity Analysis*, 19, 2, 2003, pp. 179–190.

Greene, W. "The Behavior of the Fixed Effects Estimator in Nonlinear Models." *The Econometrics Journal*, 7, 1, 2004a, pp. 98–119.

Greene, W. "Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems." *Health Economics*, 13, 2004b, pp. 959–980.

Greene, W. "Convenient Estimators for the Panel Probit Model." *Empirical Economics,* 29, 1, 2004c, pp. 21–47.

Greene, W. "Fixed Effects and Bias Due to the Incidental Parameters Problem in the Tobit Model." *Econometric Reviews,* 23, 2, 2004d, pp. 125–147.

Greene, W. "Functional Form and Heterogeneity in Models for Count Data." *Foundations and Trends in Econometrics*, 1, 2, 2005, pp. 1–110.

Greene, W. "The Econometric Approach to Efficiency Analysis." In *The Measurement of Productive Efficiency,* 2nd ed., edited by H. Fried, K. Lovell, and S. Schmidt, Oxford: Oxford University Press, 2007a.

Greene, W. *LIMDEP 9.0 Reference Guide.* Plainview, NY: Econometric Software, Inc., 2007b.

Greene, W. "A Statistical Model for Credit Scoring." In *Credit Risk: Quantitative Methods and Analysis*, D. Hensher and S. Jones, eds. Cambridge University Press, Cambridge, 2007c.

Greene, W. "Functional Form and Heterogeneity and Models for Count Data." Working paper EC-07-10, Department of Economics, Stern School of Business, New York University, 2007d.

Greene, W. "Discrete Choice Models." In *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics,* edited by T. Mills and K. Patterson, Hampshire: Palgrave, 2008a.

Greene, W. "Functional Forms for the Negative Binomial Model for Count Data." *Economics Letters*, 99, 3, 2008b, pp. 585–590.

Greene, W. *Econometric Analysis*. 6th ed., Prentice Hall, Upper Saddle River, NJ, 2008c.

Greene, W. "Discrete Choice Modeling." In *The Handbook of Econometrics:* Vol. 2, *Applied Econometrics*, T. Mills and K. Patterson, eds., Palgrave, London, 2009a.

Greene, W. "Models for Count Data with Endogenous Participation." *Empirical Economics*, 36, 1, 2009b, pp. 133–173.

Greene, W. "A Sample Selection Corrected Stochastic Frontier Model." *Journal of Productivity Analysis*, 34, 1, 2010a, pp. 15–24.

Greene, W. "Testing Hypotheses About Interaction Terms in Nonlinear Models.." *Economics Letters,* 107, 2010b, pp. 291–296.

Greene, W. "Panel Data Models for Discrete Choices." Chapter 15 in *Oxford Handbook of Panel Data*, B. Baltagi, e., 2015.

Greene, W. *LIMDEP. Version* 11, Econometric Software, Plainview, NY, 2016.

Greene, W., M. Harris, B. Hollingsworth, and P. Maitra. "A Bivariate Latent Class Correlated Generalized Ordered Probit Model with an Application to Modeling Observed Obesity Levels." Working paper EC-08-18, Stern School of Business, New York University, 2008.

Greene, W., and Hensher, D. "Specification and Estimation of Nested Logit Models." *Transportation Research,* B, 36, 1, pp. 1–18, 2002.

Greene, W., and D. Hensher. "Multinomial Logit and Discrete Choice Models." In W. Greene, *NLOGIT Version 4.0 User's Manual, Revised,* Plainview, NY: Econometric Software, Inc., 2007.

Greene, W., and D. Hensher. *Modeling Ordered Choices: A Primer*, Cambridge University Press, Cambridge, 2010a.

Greene, W., and D. Hensher. "Ordered Choices and Heterogeneity in Attribute Processing." *Journal of Transport Economics and Policy,* 44, 3, 2010b, pp. 331–364.

Greene, W., and C. McKenzie. "An LM Test for Random Effects Based on Generalized Residuals." *Economics Letters*, 127, 1, 2015, pp. 47–50.

Greene, W., and T. Seaks. "The Restricted Least Squares Estimator: A Pedagogical Note." *Review of Economics and Statistics*, 73, 1991, pp. 563–567.

Griffiths, W., C. Hill, and G. Judge. *Learning and Practicing Econometrics*. New York: John Wiley and Sons, 1993.

Griliches, Z. "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change." In *Price Statistics of the Federal Government,* prepared by the Price Statistics Review Committee of the National Bureau of Economic Research. New York: National Bureau of Economic Research, 1961.

Griliches, Z. "Economic Data Issues." In *Handbook of Econometrics,* Vol. 3, edited by Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1986.

Griliches, Z., and P. Rao. "Small Sample Properties of Several Two Stage Regression Methods in the Context of Autocorrelated Errors." *Journal of the American Statistical Association,* 64, 1969, pp. 253–272.

Grogger, J., and R. Carson. "Models for Truncated Counts." *Journal of Applied Econometrics,* 6, 1991, pp. 225–238.

Gronau, R. "Wage Comparisons: A Selectivity Bias." *Journal of Political Economy,* 82, 1974, pp. 1119–1149.

Groot, W., and H. Maassen van den Brink. "Match Specific Gains to Marriages: A Random Effects Ordered Response Model." *Quality and Quantity,* 37, 3, 2003, pp. 317–325.

Grootendorst, P. "A Review of Instrumental Variables Estimation of Treatment Effects in the Applied Health Sciences." *Health Services Outcomes Research Methods,* 7, 2007, pp. 159–179.

Grossman, M. "On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, 80, 2, 1972, pp. 223–255.

Grunfeld, Y. "The Determinants of Corporate Investment." Unpublished Ph.D. thesis, Department of Economics, University of Chicago, 1958.

Grunfeld, Y., and Z. Griliches. "Is Aggregation Necessarily Bad?" *Review of Economics and Statistics,* 42, 1960, pp. 1–13.

Guilkey, D. "Alternative Tests for a First-Order Vector Autoregressive Error Specification." *Journal of Econometrics,* 2, 1974, pp. 95–104.

Guilkey, D., and P. Schmidt. "Estimation of Seemingly Unrelated Regressions with Vector Autoregressive Errors." *Journal of the American Statistical Association,* 1973, pp. 642–647.

Gupta, K, N. Kristensen, and D. Possoli. "External Validation of the Use of Vignettes in Cross-Country Health Studies." Health Econometrics Workshop, Milan, Department of Economics, Aarhus School of Business, University of Aarhus, 2008.

Gurmu, S. "Tests for Detecting Overdispersion in the Positive Poisson Regression Model." *Journal of Business and Economic Statistics,* 9, 1991, pp. 215–222.

Gurmu, S., P. Rilstone, and S. Stern. "Semiparametric Estimation of Count Regression Models." *Journal of Econometrics,* 88, 1, 1999, pp. 123–150.

Gurmu, S., and P. Trivedi. "Recent Developments in Models of Event Counts: A Survey." Manuscript, Department of Economics, Indiana University, 1994.

Hadri, K., C. Guermat, and J. Whittaker. "Estimating Farm Efficiency in the Presence of Double Heteroscedasticity Using Panel Data." *Journal of Applied Economics,* 6, 2, 2003, pp. 255–268.

Hafner, C., H. Manner, and L. Simar. "The 'Wrong Skewness' Problem in Stochastic Frontier Models: A New Approach." *Econometric Reviews*, 2016, forthcoming.

Hahn, J. "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both *n* and *T* Are Large." *Econometrica*, 70, 2002, pp. 1639–1657.

Hahn, J., and J. Hausman. "A New Specification Test for the Validity of Instrumental Variables." *Econometrica,* 70, 2002, pp. 163–189.

Hahn, J., and J. Hausman. "Weak Instruments: Diagnosis and Cures in Empirical Econometrics." *American Economic Review,* 93, 2003, pp. 118–125.

Hahn, J., and G. Kuersteiner. "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects." Unpublished manuscript, Department of Economics, University of California, Los Angeles, 2004.

Hahn, J., and W. Newey. "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models." *Econometrica*, 72, 2004, pp. 1295–1313.

Hajivassiliou, V. "Smooth Simulation Estimation of Panel Data LDV Models." Department of Economics, Yale University, 1990.

Hajivassiliou, A. "Some Practical Issues in Maximum Simulated Likelihood." In *Simulation Based Inference in Econometrics*, edited by R. Mariano, T. Schuermann, and M. Weeks, Cambridge: Cambridge University Press, 2000.

Hall, B. *TSP Version 4.0 Reference Manual*. Palo Alto: TSP International, 1982.

Hall, B. "Software for the Computation of Tobit Model Estimates." *Journal of Econometrics,* 24, 1984, pp. 215–222.

Hall, R. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *Journal of Political Economy,* 86, 6, 1978, pp. 971–987.

Hamilton, J. *Time Series Analysis*. Princeton: Princeton University Press, 1994.

Hansen, B. "Challenges for Econometric Model Selection." *Econometric Theory,* 21, 2005, pp. 60–68.

Hansen, L. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica,* 50, 1982, pp. 1029–1054.

Hansen, L., J. Heaton, and A. Yaron. "Finite Sample Properties of Some Alternative GMM Estimators." *Journal of Business and Economic Statistics,* 14, 3, 1996, pp. 262–280.

Hansen, L., and K. Singleton. "Generalized Instrumental Variable Estimation of Nonlinear Rational Expectations Models." *Econometrica,* 50, 1982, pp. 1269–1286.

Hanushek, E. "Efficient Estimators for Regressing Regression Coefficients." *The American Statistician,* 28, 2, 1974, pp. 21–27.

Hanushek, E. "The Evidence on Class Size." In *Earning and Learning: How Schools Matter*, edited by S. Mayer and P. Peterson, Washington, DC: Brookings Institute Press, 1999. 6:37

Hanushek, E. *The Economics of Schooling and School Quality*. ed., Edward Elgar Publishing, 2002.

Hardle, W. *Applied Nonparametric Regression*. New York: Cambridge University Press, 1990.

Hardle, W., H. Liang, and J. Gao. *Partially Linear Models*. Springer-Verlag, Heidelberg, 2000.

Harris, M., and X. Zhao. "A Zero-Inflated Ordered Probit Model, with an Application to Modeling Tobacco Consumption." *Journal of Econometrics*, 141, 2, 2007, pp. 1073–1099.

Harvey, A. "Estimating Regression Models with Multiplicative Heteroscedasticity." *Econometrica,* 44, 1976, pp. 461–465.

Hausman, J. "Specification Tests in Econometrics." *Econometrica,* 46, 1978, pp. 1251–1271.

Hausman, J., B. Hall, and Z. Griliches. "Economic Models for Count Data with an Application to the Patents—R&D Relationship." *Econometrica,* 52, 1984, pp. 909–938.

Hausman, J., and A. Han. "Flexible Parametric Estimation of Duration and Competing Risk Models." *Journal of Applied Econometrics,* 5, 1990, pp. 1–28.

Hausman, J., and D. McFadden. "A Specification Test for the Multinomial Logit Model." *Econometrica,* 52, 1984, pp. 1219–1240.

Hausman, J., J. Stock, and M. Yogo. "Asymptotic Properties of the Hahn-Hausman Test for Weak Instruments." *Economics Letters,* 89, 2005, pp. 333–342.

Hausman, J., and W. Taylor. "Panel Data and Unobservable Individual Effects." *Econometrica,* 49, 1981, pp. 1377–1398.

Hausman, J., and D. Wise. "Social Experimentation, Truncated Distributions, and Efficient Estimation." *Econometrica,* 45, 1977, pp. 919–938.

Hausman, J., and D. Wise. "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences." *Econometrica,* 46, 1978, pp. 403–426.

Hausman, J. and D. Wise. "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment." *Econometrica,* 47, 2, 1979, pp. 455–573.

Hawcroft, L., and T. Milmont. "The Use (and Abuse) of the New Environmental Paradigm Scale over the Last 30 Years: A Meta-Analysis." *Journal of Environmental Psychology*, 30, 2, pp. 143–158.

Hayashi, F. *Econometrics*. Princeton: Princeton University Press, 2000.

Heckman, J. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement,* 5, 1976, pp. 475–492.

Heckman, J. "Simple Statistical Models for Discrete Panel Data Developed and Applied to the Hypothesis of True State Dependence Against the Hypothesis of Spurious State Dependence." *Annalse de l'INSEE,* 30, 1978, pp. 227–269.

Heckman, J. "Sample Selection Bias as a Specification Error." *Econometrica,* 47, 1979, pp. 153–161.

Heckman, J. "Statistical Models for Discrete Panel Data." In *Structural Analysis of Discrete Data with Econometric Applications,* edited by C. Manski and D. McFadden, Cambridge: MIT Press, 1981a.

Heckman, J. "Heterogeneity and State Dependence." In *Studies of Labor Markets*, edited by S. Rosen, NBER, Chicago: University of Chicago Press, 1981b.

Heckman, J. "Varieties of Selection Bias." *American Economic Review,* 80, 1990, pp. 313–318.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. "Characterizing Selection Bias Using Experimental Data." *Econometrica,* 66, 5, 1998, pp. 1017–1098.

Heckman, J., H. Ichimura, and P. Todd. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies,* 64, 4, 1997 pp. 605–654.

Heckman, J., H. Ichimura, and P. Todd. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies,* 65, 2, 1998, pp. 261–294.

Heckman, J., R. LaLonde, and J. Smith. "The Economics and Econometrics of Active Labour Market Programmes." In *The Handbook*

of Labor Economics, Vol. 3., edited by O. Ashenfelter and D. Card, Amsterdam: North Holland, 1999.

Heckman, J., and T. MaCurdy. "A Life Cycle Model of Female Labor Supply." *Review of Economic Studies,* 47, 1980, pp. 247–283.

Heckman, J., and B. Singer. "Econometric Duration Analysis." *Journal of Econometrics,* 24, 1984a, pp. 63–132.

Heckman, J., and B. Singer. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica,* 52, 1984b, pp. 271–320.

Heckman, J., J. Tobias, and E. Vytlacil. "Simple Estimators for Treatment Parameters in a Latent Variable Framework." *Review of Economics and Statistics,* 85, 3, 2003, pp. 748–755.

Heckman, J., and E. Vytlacil. "Instrumental Variables, Selection Models and Tight Bounds on the Average Treatment Effect." NBER Technical Working paper 0259, 2000.

Heckman, J., and E. Vytlacil. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics*, Chapter 70, edited by J. Heckman and E. E. Leamer, North Holland, Amsterdam, 2007.

Heilbron, D. "Generalized Linear Models for Altered Zero Probabilities and Overdispersion in Count Data." Technical Report, Department of Epidemiology and Biostatistics, University of California, San Francisco, 1989.

Henderson, D., and C. Parmeter. *Applied Nonparametric Econometrics*. Cambridge University Press, New York, 2015.

Hendry, D. "Monte Carlo Experimentation in Econometrics." In *Handbook of Econometrics,* Vol. 2, edited by Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1984.

Hendry, D., and N. Ericsson. "An Econometric Analysis of UK Money Demand." In M. Friedman and A. Schwartz, eds., *American Economic Review,* 81, 1991, pp. 8–38.

Hensher, D. "Dimensions of Automobile Demand–An Overview of an Australian Research Project." *Environment and Planning*, A., 18, 2010, pp. 1339–1374.

Hensher, D. "Efficient Estimation of Hierarchical Logit Mode Choice Models." *Journal of the Japanese Society of Civil Engineers,* 425/IV-14, 1991, pp. 117–128.

Hensher, D. "Sequential and Full Information Maximum Likelihood Estimation of a Nested Logit Model." *Review of Economics and Statistics,* 68, 4, 1986, pp. 657–667.

Hensher, D., and W. Greene. "The Mixed Logit Model: The State of Practice." *Transportation Research, B,* 30, 2003, pp. 133–176.

Hensher, D. and W. Greene. "Non-attendance and Dual Processing of Common-Metric Attributes in Choice Analysis: A Latent Class Specification." *Empirical Economics*, 39, 2010, pp. 413–426.

Hensher, D., and S. Jones. "Predicting Corporate Failure: Optimizing the Performance of the Mixed Logit Model." *ABACUS*, 43, 3, 2007, pp. 241–264.

Hensher, D., J. Louviere, and J. Swait. *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge University Press, 2000.

Hensher, D., J. Rose, and W. Greene. *Applied Choice Analysis.* 2nd ed., Cambridge: Cambridge University Press, 2015.

Hensher, S., Rose, J., and W. Greene. "The Implications of Willingness to Pay of Respondents Ignoring Specific Attributes." *Transportation Research Part E: Logistics and Transportation Review*, 32, 2005, pp. 203–222,

Hensher D., J. Rose, and W. Greene. "Inferring Attribute Non-attendance from Stated Choice Data: Implication for Willingness to Pay Estimates and a Warning for Stated Choice Experiment Design." *Transportation Journal*, 39, 2012, pp. 235–245.

Hess, S., and D. Hensher. "Making Use of Respondent Reported Processing Information to Understand Attribute Importance: A Latent Variable Scaling Approach." *Transportation Journal*, 40, 2, 2013, pp. 397–412.

Hilbe, J. *Negative Binomial Regression*. Cambridge University, Cambridge, 2007.

Hildebrand, G., and T. Liu. *Manufacturing Production Functions in the United States*. Ithaca, NY: Cornell University Press, 1957.

Hildreth, C., and C. Houck. "Some Estimators for a Linear Model with Random Coefficients." *Journal of the American Statistical Association,* 63, 1968, pp. 584–595.

Hill, C., and L. Adkins. "Collinearity." In *A Companion to Theoretical Econometrics*, edited by B. Baltagi, Oxford: Blackwell, 2001.

Hilts, J. "Europeans Perform Highest in Ranking of World Health." *New York Times,* June 21, 2000.

Hirano, K., G. Imbens, and G. Ridder. "Efficient Estimation of Average Treatment Effects Using Estimated Propensity Scores." *Econometrica*, 71, 2003, 1161–1189.

Hodge, A and S. Shankar. "Partial Effects in Ordered Response Models with Factor Variables." *Econometric Reviews*, 33, 8, 2014, pp. 854–868.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky. "Bayesian Model Averaging: A Tutorial." *Statistical Science,* 14, 1999, pp. 382–417.

Hole, A. "A Comparison of Approaches to Estimating Confidence Intervals for Willingness to Pay Measures." Paper CHE 8, Center for Health Economics, University of York, 2006.

Hole, A. "A Discrete Choice Model with Endogenous Attribute Attendance." *Economics Letters*, 110, 3, 2011, pp. 203–205.

Hollingshead, A. B. *Four Factor Index of Social Status*, unpublished manuscript, Department of Sociology, Yale University, New Haven, CT, 1975.

Hollingsworth, J., and B. Wildman. "The Efficiency of Health Production: Re-estimating the WHO Panel Data Using Parametric and Nonparametric Approaches to Provide Additional Information." *Health Economics* 11, 2002, pp. 1–11.

Holt, M. "Autocorrelation Specification in Singular Equation Systems: A Further Look." *Economics Letters,* 58, 1998, pp. 135–141.

Holtz-Eakin, D. "Testing for Individual Effects in Autoregressive Models." *Journal of Econometrics,* 39, 1988, pp. 297–307.

Holtz-Eakin, D., W. Newey, and H. Rosen. "Estimating Vector Autoregressions with Panel Data." *Econometrica,* 56, 6, 1988, pp. 1371–1395.

Hong, H., B. Preston, and M. Shum. "Generalized Empirical Likelihood Based Model Selection Criteria for Moment Condition Models." *Econometric Theory,* 19, 2003, pp. 923–943.

Hombrook, M. "Was David Li the Guy Who Blew Up Wall Street?" *CBC News Canada*, April 8, 2009, www.cbc.ca/news/canada/was-david-li-the-guy-who-blew-up-wall-street-1.775372.

Honorè, B., and E. Kyriazidou. "Estimation of a Panel Data Sample Selection Model." *Econometrica,* 65, 6, 1997, pp. 1335–1364.

Honorè, B., and E. Kyriazidou. "Panel Data Discrete Choice Models with Lagged Dependent Variables." *Econometrica,* 68, 4, 2000, pp. 839–874.

Horn, D., A. Horn, and G. Duncan. "Estimating Heteroscedastic Variances in Linear Models." *Journal of the American Statistical Association,* 70, 1975, pp. 380–385.

Horowitz, J. "A Smoothed Maximum Score Estimator for the Binary Response Model." *Econometrica,* 60, 1992, pp. 505–531.

Horowitz, J. "Semiparametric Estimation of a Work-Trip Mode Choice Model." *Journal of Econometrics,* 58, 1993, pp. 49–70.

Horowitz, J. "The Bootstrap." In *Handbook of Econometrics,* Vol. 5, edited by J. Heckman and E. Leamer, Amsterdam: North Holland, 2001, pp. 3159–3228.

Horowitz, J., and G. Neumann. "Specification Testing in Censored Regression Models." *Journal of Applied Econometrics,* 4(S), 1989, pp. S35–S60.

Hoxby, C. "Does Competition Among Public Schools Benefit Students and Taxpayers?" *American Economic Review,* 69, 5, 2000, pp. 1209–1238.

Hsiao, C. "Some Estimation Methods for a Random Coefficient Model." *Econometrica,* 43, 1975, pp. 305–325.

Hsiao, C. *Analysis of Panel Data.* Cambridge: Cambridge University Press, 1986.

Hsiao, C. *Analysis of Panel Data.* 2nd ed., New York: Cambridge University Press, 2003.

Hsiao, C., K. Lahiri, L. Lee, and H. Pesaran. *Analysis of Panels and Limited Dependent Variable Models*. New York: Cambridge University Press, 1999.

Hsiao, C., M. Pesaran, and A. Tahmiscioglu. "A Panel Analysis of Liquidity Constraints and Firm Investment." In *Analysis of Panels and Limited Dependent Variable Models*, edited by C. Hsiao, K. Lahiri, L. Lee, and M. Pesaran, Cambridge: Cambridge University Press, 2002, pp. 268–296.

Huang, R. "Estimation of Technical Inefficiencies with Heterogeneous Technologies."

Journal of Productivity Analysis, 21, 2003, pp. 277–296.

Huber, P. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions." In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics,* Vol. 1. Berkeley: University of California Press, 1967.

Huber, P. *Robust Statistical Procedures*. Washington, DC: National Science Foundation, 1987.

Hurd, M. "Estimation in Truncated Samples When There Is Heteroscedasticity." *Journal of Econometrics,* 11, 1979, pp. 247–258.

Hyslop, D. "State Dependence, Serial Correlation, and Heterogeneity in Labor Force Participation of Married Women." *Econometrica,* 67, 6, 1999, pp. 1255–1294.

Im, K., M. Pesaran, and Y. Shin. "Testing for Unit Roots in Heterogeneous Panels." *Journal of Econometrics,* 115, 2003, pp. 53–74.

Imbens, G. "Generalized Method of Moments and Empirical Likelihood." *Journal of Business and Economic Statistics*, 20, 2002, pp. 493–506.

Imbens, G., and J. Angrist. "Identification and Estimation of Local Average Treatment Effects." *Econometrica,* 62, 1994, pp. 467–476.

Imbens, G., and D. Hyslop. "Bias from Classical and Other Forms of Measurement Error." *Journal of Business and Economic Statistics,* 19, 2001, pp. 141–149.

Imbens, G., and J. Wooldridge. "What's New in Econometrics, Part 2: Linear Panel Data Models." NBER Econometrics Summer Institute, 2007a.

Imbens, G., and J. Wooldridge. "What's New in Econometrics, Part 4: Nonlinear Panel Data Models." NBER Econometrics Summer Institute, 2007b.

Imbens, G., and J. Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, 47, 1, 2009, pp. 5–86.

Imhof, J. "Computing the Distribution of Quadratic Forms in Normal Variables." *Biometrika,* 48, 1980, pp. 419–426.

Inkmann, J. "Misspecified Heteroscedasticity in the Panel Probit Model: A Small Sample Comparison of GMM and SML Estimators." *Journal of Econometrics,* 97, 2, 2000, pp. 227–259.

Isacsson, Gunnar. "Estimates of the Return to Schooling in Sweden from a Large Sample of Twins." *Labour Economics*, 6, 4, 1999, pp. 471–489.

Jacob, B., and S. Levitt. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." Natural Bureau of Economic Research, Working paper 9413, NBER, Cambridge, MA, 2002.

Jacob, B., and S. Levitt. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 118, 3, 2003, pp. 843–877.

Jain, D., N. Vilcassim, and P. Chintagunta. "A Random-Coefficients Logit Brand Choice Model Applied to Panel Data." *Journal of Business and Economic Statistics,* 12, 3, 1994, pp. 317–328.

Jarque, C. "An Application of LDV Models to Household Expenditure Analysis in Mexico." *Journal of Econometrics,* 36, 1987, pp. 31–54.

Jakubson, G. "The Sensitivity of Labor Supply Parameters to Unobserved Individual Effects: Fixed and Random Effects Estimates in a Nonlinear Model Using Panel Data." *Journal of Labor Economics,* 6, 1988, pp. 302–329.

Jensen, M. "A Monte Carlo Study on Two Methods of Calculating the MLE's Covariance Matrix in a Seemingly Unrelated Nonlinear Regression." *Econometric Reviews,* 14, 1995, pp. 315–330.

Johansen, S. "Statistical Analysis of Cointegration Vectors." *Journal of Economic Dynamics and Control,* 12, 1988, pp. 231–254.

Jobson, J., and W. Fuller. "Least Squares Estimation When the Covariance Matrix and Parameter Vector Are Functionally Related." *Journal of the American Statistical Association,* 75, 1980, pp. 176–181.

Johansen, S. "Estimation and Hypothesis Testing of Cointegrated Vectors in Gaussian VAR Models." *Econometrica,* 59, 6, 1991, pp. 1551–1580.

Johansen, S. "A Representation of Vector Autoregressive Processes of Order 2." *Econometric Theo*ry, 8, 1992, pp. 188–202.

Johnson, V., and J. Albert. *Ordinal Data Modeling*. New York, Springer Verlag, 1999.

Johansen, S., and K. Juselius. "Maximum Likelihood Estimation and Inference on Cointegration, with Applications for the Demand for Money." *Oxford Bulletin of Economics and Statistics,* 52, 1990, pp. 169–210.

Johnson, N., and S. Kotz. *Distributions in Statistics—Continuous Multivariate Distributions*. New York: John Wiley and Sons, 1974.

Johnson, N., S. Kotz, and A. Kemp. *Distributions in Statistics—Univariate Discrete Distributions.* 2nd ed., New York: John Wiley and Sons, 1993.

Johnson, N., S. Kotz, and A. Balakrishnan. *Distributions in Statistics, Continuous Univariate Distributions—Vol. 1.* 2nd ed., New York: John Wiley and Sons, 1994.

Johnson, N., S. Kotz, and N. Balakrishnan. *Distributions in Statistics, Continuous Univariate Distributions—Vol. 2.* 2nd ed., New York: John Wiley and Sons, 1995.

Johnson, N., S. Kotz, and N. Balakrishnan. *Distributions in Statistics, Discrete Multivariate Distributions.* New York: John Wiley and Sons, 1997.

Johnson, R., and D. Wichern. *Applied Multivariate Statistical Analysis.* 5th ed., Englewood Cliffs, NJ: Prentice Hall, 2005.

Johnson, V., and J. Albert. *Ordinal Data Modeling*, Springer-Verlag, New York, 1999.

Johnston, J. *Econometric Methods*. New York: McGraw-Hill, 1984.

Johnston, J., and J. DiNardo. *Econometric Methods.* 4th ed., New York: McGraw-Hill, 1997.

Jondrow, J., K. Lovell, I. Materov, and P. Schmidt. "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model." *Journal of Econometrics,* 19, 1982, pp. 233–238.

Jones, A. "A Double Hurdle Model of Cigarette Consumption." *Journal of Applied Econometrics*, 4, 1, 1989, pp. 23–39.

Jones, A. *Applied Econometrics for Health Economists: A Practical Guide*. 2nd ed., Taylor and Francis, London, 2007)

Jones, A., X. Koolman, and N. Rice. "Health Related Non-response in the BHPS and ECHP: Using Inverse Probability Weighted Estimators in Nonlinear Models." *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 169, 2006, pp. 543–569.

Jones, J., and J. Landwehr. "Removing Heterogeneity Bias from Logit Model Estimation." *Marketing Science,* 7, 1, 1988, pp. 41–59.

Jones, A., J. Lomas and N. Rice. "Applying Beta-type Size Distributions to Healthcare Cost Regressions." *Journal of Applied Econometrics*, 29, 4, 2014, pp. 649–670.

Jones, A., J. Lomas and N. Rice. "Healthcare Cost Regressions: Going Beyond the Mean to Estimate the Full Distribution." *Health Economics*, 24, 9, 2015, pp. 1192–1212.

Jones, A., and N. Rice. "Econometric Evaluation of Health Policies." In *The Oxford Handbook of Health Economics*, S. Glied and P. Smith, eds., Oxford: Oxford University Press, 2011.

Jones, A., and S. Schurer. "How Does Heterogeneity Shape the Socioeconomic Gradient in Health Satisfaction." *Journal of Applied Econometrics*, 26, 3, April/May 2011.

Jones, S. "The Formula that Felled Wall Street." *Financial Times Magazine*, 4/24/09

Joreskog, K., and G. Gruvaeus. "A Computer Program for Minimizing a Function of Several Variables." Educational Testing Services, Research bulletin no. 70-14, 1970.

Joreskog, K., and D. Sorbom. *LISREL V User's Guide*. Chicago: National Educational Resources, 1981.

Judd, K. *Numerical Methods in Economics*. Cambridge: MIT Press, 1998.

Judge, G., C. Hill, W. Griffiths, and T. Lee. *The Theory and Practice of Econometrics*. New York: John Wiley and Sons, 1985.

Just, R., and R. Pope. "Stochastic Specification of Production Functions and Economic Implications." *Journal of Econometrics* 7, 1, 1978, pp. 67–86.

Just, R. E., and R. D. Pope. "Production Function Estimation and Related Risk Considerations." *American Journal of Agricultural Economics*, 61, 1979, pp. 276–84.

Kalbfleisch, J., and R. Prentice. *The Statistical Analysis of Failure Time Data.,* 2nd ed., New York: John Wiley and Sons, 2002.

Kamlich, R., and S. Polachek. "Discrimination: Fact or Fiction? An Examination Using an Alternative Approach." *Southern Economic Journal,* October 1982, pp. 450–461.

Kalbfleisch, J., and D. Sprott. "Application of Likelihood Methods to Models Involving

Large Numbers of Parameters." *Journal of the Royal Statistical Society, Series B*, 32, 2, 1970, pp. 175–208.

Kao, C. "Spurious Regression and Residual Based Tests for Cointegration in Panel Data." *Journal of Econometrics,* 90, 1999, pp. 1–44.

Kaplan, E., and P. Meier. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association,* 53, 1958, pp. 457–481.

Kapteyn, A., J. Smith, and A. van Soest. "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands." *American Economic Review*, 97, 1, 2007, pp. 461–473.

Kasteridis, P., M. Munkin, and S. Yen. "Demand for Cigarettes: A Mixed Binary-Ordered Probit Approach." *Applied Economics*, 42, 4, 2010, pp. 413–426.

Katz, E. "Bias in Conditional and Unconditional Fixed Effects Logit Estimation." *Political Analysis*, 9, 2001, pp. 379–384.

Kaufman, A. "The Influence of Fannie and Freddie on Mortgage Loan Terms." *Real Estate Economics*, 42, 2, 2014, pp. 472–496.

Kay, R., and S. Little. "Assessing the Fit of the Logistic Model: A Case Study of Children with Haemolytic Uraemic Syndrome." *Applied Statistics,* 35, 1986, pp. 16–30.

Keane, M. "Simulation Estimators for Panel Data Models with Limited Dependent Variables." In, *Handbook of Statistics,* Volume 11, Chapter 20, edited by G. Maddala and C. Rao, Amsterdam: North Holland, 1993.

Keane, M. "A Computationally Practical Simulation Estimator for Panel Data." *Econometrica*, 62, 1, 1994, pp. 95–116.

Keane, M. "A Structural Perspective on the Experimentalist School." *Journal of Economic Perspectives*, 24, 2, 2010, pp. 47–58.

Keele, L., and D. Park. "Difficult Choices: An Evaluation of Heterogeneous Choice Models." presented at the 2004 Meeting of the American Political Science Association, Department of Politics and International Relations, Oxford University, manuscript, 2005.

Kelejian, H., and I. Prucha. "A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model." *International Economic Review,* 40, 1999, pp. 509–533.

Kennan, J. "The Duration of Contract Strikes in U.S. Manufacturing." *Journal of Econometrics,* 28, 1985, pp. 5–28.

Kennedy, W., and J. Gentle. *Statistical Computing*. New York: Marcel Dekker, 1980.

Kerkhofs, M., and M. Lindeboom. "Subjective Health Measures and State Dependent Reporting Errors." *Health Economics* 4, 1995, pp. 221–235.

Keuzenkamp, H., and J. Magnus. "The Significance of Testing in Econometrics." *Journal of Econometrics*, 67, 1, 1995, pp. 1–257.

Keynes, J. *The General Theory of Employment, Interest, and Money*. New York: Harcourt, Brace, and Jovanovich, 1936.

Kezde, G. "Robust Standard Error Estimation in Fixed-Effects Panel Models." Working paper, Department of Economics, Michigan State University, 2001.

Khan, S. "Distribution Free Estimation of Heteroskedastic Binary Choice Models Using Probit Criterion Functions", *Journal of Econometrics*, 172, 2013, pp. 168–182.

Kiefer, N. "Testing for Independence in Multivariate Probit Models." *Biometrika,* 69, 1982, pp. 161–166.

Kiefer, N., ed. "Econometric Analysis of Duration Data." *Journal of Econometrics,* 28, 1, 1985, pp. 1–169.

Kiefer, N. "Economic Duration Data and Hazard Functions." *Journal of Economic Literature,* 26, 1988, pp. 646–679.

King, G., C. J. Murray, J. A. Salomon, and A. Tandon. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review*, 98, 2004, pp. 191–207, gking.harvard.edu/files/abs/vign-abs.shtml.

Kingdon, G., and R. Cassen. "Explaining Low Achievement at Age 16 in England." Mimeo, Department of Economics, University of Oxford, 2007.

Kitchin, B. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data and Society*, April–June, 2014, pp. 1–12.

Kiviet, J. "On Bias, Inconsistency, and Efficiency of Some Estimators in Dynamic Panel Data Models." *Journal of Econometrics,* 68, 1, 1995, pp. 63–78.

Kiviet, J., G. Phillips, and B. Schipp. "The Bias of OLS, GLS and ZEF Estimators in Dynamic

SUR Models." *Journal of Econometrics,* 69, 1995, pp. 241–266.

Kleiber, C., and A. Zeileis. "The Grunfeld Data at 50." *German Economic Review*, 11, 4, 2010, pp. 403–546.

Kleibergen, F. "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression." *Econometrica*, 70, 2002, pp. 1781–1803.

Klein, L. *Economic Fluctuations in the United States 1921–1941*. New York: John Wiley and Sons, 1950.

Klein, R., and R. Spady. "An Efficient Semiparametric Estimator for Discrete Choice Models." *Econometrica,* 61, 1993, pp. 387–421.

Klier, T., and D. McMillen. "Clustering of Auto Supplier Plants in the United States." Journal of Business and Economic Statistics, 26. 4, 2008, pp. 460–471.

Klugman, S., and R. Parsa. "Fitting Bivariate Loss Distributions with Copulas." *Insurance: Mathematics and Economics,* 24, 2000, pp. 139–148.

Kmenta, J. *Elements of Econometrics*. New York: Macmillan, 1986.

Knapp, L., and T. Seaks. "An Analysis of the Probability of Default on Federally Guaranteed Student Loans." *Review of Economics and Statistics,* 74, 1992, pp. 404–411.

Knight, F. *The Economic Organization*. New York: Harper and Row, 1933.

Knuth, D. E. *The Art of Computer Programming*, Vol. 1, *Fundamental Algorithms*. Boston: Addison-Wesley, 1997.

Kodde, D. A., and Palm, F. C. "Wald Criteria for Jointly Testing Equality and Inequality Restrictions." *Econometrica,* 54, 5, 1986, pp. 1243–1248.

Koenker, R. "A Note on Studentizing a Test for Heteroscedasticity." *Journal of Econometrics,* 17, 1981, pp. 107–112.

Koenker, R. *Quantile Regression*, Econometric Society Monographs, Cambridge University Press, Cambridge, 2005.

Koenker, R., and G. Bassett. "Regression Quantiles." *Econometrica,* 46, 1978, pp. 107–112.

Koenker, R., and G. Bassett. "Robust Tests for Heteroscedasticity Based on Regression Quantiles." *Econometrica,* 50, 1982, pp. 43–61.

Koenker, R., and V. D'Orey. "Algorithm AS229: Computing Regresson Quantiles." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 36, 3, 1987, pp. 383–393.

Koenker, R., and K. Hallock. "Quantile Regression." *Journal of Economic Perspectives,* 15, 4, 2001, pp. 143–156.

Koop, G. *Bayesian Econometrics*. New York: John Wiley and Sons, 2003.

Koop, G., and S. Potter. "Forecasting in Large Macroeconomic Panels Using Bayesian Model Averaging." *Econometrics Journal,* 7, 2, 2004, pp. 161–185.

Koop, G., and J. Tobias. "Learning About Heterogeneity in Returns to Schooling." *Journal of Applied Econometrics,* 19, 7, 2004, pp. 827–849.

Kotz, S., N. Balakrishnan, and N. Johnson. *Continuous Multivariate Distributions, Volume 1, Models and Applications.* 2nd ed., New York, John Wiley and Sons, 2000.

Krailo, M., and M. Pike. "Conditional Multivariate Logistic Analysis of Stratified Case-Control Studies." *Applied Statistics,* 44, 1, 1984, pp. 95–103.

Krinsky, I., and L. Robb. "On Approximating the Statistical Properties of Elasticities." *Review of Economics and Statistics,* 68, 4, 1986, pp. 715–719.

Krinsky, I., and L. Robb. "On Approximating the Statistical Properties of Elasticities: Correction." *Review of Economics and Statistics,* 72, 1, 1990, pp. 189–190.

Krinsky, I., and L. Robb. "Three Methods for Calculating Statistical Properties for Elasticities." *Empirical Economics,* 16, 1991, pp. 1–11.

Kristensen, N., and E. Johansson. "New Evidence on Cross Country Differences in Job Satisfaction Using Anchoring Vignettes, *Labor Economics*, 15, 2008, pp. 96–117.

Krueger, A. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics*, 114, 2, 1999, pp. 497–532.

Krueger, A. "Economic Scene." *New York Times,* April 27, 2000, p. C2.

Kreuger, A., and S. Dale. "Estimating the Payoff to Attending a More Selective College." NBER, Cambridge, Working paper 7322, 1999.

Kruskal, W. "When are Gauss-Markov and Least Squares Estimators Identical." *Annals of Mathematical Statistics*, 39, 1968, 70–75.

Kumbhakar, S. "Efficiency Estimation with Heteroscedasticity in a Panel Data Model." *Applied Economics*, 29, 1997a, pp. 379–386.

Kumbhakar, S. "Modeling Allocative Inefficiency in a Translog Cost Function and Cost Share Equations: An Exact Relationship." *Journal of Econometrics*, 76, 1997b, pp. 351–356.

Kumbhakar, S., and K. Lovell. *Stochastic Frontier Analysis*. New York: Cambridge University Press, 2000.

Kumbhakar, S., and L. Orea. "Efficiency Measurement Using a Latent Class Stochastic Frontier Model, *Empirical Economics*, 29, 2004, pp. 169–183.

Kumbhakar, S., and C. Parmeter. "Efficiency Analysis: A Primer on Recent Advances." *Foundations and Trends in Econometrics*, 7, 2014, pp. 191–385.

Kumbhakar, S., L. Simar, T. Park, and E. Tsionas. "Nonparametric Stochastic Frontiers: A Local Maximum Likelihood Approach." *Journal of Econometrics*, 137, 2007, pp. 1–27.

Kwiatkowski, D., P. Phillips, P. Schmidt, and Y. Shin. "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root." *Journal of Econometrics,* 54, 1992, pp. 159–178.

Kyriazidou, E. "Estimation of a Panel Data Sample Selection Model." *Econometrica,* 65, 1997, pp. 1335–1364.

Kyriazidou, E. "Estimation of Dynamic Panel Data Sample Selection Models." *Review of Economic Studies,* 68, 2001, pp. 543–572.

L'Ecuyer, P. "Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators." Working paper, Department of Information Science, University of Montreal, 1998.

Lagarde, M. "Investigating Attribute Non-Attendance and Its Consequences in Choice Experiments with Latent Class Models." *Health Economics*, 22, 2013, pp. 554–567.

Lahart, J. "New Light on the Plight of Winter Babies." *Wall Street Journal*, September 22, 2009.

LaLonde, R. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American* Economic *Review,* 76, 4, 1986, pp. 604–620.

Lambert, D. "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics,* 34, 1, 1992, pp. 1–14.

Lancaster, T. *The Analysis of Transition Data*. New York: Cambridge University Press, 1990.

Lancaster, T. "The Incidental Parameters Problem since 1948." *Journal of Econometrics,* 95, 2, 2000, pp. 391–414.

Lancaster, T. *An Introduction to Modern Bayesian Inference*. Oxford: Oxford University Press, 2004.

Laporte, A., A. Karimova, and B. Ferguson. "Quantile Regression Analysis of the Rational Addiction Model: Investigating Heterogeneity in Forward-Looking Behavior." *Health Economics*, 19, 9, 2010, pp. 1063–1074.

Lawless, J. *Statistical Models and Methods for Lifetime Data.* New York: John Wiley and Sons, 1982.

Leamer, E. "A Bayesian Interpretation of Pretesting." *Journal of the Royal Statistical Society, Series B*, 38, 1, 1976, pp. 85–94.

Leamer, E. *Specification Searches: Ad Hoc Inferences with Nonexperimental Data*. New York: John Wiley and Sons, 1978.

Leamer, E. "Model Choice and Specification Analysis." In *Handbook of Econometrics,* Vol. I, S. Griliches and M. Intriligator, Amsterdam, North Holland, 1983.

Leamer, E. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives*, 24, 2, 2010, pp. 31–46.

LeCam, L. "On Some Asymptotic Properties of Maximum Likelihood Estimators and Related Bayes Estimators." *University of California Publications in Statistics,* 1, 1953, pp. 277–330.

Lechner, M. "The Estimation of Causal Effects by Difference-in-Difference Methods." *Foundations and Trends in Econometrics*, 4, 3, 2011, pp. 165–224.

Lee, K., M. Pesaran, and R. Smith. "Growth and Convergence in a Multi-country Empiridal Stochastic Solow Model." *Journal of Applied Econometrics,* 12, 1997, pp. 357–392.

Lee, L. "Generalized Econometric Models with Selectivity." *Econometrica,* 51, 1983, pp. 507–512.

Lee, L. "Specification Tests for Poisson Regression Models." *International Economic Review,* 27, 1986, pp. 689–706.

Lee, M. *Method of Moments and Semiparametric Econometrics for Limited Dependent Variables*. New York: Springer-Verlag, 1996.

Lee, M. *Limited Dependent Variable Models*. New York: Cambridge University Press, 1998.

Lee, J., and K. Seo. "A Computationally Fast Estimator for Random Coefficients Logit Demand Models Using Aggregate Data." *Rand Journal of Economics*, 46, 1, 2015, pp. 86–102.

Lee, S. "Formula from Hell." *Forbes Magazine*, May 8, 2009.

Leff, N. "Dependency Rates and Savings Rates." *American Economic Review,* 59, 5, 1969, pp. 886–896.

Leff, N. "Dependency Rates and Savings Rates: Reply." *American Economic Review,* 63, 1, 1973, p. 234.

Lemke, R. M. Leonard, and K. Tlhokwand. "Estimating Attendance at Major League Baseball Games for the 2007 Season." *Journal of Sports Economics*, August, 2009, pp. 875–886.

Lerman, R., and C. Manski. "On the Use of Simulated Frequencies to Approximate Choice Probabilities." In *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. Manski and D. McFadden, Cambridge: MIT Press, 1981.

LeSage, J. *Introduction to Spatial Econometrics*, Chapman and Hall/CRC Press, Boca Raton, FL, 2009.

Levi, M. "Errors in the Variables in the Presence of Correctly Measured Variables." *Econometrica,* 41, 1973, pp. 985–986.

Levin, A., and C. Lin. "Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties." Discussion paper 92-93, Department of Economics, University of California, San Diego, 1992.

Lewbel, A. "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables." *Journal of Econometrics,* 97, 1, 2000, pp. 145–177.

Lewbel, A. "An Overview of the Special Regressor Method." In *Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, edited by A. Ullah, J. Racine and L. Su, Oxford University Press, 2014, pp. 38–62.

Lewbel, A. "A Simple Estimator for Binary Choice Models with Endogenous Regressors." *Econometric Reviews*, 34, 2015, pp. 82–105.

Lewbel, A., Y. Dong, and T Yang. "Comparing Features of Convenient Estimators for Binary Choice Models with Endogenous Regressors." *Canadian Journal of Economics*, 45, 3, 2012, pp. 809–829.

Lewis, H. "Comments on Selectivity Biases in Wage Comparisons." *Journal of Political Economy,* 82, 1974, pp. 1149–1155.

Li, M., and J. Tobias. "Calculus Attainment and Grades Received in Intermediate Economic Theory." *Journal of Applied Economics,* 21, 9, 2006, pp. 893–896.

Li, Q., and J. Racine. *Nonparametric Econometrics*. Princeton: Princeton University Press, 2007.

Li, W., S. Ling, and M. McAleer. *A Survey of Recent Theoretical Results for Time Series Models with GARCH Errors*. Manuscript, Institute for Social and Economic Research, Osaka University, Osaka, 2001.

Liang, K., and S. Zeger. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika,* 73, 1986, pp. 13–22.

Li, D. "On Default Correlation: A Copula Approach." *Journal of Fixed Income*, 9, 4, 2000, 43–54.

Li, D. "On Default Correlation: A Copula Approach." Working paper 99-07, RiskMetrics, New York, 1999. (www.msci.com/resources/research/working_papers/defcorr.pdf).

Lindeboom, M., and E. van Doorslaer. "Cut Point Shift and Index Shift in Self-reported Health." *Equity III Project*, working paper #2, 2003.

Litman, B. "Predicting Success of Theatrical Movies: An Empirical Study." *Journal of Popular Culture*, 16, 4, 1983, pp. 159–175.

Little, R., and D. Rubin. *Statistical Analysis with Missing Data.* New York: Wiley, 1987.

Little, R., and D. Rubin, *Statistical Analysis of Missing Data*. Hoboken, NJ: John Wiley and Sons, 2002.

Loeve, M., *Probability Theory*. New York: Springer-Verlag, 1977.

Long, S. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications, 1997.

Long, S., and J. Freese, *Regression Models for Categorical Dependent Variables Using Stata*. College Station, TX: Stata Press, 2006.

Longley, J. "An Appraisal of Least Squares Programs from the Point of the User." *Journal of the American Statistical Association,* 62, 1967, pp. 819–841.

Loudermilk, M. "Estimation of Fractional Dependent Variables in Dynamic Panel Data Models with an Application to Firm Dividend Policy." *Journal of Business and Economic Statistics*, 25, 4, 2007, pp. 462–472.

Louviere, J., and J. Swait. "Discussion of Alleviating the Constant Stochastic Variance Assumption in Decision Research: Theory, Measurement and Experimental Test." *Marketing Science*, 29, 1, 2010, pp. 18–22.

Lovell, M. "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis." *Journal of the American Statistical Association,* 58, 1963, pp. 993–1010.

Lucas, R. "On the Mechanics of Economic Development." *Journal of Monetary Economics*, 22, 1988, pp. 3–42.

Machin, S., and A. Vignoles, *What's the Good of Education? The Economics of Education in the UK*. Princeton: Princeton University Press, 2005.

MacKinnon, J. "Bootstrap Inference in Econometrics." *Canadian Journal of Economics*, 35, 2002, pp. 615–645.

MacKinnon, J., and H. White. "Some Heteroscedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics*, 19, 1985, pp. 305–325.

Maddala, G. "The Use of Variance Components Models in Pooling Cross Section and Time Series Data." *Econometrica,* 39, 1971, pp. 341–358.

Maddala, G., *Econometrics*. New York, McGraw Hill, 1977.

Maddala, G. *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press, 1983.

Maddala, G. "Limited Dependent Variable Models Using Panel Data." *Journal of Human Resources,* 22, 1987, pp. 307–338.

Maddala, G. *Introduction to Econometrics,* 2nd ed. New York: Macmillan, 1992.

Maddala, G., and T. Mount. "A Comparative Study of Alternative Estimators for Variance Components Models." *Journal of the American Statistical Association,* 68, 1973, pp. 324–328.

Maddala, G., and F. Nelson. "Specification Errors in Limited Dependent Variable Models." Working paper 96, National Bureau of Economic Research, Cambridge, MA, 1975.

Maddala, G., and A. Rao. "Tests for Serial Correlation in Regression Models with Lagged Dependent Variables and Serially Correlated Errors." *Econometrica,* 41, 1973, pp. 761–774.

Maddala, G., and S. Wu. "A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test." *Oxford Bulletin of Economics and Statistics,* 61, 1999, pp. 631–652.

Magee, L., J. Burbidge, and L. Robb. "The Correlation Between Husband's and Wife's Education: Canada, 1971–1996." Social and Economic Dimensions of an Aging Population research papers, 24, McMaster University, 2000.

Magnac, T. "State Dependence and Heterogeneity in Youth Unemployment Histories." Working paper, INRA and CREST, Paris, 1997.

Magnus, J., and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* 3rd ed., New York: John Wiley and Sons, 2007.

Malinvaud, E. *Statistical Methods of Econometrics*. Amsterdam: North Holland, 1970.

Mandy, D., and C. Martins-Filho. "Seemingly Unrelated Regressions Under Additive Heteroscedasticity: Theory and Share Equation Applications." *Journal of Econometrics,* 58, 1993, pp. 315–346.

Mankiw, G. "A Letter to Ben Bernanke." *The American Economic Review*, 96, 2, 2006, pp. 182–184.

Mann, H., and A. Wald. "On the Statistical Treatment of Linear Stochastic Difference Equations." *Econometrica,* 11, 1943, pp. 173–220.

Manski, C. "The Maximum Score Estimator of the Stochastic Utility Model of Choice." *Journal of Econometrics,* 3, 1975, pp. 205–228.

Manski, C. "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator." *Journal of Econometrics,* 27, 1985, pp. 313–333.

Manski, C. "Operational Characteristics of the Maximum Score Estimator." *Journal of Econometrics,* 32, 1986, pp. 85–100.

Manski, C. "Semiparametric Analysis of the Random Effects Linear Model from Binary Response Data." *Econometrica,* 55, 1987, pp. 357–362.

Manski, C. "Anatomy of the Selection Problem." *Journal of Human Resources,* 24, 1989, pp. 343–360.

Manski, C. "Nonparametric Bounds on Treatment Effects." *American Economic Review,* 80, 1990, pp. 319–323.

Manski, C. *Analog Estimation Methods in Econometrics*. London: Chapman and Hall, 1992.

Manski, C., and S. Lerman. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica,* 45, 1977, pp. 1977–1988.

Manski, C., and D.McFadden. "Structural Analysis of Discrete Data and Econometric Applications." Cambridge, MIT Press, 1981.

Manzan, S., and D. Zerom. "A Semiparametric Analysis of Gasoline Demand in the United States Re-examining the Impact of Price." *Econometric Reviews,* 29, 4, 2010, pp. 439–468.

Marcus, A., and W. Greene. "The Determinants of Rating Assignment and Performance." Working paper CRC528, Center for Naval Analyses, 1985.

Marsaglia, G., and T. Bray. "A Convenient Method of Generating Normal Variables." *SIAM Review,* 6, 1964, pp. 260–264.

Martínez-Espinera, R., and J. Amoako-Tuffour. "Recreation Demand Analysis under Truncation, Overdispersion and Endogenous Stratification: An Application to Gros Morne National Park." *Journal of Environmental Management,* 88, 2008, pp. 1320–1332.

Martins, M. "Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal." *Journal of Applied Econometrics,* 16, 1, 2001, pp. 23–40.

Matsumoto, M., Nishimura, T. "Mersenne Twister: a 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator." *ACM Transactions on Modeling and Computer Simulation*, 8, 1 1998, pp. 3–30.

Matyas, L. *Generalized Method of Moments Estimation*. Cambridge: Cambridge University Press, 1999.

Matyas, L., and P. Sevestre, eds. *The Econometrics of Panel Data: Handbook of Theory and Applications.* 2nd ed., Dordrecht: Kluwer-Nijhoff, 1996.

McAleer, M. "The Significance of Testing Empirical Non-nested Models." *Journal of Econometrics,* 67, 1995, pp. 149–171.

McAleer, M., G. Fisher, and P. Volker. "Separate Misspecified Regressions and the U.S. Long-Run Demand for Money Function." *Review of Economics and Statistics,* 64, 1982, pp. 572–583.

McCallum, B. "A Note Concerning Covariance Expressions." *Econometrica,* 42, 1973, pp. 581–583.

McCoskey, S., and T. Selden. "Health Care Expenditures and GDP: Panel Data Unit Root Test Results." *Journal of Health Economics,* 17, 1998, pp. 369–376.

McCullagh, P., and J. Nelder. *Generalized Linear Models*. New York: Chapman and Hall, 1983.

McCullough, B. "Econometric Software Reliability: E-Views, LIMDEP, SHAZAM, and TSP." *Journal of Applied Econometrics,* 14, 2 1999, pp. 191–202.

McCullough, B., and C. Renfro. "Benchmarks and Software Standards: A Case Study of GARCH Procedures." *Journal of Economic and Social Measurement,* 25, 2, 1999, pp. 27–37.

McCullough, B., and H. Vinod. "The Numerical Reliability of Econometric Software." *Journal of Economic Literature,* 37, 2, 1999, pp. 633–665.

McDonald, J., and R. Moffitt. "The Uses of Tobit Analysis." *Review of Economics and Statistics,* 62, 1980, pp. 318–321.

McCullough, B. "Consistent Forecast Intervals When the Forecast Period Exogenous Variable Is Stochastic." *Journal of Forecasting*, 15, 4, 1996, pp. 293–304.

McDonald, J. and S.White. "A Comparison of Some Robust, Adaptive and Partially Adaptive Estimators of Regression Models." Econometric Reviews, 12, 1993, pp. 103–124.

McFadden, D. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by P. Zarembka, New York: Academic Press, 1974.

McFadden, D. "The Measurement of Urban Travel Demand." *Journal of Public Economics,* 3, 1974, pp. 303–328.

McFadden, D. "Econometric Analysis of Qualitative Response Models." In *Handbook of Econometrics,* Vol. 2, edited b y Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1984.

McFadden, D., and P. Ruud. "Estimation by Simulation." *Review of Economics and Statistics,* 76, 1994, pp. 591–608.

McFadden, D., and K. Train. "Mixed Multinomial Logit Models for Discrete Response." *Journal of Applied Econometrics,* 15, 2000, pp. 447–470.

McKenzie, C. "Microfit 4.0." *Journal of Applied Econometrics,* 13, 1998, pp. 77–90.

McKoskey, S., and C. Kao. "Testing the Stability of a Production Function with Urbanization as a Shift Factor: An Application of Nonstationary Panel Data Techniques." *Oxford Bulletin of Economics and Statistics,* 61, 1999, pp. 57–84.

McKoskey, S., and T. Selden. "Health Care Expenditures and GDP: Panel Data Unit Root Tests." *Journal of Health Economics,* 17, 1998, pp. 369–376.

McLachlan, G., and T. Krishnan. *The EM Algorithm and Extensions*. New York: John Wiley and Sons, 1997.

McLachlan, G., and D. Peel. *Finite Mixture Models*. New York: John Wiley and Sons, 2000.

McLaren, K. "Parsimonious Autocorrelation Corrections for Singular Demand Systems." *Economics Letters,* 53, 1996, pp. 115–121.

McMillen, D. "Probit with Spatial Autocorrelation." *Journal of Regional Science*, 32, 3, 1992, pp. 335–348.

Melenberg, B., and A. van Soest. "Parametric and Semi-Parametric Modelling of Vacation Expenditures." *Journal of Applied Econometrics,* 11, 1, 1996, pp. 59–76.

Messer, K., and H. White. "A Note on Computing the Heteroscedasticity Consistent Covariance Matrix Using Instrumental Variable Techniques." *Oxford Bulletin of Economics and Statistics,* 46, 1984, pp. 181–184.

Metz, A., and R. Cantor. "Moody's Credit Rating Prediction Model." Moody's, Inc., https://www.moodys.com/sites/products/DefaultResearch/2006200000425644.pdf , 2006.

Meyer, B. "Semiparametric Estimation of Hazard Models." Northwestern University, Department of Economics, 1988.

Michelsen, C., and R. Madlener. "Homeowners' Preferences for Adopting Innovative Residential Heating Systems: A Discrete Choice Analysis for Germany." *Energy Economics*, 34, 2012, pp. 1271–1283.

Miller, P., C. Mulvey, and N. Martin. "What Do Twins Studies Reveal About the Economic Returns to Education? A Comparison of Australian and U.S. Findings." *American Economic Review*, 85, 3, 1995, pp. 586–599.

Millimet, D., J. Smith, and E. Vytlacil, eds., *Modelling and Evaluating Treatment Effects in Econometrics, Advances in Econometrics*, Vol. 21, Oxford: Elsevier, 2008.

Mills, T. *Time Series Techniques for Economists*. New York: Cambridge University Press, 1990.

Mills, T. *The Econometric Modelling of Financial Time Series*. New York: Cambridge University Press, 1993.

Min, C., and A. Zellner. "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates." *Journal of Econometrics,* 56, 1993, pp. 89–118.

Mittelhammer, R., G. Judge, and D. Miller. *Econometric Foundations*. Cambridge: Cambridge University Press, 2000.

Mizon, G. "A Note to Autocorrelation Correctors: Don't." *Journal of Econometrics,* 69, 1, 1995, pp. 267–288.

Mizon, G., and J. Richard. "The Encompassing Principle and Its Application to Testing Nonnested Models." *Econometrica,* 54, 1986, pp. 657–678.

Moffitt, R., J. Fitzgerald, and P. Gottshalk. "Sample Attrition in Panel Data: The Role of Selection on Observables." *Annales d'Economie et de Statistique,* 55–56, 1999, pp. 129–152.

Mohanty, M. "A Bivariate Probit Approach to the Determination of Employment: A Study of Teen Employment Differentials in Los Angeles County." *Applied Economics,* 34, 2, 2002, pp. 143–156.

Monfardini, C., and R. Radice. "Testing Exogeneity in the Bivariate Probit Model: A Monte Carlo Study." *Oxford Bulletin of Economics and Statistics*, 70, 2, 2008, pp. 271–282.

Moran, P. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37, 1950, pp. 17–23.

Moscone, F., M. Knapp, and E. Tosetti. "Mental Health Expenditures in England: A Spatial Panel Approach." *Journal of Health Economics,* forthcoming, 2007.

Moshino, G., and D. Moro. "Autocorrelation Specification in Singular Equation Systems." *Economics Letters,* 46, 1994, pp. 303–309.

Mosteller, F. "The Tennessee Study of Class Sizes in the Early School Grades." *The Future of Children*, 5, 2, Summer/Fall, 1995, pp. 113–127.

Moulton, R. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *Review of Economics and Statistics*, 72, 1990, pp. 334–338.

Moulton, B. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics*, 32, 3, 1986, 385–97.

Mroz, T. "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions." *Econometrica,* 55, 1987, pp. 765–799.

Mullahy, J. "Specification and Testing of Some Modified Count Data Models." *Journal of Econometrics,* 33, 1986, pp. 341–365.

Mundlak, Y. "On the Pooling of Time Series and Cross Sectional Data." *Econometrica,* 56, 1978, pp. 69–86.

Munkin, M., and P. Trivedi. "Simulated Maximum Likelihood Estimation of Multivariate Mixed Poisson Regression Models with Application." *Econometric Journal,* 1, 1, 1999, pp. 1–21.

Munkin, M., and P. Trivedi. "Dental Insurance and Dental Care: The Role of Insurance and Income." HEDG working paper, University of York, 2007.

Munnell, A. "How Does Public Infrastructure Affect Regional Economic Performance." *New England Economic Review*, September/October, 1990, pp. 11–32.

Murphy, K., and R. Topel. "Estimation and Inference in Two Step Econometric Models." *Journal of Business and Economic Statistics,* 3, 1985, pp. 370–379. Reprinted, 20, 2002, pp. 88–97.

Murray, C., A. Tandon, C. Mathers, and R. Sudana. "New Approaches to Enhance Cross-Population Comparability of Survey Results." In C. Murray, A. Tandon, R. Mathers, R. Sudana, eds., *Summary Measures of Population Health*, Chapter 8.3, World Health Organization, 2002.

Nagin, D., and K. Land. "Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric Mixed Poisson Model." *Criminology,* 31, 3, 1993, pp. 327–362.

Nagler, J. "Scobit: An Alternative Estimator to Logit and Probit." *American Journal of Political Science*, 38, 1, 1994, pp. 230–255.

Nair-Reichert, U., and D. Weinhold. "Causality Tests for Cross Country Panels: A Look at FDI and Economic Growth in Less Developed Countries." *Oxford Bulletin of Economics and Statistics,* 63, 2, 2001, pp. 153–171.

Nakamura, A., and M. Nakamura. "Part-Time and Full-Time Work Behavior of Married Women: A Model with a Doubly Truncated Dependent Variable." *Canadian Journal of Economics,* 1983, pp. 229–257.

Nakosteen, R., and M. Zimmer. "Migration and Income: The Question of Self-Selection." *Southern Economic Journal,* 46, 1980, pp. 840–851.

Ndebele, T., & Marsh, D. "Consumer choice of electricity supplier: Investigating preferences for attributes of electricity services." In *New Zealand Agricultural and Economics Society 2013 Conference*. Conference held at Lincoln University, New Zealand, 2013. (http://ageconsearch.umn.edu//handle/160417)

Nelder, J., and R. Mead. "A Simplex Method for Function Minimization." *Computer Journal,* 7, 1965, pp. 308–313.

Nelson, F. "A Test for Misspecification in the Censored Normal Model." *Econometrica,* 49, 1981, pp. 1317–1329.

Nelson, C., and H. Kang. "Pitfalls in the Use of Time as an Explanatory Variable in Regression." *Journal of Business and Economic Statistics,* 2, 1984, pp. 73–82.

Nelson, C., and R. Startz. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator, *Econometrica*, 58, 4, 1990a, pp. 967–976.

Nelson, C., and R. Startz. "The Distribution of the Instrumental Variables Estimator and Its *t*-Ratio with the Instrument Is a Poor One." *Journal of Business*, 63, 1, 1990b, pp. S125–S140.

Nerlove, M. "Returns to Scale in Electricity Supply." In *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, edited

by C. Christ, Palo Alto: Stanford University Press, 1963.

Nerlove, M. "Further Evidence on the Estimation of Dynamic Relations from a Time Series of Cross Sections." *Econometrica,* 39, 1971a, pp. 359–382.

Nerlove, M. "A Note on Error Components Models." *Econometrica,* 39, 1971b, pp. 383–396.

Nerlove, M. *Essays in Panel Data Econometrics*. Cambridge: Cambridge University Press, 2002.

Nerlove, M., and S. Press. "Univariate and Multivariate Log-Linear and Logistic Models." RAND—R1306-EDA/NIH, Santa Monica, 1973.

Nerlove, M., and K. Wallis. "Use of the Durbin–Watson Statistic in Inappropriate Situations." *Econometrica,* 34, 1966, pp. 235–238.

Nevo, A. "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand." *Journal of Economics and Management Strategy*, 9, 4, 2000, pp. 513–548.

Nevo, A. "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica*, 69, 2, 2001, pp. 307–342.

Nevo, A., and M. Whinston. "Taking the Dogma out of Econometrics: Structural Modeling and Credible Evidence, *Journal of Economic Perspectives*, 24, 2, 2010, pp. 69–82.

Newey, W. "A Method of Moments Interpretation of Sequential Estimators." *Economics Letters,* 14, 1984, pp. 201–206.

Newey, W. "Maximum Likelihood Specification Testing and Conditional Moment Tests." *Econometrica,* 53, 1985a, pp. 1047–1070.

Newey, W. "Generalized Method of Moments Specification Testing." *Journal of Econometrics,* 29, 1985b, pp. 229–256.

Newey, W. "Specification Tests for Distributional Assumptions in the Tobit Model." *Journal of Econometrics,* 34, 1986, pp. 125–146.

Newey, W. "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables." *Journal of Econometrics,* 36, 1987, pp. 231–250.

Newey, W. "Two Step Series Estimation of Sample Selection Models." Manuscript, Department of Economics, MIT, 1991.

Newey, W. "The Asymptotic Variance of Semiparametric Estimators." *Econometrica,* 62, 1994, pp. 1349–1382.

Newey, W., and D. McFadden. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics,* Vol. IV, Chapter 36, edited by R. Engle and D. McFadden, 1994.

Newey, W., J. Powell, and J. Walker. "Semiparametric Estimation of Selection Models." *American Economic Review,* 80, 1990, pp. 324–328.

Newey, W., and K. West. "A Simple Positive Semi-Definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica,* 55, 1987a, pp. 703–708.

Newey, W., and K. West. "Hypothesis Testing with Efficient Method of Moments Estimation." *International Economic Review,* 28, 1987b, pp. 777–787.

New York Post. "America's New Big Wheels of Fortune." May 22, 1987, p. 3.

Neyman, J., and E. Scott. "Consistent Estimates Based on Partially Consistent Observations." *Econometrica,* 16, 1948, pp. 1–32.

Nickell, S. "Biases in Dynamic Models with Fixed Effects." *Econometrica,* 49, 1981, pp. 1417–1426.

Nicoletti, C., and F. Peracchi. "Survey Response and Survey Characteristics: Micro-level Evidence from the European Community Household Panel." *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 168, 2005, pp. 763–781.

Oaxaca, R. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review*, 14, 3, 1973, pp. 693–709.

Oberhofer, W., and J. Kmenta. "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models." *Econometrica,* 42, 1974, pp. 579–590.

Ohtani, K., and M. Kobayashi. "A Bounds Test for Equality Between Sets of Coefficients in Two Linear Regression Models Under Heteroscedasticity." *Econometric Theory,* 2, 1986, pp. 220–231.

Ohtani, K., and T. Toyoda. "Small Sample Properties of Tests of Equality Between Sets of Coefficients in Two Linear Regressions Under Heteroscedasticity." *International Economic Review,* 26, 1985, pp. 37–44.

Olsen, R. "A Note on the Uniqueness of the Maximum Likelihood Estimator in the Tobit Model." *Econometrica,* 46, 1978, pp. 1211–1215.

Orea, C., and S. Kumbhakar. "Efficiency Measurement Using a Latent Class Stochastic Frontier Model." *Empirical Economics,* 29, 2004, pp. 169–184.

Oreopoulos, P. "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter." *American Economic Review*, 96, 1, 2006, pp. 152–181.

Orcutt, G., S. Caldwell, and R. Wertheimer. *Policy Exploration through Microanalytic Simulation*. Washington, D.C.: Urban Institute, 1976.

Orme, C. "Double and Triple Length Regressions for the Information Matrix Test and Other Conditional Moment Tests." Mimeo, University of York, U.K., Department of Economics, 1990.

Osterwald-Lenum, M. "A Note on Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics." *Oxford Bulletin of Economics and Statistics,* 54, 1992, pp. 461–472.

Pagan, A., and A. Ullah. "The Econometric Analysis of Models with Risk Terms." *Journal of Applied Econometrics,* 3, 1988, pp. 87–105.

Pagan, A., and A. Ullah. *Nonparametric Econometrics*. Cambridge: Cambridge University Press, 1999.

Pagan, A., and F. Vella. "Diagnostic Tests for Models Based on Individual Data: A Survey." *Journal of Applied Econometrics,* 4, Supplement, 1989, pp. S29–S59.

Pagan, A., and M. Wickens. "A Survey of Some Recent Econometric Methods." *Economic Journal,* 99, 1989, pp. 962–1025.

Pakes, A., and D. Pollard. "Simulation and the Asymptotics of Optimization Estimators." *Econometrica,* 57, 1989, pp. 1027–1058.

Papke, L., and J. Wooldridge. "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates." *Journal of Econometrics*, 145, 2008, pp. 121–133.

Passmore, W. "The GSE Implicit Subsidy and the Value of Government Ambiguity." *FEDS* Working paper no. 2005-05, Board of Governors of the Federal Reserve—Household and Real Estate Finance Section, 2005.

Passmore, W., S. Sherlund, and G. Burgess. "The Effect of Housing Government Sponsored Enterprises on Mortgage Rates." *Real Estate Economics,* 33, 3, 2005, pp. 427–463.

Passmore, W., R. Sparks, and J. Ingpen. "GSEs, Mortgage Rates, and the Long-Run Effects of Mortgage Securitization." *Journal of Real Estate Finance and Economics*, 25, 2, 2002, pp. 215–242.

Pedroni, P. "Fully Modified OLS for Heterogeneous Cointegrated Panels." *Advances in Econometrics*, 15, 2000, pp. 93–130.

Pedroni, P. "Purchasing Power Parity Tests in Cointegrated Panels." *Review of Economics and Statistics,* 83, 2001, pp. 727–731.

Pedroni, P. "Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the PPP Hypothesis." *Econometric Theory,* 20, 2004, pp. 597–625.

Pesaran, H., and M. Weeks. "Nonnested Hypothesis Testing: An Overview." In *A Companion to Theoretical Econometrics,* edited by B. Baltagi Blackwell, Oxford, 2001.

Pesaran, M., Y. Shin, and R. Smith. "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels." *Journal of the American Statistical Association,* 94, 1999, pp. 621–634.

Pesaran, M., Y. Shin, and R. Smith. "Bounds Testing Approaches to the Analysis of Long Run Relationships." *Journal of Applied Econometrics*, 16, 3, 2001, pp. 289–326.

Pesaran, M., and R. Smith. "Estimating Long Run Relationships from Dynamic Heterogeneous Panels." *Journal of Econometrics,* 68, 1995, pp. 79–113.

Pesaresi, E., C. Flanagan, D. Scott, and P. Tragear. "Evaluating the Office of Fair Trading's 'Fee-Paying Schools' Intervention." *European Journal of Law and Economics*, 40, 3, 2015, pp. 413–429.

Petersen, D., and D. Waldman. "The Treatment of Heteroscedasticity in the Limited Dependent Variable Model." Mimeo, University of North Carolina, Chapel Hill, November 1981.

Pforr, K. "Implementation of a Multinomial Logit Model with Fixed Effects." Manuscript, Mannheim Center for European Social Research, University of Mannheim, 2011: (www.stata.com/meeting/germany11/desug11_pforr.pdf).

Phillips, A. "Stabilization Policies and the Time Form of Lagged Responses." *Economic Journal,* 67, 1957, pp. 265–277.

Phillips, P. "Exact Small Sample Theory in the Simultaneous Equations Model." In *Handbook of Econometrics,* Vol. 1, edited by Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1983.

Phillips, P. "Understanding Spurious Regressions." *Journal of Econometrics,* 33, 1986, pp. 311–340.

Phillips, P., and H. Moon. "Nonstationary Panel Data Analysis: An Overview of Some Recent Developments." *Econometric Reviews,* 19, 2000, pp. 263–286.

Phillips, P., and S. Ouliaris. "Asymptotic Properties of Residual Based Tests for Cointegration." *Econometrica,* 58, 1990, pp. 165–193.

Phillips, P., and P. Perron. "Testing for a Unit Root in Time Series Regression." *Biometrika,* 75, 1988, pp. 335–346.

Pinske, J., and Slade, M. "Contracting in Space: An Application of Spatial Statistics to Discrete Choice Models." *Journal of Econometrics*, 85, 1, 1998, pp 125–154.

Pinkse, J., M. Slade, and L. Shen. "Dynamic Spatial Discrete Choice Using One Step GMM: An Application to Mine Operating Decisions", *Spatial Economic Analysis*, 1, 1, 2006, pp. 53–99.

Pitt, M., and L. Lee. "The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry." *Journal of Development Economics*, 9, 1984, pp. 43–64.

Poirier, D. "Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics." *Journal of Economic Perspectives,* 2, 1988, pp. 121–170.

Poirier, D., ed. "Bayesian Empirical Studies in Economics and Finance." *Journal of Econometrics,* 49, 1991, pp. 1–304.

Poirier, D. *Intermediate Statistics and Econometrics*. Cambridge: MIT Press, 1995, pp. 1–217.

Poirier, D., and J. Tobias. "Bayesian Econometrics." In, *Palgrave Handbook of Econometrics, Volume 1: Theoretical Econometrics*, edited by T. Mills and K. Patterson, London: Palgrave-Macmillan, 2006.

Powell, J. "Least Absolute Deviations Estimation for Censored and Truncated Regression Models." Technical Report 356, Stanford University, IMSSS, 1981.

Powell, J. "Least Absolute Deviations Estimation for the Censored Regression Model." *Journal of Econometrics,* 25, 1984, pp. 303–325.

Powell, J. "Censored Regression Quantiles." *Journal of Econometrics,* 32, 1986a, pp. 143–155.

Powell, J. "Symmetrically Trimmed Least Squares Estimation for Tobit Models." *Econometrica,* 54, 1986b, pp. 1435–1460.

Powell, J. "Estimation of Semiparametric Models." In the *Handbook of Econometrics*, Vol. 4, edited by ed. R. Engle and D. McFadden, Amsterdam, North Holland, 1994.

Powell, M. "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives." *Computer Journal,* 1964, pp. 165–172.

Prais, S., and C. Winsten. "Trend Estimation and Serial Correlation." Cowles Commission discussion paper no. 383, Chicago, 1954.

Pratt, J. "Concavity of the Log Likelihood." *Journal of the American Statistical Association,* 76, pp. 103–106.

Prentice, R., and L. Gloeckler. "Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data." *Biometrics,* 34, 1978, pp. 57–67.

Press, W., B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes: The Art of Scientific Computing*. 3rd ed., Cambridge: Cambridge University Press, 2007.

Preston, S. "The Changing Relation Between Mortality and Level of Economic Development." *Population Studies*, 29, 1975, pp. 231–248.

Pudney, S., and M. Shields. "Gender, Race, Pay and Promotion in the British Nursing Profession: Estimation of a Generalized Ordered Probit Model." *Journal of Applied Econometrics,* 15, 4, 2000, pp. 367–399.

Quandt, R. "Computational Problems and Methods." In *Handbook of Econometrics,* Vol. 1, edited by Z. Griliches and M. Intriligator, Amsterdam: North Holland, 1983.

Quandt, R., and J. Ramsey. "Estimating Mixtures of Normal Distributions and Switching Regressions." *Journal of the American Statistical Association,* 73, December 1978, pp. 730–738.

Quester, A., and W. Greene. "Divorce Risk and Wives' Labor Supply Behavior." *Social Science Quarterly,* 63, 1982, pp. 16–27.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects." *Journal of Econometrics*, 128, 2005, pp. 301–323.

Ramsey, J. "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis." *Journal of the Royal Statistical Society,Series B*, 31, 1969, pp. 350–367.

Raj, B., and B. Baltagi, eds. *Panel Data Analysis*. Heidelberg: Physica-Verlag, 1992.

Rao, C. *Linear Statistical Inference and Its Applications*. New York: John Wiley and Sons, 1973.

Rao, C. "Information and Accuracy Attainable in Estimation of Statistical Parameters." *Bulletin of the Calcutta Mathematical Society,* 37, 1945, pp. 81–91.

Rasch, G. "Probabilistic Models for Some Intelligence and Attainment Tests." *Denmark Paedogiska,* Copenhagen, 1960.

Ravid, A. "Information, Blockbusters, and Stars: A Study of the Film Industry." *Journal of Business*, 72, 4, 1999, pp. 463–492.

Renfro, C. "Econometric Software." *Journal of Economic and Social Measurement,* 2007.

Renfro, C. "Econometric Software." in *Handbook of Computational Econometrics*, E. Kontoghiorghes and D. Belsley, eds., John Wiley and Sons, London, 2009, pp. 1–60.

Revelt, D., and K. Train. "Incentives for Appliance Efficiency: Random-Parameters Logit Models of Households' Choices." Manuscript, Department of Economics, University of California, Berkeley, 1996.

Revelt, D., and K. Train. "Customer Specific Taste Parameters and Mixed Logit: Households' Choice of Electricity Supplier." Economics Working paper, E00-274, Department of Economics, University of California at Berkeley, 2000.

Ridder, G., and T. Wansbeek. "Dynamic Models for Panel Data." In *Advanced Lectures in Quantitative Economics*, edited by R. van der Ploeg, New York: Academic Press, 1990, pp. 557–582.

Riphahn, R., A. Wambach, and A. Million. "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation." *Journal of Applied Econometrics,* 18, 4, 2003, pp. 387–405.

Rivers, D., and Q. Vuong. "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics,* 39, 1988, pp. 347–366.

Robertson, D., and J. Symons. "Some Strange Properties of Panel Data Estimators." *Journal of Applied Econometrics,* 7, 1992, pp. 175–189.

Robins, J., A. Rotnitzky, and L. Zhao. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association,* 90, 1995, pp. 106–121.

Robinson, C., and N. Tomes. "Self-Selection and Interprovincial Migration in Canada." *Canadian Journal of Economics,* 15, 1982, pp. 474–502.

Robinson, P. "Semiparametric Econometrics: A Survey." *Journal of Applied Econometrics,* 3, 1988, pp. 35–51.

Rogers, W. "Calculation of Quantile Regression Standard Errors." Stata Technical Bulletin No. 13, Stata Corporation, College Station, TX, 1993.

Rosenbaum, P., and D. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika,* 70, 1983, pp. 41–55.

Rosett, R., and F. Nelson. "Estimation of the Two-Limit Probit Regression Model." *Econometrica,* 43, 1975, pp. 141–146.

Rossi, P., and G. Allenby. "Marketing Models of Consumer Heterogeneity." *Journal of Econometrics,* 89, 1999, pp. 57–78.

Rossi, P., and G. Allenby. "Bayesian Statistics and Marketing." *Marketing Science,* 22, 2003, 304–328.

Rossi, P., G. Allenby, and R. McCulloch. *Bayesian Statistics and Marketing*. New York: John Wiley and Sons, 2005.

Rothstein, J. "Does Competition Among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000)." Working paper no. 10, Princeton University, Education Research Section, 2004.

Rotnitzky, A., and J. Robins. "Inverse Probability Weighted Estimation in Survival Analysis." In *Encyclopedia of Biostatistics*, edited

by P. Armitage and T. Coulton, New York: Wiley, 2005.

Rouse, C. "Further Estimates of the Economic Return to Schooling from a New Sample of Twins." *Economics of Education Review,* 18, 2, 1999, pp. 149–157.

Rubin, D. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology,* 55, 1974, pp. 688–701.

Rubin, D. "Inference and Missing Data." *Biometrika,* 63, 1976, pp. 581–592.

Rubin, D. "Bayesian Inference for Causal Effects." *Annals of Statistics*, 6, 1978, pp. 34–58.

Rubin, D. *Multiple Imputation for Non-response in Surveys*. New York: John Wiley and Sons, 1987.

Rubin, H. "Consistency of Maximum Likelihood Estimators in the Explosive Case." In *Statistical Inference in Dynamic Economic Models*, edited by T. Koopmans, New York: John Wiley and Sons, 1950.

Ruud, P. *An Introduction to Classical Econometric Theory*. Oxford: Oxford University Press, 2000.

Ruud, P. "A Score Test of Consistency." Manuscript, Department of Economics, University of California, Berkeley, 1982.

Ruud, P. "Tests of Specification in Econometrics." *Econometric Reviews,* 3, 1984, pp. 211–242.

Sala-i-Martin, X. "The Classical Approach to Convergence Analysis." *Economic Journal,* 106, 1996, pp. 1019–1036.

Sala-i-Martin, X. "I Just Ran Two Million Regressions." *American Economic Review,* 87, 1997, pp. 178–183.

Salisbury, L., and F. Feinberg. "Alleviating the Constant Stochastic Variance Assumption in Decision Research: Theory, Measurement and Experimental Test." *Marketing Science*, 29, 1, 2010, pp. 1–17.

Salmon, F. "Recipe for Disaster: The Formula that Killed Wall Street." *Wired Magazine*, 17, 3, 2009 (www.wired.com/2009/02/wp-quant/).

Samuelson, P. *Foundations of Economic Analysis*, Harvard University Press, Cambridge, 1938.

Saxonhouse, G. "Estimated Parameters as Dependent Variables." *American Economic Review, 66,* 1, 1976, pp. 178–183.

Scarpa R, M. Thiene, and D. Hensher. "Monitoring Choice Task Attribute Attendance in Nonmarket Valuation of Multiple Park Management Services: Does It Matter?" *Land Economics*, 86, 4, 2010, pp. 817–839.

Scarpa, R., M. Thiene, and K., Train. "Utility in Willingness To Pay Space: A Tool to Address Confounding Random Scale Effects in Destination Choice to the Alps." *American Journal of Agricultural Economics*, 90, 4, 2008, pp. 994–1010.

Scarpa, R., and K. Willis. "Willingness to Pay for Renewable Energy: Primary and Discretionary Choice of British Households' for Micro-Generation Technologies." *Energy Economics*, 32, 2010, pp. 129–136.

Schimek, M. ed. *Smoothing and Regression: Approaches, Computation, and Applications,* New York: John Wiley and Sons, 2000.

Schmidt, P., and R. Sickles. "Some Further Evidence on the Use of the Chow Test Under Heteroscedasticity." *Econometrica,* 45, 1977, pp. 1293–1298.

Schmidt, P., and R. Sickles. "Production Frontiers and Panel Data." *Journal of Business and Economic Statistics,* 2, 1984, pp. 367–374.

Schmidt, P., and R. Strauss. "The Prediction of Occupation Using Multinomial Logit Models." *International Economic Review,* 16, 1975a, pp. 471–486.

Schmidt, P., and R. Strauss. "Estimation of Models with Jointly Dependent Qualitative Variables: A Simultaneous Logit Approach." *Econometrica,* 43, 1975b, pp. 745–755.

Scott, A., S. Schurer, P. Jensen and P. Sivey. "The Effects of an Incentive Program on Quality of Care in Diabetes Management." *Health Economics*, 18, 2009, pp. 1091–1108.

Seaks, T., and K. Layson. "Box–Cox Estimation with Standard Econometric Problems." *Review of Economics and Statistics,* 65, 1983, pp. 160–164.

Sepanski, J. "On a Random Coefficients Probit Model." *Communications in Statistics—Theory and Methods,* 29, 2000, pp. 2493–2505.

Shaw, D. "On-Site Samples' Regression Problems of Nonnegative Integers, Truncation, and Endogenous Stratification." *Journal of Econometrics,* 37, 1988, pp. 211–223.

Shea, J. "Instrument Relevance in Multivariate Linear Models: A Simple Measure." *Review of Economics and Statistics,* 79, 1997, pp. 348–352.

Shephard, R. *Cost and Production Functions*. Princeton: Princeton University Press, 1953.

Shephard, R. *The Theory of Cost and Production*. Princeton: Princeton University Press, 1970.

Sherlund, S., and Gillian Burgess. "The Effect of Housing Government Sponsored Enterprises on Mortgage Rates." *Real Estate Economics*, 33, 3, 2005, pp. 427–463.

Sickles, R. "Panel Estimators and the Identification of Firm-Specific Efficiency Levels in Semiparametric and Non-Parametric Settings." *Journal of Econometrics*, 126, 2005, pp. 305–324.

Sickles, R., D. Good, and R. Johnson. "Allocative Distortions and the Regulatory Transition of the Airline Industry." *Journal of Econometrics,* 33, 1986, pp. 143–163.

Silver, J., and M. Ali. "Testing Slutsky Symmetry in Systems of Linear Demand Equations." *Journal of Econometrics,* 41, 1989, pp. 251–266.

Silverman, B. W., *Density Estimation*. London: Chapman and Hall, 1986.

Simar, L., and P. Wilson. "A General Methodology for Bootstrapping in Nonparametric Frontier Models." *Journal of Applied Statistics,* 27, 6, 2000, pp. 779–802.

Simar, L., and P. Wilson. "Estimation and Inference in Two Stage, Semiparametric Models of Production Processes." *Journal of Econometrics*, 136, 1, 2007, pp. 31–64.

Simonoff, J., and I. Sparrow. "Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers." *Chance*, 13, 3, 2000, pp. 15–24.

Somonsen, M., L. Skipper, and N. Skipper. "Price Sensitivity of Demand for Prescription Drugs:" Exploiting a Regression Kink Design." *Journal of Applied Econometrics*, 31, 2, 2016, pp. 320–337.

Sims, C. "But Economics Is Not an Experimental Science." *Journal of Economic Perspectives*, 24, 2, 2010, pp. 59–68.

Sirven, N., B. Santos-Egglmann, and J. Spagnoli. "Comparability of Health Care Responsiveness in Europe Using Anchoring Vignettes." Working paper 15, IRDES (France), 2008.

Sklar, A. "Random Variables, Joint Distributions and Copulas." *Kybernetica,* 9, 1973, pp. 449–460.

Smirnov, O. "Modeling Spatial Discrete Choice." *Regional Science and Urban Economics*, 40, 2010, pp. 292–298.

Smith, M. "Modeling Selectivity Using Archimedean Copulas." *Econometrics Journal,* 6, 2003, pp. 99–123.

Smith, M. "Using Copulas to Model Switching Regimes with an Application to Child Labour." *Economic Record,* 81, 2005, pp. S47–S57.

Smith, R. "Estimation and Inference with Nonstationary Panel Time Series Data." Manuscript, Department of Economics, Birkbeck College, 2000.

Smith, V. "Selection and Recreation Demand." *American Journal of Agricultural Economics,* 70, 1988, pp. 29–36.

Smith, M., D. Hochberg, and W. Greene. "The Effectiveness of Pre-purchase Homeownership Counseling and Financial Management Skills." Federal Reserve Bank of Philadelphia, 2014, Retrieved from www.philadelphiafed.org/communitydevelopment/homeownership-counseling-study/2014/homeownership-counselingstudy-042014.pdf.

Smith, J., and P. Todd. "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods." *Journal of Econometrics*, 125, 2005, pp. 305–353.

Snell, E. "A Scaling Procedure for Ordered Categorical Data." *Biometrics*, 20, 1964, pp. 592–607.

Snow J., *On the Mode of Communication of Cholera*, London: Churchill, 1855. [Reprinted 1965 by Hafner, New York.]

Solow, R. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics,* 39, 1957, pp. 312–320.

Sonnier, G., A. Ainslie, and T. Otter. "Heterogeneity Distributions of Willingness-To-Pay in Choice Models." *Quantitative Marketing Economics,* 5, 3, 2007, pp. 313–331.

Spector, L., and M. Mazzeo. "Probit Analysis and Economic Education." *Journal of Economic Education,* 11, 1980, pp. 37–44.

Srivistava, V., and D. Giles. *Seemingly Unrelated Regression Models: Estimation and Inference*. New York: Marcel Dekker, 1987.

Staiger, D., and J. Stock. "Instrumental Variables Regression with Weak Instruments." *Econometrica,* 65, 1997, pp. 557– 586.

Staiger, D., J. Stock, and M. Watson. "How Precise Are Estimates of the Natural Rate of

Unemployment?" NBER Working paper no. 5477, Cambridge, 1996.

Stata. *Stata User's Guide, Version 14*. College Station, TX: Stata Press, 2014.

Stern, S. "Two Dynamic Discrete Choice Estimation Problems and Simulation Method Solutions." *Review of Economics and Statistics,* 76, 1994, pp. 695–702.

Stevenson, R. 1980. "Likelihood Functions for Generalized Stochastic Frontier Estimation." *Journal of Econometrics*, 13, pp. 58–66.

Stewart, M. "Maximum Simulated Likelihood Estimation of Random Effects Dynamic Probit Models with Autocorrelated Errors." *Stata Journal*, 6, 2, 2006, pp. 256–272.

Stock, J. "The Other Transformation in Econometric Practice: Robust Tools for Inference." *Journal of Economic Perspectives*, 24, 2, 2010, pp. 83–94.

Stock, J., and M. Watson. "Forecasting Output and Inflation: The Role of Asset Prices." *NBER,* Working paper no. 8180, Cambridge, MA, 2001.

Stock, J., and M. Watson. "Combination Forecases of Output Growth in a Seven-Country Data Set." *Journal of Forecasting,* 23, 6, 2004, pp. 405–430.

Stock, J., and M. Watson. *Introduction to Econometrics.* 2nd ed., 2007.

Stock, J., J. Wright, and M. Yogo. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business and Economic Statistics,* 20, 2002, pp. 518–529.

Stoker, T. "Consistent Estimation of Scaled Coefficients." *Econometrica,* 54, 1986, pp. 1461–1482.

Stoker, T. "Lectures on Semiparametric Econometrics." Lecture Series, CORE Foundation, Louvain–la-Neuve, Belgium, 1992.

Strang, G. *Linear Algebra and Its Applications.* 5th ed., New York: Academic Press, 2016.

Stuart, A., and S. Ord. *Kendall's Advanced Theory of Statistics*. New York: Oxford University Press, 1989.

Suits, D. "Dummy Variables: Mechanics vs. Interpretation." *Review of Economics and Statistics,* 66, 1984, pp. 177–180.

Susin, S. "Hazard Hazards: The Inconsistency of the "Kaplan-Meier Empirical Hazard." and Some Alternatives." Manuscript, U.S. Census Bureau, 2001.

Swamy, P. "Efficient Inference in a Random Coefficient Regression Model." *Econometrica,* 38, 1970, pp. 311–323.

Swamy, P. *Statistical Inference in Random Coefficient Regression Models*. New York: Springer-Verlag, 1971.

Swamy, P. "Statistical Inference in Random Coefficient Models." *Springer Lecture Notes in Economic and Mathematical Systems*, Heidelberg, 1971, p. 126

Swamy, P. "Linear Models with Random Coefficients." In *Frontiers in Econometrics*, edited by P. Zarembka, New York: Academic Press, 1974.

Swamy, P., and G. Tavlas. "Random Coefficients Models: Theory and Applications." *Journal of Economic Surveys,* 9, 1995, pp. 165–182.

Swamy, P., and G. Tavlas. "Random Coefficient Models." In *A Companion to Theoretical Econometrics*, edited by B. Baltagi, Oxford: Blackwell, 2001.

Tamm, M., H. Tauchmann, J. Wasem, and S. Gress. "Elasticities of Market Shares and Social Health Insurance Choice in Germany: A Dynamic Panel Data Approach." *Health Economics*, 16, 2007, pp. 243–256.

Tandon, A., C. Murray, J. Lauer, and D. Evans. "Measuring the Overall Health System Performance for 191 Countries." World Health Organization, GPE discussion paper, EIP/GPE/EQC, no. 30, 2000, www.who.int/entity/healthinfo/paper30.pdf

Taubman, P. "The Determinates of Earnings: Genetics, Family and Other Environments, A Study of White Male Twins." *American Economic Review*, 66, 5, 1976, pp. 858–870.

Tauchen, H., A. Witte, and H. Griesinger. "Criminal Deterrence: Revisiting the Issue with a Birth Cohort." *Review of Economics and Statistics,* 3, 1994, pp. 399–412.

Taylor, L. "Estimation by Minimizing the Sum of Absolute Errors." In *Frontiers in Econometrics*, edited by P. Zarembka, New York: Academic Press, 1974.

Taylor, W. "Small Sample Properties of a Class of Two Stage Aitken Estimators." *Econometrica,* 45, 1977, pp. 497–508.

Terza, J. "Ordinal Probit: A Generalization." *Communications in Statistics,* 14, 1985a, pp. 1–12.

Terza, J. "A Tobit Type Estimator for the Censored Poisson Regression Model." *Economics Letters,* 18, 1985b, pp. 361–365.

Terza, J. "Estimating Count Data Models with Endogenous Switching and Sample Selection." Working paper IPRE-95-14, Department of Economics, Pennsylvania State University, 1995.

Terza, J. "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects." *Journal of Econometrics,* 84, 1, 1998, pp. 129–154.

Terza, J. "Parametric Nonlinear Regression with Endogenous Switching." *Econometric Reviews*, 28, 6, 2009, pp. 555–581.

Terza, J., A. Basu, and P. Rathouz. "Two State Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling." *Journal of Health Economics*, 27, 2008, pp. 531–543.

Terza, J., and D. Kenkel. "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect." *Journal of Applied Econometrics,* 16, 2, 2001, pp. 165–184.

Theil, H. *Economic Forecasts and Policy*. Amsterdam: North Holland, 1961.

Theil, H. *Principles of Econometrics*. New York: John Wiley and Sons, 1971.

Theil, H. "Linear Algebra and Matrix Methods in Econometrics." In *Handbook of Econometrics,* Vol. 1, edited by Z. Griliches and M. Intriligator, New York: North Holland, 1983.

Theil, H., and A. Goldberger. "On Pure and Mixed Estimation in Economics." *International Economic Review,* 2, 1961, pp. 65–78.

Theil, H. "Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations." *Econometrica*, 30, 1962, pp. 54–78.

Tjur, T. "Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination." *The American Statistician*, 63, 2009, pp. 366–372.

Tobin, J. "Estimation of Relationships for Limited Dependent Variables." *Econometrica,* 26, 1958, pp. 24–36.

Toyoda, T., and K. Ohtani. "Testing Equality Between Sets of Coefficients After a Preliminary Test for Equality of Disturbance Variances in Two Linear Regressions." *Journal of Econometrics,* 31, 1986, pp. 67–80.

Train, K. "Halton Sequences for Mixed Logit." Manuscript, Department of Economics, University of California, Berkeley, 1999.

Train, K. "A Comparison of Hierarchical Bayes and Maximum Simulated Likelihood for Mixed Logit." Manuscript, Department of Economics, University of California, Berkeley, 2001.

Train, K., *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press, 2003. 2nd ed., 2009.

Train, K., and G. Sonnier. "Mixed Logit with Bounded Distributions of Correlated Partworths." In *Applications of Simulation Methods in Environmental and Resource Economics*, A. Alberini and R.l Scarpa, eds., Boston: Kluwer, 2003.

Train, K., and D. McFadden. "Mixed MNL Models for Discrete Response." Journal of Applied Econometrics, 15, 2000, pp. 447–470.

Train, K., and M. Weeks. "Discrete Choice Models in Preference Space and Willingness to Pay Space." In *Applications of Simulation Methods in Environmental and Resource Economics,* R. Scarpa and A. Alberini, eds., Springer Publisher, Dordrecht, Chapter 1, pp. 1–16, 2005.

Trivedi, P., and D. Zimmer. "Copula Modeling: An Introduction for Practitioners." *Foundations and Trends in Econometrics*, 2007.

Trochim, W. *Regression Research Design for Program Evaluation: The Regression Discontinuity Approach*. Beverly Hills: Sage Publications, 1984.

Trochim, W. "The Regression Discontinuity Design." Social Research Methods, http://www.socialresearchmethods.net/kb/quasird.php, 2006.

Tsay, R., *Analysis of Financial Time Series*, 2nd ed., John Wiley and Sons, New York, 2005.

Tsionas, E. "Stochastic Frontier Models with Random Coefficients." *Journal of Applied Econometrics*, 17, 2002, pp. 127–147.

Tunali, I. "A General Structure for Models of Double Selection and an Application to a Joint Migration/Earnings Process with Remigration." *Research in Labor Economics,* 8, 1986, pp. 235–282.

UK Office of Fair Trading. "Evaluation of an OFT Intervention, Independent Fee-Paying Schools." Working paper OFT 1416, 2012.

Van der Klaauw, W. "Estimating the Effect of Financial Aid Offers on College Enrollment; A Regression-Discontinuity Approach." *International Economic Review*, 43, 2002, pp. 1249–1287.

van Ooijen, R., R. Alessie, and M. Knoef. "Health Status over the Life Cycle." Health Econometrics and Data Group, University of York, working paper 15/21, 2015.

van Soest, A., L. Delaney, C. Harmon, A. Kapteyn, and J. Smith. "Validating the Use of Vignettes for Subjective Threshold Scales." Working paper WP/14/2007, Geary Institute, University College, Dublin, 2007.

Varian, H. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives*, 28, 2, 2014, pp. 3–28.

Veall, M. "Bootstrapping the Probability Distribution of Peak Electricity Demand." *International Economic Review,* 28, 1987, pp. 203–212.

Veall, M. "Bootstrapping the Process of Model Selection: An Econometric Example." *Journal of Applied Econometrics,* 7, 1992, pp. 93–99.

Veall, M., and K. Zimmermann. "Pseudo-R^2's-in the Ordinal Probit Model." *Journal of Mathematical Sociology,* 16, 1992, pp. 333–342.

Vella, F. "Estimating Models with Sample Selection Bias: A Survey." *Journal of Human Resources,* 33, 1998, pp. 439–454.

Vella, F., and M. Verbeek. "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men." *Journal of Applied Econometrics,* 13, 2, 1998, pp. 163–184.

Vella, F., and M. Verbeek. "Two-Step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias." *Journal of Econometrics,* 90, 1999, pp. 239–263.

Verbeek, M. "On the Estimation of a Fixed Effects Model with Selectivity Bias." *Economics Letters,* 34, 1990, pp. 267–270.

Verbeek, M., and T. Nijman. "Testing for Selectivity Bias in Panel Data Models." *International Economic Review,* 33, 3, 1992, pp. 681–703.

Versaci, A. "You Talkin' to Me? Using Internet Buzz as an Early Predictor of Movie Box Office." Stern School of Business, Department of Marketing, manuscript, 2009.

Vinod, H. "Bootstrap, Jackknife, Resampling and Simulation: Applications in Econometrics." In *Handbook of Statistics: Econometrics, Vol II*., Chapter 11, edited by G. Maddala, C. Rao, and H. Vinod, Amsterdam: North Holland, 1993.

Vinod, H., and B. Raj. "Economic Issues in Bell System Divestiture: A Bootstrap Application." *Applied Statistics (Journal of the Royal Statistical Society, Series C),* 37, 2, 1994, pp. 251–261.

Vuong, Q. "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses." *Econometrica,* 57, 1989, pp. 307–334.

Vytlacil, E., A. Aakvik, and J. Heckman. "Treatment Effects for Discrete Outcomes When Responses to Treatments Vary Among Observationally Identical Persons: An Application to Norwegian Vocational Rehabilitation Programs." *Journal of Econometrics,* 125, 1/2, 2005, pp. 15–51.

Wald, A. "The Fitting of Straight Lines if Both Variables Are Subject to Error" *Annals of Mathematical Statistics*, 11, 3, 1940, pp. 284–300.

Waldman, D. "A Stationary Point for the Stochastic Frontier Likelihood." *Journal of Econometrics*, 18, 1982, pp. 275–279.

Waldman, D. "A Note on the Algebraic Equivalence of White's Test and a Variant of the Godfrey/Breusch-Pagan Test for Heteroscedasticity." *Economics Letters,* 13, 1983, pp. 197–200.

Waldman, M., S. Nicholson, N. Adilov, and J. Williams. "Autism Prevalence and Precipitation Rates in California, Oregon, and Washington Counties." *Archives of Pediatrics & Adolescent Medicine*, 162, 2008, pp. 1026–1034.

Walker, S., and D. Duncan. "Estimation of the Probability of an Event as a Function of Several Independent Variables." *Biometrika*, 54, 1967, pp. 167–179.

Wallace, T., and A. Hussain. "The Use of Error Components in Combining Cross Section with Time Series Data." *Econometrica,* 37, 1969, pp. 55–72.

Wang, P., I. Cockburn, and M. Puterman. "Analysis of Panel Data: A Mixed Poisson Regression Model Approach." *Journal of Business and Economic Statistics,* 16, 1, 1998, pp. 27–41.

Wang, H., E. Iglesias, and J. Wooldridge. "Partial Maximum Likelihood Estimation of Spatial Probit Models." *Journal of Econometrics*, 172, 1, 2013, pp. 77–89.

Wang, X., and K. Kockelman. "Application of the Dynamic Spatial Ordered Probit Model: Patterns of Ozone Concentration in Austin, Texas." Manuscript, Department of Civil Engineering, University of Texas, Austin, 2009.

Wang, C., and Y. Zhou. "Deliveries to Residential Units: A Rising Form of Freight Transportation in the U.S.." *Transportation Research Part C*, 58, 2015, pp. 46–55.

Wasi, N., and R. Carson. "The Influence of Rebate Programs on the Demand for Water Heaters: The Case of New South Wales." *Enerty Economics*, 40, 2013, pp. 645–656.

Watson, M. "Vector Autoregressions and Cointegration." In *Handbook of Econometrics,* Vol. 4., R. Engle and D. McFadden, eds., Amsterdam: North Holland, 1994.

Wedel, M., W. DeSarbo, J. Bult, and V. Ramaswamy. "A Latent Class Poisson Regression Model for Heterogeneous Count Data." *Journal of Applied Econometrics,* 8, 1993, pp. 397–411.

Weinhold, D. "A Dynamic "Fixed Effects" Model for Heterogeneous Panel Data." Manuscript, Department of Economics, London School of Economics, 1999.

Weinhold, D. "Investment, Growth and Causality Testing in Panels" (in French). *Economie et Prevision,* 126-5, 1996, pp. 163–175.

Weiss, A. "Asymptotic Theory for ARCH Models: Stability, Estimation, and Testing." Discussion paper 82-36, Department of Economics, University of California, San Diego, 1982.

West, K. "On Optimal Instrumental Variables Estimation of Stationary Time Series Models." *International Economic Review,* 42, 4, 2001, pp. 1043–1050.

White, H. "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity." *Econometrica,* 48, 1980, pp. 817–838.

White, H. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica,* 53, 1982a, pp. 1–16.

White, H., ed. "Non-nested Models." *Journal of Econometrics,* 21, 1, 1983, pp. 1–160.

White, H. *Asymptotic Theory for Econometricians, Revised*. New York: Academic Press, 2001.

Whitehouse, M. "Mind and Matter: Is an Economist Qualified To Solve Puzzle of Autism?" *Wall Street Journal*, February 27, 2007.

Wickens, M. "A Note on the Use of Proxy Variables." *Econometrica,* 40, 1972, pp. 759–760.

Wilde, J. "Identification of Multiple Equation Probit Models with Endogenous Dummy Regressors." Economics Letters, 69, 3, 2000, pp. 309–312.

Williams, R. "Logistic Regression, Part II: The Logistic Regression Model (LRM) – Interpreting Parameters." Manuscript, Department of Sociology, Notre Dame University, accessed August 15, 2016, www3.nd.edu/~rwilliam/stats2/l82.pdf.

Willis, J. "Magazine Prices Revisited." *Journal of Applied Econometrics,* 21, 3, 2006, pp. 337–344.

Willis, R., and S. Rosen. "Education and Self-Selection." *Journal of Political Economy,* 87, 1979, pp. S7–S36.

Windmeijer, F. "Goodness of Fit Measures in Binary Choice Models." *Econometric Reviews,* 14, 1995, pp. 101–116.

Winkelmann, R. "Subjective Well-Being and the Family: Results from an Ordered Probit Model with Multiple Random Effects." Discussion Paper 1016, IZA/Bonn and University of Zurich, 2002.

Winkelmann, R. *Econometric Analysis of Count Data.* 4th ed., Heidelberg: Springer-Verlag, 2003.

Winkelmann, R. "Subjective Well-being and the Family: Results from an Ordered Probit Model with Multiple Random Effects." *Empirical Economics*, 30, 3, 2005, pp. 749–761.

Winship, C., and R. Mare. "Regression Models with Ordered Variables." *American Sociological Review*, 49, 1984, pp. 512– 525.

Witte, A. "Estimating an Economic Model of Crime with Individual Data." *Quarterly Journal of Economics,* 94, 1980, pp. 57–84.

Wooldridge, J. "Selection Corrections for Panel Data Models Under Conditional Mean Assumptions." *Journal of Econometrics,* 68, 1995, pp. 115–132.

Wooldridge, J. "Asymptotic Properties of Weighted *M* Estimators for Variable Probability Samples." *Econometrica,* 67, 1999, pp. 1385–1406.

Wooldridge, J. "Inverse Probability Weighted *M*-Estimators for Sample Stratification,

Attrition and Stratification." *Portuguese Economic Journal*, 1, 2002, pp. 117–139.

Wooldridge, J. "Simple Solutions to the Initial Conditions Problem in Dynamic Nonlinear Panel Data Models with Unobserved Heterogeneity." CEMMAP Working paper CWP18/02, Centre for Microdata and Practice, IFS and University College, London, 2002c.

Wooldridge, J. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review*, 93, 2003, pp. 133–138.

Wooldridge, J. "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity." *Journal of Applied Econometrics*, 20, 1, 2005, pp. 39–54.

Wooldridge, J. *Econometric Analysis of Cross Section and Panel Data: Solutions Manual*, MIT Press, Cambridge, 2010.

Working, E. "What Do Statistical Demand Curves Show?" *Quarterly Journal of Economics,* 41, 1926, pp. 212–235.

World Health Organization. *The World Health Report, 2000, Health Systems: Improving Performance*. Geneva. 2000.

Wright, J. "Forecasting U.S. Inflation by Bayesian Model Averaging." Board of Governors, Federal Reserve System, International Finance Discussion Papers Number 780, 2003.

Wu, D. "Alternative Tests of Independence Between Stochastic Regressors and Disturbances." *Econometrica,* 41, 1973, pp. 733–750.

Wynand, P., and B. van Praag. "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection." *Journal of Econometrics,* 17, 1981, pp. 229–252.

Yatchew, A. "Nonparametric Regression Techniques in Econometrics." *Journal of Econometric Literature,* 36, 1998, pp. 669–721.

Yatchew, A. "An Elementary Estimator of the Partial Linear Model." *Economics Letters,* 57, 1997, pp. 135–143.

Yatchew, A. "Scale Economies in Electricity Distribution." *Journal of Applied Econometrics,* 15, 2, 2000, pp. 187–210.

Yatchew, A., and Z. Griliches. "Specification Error in Probit Models." *Review of Economics and Statistics,* 66, 1984, pp. 134–139.

Zabel, J. "Estimating Fixed and Random Effects Models with Selectivity." *Economics Letters,* 40, 1992, pp. 269–272.

Zaninotto, P., and E. Falischetti. "Comparison of Methods for Modelling a Count Outcome with Excess Zeros: Application to Activities of Daily Living (ADLs)." *Journal of Epidemiology and Community Health*, 65, 3, 2011.

Zarembka, P. "Transformations of Variables in Econometrics." In *Frontiers in Econometrics,* P. Zarembka, ed., Boston: Academic Press, 1974.

Zavoina, R., and W. McKelvey. "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." *Journal of Mathematical Sociology,* Summer, 1975, pp. 103–120.

Zellner, A. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias." *Journal of the American Statistical Association,* 57, 1962, pp. 500–509.

Zellner, A. "Estimators for Seemingly Unrelated Regression Equations: Some Finite Sample Results." Journal of the American Statistical Association, 57, 1963, pp. 977–992.

Zellner, A. *Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons, 1971.

Zellner, A., and D. Huang. "Further Properties of Efficient Estimators for Seemingly Unrelated Regression Equations." *International Economic Review,* 3, 1962, pp. 300–313.

Zellner, A., J. Kmenta, and J. Dreze. "Specification and Estimation of Cobb-Douglas Production Functions." *Econometrica,* 34, 1966, pp. 784–795.

Zellner, A., and N. Revankar. "Generalized Production Functions." *Review of Economic Studies,* 37, 1970, pp. 241– 250.

Zellner, A., and A. Siow. "Posterior Odds Ratios for Selected Regression Hypotheses (with Discussion)." In *Bayesian Statistics,* edited by J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, Valencia, Spain: University Press, 1980.

Zellner, A., and H. Theil. "Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations." *Econometrica,* 30, 1962, pp. 63–68.

Zigante, V. "Ever Rising Expectations—The Determinants of Subjective Welfare in Croatia." School of Economics and Management, Lund University, masters thesis www.essays.se/about/Ordered+Probit+Model, 2007.

INDEX

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z