Econometrics

Michael Creel

Department of Economics and Economic History

Universitat Autònoma de Barcelona

February 2014

Contents

List of Figures

List of Tables

Chapter 1

About this document

1.1 Prerequisites

These notes have been prepared under the assumption that the reader understands basic statistics, linear algebra, and mathematical optimization. There are many sources for this material, one are the appendices to *Introductory Econometrics: A Modern Approach* by Jeffrey Wooldridge. It is the student's resposibility to get up to speed on this material, it will not be covered in class

This document integrates lecture notes for a one year graduate level course with computer programs that illustrate and apply the methods that are studied. The immediate availability of executable (and modifiable) example programs when using the PDF version of the document is a distinguishing feature of these notes. If printed, the document is a somewhat terse approximation to a textbook. These notes are not intended to be a perfect substitute for a printed textbook. If you are a student of mine, please note that last sentence carefully. There are many good textbooks available. Students taking my courses should read the appropriate sections from at least one of the following books (or other textbooks with

similar level and content)

- Cameron, A.C. and P.K. Trivedi, *Microeconometrics - Methods and Applications*
- Davidson, R. and J.G. MacKinnon, *Econometric Theory and Methods*
- Gallant, A.R., *An Introduction to Econometric Theory*
- Hamilton, J.D., *Time Series Analysis*
- Hayashi, F., *Econometrics*

A more introductory-level reference is *Introductory Econometrics: A Modern Approach* by Jeffrey Wooldridge.

1.2 Contents

With respect to contents, the emphasis is on estimation and inference within the world of stationary data. If you take a moment to read the licensing information in the next section, you'll see that you are free to copy and modify the document. If anyone would like to contribute material that expands the contents, it would be very welcome. Error corrections and other additions are also welcome.

The integrated examples (they are on-line here and the support files are here) are an important part of these notes. GNU Octave (www.octave.org) has been used for most of the example programs, which are scattered though the document. This choice is motivated by several factors. The first is the high quality of the Octave environment for doing applied econometrics. Octave is similar to the

commercial package Matlab®, and will run scripts for that language without modification[1]. The fundamental tools (manipulation of matrices, statistical functions, minimization, *etc.*) exist and are implemented in a way that make extending them fairly easy. Second, an advantage of free software is that you don't have to pay for it. This can be an important consideration if you are at a university with a tight budget or if need to run many copies, as can be the case if you do parallel computing (discussed in Chapter 23). Third, Octave runs on GNU/Linux, Windows and MacOS. Figure 1.1 shows a sample GNU/Linux work environment, with an Octave script being edited, and the results are visible in an embedded shell window. As of 2011, some examples are being added using Gretl, the Gnu Regression, Econometrics, and Time-Series Library. This is an easy to use program, available in a number of languages, and it comes with a lot of data ready to use. It runs on the major operating systems. As of 2012, I am increasingly trying to make examples run on Matlab, though the need for add-on toolboxes for tasks as simple as generating random numbers limits what can be done.

The main document was prepared using LyX (www.lyx.org). LyX is a free[2] "what you see is what you mean" word processor, basically working as a graphical frontend to LaTeX. It (with help from other applications) can export your work in LaTeX, HTML, PDF and several other forms. It will run on Linux, Windows, and MacOS systems. Figure 1.2 shows LyX editing this document.

[1] Matlab®is a trademark of The Mathworks, Inc. Octave will run pure Matlab scripts. If a Matlab script calls an extension, such as a toolbox function, then it is necessary to make a similar extension available to Octave. The examples discussed in this document call a number of functions, such as a BFGS minimizer, a program for ML estimation, etc. All of this code is provided with the examples, as well as on the PelicanHPC live CD image.

[2] "Free" is used in the sense of "freedom", but LyX is also free of charge (free as in "free beer").

Figure 1.1: Octave

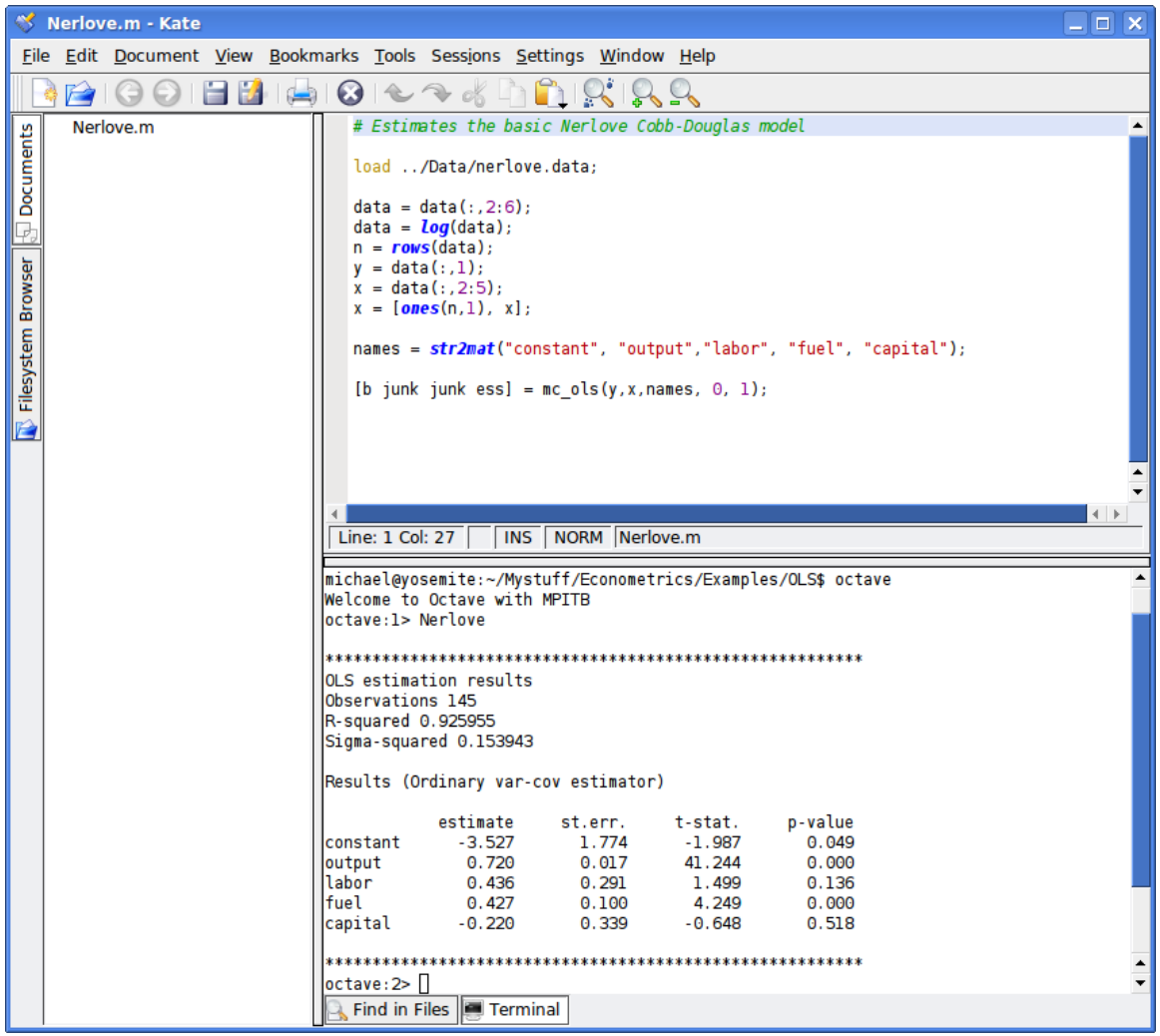
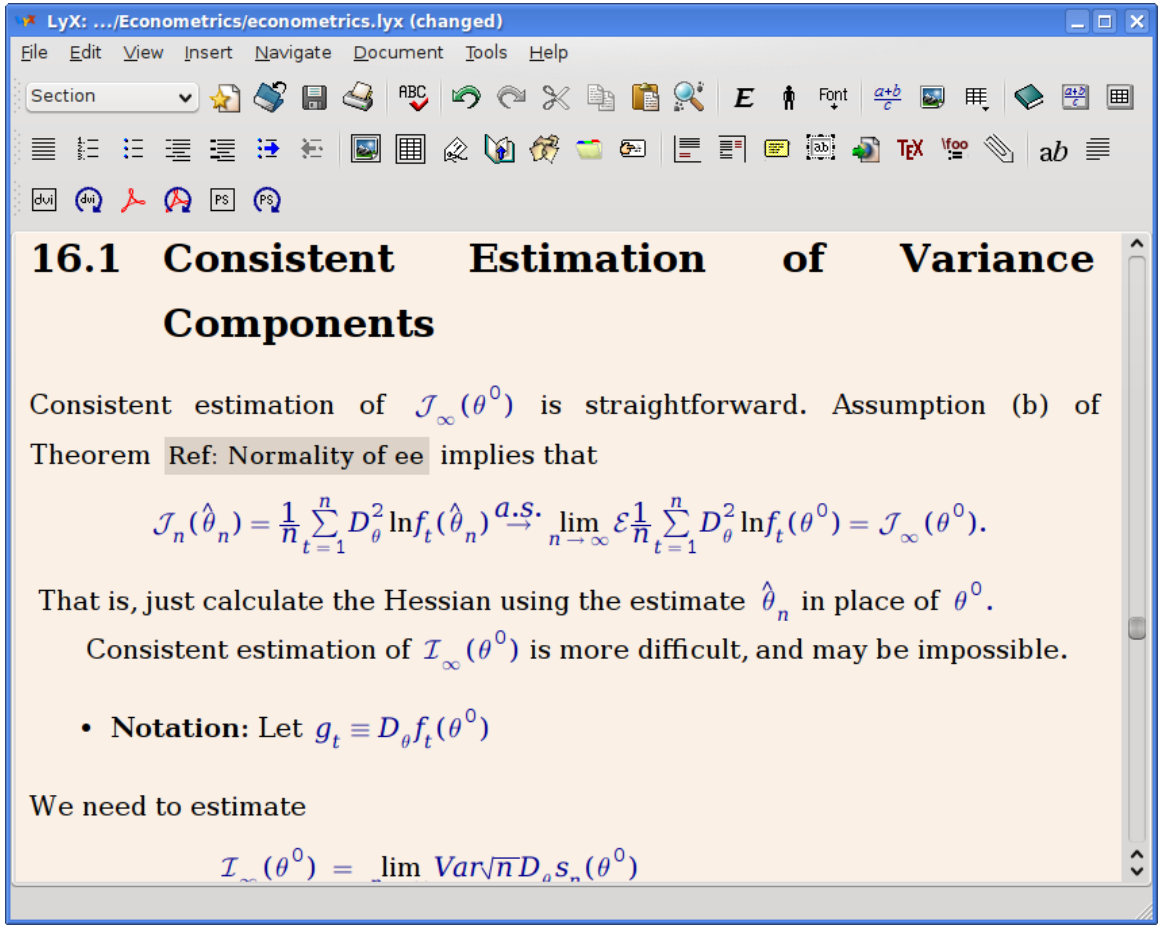


Figure 1.2: LyX



1.3 Licenses

All materials are copyrighted by Michael Creel with the date that appears above. They are provided under the terms of the GNU General Public License, ver. 2, which forms Section 26.1 of the notes, or, at your option, under the Creative Commons Attribution-Share Alike 2.5 license, which forms Section 26.2 of the notes. The main thing you need to know is that you are free to modify and distribute these materials in any way you like, as long as you share your contributions in the same way the materials are made available to you. In particular, you must make available the source files, in editable form, for your modified version of the materials.

1.4 Obtaining the materials

The materials are available on my web page. In addition to the final product, which you're probably looking at in some form now, you can obtain the editable LyX sources, which will allow you to create your own version, if you like, or send error corrections and contributions.

1.5 An easy way run the examples

Octave is available from the Octave home page, www.octave.org. Also, some updated links to packages for Windows and MacOS are at http://www.dynare.org/download/octave. The example programs are available as links to files on my web page in the PDF version, and here. Support files needed to run these are available here. The files won't run properly from your browser, since there are dependencies between files - they are only illustrative when browsing. To see how to use these files (edit and run

them), you should go to the home page of this document, since you will probably want to download the pdf version together with all the support files and examples. Then set the base URL of the PDF file to point to wherever the Octave files are installed. Then you need to install Octave and the support files. All of this may sound a bit complicated, because it is. An easier solution is available:

The Linux OS image file econometrics.iso an ISO image file that may be copied to USB or burnt to CDROM. It contains a bootable-from-CD or USB GNU/Linux system. These notes, in source form and as a PDF, together with all of the examples and the software needed to run them are available on econometrics.iso. I recommend starting off by using virtualization, to run the Linux system with all of the materials inside of a virtual computer, while still running your normal operating system. Various virtualization platforms are available. I recommend Virtualbox [3], which runs on Windows, Linux, and Mac OS.

[3] Virtualbox is free software (GPL v2). That, and the fact that it works very well, is the reason it is recommended here. There are a number of similar products available. It is possible to run PelicanHPC as a virtual machine, and to communicate with the installed operating system using a private network. Learning how to do this is not too difficult, and it is very convenient.

Chapter 2

Introduction: Economic and econometric models

Here's some data: 100 observations on 3 economic variables. Let's do some exploratory analysis using Gretl:

- histograms
- correlations
- x-y scatterplots

So, what can we say? Correlations? Yes. Causality? Who knows? This is economic data, generated by economic agents, following their own beliefs, technologies and preferences. It is not experimental data generated under controlled conditions. How can we determine causality if we don't have experimental data?

Without a model, we can't distinguish correlation from causality. It turns out that the variables we're looking at are QUANTITY (q), PRICE (p), and INCOME (m). Economic theory tells us that the quantity of a good that consumers will puchase (the demand function) is something like:

$$q = f(p, m, z)$$

- q is the quantity demanded
- p is the price of the good
- m is income
- z is a vector of other variables that may affect demand

The supply of the good to the market is the aggregation of the firms' supply functions. The market supply function is something like

$$q = g(p, z)$$

Suppose we have a sample consisting of a number of observations on q p and m at different time periods $t = 1, 2, ..., n$. Supply and demand in each period is

$$q_t = f(p_t, m_t, z_t)$$
$$q_t = g(p_t, z_t)$$

(draw some graphs showing roles of m and z)

This is the basic economic model of supply and demand: q and p are determined in the market equilibrium, given by the intersection of the two curves. These two variables are determined jointly by

the model, and are the *endogenous variables*. Income (m) is not determined by this model, its value is determined independently of q and p by some other process. m is an *exogenous variable*. So, m causes q, though the demand function. Because q and p are jointly determined, m also causes p. p and q do not cause m, according to this theoretical model. q and p have a joint causal relationship.

- Economic theory can help us to determine the causality relationships between correlated variables.

- If we had experimental data, we could control certain variables and observe the outcomes for other variables. If we see that variable x changes as the controlled value of variable y is changed, then we know that y causes x. With economic data, we are unable to control the values of the variables: for example in supply and demand, if price changes, then quantity changes, but quantity also affect price. We can't control the market price, because the market price changes as quantity adjusts. This is the reason we need a theoretical model to help us distinguish correlation and causality.

The model is essentially a theoretical construct up to now:

- We don't know the forms of the functions f and g.

- Some components of z_t may not be observable. For example, people don't eat the same lunch every day, and you can't tell what they will order just by looking at them. There are unobservable components to supply and demand, and we can model them as random variables. Suppose we can break z_t into two unobservable components ε_{t1} and ϵ_{t2}.

An econometric model attempts to quantify the relationship more precisely. A step toward an estimable econometric model is to suppose that the model may be written as

$$q_t = \alpha_1 + \alpha_2 p_t + \alpha_3 m_t + \varepsilon_{t1}$$
$$q_t = \beta_1 + \beta_2 p_t + \varepsilon_{t1}$$

We have imposed a number of restrictions on the theoretical model:

- The functions f and g have been specified to be linear functions
- The parameters (α_1, β_2, etc.) are constant over time.
- There is a single unobservable component in each equation, and we assume it is additive.

If we assume nothing about the error terms ϵ_{t1} and ϵ_{t2}, we can always write the last two equations, as the errors simply make up the difference between the true demand and supply functions and the assumed forms. But in order for the β coefficients to exist in a sense that has economic meaning, and in order to be able to use sample data to make reliable inferences about their values, we need to make additional assumptions. Such assumptions might be something like:

- $E(\epsilon_{tj}) = 0$, $j = 1, 2$
- $E(p_t \epsilon_{tj}) = 0$, $j = 1, 2$
- $E(m_t \epsilon_{tj}) = 0$, $j = 1, 2$

These are assertions that the errors are uncorrelated with the variables, and such assertions may or may not be reasonable. Later we will see how such assumption may be used and/or tested.

All of the last six bulleted points have **no theoretical basis**, in that the theory of supply and demand doesn't imply these conditions. The validity of any results we obtain using this model will be contingent on these additional restrictions being at least approximately correct. For this reason, *specification testing* will be needed, to check that the model seems to be reasonable. Only when we are convinced that the model is at least approximately correct should we use it for economic analysis.

When testing a hypothesis using an econometric model, at least three factors can cause a statistical test to reject the null hypothesis:

1. the hypothesis is false
2. a type I error has occured
3. the econometric model is not correctly specified, and thus the test does not have the assumed distribution

To be able to make scientific progress, we would like to ensure that the third reason is not contributing in a major way to rejections, so that rejection will be most likely due to either the first or second reasons. Hopefully the above example makes it clear that econometric models are necessarily more detailed than what we can obtain from economic theory, and that this additional detail introduces many possible sources of misspecification of econometric models. In the next few sections we will obtain results supposing that the econometric model is entirely correctly specified. Later we will examine the consequences of misspecification and see some methods for determining if a model is correctly specified. Later on, econometric methods that seek to minimize maintained assumptions are introduced.

Chapter 3

Ordinary Least Squares

3.1 The Linear Model

Consider approximating a variable y using the variables $x_1, x_2, ..., x_k$. We can consider a model that is a linear approximation:

Linearity: the model is a linear function of the parameter vector β^0 :

$$y = \beta_1^0 x_1 + \beta_2^0 x_2 + ... + \beta_k^0 x_k + \epsilon$$

or, using vector notation:

$$y = \mathbf{x}'\beta^0 + \epsilon$$

The dependent variable y is a scalar random variable, $\mathbf{x} = (\ x_1 \ \ x_2 \ \ \cdots \ \ x_k)'$ is a k-vector of explanatory variables, and $\beta^0 = (\ \beta_1^0 \ \ \beta_2^0 \ \ \cdots \ \ \beta_k^0)'$. The superscript "0" in β^0 means this is the "true value" of the unknown parameter. It will be defined more precisely later, and usually suppressed when it's

not necessary for clarity.

Suppose that we want to use data to try to determine the best linear approximation to y using the variables $\mathbf{x}$. The data $\{(y_t, \mathbf{x}_t)\}, t = 1, 2, ..., n$ are obtained by some form of sampling[1]. An individual observation is

$$y_t = \mathbf{x}_t'\beta + \varepsilon_t$$

The n observations can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{3.1}$$

where $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}'$ is $n \times 1$ and $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix}'$.

Linear models are more general than they might first appear, since one can employ nonlinear transformations of the variables:

$$\varphi_0(z) = \begin{bmatrix} \varphi_1(w) & \varphi_2(w) & \cdots & \varphi_p(w) \end{bmatrix} \beta + \varepsilon$$

where the $\phi_i()$ are known functions. Defining $y = \varphi_0(z)$, $x_1 = \varphi_1(w)$, *etc.* leads to a model in the form of equation 3.4. For example, the Cobb-Douglas model

$$z = A w_2^{\beta_2} w_3^{\beta_3} \exp(\varepsilon)$$

can be transformed logarithmically to obtain

$$\ln z = \ln A + \beta_2 \ln w_2 + \beta_3 \ln w_3 + \varepsilon.$$

[1]For example, cross-sectional data may be obtained by random sampling. Time series data accumulate historically.

If we define $y = \ln z$, $\beta_1 = \ln A$, *etc.*, we can put the model in the form needed. The approximation is linear in the parameters, but not necessarily linear in the variables.

3.2 Estimation by least squares

Figure 3.1, obtained by running TypicalData.m shows some data that follows the linear model $y_t = \beta_1 + \beta_2 x_{t2} + \epsilon_t$. The green line is the "true" regression line $\beta_1 + \beta_2 x_{t2}$, and the red crosses are the data points (x_{t2}, y_t), where ϵ_t is a random error that has mean zero and is independent of x_{t2}. Exactly how the green line is defined will become clear later. In practice, we only have the data, and we don't know where the green line lies. We need to gain information about the straight line that best fits the data points.

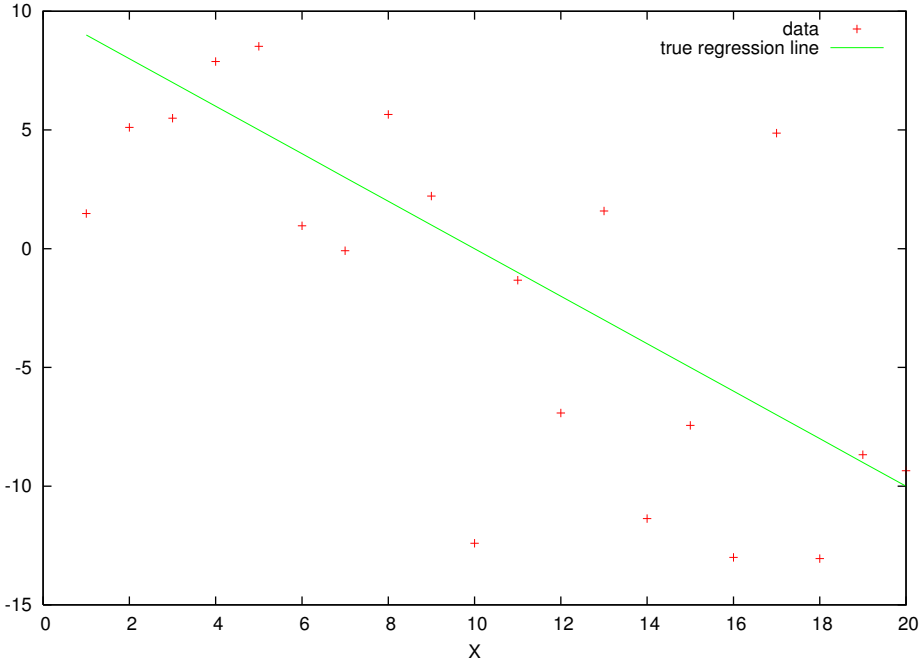
The *ordinary least squares* (OLS) estimator is defined as the value that minimizes the sum of the squared errors:

$$\hat{\beta} = \arg\min s(\beta)$$

where

$$\begin{aligned} s(\beta) &= \sum_{t=1}^{n} (y_t - \mathbf{x}_t'\beta)^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ &= \| \mathbf{y} - \mathbf{X}\beta \|^2 \end{aligned} \tag{3.2}$$

Figure 3.1: Typical data, Classical Model



This last expression makes it clear how the OLS estimator is defined: it minimizes the Euclidean distance between y and $X\beta$. The fitted OLS coefficients are those that give the best linear approximation to y using $\mathbf{x}$ as basis functions, where "best" means minimum Euclidean distance. One could think of other estimators based upon other metrics. For example, the *minimum absolute distance* (MAD) minimizes $\sum_{t=1}^{n} |y_t - \mathbf{x}_t'\beta|$. Later, we will see that which estimator is best in terms of their statistical properties, rather than in terms of the metrics that define them, depends upon the properties of ϵ, about which we have as yet made no assumptions.

- To minimize the criterion $s(\beta)$, find the derivative with respect to β:

$$D_\beta s(\beta) = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta$$

  Then setting it to zeros gives

$$D_\beta s(\hat{\beta}) = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \equiv 0$$

  so

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- To verify that this is a minimum, check the second order sufficient condition:

$$D_\beta^2 s(\hat{\beta}) = 2\mathbf{X}'\mathbf{X}$$

  Since $\rho(\mathbf{X}) = K$, this matrix is positive definite, since it's a quadratic form in a p.d. matrix (identity matrix of order n), so $\hat{\beta}$ is in fact a minimizer.

- The *fitted values* are the vector $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$.
- The *residuals* are the vector $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$
- Note that

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \varepsilon \\ &= \mathbf{X}\hat{\beta} + \hat{\varepsilon} \end{aligned}$$

- Also, the first order conditions can be written as

$$\begin{aligned} \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta} &= 0 \\ \mathbf{X}'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) &= 0 \\ \mathbf{X}'\hat{\varepsilon} &= 0 \end{aligned}$$

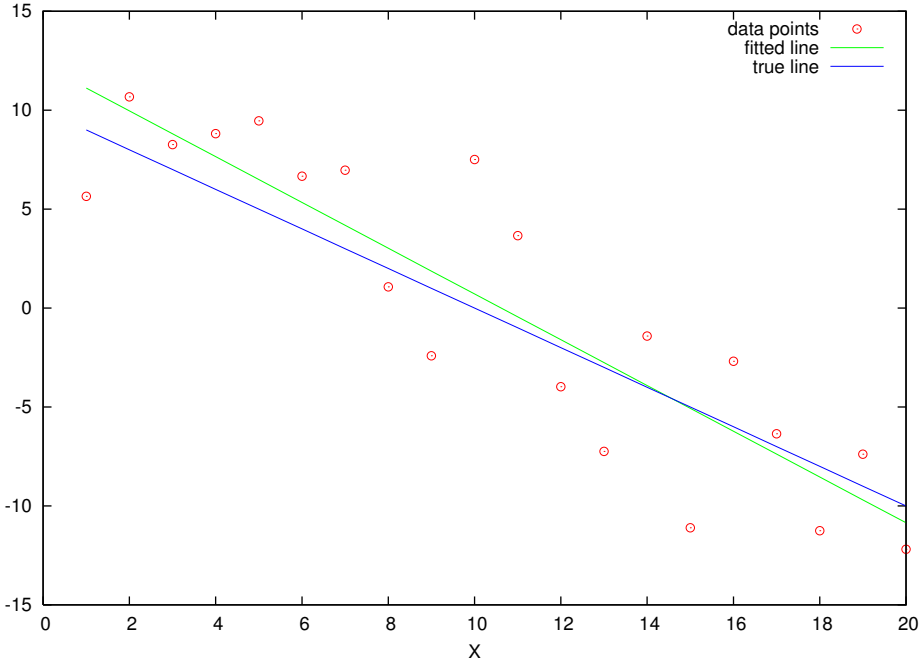
  which is to say, the OLS residuals are orthogonal to $\mathbf{X}$. Let's look at this more carefully.

3.3 Geometric interpretation of least squares estimation

In X, Y Space

Figure 3.2 shows a typical fit to data, along with the true regression line. Note that the true line and the estimated line are different. This figure was created by running the Octave program OlsFit.m . You can experiment with changing the parameter values to see how this affects the fit, and to see how the fitted line will sometimes be close to the true line, and sometimes rather far away.

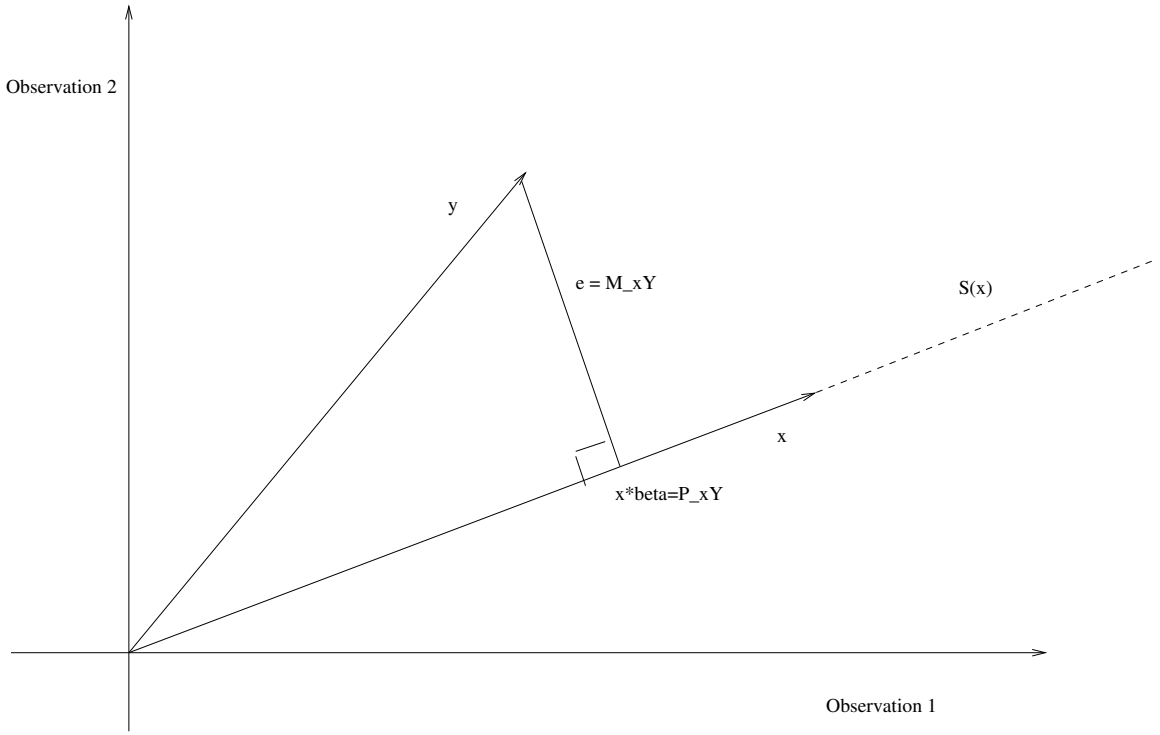
Figure 3.2: Example OLS Fit



In Observation Space

If we want to plot in observation space, we'll need to use only two or three observations, or we'll encounter some limitations of the blackboard. If we try to use 3, we'll encounter the limits of my artistic ability, so let's use two. With only two observations, we can't have $K > 1$.

Figure 3.3: The fit in observation space



- We can decompose y into two components: the orthogonal projection onto the $K-$dimensional space spanned by X, $X\hat{\beta}$, and the component that is the orthogonal projection onto the $n - K$ subpace that is orthogonal to the span of X, $\hat{\varepsilon}$.

- Since $\hat{\beta}$ is chosen to make $\hat{\varepsilon}$ as short as possible, $\hat{\varepsilon}$ will be orthogonal to the space spanned by X. Since X is in this space, $X'\hat{\varepsilon} = 0$. Note that the f.o.c. that define the least squares estimator imply that this is so.

Projection Matrices

$X\hat{\beta}$ is the projection of y onto the span of X, or

$$X\hat{\beta} = X\left(X'X\right)^{-1}X'y$$

Therefore, the matrix that projects y onto the span of X is

$$P_X = X(X'X)^{-1}X'$$

since

$$X\hat{\beta} = P_X y.$$

$\hat{\varepsilon}$ is the projection of y onto the $N - K$ dimensional space that is orthogonal to the span of X. We have that

$$\begin{aligned}\hat{\varepsilon} &= y - X\hat{\beta} \\ &= y - X(X'X)^{-1}X'y \\ &= \left[I_n - X(X'X)^{-1}X'\right]y.\end{aligned}$$

So the matrix that projects y onto the space orthogonal to the span of X is

$$\begin{aligned} M_X &= I_n - X(X'X)^{-1}X' \\ &= I_n - P_X. \end{aligned}$$

We have

$$\hat{\varepsilon} = M_X y.$$

Therefore

$$\begin{aligned} y &= P_X y + M_X y \\ &= X\hat{\beta} + \hat{\varepsilon}. \end{aligned}$$

These two projection matrices decompose the n dimensional vector y into two orthogonal components - the portion that lies in the K dimensional space defined by X, and the portion that lies in the orthogonal $n - K$ dimensional space.

- Note that both P_X and M_X are *symmetric* and *idempotent*.
  - A symmetric matrix A is one such that $A = A'$.
  - An idempotent matrix A is one such that $A = AA$.
  - The only nonsingular idempotent matrix is the identity matrix.

3.4 Influential observations and outliers

The OLS estimator of the i^{th} element of the vector β_0 is simply

$$\begin{aligned} \hat{\beta}_i &= \left[(X'X)^{-1}X'\right]_{i\cdot} y \\ &= c_i' y \end{aligned}$$

This is how we define a linear estimator - it's a linear function of the dependent variable. Since it's a linear combination of the observations on the dependent variable, where the weights are determined by the observations on the regressors, some observations may have more influence than others.

To investigate this, let e_t be an n vector of zeros with a 1 in the tth position, *i.e.,* it's the tth column of the matrix I_n. Define

$$\begin{aligned} h_t &= (P_X)_{tt} \\ &= e_t' P_X e_t \end{aligned}$$

so h_t is the tth element on the main diagonal of P_X. Note that

$$h_t = \| P_X e_t \|^2$$

so

$$h_t \leq \| e_t \|^2 = 1$$

So $0 < h_t < 1$. Also,

$$Tr P_X = K \Rightarrow \overline{h} = K/n.$$

So the average of the h_t is K/n. The value h_t is referred to as the *leverage* of the observation. If the leverage is much higher than average, the observation has the potential to affect the OLS fit importantly. However, an observation may also be influential due to the value of y_t, rather than the weight it is multiplied by, which only depends on the x_t's.

To account for this, consider estimation of β without using the t^{th} observation (designate this estimator as $\hat{\beta}^{(t)}$). One can show (see Davidson and MacKinnon, pp. 32-5 for proof) that

$$\hat{\beta}^{(t)} = \hat{\beta} - \left(\frac{1}{1-h_t}\right)(X'X)^{-1}X_t'\hat{\varepsilon}_t$$

so the change in the t^{th} observations fitted value is

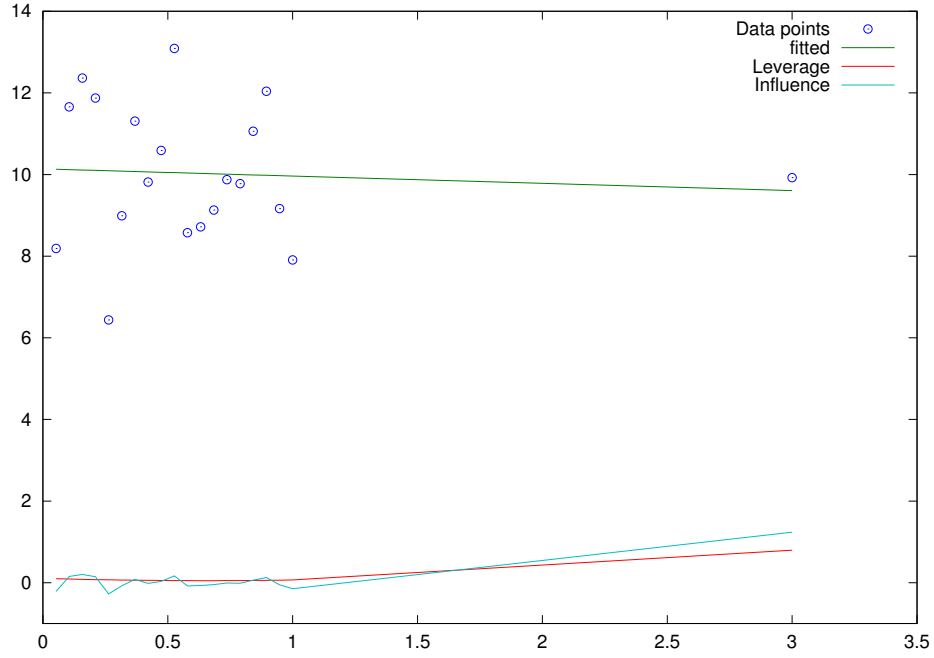
$$\mathbf{x}_t'\hat{\beta} - \mathbf{x}_t'\hat{\beta}^{(t)} = \left(\frac{h_t}{1-h_t}\right)\hat{\varepsilon}_t$$

While an observation may be influential if it doesn't affect its own fitted value, it certainly *is* influential if it does. A fast means of identifying influential observations is to plot $\left(\frac{h_t}{1-h_t}\right)\hat{\varepsilon}_t$ (which I will refer to as the *own influence* of the observation) as a function of t. Figure 3.4 gives an example plot of data, fit, leverage and influence. The Octave program is InfluentialObservation.m. **(note to self when lecturing: load the data ../OLS/influencedata into Gretl and reproduce this).** If you re-run the program you will see that the leverage of the last observation (an outlying value of x) is always high, and the influence is sometimes high.

After influential observations are detected, one needs to determine *why* they are influential. Possible causes include:

- data entry error, which can easily be corrected once detected. Data entry errors *are very common.*

Figure 3.4: Detection of influential observations



- special economic factors that affect some observations. These would need to be identified and incorporated in the model. This is the idea behind *structural change*: the parameters may not be constant across all observations.
- pure randomness may have caused us to sample a low-probability observation.

There exist *robust* estimation methods that downweight outliers.

3.5 Goodness of fit

The fitted model is

$$y = X\hat{\beta} + \hat{\varepsilon}$$

Take the inner product:

$$y'y = \hat{\beta}'X'X\hat{\beta} + 2\hat{\beta}'X'\hat{\varepsilon} + \hat{\varepsilon}'\hat{\varepsilon}$$

But the middle term of the RHS is zero since $X'\hat{\varepsilon} = 0$, so

$$y'y = \hat{\beta}'X'X\hat{\beta} + \hat{\varepsilon}'\hat{\varepsilon} \tag{3.3}$$

The *uncentered* R_u^2 is defined as

$$\begin{aligned} R_u^2 &= 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y} \\ &= \frac{\hat{\beta}'X'X\hat{\beta}}{y'y} \\ &= \frac{\| P_X y \|^2}{\| y \|^2} \\ &= \cos^2(\phi), \end{aligned}$$

where ϕ is the angle between y and the span of X .

- The uncentered R^2 changes if we add a constant to y, since this changes ϕ (see Figure 3.5, the yellow vector is a constant, since it's on the 45 degree line in observation space). Another, more common definition measures the contribution of the variables, other than the constant term, to explaining the variation in y. Thus it measures the ability of the model to explain the variation of y about its unconditional sample mean.

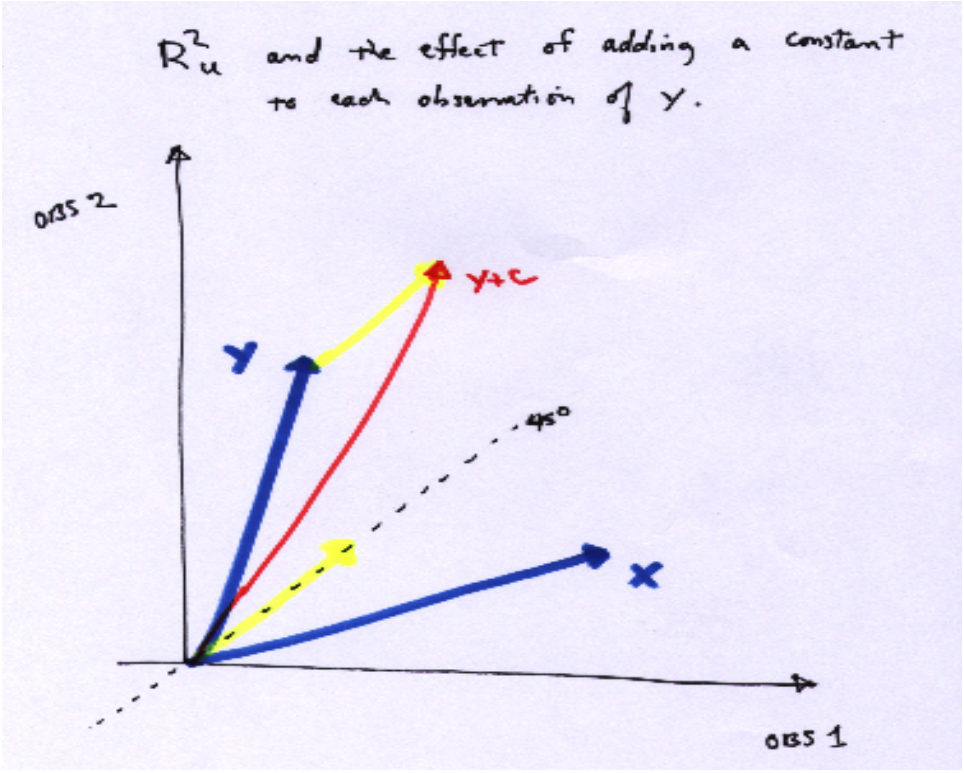
Let $\iota = (1, 1, ..., 1)'$, a n -vector. So

$$\begin{aligned} M_\iota &= I_n - \iota(\iota'\iota)^{-1}\iota' \\ &= I_n - \iota\iota'/n \end{aligned}$$

$M_\iota y$ just returns the vector of deviations from the mean. In terms of deviations from the mean, equation 3.3 becomes

$$y'M_\iota y = \hat{\beta}'X'M_\iota X\hat{\beta} + \hat{\varepsilon}'M_\iota\hat{\varepsilon}$$

Figure 3.5: Uncentered R^2



The *centered* R_c^2 is defined as

$$R_c^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'M_\iota y} = 1 - \frac{ESS}{TSS}$$

where $ESS = \hat{\varepsilon}'\hat{\varepsilon}$ and $TSS = y'M_\iota y = \sum_{t=1}^{n}(y_t - \bar{y})^2$.

Supposing that X contains a column of ones (*i.e.,* there is a constant term),

$$X'\hat{\varepsilon} = 0 \Rightarrow \sum_t \hat{\varepsilon}_t = 0$$

so $M_\iota\hat{\varepsilon} = \hat{\varepsilon}$. In this case

$$y'M_\iota y = \hat{\beta}'X'M_\iota X\hat{\beta} + \hat{\varepsilon}'\hat{\varepsilon}$$

So

$$R_c^2 = \frac{RSS}{TSS}$$

where $RSS = \hat{\beta}'X'M_\iota X\hat{\beta}$

- Supposing that a column of ones is in the space spanned by X ($P_X\iota = \iota$), then one can show that $0 \leq R_c^2 \leq 1$.

3.6 The classical linear regression model

Up to this point the model is empty of content beyond the definition of a best linear approximation to y and some geometrical properties. There is no economic content to the model, and the regression parameters have no economic interpretation. For example, what is the partial derivative of y with

respect to x_j? The linear approximation is

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

The partial derivative is

$$\frac{\partial y}{\partial x_j} = \beta_j + \frac{\partial \epsilon}{\partial x_j}$$

Up to now, there's no guarantee that $\frac{\partial \epsilon}{\partial x_j}$=0. For the β to have an economic meaning, we need to make additional assumptions. The assumptions that are appropriate to make depend on the data under consideration. We'll start with the classical linear regression model, which incorporates some assumptions that are clearly not realistic for economic data. This is to be able to explain some concepts with a minimum of confusion and notational clutter. Later we'll adapt the results to what we can get with more realistic assumptions.

Linearity: the model is a linear function of the parameter vector β^0 :

$$y = \beta_1^0 x_1 + \beta_2^0 x_2 + ... + \beta_k^0 x_k + \epsilon \tag{3.4}$$

or, using vector notation:

$$y = \mathbf{x}'\beta^0 + \epsilon$$

Nonstochastic linearly independent regressors: $\mathbf{X}$ is a fixed matrix of constants, it has rank K equal to its number of columns, and

$$\lim \frac{1}{n}\mathbf{X}'\mathbf{X} = Q_X \tag{3.5}$$

where Q_X is a finite positive definite matrix. This is needed to be able to identify the individual effects of the explanatory variables.

Independently and identically distributed errors:

$$\epsilon \sim IID(0, \sigma^2 I_n) \tag{3.6}$$

ε is jointly distributed IID. This implies the following two properties:

Homoscedastic errors:

$$V(\varepsilon_t) = \sigma_0^2, \forall t \tag{3.7}$$

Nonautocorrelated errors:

$$\mathcal{E}(\varepsilon_t \epsilon_s) = 0, \forall t \neq s \tag{3.8}$$

Optionally, we will sometimes assume that the errors are normally distributed.

Normally distributed errors:

$$\epsilon \sim N(0, \sigma^2 I_n) \tag{3.9}$$

3.7 Small sample statistical properties of the least squares estimator

Up to now, we have only examined numeric properties of the OLS estimator, that always hold. Now we will examine statistical properties. The statistical properties depend upon the assumptions we make.

Unbiasedness

We have $\hat{\beta} = (X'X)^{-1}X'y$. By linearity,

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon\end{aligned}$$

By 3.5 and 3.6

$$\begin{aligned}E(X'X)^{-1}X'\varepsilon &= E(X'X)^{-1}X'\varepsilon \\ &= (X'X)^{-1}X'E\varepsilon \\ &= 0\end{aligned}$$

so the OLS estimator is unbiased under the assumptions of the classical model.

Figure 3.6 shows the results of a small Monte Carlo experiment where the OLS estimator was calculated for 10000 samples from the classical model with $y = 1 + 2x + \varepsilon$, where $n = 20$, $\sigma_\varepsilon^2 = 9$, and x is fixed across samples. We can see that the β_2 appears to be estimated without bias. The program that generates the plot is Unbiased.m , if you would like to experiment with this.

With time series data, the OLS estimator will often be biased. Figure 3.7 shows the results of a small Monte Carlo experiment where the OLS estimator was calculated for 1000 samples from the AR(1) model with $y_t = 0 + 0.9y_{t-1} + \varepsilon_t$, where $n = 20$ and $\sigma_\varepsilon^2 = 1$. In this case, assumption 3.5 does not hold: the regressors are stochastic. We can see that the bias in the estimation of β_2 is about -0.2.

The program that generates the plot is Biased.m , if you would like to experiment with this.

Figure 3.6: Unbiasedness of OLS under classical assumptions

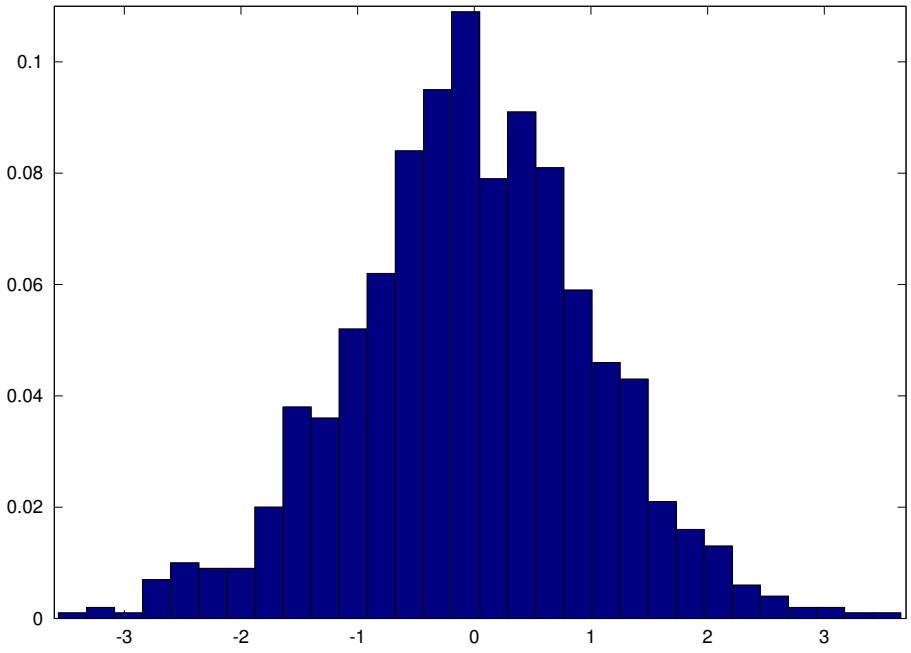
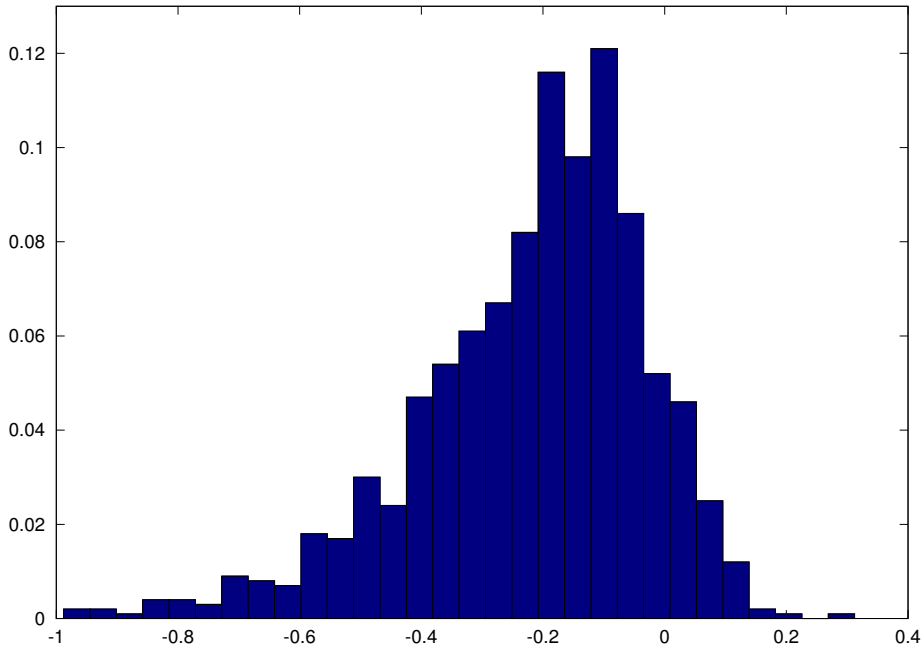


Figure 3.7: Biasedness of OLS when an assumption fails



Normality

With the linearity assumption, we have $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$. This is a linear function of ε. Adding the assumption of normality (3.9, which implies strong exogeneity), then

$$\hat{\beta} \sim N\left(\beta, (X'X)^{-1}\sigma_0^2\right)$$

since a linear function of a normal random vector is also normally distributed. In Figure 3.6 you can see that the estimator appears to be normally distributed. It in fact is normally distributed, since the DGP (see the Octave program) has normal errors. Even when the data may be taken to be IID, the assumption of normality is often questionable or simply untenable. For example, if the dependent variable is the number of automobile trips per week, it is a count variable with a discrete distribution, and is thus not normally distributed. Many variables in economics can take on only nonnegative values, which, strictly speaking, rules out normality.[2]

The variance of the OLS estimator and the Gauss-Markov theorem

Now let's make all the classical assumptions except the assumption of normality. We have $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$ and we know that $E(\hat{\beta}) = \beta$. So

$$\begin{aligned} Var(\hat{\beta}) &= E\left\{\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\right\} \\ &= E\left\{(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right\} \\ &= (X'X)^{-1}\sigma_0^2 \end{aligned}$$

[2]Normality may be a good model nonetheless, as long as the probability of a negative value occuring is negligable under the model. This depends upon the mean being large enough in relation to the variance.

The OLS estimator is a *linear estimator*, which means that it is a linear function of the dependent variable, y.

$$\begin{aligned}\hat{\beta} &= \left[(X'X)^{-1}X'\right]y \\ &= Cy\end{aligned}$$

where C is a function of the explanatory variables only, not the dependent variable. It is also *unbiased* under the present assumptions, as we proved above. One could consider other weights W that are a function of X that define some other linear estimator. We'll still insist upon unbiasedness. Consider $\tilde{\beta} = Wy$, where $W = W(X)$ is some $k \times n$ matrix function of X. Note that since W is a function of X, it is nonstochastic, too. If the estimator is unbiased, then we must have $WX = I_K$:

$$\begin{aligned}\mathcal{E}(Wy) &= \mathcal{E}(WX\beta_0 + W\varepsilon) \\ &= WX\beta_0 \\ &= \beta_0 \\ &\Rightarrow \\ WX &= I_K\end{aligned}$$

The variance of $\tilde{\beta}$ is

$$V(\tilde{\beta}) = WW'\sigma_0^2.$$

Define

$$D = W - (X'X)^{-1}X'$$

so

$$W = D + (X'X)^{-1}X'$$

Since $WX = I_K$, $DX = 0$, so

$$\begin{aligned} V(\tilde{\beta}) &= \left(D + (X'X)^{-1}X'\right)\left(D + (X'X)^{-1}X'\right)'\sigma_0^2 \\ &= \left(DD' + (X'X)^{-1}\right)\sigma_0^2 \end{aligned}$$

So

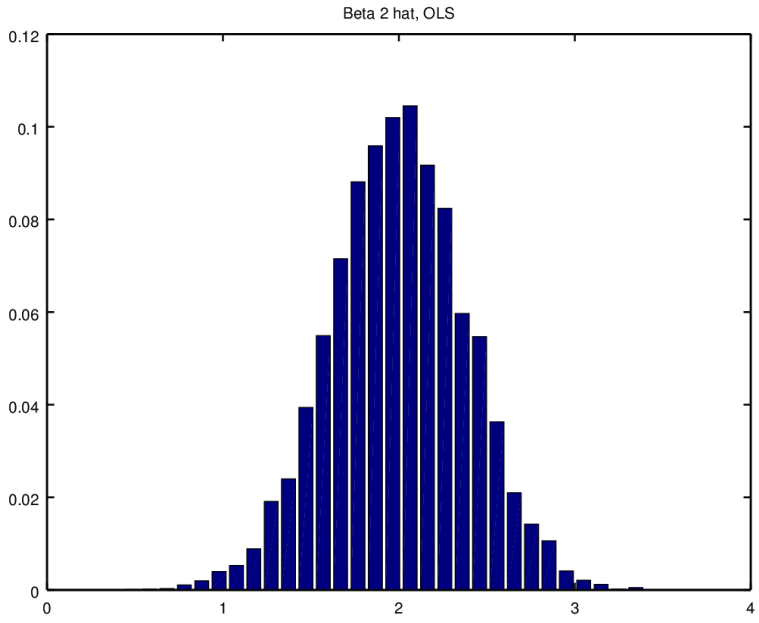
$$V(\tilde{\beta}) \geq V(\hat{\beta})$$

The inequality is a shorthand means of expressing, more formally, that $V(\tilde{\beta}) - V(\hat{\beta})$ is a positive semi-definite matrix. This is a proof of the Gauss-Markov Theorem. The OLS estimator is the "best linear unbiased estimator" (BLUE).

- It is worth emphasizing again that we have not used the normality assumption in any way to prove the Gauss-Markov theorem, so it is valid if the errors are not normally distributed, as long as the other assumptions hold.

To illustrate the Gauss-Markov result, consider the estimator that results from splitting the sample into p equally-sized parts, estimating using each part of the data separately by OLS, then averaging the p resulting estimators. You should be able to show that this estimator is unbiased, but inefficient with respect to the OLS estimator. The program Efficiency.m illustrates this using a small Monte Carlo experiment, which compares the OLS estimator and a 3-way split sample estimator. The data generating process follows the classical model, with $n = 21$. The true parameter value is $\beta = 2$. In Figures 3.8 and 3.9 we can see that the OLS estimator is more efficient, since the tails of its histogram

Figure 3.8: Gauss-Markov Result: The OLS estimator



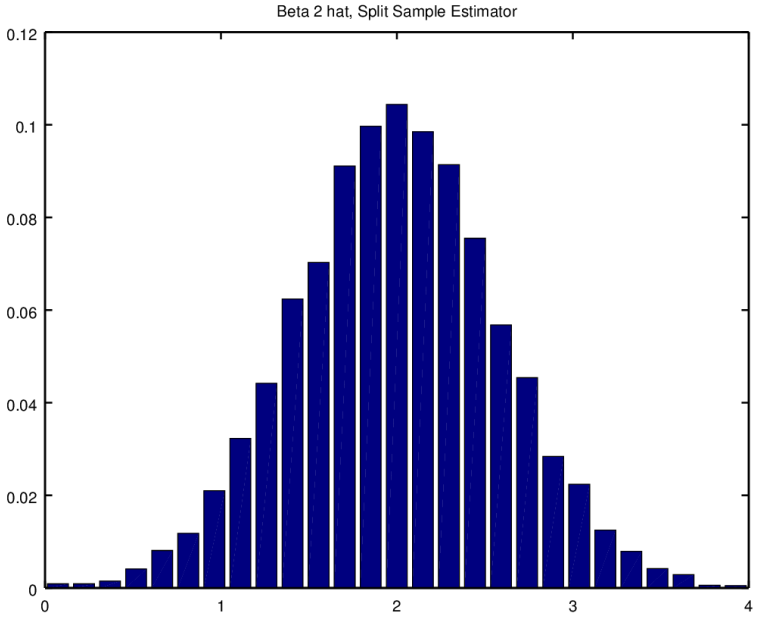
are more narrow.

We have that $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = \left(X^{'}X\right)^{-1}\sigma_0^2$, but we still need to estimate the variance of ϵ, σ_0^2, in order to have an idea of the precision of the estimates of β. A commonly used estimator of σ_0^2 is

$$\widehat{\sigma_0^2} = \frac{1}{n-K}\hat{\varepsilon}'\hat{\varepsilon}$$

This estimator is unbiased:

Figure 3.9: Gauss-Markov Resul: The split sample estimator



$$
\begin{aligned}
\widehat{\sigma_0^2} &= \frac{1}{n-K}\hat{\varepsilon}'\hat{\varepsilon} \\
&= \frac{1}{n-K}\varepsilon' M \varepsilon \\
\mathcal{E}(\widehat{\sigma_0^2}) &= \frac{1}{n-K}E(Tr\varepsilon' M \varepsilon) \\
&= \frac{1}{n-K}E(TrM\varepsilon\varepsilon') \\
&= \frac{1}{n-K}TrE(M\varepsilon\varepsilon') \\
&= \frac{1}{n-K}\sigma_0^2 TrM \\
&= \frac{1}{n-K}\sigma_0^2 (n-k) \\
&= \sigma_0^2
\end{aligned}
$$

where we use the fact that $Tr(AB) = Tr(BA)$ when both products are conformable. Thus, this estimator is also unbiased under these assumptions.

3.8 Example: The Nerlove model

Theoretical background

For a firm that takes input prices w and the output level q as given, the cost minimization problem is to choose the quantities of inputs x to solve the problem

$$\min_x w'x$$

subject to the restriction

$$f(x) = q.$$

The solution is the vector of factor demands $x(w, q)$. The *cost function* is obtained by substituting the factor demands into the criterion function:

$$Cw, q) = w'x(w, q).$$

- **Monotonicity** Increasing factor prices cannot decrease cost, so

$$\frac{\partial C(w, q)}{\partial w} \geq 0$$

  Remember that these derivatives give the conditional factor demands (Shephard's Lemma).

- **Homogeneity** The cost function is homogeneous of degree 1 in input prices: $C(tw, q) = tC(w, q)$ where t is a scalar constant. This is because the factor demands are homogeneous of degree zero in factor prices - they only depend upon relative prices.

- **Returns to scale** The *returns to scale* parameter γ is defined as the inverse of the elasticity of cost with respect to output:

$$\gamma = \left(\frac{\partial C(w, q)}{\partial q} \frac{q}{C(w, q)} \right)^{-1}$$

Constant returns to scale is the case where increasing production q implies that cost increases in the proportion 1:1. If this is the case, then $\gamma = 1$.

Cobb-Douglas functional form

The Cobb-Douglas functional form is linear in the logarithms of the regressors and the dependent variable. For a cost function, if there are g factors, the Cobb-Douglas cost function has the form

$$C = Aw_1^{\beta_1}...w_g^{\beta_g}q^{\beta_q}e^{\varepsilon}$$

What is the elasticity of C with respect to w_j?

$$\begin{aligned} e_{w_j}^C &= \left(\frac{\partial C}{\partial_{W_J}}\right)\left(\frac{w_j}{C}\right) \\ &= \beta_j A w_1^{\beta_1}.w_j^{\beta_j-1}..w_g^{\beta_g}q^{\beta_q}e^{\varepsilon}\frac{w_j}{Aw_1^{\beta_1}...w_g^{\beta_g}q^{\beta_q}e^{\varepsilon}} \\ &= \beta_j \end{aligned}$$

This is one of the reasons the Cobb-Douglas form is popular - the coefficients are easy to interpret, since they are the elasticities of the dependent variable with respect to the explanatory variable. Not that in this case,

$$\begin{aligned} e_{w_j}^C &= \left(\frac{\partial C}{\partial_{W_J}}\right)\left(\frac{w_j}{C}\right) \\ &= x_j(w,q)\frac{w_j}{C} \\ &\equiv s_j(w,q) \end{aligned}$$

the *cost share* of the j^{th} input. So with a Cobb-Douglas cost function, $\beta_j = s_j(w, q)$. The cost shares are constants.

Note that after a logarithmic transformation we obtain

$$\ln C = \alpha + \beta_1 \ln w_1 + ... + \beta_g \ln w_g + \beta_q \ln q + \epsilon$$

where $\alpha = \ln A$. So we see that the transformed model is linear in the logs of the data.

One can verify that the property of HOD1 implies that

$$\sum_{i=1}^{g} \beta_g = 1$$

In other words, the cost shares add up to 1.

The hypothesis that the technology exhibits CRTS implies that

$$\gamma = \frac{1}{\beta_q} = 1$$

so $\beta_q = 1$. Likewise, monotonicity implies that the coefficients $\beta_i \geq 0, i = 1, ..., g$.

The Nerlove data and OLS

The file nerlove.data contains data on 145 electric utility companies' cost of production, output and input prices. The data are for the U.S., and were collected by M. Nerlove. The observations are by row, and the columns are **COMPANY, COST** (C)**, OUTPUT** (Q), **PRICE OF LABOR** (P_L), **PRICE OF FUEL** (P_F) and **PRICE OF CAPITAL** (P_K). Note that the data are sorted by output

level (the third column).

We will estimate the Cobb-Douglas model

$$\ln C = \beta_1 + \beta_2 \ln Q + \beta_3 \ln P_L + \beta_4 \ln P_F + \beta_5 \ln P_K + \epsilon \tag{3.10}$$

using OLS. To do this yourself, you need the data file mentioned above, as well as Nerlove.m (the estimation program), and the library of Octave functions mentioned in the introduction to Octave that forms section 24 of this document.[3]

The results are

```
*********************************************************
OLS estimation results
Observations 145
R-squared 0.925955
Sigma-squared 0.153943

Results (Ordinary var-cov estimator)

            estimate      st.err.      t-stat.      p-value
constant      -3.527        1.774       -1.987        0.049
output         0.720        0.017       41.244        0.000
labor          0.436        0.291        1.499        0.136
fuel           0.427        0.100        4.249        0.000
capital       -0.220        0.339       -0.648        0.518

*********************************************************
```

[3] If you are running the bootable CD, you have all of this installed and ready to run.

- Do the theoretical restrictions hold?
- Does the model fit well?
- What do you think about RTS?

While we will most often use Octave programs as examples in this document, since following the programming statements is a useful way of learning how theory is put into practice, you may be interested in a more "user-friendly" environment for doing econometrics. I heartily recommend Gretl, the Gnu Regression, Econometrics, and Time-Series Library. This is an easy to use program, available in English, French, and Spanish, and it comes with a lot of data ready to use. It even has an option to save output as LaTeX fragments, so that I can just include the results into this document, no muss, no fuss. Here is the Nerlove data in the form of a GRETL data set: nerlove.gdt . Here the results of the Nerlove model from GRETL:

Model 2: OLS estimates using the 145 observations 1–145
Dependent variable: l_cost

Variable	Coefficient	Std. Error	t-statistic	p-value
const	−3.5265	1.77437	−1.9875	0.0488
l_output	0.720394	0.0174664	41.2445	0.0000
l_labor	0.436341	0.291048	1.4992	0.1361
l_fuel	0.426517	0.100369	4.2495	0.0000
l_capita	−0.219888	0.339429	−0.6478	0.5182

Mean of dependent variable	1.72466
S.D. of dependent variable	1.42172
Sum of squared residuals	21.5520
Standard error of residuals ($\hat{\sigma}$)	0.392356
Unadjusted R^2	0.925955
Adjusted $\bar{R}^2$	0.923840
$F(4, 140)$	437.686
Akaike information criterion	145.084
Schwarz Bayesian criterion	159.967

Fortunately, Gretl and my OLS program agree upon the results. Gretl is included in the bootable CD mentioned in the introduction. I recommend using GRETL to repeat the examples that are done using Octave.

The previous properties hold for finite sample sizes. Before considering the asymptotic properties of the OLS estimator it is useful to review the MLE estimator, since under the assumption of normal errors the two estimators coincide.

3.9 Exercises

1. Prove that the split sample estimator used to generate figure 3.9 is unbiased.

2. Calculate the OLS estimates of the Nerlove model using Octave and GRETL, and provide printouts of the results. Interpret the results.

3. Do an analysis of whether or not there are influential observations for OLS estimation of the

Nerlove model. Discuss.

4. Using GRETL, examine the residuals after OLS estimation and tell me whether or not you believe that the assumption of independent identically distributed normal errors is warranted. No need to do formal tests, just look at the plots. Print out any that you think are relevant, and interpret them.

5. For a random vector $X \sim N(\mu_x, \Sigma)$, what is the distribution of $AX + b$, where A and b are conformable matrices of constants?

6. Using Octave, write a little program that verifies that $Tr(AB) = Tr(BA)$ for A and B 4x4 matrices of random numbers. Note: there is an Octave function `trace`.

7. For the model with a constant and a single regressor, $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$, which satisfies the classical assumptions, prove that the variance of the OLS estimator declines to zero as the sample size increases.

Chapter 4

Asymptotic properties of the least squares estimator

The OLS estimator under the classical assumptions is BLUE[1], for all sample sizes. Now let's see what happens when the sample size tends to infinity.

[1]BLUE ≡ best linear unbiased estimator if I haven't defined it before

4.1 Consistency

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta + \varepsilon) \\
&= \beta_0 + (X'X)^{-1}X'\varepsilon \\
&= \beta_0 + \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}
\end{aligned}$$

Consider the last two terms. By assumption $\lim_{n\to\infty}\left(\frac{X'X}{n}\right) = Q_X \Rightarrow \lim_{n\to\infty}\left(\frac{X'X}{n}\right)^{-1} = Q_X^{-1}$, since the inverse of a nonsingular matrix is a continuous function of the elements of the matrix. Considering $\frac{X'\varepsilon}{n}$,

$$\frac{X'\varepsilon}{n} = \frac{1}{n}\sum_{t=1}^{n} x_t\varepsilon_t$$

Each $x_t\varepsilon_t$ has expectation zero, so

$$E\left(\frac{X'\varepsilon}{n}\right) = 0$$

The variance of each term is

$$V(x_t\epsilon_t) = x_t x_t'\sigma^2.$$

As long as these are finite, and given a technical condition[2], the Kolmogorov SLLN applies, so

$$\frac{1}{n}\sum_{t=1}^{n} x_t\varepsilon_t \xrightarrow{a.s.} 0.$$

This implies that

$$\hat{\beta} \xrightarrow{a.s.} \beta_0.$$

This is the property of *strong consistency:* the estimator converges in almost surely to the true value.

- The consistency proof does not use the normality assumption.
- Remember that almost sure convergence implies convergence in probability.

4.2 Asymptotic normality

We've seen that the OLS estimator is normally distributed *under the assumption of normal errors.* If the error distribution is unknown, we of course don't know the distribution of the estimator. However, we can get asymptotic results. *Assuming the distribution of* ε is unknown, but the the other classical assumptions hold:

[2]For application of LLN's and CLT's, of which there are very many to choose from, I'm going to avoid the technicalities. Basically, as long as terms that make up an average have finite variances and are not too strongly dependent, one will be able to find a LLN or CLT to apply. Which one it is doesn't matter, we only need the result.

$$\begin{aligned}
\hat{\beta} &= \beta_0 + (X'X)^{-1}X'\varepsilon \\
\hat{\beta} - \beta_0 &= (X'X)^{-1}X'\varepsilon \\
\sqrt{n}\left(\hat{\beta} - \beta_0\right) &= \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{\sqrt{n}}
\end{aligned}$$

- Now as before, $\left(\frac{X'X}{n}\right)^{-1} \to Q_X^{-1}$.
- Considering $\frac{X'\varepsilon}{\sqrt{n}}$, the limit of the variance is

$$\begin{aligned}
\lim_{n\to\infty} V\left(\frac{X'\varepsilon}{\sqrt{n}}\right) &= \lim_{n\to\infty} E\left(\frac{X'\epsilon\epsilon'X}{n}\right) \\
&= \sigma_0^2 Q_X
\end{aligned}$$

  The mean is of course zero. To get asymptotic normality, we need to apply a CLT. We assume one (for instance, the Lindeberg-Feller CLT) holds, so

$$\frac{X'\varepsilon}{\sqrt{n}} \xrightarrow{d} N\left(0, \sigma_0^2 Q_X\right)$$

  Therefore,

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \sigma_0^2 Q_X^{-1}\right) \tag{4.1}$$

- In summary, the OLS estimator is normally distributed in small and large samples if ε is normally distributed. If ε is not normally distributed, $\hat{\beta}$ is asymptotically normally distributed when a

CLT can be applied.

4.3 Asymptotic efficiency

The least squares objective function is

$$s(\beta) = \sum_{t=1}^{n} (y_t - x_t'\beta)^2$$

Supposing that ε is normally distributed, the model is

$$y = X\beta_0 + \varepsilon,$$

$$\begin{aligned} \varepsilon &\sim N(0, \sigma_0^2 I_n), \text{ so} \\ f(\varepsilon) &= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right) \end{aligned}$$

The joint density for y can be constructed using a change of variables. We have $\varepsilon = y - X\beta$, so $\frac{\partial \varepsilon}{\partial y'} = I_n$ and $|\frac{\partial \varepsilon}{\partial y'}| = 1$, so

$$f(y) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - x_t'\beta)^2}{2\sigma^2}\right).$$

Taking logs,

$$\ln L(\beta, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \sum_{t=1}^{n} \frac{(y_t - x_t'\beta)^2}{2\sigma^2}.$$

Maximizing this function with respect to β and σ gives what is known as the maximum likelihood (ML) estimator. It turns out that ML estimators are asymptotically efficient, a concept that will be explained in detail later. It's clear that the first order conditions for the MLE of β_0 are the same as the first order conditions that define the OLS estimator (up to multiplication by a constant), so the OLS estimator of β is also the ML estimator. *The estimators are the same, under the present assumptions.* Therefore, their properties are the same. *In particular, under the classical assumptions with normality, the OLS estimator $\hat{\beta}$ is asymptotically efficient.* Note that one needs to make an assumption about the distribution of the errors to compute the ML estimator. If the errors had a distribution other than the normal, then the OLS estimator and the ML estimator would not coincide.

As we'll see later, it will be possible to use (iterated) linear estimation methods and still achieve asymptotic efficiency even if the assumption that $Var(\varepsilon) \neq \sigma^2 I_n$, as long as ε is still normally distributed. This is **not** the case if ε is nonnormal. In general with nonnormal errors it will be necessary to use nonlinear estimation methods to achieve asymptotically efficient estimation.

4.4 Exercises

1. Write an Octave program that generates a histogram for R Monte Carlo replications of $\sqrt{n}\left(\hat{\beta}_j - \beta_j\right)$, where $\hat{\beta}$ is the OLS estimator and β_j is one of the k slope parameters. R should be a large number, at least 1000. The model used to generate data should follow the classical assumptions, except that the errors should not be normally distributed (try $U(-a, a)$, $t(p)$, $\chi^2(p) - p$, etc). Generate histograms for $n \in \{20, 50, 100, 1000\}$. Do you observe evidence of asymptotic normality? Comment.

Chapter 5

Restrictions and hypothesis tests

5.1 Exact linear restrictions

In many cases, economic theory suggests restrictions on the parameters of a model. For example, a demand function is supposed to be homogeneous of degree zero in prices and income. If we have a Cobb-Douglas (log-linear) model,

$$\ln q = \beta_0 + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m + \varepsilon,$$

then we need that

$$k^0 \ln q = \beta_0 + \beta_1 \ln kp_1 + \beta_2 \ln kp_2 + \beta_3 \ln km + \varepsilon,$$

so

$$\begin{aligned}\beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m &= \beta_1 \ln kp_1 + \beta_2 \ln kp_2 + \beta_3 \ln km \\ &= (\ln k)(\beta_1 + \beta_2 + \beta_3) + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m.\end{aligned}$$

The only way to guarantee this for arbitrary k is to set

$$\beta_1 + \beta_2 + \beta_3 = 0,$$

which is a *parameter restriction.* In particular, this is a linear equality restriction, which is probably the most commonly encountered case.

Imposition

The general formulation of linear equality restrictions is the model

$$\begin{aligned}y &= X\beta + \varepsilon \\ R\beta &= r\end{aligned}$$

where R is a $Q \times K$ matrix, $Q < K$ and r is a $Q \times 1$ vector of constants.

- We assume R is of rank Q, so that there are no redundant restrictions.
- We also assume that $\exists\beta$ that satisfies the restrictions: they aren't infeasible.

Let's consider how to estimate β subject to the restrictions $R\beta = r$. The most obvious approach is to set up the Lagrangean

$$\min_{\beta} s(\beta) = \frac{1}{n}(y - X\beta)'(y - X\beta) + 2\lambda'(R\beta - r).$$

The Lagrange multipliers are scaled by 2, which makes things less messy. The fonc are

$$\begin{aligned} D_\beta s(\hat{\beta}, \hat{\lambda}) &= -2X'y + 2X'X\hat{\beta}_R + 2R'\hat{\lambda} \equiv 0 \\ D_\lambda s(\hat{\beta}, \hat{\lambda}) &= R\hat{\beta}_R - r \equiv 0, \end{aligned}$$

which can be written as

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'y \\ r \end{bmatrix}.$$

We get

$$\begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ r \end{bmatrix}.$$

Maybe you're curious about how to invert a partitioned matrix? I can help you with that:

Note that

$$\begin{aligned}
\begin{bmatrix} (X'X)^{-1} & 0 \\ -R(X'X)^{-1} & I_Q \end{bmatrix} \begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} &\equiv AB \\
&= \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -R(X'X)^{-1}R' \end{bmatrix} \\
&\equiv \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -P \end{bmatrix} \\
&\equiv C,
\end{aligned}$$

and

$$\begin{aligned}
\begin{bmatrix} I_K & (X'X)^{-1}R'P^{-1} \\ 0 & -P^{-1} \end{bmatrix} \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -P \end{bmatrix} &\equiv DC \\
&= I_{K+Q},
\end{aligned}$$

so

$$\begin{aligned}
DAB &= I_{K+Q} \\
DA &= B^{-1} \\
B^{-1} &= \begin{bmatrix} I_K & (X'X)^{-1}R'P^{-1} \\ 0 & -P^{-1} \end{bmatrix} \begin{bmatrix} (X'X)^{-1} & 0 \\ -R(X'X)^{-1} & I_Q \end{bmatrix} \\
&= \begin{bmatrix} (X'X)^{-1} - (X'X)^{-1}R'P^{-1}R(X'X)^{-1} & (X'X)^{-1}R'P^{-1} \\ P^{-1}R(X'X)^{-1} & -P^{-1} \end{bmatrix},
\end{aligned}$$

If you weren't curious about that, please start paying attention again. Also, note that we have made the definition $P = R\left(X'X\right)^{-1}R'$)

$$\begin{aligned}
\begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} &= \begin{bmatrix} (X'X)^{-1} - (X'X)^{-1}R'P^{-1}R\left(X'X\right)^{-1} & (X'X)^{-1}R'P^{-1} \\ P^{-1}R\left(X'X\right)^{-1} & -P^{-1} \end{bmatrix} \begin{bmatrix} X'y \\ r \end{bmatrix} \\
&= \begin{bmatrix} \hat{\beta} - (X'X)^{-1}R'P^{-1}\left(R\hat{\beta} - r\right) \\ P^{-1}\left(R\hat{\beta} - r\right) \end{bmatrix} \\
&= \begin{bmatrix} (I_K - (X'X)^{-1}R'P^{-1}R) \\ P^{-1}R \end{bmatrix} \hat{\beta} + \begin{bmatrix} (X'X)^{-1}R'P^{-1}r \\ -P^{-1}r \end{bmatrix}
\end{aligned}$$

The fact that $\hat{\beta}_R$ and $\hat{\lambda}$ are linear functions of $\hat{\beta}$ makes it easy to determine their distributions, since the distribution of $\hat{\beta}$ is already known. Recall that for x a random vector, and for A and b a matrix and vector of constants, respectively, $Var\left(Ax + b\right) = AVar(x)A'$.

Though this is the obvious way to go about finding the restricted estimator, an easier way, if the number of restrictions is small, is to impose them by substitution. Write

$$\begin{aligned}
y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\
\begin{bmatrix} R_1 & R_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} &= r
\end{aligned}$$

where R_1 is $Q \times Q$ nonsingular. Supposing the Q restrictions are linearly independent, one can always make R_1 nonsingular by reorganizing the columns of X. Then

$$\beta_1 = R_1^{-1}r - R_1^{-1}R_2\beta_2.$$

Substitute this into the model

$$\begin{aligned} y &= X_1 R_1^{-1} r - X_1 R_1^{-1} R_2 \beta_2 + X_2 \beta_2 + \varepsilon \\ y - X_1 R_1^{-1} r &= \left[X_2 - X_1 R_1^{-1} R_2 \right] \beta_2 + \varepsilon \end{aligned}$$

or with the appropriate definitions,

$$y_R = X_R \beta_2 + \varepsilon.$$

This model satisfies the classical assumptions, *supposing the restriction is true.* One can estimate by OLS. The variance of $\hat{\beta}_2$ is as before

$$V(\hat{\beta}_2) = \left(X_R' X_R \right)^{-1} \sigma_0^2$$

and the estimator is

$$\hat{V}(\hat{\beta}_2) = \left(X_R' X_R \right)^{-1} \hat{\sigma}^2$$

where one estimates σ_0^2 in the normal way, using the restricted model, *i.e.,*

$$\widehat{\sigma_0^2} = \frac{\left(y_R - X_R \widehat{\beta_2} \right)' \left(y_R - X_R \widehat{\beta_2} \right)}{n - (K - Q)}$$

To recover $\hat{\beta}_1$, use the restriction. To find the variance of $\hat{\beta}_1$, use the fact that it is a linear function of $\hat{\beta}_2$, so

$$\begin{aligned} V(\hat{\beta}_1) &= R_1^{-1} R_2 V(\hat{\beta}_2) R_2' \left(R_1^{-1} \right)' \\ &= R_1^{-1} R_2 \left(X_2' X_2 \right)^{-1} R_2' \left(R_1^{-1} \right)' \sigma_0^2 \end{aligned}$$

Properties of the restricted estimator

We have that

$$\begin{aligned}
\hat{\beta}_R &= \hat{\beta} - (X'X)^{-1}R'P^{-1}\left(R\hat{\beta} - r\right) \\
&= \hat{\beta} + (X'X)^{-1}R'P^{-1}r - (X'X)^{-1}R'P^{-1}R(X'X)^{-1}X'y \\
&= \beta + (X'X)^{-1}X'\varepsilon + (X'X)^{-1}R'P^{-1}\left[r - R\beta\right] - (X'X)^{-1}R'P^{-1}R(X'X)^{-1}X'\varepsilon \\
\hat{\beta}_R - \beta &= (X'X)^{-1}X'\varepsilon \\
&+ (X'X)^{-1}R'P^{-1}\left[r - R\beta\right] \\
&- (X'X)^{-1}R'P^{-1}R(X'X)^{-1}X'\varepsilon
\end{aligned}$$

Mean squared error is

$$MSE(\hat{\beta}_R) = \mathcal{E}(\hat{\beta}_R - \beta)(\hat{\beta}_R - \beta)'$$

Noting that the crosses between the second term and the other terms expect to zero, and that the cross of the first and third has a cancellation with the square of the third, we obtain

$$\begin{aligned}
MSE(\hat{\beta}_R) &= (X'X)^{-1}\sigma^2 \\
&+ (X'X)^{-1}R'P^{-1}\left[r - R\beta\right]\left[r - R\beta\right]'P^{-1}R(X'X)^{-1} \\
&- (X'X)^{-1}R'P^{-1}R(X'X)^{-1}\sigma^2
\end{aligned}$$

So, the first term is the OLS covariance. The second term is PSD, and the third term is NSD.

- If the restriction is true, the second term is 0, so we are better off. *True restrictions improve efficiency of estimation.*

- If the restriction is false, we may be better or worse off, in terms of MSE, depending on the magnitudes of $r - R\beta$ and σ^2.

5.2 Testing

In many cases, one wishes to test economic theories. If theory suggests parameter restrictions, as in the above homogeneity example, one can test theory by testing parameter restrictions. A number of tests are available. The first two (t and F) have a known small sample distributions, when the errors are normally distributed. The third and fourth (Wald and score) do not require normality of the errors, but their distributions are known only approximately, so that they are not exactly valid with finite samples.

t-test

Suppose one has the model

$$y = X\beta + \varepsilon$$

and one wishes to test the *single restriction* $H_0 : R\beta = r$ vs. $H_A : R\beta \neq r$. Under H_0, with normality of the errors,

$$R\hat{\beta} - r \sim N\left(0, R(X'X)^{-1}R'\sigma_0^2\right)$$

so

$$\frac{R\hat{\beta} - r}{\sqrt{R(X'X)^{-1}R'\sigma_0^2}} = \frac{R\hat{\beta} - r}{\sigma_0\sqrt{R(X'X)^{-1}R'}} \sim N(0,1).$$

The problem is that σ_0^2 is unknown. One could use the consistent estimator $\widehat{\sigma_0^2}$ in place of σ_0^2, but the test would only be valid asymptotically in this case.

Proposition 1. $\frac{N(0,1)}{\sqrt{\frac{\chi^2(q)}{q}}} \sim t(q)$

as long as the $N(0,1)$ *and the* $\chi^2(q)$ *are independent.*

We need a few results on the χ^2 distribution.

Proposition 2. *If* $x \sim N(\mu, I_n)$ *is a vector of* n *independent r.v.'s., then* $x'x \sim \chi^2(n, \lambda)$ *where* $\lambda = \sum_i \mu_i^2 = \mu'\mu$ *is the* noncentrality parameter.

When a χ^2 r.v. has the noncentrality parameter equal to zero, it is referred to as a central χ^2 r.v., and it's distribution is written as $\chi^2(n)$, suppressing the noncentrality parameter.

Proposition 3. *If the* n *dimensional random vector* $x \sim N(0, V)$, *then* $x'V^{-1}x \sim \chi^2(n)$.

We'll prove this one as an indication of how the following unproven propositions could be proved.

Proof: Factor V^{-1} as $P'P$ (this is the Cholesky factorization, where P is defined to be upper triangular). Then consider $y = Px$. We have

$$y \sim N(0, PVP')$$

but

$$\begin{aligned} VP'P &= I_n \\ PVP'P &= P \end{aligned}$$

so $PVP' = I_n$ and thus $y \sim N(0, I_n)$. Thus $y'y \sim \chi^2(n)$ but

$$y'y = x'P'Px = xV^{-1}x$$

and we get the result we wanted.

A more general proposition which implies this result is

Proposition 4. *If the n dimensional random vector $x \sim N(0, V)$, then $x'Bx \sim \chi^2(\rho(B))$ if and only if BV is idempotent.*

An immediate consequence is

Proposition 5. *If the random vector (of dimension n) $x \sim N(0, I)$, and B is idempotent with rank r, then $x'Bx \sim \chi^2(r)$.*

Consider the random variable

$$\begin{aligned} \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma_0^2} &= \frac{\varepsilon' M_X \varepsilon}{\sigma_0^2} \\ &= \left(\frac{\varepsilon}{\sigma_0}\right)' M_X \left(\frac{\varepsilon}{\sigma_0}\right) \\ &\sim \chi^2(n-K) \end{aligned}$$

Proposition 6. *If the random vector (of dimension n) $x \sim N(0, I)$, then Ax and $x'Bx$ are independent if $AB = 0$.*

Now consider (remember that we have only one restriction in this case)

$$\frac{\frac{R\hat{\beta}-r}{\sigma_0\sqrt{R(X'X)^{-1}R'}}}{\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{(n-K)\sigma_0^2}}} = \frac{R\hat{\beta}-r}{\widehat{\sigma_0}\sqrt{R(X'X)^{-1}R'}}$$

This will have the $t(n-K)$ distribution if $\hat{\beta}$ and $\hat{\varepsilon}'\hat{\varepsilon}$ are independent. But $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$ and

$$(X'X)^{-1}X'M_X = 0,$$

so

$$\frac{R\hat{\beta}-r}{\widehat{\sigma_0}\sqrt{R(X'X)^{-1}R'}} = \frac{R\hat{\beta}-r}{\hat{\sigma}_{R\hat{\beta}}} \sim t(n-K)$$

In particular, for the commonly encountered *test of significance* of an individual coefficient, for which $H_0 : \beta_i = 0$ vs. $H_0 : \beta_i \neq 0$, the test statistic is

$$\frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}i}} \sim t(n-K)$$

- **Note**: the $t-$ test is strictly valid only if the errors are actually normally distributed. If one has nonnormal errors, one could use the above asymptotic result to justify taking critical values from the $N(0,1)$ distribution, since $t(n-K) \xrightarrow{d} N(0,1)$ as $n \to \infty$. In practice, a conservative procedure is to take critical values from the t distribution if nonnormality is suspected. This will reject H_0 less often since the t distribution is fatter-tailed than is the normal.

F test

The F test allows testing multiple restrictions jointly.

Proposition 7. *If $x \sim \chi^2(r)$ and $y \sim \chi^2(s)$, then $\frac{x/r}{y/s} \sim F(r,s)$, provided that x and y are independent.*

Proposition 8. *If the random vector (of dimension n) $x \sim N(0,I)$, then $x'Ax$ and $x'Bx$ are independent if $AB = 0$.*

Using these results, and previous results on the χ^2 distribution, it is simple to show that the following statistic has the F distribution:

$$F = \frac{\left(R\hat{\beta} - r\right)' \left(R\left(X'X\right)^{-1} R'\right)^{-1} \left(R\hat{\beta} - r\right)}{q\hat{\sigma}^2} \sim F(q, n-K).$$

A numerically equivalent expression is

$$\frac{\left(ESS_R - ESS_U\right)/q}{ESS_U/(n-K)} \sim F(q, n-K).$$

- **Note:** The F test is strictly valid only if the errors are truly normally distributed. The following tests will be appropriate when one cannot assume normally distributed errors.

Wald-type tests

The t and F tests require normality of the errors. The Wald test does not, but it is an asymptotic test - it is only approximately valid in finite samples.

The Wald principle is based on the idea that if a restriction is true, the unrestricted model should "approximately" satisfy the restriction. Given that the least squares estimator is asymptotically normally distributed:

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N\left(0, \sigma_0^2 Q_X^{-1}\right)$$

then under $H_0 : R\beta_0 = r$, we have

$$\sqrt{n}\left(R\hat{\beta}-r\right) \xrightarrow{d} N\left(0, \sigma_0^2 R Q_X^{-1} R'\right)$$

so by Proposition [3]

$$n\left(R\hat{\beta}-r\right)'\left(\sigma_0^2 R Q_X^{-1} R'\right)^{-1}\left(R\hat{\beta}-r\right) \xrightarrow{d} \chi^2(q)$$

Note that Q_X^{-1} or σ_0^2 are not observable. The test statistic we use substitutes the consistent estimators. Use $(X'X/n)^{-1}$ as the consistent estimator of Q_X^{-1}. With this, there is a cancellation of $n's$, and the statistic to use is

$$\left(R\hat{\beta}-r\right)'\left(\widehat{\sigma_0^2} R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right) \xrightarrow{d} \chi^2(q)$$

- The Wald test is a simple way to test restrictions without having to estimate the restricted model.
- Note that this formula is similar to one of the formulae provided for the F test.

Score-type tests (Rao tests, Lagrange multiplier tests)

The score test is another asymptotically valid test that does not require normality of the errors.

In some cases, an unrestricted model may be nonlinear in the parameters, but the model is linear

in the parameters under the null hypothesis. For example, the model

$$y = (X\beta)^{\gamma} + \varepsilon$$

is nonlinear in β and γ, but is linear in β under $H_0 : \gamma = 1$. Estimation of nonlinear models is a bit more complicated, so one might prefer to have a test based upon the restricted, linear model. The score test is useful in this situation.

- Score-type tests are based upon the general principle that the gradient vector of the unrestricted model, evaluated at the restricted estimate, should be asymptotically normally distributed with mean zero, if the restrictions are true. The original development was for ML estimation, but the principle is valid for a wide variety of estimation methods.

We have seen that

$$\begin{aligned} \hat{\lambda} &= \left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta} - r\right) \\ &= P^{-1}\left(R\hat{\beta} - r\right) \end{aligned}$$

so

$$\sqrt{n}\hat{P\lambda} = \sqrt{n}\left(R\hat{\beta} - r\right)$$

Given that

$$\sqrt{n}\left(R\hat{\beta} - r\right) \xrightarrow{d} N\left(0, \sigma_0^2 R Q_X^{-1} R'\right)$$

under the null hypothesis, we obtain

$$\sqrt{n}\hat{P\lambda} \xrightarrow{d} N\left(0, \sigma_0^2 R Q_X^{-1} R'\right)$$

So

$$\left(\sqrt{n}\hat{P}\hat{\lambda}\right)'\left(\sigma_0^2 R Q_X^{-1} R'\right)^{-1}\left(\sqrt{n}\hat{P}\hat{\lambda}\right) \xrightarrow{d} \chi^2(q)$$

Noting that $\lim nP = RQ_X^{-1}R'$, we obtain,

$$\hat{\lambda}'\left(\frac{R(X'X)^{-1}R'}{\sigma_0^2}\right)\hat{\lambda} \xrightarrow{d} \chi^2(q)$$

since the powers of n cancel. To get a usable test statistic substitute a consistent estimator of σ_0^2.

- This makes it clear why the test is sometimes referred to as a Lagrange multiplier test. It may seem that one needs the actual Lagrange multipliers to calculate this. If we impose the restrictions by substitution, these are not available. Note that the test can be written as

  $$\frac{\left(R'\hat{\lambda}\right)'(X'X)^{-1}R'\hat{\lambda}}{\sigma_0^2} \xrightarrow{d} \chi^2(q)$$

  However, we can use the fonc for the restricted estimator:

  $$-X'y + X'X\hat{\beta}_R + R'\hat{\lambda}$$

  to get that

  $$\begin{aligned} R'\hat{\lambda} &= X'(y - X\hat{\beta}_R) \\ &= X'\hat{\varepsilon}_R \end{aligned}$$

Substituting this into the above, we get

$$\frac{\hat{\varepsilon}_R' X (X'X)^{-1} X' \hat{\varepsilon}_R}{\sigma_0^2} \xrightarrow{d} \chi^2(q)$$

but this is simply

$$\hat{\varepsilon}_R' \frac{P_X}{\sigma_0^2} \hat{\varepsilon}_R \xrightarrow{d} \chi^2(q).$$

To see why the test is also known as a score test, note that the fonc for restricted least squares

$$-X'y + X'X\hat{\beta}_R + R'\hat{\lambda}$$

give us

$$R'\hat{\lambda} = X'y - X'X\hat{\beta}_R$$

and the rhs is simply the gradient (score) of the unrestricted model, evaluated at the restricted estimator. The scores evaluated at the unrestricted estimate are identically zero. The logic behind the score test is that the scores evaluated at the restricted estimate should be approximately zero, if the restriction is true. The test is also known as a Rao test, since P. Rao first proposed it in 1948.

5.3 The asymptotic equivalence of the LR, Wald and score tests

Note: the discussion of the LR test has been moved forward in these notes. I no longer teach the material in this section, but I'm leaving it here for reference.

We have seen that the three tests all converge to χ^2 random variables. In fact, they all converge to the *same* χ^2 rv, under the null hypothesis. We'll show that the Wald and LR tests are asymptotically equivalent. We have seen that the Wald test is asymptotically equivalent to

$$W \stackrel{a}{=} n\left(R\hat{\beta} - r\right)' \left(\sigma_0^2 R Q_X^{-1} R'\right)^{-1} \left(R\hat{\beta} - r\right) \stackrel{d}{\to} \chi^2(q) \tag{5.1}$$

Using

$$\hat{\beta} - \beta_0 = (X'X)^{-1}X'\varepsilon$$

and

$$R\hat{\beta} - r = R(\hat{\beta} - \beta_0)$$

we get

$$\begin{aligned} \sqrt{n}R(\hat{\beta} - \beta_0) &= \sqrt{n}R(X'X)^{-1}X'\varepsilon \\ &= R\left(\frac{X'X}{n}\right)^{-1} n^{-1/2}X'\varepsilon \end{aligned}$$

Substitute this into [5.1] to get

$$\begin{aligned} W &\stackrel{a}{=} n^{-1}\varepsilon' X Q_X^{-1} R' \left(\sigma_0^2 R Q_X^{-1} R'\right)^{-1} R Q_X^{-1} X'\varepsilon \\ &\stackrel{a}{=} \varepsilon' X (X'X)^{-1} R' \left(\sigma_0^2 R (X'X)^{-1} R'\right)^{-1} R (X'X)^{-1} X'\varepsilon \\ &\stackrel{a}{=} \frac{\varepsilon' A (A'A)^{-1} A'\varepsilon}{\sigma_0^2} \\ &\stackrel{a}{=} \frac{\varepsilon' P_R \varepsilon}{\sigma_0^2} \end{aligned}$$

where P_R is the projection matrix formed by the matrix $X(X'X)^{-1}R'$.

- Note that this matrix is idempotent and has q columns, so the projection matrix has rank q.

Now consider the likelihood ratio statistic

$$LR \stackrel{a}{=} n^{1/2} g(\theta_0)' \mathcal{I}(\theta_0)^{-1} R' \left(R \mathcal{I}(\theta_0)^{-1} R'\right)^{-1} R \mathcal{I}(\theta_0)^{-1} n^{1/2} g(\theta_0) \tag{5.2}$$

Under normality, we have seen that the likelihood function is

$$\ln L(\beta, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \frac{1}{2} \frac{(y - X\beta)'(y - X\beta)}{\sigma^2}.$$

Using this,

$$\begin{aligned} g(\beta_0) &\equiv D_\beta \frac{1}{n} \ln L(\beta, \sigma) \\ &= \frac{X'(y - X\beta_0)}{n\sigma^2} \\ &= \frac{X'\varepsilon}{n\sigma^2} \end{aligned}$$

Also, by the information matrix equality:

$$\begin{aligned} \mathcal{I}(\theta_0) &= -H_\infty(\theta_0) \\ &= \lim -D_{\beta'} g(\beta_0) \\ &= \lim -D_{\beta'} \frac{X'(y - X\beta_0)}{n\sigma^2} \\ &= \lim \frac{X'X}{n\sigma^2} \\ &= \frac{Q_X}{\sigma^2} \end{aligned}$$

so

$$\mathcal{I}(\theta_0)^{-1} = \sigma^2 Q_X^{-1}$$

Substituting these last expressions into [5.2], we get

$$\begin{aligned} LR &\stackrel{a}{=} \varepsilon' X'(X'X)^{-1}R' \left(\sigma_0^2 R(X'X)^{-1}R'\right)^{-1} R(X'X)^{-1}X'\varepsilon \\ &\stackrel{a}{=} \frac{\varepsilon' P_R \varepsilon}{\sigma_0^2} \\ &\stackrel{a}{=} W \end{aligned}$$

This completes the proof that the Wald and LR tests are asymptotically equivalent. Similarly, one can show that, *under the null hypothesis*,

$$qF \stackrel{a}{=} W \stackrel{a}{=} LM \stackrel{a}{=} LR$$

- The proof for the statistics except for LR does not depend upon normality of the errors, as can be verified by examining the expressions for the statistics.
- The LR statistic *is* based upon distributional assumptions, since one can't write the likelihood function without them.
- However, due to the close relationship between the statistics qF and LR, supposing normality, the qF statistic can be thought of as a *pseudo-LR statistic,* in that it's like a LR statistic in that it uses the value of the objective functions of the restricted and unrestricted models, but it doesn't require distributional assumptions.
- The presentation of the score and Wald tests has been done in the context of the linear model. This is readily generalizable to nonlinear models and/or other estimation methods.

Though the four statistics *are* asymptotically equivalent, they are numerically different in small sam-

ples. The numeric values of the tests also depend upon how σ^2 is estimated, and we've already seen than there are several ways to do this. For example all of the following are consistent for σ^2 under H_0

$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$
$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$
$$\frac{\hat{\varepsilon}_R'\hat{\varepsilon}_R}{n-k+q}$$
$$\frac{\hat{\varepsilon}_R'\hat{\varepsilon}_R}{n}$$

and in general the denominator call be replaced with any quantity a such that $\lim a/n = 1$.

It can be shown, for linear regression models subject to linear restrictions, and if $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$ is used to calculate the Wald test and $\frac{\hat{\varepsilon}_R'\hat{\varepsilon}_R}{n}$ is used for the score test, that

$$W > LR > LM.$$

For this reason, the Wald test will always reject if the LR test rejects, and in turn the LR test rejects if the LM test rejects. This is a bit problematic: there is the possibility that by careful choice of the statistic used, one can manipulate reported results to favor or disfavor a hypothesis. A conservative/honest approach would be to report all three test statistics when they are available. In the case of linear models with normal errors the F test is to be preferred, since asymptotic approximations are not an issue.

The small sample behavior of the tests can be quite different. The true size (probability of rejection of the null when the null is true) of the Wald test is often dramatically higher than the nominal size associated with the asymptotic distribution. Likewise, the true size of the score test is often smaller

than the nominal size.

5.4 Interpretation of test statistics

Now that we have a menu of test statistics, we need to know how to use them.

5.5 Confidence intervals

Confidence intervals for single coefficients are generated in the normal manner. Given the t statistic

$$t(\beta) = \frac{\hat{\beta} - \beta}{\widehat{\sigma_{\hat{\beta}}}}$$

a $100\,(1-\alpha)\,\%$ confidence interval for β_0 is defined by the bounds of the set of β such that $t(\beta)$ does not reject $H_0 : \beta_0 = \beta$, using a α significance level:

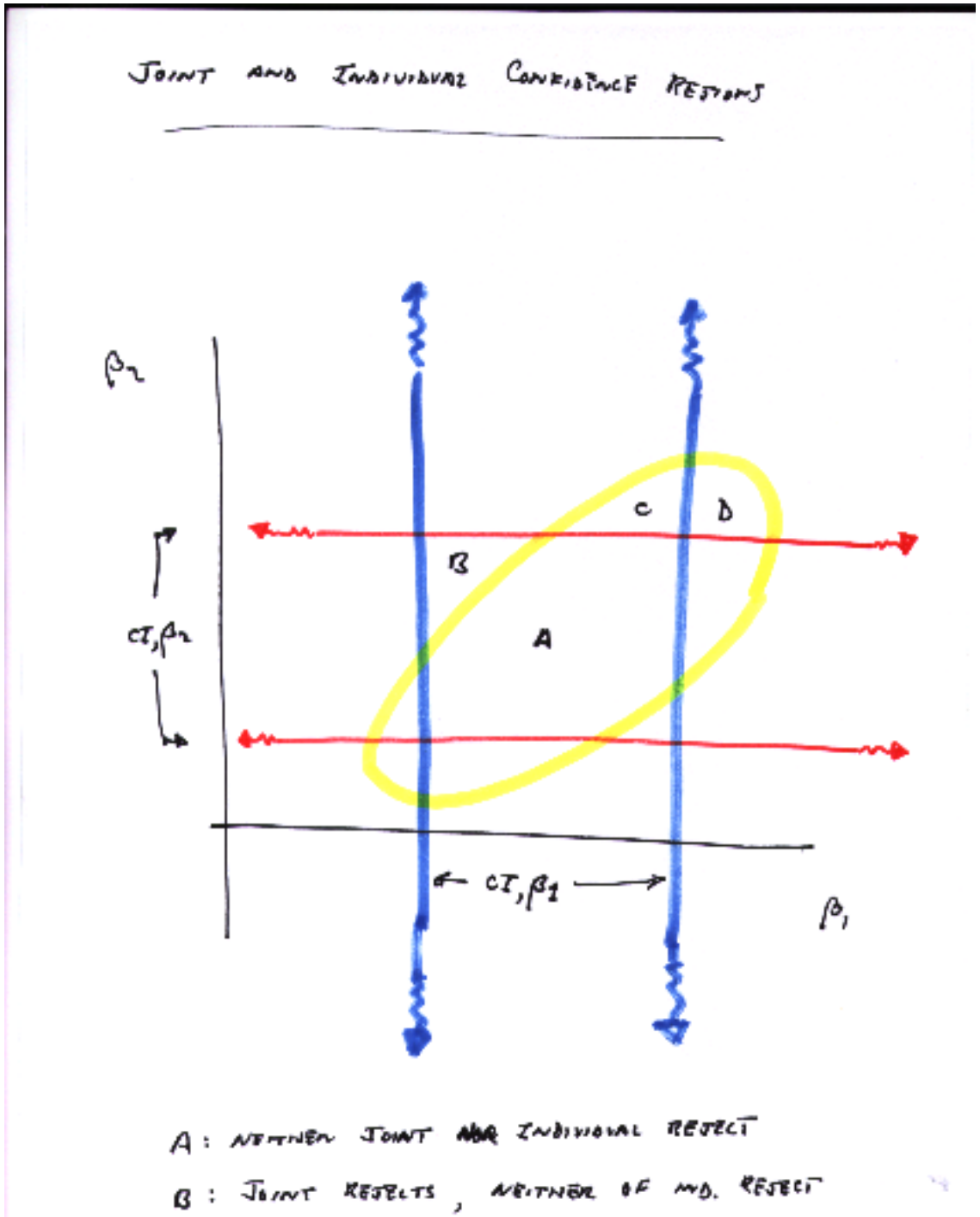
$$C(\alpha) = \{\beta : -c_{\alpha/2} < \frac{\hat{\beta} - \beta}{\widehat{\sigma_{\hat{\beta}}}} < c_{\alpha/2}\}$$

The set of such β is the interval

$$\hat{\beta} \pm \widehat{\sigma_{\hat{\beta}}} c_{\alpha/2}$$

A confidence ellipse for two coefficients jointly would be, analogously, the set of $\{\beta_1, \beta_2\}$ such that the F (or some other test statistic) doesn't reject at the specified critical value. This generates an ellipse, if the estimators are correlated.

Figure 5.1: Joint and Individual Confidence Regions



- The region is an ellipse, since the CI for an individual coefficient defines a (infinitely long) rectangle with total prob. mass $1-\alpha$, since the other coefficient is marginalized (e.g., can take on any value). Since the ellipse is bounded in both dimensions but also contains mass $1-\alpha$, it must extend beyond the bounds of the individual CI.

- From the pictue we can see that:

  - Rejection of hypotheses individually does not imply that the joint test will reject.
  - Joint rejection does not imply individal tests will reject.

5.6 Bootstrapping

When we rely on asymptotic theory to use the normal distribution-based tests and confidence intervals, we're often at serious risk of making important errors. If the sample size is small and errors are highly nonnormal, the small sample distribution of $\sqrt{n}\left(\hat{\beta}-\beta_0\right)$ may be very different than its large sample distribution. Also, the distributions of test statistics may not resemble their limiting distributions at all. A means of trying to gain information on the small sample distribution of test statistics and estimators is the *bootstrap.* We'll consider a simple example, just to get the main idea.

Suppose that

$$\begin{array}{ccc} y & = & X\beta_0+\varepsilon \\ \varepsilon & \sim & IID(0,\sigma_0^2) \end{array}$$

X is nonstochastic

Given that the distribution of ε is unknown, the distribution of $\hat{\beta}$ will be unknown in small samples. However, since we have random sampling, we could generate *artificial data.* The steps are:

1. Draw n observations from $\hat{\varepsilon}$ **with replacement**. Call this vector $\tilde{\varepsilon}^j$ (it's a $n \times 1$).
2. Then generate the data by $\tilde{y}^j = X\hat{\beta} + \tilde{\varepsilon}^j$
3. Now take this and estimate
$$\tilde{\beta}^j = (X'X)^{-1}X'\tilde{y}^j.$$
4. Save $\tilde{\beta}^j$
5. Repeat steps 1-4, until we have a large number, J, of $\tilde{\beta}^j$.

With this, we can use the replications to calculate the *empirical distribution of* $\tilde{\beta}_j$. One way to form a 100(1-α)% confidence interval for β_0 would be to order the $\tilde{\beta}^j$ from smallest to largest, and drop the first and last $J\alpha/2$ of the replications, and use the remaining endpoints as the limits of the CI. Note that this will not give the shortest CI if the empirical distribution is skewed.

- Suppose one was interested in the distribution of some function of $\hat{\beta}$, for example a test statistic. Simple: just calculate the transformation for each j, and work with the empirical distribution of the transformation.
- If the assumption of iid errors is too strong (for example if there is heteroscedasticity or autocorrelation, see below) one can work with a bootstrap defined by sampling from (y, x) with replacement.

- How to choose J: J should be large enough that the results don't change with repetition of the entire bootstrap. This is easy to check. If you find the results change a lot, increase J and try again.

- The bootstrap is based fundamentally on the idea that the empirical distribution of the sample data converges to the actual sampling distribution as n becomes large, so statistics based on sampling from the empirical distribution should converge in distribution to statistics based on sampling from the actual sampling distribution.

- In finite samples, this doesn't hold. At a minimum, the bootstrap is a good way to check if asymptotic theory results offer a decent approximation to the small sample distribution.

- Bootstrapping can be used to test hypotheses. Basically, use the bootstrap to get an approximation to the empirical distribution of the test statistic under the alternative hypothesis, and use this to get critical values. Compare the test statistic calculated using the real data, under the null, to the bootstrap critical values. There are many variations on this theme, which we won't go into here.

5.7 Wald test for nonlinear restrictions: the delta method

Testing nonlinear restrictions of a linear model is not much more difficult, at least when the model is linear. Since estimation subject to nonlinear restrictions requires nonlinear estimation methods, which are beyond the score of this course, we'll just consider the Wald test for nonlinear restrictions on a linear model.

Consider the q nonlinear restrictions

$$r(\beta_0) = 0.$$

where $r(\cdot)$ is a q-vector valued function. Write the derivative of the restriction evaluated at β as

$$D_{\beta'} r(\beta)|_{\beta} = R(\beta)$$

We suppose that the restrictions are not redundant in a neighborhood of β_0, so that

$$\rho(R(\beta)) = q$$

in a neighborhood of β_0. Take a first order Taylor's series expansion of $r(\hat{\beta})$ about β_0:

$$r(\hat{\beta}) = r(\beta_0) + R(\beta^*)(\hat{\beta} - \beta_0)$$

where β^* is a convex combination of $\hat{\beta}$ and β_0. Under the null hypothesis we have

$$r(\hat{\beta}) = R(\beta^*)(\hat{\beta} - \beta_0)$$

Due to consistency of $\hat{\beta}$ we can replace β^* by β_0, asymptotically, so

$$\sqrt{n} r(\hat{\beta}) \stackrel{a}{=} \sqrt{n} R(\beta_0)(\hat{\beta} - \beta_0)$$

We've already seen the distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$. Using this we get

$$\sqrt{n} r(\hat{\beta}) \stackrel{d}{\to} N\left(0, R(\beta_0) Q_X^{-1} R(\beta_0)' \sigma_0^2\right).$$

Considering the quadratic form

$$\frac{nr(\hat{\beta})' \left(R(\beta_0)Q_X^{-1}R(\beta_0)'\right)^{-1} r(\hat{\beta})}{\sigma_0^2} \xrightarrow{d} \chi^2(q)$$

under the null hypothesis. Substituting consistent estimators for $\beta_{0,}Q_X$ and σ_0^2, the resulting statistic is

$$\frac{r(\hat{\beta})' \left(R(\hat{\beta})(X'X)^{-1}R(\hat{\beta})'\right)^{-1} r(\hat{\beta})}{\widehat{\sigma^2}} \xrightarrow{d} \chi^2(q)$$

under the null hypothesis.

- This is known in the literature as the *delta method*, or as *Klein's approximation.*
- Since this is a Wald test, it will tend to over-reject in finite samples. The score and LR tests are also possibilities, but they require estimation methods for nonlinear models, which aren't in the scope of this course.

Note that this also gives a convenient way to estimate nonlinear functions and associated asymptotic confidence intervals. If the nonlinear function $r(\beta_0)$ is not hypothesized to be zero, we just have

$$\sqrt{n}\left(r(\hat{\beta}) - r(\beta_0)\right) \xrightarrow{d} N\left(0, R(\beta_0)Q_X^{-1}R(\beta_0)'\sigma_0^2\right)$$

so an approximation to the distribution of the function of the estimator is

$$r(\hat{\beta}) \approx N(r(\beta_0), R(\beta_0)(X'X)^{-1}R(\beta_0)'\sigma_0^2)$$

For example, the vector of elasticities of a function $f(x)$ is

$$\eta(x) = \frac{\partial f(x)}{\partial x} \odot \frac{x}{f(x)}$$

where $\odot$ means element-by-element multiplication. Suppose we estimate a linear function

$$y = x'\beta + \varepsilon.$$

The elasticities of y w.r.t. x are

$$\eta(x) = \frac{\beta}{x'\beta} \odot x$$

(note that this is the entire vector of elasticities). The estimated elasticities are

$$\widehat{\eta}(x) = \frac{\hat{\beta}}{x'\hat{\beta}} \odot x$$

To calculate the estimated standard errors of all five elasticites, use

$$\begin{aligned} R(\beta) &= \frac{\partial \eta(x)}{\partial \beta'} \\ &= \frac{\begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & x_k \end{bmatrix} x'\beta - \begin{bmatrix} \beta_1 x_1^2 & 0 & \cdots & 0 \\ 0 & \beta_2 x_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \beta_k x_k^2 \end{bmatrix}}{(x'\beta)^2}. \end{aligned}$$

To get a consistent estimator just substitute in $\hat{\beta}$. Note that the elasticity and the standard error are functions of x. The program ExampleDeltaMethod.m shows how this can be done.

In many cases, nonlinear restrictions can also involve the data, not just the parameters. For example, consider a model of expenditure shares. Let $x(p, m)$ be a demand funcion, where p is prices and m is income. An expenditure share system for G goods is

$$s_i(p, m) = \frac{p_i x_i(p, m)}{m}, i = 1, 2, ..., G.$$

Now demand must be positive, and we assume that expenditures sum to income, so we have the restrictions

$$\begin{aligned} 0 \leq s_i(p, m) &\leq 1, \quad \forall i \\ \sum_{i=1}^{G} s_i(p, m) &= 1 \end{aligned}$$

Suppose we postulate a linear model for the expenditure shares:

$$s_i(p, m) = \beta_1^i + p'\beta_p^i + m\beta_m^i + \varepsilon^i$$

It is fairly easy to write restrictions such that the shares sum to one, but the restriction that the shares lie in the $[0, 1]$ interval depends on both parameters and the values of p and m. It is impossible to impose the restriction that $0 \leq s_i(p, m) \leq 1$ for all possible p and m. In such cases, one might consider whether or not a linear model is a reasonable specification.

5.8 Example: the Nerlove data

Remember that we in a previous example (section 3.8) that the OLS results for the Nerlove model are

```
*******************************************************
OLS estimation results
Observations 145
R-squared 0.925955
Sigma-squared 0.153943

Results (Ordinary var-cov estimator)

            estimate     st.err.     t-stat.     p-value
constant      -3.527       1.774      -1.987       0.049
output         0.720       0.017      41.244       0.000
labor          0.436       0.291       1.499       0.136
fuel           0.427       0.100       4.249       0.000
capital       -0.220       0.339      -0.648       0.518

*******************************************************
```

Note that $s_K = \beta_K < 0$, and that $\beta_L + \beta_F + \beta_K \neq 1$.

Remember that if we have constant returns to scale, then $\beta_Q = 1$, and if there is homogeneity of degree 1 then $\beta_L + \beta_F + \beta_K = 1$. We can test these hypotheses either separately or jointly. NerloveRestrictions.m imposes and tests CRTS and then HOD1. From it we obtain the results that follow:

Imposing and testing HOD1

****
Restricted LS estimation results
Observations 145
R-squared 0.925652
Sigma-squared 0.155686

	estimate	st.err.	t-stat.	p-value
constant	-4.691	0.891	-5.263	0.000
output	0.721	0.018	41.040	0.000
labor	0.593	0.206	2.878	0.005
fuel	0.414	0.100	4.159	0.000
capital	-0.007	0.192	-0.038	0.969

****

	Value	p-value
F	0.574	0.450
Wald	0.594	0.441
LR	0.593	0.441
Score	0.592	0.442

Imposing and testing CRTS

```
*******************************************************
Restricted LS estimation results
Observations 145
R-squared 0.790420
Sigma-squared 0.438861

            estimate     st.err.     t-stat.     p-value
constant      -7.530       2.966      -2.539       0.012
output         1.000       0.000         Inf       0.000
labor          0.020       0.489       0.040       0.968
fuel           0.715       0.167       4.289       0.000
capital        0.076       0.572       0.132       0.895

*******************************************************
             Value     p-value
F          256.262       0.000
Wald       265.414       0.000
LR         150.863       0.000
Score       93.771       0.000
```

Notice that the input price coefficients in fact sum to 1 when HOD1 is imposed. HOD1 is not rejected at usual significance levels (*e.g.*, $\alpha = 0.10$). Also, R^2 does not drop much when the restriction

is imposed, compared to the unrestricted results. For CRTS, you should note that $\beta_Q = 1$, so the restriction is satisfied. Also note that the hypothesis that $\beta_Q = 1$ is rejected by the test statistics at all reasonable significance levels. Note that R^2 drops quite a bit when imposing CRTS. If you look at the unrestricted estimation results, you can see that a t-test for $\beta_Q = 1$ also rejects, and that a confidence interval for β_Q does not overlap 1.

From the point of view of neoclassical economic theory, these results are not anomalous: HOD1 is an implication of the theory, but CRTS is not.

Exercise 9. Modify the NerloveRestrictions.m program to impose and test the restrictions jointly.

The Chow test Since CRTS is rejected, let's examine the possibilities more carefully. Recall that the data is sorted by output (the third column). Define 5 subsamples of firms, with the first group being the 29 firms with the lowest output levels, then the next 29 firms, etc. The five subsamples can be indexed by $j = 1, 2, ..., 5$, where $j = 1$ for $t = 1, 2, ...29$, $j = 2$ for $t = 30, 31, ...58$, etc. Define

dummy variables $D_1, D_2, ..., D_5$ where

$$D_1 = \begin{cases} 1 & t \in \{1, 2, ...29\} \\ 0 & t \notin \{1, 2, ...29\} \end{cases}$$

$$D_2 = \begin{cases} 1 & t \in \{30, 31, ...58\} \\ 0 & t \notin \{30, 31, ...58\} \end{cases}$$

$$\vdots$$

$$D_5 = \begin{cases} 1 & t \in \{117, 118, ..., 145\} \\ 0 & t \notin \{117, 118, ..., 145\} \end{cases}$$

Define the model

$$\ln C_t = \sum_{j=1}^{5} \alpha_1 D_j + \sum_{j=1}^{5} \gamma_j D_j \ln Q_t + \sum_{j=1}^{5} \beta_{Lj} D_j \ln P_{Lt} + \sum_{j=1}^{5} \beta_{Fj} D_j \ln P_{Ft} + \sum_{j=1}^{5} \beta_{Kj} D_j \ln P_{Kt} + \epsilon_t \quad (5.3)$$

Note that the first column of nerlove.data indicates this way of breaking up the sample, and provides and easy way of defining the dummy variables. The new model may be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \\ y_5 \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & & 0 \\ 0 & X_2 & & & \\ \vdots & & X_3 & & \\ & & & X_4 & 0 \\ 0 & & & & X_5 \end{bmatrix} \begin{bmatrix} \beta^1 \\ \beta^2 \\ \\ \\ \beta^5 \end{bmatrix} + \begin{bmatrix} \epsilon^1 \\ \epsilon^2 \\ \vdots \\ \\ \epsilon^5 \end{bmatrix} \quad (5.4)$$

where y_1 is 29×1, X_1 is 29×5, β^j is the 5×1 vector of coefficients for the j^{th} subsample (e.g., $\beta^1 = (\alpha_1, \gamma_1, \beta_{L1}, \beta_{F1}, \beta_{K1})'$), and ϵ^j is the 29×1 vector of errors for the j^{th} subsample.

The Octave program Restrictions/ChowTest.m estimates the above model. It also tests the hypothesis that the five subsamples share the same parameter vector, or in other words, that there is coefficient stability across the five subsamples. The null to test is that the parameter vectors for the separate groups are all the same, that is,

$$\beta^1 = \beta^2 = \beta^3 = \beta^4 = \beta^5$$

This type of test, that parameters are constant across different sets of data, is sometimes referred to as a *Chow test.*

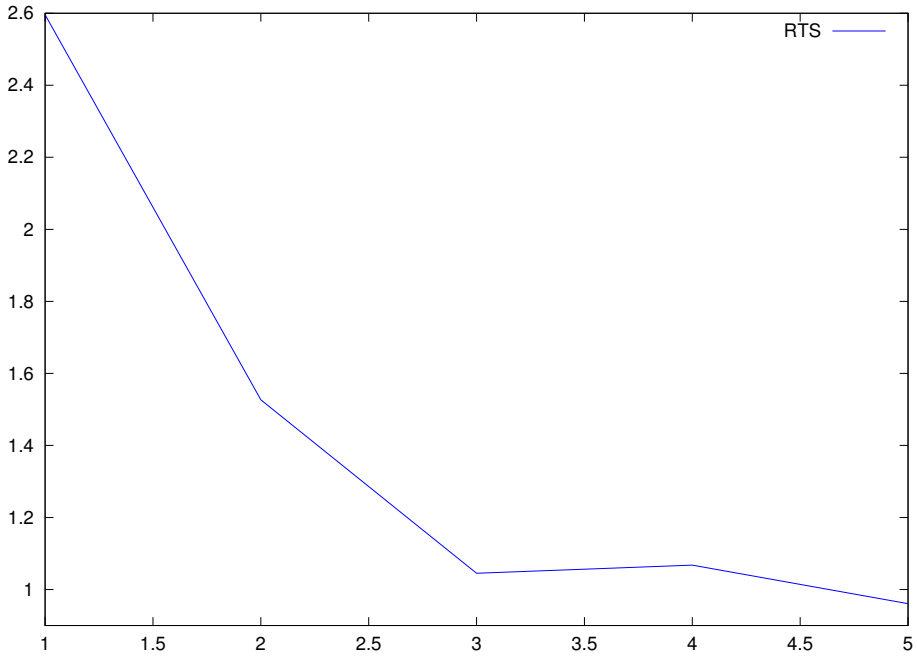
- There are 20 restrictions. If that's not clear to you, look at the Octave program.
- The restrictions are rejected at all conventional significance levels.

Since the restrictions are rejected, we should probably use the unrestricted model for analysis. What is the pattern of RTS as a function of the output group (small to large)? Figure 5.2 plots RTS. We can see that there is increasing RTS for small firms, but that RTS is approximately constant for large firms.

5.9 Exercises

1. Using the Chow test on the Nerlove model, we reject that there is coefficient stability across the 5 groups. But perhaps we could restrict the input price coefficients to be the same but let the

Figure 5.2: RTS as a function of firm size



constant and output coefficients vary by group size. This new model is

$$\ln C = \sum_{j=1}^{5} \alpha_j D_j + \sum_{j=1}^{5} \gamma_j D_j \ln Q + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon \tag{5.5}$$

(a) estimate this model by OLS, giving R^2, estimated standard errors for coefficients, t-statistics for tests of significance, and the associated p-values. Interpret the results in detail.

(b) Test the restrictions implied by this model (relative to the model that lets all coefficients vary across groups) using the F, qF, Wald, score and likelihood ratio tests. Comment on the results.

(c) Estimate this model but imposing the HOD1 restriction, *using an OLS* estimation program. Don't use mc_olsr or any other restricted OLS estimation program. Give estimated standard errors for all coefficients.

(d) Plot the estimated RTS parameters as a function of firm size. Compare the plot to that given in the notes for the unrestricted model. Comment on the results.

2. For the model of the above question, compute 95% confidence intervals for RTS for each of the 5 groups of firms, using the delta method to compute standard errors. Comment on the results.

3. Perform a Monte Carlo study that generates data from the model

$$y = -2 + 1x_2 + 1x_3 + \epsilon$$

where the sample size is 30, x_2 and x_3 are independently uniformly distributed on $[0, 1]$ and $\epsilon \sim IIN(0, 1)$

(a) Compare the means and standard errors of the estimated coefficients using OLS and restricted OLS, imposing the restriction that $\beta_2 + \beta_3 = 2$.

(b) Compare the means and standard errors of the estimated coefficients using OLS and restricted OLS, imposing the restriction that $\beta_2 + \beta_3 = 1$.

(c) Discuss the results.

Chapter 6

Stochastic regressors

Up to now we have treated the regressors as fixed, which is clearly unrealistic. Now we will assume they are random. There are several ways to think of the problem. First, if we are interested in an analysis *conditional* on the explanatory variables, then it is irrelevant if they are stochastic or not, since conditional on the values of they regressors take on, they are nonstochastic, which is the case already considered.

- In cross-sectional analysis it is usually reasonable to make the analysis conditional on the regressors.
- In dynamic models, where y_t may depend on y_{t-1}, a conditional analysis is not sufficiently general, since we may want to predict into the future many periods out, so we need to consider the behavior of $\hat{\beta}$ and the relevant test statistics unconditional on X.

The model we'll deal will involve a combination of the following assumptions

Assumption 10. ***Linearity****: the model is a linear function of the parameter vector* β_0 *:*

$$y_t = x_t'\beta_0 + \varepsilon_t,$$

or in matrix form,

$$y = X\beta_0 + \varepsilon,$$

where y *is* $n \times 1$, $X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}'$, *where* x_t *is* $K \times 1$, *and* β_0 *and* ε *are conformable.*

Assumption 11. ***Stochastic, linearly independent regressors***

X *has rank* K *with probability 1*

X *is stochastic*

$\lim_{n\to\infty} \Pr\left(\frac{1}{n}X'X = Q_X\right) = 1$, *where* Q_X *is a finite positive definite matrix.*

Assumption 12. ***Central limit theorem***

$n^{-1/2}X'\varepsilon \xrightarrow{d} N(0, Q_X\sigma_0^2)$

Assumption 13. ***Normality (Optional)****:* $\varepsilon|X \sim N(0, \sigma^2 I_n)$*:* ϵ *is normally distributed*

Assumption 14. ***Strongly exogenous regressors****. The regressors* $\mathbf{X}$ *are strongly exogenous if*

$$\mathcal{E}(\varepsilon_t|\mathbf{X}) \;=\; 0, \forall t \tag{6.1}$$

Assumption 15. ***Weakly exogenous regressors****: The regressors are weakly exogenous if*

$$\mathcal{E}(\varepsilon_t|\mathbf{x}_t) = 0, \forall t$$

In both cases, $\mathbf{x}_t'\beta$ is the conditional mean of y_t given $\mathbf{x}_t$: $E(y_t|\mathbf{x}_t) = \mathbf{x}_t'\beta$

6.1 Case 1

Normality of ε, strongly exogenous regressors

In this case,

$$\hat{\beta} = \beta_0 + (X'X)^{-1}X'\varepsilon$$

$$\begin{aligned}\mathcal{E}(\hat{\beta}|X) &= \beta_0 + (X'X)^{-1}X'\mathcal{E}(\varepsilon|X) \\ &= \beta_0\end{aligned}$$

and since this holds for all X, $E(\hat{\beta}) = \beta$, unconditional on X. Likewise,

$$\hat{\beta}|X \sim N\left(\beta, (X'X)^{-1}\sigma_0^2\right)$$

- If the density of X is $d\mu(X)$, the marginal density of $\hat{\beta}$ is obtained by multiplying the conditional density by $d\mu(X)$ and integrating over X. Doing this leads to a nonnormal density for $\hat{\beta}$, in small samples.

- However, conditional on X, the usual test statistics have the t, F and χ^2 distributions. *Importantly,* these distributions don't depend on X, so when marginalizing to obtain the unconditional distribution, nothing changes. The tests are valid in small samples.
- Summary: When X is stochastic but strongly exogenous and ε is normally distributed:
  1. $\hat{\beta}$ is unbiased
  2. $\hat{\beta}$ is nonnormally distributed
  3. The usual test statistics have the same distribution as with nonstochastic X.
  4. The Gauss-Markov theorem still holds, since it holds conditionally on X, and this is true for all X.
  5. Asymptotic properties are treated in the next section.

6.2 Case 2

ε nonnormally distributed, strongly exogenous regressors

The unbiasedness of $\hat{\beta}$ carries through as before. However, the argument regarding test statistics doesn't hold, due to nonnormality of ε. Still, we have

$$\begin{aligned} \hat{\beta} &= \beta_0 + (X'X)^{-1}X'\varepsilon \\ &= \beta_0 + \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n} \end{aligned}$$

Now

$$\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{p} Q_X^{-1}$$

by assumption, and

$$\frac{X'\varepsilon}{n} = \frac{n^{-1/2}X'\varepsilon}{\sqrt{n}} \xrightarrow{p} 0$$

since the numerator converges to a $N(0, Q_X\sigma^2)$ r.v. and the denominator still goes to infinity. We have unbiasedness and the variance disappearing, so, *the estimator is consistent*:

$$\hat{\beta} \xrightarrow{p} \beta_0.$$

Considering the asymptotic distribution

$$\begin{aligned}\sqrt{n}\left(\hat{\beta} - \beta_0\right) &= \sqrt{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n} \\ &= \left(\frac{X'X}{n}\right)^{-1} n^{-1/2}X'\varepsilon\end{aligned}$$

so

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, Q_X^{-1}\sigma_0^2)$$

directly following the assumptions. *Asymptotic normality of the estimator still holds.* Since the asymptotic results on all test statistics only require this, all the previous asymptotic results on test statistics are also valid in this case.

- Summary: Under strongly exogenous regressors, with ε normal or nonnormal, $\hat{\beta}$ has the proper-

ties:

1. Unbiasedness
2. Consistency
3. Gauss-Markov theorem holds, since it holds in the previous case and doesn't depend on normality.
4. Asymptotic normality
5. Tests are asymptotically valid
6. Tests are not valid in small samples if the error is normally distributed

6.3 Case 3

Weakly exogenous regressors

An important class of models are *dynamic models*, where lagged dependent variables have an impact on the current value. A simple version of these models that captures the important points is

$$\begin{aligned} y_t &= z_t'\alpha + \sum_{s=1}^{p} \gamma_s y_{t-s} + \varepsilon_t \\ &= x_t'\beta + \varepsilon_t \end{aligned}$$

where now x_t contains lagged dependent variables. Clearly, even with $E(\epsilon_t|\mathbf{x}_t) = 0$, X and ε are not uncorrelated, so one can't show unbiasedness. For example,

$$\mathcal{E}(\varepsilon_{t-1}x_t) \neq 0$$

since x_t contains y_{t-1} (which is a function of ε_{t-1}) as an element.

- This fact implies that all of the small sample properties such as unbiasedness, Gauss-Markov theorem, and small sample validity of test statistics *do not hold* in this case. Recall Figure 3.7. This is a case of weakly exogenous regressors, and we see that the OLS estimator is biased in this case.

- Nevertheless, under the above assumptions, all asymptotic properties continue to hold, using the same arguments as before.

6.4 When are the assumptions reasonable?

The two assumptions we've added are

1. $\lim_{n\to\infty} \Pr\left(\frac{1}{n}X'X = Q_X\right) = 1$, a Q_X finite positive definite matrix.
2. $n^{-1/2}X'\varepsilon \xrightarrow{d} N(0, Q_X\sigma_0^2)$

The most complicated case is that of dynamic models, since the other cases can be treated as nested in this case. There exist a number of central limit theorems for dependent processes, many of which are fairly technical. We won't enter into details (see Hamilton, Chapter 7 if you're interested). A main requirement for use of standard asymptotics for a dependent sequence

$$\{s_t\} = \{\frac{1}{n}\sum_{t=1}^{n} z_t\}$$

to converge in probability to a finite limit is that z_t be *stationary*, in some sense.

- Strong stationarity requires that the joint distribution of the set

$$\{z_t, z_{t+s}, z_{t-q}, ...\}$$

not depend on t.

- Covariance (weak) stationarity requires that the first and second moments of this set not depend on t.

- An example of a sequence that doesn't satisfy this is an AR(1) process with a unit root (a *random walk)*:

$$\begin{aligned} x_t &= x_{t-1} + \varepsilon_t \\ \varepsilon_t &\sim IIN(0, \sigma^2) \end{aligned}$$

One can show that the variance of x_t depends upon t in this case, so it's not weakly stationary.

- The series $\sin t + \epsilon_t$ has a first moment that depends upon t, so it's not weakly stationary either.

Stationarity prevents the process from trending off to plus or minus infinity, and prevents cyclical behavior which would allow correlations between far removed z_t znd z_s to be high. *Draw a picture here.*

- In summary, the assumptions are reasonable when the stochastic conditioning variables have variances that are finite, and are not too strongly dependent. The AR(1) model with unit root is an example of a case where the dependence is too strong for standard asymptotics to apply.

- The study of nonstationary processes is an important part of econometrics, but it isn't in the scope of this course.

6.5 Exercises

1. Show that for two random variables A and B, if $E(A|B) = 0$, then $E\left(Af(B)\right) = 0$. How is this used in the proof of the Gauss-Markov theorem?

2. Is it possible for an AR(1) model for time series data, *e.g.*, $y_t = 0 + 0.9y_{t-1} + \varepsilon_t$ satisfy weak exogeneity? Strong exogeneity? Discuss.

Chapter 7

Data problems

In this section we'll consider problems associated with the regressor matrix: collinearity, missing observations and measurement error.

7.1 Collinearity

Motivation: Data on Mortality and Related Factors

The data set mortality.data contains annual data from 1947 - 1980 on death rates in the U.S., along with data on factors like smoking and consumption of alcohol. The data description is:

> DATA4-7: Death rates in the U.S. due to coronary heart disease and their determinants. Data compiled by Jennifer Whisenand

- chd = death rate per 100,000 population (Range 321.2 - 375.4)

- cal = Per capita consumption of calcium per day in grams (Range 0.9 - 1.06)
- unemp = Percent of civilian labor force unemployed in 1,000 of persons 16 years and older (Range 2.9 - 8.5)
- cig = Per capita consumption of cigarettes in pounds of tobacco by persons 18 years and older– approx. 339 cigarettes per pound of tobacco (Range 6.75 - 10.46)
- edfat = Per capita intake of edible fats and oil in pounds–includes lard, margarine and butter (Range 42 - 56.5)
- meat = Per capita intake of meat in pounds–includes beef, veal, pork, lamb and mutton (Range 138 - 194.8)
- spirits = Per capita consumption of distilled spirits in taxed gallons for individuals 18 and older (Range 1 - 2.9)
- beer = Per capita consumption of malted liquor in taxed gallons for individuals 18 and older (Range 15.04 - 34.9)
- wine = Per capita consumption of wine measured in taxed gallons for individuals 18 and older (Range 0.77 - 2.65)

Consider estimation results for several models:

$$\widehat{\text{chd}} = \underset{(58.939)}{334.914} + \underset{(5.156)}{5.41216}\,\text{cig} + \underset{(7.373)}{36.8783}\,\text{spirits} - \underset{(1.2513)}{5.10365}\,\text{beer} + \underset{(12.735)}{13.9764}\,\text{wine}$$

$$T = 34 \quad \bar{R}^2 = 0.5528 \quad F(4, 29) = 11.2 \quad \hat{\sigma} = 9.9945$$

(standard errors in parentheses)

$$\widehat{\text{chd}} = \underset{(56.624)}{353.581} + \underset{(4.7523)}{3.17560}\,\text{cig} + \underset{(7.275)}{38.3481}\,\text{spirits} - \underset{(1.0102)}{4.28816}\,\text{beer}$$

$$T = 34 \quad \bar{R}^2 = 0.5498 \quad F(3, 30) = 14.433 \quad \hat{\sigma} = 10.028$$

(standard errors in parentheses)

$$\widehat{\text{chd}} = \underset{(67.21)}{243.310} + \underset{(6.1508)}{10.7535}\,\text{cig} + \underset{(8.0359)}{22.8012}\,\text{spirits} - \underset{(12.638)}{16.8689}\,\text{wine}$$

$$T = 34 \quad \bar{R}^2 = 0.3198 \quad F(3, 30) = 6.1709 \quad \hat{\sigma} = 12.327$$

(standard errors in parentheses)

$$\widehat{\text{chd}} = \underset{(49.119)}{181.219} + \underset{(4.4371)}{16.5146}\,\text{cig} + \underset{(6.2079)}{15.8672}\,\text{spirits}$$

$$T = 34 \quad \bar{R}^2 = 0.3026 \quad F(2, 31) = 8.1598 \quad \hat{\sigma} = 12.481$$

(standard errors in parentheses)

Note how the signs of the coefficients change depending on the model, and that the magnitudes of the parameter estimates vary a lot, too. The parameter estimates are highly sensitive to the particular model we estimate. Why? We'll see that the problem is that the data exhibit *collinearity*.

Collinearity: definition

Collinearity is the existence of linear relationships amongst the regressors. We can always write

$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_K \mathbf{x}_K + v = 0$$

where $\mathbf{x}_i$ is the i^{th} column of the regressor matrix X, and v is an $n \times 1$ vector. In the case that there exists collinearity, the variation in v is relatively small, so that there is an approximately exact linear relation between the regressors.

- "relative" and "approximate" are imprecise, so it's difficult to define when collinearilty exists.

In the extreme, if there are exact linear relationships (every element of v equal) then $\rho(X) < K$, so $\rho(X'X) < K$, so $X'X$ is not invertible and the OLS estimator is not uniquely defined. For example,

if the model is

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t \\ x_{2t} &= \alpha_1 + \alpha_2 x_{3t} \end{aligned}$$

then we can write

$$\begin{aligned} y_t &= \beta_1 + \beta_2 (\alpha_1 + \alpha_2 x_{3t}) + \beta_3 x_{3t} + \varepsilon_t \\ &= \beta_1 + \beta_2 \alpha_1 + \beta_2 \alpha_2 x_{3t} + \beta_3 x_{3t} + \varepsilon_t \\ &= (\beta_1 + \beta_2 \alpha_1) + (\beta_2 \alpha_2 + \beta_3) x_{3t} \\ &= \gamma_1 + \gamma_2 x_{3t} + \varepsilon_t \end{aligned}$$

- The $\gamma' s$ can be consistently estimated, but since the γ's define two equations in three $\beta' s$, the $\beta' s$ can't be consistently estimated (there are multiple values of β that solve the first order conditions). The $\beta' s$ are *unidentified* in the case of perfect collinearity.
- Perfect collinearity is unusual, except in the case of an error in construction of the regressor matrix, such as including the same regressor twice.

Another case where perfect collinearity may be encountered is with models with dummy variables, if one is not careful. Consider a model of rental price (y_i) of an apartment. This could depend factors such as size, quality etc., collected in x_i, as well as on the location of the apartment. Let $B_i = 1$ if the i^{th} apartment is in Barcelona, $B_i = 0$ otherwise. Similarly, define G_i, T_i and L_i for Girona, Tarragona

and Lleida. One could use a model such as

$$y_i = \beta_1 + \beta_2 B_i + \beta_3 G_i + \beta_4 T_i + \beta_5 L_i + x_i'\gamma + \varepsilon_i$$

In this model, $B_i + G_i + T_i + L_i = 1$, $\forall i$, so there is an exact relationship between these variables and the column of ones corresponding to the constant. One must either drop the constant, or one of the qualitative variables.

A brief aside on dummy variables

Dummy variable: A dummy variable is a binary-valued variable that indicates whether or not some condition is true. It is customary to assign the value 1 if the condition is true, and 0 if the condition is false.

Dummy variables are used essentially like any other regressor. Use d to indicate that a variable is a dummy, so that variables like d_t and d_{t2} are understood to be dummy variables. Variables like x_t and x_{t3} are ordinary continuous regressors. You know how to interpret the following models:

$$y_t = \beta_1 + \beta_2 d_t + \epsilon_t$$

$$y_t = \beta_1 d_t + \beta_2 (1 - d_t) + \epsilon_t$$

$$y_t = \beta_1 + \beta_2 d_t + \beta_3 x_t + \epsilon_t$$

Interaction terms: an interaction term is the product of two variables, so that the effect of one

variable on the dependent variable depends on the value of the other. The following model has an interaction term. Note that $\frac{\partial E(y|x)}{\partial x} = \beta_3 + \beta_4 d_t$. The slope depends on the value of d_t.

$$y_t = \beta_1 + \beta_2 d_t + \beta_3 x_t + \beta_4 d_t x_t + \epsilon_t$$

Multiple dummy variables: we can use more than one dummy variable in a model. We will study models of the form

$$y_t = \beta_1 + \beta_2 d_{t1} + \beta_3 d_{t2} + \beta_4 x_t + \epsilon_t$$

$$y_t = \beta_1 + \beta_2 d_{t1} + \beta_3 d_{t2} + \beta_4 d_{t1} d_{t2} + \beta_5 x_t + \epsilon_t$$

Incorrect usage: You should understand why the following models are not correct usages of dummy variables:

1. overparameterization:
$$y_t = \beta_1 + \beta_2 d_t + \beta_3 (1 - d_t) + \epsilon_t$$

2. multiple values assigned to multiple categories. Suppose that we a condition that defines 4 possible categories, and we create a variable $d = 1$ if the observation is in the first category, $d = 2$ if in the second, etc. (This is not strictly speaking a dummy variable, according to our definition). Why is the following model not a good one?

$$y_t = \beta_1 + \beta_2 d + \epsilon$$

What is the correct way to deal with this situation?

Multiple parameterizations. To formulate a model that conditions on a given set of categorical information, there are multiple ways to use dummy variables. For example, the two models

$$y_t = \beta_1 d_t + \beta_2(1 - d_t) + \beta_3 x_t + \beta_4 d_t x_t + \epsilon_t$$

and

$$y_t = \alpha_1 + \alpha_2 d_t + \alpha_3 x_t d_t + \alpha_4 x_t(1 - d_t) + \epsilon_t$$

are equivalent. You should know what are the 4 equations that relate the β_j parameters to the α_j parameters, $j = 1, 2, 3, 4$. You should know how to interpret the parameters of both models.

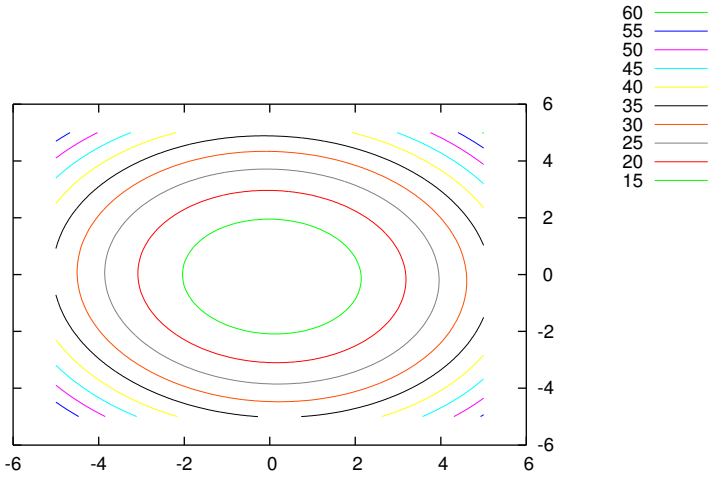
Back to collinearity

The more common case, if one doesn't make mistakes such as these, is the existence of inexact linear relationships, *i.e.*, correlations between the regressors that are less than one in absolute value, but not zero. The basic problem is that when two (or more) variables move together, it is difficult to determine their separate influences.

Example 16. Two children are in a room, along with a broken lamp. Both say "I didn't do it!". How can we tell who broke the lamp?

Lack of knowledge about the separate influences of variables is reflected in imprecise estimates, *i.e.*, estimates with high variances. *With economic data, collinearity is commonly encountered, and is often a severe problem.*

Figure 7.1: $s(\beta)$ when there is no collinearity



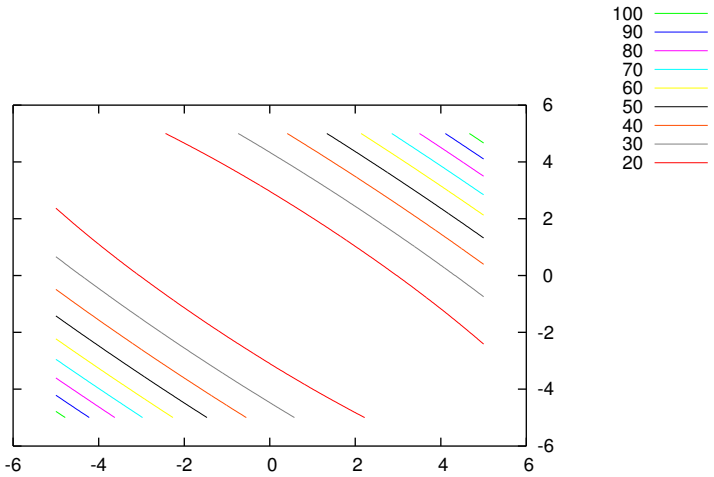
When there is collinearity, the minimizing point of the objective function that defines the OLS estimator ($s(\beta)$, the sum of squared errors) is relatively poorly defined. This is seen in Figures 7.1 and 7.2.

To see the effect of collinearity on variances, partition the regressor matrix as

$$X = \begin{bmatrix} \mathbf{x} & W \end{bmatrix}$$

where $\mathbf{x}$ is the first column of X (note: we can interchange the columns of X isf we like, so there's no loss of generality in considering the first column). Now, the variance of $\hat{\beta}$, under the classical

Figure 7.2: $s(\beta)$ when there is collinearity



assumptions, is

$$V(\hat{\beta}) = (X'X)^{-1}\sigma^2$$

Using the partition,

$$X'X = \begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'W \\ W'\mathbf{x} & W'W \end{bmatrix}$$

and following a rule for partitioned inversion,

$$\begin{aligned} (X'X)^{-1}_{1,1} &= \left(\mathbf{x}'\mathbf{x} - \mathbf{x}'W(W'W)^{-1}W'\mathbf{x}\right)^{-1} \\ &= \left(\mathbf{x}'\left(I_n - W(W'W)'^1W'\right)\mathbf{x}\right)^{-1} \\ &= \left(ESS_{\mathbf{x}|W}\right)^{-1} \end{aligned}$$

where by $ESS_{\mathbf{x}|W}$ we mean the error sum of squares obtained from the regression

$$\mathbf{x} = W\lambda + v.$$

Since

$$R^2 = 1 - ESS/TSS,$$

we have

$$ESS = TSS(1 - R^2)$$

so the variance of the coefficient corresponding to $\mathbf{x}$ is

$$V(\hat{\beta}_{\mathbf{x}}) = \frac{\sigma^2}{TSS_{\mathbf{x}}(1 - R^2_{\mathbf{x}|W})} \tag{7.1}$$

We see three factors influence the variance of this coefficient. It will be high if

1. σ^2 is large
2. There is little variation in $\mathbf{x}$. *Draw a picture here.*
3. There is a strong linear relationship between x and the other regressors, so that W can explain the movement in $\mathbf{x}$ well. In this case, $R^2_{\mathbf{x}|W}$ will be close to 1. As $R^2_{\mathbf{x}|W} \to 1, V(\hat{\beta}_{\mathbf{x}}) \to \infty$.

The last of these cases is collinearity.

Intuitively, when there are strong linear relations between the regressors, it is difficult to determine the separate influence of the regressors on the dependent variable. This can be seen by comparing the OLS objective function in the case of no correlation between regressors with the objective function with correlation between the regressors. See the figures nocollin.ps (no correlation) and collin.ps (correlation), available on the web site.

Example 17. The Octave script DataProblems/collinearity.m performs a Monte Carlo study with correlated regressors. The model is $y = 1 + x_2 + x_3 + \epsilon$, where the correlation between x_2 and $x_3$can be set. Three estimators are used: OLS, OLS dropping x_3 (a false restriction), and restricted LS using $\beta_2 = \beta_3$ (a true restriction). The output when the correlation between the two regressors is 0.9 is

```
octave:1> collinearity

Contribution received from node 0.  Received so far: 500

Contribution received from node 0.  Received so far: 1000
```

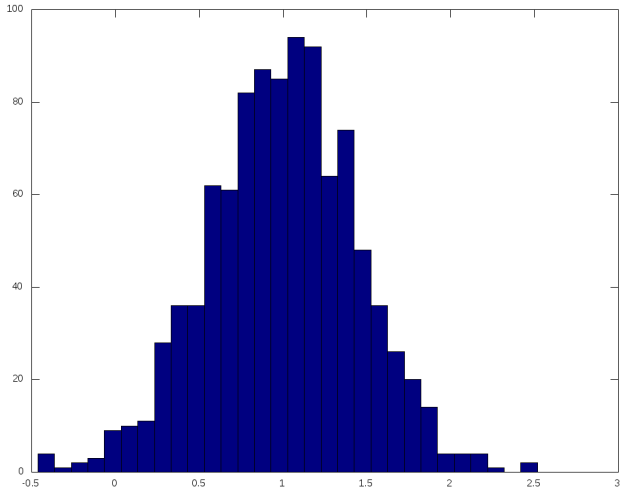
```
correlation between x2 and x3: 0.900000

descriptive statistics for 1000 OLS replications
      mean  st. dev.       min       max
     0.996     0.182     0.395     1.574
     0.996     0.444    -0.463     2.517
     1.008     0.436    -0.342     2.301
descriptive statistics for 1000 OLS replications, dropping x3
      mean  st. dev.       min       max
     0.999     0.198     0.330     1.696
     1.905     0.207     1.202     2.651
descriptive statistics for 1000 Restricted OLS replications, b2=b3
      mean  st. dev.       min       max
     0.998     0.179     0.433     1.574
     1.002     0.096     0.663     1.339
     1.002     0.096     0.663     1.339
octave:2>
```

Figure 7.3 shows histograms for the estimated β_2, for each of the three estimators.

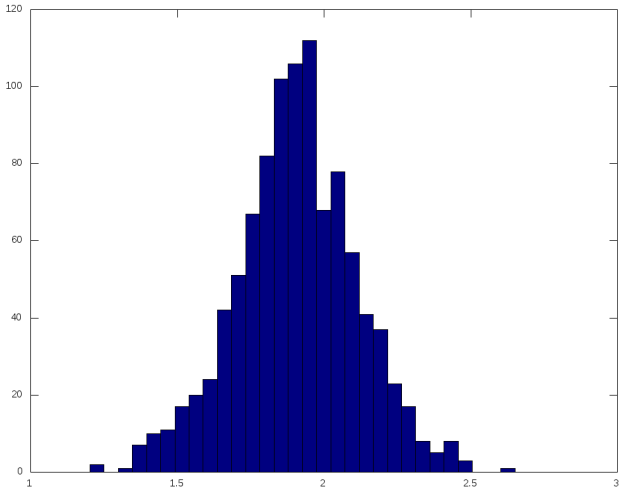
- repeat the experiment with a lower value of rho, and note how the standard errors of the OLS estimator change.

Figure 7.3: Collinearity: Monte Carlo results

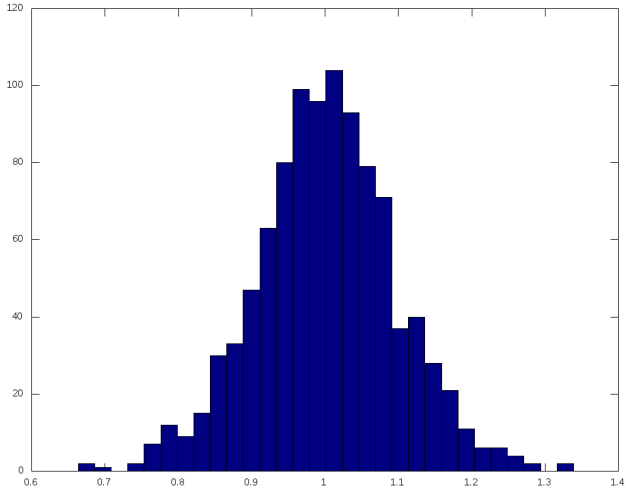
(a) OLS,$\hat{\beta}_2$



(b) OLS,$\hat{\beta}_2$, dropping x3



(c) Restricted LS,$\hat{\beta}_2$, with true restriction$\beta_2 = \beta_3$



Detection of collinearity

The best way is simply to regress each explanatory variable in turn on the remaining regressors. If any of these auxiliary regressions has a high R^2, there is a problem of collinearity. Furthermore, this procedure identifies which parameters are affected.

- Sometimes, we're only interested in certain parameters. Collinearity isn't a problem if it doesn't affect what we're interested in estimating.

An alternative is to examine the matrix of correlations between the regressors. High correlations are sufficient but not necessary for severe collinearity.

Also indicative of collinearity is that the model fits well (high R^2), but none of the variables is significantly different from zero (e.g., their separate influences aren't well determined).

In summary, the artificial regressions are the best approach if one wants to be careful.

Example 18. Nerlove data and collinearity. The simple Nerlove model is

$$\ln C = \beta_1 + \beta_2 \ln Q + \beta_3 \ln P_L + \beta_4 \ln P_F + \beta_5 \ln P_K + \epsilon$$

When this model is estimated by OLS, some coefficients are not significant (see subsection 3.8). This may be due to collinearity.The Octave script DataProblems/NerloveCollinearity.m checks the regressors for collinearity. If you run this, you will see that collinearity is not a problem with this data. Why is the coefficient of $\ln P_K$ not significantly different from zero?

Dealing with collinearity

More information

Collinearity is a problem of an uninformative sample. The first question is: is all the available information being used? Is more data available? Are there coefficient restrictions that have been neglected? *Picture illustrating how a restriction can solve problem of perfect collinearity.*

Stochastic restrictions and ridge regression

Supposing that there is no more data or neglected restrictions, one possibility is to change perspectives, to Bayesian econometrics. One can express prior beliefs regarding the coefficients using stochastic restrictions. A stochastic linear restriction would be something of the form

$$R\beta = r + v$$

where R and r are as in the case of exact linear restrictions, but v is a random vector. For example, the model could be

$$\begin{aligned} y &= X\beta + \varepsilon \\ R\beta &= r + v \\ \begin{pmatrix} \varepsilon \\ v \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 I_n & 0_{n\times q} \\ 0_{q\times n} & \sigma_v^2 I_q \end{pmatrix} \right) \end{aligned}$$

This sort of model isn't in line with the classical interpretation of parameters as constants: according to this interpretation the left hand side of $R\beta = r + v$ is constant but the right is random. This

model does fit the Bayesian perspective: we combine information coming from the model and the data, summarized in

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &\sim N(0, \sigma_\varepsilon^2 I_n) \end{aligned}$$

with prior beliefs regarding the distribution of the parameter, summarized in

$$R\beta \sim N(r, \sigma_v^2 I_q)$$

Since the sample is random it is reasonable to suppose that $\mathcal{E}(\varepsilon v') = 0$, which is the last piece of information in the specification. How can you estimate using this model? The solution is to treat the restrictions as artificial data. Write

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ v \end{bmatrix}$$

This model is heteroscedastic, since $\sigma_\varepsilon^2 \neq \sigma_v^2$. Define the *prior precision* $k = \sigma_\varepsilon / \sigma_v$. This expresses the degree of belief in the restriction relative to the variability of the data. Supposing that we specify k, then the model

$$\begin{bmatrix} y \\ kr \end{bmatrix} = \begin{bmatrix} X \\ kR \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ kv \end{bmatrix}$$

is homoscedastic and can be estimated by OLS. Note that this estimator is biased. It is consistent, however, given that k is a fixed constant, even if the restriction is false (this is in contrast to the case of false exact restrictions). To see this, note that there are Q restrictions, where Q is the number of

rows of R. As $n \to \infty$, these Q artificial observations have no weight in the objective function, so the estimator has the same limiting objective function as the OLS estimator, and is therefore consistent.

To motivate the use of stochastic restrictions, consider the expectation of the squared length of $\hat{\beta}$:

$$\begin{aligned}
\mathcal{E}(\hat{\beta}'\hat{\beta}) &= \mathcal{E}\left\{\left(\beta + (X'X)^{-1}X'\varepsilon\right)'\left(\beta + (X'X)^{-1}X'\varepsilon\right)\right\} \\
&= \beta'\beta + \mathcal{E}\left(\varepsilon'X(X'X)^{-1}(X'X)^{-1}X'\varepsilon\right) \\
&= \beta'\beta + Tr\,(X'X)^{-1}\sigma^2 \\
&= \beta'\beta + \sigma^2 \sum_{i=1}^{K} \lambda_i (\text{the trace is the sum of eigenvalues}) \\
&> \beta'\beta + \lambda_{\max(X'X^{-1})}\sigma^2 (\text{the eigenvalues are all positive, since} X'X \text{ is p.d.}
\end{aligned}$$

so

$$\mathcal{E}(\hat{\beta}'\hat{\beta}) > \beta'\beta + \frac{\sigma^2}{\lambda_{\min(X'X)}}$$

where $\lambda_{\min(X'X)}$ is the minimum eigenvalue of $X'X$ (which is the inverse of the maximum eigenvalue of $(X'X)^{-1}$). As collinearity becomes worse and worse, $X'X$ becomes more nearly singular, so $\lambda_{\min(X'X)}$ tends to zero (recall that the determinant is the product of the eigenvalues) and $\mathcal{E}(\hat{\beta}'\hat{\beta})$ tends to infinite. On the other hand, $\beta'\beta$ is finite.

Now considering the restriction $I_K\beta = 0 + v$. With this restriction the model becomes

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ kI_K \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ kv \end{bmatrix}$$

and the estimator is

$$\begin{aligned}\hat{\beta}_{ridge} &= \left(\begin{bmatrix} X' & kI_K \end{bmatrix}\begin{bmatrix} X \\ kI_K \end{bmatrix}\right)^{-1}\begin{bmatrix} X' & I_K \end{bmatrix}\begin{bmatrix} y \\ 0 \end{bmatrix} \\ &= \left(X'X + k^2 I_K\right)^{-1} X'y\end{aligned}$$

This is the ordinary *ridge regression* estimator. The ridge regression estimator can be seen to add $k^2 I_K$, which is nonsingular, to $X'X$, which is more and more nearly singular as collinearity becomes worse and worse. As $k \to \infty$, the restrictions tend to $\beta = 0$, that is, the coefficients are shrunken toward zero. Also, the estimator tends to

$$\hat{\beta}_{ridge} = \left(X'X + k^2 I_K\right)^{-1} X'y \to \left(k^2 I_K\right)^{-1} X'y = \frac{X'y}{k^2} \to 0$$

so $\hat{\beta}'_{ridge}\hat{\beta}_{ridge} \to 0$. This is clearly a false restriction in the limit, if our original model is at all sensible.

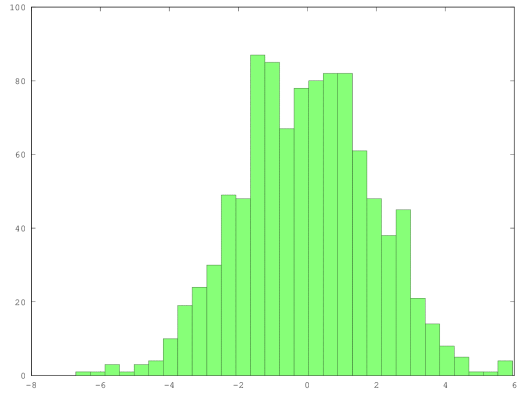
There should be some amount of shrinkage that is in fact a true restriction. The problem is to determine the k such that the restriction is correct. The interest in ridge regression centers on the fact that it can be shown that there exists a k such that $MSE(\hat{\beta}_{ridge}) < \hat{\beta}_{OLS}$. The problem is that this k depends on β and σ^2, which are unknown.

The ridge trace method plots $\hat{\beta}'_{ridge}\hat{\beta}_{ridge}$ as a function of k, and chooses the value of k that "artistically" seems appropriate (e.g., where the effect of increasing k dies off). *Draw picture here.* This means of choosing k is obviously subjective. This is not a problem from the Bayesian perspective: the choice of k reflects prior beliefs about the length of β.

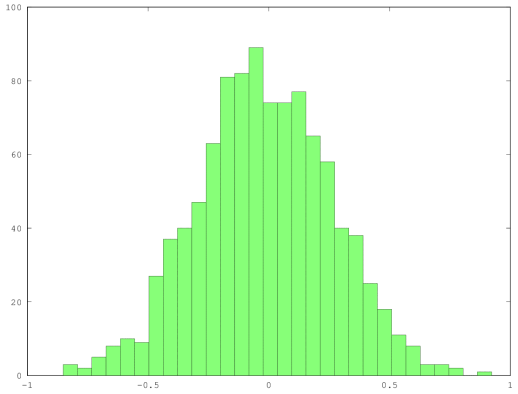
In summary, the ridge estimator offers some hope, but it is impossible to guarantee that it will outperform the OLS estimator. Collinearity is a fact of life in econometrics, and there is no clear

Figure 7.4: OLS and Ridge regression

(a) OLS



(b) Ridge



solution to the problem.

The Octave script DataProblems/RidgeRegression.m does a Monte Carlo study that shows that ridge regression can help to deal with collinearity. This script generates Figures and, which show the Monte Carlo sampling frequency of the OLS and ridge estimators, after subtracting the true parameter values. You can see that the ridge estimator has much lower RMSE.

7.2 Measurement error

Measurement error is exactly what it says, either the dependent variable or the regressors are measured with error. Thinking about the way economic data are reported, measurement error is probably quite prevalent. For example, estimates of growth of GDP, inflation, etc. are commonly revised several times. Why should the last revision necessarily be correct?

Error of measurement of the dependent variable

Measurement errors in the dependent variable and the regressors have important differences. First consider error in measurement of the dependent variable. The data generating process is presumed to be

$$\begin{aligned} y^* &= X\beta + \varepsilon \\ y &= y^* + v \\ v_t &\sim iid(0, \sigma_v^2) \end{aligned}$$

where $y^* = y + v$ is the unobservable true dependent variable, and y is what is observed. We assume that ε and v are independent and that $y^* = X\beta + \varepsilon$ satisfies the classical assumptions. Given this, we have

$$y + v = X\beta + \varepsilon$$

so

$$\begin{aligned} y &= X\beta + \varepsilon - v \\ &= X\beta + \omega \\ \omega_t &\sim iid(0, \sigma_\varepsilon^2 + \sigma_v^2) \end{aligned}$$

- As long as v is uncorrelated with X, this model satisfies the classical assumptions and can be estimated by OLS. This type of measurement error isn't a problem, then, except in that the increased variability of the error term causes an increase in the variance of the OLS estimator (see equation 7.1).

Error of measurement of the regressors

The situation isn't so good in this case. The DGP is

$$\begin{aligned} y_t &= x_t^{*\prime}\beta + \varepsilon_t \\ x_t &= x_t^* + v_t \\ v_t &\sim iid(0, \Sigma_v) \end{aligned}$$

where Σ_v is a $K \times K$ matrix. Now X^* contains the true, unobserved regressors, and X is what is observed. Again assume that v is independent of ε, and that the model $y = X^*\beta + \varepsilon$ satisfies the classical assumptions. Now we have

$$\begin{aligned} y_t &= (x_t - v_t)' \beta + \varepsilon_t \\ &= x_t'\beta - v_t'\beta + \varepsilon_t \\ &= x_t'\beta + \omega_t \end{aligned}$$

The problem is that now there is a correlation between x_t and ω_t, since

$$\begin{aligned} \mathcal{E}(x_t\omega_t) &= \mathcal{E}\left((x_t^* + v_t)\left(-v_t'\beta + \varepsilon_t\right)\right) \\ &= -\Sigma_v\beta \end{aligned}$$

where

$$\Sigma_v = \mathcal{E}\left(v_t v_t'\right).$$

Because of this correlation, the OLS estimator is biased and inconsistent, just as in the case of autocorrelated errors with lagged dependent variables. In matrix notation, write the estimated model as

$$y = X\beta + \omega$$

We have that

$$\hat{\beta} = \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'y}{n}\right)$$

and

$$\begin{aligned} plim\left(\frac{X'X}{n}\right)^{-1} &= plim\frac{\left(X^{*\prime}+V'\right)\left(X^{*}+V\right)}{n} \\ &= \left(Q_{X^*}+\Sigma_v\right)^{-1} \end{aligned}$$

since X^* and V are independent, and

$$\begin{aligned} plim\frac{V'V}{n} &= \lim \mathcal{E}\frac{1}{n}\sum_{t=1}^{n} v_t v_t' \\ &= \Sigma_v \end{aligned}$$

Likewise,

$$\begin{aligned} plim\left(\frac{X'y}{n}\right) &= plim\frac{\left(X^{*\prime}+V'\right)\left(X^{*}\beta+\varepsilon\right)}{n} \\ &= Q_{X^*}\beta \end{aligned}$$

so

$$plim\hat{\beta} = \left(Q_{X^*} + \Sigma_v\right)^{-1} Q_{X^*}\beta$$

So we see that the least squares estimator is inconsistent when the regressors are measured with error.

- A potential solution to this problem is the instrumental variables (IV) estimator, which we'll discuss shortly.

Example 19. Measurement error in a dynamic model. Consider the model

$$\begin{aligned} y_t^* &= \alpha + \rho y_{t-1}^* + \beta x_t + \epsilon_t \\ y_t &= y_t^* + \upsilon_t \end{aligned}$$

where ϵ_t and υ_t are independent Gaussian white noise errors. Suppose that y_t^* is not observed, and instead we observe y_t. What are the properties of the OLS regression on the equation

$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \nu_t$$

? The error is

$$\begin{aligned} \nu_t &= y_t - \alpha - \rho y_{t-1} - \beta x_t \\ &= y_t^* + \upsilon_t - \alpha - \rho y_{t-1}^* - \rho\upsilon_{t-1} - \beta x_t \\ &= \alpha + \rho y_{t-1}^* + \beta x_t + \epsilon_t + \upsilon_t - \alpha - \rho y_{t-1}^* - \rho\upsilon_{t-1} - \beta x_t \\ &= \epsilon_t + \upsilon_t - \rho\upsilon_{t-1} \end{aligned}$$

So the error term is autocorrelated. Note that $y_{t-1} = \alpha + \rho y_{t-2} + \beta x_{t-1} + \nu_{t-1}$, so we the error ν_t

and the regressor y_{t-1} are correlated, because they share the common term v_{t-1}. This means that the equation

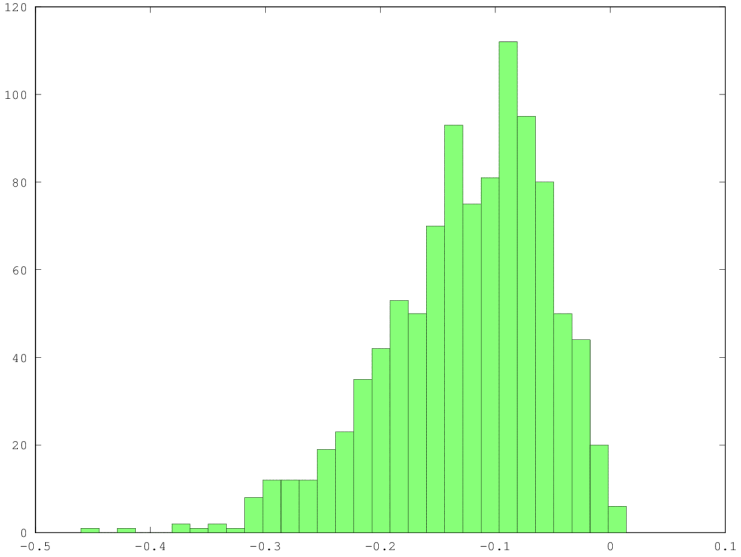
$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \nu_t$$

does not satisfy weak exogeneity, and the OLS estimator will be biased and inconsistent.

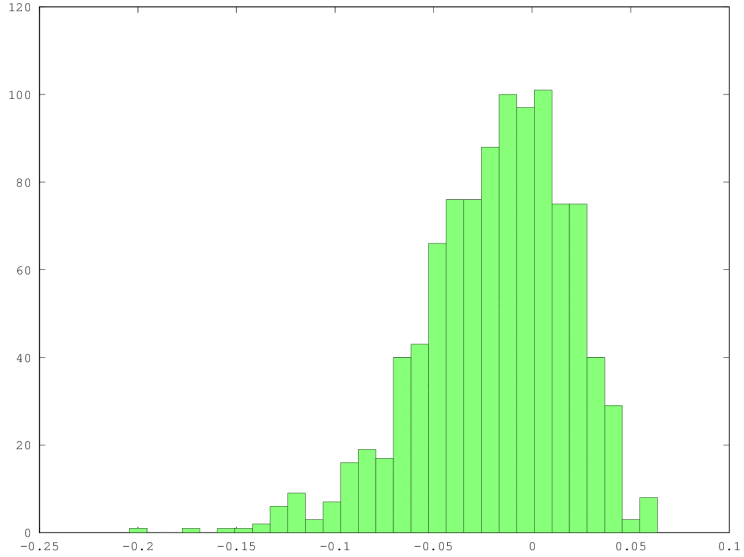
The Octave script DataProblems/MeasurementError.m does a Monte Carlo study. The sample size is $n = 100$. Figure 7.5 gives the results. The first panel shows a histogram for 1000 replications of $\hat{\rho} - \rho$, when $\sigma_\nu = 1$, so that there is significant measurement error. The second panel repeats this with $\sigma_\nu = 0$, so that there is not measurement error. Note that there is much more bias with measurement error. There is also bias without measurement error. This is due to the same reason that we saw bias in Figure 3.7: one of the classical assumptions (nonstochastic regressors) that guarantees unbiasedness of OLS does not hold for this model. Without measurement error, the OLS estimator *is* consistent. By re-running the script with larger n, you can verify that the bias disappears when $\sigma_\nu = 0$, but not when $\sigma_\nu > 0$.

Figure 7.5: $\hat{\rho} - \rho$ with and without measurement error

(a) with measurement error: $\sigma_\nu = 1$



(b) without measurement error: $\sigma_\nu = 0$



7.3 Missing observations

Missing observations occur quite frequently: time series data may not be gathered in a certain year, or respondents to a survey may not answer all questions. We'll consider two cases: missing observations on the dependent variable and missing observations on the regressors.

Missing observations on the dependent variable

In this case, we have

$$y = X\beta + \varepsilon$$

or

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

where y_2 is not observed. Otherwise, we assume the classical assumptions hold.

- A clear alternative is to simply estimate using the compete observations

$$y_1 = X_1\beta + \varepsilon_1$$

  Since these observations satisfy the classical assumptions, one could estimate by OLS.

- The question remains whether or not one could somehow replace the unobserved y_2 by a predictor, and improve over OLS in some sense. Let $\hat{y}_2$ be the predictor of y_2. Now

$$\begin{aligned} \hat{\beta} &= \left\{ \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}' \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right\}^{-1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ \hat{y}_2 \end{bmatrix} \\ &= \left[X_1'X_1 + X_2'X_2\right]^{-1} \left[X_1'y_1 + X_2'\hat{y}_2\right] \end{aligned}$$

Recall that the OLS fonc are

$$X'X\hat{\beta} = X'y$$

so if we regressed using only the first (complete) observations, we would have

$$X_1'X_1\hat{\beta}_1 = X_1'y_{1.}$$

Likewise, an OLS regression using only the second (filled in) observations would give

$$X_2'X_2\hat{\beta}_2 = X_2'\hat{y}_2.$$

Substituting these into the equation for the overall combined estimator gives

$$\begin{aligned}
\hat{\beta} &= [X_1'X_1 + X_2'X_2]^{-1}\left[X_1'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2\right] \\
&= [X_1'X_1 + X_2'X_2]^{-1} X_1'X_1\hat{\beta}_1 + [X_1'X_1 + X_2'X_2]^{-1} X_2'X_2\hat{\beta}_2 \\
&\equiv A\hat{\beta}_1 + (I_K - A)\hat{\beta}_2
\end{aligned}$$

where

$$A \equiv [X_1'X_1 + X_2'X_2]^{-1} X_1'X_1$$

and we use

$$\begin{aligned}
[X_1'X_1 + X_2'X_2]^{-1} X_2'X_2 &= [X_1'X_1 + X_2'X_2]^{-1}[(X_1'X_1 + X_2'X_2) - X_1'X_1] \\
&= I_K - [X_1'X_1 + X_2'X_2]^{-1} X_1'X_1 \\
&= I_K - A.
\end{aligned}$$

Now,

$$\mathcal{E}(\hat{\beta}) = A\beta + (I_K - A)\mathcal{E}\left(\hat{\beta}_2\right)$$

and this will be unbiased only if $\mathcal{E}\left(\hat{\beta}_2\right) = \beta$.

- The conclusion is that the filled in observations alone would need to define an unbiased estimator.

This will be the case only if

$$\hat{y}_2 = X_2\beta + \hat{\varepsilon}_2$$

where $\hat{\varepsilon}_2$ has mean zero. Clearly, it is difficult to satisfy this condition without knowledge of β.

- Note that putting $\hat{y}_2 = \bar{y}_1$ does not satisfy the condition and therefore leads to a biased estimator.

Exercise 20. Formally prove this last statement.

The sample selection problem

In the above discussion we assumed that the missing observations are random. The sample selection problem is a case where the missing observations are not random. Consider the model

$$y_t^* = x_t'\beta + \varepsilon_t$$

which is assumed to satisfy the classical assumptions. However, y_t^* is not always observed. What is observed is y_t defined as

$$y_t = y_t^* \text{ if } y_t^* \geq 0$$

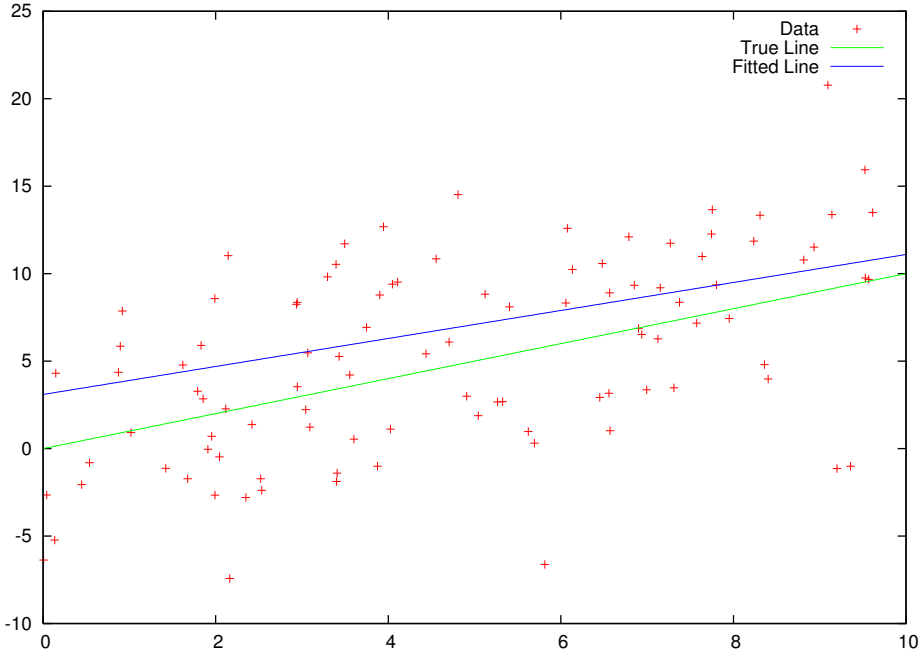
Or, in other words, y_t^* is missing when it is less than zero.

The difference in this case is that the missing values are not random: they are correlated with the x_t. Consider the case

$$y^* = x + \varepsilon$$

with $V(\varepsilon) = 25$, but using only the observations for which $y^* > 0$ to estimate. Figure 7.6 illustrates the bias. The Octave program is sampsel.m

Figure 7.6: Sample selection bias



There are means of dealing with sample selection bias, but we will not go into it here. One should at least be aware that nonrandom selection of the sample will normally lead to bias and inconsistency if the problem is not taken into account.

Missing observations on the regressors

Again the model is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

but we assume now that each row of X_2 has an unobserved component(s). Again, one could just estimate using the complete observations, but it may seem frustrating to have to drop observations simply because of a single missing variable. In general, if the unobserved X_2 is replaced by some prediction, X_2^*, then we are in the case of errors of observation. As before, this means that the OLS estimator is biased when X_2^* is used instead of X_2. Consistency is salvaged, however, as long as the number of missing observations doesn't increase with n.

- Including observations that have missing values replaced by *ad hoc* values can be interpreted as introducing false stochastic restrictions. In general, this introduces bias. It is difficult to determine whether MSE increases or decreases. Monte Carlo studies suggest that it is dangerous to simply substitute the mean, for example.

- In the case that there is only one regressor other than the constant, subtitution of $\bar{x}$ for the missing x_t *does not lead to bias*. This is a special case that doesn't hold for $K > 2$.

Exercise 21. Prove this last statement.

- In summary, if one is strongly concerned with bias, it is best to drop observations that have missing components. There is potential for reduction of MSE through filling in missing elements with intelligent guesses, but this could also increase MSE.

7.4 Missing regressors

Suppose that the model $y = X\beta + W\gamma + \epsilon$ satisfies the classical assumptions, so OLS would be a consistent estimator. However, let's suppose that the regressors W are not available in the sample. What are the properties of the OLS estimator of the model $y = X\beta + \omega$? We can think of this as a case of imposing false restrictions: $\gamma = 0$ when in fact $\gamma \neq 0$. We know that the restricted least squares estimator is biased and inconsistent, in general, when we impose false restrictions. Another way of thinking of this is to look to see if X and ω are correlated. We have

$$\begin{aligned} E(X_t\omega_t) &= E\left(X_t\left(W_t'\gamma + \epsilon_t\right)\right) \\ &= E(X_tW_t'\gamma) + E(X_t\epsilon_t) \\ &= E(X_tW_t'\gamma) \end{aligned}$$

where the last line follows because $E(X_t\epsilon_t) = 0$ by assumption. So, there will be correlation between the error and the regressors if there is collinearity between the included regressors X_t and the missing regressors W_t. If there is not, the OLS estimator will be consistent. Because the normal thing is to have collinearity between regressors, we expect that missing regressors will lead to bias and inconsistency of the OLS estimator.

7.5 Exercises

1. Consider the simple Nerlove model

$$\ln C = \beta_1 + \beta_2 \ln Q + \beta_3 \ln P_L + \beta_4 \ln P_F + \beta_5 \ln P_K + \epsilon$$

   When this model is estimated by OLS, some coefficients are not significant. We have seen that collinearity is not an important problem. Why is β_5 not significantly different from zero? Give an economic explanation.

2. For the model $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$,

   (a) verify that the level sets of the OLS criterion function (defined in equation 3.2) are straight lines when there is perfect collinearity

   (b) For this model with perfect collinearity, the OLS estimator does not exist. Depict what this statement means using a drawing.

   (c) Show how a restriction $R_1\beta_1 + R_2\beta_2 = r$ causes the restricted least squares estimator to exist, using a drawing.

Chapter 8

Functional form and nonnested tests

Though theory often suggests which conditioning variables should be included, and suggests the signs of certain derivatives, it is usually silent regarding the functional form of the relationship between the dependent variable and the regressors. For example, considering a cost function, one could have a Cobb-Douglas model

$$c = Aw_1^{\beta_1} w_2^{\beta_2} q^{\beta_q} e^{\varepsilon}$$

This model, after taking logarithms, gives

$$\ln c = \beta_0 + \beta_1 \ln w_1 + \beta_2 \ln w_2 + \beta_q \ln q + \varepsilon$$

where $\beta_0 = \ln A$. Theory suggests that $A > 0, \beta_1 > 0, \beta_2 > 0, \beta_3 > 0$. This model isn't compatible with a fixed cost of production since $c = 0$ when $q = 0$. Homogeneity of degree one in input prices suggests that $\beta_1 + \beta_2 = 1$, while constant returns to scale implies $\beta_q = 1$.

While this model may be reasonable in some cases, an alternative

$$\sqrt{c} = \beta_0 + \beta_1\sqrt{w_1} + \beta_2\sqrt{w_2} + \beta_q\sqrt{q} + \varepsilon$$

may be just as plausible. Note that $\sqrt{x}$ and $\ln(x)$ look quite alike, for certain values of the regressors, and up to a linear transformation, so it may be difficult to choose between these models.

The basic point is that many functional forms are compatible with the linear-in-parameters model, since this model can incorporate a wide variety of nonlinear transformations of the dependent variable and the regressors. For example, suppose that $g(\cdot)$ is a real valued function and that $x(\cdot)$ is a $K-$ vector-valued function. The following model is linear in the parameters but nonlinear in the variables:

$$\begin{aligned} x_t &= x(z_t) \\ y_t &= x_t'\beta + \varepsilon_t \end{aligned}$$

There may be P fundamental conditioning variables z_t, but there may be K regressors, where K may be smaller than, equal to or larger than P. For example, x_t could include squares and cross products of the conditioning variables in z_t.

8.1 Flexible functional forms

Given that the functional form of the relationship between the dependent variable and the regressors is in general unknown, one might wonder if there exist parametric models that can closely approximate a wide variety of functional relationships. A "Diewert-Flexible" functional form is defined as one such that the function, the vector of first derivatives and the matrix of second derivatives can take on an arbitrary value *at a single data point.* Flexibility in this sense clearly requires that there be at least

$$K = 1 + P + \left(P^2 - P\right)/2 + P$$

free parameters: one for each independent effect that we wish to model.

Suppose that the model is

$$y = g(x) + \varepsilon$$

A second-order Taylor's series expansion (with remainder term) of the function $g(x)$ about the point $x = 0$ is

$$g(x) = g(0) + x' D_x g(0) + \frac{x' D_x^2 g(0) x}{2} + R$$

Use the approximation, which simply drops the remainder term, as an approximation to $g(x)$:

$$g(x) \simeq g_K(x) = g(0) + x' D_x g(0) + \frac{x' D_x^2 g(0) x}{2}$$

As $x \to 0$, the approximation becomes more and more exact, in the sense that $g_K(x) \to g(x)$, $D_x g_K(x) \to D_x g(x)$ and $D_x^2 g_K(x) \to D_x^2 g(x)$. For $x = 0$, the approximation is exact, up to the second order. The idea behind many flexible functional forms is to note that $g(0)$, $D_x g(0)$ and $D_x^2 g(0)$

are all constants. If we treat them as parameters, the approximation will have exactly enough free parameters to approximate the function $g(x)$, which is of unknown form, exactly, up to second order, at the point $x = 0$. The model is

$$g_K(x) = \alpha + x'\beta + 1/2x'\Gamma x$$

so the regression model to fit is

$$y = \alpha + x'\beta + 1/2x'\Gamma x + \varepsilon$$

- While the regression model has enough free parameters to be Diewert-flexible, the question remains: is $plim\hat{\alpha} = g(0)$? Is $plim\hat{\beta} = D_x g(0)$? Is $plim\hat{\Gamma} = D_x^2 g(0)$?
- The answer is no, in general. The reason is that if we treat the true values of the parameters as these derivatives, then ε is forced to play the part of the remainder term, which is a function of x, so that x and ε are correlated in this case. As before, the estimator is biased in this case.
- A simpler example would be to consider a first-order T.S. approximation to a quadratic function. *Draw picture.*
- The conclusion is that "flexible functional forms" aren't really flexible in a useful statistical sense, in that neither the function itself nor its derivatives are consistently estimated, unless the function belongs to the parametric family of the specified functional form. In order to lead to consistent inferences, the regression model must be correctly specified.

The translog form

In spite of the fact that FFF's aren't really flexible for the purposes of econometric estimation and inference, they are useful, and they are certainly subject to less bias due to misspecification of the functional form than are many popular forms, such as the Cobb-Douglas or the simple linear in the variables model. The translog model is probably the most widely used FFF. This model is as above, except that the variables are subjected to a logarithmic tranformation. Also, the expansion point is usually taken to be the sample mean of the data, after the logarithmic transformation. The model is defined by

$$\begin{aligned} y &= \ln(c) \\ x &= \ln\left(\frac{z}{\bar{z}}\right) \\ &= \ln(z) - \ln(\bar{z}) \\ y &= \alpha + x'\beta + 1/2x'\Gamma x + \varepsilon \end{aligned}$$

In this presentation, the t subscript that distinguishes observations is suppressed for simplicity. Note that

$$\begin{aligned} \frac{\partial y}{\partial x} &= \beta + \Gamma x \\ &= \frac{\partial \ln(c)}{\partial \ln(z)} \text{ (the other part of } x \text{ is constant)} \\ &= \frac{\partial c}{\partial z}\frac{z}{c} \end{aligned}$$

which is the elasticity of c with respect to z. This is a convenient feature of the translog model. Note that at the means of the conditioning variables, $\bar{z}$, $x = 0$, so

$$\left.\frac{\partial y}{\partial x}\right|_{z=\bar{z}} = \beta$$

so the β are the first-order elasticities, at the means of the data.

To illustrate, consider that y is cost of production:

$$y = c(w, q)$$

where w is a vector of input prices and q is output. We could add other variables by extending q in the obvious manner, but this is supressed for simplicity. By Shephard's lemma, the conditional factor demands are

$$x = \frac{\partial c(w, q)}{\partial w}$$

and the cost shares of the factors are therefore

$$s = \frac{wx}{c} = \frac{\partial c(w, q)}{\partial w}\frac{w}{c}$$

which is simply the vector of elasticities of cost with respect to input prices. If the cost function is modeled using a translog function, we have

$$\begin{aligned} \ln(c) &= \alpha + x'\beta + z'\delta + 1/2 \left[\begin{array}{cc} x' & z \end{array} \right] \left[\begin{array}{cc} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{12}' & \Gamma_{22} \end{array} \right] \left[\begin{array}{c} x \\ z \end{array} \right] \\ &= \alpha + x'\beta + z'\delta + 1/2 x'\Gamma_{11}x + x'\Gamma_{12}z + 1/2 z^2\gamma_{22} \end{aligned}$$

where $x = \ln(w/\bar{w})$ (element-by-element division) and $z = \ln(q/\bar{q})$, and

$$\begin{aligned}
\Gamma_{11} &= \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{12} & \gamma_{22} \end{bmatrix} \\
\Gamma_{12} &= \begin{bmatrix} \gamma_{13} \\ \gamma_{23} \end{bmatrix} \\
\Gamma_{22} &= \gamma_{33}.
\end{aligned}$$

Note that symmetry of the second derivatives has been imposed.

Then the share equations are just

$$s = \beta + \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}$$

Therefore, the share equations and the cost equation have parameters in common. By pooling the equations together and imposing the (true) restriction that the parameters of the equations be the same, we can gain efficiency.

To illustrate in more detail, consider the case of two inputs, so

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

In this case the translog model of the logarithmic cost function is

$$\ln c = \alpha + \beta_1 x_1 + \beta_2 x_2 + \delta z + \frac{\gamma_{11}}{2} x_1^2 + \frac{\gamma_{22}}{2} x_2^2 + \frac{\gamma_{33}}{2} z^2 + \gamma_{12} x_1 x_2 + \gamma_{13} x_1 z + \gamma_{23} x_2 z$$

The two cost shares of the inputs are the derivatives of $\ln c$ with respect to x_1 and x_2:

$$\begin{aligned} s_1 &= \beta_1 + \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}z \\ s_2 &= \beta_2 + \gamma_{12}x_1 + \gamma_{22}x_2 + \gamma_{13}z \end{aligned}$$

Note that the share equations and the cost equation have parameters in common. One can do a pooled estimation of the three equations at once, imposing that the parameters are the same. In this way we're using more observations and therefore more information, which will lead to imporved efficiency. Note that this does assume that the cost equation is correctly specified (*i.e.,* not an approximation), since otherwise the derivatives would not be the true derivatives of the log cost function, and would then be misspecified for the shares. To pool the equations, write the model in matrix form (adding in error terms)

$$\begin{bmatrix} \ln c \\ s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2 & z & \frac{x_1^2}{2} & \frac{x_2^2}{2} & \frac{z^2}{2} & x_1x_2 & x_1z & x_2z \\ 0 & 1 & 0 & 0 & x_1 & 0 & 0 & x_2 & z & 0 \\ 0 & 0 & 1 & 0 & 0 & x_2 & 0 & x_1 & 0 & z \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \delta \\ \gamma_{11} \\ \gamma_{22} \\ \gamma_{33} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{23} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

This is *one* observation on the three equations. With the appropriate notation, a single observation can be written as

$$y_t = X_t\theta + \varepsilon_t$$

The overall model would stack n observations on the three equations for a total of $3n$ observations:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Next we need to consider the errors. For observation t the errors can be placed in a vector

$$\varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}$$

First consider the covariance matrix of this vector: the shares are certainly correlated since they must sum to one. (In fact, with 2 shares the variances are equal and the covariance is -1 times the variance. General notation is used to allow easy extension to the case of more than 2 inputs). Also, it's likely that the shares and the cost equation have different variances. Supposing that the model is covariance stationary, the variance of ε_t won$'$t depend upon t:

$$Var\varepsilon_t = \Sigma_0 = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \cdot & \sigma_{22} & \sigma_{23} \\ \cdot & \cdot & \sigma_{33} \end{bmatrix}$$

Note that this matrix is singular, since the shares sum to 1. Assuming that there is no autocorrelation, the overall covariance matrix has the *seemingly unrelated regressions* (SUR) structure.

$$
\begin{aligned}
Var \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} &= \Sigma \\
&= \begin{bmatrix} \Sigma_0 & 0 & \cdots & 0 \\ 0 & \Sigma_0 & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & \cdots & 0 & \Sigma_0 \end{bmatrix} \\
&= I_n \otimes \Sigma_0
\end{aligned}
$$

where the symbol $\otimes$ indicates the *Kronecker product*. The Kronecker product of two matrices A and B is

$$
A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & \ddots & & \vdots \\ \vdots & & & \\ a_{pq}B & \cdots & & a_{pq}B \end{bmatrix}.
$$

FGLS estimation of a translog model

So, this model has heteroscedasticity and autocorrelation, so OLS won't be efficient. The next question is: how do we estimate efficiently using FGLS? FGLS is based upon inverting the estimated error covariance $\hat{\Sigma}$. So we need to estimate Σ.

An asymptotically efficient procedure is (supposing normality of the errors)

1. Estimate each equation by OLS

2. Estimate Σ_0 using
$$\hat{\Sigma}_0 = \frac{1}{n}\sum_{t=1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_t'$$

3. Next we need to account for the singularity of Σ_0. It can be shown that $\hat{\Sigma}_0$ will be singular when the shares sum to one, so FGLS won't work. The solution is to drop one of the share equations, for example the second. The model becomes

$$\begin{bmatrix} \ln c \\ s_1 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2 & z & \frac{x_1^2}{2} & \frac{x_2^2}{2} & \frac{z^2}{2} & x_1x_2 & x_1z & x_2z \\ 0 & 1 & 0 & 0 & x_1 & 0 & 0 & x_2 & z & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \delta \\ \gamma_{11} \\ \gamma_{22} \\ \gamma_{33} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{23} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

or in matrix notation for the observation:

$$y_t^* = X_t^* \theta + \varepsilon_t^*$$

and in stacked notation for all observations we have the $2n$ observations:

$$\begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{bmatrix} = \begin{bmatrix} X_1^* \\ X_2^* \\ \vdots \\ X_n^* \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1^* \\ \varepsilon_2^* \\ \vdots \\ \varepsilon_n^* \end{bmatrix}$$

or, finally in matrix notation for all observations:

$$y^* = X^*\theta + \varepsilon^*$$

Considering the error covariance, we can define

$$\begin{aligned} \Sigma_0^* &= Var \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \\ \Sigma^* &= I_n \otimes \Sigma_0^* \end{aligned}$$

Define $\hat{\Sigma}_0^*$ as the leading 2×2 block of $\hat{\Sigma}_0$, and form

$$\hat{\Sigma}^* = I_n \otimes \hat{\Sigma}_0^*.$$

This is a consistent estimator, following the consistency of OLS and applying a LLN.

4. Next compute the Cholesky factorization

$$\hat{P}_0 = Chol\left(\hat{\Sigma}_0^*\right)^{-1}$$

(I am assuming this is defined as an upper triangular matrix, which is consistent with the way Octave does it) and the Cholesky factorization of the overall covariance matrix of the 2 equation model, which can be calculated as

$$\hat{P} = Chol\hat{\Sigma}^* = I_n \otimes \hat{P}_0$$

5. Finally the FGLS estimator can be calculated by applying OLS to the transformed model

$$\hat{P}'y^* = \hat{P}'X^*\theta + \hat{\hat{P}}'\varepsilon^*$$

or by directly using the GLS formula

$$\hat{\theta}_{FGLS} = \left(X^{*\prime}\left(\hat{\Sigma}_0^*\right)^{-1}X^*\right)^{-1}X^{*\prime}\left(\hat{\Sigma}_0^*\right)^{-1}y^*$$

It is equivalent to transform each observation individually:

$$\hat{P}_0'y_y^* = \hat{P}_0'X_t^*\theta + \hat{P}_0'\varepsilon^*$$

and then apply OLS. This is probably the simplest approach.

A few last comments.

1. We have assumed no autocorrelation across time. This is clearly restrictive. It is relatively simple to relax this, but we won't go into it here.

2. Also, we have only imposed symmetry of the second derivatives. Another restriction that the

model should satisfy is that the estimated shares should sum to 1. This can be accomplished by imposing

$$\begin{aligned} \beta_1 + \beta_2 &= 1 \\ \sum_{i=1}^{3} \gamma_{ij} &= 0, \ j = 1, 2, 3. \end{aligned}$$

These are linear parameter restrictions, so they are easy to impose and will improve efficiency if they are true.

3. The estimation procedure outlined above can be *iterated.* That is, estimate $\hat{\theta}_{FGLS}$ as above, then re-estimate Σ_0^* using errors calculated as

$$\hat{\varepsilon} = y - X\hat{\theta}_{FGLS}$$

These might be expected to lead to a better estimate than the estimator based on $\hat{\theta}_{OLS}$, since FGLS is asymptotically more efficient. Then re-estimate θ using the new estimated error covariance. It can be shown that if this is repeated until the estimates don't change (*i.e.,* iterated to convergence) then the resulting estimator is the MLE. At any rate, the asymptotic properties of the iterated and uniterated estimators are the same, since both are based upon a consistent estimator of the error covariance.

8.2 Testing nonnested hypotheses

Given that the choice of functional form isn't perfectly clear, in that many possibilities exist, how can one choose between forms? When one form is a parametric restriction of another, the previously studied tests such as Wald, LR, score or qF are all possibilities. For example, the Cobb-Douglas model is a parametric restriction of the translog: The translog is

$$y_t = \alpha + x_t'\beta + 1/2x_t'\Gamma x_t + \varepsilon$$

where the variables are in logarithms, while the Cobb-Douglas is

$$y_t = \alpha + x_t'\beta + \varepsilon$$

so a test of the Cobb-Douglas versus the translog is simply a test that $\Gamma = 0$.

The situation is more complicated when we want to test *non-nested hypotheses.* If the two functional forms are linear in the parameters, and use the same transformation of the dependent variable, then they may be written as

$$\begin{aligned} M_1 : y &= X\beta + \varepsilon \\ \varepsilon_t &\sim iid(0, \sigma_\varepsilon^2) \\ M_2 : y &= Z\gamma + \eta \\ \eta &\sim iid(0, \sigma_\eta^2) \end{aligned}$$

We wish to test hypotheses of the form: $H_0 : M_i$ *is correctly specified* versus $H_A : M_i$ *is misspecified*,

for $i = 1, 2$.

- One could account for non-iid errors, but we'll suppress this for simplicity.
- There are a number of ways to proceed. We'll consider the J test, proposed by Davidson and MacKinnon, *Econometrica* (1981). The idea is to artificially nest the two models, e.g.,

$$y = (1 - \alpha)X\beta + \alpha(Z\gamma) + \omega$$

If the first model is correctly specified, then the true value of α is zero. On the other hand, if the second model is correctly specified then $\alpha = 1$.

  - The problem is that this model is not identified in general. For example, if the models share some regressors, as in

$$\begin{aligned} M_1 : y_t &= \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t \\ M_2 : y_t &= \gamma_1 + \gamma_2 x_{2t} + \gamma_3 x_{4t} + \eta_t \end{aligned}$$

then the composite model is

$$y_t = (1 - \alpha)\beta_1 + (1 - \alpha)\beta_2 x_{2t} + (1 - \alpha)\beta_3 x_{3t} + \alpha\gamma_1 + \alpha\gamma_2 x_{2t} + \alpha\gamma_3 x_{4t} + \omega_t$$

Combining terms we get

$$\begin{aligned} y_t &= ((1 - \alpha)\beta_1 + \alpha\gamma_1) + ((1 - \alpha)\beta_2 + \alpha\gamma_2)\, x_{2t} + (1 - \alpha)\beta_3 x_{3t} + \alpha\gamma_3 x_{4t} + \omega_t \\ &= \delta_1 + \delta_2 x_{2t} + \delta_3 x_{3t} + \delta_4 x_{4t} + \omega_t \end{aligned}$$

The four $\delta's$ are consistently estimable, but α is not, since we have four equations in 7 unknowns, so one can't test the hypothesis that $\alpha = 0$.

The idea of the J test is to substitute $\hat{\gamma}$ in place of γ. This is a consistent estimator supposing that the second model is correctly specified. It will tend to a finite probability limit even if the second model is misspecified. Then estimate the model

$$\begin{aligned} y &= (1-\alpha)X\beta + \alpha(Z\hat{\gamma}) + \omega \\ &= X\theta + \alpha\hat{y} + \omega \end{aligned}$$

where $\hat{y} = Z(Z'Z)^{-1}Z'y = P_Z y$. In this model, α is consistently estimable, and one can show that, under the hypothesis that the first model is correct, $\alpha \xrightarrow{p} 0$ and that the ordinary t -statistic for $\alpha = 0$ is asymptotically normal:

$$t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}} \overset{a}{\sim} N(0,1)$$

- If the second model is correctly specified, then $t \xrightarrow{p} \infty$, since $\hat{\alpha}$ tends in probability to 1, while it's estimated standard error tends to zero. Thus the test will always reject the false null model, asymptotically, since the statistic will eventually exceed any critical value with probability one.
- We can reverse the roles of the models, testing the second against the first.
- It may be the case that *neither* model is correctly specified. In this case, the test will still reject the null hypothesis, asymptotically, if we use critical values from the $N(0,1)$ distribution, since as long as $\hat{\alpha}$ tends to something different from zero, $|t| \xrightarrow{p} \infty$. Of course, when we switch the roles of the models the other will also be rejected asymptotically.

- In summary, there are 4 possible outcomes when we test two models, each against the other. Both may be rejected, neither may be rejected, or one of the two may be rejected.

- There are other tests available for non-nested models. The $J-$ test is simple to apply when both models are linear in the parameters. The P-test is similar, but easier to apply when M_1 is nonlinear.

- The above presentation assumes that the same transformation of the dependent variable is used by both models. MacKinnon, White and Davidson, *Journal of Econometrics*, (1983) shows how to deal with the case of different transformations.

- Monte-Carlo evidence shows that these tests often over-reject a correctly specified model. Can use bootstrap critical values to get better-performing tests.

Chapter 9

Generalized least squares

Recall the assumptions of the classical linear regression model, in Section 3.6. One of the assumptions we've made up to now is that

$$\varepsilon_t \sim IID(0, \sigma^2)$$

or occasionally

$$\varepsilon_t \sim IIN(0, \sigma^2).$$

Now we'll investigate the consequences of nonidentically and/or dependently distributed errors. We'll assume fixed regressors for now, to keep the presentation simple, and later we'll look at the consequences of relaxing this admittedly unrealistic assumption. The model is

$$\begin{aligned} y &= X\beta + \varepsilon \\ \mathcal{E}(\varepsilon) &= 0 \\ V(\varepsilon) &= \Sigma \end{aligned}$$

where Σ is a general symmetric positive definite matrix (we'll write β in place of β_0 to simplify the typing of these notes).

- The case where Σ is a diagonal matrix gives uncorrelated, nonidentically distributed errors. This is known as *heteroscedasticity*: $\exists i, j \; s.t. \; V(\epsilon_i) \neq V(\epsilon_j)$
- The case where Σ has the same number on the main diagonal but nonzero elements off the main diagonal gives identically (assuming higher moments are also the same) dependently distributed errors. This is known as *autocorrelation*: $\exists i \neq j \; s.t. \; E(\epsilon_i \epsilon_j) \neq 0)$
- The general case combines heteroscedasticity and autocorrelation. This is known as "nonspherical" disturbances, though why this term is used, I have no idea. Perhaps it's because under the classical assumptions, a joint confidence region for ε would be an $n-$ dimensional hypersphere.

9.1 Effects of nonspherical disturbances on the OLS estimator

The least square estimator is

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ &= \beta + (X'X)^{-1}X'\varepsilon \end{aligned}$$

- We have unbiasedness, as before.

- The variance of $\hat{\beta}$ is

$$\begin{aligned} \mathcal{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] &= \mathcal{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'\Sigma X(X'X)^{-1} \end{aligned} \tag{9.1}$$

  Due to this, any test statistic that is based upon an estimator of σ^2 is invalid, since there *isn't* any σ^2, it doesn't exist as a feature of the true d.g.p. In particular, the formulas for the t, F, χ^2 based tests given above do not lead to statistics with these distributions.

- $\hat{\beta}$ is still consistent, following exactly the same argument given before.

- If ε is normally distributed, then

$$\hat{\beta} \sim N\left(\beta, (X'X)^{-1}X'\Sigma X(X'X)^{-1}\right)$$

  The problem is that Σ is unknown in general, so this distribution won't be useful for testing hypotheses.

- Without normality, and with stochastic X (e.g., weakly exogenous regressors) we still have

$$\begin{aligned} \sqrt{n}\left(\hat{\beta}-\beta\right) &= \sqrt{n}(X'X)^{-1}X'\varepsilon \\ &= \left(\frac{X'X}{n}\right)^{-1} n^{-1/2}X'\varepsilon \end{aligned}$$

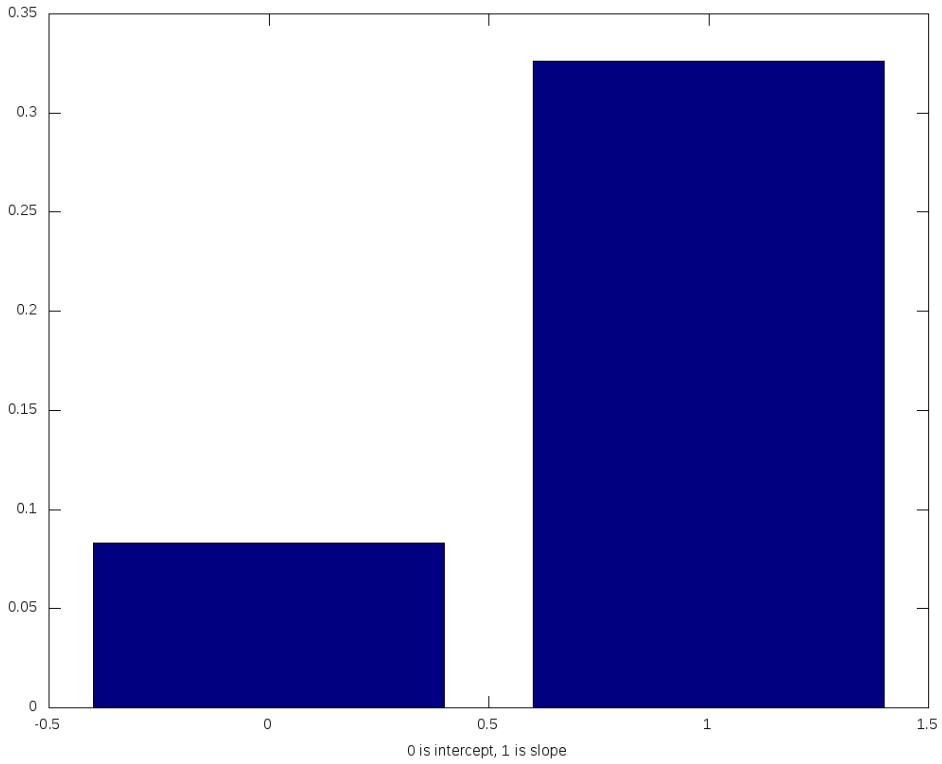
Define the limiting variance of $n^{-1/2}X'\varepsilon$ (supposing a CLT applies) as

$$\lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon'X}{n}\right) = \Omega, \text{ a.s.}$$

so we obtain $\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right)$. Note that the true asymptotic distribution of the OLS has changed with respect to the results under the classical assumptions. If we neglect to take this into account, the Wald and score tests will not be asymptotically valid. So we need to figure out *how* to take it into account.

To see the invalidity of test procedures that are correct under the classical assumptions, when we have nonspherical errors, consider the Octave script GLS/EffectsOLS.m. This script does a Monte Carlo study, generating data that are either heteroscedastic or homoscedastic, and then computes the empirical rejection frequency of a nominally 10% t-test. When the data are heteroscedastic, we obtain something like what we see in Figure 9.1. This sort of heteroscedasticity causes us to reject a true null hypothesis regarding the slope parameter much too often. You can experiment with the script to look at the effects of other sorts of HET, and to vary the sample size.

Figure 9.1: Rejection frequency of 10% t-test, H0 is true.



Summary: OLS with heteroscedasticity and/or autocorrelation is:

- unbiased with fixed or strongly exogenous regressors
- biased with weakly exogenous regressors
- has a different variance than before, so the previous test statistics aren't valid
- is consistent
- is asymptotically normally distributed, but with a different limiting covariance matrix. Previous test statistics aren't valid in this case for this reason.
- is inefficient, as is shown below.

9.2 The GLS estimator

Suppose Σ were known. Then one could form the Cholesky decomposition

$$P'P = \Sigma^{-1}$$

Here, P is an upper triangular matrix. We have

$$P'P\Sigma = I_n$$

so

$$P'P\Sigma P' = P',$$

which implies that

$$P\Sigma P' = I_n$$

Let's take some time to play with the Cholesky decomposition. Try out the GLS/cholesky.m Octave script to see that the above claims are true, and also to see how one can generate data from a $N(0, V)$ distribition.

Consider the model

$$Py = PX\beta + P\varepsilon,$$

or, making the obvious definitions,

$$y^* = X^*\beta + \varepsilon^*.$$

This variance of $\varepsilon^* = P\varepsilon$ is

$$\begin{aligned} \mathcal{E}(P\varepsilon\varepsilon' P') &= P\Sigma P' \\ &= I_n \end{aligned}$$

Therefore, the model

$$\begin{aligned} y^* &= X^*\beta + \varepsilon^* \\ \mathcal{E}(\varepsilon^*) &= 0 \\ V(\varepsilon^*) &= I_n \end{aligned}$$

satisfies the classical assumptions. The GLS estimator is simply OLS applied to the transformed

model:

$$\begin{aligned}\hat{\beta}_{GLS} &= (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*} \\ &= (X'P'PX)^{-1}X'P'Py \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y\end{aligned}$$

The GLS estimator is unbiased in the same circumstances under which the OLS estimator is unbiased. For example, assuming X is nonstochastic

$$\begin{aligned}\mathcal{E}(\hat{\beta}_{GLS}) &= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y\right\} \\ &= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(X\beta+\varepsilon\right\} \\ &= \beta.\end{aligned}$$

To get the variance of the estimator, we have

$$\begin{aligned}\hat{\beta}_{GLS} &= (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*} \\ &= (X^{*\prime}X^{*})^{-1}X^{*\prime}\left(X^{*}\beta+\varepsilon^{*}\right) \\ &= \beta+(X^{*\prime}X^{*})^{-1}X^{*\prime}\varepsilon^{*}\end{aligned}$$

so

$$\begin{aligned}\mathcal{E}\left\{\left(\hat{\beta}_{GLS}-\beta\right)\left(\hat{\beta}_{GLS}-\beta\right)'\right\} &= \mathcal{E}\left\{(X^{*\prime}X^{*})^{-1}X^{*\prime}\varepsilon^{*}\varepsilon^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1}\right\} \\ &= (X^{*\prime}X^{*})^{-1}X^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1} \\ &= (X^{*\prime}X^{*})^{-1} \\ &= (X'\Sigma^{-1}X)^{-1}\end{aligned}$$

Either of these last formulas can be used.

- All the previous results regarding the desirable properties of the least squares estimator hold, when dealing with the transformed model, since the transformed model satisfies the classical assumptions..
- Tests are valid, using the previous formulas, as long as we substitute X^{*} in place of X. Furthermore, any test that involves σ^{2} can set it to 1. This is preferable to re-deriving the appropriate formulas.
- The GLS estimator is more efficient than the OLS estimator. This is a consequence of the Gauss-Markov theorem, since the GLS estimator is based on a model that satisfies the classical assumptions but the OLS estimator is not. To see this directly, note that

$$\begin{aligned}Var(\hat{\beta})-Var(\hat{\beta}_{GLS}) &= (X'X)^{-1}X'\Sigma X(X'X)^{-1}-(X'\Sigma^{-1}X)^{-1} \\ &= A\Sigma A^{'}\end{aligned}$$

where $A=\left[(X'X)^{-1}X'-(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\right]$. This may not seem obvious, but it is true, as you

can verify for yourself. Then noting that $A\Sigma A'$ is a quadratic form in a positive definite matrix, we conclude that $A\Sigma A'$ is positive semi-definite, and that GLS is efficient relative to OLS.

- As one can verify by calculating first order conditions, the GLS estimator is the solution to the minimization problem

$$\hat{\beta}_{GLS} = \arg\min(y - X\beta)'\Sigma^{-1}(y - X\beta)$$

so the *metric* Σ^{-1} is used to weight the residuals.

9.3 Feasible GLS

The problem is that Σ ordinarily isn't known, so this estimator isn't available.

- Consider the dimension of Σ : it's an $n \times n$ matrix with $(n^2 - n)/2 + n = (n^2 + n)/2$ unique elements (remember - it is symmetric, because it's a covariance matrix).
- The number of parameters to estimate is larger than n and increases faster than n. There's no way to devise an estimator that satisfies a LLN without adding restrictions.
- The *feasible GLS estimator* is based upon making sufficient assumptions regarding the form of Σ so that a consistent estimator can be devised.

Suppose that we *parameterize* Σ as a function of X and θ, where θ may include β as well as other parameters, so that

$$\Sigma = \Sigma(X, \theta)$$

where θ is of fixed dimension. If we can consistently estimate θ, we can consistently estimate Σ, as long as the elements of $\Sigma(X, \theta)$ are continuous functions of θ (by the Slutsky theorem). In this case,

$$\widehat{\Sigma} = \Sigma(X, \hat{\theta}) \xrightarrow{p} \Sigma(X, \theta)$$

If we replace Σ in the formulas for the GLS estimator with $\widehat{\Sigma}$, we obtain the FGLS estimator. **The FGLS estimator shares the same asymptotic properties as GLS. These are**

1. Consistency
2. Asymptotic normality
3. Asymptotic efficiency *if* the errors are normally distributed. (Cramer-Rao).
4. Test procedures are asymptotically valid.

In practice, the usual way to proceed is

1. Define a consistent estimator of θ. This is a case-by-case proposition, depending on the parameterization $\Sigma(\theta)$. We'll see examples below.
2. Form $\widehat{\Sigma} = \Sigma(X, \hat{\theta})$
3. Calculate the Cholesky factorization $\widehat{P} = Chol(\hat{\Sigma}^{-1})$.
4. Transform the model using
$$\hat{P}y = \hat{P}X\beta + \hat{P}\varepsilon$$
5. Estimate using OLS on the transformed model.

9.4 Heteroscedasticity

Heteroscedasticity is the case where

$$\mathcal{E}(\varepsilon\varepsilon') = \Sigma$$

is a diagonal matrix, so that the errors are uncorrelated, but have different variances. Heteroscedasticity is usually thought of as associated with cross sectional data, though there is absolutely no reason why time series data cannot also be heteroscedastic. Actually, the popular ARCH (autoregressive conditionally heteroscedastic) models explicitly assume that a time series is heteroscedastic.

Consider a supply function

$$q_i = \beta_1 + \beta_p P_i + \beta_s S_i + \varepsilon_i$$

where P_i is price and S_i is some measure of size of the i^{th} firm. One might suppose that unobservable factors (e.g., talent of managers, degree of coordination between production units, *etc.*) account for the error term ε_i. If there is more variability in these factors for large firms than for small firms, then ε_i may have a higher variance when S_i is high than when it is low.

Another example, individual demand.

$$q_i = \beta_1 + \beta_p P_i + \beta_m M_i + \varepsilon_i$$

where P is price and M is income. In this case, ε_i can reflect variations in preferences. There are more possibilities for expression of preferences when one is rich, so it is possible that the variance of ε_i could be higher when M is high.

Add example of group means.

OLS with heteroscedastic consistent varcov estimation

Eicker (1967) and White (1980) showed how to modify test statistics to account for heteroscedasticity of unknown form. The OLS estimator has asymptotic distribution

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right)$$

as we've already seen. Recall that we defined

$$\lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon'X}{n}\right) = \Omega$$

This matrix has dimension $K \times K$ and can be consistently estimated, even if we can't estimate Σ consistently. The consistent estimator, under heteroscedasticity but no autocorrelation is

$$\widehat{\Omega} = \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \hat{\varepsilon}_t^2$$

One can then modify the previous test statistics to obtain tests that are valid when there is heteroscedasticity of unknown form. For example, the Wald test for $H_0 : R\beta - r = 0$ would be

$$n\left(R\hat{\beta}-r\right)'\left(R\left(\frac{X'X}{n}\right)^{-1}\hat{\Omega}\left(\frac{X'X}{n}\right)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right) \overset{a}{\sim} \chi^2(q)$$

To see the effects of ignoring HET when doing OLS, and the good effect of using a HET consistent covariance estimator, consider the script bootstrap_example1.m. This script generates data from a linear model with HET, then computes standard errors using the ordinary OLS formula, the Eicker-

White formula, and also bootstrap standard errors. Note that Eicker-White and bootstrap pretty much agree, while the OLS formula gives standard errors that are quite different. Typical output of this script follows:

```
octave:1> bootstrap_example1
Bootstrap standard errors
   0.083376   0.090719   0.143284
*********************************************************
OLS estimation results
Observations 100
R-squared 0.014674
Sigma-squared 0.695267
Results (Ordinary var-cov estimator)
     estimate     st.err.     t-stat.     p-value
1      -0.115       0.084      -1.369       0.174
2      -0.016       0.083      -0.197       0.845
3      -0.105       0.088      -1.189       0.237
*********************************************************
OLS estimation results
Observations 100
R-squared 0.014674
Sigma-squared 0.695267
Results (Het. consistent var-cov estimator)
     estimate     st.err.     t-stat.     p-value
1      -0.115       0.084      -1.381       0.170
2      -0.016       0.090      -0.182       0.856
3      -0.105       0.140      -0.751       0.454
```

- If you run this several times, you will notice that the OLS standard error for the last parameter appears to be biased downward, at least comparing to the other two methods, which are asymptotically valid.
- The true coefficients *are* zero. With a standard error biased downward, the t-test for lack of significance will reject more often than it should (the variables really are not significant, but we will find that they seem to be more often than is due to Type-I error.
- For example, you should see that the p-value for the last coefficient is smaller than 0.10 more than 10% of the time. Run the script 20 times and you'll see.

Detection

There exist many tests for the presence of heteroscedasticity. We'll discuss three methods.

Goldfeld-Quandt The sample is divided in to three parts, with n_1, n_2 and n_3 observations, where $n_1 + n_2 + n_3 = n$. The model is estimated using the first and third parts of the sample, separately, so that $\hat{\beta}^1$ and $\hat{\beta}^3$ will be independent. Then we have

$$\frac{\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^1}{\sigma^2} = \frac{\varepsilon^{1\prime} M^1 \varepsilon^1}{\sigma^2} \xrightarrow{d} \chi^2(n_1 - K)$$

and

$$\frac{\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^3}{\sigma^2} = \frac{\varepsilon^{3\prime} M^3 \varepsilon^3}{\sigma^2} \xrightarrow{d} \chi^2(n_3 - K)$$

so

$$\frac{\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^{1}/(n_1 - K)}{\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^{3}/(n_3 - K)} \xrightarrow{d} F(n_1 - K, n_3 - K).$$

The distributional result is exact if the errors are normally distributed. This test is a two-tailed test. Alternatively, and probably more conventionally, if one has prior ideas about the possible magnitudes of the variances of the observations, one could order the observations accordingly, from largest to smallest. In this case, one would use a conventional one-tailed F-test. *Draw picture.*

- Ordering the observations is an important step if the test is to have any power.
- The motive for dropping the middle observations is to increase the difference between the average variance in the subsamples, supposing that there exists heteroscedasticity. This can increase the power of the test. On the other hand, dropping too many observations will substantially increase the variance of the statistics $\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^{1}$ and $\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^{3}$. A rule of thumb, based on Monte Carlo experiments is to drop around 25% of the observations.
- If one doesn't have any ideas about the form of the het. the test will probably have low power since a sensible data ordering isn't available.

White's test When one has little idea if there exists heteroscedasticity, and no idea of its potential form, the White test is a possibility. The idea is that if there is homoscedasticity, then

$$\mathcal{E}(\varepsilon_t^2 | x_t) = \sigma^2, \forall t$$

so that x_t or functions of x_t shouldn't help to explain $\mathcal{E}(\varepsilon_t^2)$. The test works as follows:

1. Since ε_t isn't available, use the consistent estimator $\hat{\varepsilon}_t$ instead.

2. Regress

$$\hat{\varepsilon}_t^2 = \sigma^2 + z_t'\gamma + v_t$$

where z_t is a P-vector. z_t may include some or all of the variables in x_t, as well as other variables. White's original suggestion was to use x_t, plus the set of all unique squares and cross products of variables in x_t.

3. Test the hypothesis that $\gamma = 0$. The qF statistic in this case is

$$qF = \frac{P\left(ESS_R - ESS_U\right)/P}{ESS_U/\left(n - P - 1\right)}$$

Note that $ESS_R = TSS_U$, so dividing both numerator and denominator by this we get

$$qF = \left(n - P - 1\right)\frac{R^2}{1 - R^2}$$

Note that this is the R^2 of the artificial regression used to test for heteroscedasticity, not the R^2 of the original model.

An asymptotically equivalent statistic, under the null of no heteroscedasticity (so that R^2 should tend to zero), is

$$nR^2 \overset{a}{\sim} \chi^2(P).$$

This doesn't require normality of the errors, though it does assume that the fourth moment of ε_t is constant, under the null. **Question**: why is this necessary?

- The White test has the disadvantage that it may not be very powerful unless the z_t vector is

chosen well, and this is hard to do without knowledge of the form of heteroscedasticity.

- It also has the problem that specification errors other than heteroscedasticity may lead to rejection.

- Note: the null hypothesis of this test may be interpreted as $\theta = 0$ for the variance model $V(\varepsilon_t^2) = h(\alpha + z_t'\theta)$, where $h(\cdot)$ is an arbitrary function of unknown form. The test is more general than is may appear from the regression that is used.

Plotting the residuals A very simple method is to simply plot the residuals (or their squares). *Draw pictures here.* Like the Goldfeld-Quandt test, this will be more informative if the observations are ordered according to the suspected form of the heteroscedasticity.

Correction

Correcting for heteroscedasticity requires that a parametric form for $\Sigma(\theta)$ be supplied, and that a means for estimating θ consistently be determined. The estimation method will be specific to the for supplied for $\Sigma(\theta)$. We'll consider two examples. Before this, let's consider the general nature of GLS when there is heteroscedasticity.

When we have HET but no AUT, Σ is a diagonal matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ \vdots & \sigma_2^2 & & \vdots \\ & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}$$

Likewise, Σ^{-1} is diagonal

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ \vdots & \frac{1}{\sigma_2^2} & & \vdots \\ & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_n^2} \end{bmatrix}$$

and so is the Cholesky decomposition $P = chol(\Sigma^{-1})$

$$P = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ \vdots & \frac{1}{\sigma_2} & & \vdots \\ & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_n} \end{bmatrix}$$

We need to transform the model, just as before, in the general case:

$$Py = PX\beta + P\varepsilon,$$

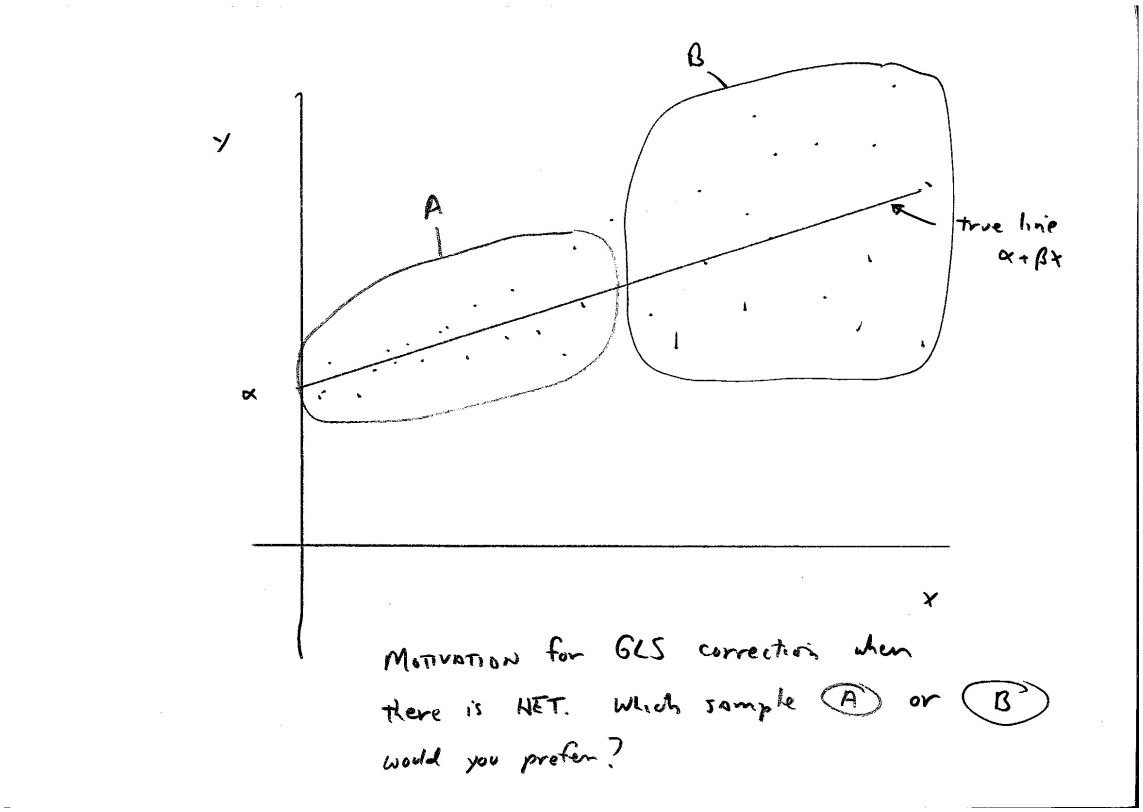
or, making the obvious definitions,

$$y^* = X^*\beta + \varepsilon^*.$$

Note that multiplying by P just divides the data for each observation (y_i, x_i) by the corresponding standard error of the error term, σ_i. That is, $y_i^* = y_i/\sigma_i$ and $x_i^* = x_i/\sigma_i$ (note that x_i is a K-vector: we divided each element, including the 1 corresponding to the constant).

This makes sense. Consider Figure 9.2, which shows a true regression line with heteroscedastic errors. Which sample is more informative about the location of the line? The ones with observations

Figure 9.2: Motivation for GLS correction when there is HET



with smaller variances. So, the GLS solution is equivalent to OLS on the transformed data. By the transformed data is the original data, weighted by the inverse of the standard error of the observation's error term. When the standard error is small, the weight is high, and vice versa. The GLS correction for the case of HET is also known as weighted least squares, for this reason.

Multiplicative heteroscedasticity

Suppose the model is

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \sigma_t^2 &= \mathcal{E}(\varepsilon_t^2) = (z_t'\gamma)^\delta \end{aligned}$$

but the other classical assumptions hold. In this case

$$\varepsilon_t^2 = (z_t'\gamma)^\delta + v_t$$

and v_t has mean zero. Nonlinear least squares could be used to estimate γ and δ consistently, were ε_t observable. The solution is to substitute the squared OLS residuals $\hat{\varepsilon}_t^2$ in place of ε_t^2, since it is consistent by the Slutsky theorem. Once we have $\hat{\gamma}$ and $\hat{\delta}$, we can estimate σ_t^2 consistently using

$$\hat{\sigma}_t^2 = (z_t'\hat{\gamma})^{\hat{\delta}} \stackrel{p}{\rightarrow} \sigma_t^2 \ .$$

In the second step, we transform the model by dividing by the standard deviation:

$$\frac{y_t}{\hat{\sigma}_t} = \frac{x_t'\beta}{\hat{\sigma}_t} + \frac{\varepsilon_t}{\hat{\sigma}_t}$$

or

$$y_t^* = x_t^{*\prime}\beta + \varepsilon_t^*.$$

Asymptotically, this model satisfies the classical assumptions.

- This model is a bit complex in that NLS is required to estimate the model of the variance. A simpler version would be

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \sigma_t^2 &= \mathcal{E}(\varepsilon_t^2) = \sigma^2 z_t^\delta \end{aligned}$$

  where z_t is a single variable. There are still two parameters to be estimated, and the model of the variance is still nonlinear in the parameters. However, the *search method* can be used in this case to reduce the estimation problem to repeated applications of OLS.

- First, we define an interval of reasonable values for δ, e.g., $\delta \in [0, 3]$.

- Partition this interval into M equally spaced values, e.g., $\{0, .1, .2, ..., 2.9, 3\}$.

- For each of these values, calculate the variable $z_t^{\delta_m}$.

- The regression

$$\hat{\varepsilon}_t^2 = \sigma^2 z_t^{\delta_m} + v_t$$

  is linear in the parameters, conditional on δ_m, so one can estimate σ^2 by OLS.

- Save the pairs (σ_m^2, δ_m), and the corresponding ESS_m. Choose the pair with the minimum ESS_m as the estimate.

- Next, divide the model by the estimated standard deviations.

- Can refine. *Draw picture.*

- Works well when the parameter to be searched over is low dimensional, as in this case.

Groupwise heteroscedasticity

A common case is where we have repeated observations on each of a number of economic agents: e.g., 10 years of macroeconomic data on each of a set of countries or regions, or daily observations of transactions of 200 banks. This sort of data is a *pooled cross-section time-series model.* It may be reasonable to presume that the variance is constant over time within the cross-sectional units, but that it differs across them (e.g., firms or countries of different sizes...). The model is

$$\begin{aligned} y_{it} &= x'_{it}\beta + \varepsilon_{it} \\ \mathcal{E}(\varepsilon_{it}^2) &= \sigma_i^2, \forall t \end{aligned}$$

where $i = 1, 2, ..., G$ are the agents, and $t = 1, 2, ..., n$ are the observations on each agent.

- The other classical assumptions are presumed to hold.
- In this case, the variance σ_i^2 is specific to each agent, but constant over the n observations for that agent.
- In this model, we assume that $\mathcal{E}(\varepsilon_{it}\varepsilon_{is}) = 0$. This is a strong assumption that we'll relax later.

To correct for heteroscedasticity, just estimate each σ_i^2 using the natural estimator:

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{t=1}^{n} \hat{\varepsilon}_{it}^2$$

- Note that we use $1/n$ here since it's possible that there are more than n regressors, so $n - K$ could be negative. Asymptotically the difference is unimportant.

- With each of these, transform the model as usual:

$$\frac{y_{it}}{\hat{\sigma}_i} = \frac{x'_{it}\beta}{\hat{\sigma}_i} + \frac{\varepsilon_{it}}{\hat{\sigma}_i}$$

  Do this for each cross-sectional group. This transformed model satisfies the classical assumptions, asymptotically.

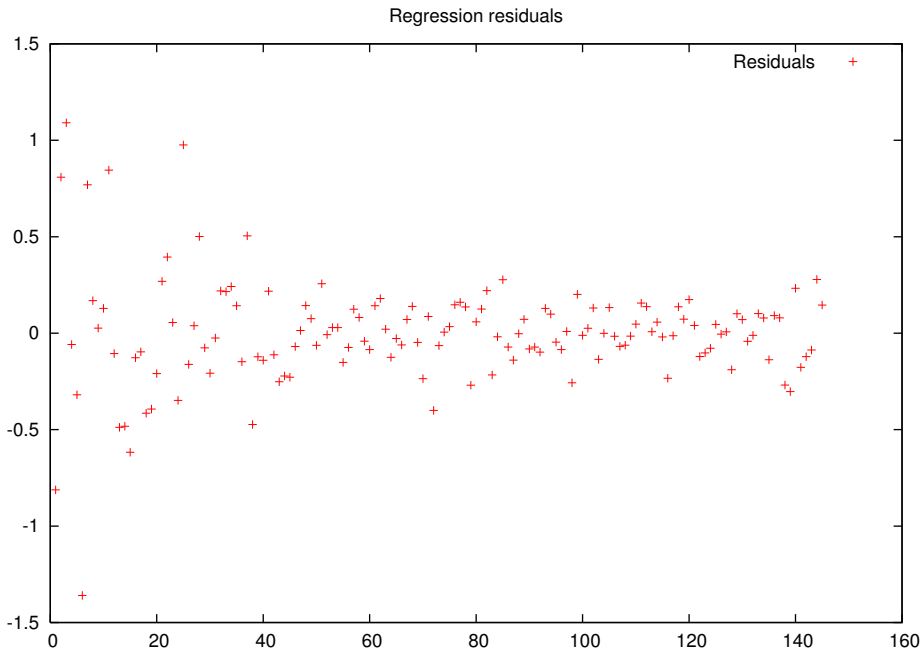
Example: the Nerlove model (again!)

Remember the Nerlove data - see sections 3.8 and 5.8. Let's check the Nerlove data for evidence of heteroscedasticity. In what follows, we're going to use the model with the constant and output coefficient varying across 5 groups, but with the input price coefficients fixed (see Equation 5.5 for the rationale behind this). Figure 9.3, which is generated by the Octave program GLS/NerloveResiduals.m plots the residuals. We can see pretty clearly that the error variance is larger for small firms than for larger firms.

Now let's try out some tests to formally check for heteroscedasticity. The Octave program GLS/HetTests.m performs the White and Goldfeld-Quandt tests, using the above model. The results are

```
                  Value     p-value
White's test     61.903       0.000
               Value     p-value
GQ test       10.886       0.000
```

Figure 9.3: Residuals, Nerlove model, sorted by firm size



All in all, it is very clear that the data are heteroscedastic. That means that OLS estimation is not efficient, and tests of restrictions that ignore heteroscedasticity are not valid. The previous tests (CRTS, HOD1 and the Chow test) were calculated assuming homoscedasticity. The Octave program GLS/NerloveRestrictions-Het.m uses the Wald test to check for CRTS and HOD1, but using a heteroscedastic-consistent covariance estimator.[1] The results are

```
Testing HOD1
                Value     p-value
Wald test       6.161       0.013

Testing CRTS
                Value     p-value
Wald test      20.169       0.001
```

We see that the previous conclusions are altered - both CRTS is and HOD1 are rejected at the 5% level. Maybe the rejection of HOD1 is due to to Wald test's tendency to over-reject?

From the previous plot, it seems that the variance of ϵ is a decreasing function of output. Suppose that the 5 size groups have different error variances (heteroscedasticity by groups):

$$Var(\epsilon_i) = \sigma_j^2,$$

[1] By the way, notice that GLS/NerloveResiduals.m and GLS/HetTests.m use the restricted LS estimator directly to restrict the fully general model with all coefficients varying to the model with only the constant and the output coefficient varying. But GLS/NerloveRestrictions-Het.m estimates the model by substituting the restrictions into the model. The methods are equivalent, but the second is more convenient and easier to understand.

where $j = 1$ if $i = 1, 2, ..., 29$, *etc.*, as before. The Octave script GLS/NerloveGLS.m estimates the model using GLS (through a transformation of the model so that OLS can be applied). The estimation results are i

```
**********************************************************
OLS estimation results
Observations 145
R-squared 0.958822
Sigma-squared 0.090800

Results (Het. consistent var-cov estimator)
```

	estimate	st.err.	t-stat.	p-value
constant1	-1.046	1.276	-0.820	0.414
constant2	-1.977	1.364	-1.450	0.149
constant3	-3.616	1.656	-2.184	0.031
constant4	-4.052	1.462	-2.771	0.006
constant5	-5.308	1.586	-3.346	0.001
output1	0.391	0.090	4.363	0.000
output2	0.649	0.090	7.184	0.000
output3	0.897	0.134	6.688	0.000
output4	0.962	0.112	8.612	0.000
output5	1.101	0.090	12.237	0.000
labor	0.007	0.208	0.032	0.975

fuel	0.498	0.081	6.149	0.000
capital	-0.460	0.253	-1.818	0.071

****

****
OLS estimation results
Observations 145
R-squared 0.987429
Sigma-squared 1.092393

Results (Het. consistent var-cov estimator)

	estimate	st.err.	t-stat.	p-value
constant1	-1.580	0.917	-1.723	0.087
constant2	-2.497	0.988	-2.528	0.013
constant3	-4.108	1.327	-3.097	0.002
constant4	-4.494	1.180	-3.808	0.000
constant5	-5.765	1.274	-4.525	0.000
output1	0.392	0.090	4.346	0.000
output2	0.648	0.094	6.917	0.000
output3	0.892	0.138	6.474	0.000
output4	0.951	0.109	8.755	0.000

```
output5          1.093       0.086      12.684       0.000
labor            0.103       0.141       0.733       0.465
fuel             0.492       0.044      11.294       0.000
capital         -0.366       0.165      -2.217       0.028

*******************************************************

Testing HOD1
                 Value     p-value
Wald test        9.312       0.002
```

The first panel of output are the OLS estimation results, which are used to consistently estimate the σ_j^2. The second panel of results are the GLS estimation results. Some comments:

- The R^2 measures are not comparable - the dependent variables are not the same. The measure for the GLS results uses the transformed dependent variable. One could calculate a comparable R^2 measure, but I have not done so.
- The differences in estimated standard errors (smaller in general for GLS) *can* be interpreted as evidence of improved efficiency of GLS, since the OLS standard errors are calculated using the Huber-White estimator. They would not be comparable if the ordinary (inconsistent) estimator had been used.
- Note that the previously noted pattern in the output coefficients persists. The nonconstant CRTS result is robust.

- The coefficient on capital is now negative and significant at the 3% level. That seems to indicate some kind of problem with the model or the data, or economic theory.
- Note that HOD1 is now rejected. Problem of Wald test over-rejecting? Specification error in model?

9.5 Autocorrelation

Autocorrelation, which is the serial correlation of the error term, is a problem that is usually associated with time series data, but also can affect cross-sectional data. For example, a shock to oil prices will simultaneously affect all countries, so one could expect contemporaneous correlation of macroeconomic variables across countries.

Example

Consider the Keeling-Whorf data on atmospheric CO2 concentrations an Mauna Loa, Hawaii (see http://en.wikipedia.org/wiki/Keeling_Curve and http://cdiac.ornl.gov/ftp/ndp001/maunaloa.txt).

From the file maunaloa.txt: "THE DATA FILE PRESENTED IN THIS SUBDIRECTORY CONTAINS MONTHLY AND ANNUAL ATMOSPHERIC CO2 CONCENTRATIONS DERIVED FROM THE SCRIPPS INSTITUTION OF OCEANOGRAPHY'S (SIO's) CONTINUOUS MONITORING PROGRAM AT MAUNA LOA OBSERVATORY, HAWAII. THIS RECORD CONSTITUTES THE LONGEST CONTINUOUS RECORD OF ATMOSPHERIC CO2 CONCENTRATIONS AVAILABLE IN THE WORLD. MONTHLY AND ANNUAL AVERAGE MOLE FRACTIONS OF CO2 IN WATER-

VAPOR-FREE AIR ARE GIVEN FROM MARCH 1958 THROUGH DECEMBER 2003, EXCEPT FOR A FEW INTERRUPTIONS."

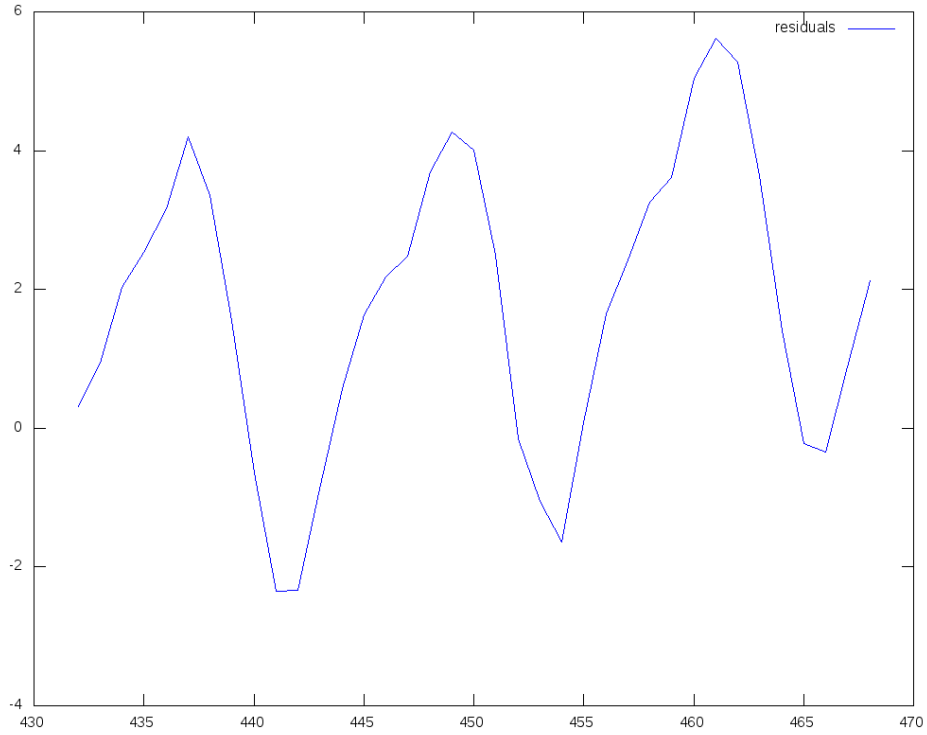
The data is available in Octave format at CO2.data .

If we fit the model $CO2_t = \beta_1 + \beta_2 t + \epsilon_t$, we get the results

```
octave:8> CO2Example
warning: load: file found in load path

*********************************************************
OLS estimation results
Observations 468
R-squared 0.979239
Sigma-squared 5.696791

Results (Het. consistent var-cov estimator)

     estimate      st.err.      t-stat.      p-value
1     316.918        0.227     1394.406        0.000
2       0.121        0.001      141.521        0.000

*********************************************************
```

It seems pretty clear that CO2 concentrations have been going up in the last 50 years, surprise, surprise. Let's look at a residual plot for the last 3 years of the data, see Figure 9.4. Note that there is a very predictable pattern. This is pretty strong evidence that the errors of the model are not independent of one another, which means there seems to be autocorrelation.

Figure 9.4: Residuals from time trend for CO2 data



Causes

Autocorrelation is the existence of correlation across the error term:

$$\mathcal{E}(\varepsilon_t \varepsilon_s) \neq 0, t \neq s.$$

Why might this occur? Plausible explanations include

1. Lags in adjustment to shocks. In a model such as

   $$y_t = x_t' \beta + \varepsilon_t,$$

   one could interpret $x_t' \beta$ as the equilibrium value. Suppose x_t is constant over a number of observations. One can interpret ε_t as a shock that moves the system away from equilibrium. If the time needed to return to equilibrium is long with respect to the observation frequency, one could expect ε_{t+1} to be positive, conditional on ε_t positive, which induces a correlation.

2. Unobserved factors that are correlated over time. The error term is often assumed to correspond to unobservable factors. If these factors are correlated, there will be autocorrelation.

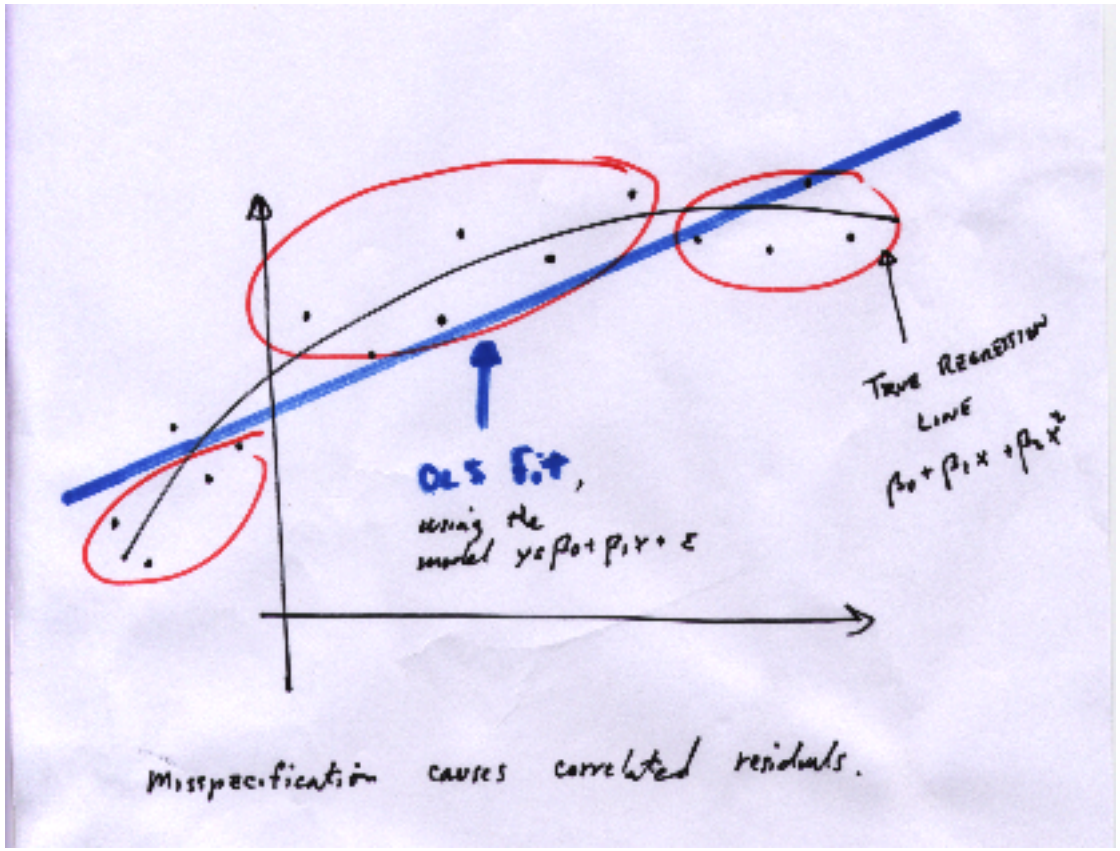
3. Misspecification of the model. Suppose that the DGP is

   $$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \varepsilon_t$$

   but we estimate

   $$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

Figure 9.5: Autocorrelation induced by misspecification



The effects are illustrated in Figure 9.5.

Effects on the OLS estimator

The variance of the OLS estimator is the same as in the case of heteroscedasticity - the standard formula does not apply. The correct formula is given in equation 9.1. Next we discuss two GLS corrections for OLS. These will potentially induce inconsistency when the regressors are nonstochastic

(see Chapter 6) and should either not be used in that case (which is usually the relevant case) or used with caution. The more recommended procedure is discussed in section 9.5.

AR(1)

There are many types of autocorrelation. We'll consider two examples. The first is the most commonly encountered case: autoregressive order 1 (AR(1) errors. The model is

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t \\ u_t &\sim iid(0, \sigma_u^2) \\ \mathcal{E}(\varepsilon_t u_s) &= 0, t < s \end{aligned}$$

We assume that the model satisfies the other classical assumptions.

- We need a stationarity assumption: $|\rho| < 1$. Otherwise the variance of ε_t explodes as t increases, so standard asymptotics will not apply.
- By recursive substitution we obtain

$$\begin{aligned} \varepsilon_t &= \rho\varepsilon_{t-1} + u_t \\ &= \rho\left(\rho\varepsilon_{t-2} + u_{t-1}\right) + u_t \\ &= \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t \\ &= \rho^2\left(\rho\varepsilon_{t-3} + u_{t-2}\right) + \rho u_{t-1} + u_t \end{aligned}$$

In the limit the lagged ε drops out, since $\rho^m \to 0$ as $m \to \infty$, so we obtain

$$\varepsilon_t = \sum_{m=0}^{\infty} \rho^m u_{t-m}$$

With this, the variance of ε_t is found as

$$\begin{aligned} \mathcal{E}(\varepsilon_t^2) &= \sigma_u^2 \sum_{m=0}^{\infty} \rho^{2m} \\ &= \frac{\sigma_u^2}{1-\rho^2} \end{aligned}$$

- If we had directly assumed that ε_t were covariance stationary, we could obtain this using

$$\begin{aligned} V(\varepsilon_t) &= \rho^2 \mathcal{E}(\varepsilon_{t-1}^2) + 2\rho \mathcal{E}(\varepsilon_{t-1} u_t) + \mathcal{E}(u_t^2) \\ &= \rho^2 V(\varepsilon_t) + \sigma_u^2, \end{aligned}$$

so

$$V(\varepsilon_t) = \frac{\sigma_u^2}{1-\rho^2}$$

- The variance is the 0^{th} order autocovariance: $\gamma_0 = V(\varepsilon_t)$
- Note that the variance does not depend on t

Likewise, the first order autocovariance γ_1 is

$$\begin{aligned} Cov(\varepsilon_t, \varepsilon_{t-1}) &= \gamma_s = \mathcal{E}((\rho\varepsilon_{t-1} + u_t)\,\varepsilon_{t-1}) \\ &= \rho V(\varepsilon_t) \\ &= \frac{\rho\sigma_u^2}{1-\rho^2} \end{aligned}$$

- Using the same method, we find that for $s < t$

$$Cov(\varepsilon_t, \varepsilon_{t-s}) = \gamma_s = \frac{\rho^s\sigma_u^2}{1-\rho^2}$$

- The autocovariances don't depend on t: the process $\{\varepsilon_t\}$ is *covariance stationary*

The *correlation (*in general, for r.v.'s x and y) is defined as

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\text{se}(x)\text{se}(y)}$$

but in this case, the two standard errors are the same, so the s-order autocorrelation ρ_s is

$$\rho_s = \rho^s$$

- All this means that the overall matrix Σ has the form

$$\Sigma = \underbrace{\frac{\sigma_u^2}{1-\rho^2}}_{\text{this is the variance}} \underbrace{\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & & \ddots & & \vdots \\ & & & \ddots & \rho \\ \rho^{n-1} & \cdots & & & 1 \end{bmatrix}}_{\text{this is the correlation matrix}}$$

  So we have homoscedasticity, but elements off the main diagonal are not zero. All of this depends only on two parameters, ρ and σ_u^2. If we can estimate these consistently, we can apply FGLS.

It turns out that it's easy to estimate these consistently. The steps are

1. Estimate the model $y_t = x_t'\beta + \varepsilon_t$ by OLS.

2. Take the residuals, and estimate the model

$$\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + u_t^*$$

   Since $\hat{\varepsilon}_t \xrightarrow{p} \varepsilon_t$, this regression is asymptotically equivalent to the regression

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

   which satisfies the classical assumptions. Therefore, $\hat{\rho}$ obtained by applying OLS to $\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + u_t^*$

is consistent. Also, since $u_t^* \xrightarrow{p} u_t$, the estimator

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{t=2}^{n} (\hat{u}_t^*)^2 \xrightarrow{p} \sigma_u^2$$

3. With the consistent estimators $\hat{\sigma}_u^2$ and $\hat{\rho}$, form $\hat{\Sigma} = \Sigma(\hat{\sigma}_u^2, \hat{\rho})$ using the previous structure of Σ, and estimate by FGLS. Actually, one can omit the factor $\hat{\sigma}_u^2/(1-\rho^2)$, since it cancels out in the formula

$$\hat{\beta}_{FGLS} = \left(X'\hat{\Sigma}^{-1}X\right)^{-1}(X'\hat{\Sigma}^{-1}y).$$

- One can iterate the process, by taking the first FGLS estimator of β, re-estimating ρ and σ_u^2, etc. If one iterates to convergences it's equivalent to MLE (supposing normal errors).

- An asymptotically equivalent approach is to simply estimate the transformed model

$$y_t - \hat{\rho}y_{t-1} = (x_t - \hat{\rho}x_{t-1})'\beta + u_t^*$$

using $n-1$ observations (since y_0 and x_0 aren't available). This is the method of Cochrane and Orcutt. Dropping the first observation is asymptotically irrelevant, but *it can be very important in small samples.* One can recuperate the first observation by putting

$$\begin{aligned} y_1^* &= y_1 \sqrt{1-\hat{\rho}^2} \\ x_1^* &= x_1 \sqrt{1-\hat{\rho}^2} \end{aligned}$$

This somewhat odd-looking result is related to the Cholesky factorization of Σ^{-1}. See Davidson and MacKinnon, pg. 348-49 for more discussion. Note that the variance of y_1^* is σ_u^2, asymptoti-

cally, so we see that the transformed model will be homoscedastic (and nonautocorrelated, since the $u's$ are uncorrelated with the $y's$, in different time periods.

MA(1)

The linear regression model with moving average order 1 errors is

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \varepsilon_t &= u_t + \phi u_{t-1} \\ u_t &\sim iid(0, \sigma_u^2) \\ \mathcal{E}(\varepsilon_t u_s) &= 0, t < s \end{aligned}$$

In this case,

$$\begin{aligned} V(\varepsilon_t) = \gamma_0 = & \mathcal{E}\left[(u_t + \phi u_{t-1})^2\right] \\ = & \sigma_u^2 + \phi^2 \sigma_u^2 \\ = & \sigma_u^2(1 + \phi^2) \end{aligned}$$

Similarly

$$\begin{aligned} \gamma_1 &= \mathcal{E}\left[(u_t + \phi u_{t-1})(u_{t-1} + \phi u_{t-2})\right] \\ &= \phi \sigma_u^2 \end{aligned}$$

and

$$\begin{aligned}\gamma_2 &= [(u_t + \phi u_{t-1})(u_{t-2} + \phi u_{t-3})] \\ &= 0\end{aligned}$$

so in this case

$$\Sigma = \sigma_u^2 \begin{bmatrix} 1+\phi^2 & \phi & 0 & \cdots & 0 \\ \phi & 1+\phi^2 & \phi & & \\ 0 & \phi & \ddots & & \vdots \\ \vdots & & & \ddots & \phi \\ 0 & \cdots & & \phi & 1+\phi^2 \end{bmatrix}$$

Note that the first order autocorrelation is

$$\begin{aligned}\rho_1 &= \frac{\phi\sigma_u^2}{\sigma_u^2(1+\phi^2)} &= \frac{\gamma_1}{\gamma_0} \\ &= & \frac{\phi}{(1+\phi^2)}\end{aligned}$$

- This achieves a maximum at $\phi = 1$ and a minimum at $\phi = -1$, and the maximal and minimal autocorrelations are 1/2 and -1/2. Therefore, series that are more strongly autocorrelated can't be MA(1) processes.

Again the covariance matrix has a simple structure that depends on only two parameters. The problem in this case is that one can't estimate ϕ using OLS on

$$\hat{\varepsilon}_t = u_t + \phi u_{t-1}$$

because the u_t are unobservable and they can't be estimated consistently. However, there is a simple way to estimate the parameters.

- Since the model is homoscedastic, we can estimate

$$V(\varepsilon_t) = \sigma_\varepsilon^2 = \sigma_u^2(1+\phi^2)$$

  using the typical estimator:

$$\widehat{\sigma_\varepsilon^2} = \sigma_u^2\widehat{(1+\phi^2)} = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^2$$

- By the Slutsky theorem, we can interpret this as defining an (unidentified) estimator of both σ_u^2 and ϕ, e.g., use this as

$$\widehat{\sigma_u^2}(1+\widehat{\phi^2}) = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^2$$

  However, this isn't sufficient to define consistent estimators of the parameters, since it's unidentified - two unknowns, one equation.

- To solve this problem, estimate the covariance of ε_t and ε_{t-1} using

$$\widehat{Cov}(\varepsilon_t, \varepsilon_{t-1}) = \widehat{\phi\sigma_u^2} = \frac{1}{n}\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}$$

  This is a consistent estimator, following a LLN (and given that the epsilon hats are consistent for the epsilons). As above, this can be interpreted as defining an unidentified estimator of the two parameters:

$$\hat{\phi}\widehat{\sigma_u^2} = \frac{1}{n}\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}$$

- Now solve these two equations to obtain identified (and therefore consistent) estimators of both ϕ and σ_u^2. Define the consistent estimator

$$\hat{\Sigma} = \Sigma(\hat{\phi}, \widehat{\sigma_u^2})$$

following the form we've seen above, and transform the model using the Cholesky decomposition. The transformed model satisfies the classical assumptions asymptotically.

- Note: there is no guarantee that Σ estimated using the above method will be positive definite, which may pose a problem. Another method would be to use ML estimation, if one is willing to make distributional assumptions regarding the white noise errors.

Monte Carlo example: AR1

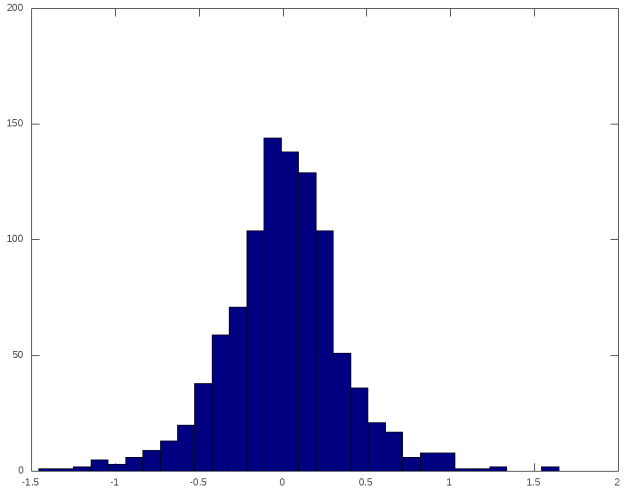
Let's look at a Monte Carlo study that compares OLS and GLS when we have AR1 errors. The model is

$$y_t = 1 + x_t + \epsilon_t$$
$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

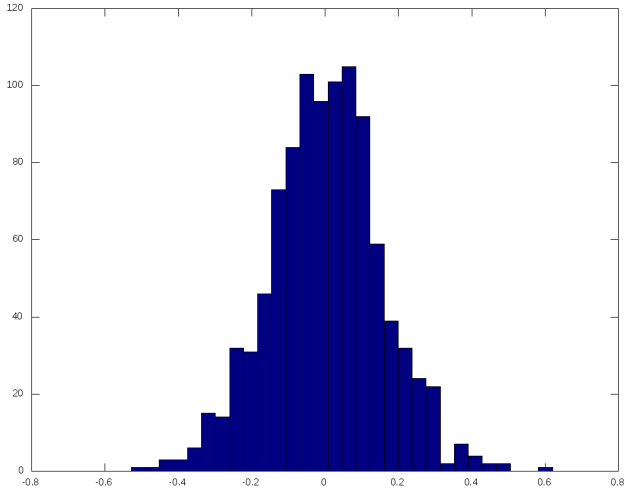
with $\rho = 0.9$. The sample size is $n = 30$, and 1000 Monte Carlo replications are done. The Octave script is GLS/AR1Errors.m. Figure 9.6 shows histograms of the estimated coefficient of x minus the true value. We can see that the GLS histogram is much more concentrated about 0, which is indicative of the efficiency of GLS relative to OLS.

Figure 9.6: Efficiency of OLS and FGLS, AR1 errors

(a) OLS



(b) GLS



Asymptotically valid inferences with autocorrelation of unknown form

See Hamilton Ch. 10, pp. 261-2 and 280-84.

When the form of autocorrelation is unknown, one may decide to use the OLS estimator, without correction. We've seen that this estimator has the limiting distribution

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right)$$

where, as before, Ω is

$$\Omega = \lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon' X}{n}\right)$$

We need a consistent estimate of Ω. Define $m_t = x_t\varepsilon_t$ (recall that x_t is defined as a $K \times 1$ vector). Note that

$$\begin{aligned} X'\varepsilon &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= \sum_{t=1}^{n} x_t\varepsilon_t \\ &= \sum_{t=1}^{n} m_t \end{aligned}$$

so that

$$\Omega = \lim_{n\to\infty} \frac{1}{n}\mathcal{E}\left[\left(\sum_{t=1}^{n} m_t\right)\left(\sum_{t=1}^{n} m_t'\right)\right]$$

We assume that m_t is covariance stationary (so that the covariance between m_t and m_{t-s} does not depend on t).

Define the $v-th$ autocovariance of m_t as

$$\Gamma_v = \mathcal{E}(m_t m_{t-v}').$$

Note that $\mathcal{E}(m_t m_{t+v}') = \Gamma_v'$. *(show this with an example).* In general, we expect that:

- m_t will be autocorrelated, since ε_t is potentially autocorrelated:

$$\Gamma_v = \mathcal{E}(m_t m_{t-v}') \neq 0$$

  Note that this autocovariance does not depend on t, due to covariance stationarity.

- contemporaneously correlated ($\mathcal{E}(m_{it} m_{jt}) \neq 0$), since the regressors in x_t will in general be correlated (more on this later).

- and heteroscedastic ($\mathcal{E}(m_{it}^2) = \sigma_i^2$, which depends upon i), again since the regressors will have different variances.

While one could estimate Ω parametrically, we in general have little information upon which to base a parametric specification. Recent research has focused on consistent nonparametric estimators of Ω.

Now define

$$\Omega_n = \mathcal{E}\frac{1}{n}\left[\left(\sum_{t=1}^{n} m_t\right)\left(\sum_{t=1}^{n} m_t'\right)\right]$$

We have (*show that the following is true, by expanding sum and shifting rows to left)*

$$\Omega_n = \Gamma_0 + \frac{n-1}{n}\left(\Gamma_1 + \Gamma_1'\right) + \frac{n-2}{n}\left(\Gamma_2 + \Gamma_2'\right) \cdots + \frac{1}{n}\left(\Gamma_{n-1} + \Gamma_{n-1}'\right)$$

The natural, consistent estimator of Γ_v is

$$\widehat{\Gamma_v} = \frac{1}{n}\sum_{t=v+1}^{n} \hat{m}_t \hat{m}_{t-v}'.$$

where

$$\hat{m}_t = x_t \hat{\varepsilon}_t$$

(note: one could put $1/(n-v)$ instead of $1/n$ here). So, a natural, but inconsistent, estimator of Ω_n would be

$$\begin{aligned}
\hat{\Omega}_n &= \widehat{\Gamma_0} + \frac{n-1}{n}\left(\widehat{\Gamma_1} + \widehat{\Gamma_1'}\right) + \frac{n-2}{n}\left(\widehat{\Gamma_2} + \widehat{\Gamma_2'}\right) + \cdots + \frac{1}{n}\left(\widehat{\Gamma_{n-1}} + \widehat{\Gamma_{n-1}'}\right) \\
&= \widehat{\Gamma_0} + \sum_{v=1}^{n-1} \frac{n-v}{n}\left(\widehat{\Gamma_v} + \widehat{\Gamma_v'}\right).
\end{aligned}$$

This estimator is inconsistent in general, since the number of parameters to estimate is more than the number of observations, and increases more rapidly than n, so information does not build up as $n \to \infty$.

On the other hand, supposing that Γ_v tends to zero sufficiently rapidly as v tends to ∞, a modified estimator

$$\hat{\Omega}_n = \widehat{\Gamma_0} + \sum_{v=1}^{q(n)} \left(\widehat{\Gamma_v} + \widehat{\Gamma_v'}\right),$$

where $q(n) \xrightarrow{p} \infty$ as $n \to \infty$ will be consistent, provided $q(n)$ grows sufficiently slowly.

- The assumption that autocorrelations die off is reasonable in many cases. For example, the AR(1) model with $|\rho| < 1$ has autocorrelations that die off.
- The term $\frac{n-v}{n}$ can be dropped because it tends to one for $v < q(n)$, given that $q(n)$ increases slowly relative to n.
- A disadvantage of this estimator is that is may not be positive definite. This could cause one to calculate a negative χ^2 statistic, for example!
- Newey and West proposed and estimator (*Econometrica*, 1987) that solves the problem of possible nonpositive definiteness of the above estimator. Their estimator is

$$\hat{\Omega}_n = \widehat{\Gamma_0} + \sum_{v=1}^{q(n)} \left[1 - \frac{v}{q+1}\right] \left(\widehat{\Gamma_v} + \widehat{\Gamma_v'}\right).$$

  This estimator is p.d. by construction. The condition for consistency is that $n^{-1/4}q(n) \to 0$. Note that this is a very slow rate of growth for q. This estimator is nonparametric - we've placed no parametric restrictions on the form of Ω. It is an example of a *kernel* estimator.

Finally, since Ω_n has Ω as its limit, $\hat{\Omega}_n \xrightarrow{p} \Omega$. We can now use $\hat{\Omega}_n$ and $\widehat{Q_X} = \frac{1}{n}X'X$ to consistently estimate the limiting distribution of the OLS estimator under heteroscedasticity and autocorrelation of unknown form. With this, asymptotically valid tests are constructed in the usual way.

Testing for autocorrelation

Durbin-Watson test

The Durbin-Watson test is not strictly valid in most situations where we would like to use it. Nevertheless, it is encountered often enough so that one should know something about it. The Durbin-Watson test statistic is

$$\begin{aligned} DW &= \frac{\sum_{t=2}^{n}\left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} \\ &= \frac{\sum_{t=2}^{n}\left(\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2\right)}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} \end{aligned}$$

- The null hypothesis is that the first order autocorrelation of the errors is zero: $H_0 : \rho_1 = 0$. The alternative is of course $H_A : \rho_1 \neq 0$. Note that the alternative is not that the errors are AR(1), since many general patterns of autocorrelation will have the first order autocorrelation different than zero. For this reason the test is useful for detecting autocorrelation in general. For the same reason, one shouldn't just assume that an AR(1) model is appropriate when the DW test rejects the null.

- Under the null, the middle term tends to zero, and the other two tend to one, so $DW \xrightarrow{p} 2$.

- Supposing that we had an AR(1) error process with $\rho = 1$. In this case the middle term tends to -2, so $DW \xrightarrow{p} 0$

- Supposing that we had an AR(1) error process with $\rho = -1$. In this case the middle term tends to 2, so $DW \xrightarrow{p} 4$

- These are the extremes: DW always lies between 0 and 4.

- The distribution of the test statistic depends on the matrix of regressors, X, so tables can't give

exact critical values. The give upper and lower bounds, which correspond to the extremes that are possible. See Figure 9.7. There are means of determining exact critical values conditional on X.

- Note that DW can be used to test for nonlinearity (add discussion).
- The DW test is based upon the assumption that the matrix X is fixed in repeated samples. This is often unreasonable in the context of economic time series, which is precisely the context where the test would have application. It is possible to relate the DW test to other test statistics which are valid without strict exogeneity.

Breusch-Godfrey test

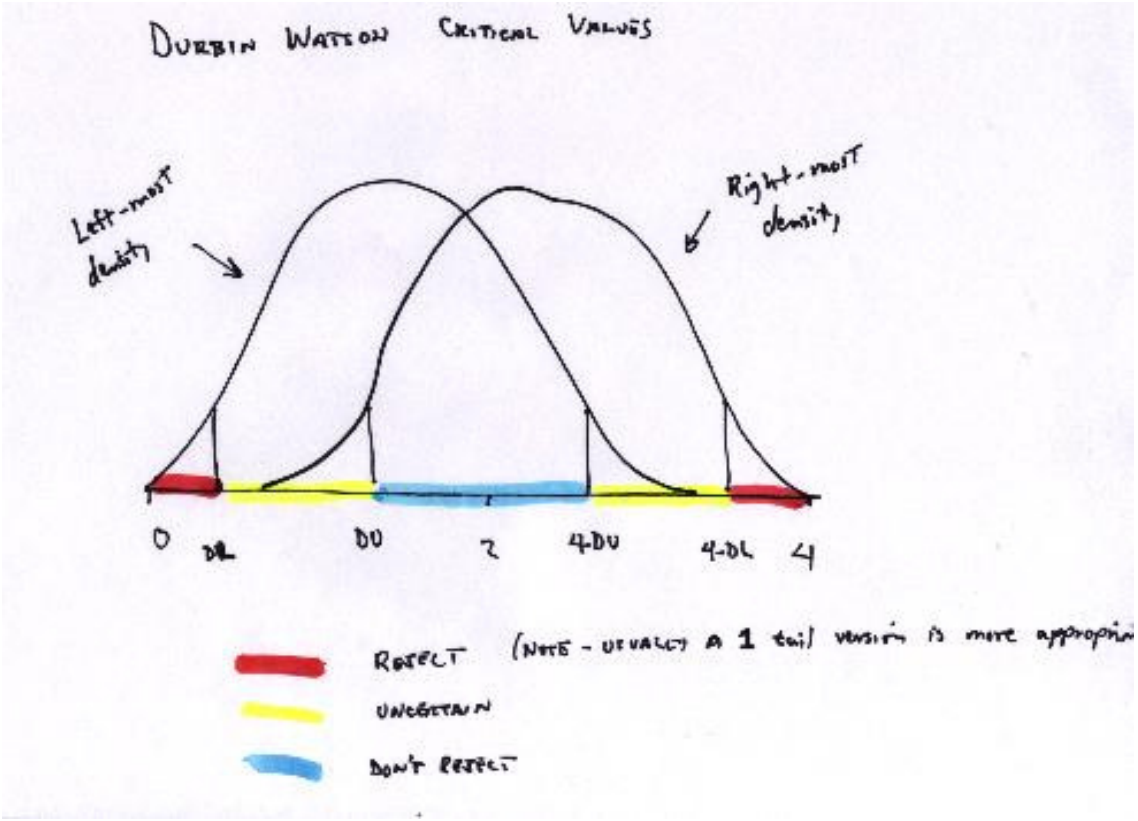
This test uses an auxiliary regression, as does the White test for heteroscedasticity. The regression is

$$\hat{\varepsilon}_t = x_t'\delta + \gamma_1\hat{\varepsilon}_{t-1} + \gamma_2\hat{\varepsilon}_{t-2} + \cdots + \gamma_P\hat{\varepsilon}_{t-P} + v_t$$

and the test statistic is the nR^2 statistic, just as in the White test. There are P restrictions, so the test statistic is asymptotically distributed as a $\chi^2(P)$.

- The intuition is that the lagged errors shouldn't contribute to explaining the current error if there is no autocorrelation.
- x_t is included as a regressor to account for the fact that the $\hat{\varepsilon}_t$ are not independent even if the ε_t are. This is a technicality that we won't go into here.
- This test is valid even if the regressors are stochastic and contain lagged dependent variables, so it is considerably more useful than the DW test for typical time series data.

Figure 9.7: Durbin-Watson critical values



- The alternative is not that the model is an AR(P), following the argument above. The alternative is simply that some or all of the first P autocorrelations are different from zero. This is compatible with many specific forms of autocorrelation.

Lagged dependent variables and autocorrelation

We've seen that the OLS estimator is consistent under autocorrelation, as long as $plim\frac{X'\varepsilon}{n} = 0$. This will be the case when $\mathcal{E}(X'\varepsilon) = 0$, following a LLN. An important exception is the case where X contains lagged $y's$ and the errors are autocorrelated.

Example 22. Dynamic model with MA1 errors. Consider the model

$$\begin{aligned} y_t &= \alpha + \rho y_{t-1} + \beta x_t + \epsilon_t \\ \epsilon_t &= \upsilon_t + \phi \upsilon_{t-1} \end{aligned}$$

We can easily see that a regressor is not weakly exogenous:

$$\begin{aligned} \mathcal{E}(y_{t-1}\varepsilon_t) &= \mathcal{E}\left\{(\alpha + \rho y_{t-2} + \beta x_{t-1} + \upsilon_{t-1} + \phi \upsilon_{t-2})(\upsilon_t + \phi \upsilon_{t-1})\right\} \\ &\neq 0 \end{aligned}$$

since one of the terms is $\mathcal{E}(\phi \upsilon_{t-1}^2)$ which is clearly nonzero. In this case $\mathcal{E}(\mathbf{x}_t\varepsilon_t) \neq 0$, and therefore $plim\frac{X'\varepsilon}{n} \neq 0$. Since

$$plim\hat{\beta} = \beta + plim\frac{X'\varepsilon}{n}$$

the OLS estimator is inconsistent in this case. One needs to estimate by instrumental variables (IV), which we'll get to later

The Octave script GLS/DynamicMA.m does a Monte Carlo study. The sample size is $n = 100$. The true coefficients are $\alpha = 1$ $\rho = 0.9$ and $\beta = 1$. The MA parameter is $\phi = -0.95$. Figure 9.8 gives the results. You can see that the constant and the autoregressive parameter have a lot of bias. By re-running the script with $\phi = 0$, you will see that much of the bias disappears (not all - why?).

Examples

Nerlove model, yet again The Nerlove model uses cross-sectional data, so one may not think of performing tests for autocorrelation. However, specification error can induce autocorrelated errors. Consider the simple Nerlove model

$$\ln C = \beta_1 + \beta_2 \ln Q + \beta_3 \ln P_L + \beta_4 \ln P_F + \beta_5 \ln P_K + \epsilon$$

and the extended Nerlove model

$$\ln C = \sum_{j=1}^{5} \alpha_j D_j + \sum_{j=1}^{5} \gamma_j D_j \ln Q + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon$$

discussed around equation 5.5. If you have done the exercises, you have seen evidence that the extended model is preferred. So if it is in fact the proper model, the simple model is misspecified. Let's check if this misspecification might induce autocorrelated errors.

The Octave program GLS/NerloveAR.m estimates the simple Nerlove model, and plots the residuals as a function of $\ln Q$, and it calculates a Breusch-Godfrey test statistic. The residual plot is in Figure 9.9 , and the test results are:

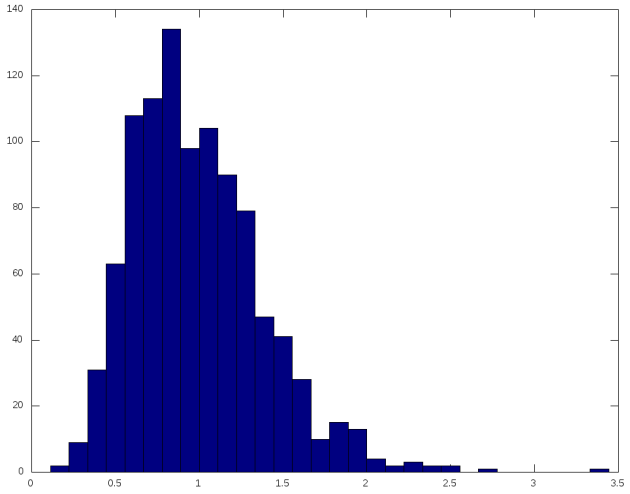
```
                          Value      p-value
Breusch-Godfrey test      34.930       0.000
```

Clearly, there is a problem of autocorrelated residuals.

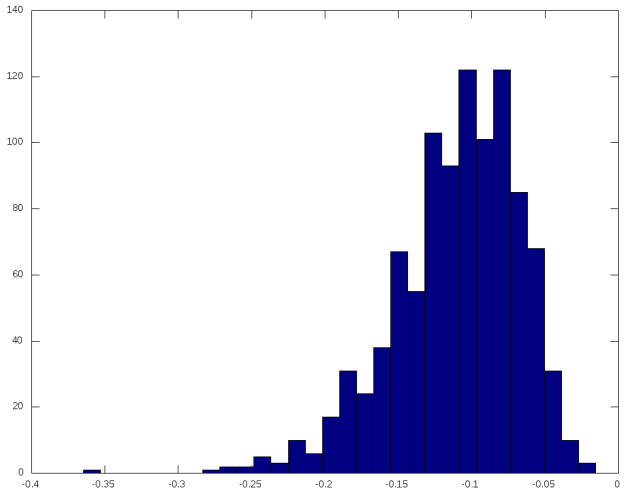
Repeat the autocorrelation tests using the extended Nerlove model (Equation 5.5) to see the problem is solved.

Figure 9.8: Dynamic model with MA(1) errors

(a) $\hat{\alpha} - \alpha$

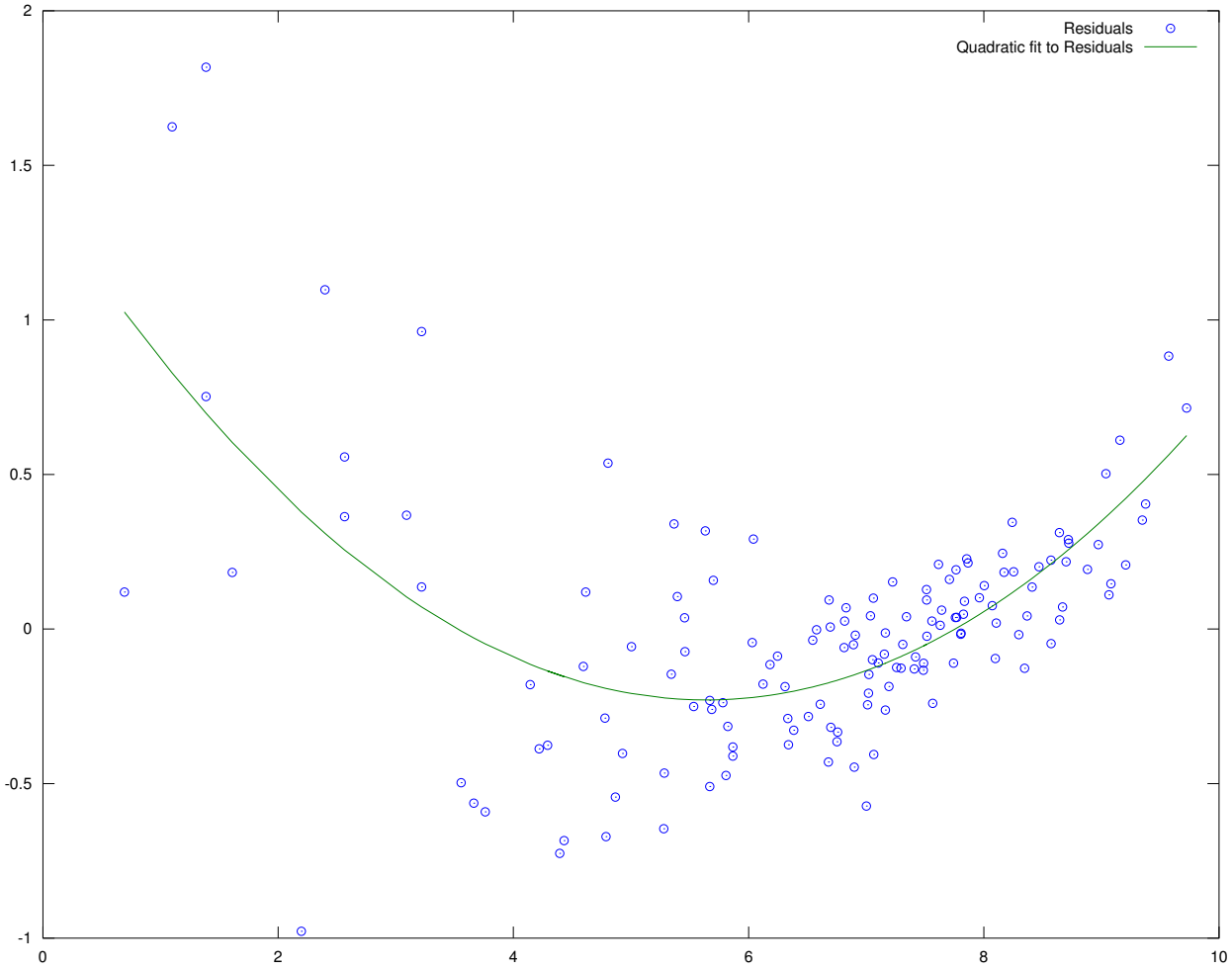


(b) $\hat{\rho} - \rho$



(c) $\hat{\beta} - \beta$

Figure 9.9: Residuals of simple Nerlove model



Klein model Klein's Model I is a simple macroeconometric model. One of the equations in the model explains consumption (C) as a function of profits (P), both current and lagged, as well as the sum of wages in the private sector (W^p) and wages in the government sector (W^g). Have a look at the README file for this data set. This gives the variable names and other information.

Consider the model

$$C_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^p + W_t^g) + \epsilon_{1t}$$

The Octave program GLS/Klein.m estimates this model by OLS, plots the residuals, and performs the Breusch-Godfrey test, using 1 lag of the residuals. The estimation and test results are:

```
**********************************************************
OLS estimation results
Observations 21
R-squared 0.981008
Sigma-squared 1.051732

Results (Ordinary var-cov estimator)

                 estimate      st.err.      t-stat.      p-value
Constant           16.237        1.303       12.464        0.000
Profits             0.193        0.091        2.115        0.049
Lagged Profits      0.090        0.091        0.992        0.335
Wages               0.796        0.040       19.933        0.000

**********************************************************
```

	Value	p-value
Breusch-Godfrey test	1.539	0.215

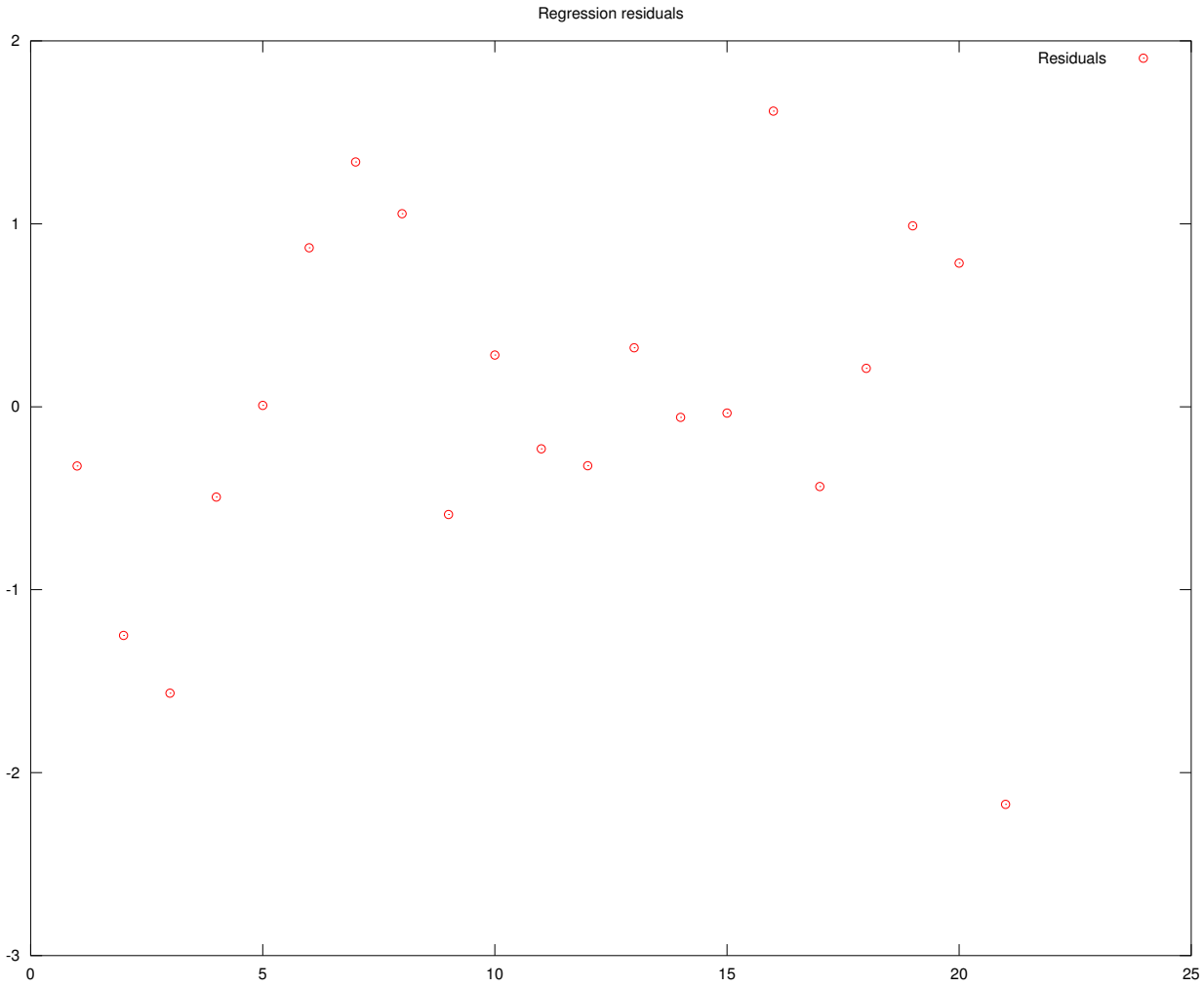
and the residual plot is in Figure 9.10. The test does not reject the null of nonautocorrelatetd errors, but we should remember that we have only 21 observations, so power is likely to be fairly low. The residual plot leads me to suspect that there may be autocorrelation - there are some significant runs below and above the x-axis. Your opinion may differ.

Since it seems that there *may* be autocorrelation, lets's try an AR(1) correction. The Octave program GLS/KleinAR1.m estimates the Klein consumption equation assuming that the errors follow the AR(1) pattern. The results, with the Breusch-Godfrey test for remaining autocorrelation are:

```
*******************************************************
OLS estimation results
Observations 21
R-squared 0.967090
Sigma-squared 0.983171

Results (Ordinary var-cov estimator)
```

	estimate	st.err.	t-stat.	p-value
Constant	16.992	1.492	11.388	0.000
Profits	0.215	0.096	2.232	0.039
Lagged Profits	0.076	0.094	0.806	0.431
Wages	0.774	0.048	16.234	0.000

Figure 9.10: OLS residuals, Klein consumption equation



```
*******************************************************
                           Value     p-value
Breusch-Godfrey test       2.129       0.345
```

- The test is farther away from the rejection region than before, and the residual plot is a bit more favorable for the hypothesis of nonautocorrelated residuals, IMHO. For this reason, it seems that the AR(1) correction might have improved the estimation.
- Nevertheless, there has not been much of an effect on the estimated coefficients nor on their estimated standard errors. This is probably because the estimated AR(1) coefficient is not very large (around 0.2)
- The existence or not of autocorrelation in this model will be important later, in the section on simultaneous equations.

9.6 Exercises

1. Comparing the variances of the OLS and GLS estimators, I claimed that the following holds:

$$Var(\hat{\beta}) - Var(\hat{\beta}_{GLS}) \;=\; A\Sigma A^{'}$$

   Verify that this is true.

2. Show that the GLS estimator can be defined as

$$\hat{\beta}_{GLS} = \arg\min(y - X\beta)'\Sigma^{-1}(y - X\beta)$$

3. The limiting distribution of the OLS estimator with heteroscedasticity of unknown form is

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right),$$

where

$$\lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon'X}{n}\right) = \Omega$$

Explain why

$$\widehat{\Omega} = \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \hat{\varepsilon}_t^2$$

is a consistent estimator of this matrix.

4. Define the $v - th$ autocovariance of a covariance stationary process m_t, where $E(m_t) = 0$ as

$$\Gamma_v = \mathcal{E}(m_t m_{t-v}').$$

Show that $\mathcal{E}(m_t m_{t+v}') = \Gamma_v'$.

5. For the Nerlove model with dummies and interactions discussed above (see Section 9.4 and equation 5.5)

$$\ln C = \sum_{j=1}^{5} \alpha_j D_j + \sum_{j=1}^{5} \gamma_j D_j \ln Q + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon$$

above, we did a GLS correction based on the assumption that there is HET by groups ($V(\epsilon_t|x_t) = \sigma_j^2$). Let's assume that this model is correctly specified, except that there may or may not be HET, and if it is present it may be of the form assumed, or perhaps of some other form. What happens if the assumed form of HET is incorrect?

(a) Is the "FGLS" based on the assumed form of HET consistent?

(b) Is it efficient? Is it likely to be efficient with respect to OLS?

(c) Are hypothesis tests using the "FGLS" estimator valid? If not, can they be made valid following some procedure? Explain.

(d) Are the t-statistics reported in Section 9.4 valid?

(e) Which estimator do you prefer, the OLS estimator or the FGLS estimator? Discuss.

6. Perhaps we can be a little more parsimonious with the Nerlove data (nerlove.data), rather than using so many parameters to account for non-constant returns to scale, and to account for heteroscedasticity. Consider the original model

$$\ln C = \beta + \beta_Q \ln Q + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon$$

(a) Estimate by OLS, plot the residuals, and test for autocorrelation and heteroscedasticity. Explain your findings.

(b) Consider the model

$$\ln C = \beta + \beta_Q \ln Q + \gamma_Q (\ln Q)^2 + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon$$

i. Explain how this model can account for non-constant returns to scale.

ii. estimate this model, and test for autocorrelation and heteroscedasticity. You should find that there is HET, but no strong evidence of AUT. Why is this the case?

iii. Do a GLS correction where it is assumed that $V(\epsilon_i) = \frac{\sigma^2}{(\ln Q_i)^2}$. In GRETL, there is a weighted least squares option that you can use. Why does this assumed form of HET make sense?

iv. plot the weighted residuals versus output. Is there evidence of HET, or has the correction eliminated the problem?

v. plot the fitted values for returns to scale, for all of the firms.

7. The hall.csv or hall.gdt dataset contains monthly observation on 3 variables: the consumption ratio c_t/c_{t-1}; the gross return of an equally weighted index of assets ewr_t; and the gross return of the same index, but weighted by value, vwr_t. The idea is that a representative consumer may finance consumption by investing in assets. Present wealth is used for two things: consumption and investment. The return on investment defines wealth in the next period, and the process repeats. For the moment, explore the properties of the variables.

   (a) Are the variances constant over time?

   (b) Do the variables appear to be autocorrelated? Hint: regress a variable on its own lags.

   (c) Do the variable seem to be normally distributed?

   (d) Look at the properties of the growth rates of the variables: repeat a-c for growth rates. The growth rate of a variable x_t is given by $\log(x_t/x_{t-1})$.

8. Consider the model

$$\begin{aligned} y_t &= C + A_1 y_{t-1} + \epsilon_t \\ E(\epsilon_t \epsilon_t') &= \Sigma \\ E(\epsilon_t \epsilon_s') &= 0, t \neq s \end{aligned}$$

where y_t and ϵ_t are $G \times 1$ vectors, C is a $G \times 1$ of constants, and $A_1$is a $G \times G$ matrix of parameters. The matrix Σ is a $G \times G$ covariance matrix. Assume that we have n observations. This is a *vector autoregressive* model, of order 1 - commonly referred to as a VAR(1) model.

(a) Show how the model can be written in the form $Y = X\beta + \nu$, where Y is a $Gn \times 1$ vector, β is a $(G + G^2)\times 1$ parameter vector, and the other items are conformable. What is the structure of X? What is the structure of the covariance matrix of ν?

(b) This model has HET and AUT. Verify this statement.

(c) Set $G = 2$,$C = (0\,0)'A = \begin{bmatrix} 0.8 & -0.1 \\ 0.2 & 0.5 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. Simulate data from this model, then estimate the model using OLS and feasible GLS. You should find that the two estimators are identical, which might seem surprising, given that there is HET and AUT.

(d) (optional, and advanced). Prove analytically that the OLS and GLS estimators are identical. Hint: this model is of the form of *seemingly unrelated regressions*.

9. Consider the model

$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t$$

where ϵ_t is a $N(0,1)$ white noise error. This is an autogressive model of order 2 (AR2) model. Suppose that data is generated from the AR2 model, but the econometrician mistakenly decides to estimate an AR1 model ($y_t = \alpha + \rho_1 y_{t-1} + \epsilon_t$).

(a) simulate data from the AR2 model, setting $\rho_1 = 0.5$ and $\rho_2 = 0.4$, using a sample size of $n = 30$.

(b) Estimate the AR1 model by OLS, using the simulated data

(c) test the hypothesis that $\rho_1 = 0.5$

(d) test for autocorrelation using the test of your choice

(e) repeat the above steps 10000 times.

i. What percentage of the time does a t-test reject the hypothesis that $\rho_1 = 0.5$?

ii. What percentage of the time is the hypothesis of no autocorrelation rejected?

(f) discuss your findings. Include a residual plot for a representative sample.

10. Modify the script given in Subsection 9.5 so that the first observation is dropped, rather than given special treatment. This corresponds to using the Cochrane-Orcutt method, whereas the script as provided implements the Prais-Winsten method. Check if there is an efficiency loss when the first observation is dropped.

Chapter 10

Endogeneity and simultaneity

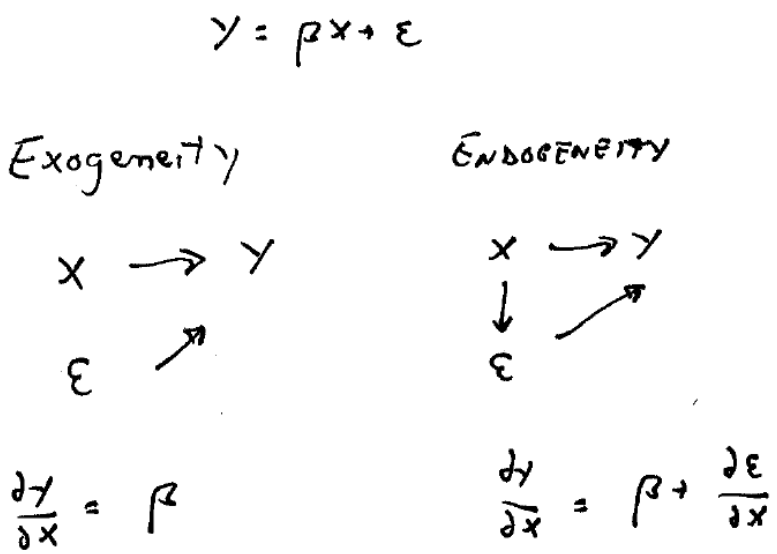
Several times we've encountered cases where correlation between regressors and the error term lead to biasedness and inconsistency of the OLS estimator. Cases include autocorrelation with lagged dependent variables (Example 22), measurement error in the regressors (Example 19) and missing regressors (Section 7.4). Another important case we have not seen yet is that of simultaneous equations. The cause is different, but the effect is the same: bias and inconsistency when OLS is applied to a single equation. The basic idea is presented in Figure 10.1. A simple regression will estimate the overall effect of x on y. If we're interested in the direct effect, β, then we have a problem when the overall effect and the direct effect differ.

10.1 Simultaneous equations

Up until now our model is

$$y = X\beta + \varepsilon$$

Figure 10.1: Exogeneity and Endogeneity (adapted from Cameron and Trivedi)



where we assume weak exogeneity of the regressors, so that $E(x_t\epsilon_t) = 0$. With weak exogeneity, the OLS estimator has desirable large sample properties (consistency, asymptotic normality).

Simultaneous equations is a different prospect. An example of a simultaneous equation system is a simple supply-demand system:

$$\begin{aligned}
\text{Demand: } q_t &= \alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t} \\
\text{Supply: } q_t &= \beta_1 + \beta_2 p_t + \varepsilon_{2t} \\
\mathcal{E}\left(\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} & \varepsilon_{2t} \end{bmatrix}\right) &= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \cdot & \sigma_{22} \end{bmatrix} \\
&\equiv \Sigma, \forall t
\end{aligned}$$

The presumption is that q_t and p_t are jointly determined at the same time by the intersection of these equations. We'll assume that y_t is determined by some unrelated process. It's easy to see that we have correlation between regressors and errors. Solving for p_t :

$$\begin{aligned}
\alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t} &= \beta_1 + \beta_2 p_t + \varepsilon_{2t} \\
\beta_2 p_t - \alpha_2 p_t &= \alpha_1 - \beta_1 + \alpha_3 y_t + \varepsilon_{1t} - \varepsilon_{2t} \\
p_t &= \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}
\end{aligned}$$

Now consider whether p_t is uncorrelated with ε_{1t} :

$$\begin{aligned}
\mathcal{E}(p_t \varepsilon_{1t}) &= \mathcal{E}\left\{\left(\frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\varepsilon_{1t}\right\} \\
&= \frac{\sigma_{11} - \sigma_{12}}{\beta_2 - \alpha_2}
\end{aligned}$$

Because of this correlation, weak exogeneity does not hold, and OLS estimation of the demand equation will be biased and inconsistent. The same applies to the supply equation, for the same reason.

In this model, q_t and p_t are the *endogenous* varibles (endogs), that are determined within the system. y_t is an *exogenous* variable (exogs). These concepts are a bit tricky, and we'll return to it in a minute. First, some notation. Suppose we group together current endogs in the vector Y_t. If there are G endogs, Y_t is $G \times 1$. Group current and lagged exogs, as well as lagged endogs in the vector X_t , which is $K \times 1$. Stack the errors of the G equations into the error vector E_t. The model, with additional assumtions, can be written as

$$\begin{aligned} Y_t'\Gamma &= X_t'B + E_t' \\ E_t &\sim N(0, \Sigma), \forall t \\ \mathcal{E}(E_t E_s') &= 0, t \neq s \end{aligned} \tag{10.1}$$

There are G equations here, and the parameters that enter into each equation are contained in the *columns* of the matrices Γ and B. We can stack all n observations and write the model as

$$\begin{aligned} Y\Gamma &= XB + E \\ \mathcal{E}(X'E) &= 0_{(K \times G)} \\ vec(E) &\sim N(0, \Psi) \end{aligned}$$

where

$$Y = \begin{bmatrix} Y_1' \\ Y_2' \\ \vdots \\ Y_n' \end{bmatrix}, X = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}, E = \begin{bmatrix} E_1' \\ E_2' \\ \vdots \\ E_n' \end{bmatrix}$$

Y is $n \times G$, X is $n \times K$, and E is $n \times G$.

- This system is *complete*, in that there are as many equations as endogs.
- There is a normality assumption. This isn't necessary, but allows us to consider the relationship between least squares and ML estimators.
- Since there is no autocorrelation of the E_t 's, and since the columns of E are individually homoscedastic, then

$$\begin{aligned} \Psi &= \begin{bmatrix} \sigma_{11} I_n & \sigma_{12} I_n & \cdots & \sigma_{1G} I_n \\ & \sigma_{22} I_n & & \vdots \\ & & \ddots & \vdots \\ \cdot & & & \sigma_{GG} I_n \end{bmatrix} \\ &= I_n \otimes \Sigma \end{aligned}$$

- X may contain lagged endogenous and exogenous variables. These variables are *predetermined.*
- We need to define what is meant by "endogenous" and "exogenous" when classifying the current period variables. Remember the definition of weak exogeneity Assumption 15, the regressors are weakly exogenous if $E(E_t|X_t) = 0$. Endogenous regressors are those for which this assumption

does not hold. As long as there is no autocorrelation, lagged endogenous variables are weakly exogenous.

10.2 Reduced form

Recall that the model is

$$\begin{aligned} Y_t'\Gamma &= X_t'B + E_t' \\ V(E_t) &= \Sigma \end{aligned}$$

This is the model in *structural form.*

Definition 23. [Structural form] An equation is in structural form when more than one current period endogenous variable is included.

The solution for the current period endogs is easy to find. It is

$$\begin{aligned} Y_t' &= X_t'B\Gamma^{-1} + E_t'\Gamma^{-1} \\ &= X_t'\Pi + V_t' \end{aligned}$$

Now only one current period endog appears in each equation. This is the *reduced form.*

Definition 24. [Reduced form] An equation is in reduced form if only one current period endog is included.

An example is our supply/demand system. The reduced form for quantity is obtained by solving the supply equation for price and substituting into demand:

$$\begin{aligned} q_t &= \alpha_1 + \alpha_2\left(\frac{q_t - \beta_1 - \varepsilon_{2t}}{\beta_2}\right) + \alpha_3 y_t + \varepsilon_{1t} \\ \beta_2 q_t - \alpha_2 q_t &= \beta_2\alpha_1 - \alpha_2(\beta_1 + \varepsilon_{2t}) + \beta_2\alpha_3 y_t + \beta_2\varepsilon_{1t} \\ q_t &= \frac{\beta_2\alpha_1 - \alpha_2\beta_1}{\beta_2 - \alpha_2} + \frac{\beta_2\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\beta_2\varepsilon_{1t} - \alpha_2\varepsilon_{2t}}{\beta_2 - \alpha_2} \\ &= \pi_{11} + \pi_{21} y_t + V_{1t} \end{aligned}$$

Similarly, the rf for price is

$$\begin{aligned} \beta_1 + \beta_2 p_t + \varepsilon_{2t} &= \alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t} \\ \beta_2 p_t - \alpha_2 p_t &= \alpha_1 - \beta_1 + \alpha_3 y_t + \varepsilon_{1t} - \varepsilon_{2t} \\ p_t &= \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2} \\ &= \pi_{12} + \pi_{22} y_t + V_{2t} \end{aligned}$$

The interesting thing about the rf is that the equations individually satisfy the classical assumptions, since y_t is uncorrelated with ε_{1t} and ε_{2t} by assumption, and therefore $\mathcal{E}(y_t V_{it}) = 0$, i=1,2, $\forall t$. The errors of the rf are

$$\begin{bmatrix} V_{1t} \\ V_{2t} \end{bmatrix} = \begin{bmatrix} \frac{\beta_2\varepsilon_{1t} - \alpha_2\varepsilon_{2t}}{\beta_2 - \alpha_2} \\ \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2} \end{bmatrix}$$

The variance of V_{1t} is

$$\begin{aligned} V(V_{1t}) &= \mathcal{E}\left[\left(\frac{\beta_2\varepsilon_{1t} - \alpha_2\varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\left(\frac{\beta_2\varepsilon_{1t} - \alpha_2\varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\right] \\ &= \frac{\beta_2^2\sigma_{11} - 2\beta_2\alpha_2\sigma_{12} + \alpha_2\sigma_{22}}{(\beta_2 - \alpha_2)^2} \end{aligned}$$

- This is constant over time, so the first rf equation is homoscedastic.
- Likewise, since the ε_t are independent over time, so are the V_t.

The variance of the second rf error is

$$\begin{aligned} V(V_{2t}) &= \mathcal{E}\left[\left(\frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\left(\frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\right] \\ &= \frac{\sigma_{11} - 2\sigma_{12} + \sigma_{22}}{(\beta_2 - \alpha_2)^2} \end{aligned}$$

and the contemporaneous covariance of the errors across equations is

$$\begin{aligned} \mathcal{E}(V_{1t}V_{2t}) &= \mathcal{E}\left[\left(\frac{\beta_2\varepsilon_{1t} - \alpha_2\varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\left(\frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\right] \\ &= \frac{\beta_2\sigma_{11} - (\beta_2 + \alpha_2)\,\sigma_{12} + \sigma_{22}}{(\beta_2 - \alpha_2)^2} \end{aligned}$$

- In summary the rf equations individually satisfy the classical assumptions, under the assumtions we've made, but they are contemporaneously correlated.

The general form of the rf is

$$\begin{aligned} Y_t' &= X_t' B\Gamma^{-1} + E_t'\Gamma^{-1} \\ &= X_t'\Pi + V_t' \end{aligned}$$

so we have that

$$V_t = \left(\Gamma^{-1}\right)' E_t \sim N\left(0, \left(\Gamma^{-1}\right)' \Sigma\Gamma^{-1}\right), \forall t$$

and that the V_t are timewise independent (note that this wouldn't be the case if the E_t were autocorrelated).

From the reduced form, we can easily see that the endogenous variables are correlated with the structural errors:

$$\begin{aligned} E(E_t Y_t') &= E\left(E_t\left(X_t' B\Gamma^{-1} + E_t'\Gamma^{-1}\right)\right) \\ &= E\left(E_t X_t' B\Gamma^{-1} + E_t E_t'\Gamma^{-1}\right) \\ &= \Sigma\Gamma^{-1} \end{aligned} \tag{10.2}$$

10.3 Estimation of the reduced form equations

From above, the RF equations are

$$\begin{aligned} Y_t' &= X_t' B\Gamma^{-1} + E_t'\Gamma^{-1} \\ &= X_t'\Pi + V_t' \end{aligned}$$

and

$$V_t \sim N\left(0, \Xi\right), \forall t$$

where we define $\Xi \equiv \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1}$. The rf parameter estimator $\hat{\Pi}$, is simply OLS applied to this model, equation by equation::

$$\hat{\Pi} = (X'X)^{-1} X'Y$$

which is simply

$$\hat{\Pi} = (X'X)^{-1} X' \left[\begin{array}{cccc} y_1 & y_2 & \cdots & y_G \end{array} \right]$$

that is, OLS equation by equation using *all* the exogs in the estimation of each column of Π.

It may seem odd that we use OLS on the reduced form, since the rf equations are correlated, because $\Xi \equiv \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1}$ is a full matrix. Why don't we do GLS to improve efficiency of estimation of the RF parameters?

OLS equation by equation to get the rf is equivalent to

$$\left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_G \end{array} \right] = \left[\begin{array}{cccc} X & 0 & \cdots & 0 \\ 0 & X & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & X \end{array} \right] \left[\begin{array}{c} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_G \end{array} \right] + \left[\begin{array}{c} v_1 \\ v_2 \\ \vdots \\ v_G \end{array} \right]$$

where y_i is the $n \times 1$ vector of observations of the i^{th} endog, X is the entire $n \times K$ matrix of exogs, π_i is the i^{th} column of Π, and v_i is the i^{th} column of V. Use the notation

$$y = \mathbf{X}\pi + v$$

to indicate the pooled model. Following this notation, the error covariance matrix is

$$V(v) = \Xi \otimes I_n$$

- This is a special case of a type of model known as a set of *seemingly unrelated equations (SUR)* since the parameter vector of each equation is different. The important feature of this special case is that *the regressors are the same in each equation*. The equations are contemporanously correlated, because of the non-zero off diagonal elements in Ξ.
- Note that each equation of the system individually satisfies the classical assumptions.
- Normally when doing SUR, one simply does GLS on the whole system $y = \mathbf{X}\pi + v$, where $V(v) = \Xi \otimes I_n$, which is in general more efficient than OLS on each equation.
- However, when the regressors are the same in all equations, as is true in the present case of estimation of the RF parameters, SUR $\equiv$OLS. To show this note that in this case $\mathbf{X} = I_n \otimes X$. Using the rules

  1. $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$
  2. $(A \otimes B)' = (A' \otimes B')$ and

3. $(A \otimes B)(C \otimes D) = (AC \otimes BD)$, we get

$$\begin{aligned}
\hat{\pi}_{SUR} &= \left((I_n \otimes X)' (\Xi \otimes I_n)^{-1} (I_n \otimes X)\right)^{-1} (I_n \otimes X)' (\Xi \otimes I_n)^{-1} y \\
&= \left(\left(\Xi^{-1} \otimes X'\right)(I_n \otimes X)\right)^{-1} \left(\Xi^{-1} \otimes X'\right) y \\
&= \left(\Xi \otimes (X'X)^{-1}\right)\left(\Xi^{-1} \otimes X'\right) y \\
&= \left[I_G \otimes (X'X)^{-1} X'\right] y \\
&= \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \vdots \\ \hat{\pi}_G \end{bmatrix}
\end{aligned}$$

- Note that this provides the answer to the exercise 8d in the chapter on GLS.
- So the unrestricted rf coefficients can be estimated efficiently (assuming normality) by OLS, even if the equations are correlated.
- We have ignored any potential zeros in the matrix Π, which if they exist could potentially increase the efficiency of estimation of the rf.
- Another example where SUR≡OLS is in estimation of vector autoregressions which is discussed in Section 15.2.

10.4 Bias and inconsistency of OLS estimation of a structural equation

Considering the first equation (this is without loss of generality, since we can always reorder the equations) we can partition the Y matrix as

$$Y = \begin{bmatrix} y & Y_1 & Y_2 \end{bmatrix}$$

- y is the first column
- Y_1 are the other endogenous variables that enter the first equation
- Y_2 are endogs that are excluded from this equation

Similarly, partition X as

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

- X_1 are the included exogs, and X_2 are the excluded exogs.

Finally, partition the error matrix as

$$E = \begin{bmatrix} \varepsilon & E_{12} \end{bmatrix}$$

Assume that Γ has ones on the main diagonal. These are normalization restrictions that simply scale the remaining coefficients on each equation, and which scale the variances of the error terms.

Given this scaling and our partitioning, the coefficient matrices can be written as

$$\begin{aligned}\Gamma &= \begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \end{bmatrix} \\ B &= \begin{bmatrix} \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix}\end{aligned}$$

With this, the first equation can be written as

$$\begin{aligned} y &= Y_1\gamma_1 + X_1\beta_1 + \varepsilon \\ &= Z\delta + \varepsilon \end{aligned} \tag{10.3}$$

The problem, as we've seen, is that the columns of Z corresponding to Y_1 are correlated with ε, because these are endogenous variables, and as we saw in equation 10.2, the endogenous variables are correlated with the structural errors, so they don't satisfy weak exogeneity. So, $E(Z'\epsilon) \neq 0$. What are the properties of the OLS estimator in this situation?

$$\begin{aligned} \hat{\delta} &= (Z'Z)^{-1} Z'y \\ &= (Z'Z)^{-1} Z' \left(Z\delta^0 + \varepsilon\right) \\ &= \delta^0 + (Z'Z)^{-1} Z'\epsilon \end{aligned}$$

It's clear that the OLS estimator is biased in general. Also,

$$\hat{\delta} - \delta^0 = \left(\frac{Z'Z}{n}\right)^{-1} \frac{Z'\epsilon}{n}$$

Say that $\lim \frac{Z'\epsilon}{n} = A$,a.s., and $\lim \frac{Z'Z}{n} = Q_Z$, $a.s.$ Then

$$\lim \left(\hat{\delta} - \delta^0\right) = Q_Z^{-1} A \neq 0, \; a.s.$$

So the OLS estimator of a structural equation is inconsistent. In general, correlation between regressors and errors leads to this problem, whether due to measurement error, simultaneity, or omitted regressors.

10.5 Note about the rest of this chaper

In class, I will not teach the material in the rest of this chapter at this time, but instead we will go on to GMM. The material that follows is easier to understand in the context of GMM, where we get a nice unified theory.

10.6 Identification by exclusion restrictions

The material in the rest of this chapter is no longer used in classes, but I'm leaving it in the notes for reference.

The identification problem in simultaneous equations is in fact of the same nature as the identification problem in any estimation setting: does the limiting objective function have the proper curvature

so that there is a unique global minimum or maximum at the true parameter value? In the context of IV estimation, this is the case if the limiting covariance of the IV estimator is positive definite and $plim\frac{1}{n}W'\varepsilon = 0$. This matrix is

$$V_\infty(\hat{\beta}_{IV}) = (Q_{XW}Q_{WW}^{-1}Q'_{XW})^{-1}\sigma^2$$

- The necessary and sufficient condition for identification is simply that this matrix be positive definite, and that the instruments be (asymptotically) uncorrelated with ε.
- For this matrix to be positive definite, we need that the conditions noted above hold: Q_{WW} must be positive definite and Q_{XW} must be of full rank (K).
- These identification conditions are not that intuitive nor is it very obvious how to check them.

Necessary conditions

If we use IV estimation for a single equation of the system, the equation can be written as

$$y = Z\delta + \varepsilon$$

where

$$Z = \begin{bmatrix} Y_1 & X_1 \end{bmatrix}$$

Notation:

- Let K be the total numer of weakly exogenous variables.

- Let $K^* = cols(X_1)$ be the number of included exogs, and let $K^{**} = K - K^*$ be the number of excluded exogs (in this equation).
- Let $G^* = cols(Y_1) + 1$ be the total number of included endogs, and let $G^{**} = G - G^*$ be the number of excluded endogs.

Using this notation, consider the selection of instruments.

- Now the X_1 are weakly exogenous and can serve as their own instruments.
- It turns out that X exhausts the set of possible instruments, in that if the variables in X don't lead to an identified model then no other instruments will identify the model either. Assuming this is true (we'll prove it in a moment), then a necessary condition for identification is that $cols(X_2) \geq cols(Y_1)$ since if not then at least one instrument must be used twice, so W will not have full column rank:

  $$\rho(W) < K^* + G^* - 1 \Rightarrow \rho(Q_{ZW}) < K^* + G^* - 1$$

  This is the *order condition* for identification in a set of simultaneous equations. When the only identifying information is exclusion restrictions on the variables that enter an equation, then the number of excluded exogs must be greater than or equal to the number of included endogs, minus 1 (the normalized lhs endog), e.g.,

  $$K^{**} \geq G^* - 1$$

- To show that this is in fact a necessary condition consider some arbitrary set of instruments W.

A necessary condition for identification is that

$$\rho\left(plim\frac{1}{n}W'Z\right) = K^* + G^* - 1$$

where

$$Z = \left[\begin{array}{cc} Y_1 & X_1 \end{array}\right]$$

Recall that we've partitioned the model

$$Y\Gamma = XB + E$$

as

$$Y = \left[\begin{array}{ccc} y & Y_1 & Y_2 \end{array}\right]$$

$$X = \left[\begin{array}{cc} X_1 & X_2 \end{array}\right]$$

Given the reduced form

$$Y = X\Pi + V$$

we can write the reduced form using the same partition

$$\left[\begin{array}{ccc} y & Y_1 & Y_2 \end{array}\right] = \left[\begin{array}{cc} X_1 & X_2 \end{array}\right] \left[\begin{array}{ccc} \pi_{11} & \Pi_{12} & \Pi_{13} \\ \pi_{21} & \Pi_{22} & \Pi_{23} \end{array}\right] + \left[\begin{array}{ccc} v & V_1 & V_2 \end{array}\right]$$

so we have

$$Y_1 = X_1\Pi_{12} + X_2\Pi_{22} + V_1$$

so

$$\frac{1}{n}W'Z = \frac{1}{n}W'\left[\begin{array}{cc} X_1\Pi_{12} + X_2\Pi_{22} + V_1 & X_1 \end{array}\right]$$

Because the W 's are uncorrelated with the V_1 's, by assumption, the cross between W and V_1 converges in probability to zero, so

$$plim\frac{1}{n}W'Z = plim\frac{1}{n}W'\left[\begin{array}{cc} X_1\Pi_{12} + X_2\Pi_{22} & X_1 \end{array}\right]$$

Since the far rhs term is formed only of linear combinations of columns of X, the rank of this matrix can never be greater than K, regardless of the choice of instruments. If Z has more than K columns, then it is not of full column rank. When Z has more than K columns we have

$$G^* - 1 + K^* > K$$

or noting that $K^{**} = K - K^*$,

$$G^* - 1 > K^{**}$$

In this case, the limiting matrix is not of full column rank, and the identification condition fails.

Sufficient conditions

Identification essentially requires that the structural parameters be recoverable from the data. This won't be the case, in general, unless the structural model is subject to some restrictions. We've already identified necessary conditions. Turning to sufficient conditions (again, we're only considering identification through zero restricitions on the parameters, for the moment).

The model is

$$\begin{aligned} Y_t'\Gamma &= X_t'B + E_t \\ V(E_t) &= \Sigma \end{aligned}$$

This leads to the reduced form

$$\begin{aligned} Y_t' &= X_t'B\Gamma^{-1} + E_t\Gamma^{-1} \\ &= X_t'\Pi + V_t \\ V(V_t) &= \left(\Gamma^{-1}\right)'\Sigma\Gamma^{-1} \\ &= \Omega \end{aligned}$$

The reduced form parameters are consistently estimable, but none of them are known *a priori,* and there are no restrictions on their values. The problem is that more than one structural form has the same reduced form, so knowledge of the reduced form parameters alone isn't enough to determine the structural parameters. To see this, consider the model

$$\begin{aligned} Y_t'\Gamma F &= X_t'BF + E_tF \\ V(E_tF) &= F'\Sigma F \end{aligned}$$

where F is some arbirary nonsingular $G \times G$ matrix. The rf of this new model is

$$\begin{aligned} Y_t' &= X_t' B F \left(\Gamma F\right)^{-1} + E_t F \left(\Gamma F\right)^{-1} \\ &= X_t' B F F^{-1} \Gamma^{-1} + E_t F F^{-1} \Gamma^{-1} \\ &= X_t' B \Gamma^{-1} + E_t \Gamma^{-1} \\ &= X_t' \Pi + V_t \end{aligned}$$

Likewise, the covariance of the rf of the transformed model is

$$\begin{aligned} V(E_t F \left(\Gamma F\right)^{-1}) &= V(E_t \Gamma^{-1}) \\ &= \Omega \end{aligned}$$

Since the two structural forms lead to the same rf, and the rf is all that is directly estimable, the models are said to be *observationally equivalent.* What we need for identification are restrictions on Γ and B such that the only admissible F is an identity matrix (if all of the equations are to be identified). Take the coefficient matrices as partitioned before:

$$\begin{bmatrix} \Gamma \\ B \end{bmatrix} = \begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \\ \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

The coefficients of the first equation of the transformed model are simply these coefficients multiplied

by the first column of F. This gives

$$\begin{bmatrix} \Gamma \\ B \end{bmatrix} \begin{bmatrix} f_{11} \\ F_2 \end{bmatrix} = \begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \\ \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} f_{11} \\ F_2 \end{bmatrix}$$

For identification of the first equation we need that there be enough restrictions so that the only admissible

$$\begin{bmatrix} f_{11} \\ F_2 \end{bmatrix}$$

be the leading column of an identity matrix, so that

$$\begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \\ \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} f_{11} \\ F_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -\gamma_1 \\ 0 \\ \beta_1 \\ 0 \end{bmatrix}$$

Note that the third and fifth rows are

$$\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix} F_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Supposing that the leading matrix is of full column rank, e.g.,

$$\rho\left(\left[\begin{array}{c}\Gamma_{32}\\ B_{22}\end{array}\right]\right) = cols\left(\left[\begin{array}{c}\Gamma_{32}\\ B_{22}\end{array}\right]\right) = G-1$$

then the only way this can hold, without additional restrictions on the model's parameters, is if F_2 is a vector of zeros. Given that F_2 is a vector of zeros, then the first equation

$$\left[\begin{array}{cc}1 & \Gamma_{12}\end{array}\right]\left[\begin{array}{c}f_{11}\\ F_2\end{array}\right] = 1 \Rightarrow f_{11} = 1$$

Therefore, as long as

$$\rho\left(\left[\begin{array}{c}\Gamma_{32}\\ B_{22}\end{array}\right]\right) = G-1$$

then

$$\left[\begin{array}{c}f_{11}\\ F_2\end{array}\right] = \left[\begin{array}{c}1\\ 0_{G-1}\end{array}\right]$$

The first equation is identified in this case, so the condition is sufficient for identification. It is also necessary, since the condition implies that this submatrix must have at least $G-1$ rows. Since this matrix has

$$G^{**} + K^{**} = G - G^{*} + K^{**}$$

rows, we obtain

$$G - G^{*} + K^{**} \geq G - 1$$

or

$$K^{**} \geq G^* - 1$$

which is the previously derived necessary condition.

The above result is fairly intuitive (draw picture here). The necessary condition ensures that there are enough variables not in the equation of interest to potentially move the other equations, so as to trace out the equation of interest. The sufficient condition ensures that those other equations in fact do move around as the variables change their values. Some points:

- When an equation has $K^{**} = G^* - 1$, is is *exactly identified*, in that omission of an identifiying restriction is not possible without loosing consistency.
- When $K^{**} > G^* - 1$, the equation is *overidentified*, since one could drop a restriction and still retain consistency. Overidentifying restrictions are therefore testable. When an equation is overidentified we have more instruments than are strictly necessary for consistent estimation. Since estimation by IV with more instruments is more efficient asymptotically, one should employ overidentifying restrictions if one is confident that they're true.
- We can repeat this partition for each equation in the system, to see which equations are identified and which aren't.
- These results are valid assuming that the only identifying information comes from knowing which variables appear in which equations, e.g., by exclusion restrictions, and through the use of a normalization. There are other sorts of identifying information that can be used. These include

    1. Cross equation restrictions

2. Additional restrictions on parameters within equations (as in the Klein model discussed below)
3. Restrictions on the covariance matrix of the errors
4. Nonlinearities in variables

- When these sorts of information are available, the above conditions aren't necessary for identification, though they are of course still sufficient.

To give an example of how other information can be used, consider the model

$$Y\Gamma = XB + E$$

where Γ is an upper triangular matrix with 1's on the main diagonal. This is a *triangular system* of equations. In this case, the first equation is

$$y_1 = XB_{\cdot 1} + E_{\cdot 1}$$

Since only exogs appear on the rhs, this equation is identified.

The second equation is

$$y_2 = -\gamma_{21}y_1 + XB_{\cdot 2} + E_{\cdot 2}$$

This equation has $K^{**} = 0$ excluded exogs, and $G^* = 2$ included endogs, so it fails the order (necessary) condition for identification.

- However, suppose that we have the restriction $\Sigma_{21} = 0$, so that the first and second structural

errors are uncorrelated. In this case

$$\mathcal{E}(y_{1t}\varepsilon_{2t}) = \mathcal{E}\left\{(X_t'B_{\cdot 1} + \varepsilon_{1t})\varepsilon_{2t}\right\} = 0$$

so there's no problem of simultaneity. If the entire Σ matrix is diagonal, then following the same logic, all of the equations are identified. This is known as a *fully recursive* model.

10.7 2SLS

When we have no information regarding cross-equation restrictions or the structure of the error covariance matrix, one can estimate the parameters of a single equation of the system without regard to the other equations.

- This isn't always efficient, as we'll see, but it has the advantage that misspecifications in other equations will not affect the consistency of the estimator of the parameters of the equation of interest.
- Also, estimation of the equation won't be affected by identification problems in other equations.

The 2SLS estimator is very simple: it is the GIV estimator, using all of the weakly exogenous variables as instruments. In the first stage, each column of Y_1 is regressed on *all* the weakly exogenous variables in the system, e.g., the entire X matrix. The fitted values are

$$\begin{aligned}
\hat{Y}_1 &= X(X'X)^{-1}X'Y_1 \\
&= P_X Y_1 \\
&= X\hat{\Pi}_1
\end{aligned}$$

Since these fitted values are the projection of Y_1 on the space spanned by X, and since any vector in this space is uncorrelated with ε by assumption, $\hat{Y}_1$ is uncorrelated with ε. Since $\hat{Y}_1$ is simply the reduced-form prediction, it is correlated with Y_1, The only other requirement is that the instruments be linearly independent. This should be the case when the order condition is satisfied, since there are more columns in X_2 than in Y_1 in this case.

The second stage substitutes $\hat{Y}_1$ in place of Y_1, and estimates by OLS. This original model is

$$\begin{aligned} y &= Y_1\gamma_1 + X_1\beta_1 + \varepsilon \\ &= Z\delta + \varepsilon \end{aligned}$$

and the second stage model is

$$y = \hat{Y}_1\gamma_1 + X_1\beta_1 + \varepsilon.$$

Since X_1 is in the space spanned by X, $P_X X_1 = X_1$, so we can write the second stage model as

$$\begin{aligned} y &= P_X Y_1\gamma_1 + P_X X_1\beta_1 + \varepsilon \\ &\equiv P_X Z\delta + \varepsilon \end{aligned}$$

The OLS estimator applied to this model is

$$\hat{\delta} = (Z'P_X Z)^{-1} Z'P_X y$$

which is exactly what we get if we estimate using IV, with the reduced form predictions of the endogs used as instruments. Note that if we define

$$\begin{aligned} \hat{Z} &= P_X Z \\ &= \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix} \end{aligned}$$

so that $\hat{Z}$ are the instruments for Z, then we can write

$$\hat{\delta} = (\hat{Z}'Z)^{-1}\hat{Z}'y$$

- Important note: OLS on the transformed model can be used to calculate the 2SLS estimate of δ, since we see that it's equivalent to IV using a particular set of instruments. However *the OLS covariance formula is not valid.* We need to apply the IV covariance formula already seen above.

Actually, there is also a simplification of the general IV variance formula. Define

$$\begin{aligned} \hat{Z} &= P_X Z \\ &= \begin{bmatrix} \hat{Y} & X \end{bmatrix} \end{aligned}$$

The IV covariance estimator would ordinarily be

$$\hat{V}(\hat{\delta}) = \left(Z'\hat{Z}\right)^{-1}\left(\hat{Z}'\hat{Z}\right)\left(\hat{Z}'Z\right)^{-1}\hat{\sigma}^2_{IV}$$

However, looking at the last term in brackets

$$\hat{Z}'Z = \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix}' \begin{bmatrix} Y_1 & X_1 \end{bmatrix} = \begin{bmatrix} Y_1'(P_X)Y_1 & Y_1'(P_X)X_1 \\ X_1'Y_1 & X_1'X_1 \end{bmatrix}$$

but since P_X is idempotent and since $P_X X = X$, we can write

$$\begin{aligned} \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix}' \begin{bmatrix} Y_1 & X_1 \end{bmatrix} &= \begin{bmatrix} Y_1'P_XP_XY_1 & Y_1'P_XX_1 \\ X_1'P_XY_1 & X_1'X_1 \end{bmatrix} \\ &= \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix}' \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix} \\ &= \hat{Z}'\hat{Z} \end{aligned}$$

Therefore, the second and last term in the variance formula cancel, so the 2SLS varcov estimator simplifies to

$$\hat{V}(\hat{\delta}) = \left(Z'\hat{Z}\right)^{-1} \hat{\sigma}^2_{IV}$$

which, following some algebra similar to the above, can also be written as

$$\hat{V}(\hat{\delta}) = \left(\hat{Z}'\hat{Z}\right)^{-1} \hat{\sigma}^2_{IV} \tag{10.4}$$

Finally, recall that though this is presented in terms of the first equation, it is general since any equation can be placed first.

Properties of 2SLS:

1. Consistent
2. Asymptotically normal

3. Biased when the mean esists (the existence of moments is a technical issue we won't go into here).

4. Asymptotically inefficient, except in special circumstances (more on this later).

10.8 Testing the overidentifying restrictions

The selection of which variables are endogs and which are exogs *is part of the specification of the model*. As such, there is room for error here: one might erroneously classify a variable as exog when it is in fact correlated with the error term. A general test for the specification on the model can be formulated as follows:

The IV estimator can be calculated by applying OLS to the transformed model, so the IV objective function at the minimized value is

$$s(\hat{\beta}_{IV}) = \left(y - X\hat{\beta}_{IV}\right)' P_W \left(y - X\hat{\beta}_{IV}\right),$$

but

$$\begin{aligned}
\hat{\varepsilon}_{IV} &= y - X\hat{\beta}_{IV} \\
&= y - X(X'P_WX)^{-1}X'P_Wy \\
&= \left(I - X(X'P_WX)^{-1}X'P_W\right)y \\
&= \left(I - X(X'P_WX)^{-1}X'P_W\right)(X\beta + \varepsilon) \\
&= A\,(X\beta + \varepsilon)
\end{aligned}$$

where

$$A \equiv I - X(X'P_WX)^{-1}X'P_W$$

so

$$s(\hat{\beta}_{IV}) = (\varepsilon' + \beta'X')\, A'P_WA\,(X\beta + \varepsilon)$$

Moreover, $A'P_WA$ is idempotent, as can be verified by multiplication:

$$\begin{aligned} A'P_WA &= \left(I - P_WX(X'P_WX)^{-1}X'\right) P_W \left(I - X(X'P_WX)^{-1}X'P_W\right) \\ &= \left(P_W - P_WX(X'P_WX)^{-1}X'P_W\right)\left(P_W - P_WX(X'P_WX)^{-1}X'P_W\right) \\ &= \left(I - P_WX(X'P_WX)^{-1}X'\right) P_W. \end{aligned}$$

Furthermore, A is orthogonal to X

$$\begin{aligned} AX &= \left(I - X(X'P_WX)^{-1}X'P_W\right) X \\ &= X - X \\ &= 0 \end{aligned}$$

so

$$s(\hat{\beta}_{IV}) = \varepsilon' A'P_WA\varepsilon$$

Supposing the ε are normally distributed, with variance σ^2, then the random variable

$$\frac{s(\hat{\beta}_{IV})}{\sigma^2} = \frac{\varepsilon' A'P_WA\varepsilon}{\sigma^2}$$

is a quadratic form of a $N(0,1)$ random variable with an idempotent matrix in the middle, so

$$\frac{s(\hat{\beta}_{IV})}{\sigma^2} \sim \chi^2(\rho(A'P_W A))$$

This isn't available, since we need to estimate σ^2. Substituting a consistent estimator,

$$\frac{s(\hat{\beta}_{IV})}{\widehat{\sigma^2}} \overset{a}{\sim} \chi^2(\rho(A'P_W A))$$

- Even if the ε aren't normally distributed, the asymptotic result still holds. The last thing we need to determine is the rank of the idempotent matrix. We have

$$A'P_W A = \left(P_W - P_W X(X'P_W X)^{-1}X'P_W\right)$$

so

$$\begin{aligned}
\rho(A'P_W A) &= Tr\left(P_W - P_W X(X'P_W X)^{-1}X'P_W\right) \\
&= TrP_W - TrX'P_W P_W X(X'P_W X)^{-1} \\
&= TrW(W'W)^{-1}W' - K_X \\
&= TrW'W(W'W)^{-1} - K_X \\
&= K_W - K_X
\end{aligned}$$

where K_W is the number of columns of W and K_X is the number of columns of X. The degrees of freedom of the test is simply the number of overidentifying restrictions: the number of instruments we have beyond the number that is strictly necessary for consistent estimation.

- This test is an overall specification test: the joint null hypothesis is that the model is correctly specified *and* that the W form valid instruments (e.g., that the variables classified as exogs really are uncorrelated with ε. Rejection can mean that either the model $y = Z\delta + \varepsilon$ is misspecified, or that there is correlation between X and ε.

- This is a particular case of the GMM criterion test, which is covered in the second half of the course. See Section 14.8.

- Note that since

$$\hat{\varepsilon}_{IV} = A\varepsilon$$

and

$$s(\hat{\beta}_{IV}) = \varepsilon' A' P_W A \varepsilon$$

we can write

$$\begin{aligned} \frac{s(\hat{\beta}_{IV})}{\widehat{\sigma^2}} &= \frac{(\hat{\varepsilon}'W(W'W)^{-1}W')\,(W(W'W)^{-1}W'\hat{\varepsilon})}{\hat{\varepsilon}'\hat{\varepsilon}/n} \\ &= n(RSS_{\hat{\varepsilon}_{IV}|W}/TSS_{\hat{\varepsilon}_{IV}}) \\ &= nR_u^2 \end{aligned}$$

where R_u^2 is the uncentered R^2 from a regression of the IV residuals on all of the instruments W. This is a convenient way to calculate the test statistic.

On an aside, consider IV estimation of a just-identified model, using the standard notation

$$y = X\beta + \varepsilon$$

and W is the matrix of instruments. If we have exact identification then $cols(W) = cols(X)$, so $W^{'}X$ is a square matrix. The transformed model is

$$P_W y = P_W X\beta + P_W \varepsilon$$

and the fonc are

$$X'P_W(y - X\hat{\beta}_{IV}) = 0$$

The IV estimator is

$$\hat{\beta}_{IV} = (X'P_W X)^{-1} X'P_W y$$

Considering the inverse here

$$\begin{aligned}
(X'P_W X)^{-1} &= \left(X'W(W'W)^{-1}W'X\right)^{-1} \\
&= (W'X)^{-1}\left(X'W(W'W)^{-1}\right)^{-1} \\
&= (W'X)^{-1}(W'W)\,(X'W)^{-1}
\end{aligned}$$

Now multiplying this by $X'P_W y$, we obtain

$$\begin{aligned}
\hat{\beta}_{IV} &= (W'X)^{-1}(W'W)\,(X'W)^{-1}\,X'P_W y \\
&= (W'X)^{-1}(W'W)\,(X'W)^{-1}\,X'W(W'W)^{-1}W'y \\
&= (W'X)^{-1}W'y
\end{aligned}$$

The objective function for the generalized IV estimator is

$$\begin{aligned} s(\hat{\beta}_{IV}) &= \left(y - X\hat{\beta}_{IV}\right)' P_W \left(y - X\hat{\beta}_{IV}\right) \\ &= y'P_W \left(y - X\hat{\beta}_{IV}\right) - \hat{\beta}_{IV}'X'P_W \left(y - X\hat{\beta}_{IV}\right) \\ &= y'P_W \left(y - X\hat{\beta}_{IV}\right) - \hat{\beta}_{IV}'X'P_W y + \hat{\beta}_{IV}'X'P_W X\hat{\beta}_{IV} \\ &= y'P_W \left(y - X\hat{\beta}_{IV}\right) - \hat{\beta}_{IV}' \left(X'P_W y + X'P_W X\hat{\beta}_{IV}\right) \\ &= y'P_W \left(y - X\hat{\beta}_{IV}\right) \end{aligned}$$

by the fonc for generalized IV. However, when we're in the just indentified case, this is

$$\begin{aligned} s(\hat{\beta}_{IV}) &= y'P_W \left(y - X(W'X)^{-1}W'y\right) \\ &= y'P_W \left(I - X(W'X)^{-1}W'\right) y \\ &= y' \left(W(W'W)^{-1}W' - W(W'W)^{-1}W'X(W'X)^{-1}W'\right) y \\ &= 0 \end{aligned}$$

The value of the objective function of the IV estimator is zero in the just identified case. This makes sense, since we've already shown that the objective function after dividing by σ^2 is asymptotically χ^2 with degrees of freedom equal to the number of overidentifying restrictions. In the present case, there are no overidentifying restrictions, so we have a $\chi^2(0)$ rv, which has mean 0 and variance 0, e.g., it's simply 0. This means we're not able to test the identifying restrictions in the case of exact identification.

10.9 System methods of estimation

2SLS is a single equation method of estimation, as noted above. The advantage of a single equation method is that it's unaffected by the other equations of the system, so they don't need to be specified (except for defining what are the exogs, so 2SLS can use the complete set of instruments). The disadvantage of 2SLS is that it's inefficient, in general.

- Recall that overidentification improves efficiency of estimation, since an overidentified equation can use more instruments than are necessary for consistent estimation.
- Secondly, the assumption is that

$$\begin{aligned} Y\Gamma &= XB + E \\ \mathcal{E}(X'E) &= 0_{(K\times G)} \\ vec(E) &\sim N(0, \Psi) \end{aligned}$$

- Since there is no autocorrelation of the E_t 's, and since the columns of E are individually homoscedastic, then

$$\begin{aligned} \Psi &= \begin{bmatrix} \sigma_{11}I_n & \sigma_{12}I_n & \cdots & \sigma_{1G}I_n \\ & \sigma_{22}I_n & & \vdots \\ & & \ddots & \vdots \\ \cdot & & & \sigma_{GG}I_n \end{bmatrix} \\ &= \Sigma \otimes I_n \end{aligned}$$

This means that the structural equations are heteroscedastic and correlated with one another

- In general, ignoring this will lead to inefficient estimation, following the section on GLS. When equations are correlated with one another estimation should account for the correlation in order to obtain efficiency.
- Also, since the equations are correlated, information about one equation is implicitly information about all equations. Therefore, overidentification restrictions in any equation improve efficiency for *all* equations, even the just identified equations.
- Single equation methods can't use these types of information, and are therefore inefficient (in general).

3SLS

Note: It is easier and more practical to treat the 3SLS estimator as a generalized method of moments estimator (see Chapter 14). I no longer teach the following section, but it is retained for its possible historical interest. Another alternative is to use FIML (Subsection 10.9), if you are willing to make distributional assumptions on the errors. This is computationally feasible with modern computers.

Following our above notation, each structural equation can be written as

$$\begin{aligned} y_i &= Y_i\gamma_1 + X_i\beta_1 + \varepsilon_i \\ &= Z_i\delta_i + \varepsilon_i \end{aligned}$$

Grouping the G equations together we get

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & Z_G \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_G \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_G \end{bmatrix}$$

or

$$y = Z\delta + \varepsilon$$

where we already have that

$$\begin{aligned} \mathcal{E}(\varepsilon\varepsilon') &= \Psi \\ &= \Sigma \otimes I_n \end{aligned}$$

The 3SLS estimator is just 2SLS combined with a GLS correction that takes advantage of the structure

of Ψ. Define $\hat{Z}$ as

$$\hat{Z} = \begin{bmatrix} X(X'X)^{-1}X'Z_1 & 0 & \cdots & 0 \\ 0 & X(X'X)^{-1}X'Z_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & X(X'X)^{-1}X'Z_G \end{bmatrix}$$

$$= \begin{bmatrix} \hat{Y}_1 & X_1 & 0 & \cdots & 0 \\ 0 & & \hat{Y}_2 & X_2 & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & & \cdots & 0 & \hat{Y}_G & X_G \end{bmatrix}$$

These instruments are simply the *unrestricted* rf predicitions of the endogs, combined with the exogs. The distinction is that if the model is overidentified, then

$$\Pi = B\Gamma^{-1}$$

may be subject to some zero restrictions, depending on the restrictions on Γ and B, and $\hat{\Pi}$ does not impose these restrictions. Also, note that $\hat{\Pi}$ is calculated using OLS equation by equation, as was discussed in Section 10.3.

The 2SLS estimator would be

$$\hat{\delta} = (\hat{Z}'Z)^{-1}\hat{Z}'y$$

as can be verified by simple multiplication, and noting that the inverse of a block-diagonal matrix is just the matrix with the inverses of the blocks on the main diagonal. This IV estimator still ignores the covariance information. The natural extension is to add the GLS transformation, putting the inverse

of the error covariance into the formula, which gives the 3SLS estimator

$$\begin{aligned}\hat{\delta}_{3SLS} &= \left(\hat{Z}'\left(\Sigma \otimes I_n\right)^{-1} Z\right)^{-1} \hat{Z}'\left(\Sigma \otimes I_n\right)^{-1} y \\ &= \left(\hat{Z}'\left(\Sigma^{-1} \otimes I_n\right) Z\right)^{-1} \hat{Z}'\left(\Sigma^{-1} \otimes I_n\right) y\end{aligned}$$

This estimator requires knowledge of Σ. The solution is to define a feasible estimator using a consistent estimator of Σ. The obvious solution is to use an estimator based on the 2SLS residuals:

$$\hat{\varepsilon}_i = y_i - Z_i \hat{\delta}_{i,2SLS}$$

(IMPORTANT NOTE: this is calculated using Z_i, not $\hat{Z}_i$). Then the element i, j of Σ is estimated by

$$\hat{\sigma}_{ij} = \frac{\hat{\varepsilon}_i' \hat{\varepsilon}_j}{n}$$

Substitute $\hat{\Sigma}$ into the formula above to get the feasible 3SLS estimator.

Analogously to what we did in the case of 2SLS, the asymptotic distribution of the 3SLS estimator can be shown to be

$$\sqrt{n}\left(\hat{\delta}_{3SLS} - \delta\right) \overset{a}{\sim} N\left(0, \lim_{n\to\infty} \mathcal{E}\left\{\left(\frac{\hat{Z}'\left(\Sigma \otimes I_n\right)^{-1} \hat{Z}}{n}\right)^{-1}\right\}\right)$$

A formula for estimating the variance of the 3SLS estimator in finite samples (cancelling out the powers of n) is

$$\hat{V}\left(\hat{\delta}_{3SLS}\right) = \left(\hat{Z}'\left(\hat{\Sigma}^{-1} \otimes I_n\right) \hat{Z}\right)^{-1}$$

- This is analogous to the 2SLS formula in equation (10.4), combined with the GLS correction.
- In the case that all equations are just identified, 3SLS is numerically equivalent to 2SLS. Proving this is easiest if we use a GMM interpretation of 2SLS and 3SLS. GMM is presented in the next econometrics course. For now, take it on faith.

FIML

Full information maximum likelihood is an alternative estimation method. FIML will be asymptotically efficient, since ML estimators based on a given information set are asymptotically efficient w.r.t. all other estimators that use the same information set, and in the case of the full-information ML estimator we use the entire information set. The 2SLS and 3SLS estimators don't require distributional assumptions, while FIML of course does. Our model is, recall

$$\begin{aligned} Y_t'\Gamma &= X_t'B + E_t' \\ E_t &\sim N(0, \Sigma), \forall t \\ \mathcal{E}(E_t E_s') &= 0, t \neq s \end{aligned}$$

The joint normality of E_t means that the density for E_t is the multivariate normal, which is

$$(2\pi)^{-g/2} \left(\det \Sigma^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2} E_t' \Sigma^{-1} E_t\right)$$

The transformation from E_t to Y_t requires the Jacobian

$$|\det \frac{dE_t}{dY_t'}| = |\det \Gamma|$$

so the density for Y_t is

$$(2\pi)^{-G/2}|\det \Gamma| \left(\det \Sigma^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}\left(Y_t'\Gamma - X_t'B\right)\Sigma^{-1}\left(Y_t'\Gamma - X_t'B\right)'\right)$$

Given the assumption of independence over time, the joint log-likelihood function is

$$\ln L(B,\Gamma,\Sigma) = -\frac{nG}{2}\ln(2\pi) + n\ln(|\det \Gamma|) - \frac{n}{2}\ln\det\Sigma^{-1} - \frac{1}{2}\sum_{t=1}^{n}\left(Y_t'\Gamma - X_t'B\right)\Sigma^{-1}\left(Y_t'\Gamma - X_t'B\right)'$$

- This is a nonlinear in the parameters objective function. Maximixation of this can be done using iterative numeric methods. We'll see how to do this in the next section.
- It turns out that the asymptotic distribution of 3SLS and FIML are the same, *assuming normality of the errors*.
- One can calculate the FIML estimator by iterating the 3SLS estimator, thus avoiding the use of a nonlinear optimizer. The steps are

  1. Calculate $\hat{\Gamma}_{3SLS}$ and $\hat{B}_{3SLS}$ as normal.
  2. Calculate $\hat{\Pi} = \hat{B}_{3SLS}\hat{\Gamma}_{3SLS}^{-1}$. This is new, we didn't estimate Π in this way before. This estimator may have some zeros in it. When Greene says iterated 3SLS doesn't lead to FIML, he means this for a procedure that doesn't update $\hat{\Pi}$, but only updates $\hat{\Sigma}$ and $\hat{B}$

and $\hat{\Gamma}$. If you update $\hat{\Pi}$ you *do* converge to FIML.

3. Calculate the instruments $\hat{Y} = X\hat{\Pi}$ and calculate $\hat{\Sigma}$ using $\hat{\Gamma}$ and $\hat{B}$ to get the estimated errors, applying the usual estimator.
4. Apply 3SLS using these new instruments and the estimate of Σ.
5. Repeat steps 2-4 until there is no change in the parameters.

- FIML is fully efficient, since it's an ML estimator that uses all information. This implies that 3SLS is fully efficient *when the errors are normally distributed.* Also, if each equation is just identified and the errors are normal, then 2SLS will be fully efficient, since in this case 2SLS≡3SLS.
- When the errors aren't normally distributed, the likelihood function is of course different than what's written above.

10.10 Example: Klein's Model 1

To give a practical example, consider the following (old-fashioned, but illustrative) macro model (this is the widely known Klein's Model 1)

$$\begin{aligned}
\text{Consumption: } C_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^p + W_t^g) + \varepsilon_{1t} \\
\text{Investment: } I_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} \\
\text{Private Wages: } W_t^p &= \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t} \\
\text{Output: } X_t &= C_t + I_t + G_t \\
\text{Profits: } P_t &= X_t - T_t - W_t^p \\
\text{Capital Stock: } K_t &= K_{t-1} + I_t \\
\begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{pmatrix} &\sim IID \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{pmatrix} \right)
\end{aligned}$$

The other variables are the government wage bill, W_t^g, taxes, T_t, government nonwage spending, G_t,and a time trend, A_t. The endogenous variables are the lhs variables,

$$Y_t' = \begin{bmatrix} C_t & I_t & W_t^p & X_t & P_t & K_t \end{bmatrix}$$

and the predetermined variables are all others:

$$X_t' = \begin{bmatrix} 1 & W_t^g & G_t & T_t & A_t & P_{t-1} & K_{t-1} & X_{t-1} \end{bmatrix}.$$

The model assumes that the errors of the equations are contemporaneously correlated, but nonautocorrelated. The model written as $Y\Gamma = XB + E$ gives

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 \\ -\alpha_3 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & -\gamma_1 & 1 & -1 & 0 \\ -\alpha_1 & -\beta_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} \alpha_0 & \beta_0 & \gamma_0 & 0 & 0 & 0 \\ \alpha_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & \gamma_3 & 0 & 0 & 0 \\ \alpha_2 & \beta_2 & 0 & 0 & 0 & 0 \\ 0 & \beta_3 & 0 & 0 & 0 & 1 \\ 0 & 0 & \gamma_2 & 0 & 0 & 0 \end{bmatrix}$$

To check this identification of the consumption equation, we need to extract Γ_{32} and B_{22}, the submatrices of coefficients of endogs and exogs that *don't* appear in this equation. These are the rows that

have zeros in the first column, and we need to drop the first column. We get

$$\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 & 0 & -1 \\ 0 & -\gamma_1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & \gamma_3 & 0 & 0 & 0 \\ \beta_3 & 0 & 0 & 0 & 1 \\ 0 & \gamma_2 & 0 & 0 & 0 \end{bmatrix}$$

We need to find a set of 5 rows of this matrix gives a full-rank 5×5 matrix. For example, selecting rows 3,4,5,6, and 7 we obtain the matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & \gamma_3 & 0 & 0 & 0 \\ \beta_3 & 0 & 0 & 0 & 1 \end{bmatrix}$$

This matrix is of full rank, so the sufficient condition for identification is met. Counting included endogs, $G^* = 3$, and counting excluded exogs, $K^{**} = 5$, so

$$\begin{aligned} K^{**} - L &= G^* - 1 \\ 5 - L &= 3 - 1 \\ L &= 3 \end{aligned}$$

- The equation is over-identified by three restrictions, according to the counting rules, which are correct when the only identifying information are the exclusion restrictions. However, there is additional information in this case. Both W_t^p and W_t^g enter the consumption equation, and their coefficients are restricted to be the same. For this reason the consumption equation is in fact overidentified by four restrictions.

The Octave program Simeq/Klein2SLS.m performs 2SLS estimation for the 3 equations of Klein's model 1, assuming nonautocorrelated errors, so that lagged endogenous variables can be used as instruments. The results are:

```
CONSUMPTION EQUATION


*********************************************************
2SLS estimation results
Observations 21
R-squared 0.976711
Sigma-squared 1.044059
```

	estimate	st.err.	t-stat.	p-value
Constant	16.555	1.321	12.534	0.000
Profits	0.017	0.118	0.147	0.885
Lagged Profits	0.216	0.107	2.016	0.060
Wages	0.810	0.040	20.129	0.000

****

INVESTMENT EQUATION

****

2SLS estimation results

Observations 21

R-squared 0.884884

Sigma-squared 1.383184

	estimate	st.err.	t-stat.	p-value
Constant	20.278	7.543	2.688	0.016
Profits	0.150	0.173	0.867	0.398
Lagged Profits	0.616	0.163	3.784	0.001
Lagged Capital	-0.158	0.036	-4.368	0.000

****

WAGES EQUATION

```
**********************************************************
2SLS estimation results
Observations 21
R-squared 0.987414
Sigma-squared 0.476427

                 estimate     st.err.     t-stat.     p-value
Constant            1.500       1.148       1.307       0.209
Output              0.439       0.036      12.316       0.000
Lagged Output       0.147       0.039       3.777       0.002
Trend               0.130       0.029       4.475       0.000

**********************************************************
```

The above results are not valid (specifically, they are inconsistent) if the errors are autocorrelated, since lagged endogenous variables will not be valid instruments in that case. You might consider eliminating the lagged endogenous variables as instruments, and re-estimating by 2SLS, to obtain consistent parameter estimates in this more complex case. Standard errors will still be estimated inconsistently, unless use a Newey-West type covariance estimator. Food for thought...

Chapter 11

Numeric optimization methods

Readings: Hamilton, ch. 5, section 7 (pp. 133-139)*; Gourieroux and Monfort, Vol. 1, ch. 13, pp. 443-60*; Goffe, et. al. (1994).

The next chapter introduces extremum estimators, which are minimizers or maximizers of objective functions. If we're going to be applying extremum estimators, we'll need to know how to find an extremum. This section gives a very brief introduction to what is a large literature on numeric optimization methods. We'll consider a few well-known techniques, and one fairly new technique that may allow one to solve difficult problems. The main objective is to become familiar with the issues, and to learn how to use the BFGS algorithm at the practical level.

The general problem we consider is how to find the maximizing element $\hat{\theta}$ (a K -vector) of a function $s(\theta)$. This function may not be continuous, and it may not be differentiable. Even if it is twice continuously differentiable, it may not be globally concave, so local maxima, minima and

saddlepoints may all exist. Supposing $s(\theta)$ were a quadratic function of θ, e.g.,

$$s(\theta) = a + b'\theta + \frac{1}{2}\theta' C\theta,$$

the first order conditions would be linear:

$$D_\theta s(\theta) = b + C\theta$$

so the maximizing (minimizing) element would be $\hat{\theta} = -C^{-1}b$. This is the sort of problem we have with linear models estimated by OLS. It's also the case for feasible GLS, since conditional on the estimate of the varcov matrix, we have a quadratic objective function in the remaining parameters.

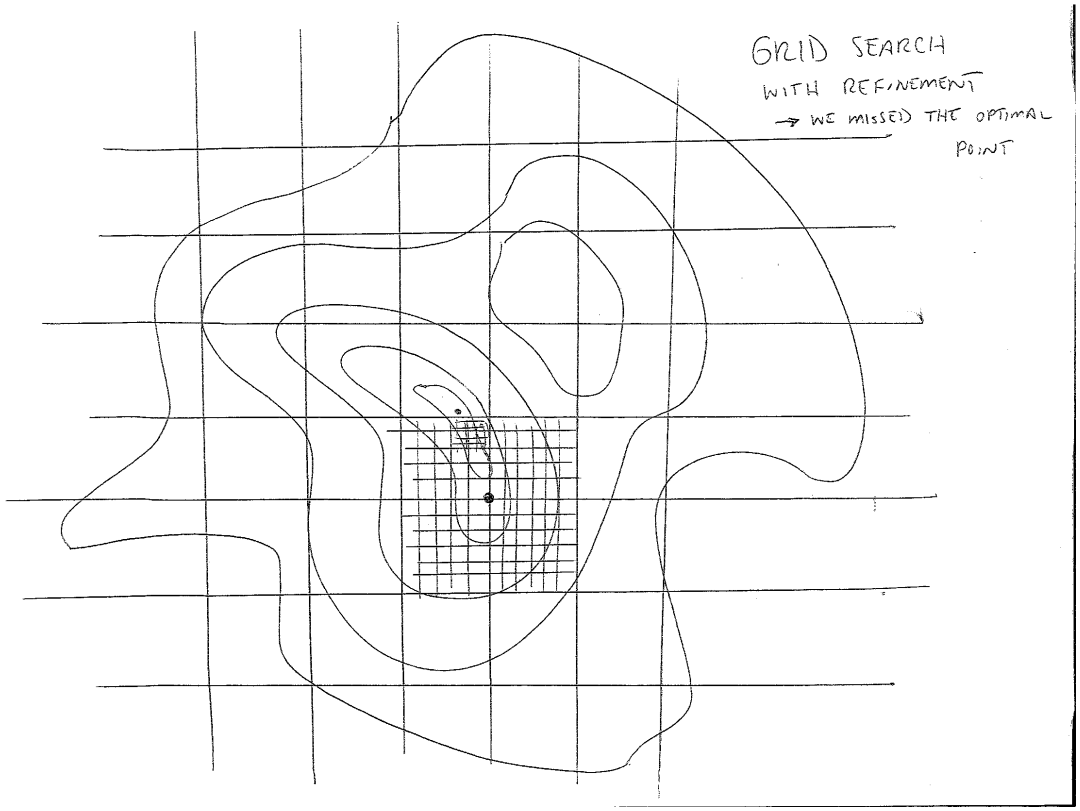
More general problems will not have linear f.o.c., and we will not be able to solve for the maximizer analytically. This is when we need a numeric optimization method.

11.1 Search

The idea is to create a grid over the parameter space and evaluate the function at each point on the grid. Select the best point. Then refine the grid in the neighborhood of the best point, and continue until the accuracy is "good enough". See Figure 11.1. One has to be careful that the grid is fine enough in relationship to the irregularity of the function to ensure that sharp peaks are not missed entirely.

To check q values in each dimension of a K dimensional parameter space, we need to check q^K points. For example, if $q = 100$ and $K = 10$, there would be 100^{10} points to check. If 1000 points can be checked in a second, it would take 3.171×10^9 years to perform the calculations, which is approximately 2/3 the age of the earth. The search method is a very reasonable choice if K is small,

Figure 11.1: Search method



but it quickly becomes infeasible if K is moderate or large.

11.2 Derivative-based methods

Introduction

Derivative-based methods are defined by

1. the method for choosing the initial value, θ^1
2. the iteration method for choosing θ^{k+1} given θ^k (based upon derivatives)
3. the stopping criterion.

The iteration method can be broken into two problems: choosing the stepsize a^k (a scalar) and choosing the direction of movement, d^k, which is of the same dimension of θ, so that

$$\theta^{(k+1)} = \theta^{(k)} + a^k d^k.$$

A locally increasing direction of search d is a direction such that

$$\frac{\partial s(\theta + ad)}{\partial a} > 0$$

for a positive but small. That is, if we go in direction d, we will improve on the objective function, at least if we don't go too far in that direction.

- As long as the gradient at θ is not zero there exist increasing directions, and they can all be represented as $Q^k g(\theta^k)$ where Q^k is a symmetric pd matrix and $g(\theta) = D_\theta s(\theta)$ is the gradient at θ. To see this, take a T.S. expansion around $a^0 = 0$

$$\begin{aligned} s(\theta + ad) &= s(\theta + 0d) + (a - 0)\, g(\theta + 0d)'d + o(1) \\ &= s(\theta) + ag(\theta)'d + o(1) \end{aligned}$$

  For small enough a the $o(1)$ term can be ignored. If d is to be an increasing direction, we need $g(\theta)'d > 0$. Defining $d = Qg(\theta)$, where Q is positive definite, we guarantee that

$$g(\theta)'d = g(\theta)'Qg(\theta) > 0$$

  unless $g(\theta) = 0$. Every increasing direction can be represented in this way (p.d. matrices are those such that the angle between g and $Qg(\theta)$ is less that 90 degrees). See Figure 11.2.

- With this, the iteration rule becomes

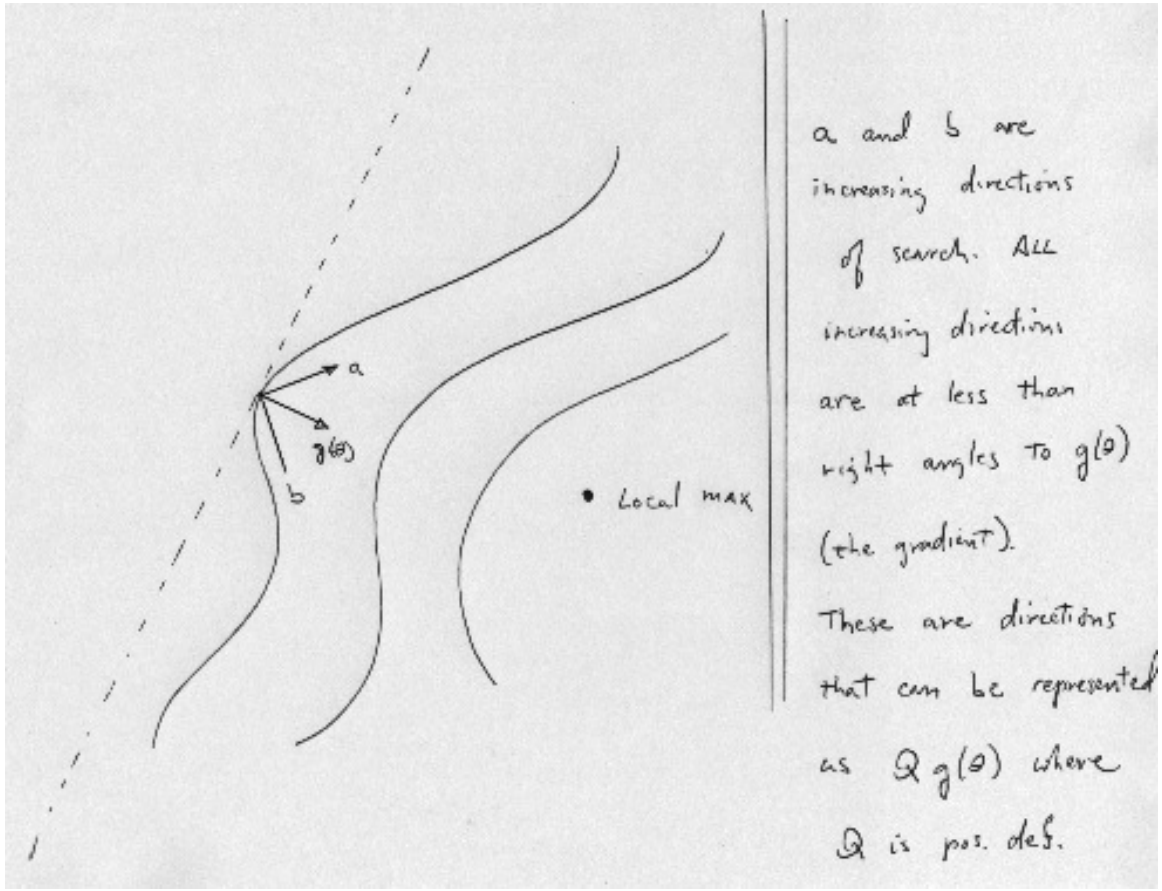
$$\theta^{(k+1)} = \theta^{(k)} + a^k Q^k g(\theta^k)$$

and we keep going until the gradient becomes zero, so that there is no increasing direction. The problem is how to choose a and Q.

- **Conditional on** Q, choosing a is fairly straightforward. A simple line search is an attractive possibility, since a is a scalar.

- The remaining problem is how to choose Q.

Figure 11.2: Increasing directions of search



- Note also that this gives no guarantees to find a global maximum.

Steepest descent

Steepest descent (ascent if we're maximizing) just sets Q to and identity matrix, since the gradient provides the direction of maximum rate of change of the objective function.

- Advantages: fast - doesn't require anything more than first derivatives.
- Disadvantages: This doesn't always work too well however: see the Rosenbrock, or "banana" function: http://en.wikipedia.org/wiki/Rosenbrock_function.

Newton' s method

Newton's method uses information about the slope and curvature of the objective function to determine which direction and how far to move from an initial point. Supposing we're trying to maximize $s_n(\theta)$. Take a second order Taylor's series approximation of $s_n(\theta)$ about θ^k (an initial guess).

$$s_n(\theta) \approx s_n(\theta^k) + g(\theta^k)' \left(\theta - \theta^k\right) + 1/2 \left(\theta - \theta^k\right)' H(\theta^k) \left(\theta - \theta^k\right)$$

To attempt to maximize $s_n(\theta)$, we can maximize the portion of the right-hand side that depends on θ, *i.e.*, we can maximize

$$\tilde{s}(\theta) = g(\theta^k)'\theta + 1/2 \left(\theta - \theta^k\right)' H(\theta^k) \left(\theta - \theta^k\right)$$

with respect to θ. This is a much easier problem, since it is a quadratic function in θ, so it has linear first order conditions. These are

$$D_\theta \tilde{s}(\theta) = g(\theta^k) + H(\theta^k)\left(\theta - \theta^k\right)$$

So the solution for the next round estimate is

$$\theta^{k+1} = \theta^k - H(\theta^k)^{-1} g(\theta^k)$$

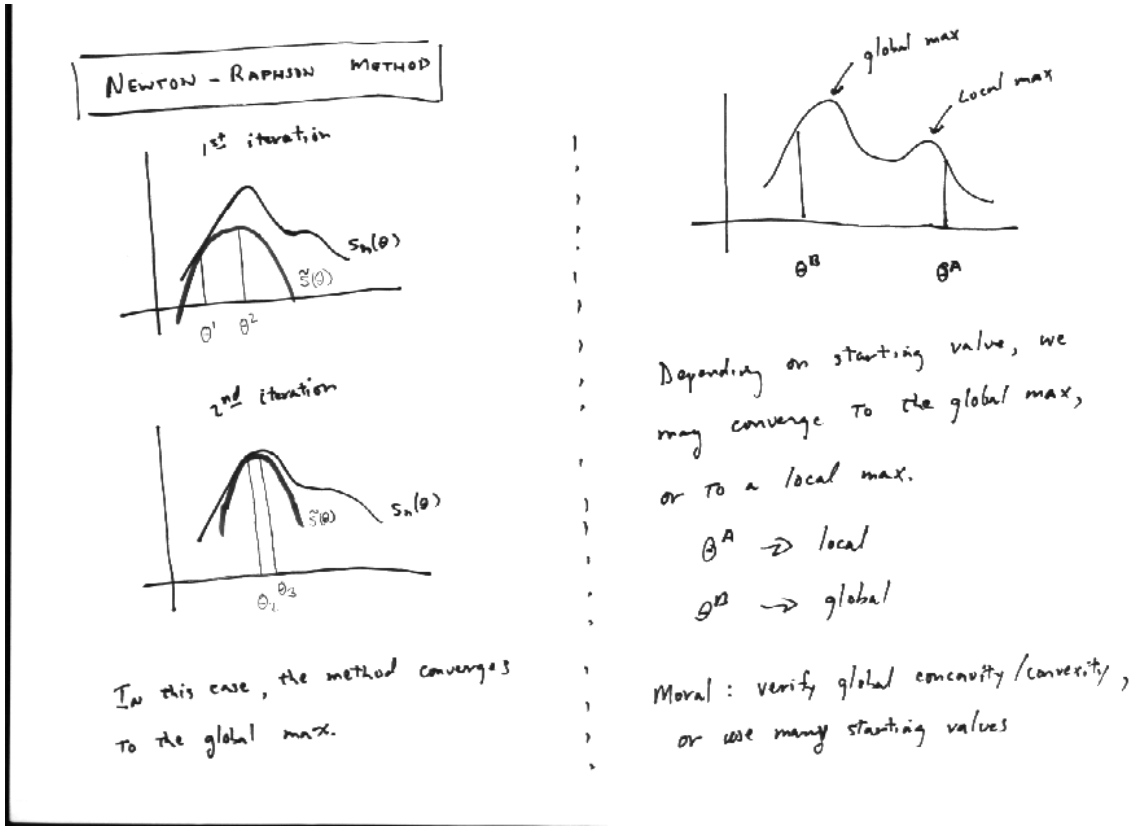
See http://en.wikipedia.org/wiki/Newton%27s_method_in_optimization for more information. This is illustrated in Figure 11.3.

However, it's good to include a stepsize, since the approximation to $s_n(\theta)$ may be bad far away from the maximizer $\hat{\theta}$, so the actual iteration formula is

$$\theta^{k+1} = \theta^k - a^k H(\theta^k)^{-1} g(\theta^k)$$

- A potential problem is that the Hessian may not be negative definite when we're far from the maximizing point. So $-H(\theta^k)^{-1}$ may not be positive definite, and $-H(\theta^k)^{-1} g(\theta^k)$ may not define an increasing direction of search. This can happen when the objective function has flat regions, in which case the Hessian matrix is very ill-conditioned (e.g., is nearly singular), or when we're in the vicinity of a local minimum, $H(\theta^k)$ is positive definite, and our direction is a *decreasing* direction of search. Matrix inverses by computers are subject to large errors when the matrix is ill-conditioned. Also, we certainly don't want to go in the direction of a minimum when we're maximizing. To solve this problem, *Quasi-Newton* methods simply add a positive definite component to $H(\theta)$ to ensure that the resulting matrix is positive definite, *e.g.*, $Q = -H(\theta) + b\mathbf{I}$, where b is chosen large enough so that Q is well-conditioned and positive definite. This has the benefit that improvement in the objective function is guaranteed. See

Figure 11.3: Newton iteration



http://en.wikipedia.org/wiki/Quasi-Newton_method.

- Another variation of quasi-Newton methods is to approximate the Hessian by using successive gradient evaluations. This avoids actual calculation of the Hessian, which is an order of magnitude (in the dimension of the parameter vector) more costly than calculation of the gradient. They can be done to ensure that the approximation is p.d. DFP and BFGS are two well-known examples.

- show bfgsmin_example.m to optimize Rosenbrock function

Stopping criteria

The last thing we need is to decide when to stop. A digital computer is subject to limited machine precision and round-off errors. For these reasons, it is unreasonable to hope that a program can **exactly** find the point that maximizes a function. We need to define acceptable tolerances. Some stopping criteria are:

- Negligable change in parameters:
$$|\theta_j^k - \theta_j^{k-1}| < \varepsilon_1, \forall j$$

- Negligable relative change:
$$|\frac{\theta_j^k - \theta_j^{k-1}}{\theta_j^{k-1}}| < \varepsilon_2, \forall j$$

- Negligable change of function:
$$|s(\theta^k) - s(\theta^{k-1})| < \varepsilon_3$$

- Gradient negligibly different from zero:

$$|g_j(\theta^k)| < \varepsilon_4, \forall j$$

- Or, even better, check all of these.
- Also, if we're maximizing, it's good to check that the last round (real, not approximate) Hessian is negative definite.

Starting values

The Newton-Raphson and related algorithms work well if the objective function is concave (when maximizing), but not so well if there are convex regions and local minima or multiple local maxima. The algorithm may converge to a local minimum or to a local maximum that is not optimal. The algorithm may also have difficulties converging at all.

- The usual way to "ensure" that a global maximum has been found is to use many different starting values, and choose the solution that returns the highest objective function value. **THIS IS IMPORTANT in practice.** More on this later.

Calculating derivatives

The Newton-Raphson algorithm requires first and second derivatives. It is often difficult to calculate derivatives (especially the Hessian) analytically if the function $s_n(\cdot)$ is complicated. Possible solutions are to calculate derivatives numerically, or to use programs such as MuPAD or Mathematica to calculate analytic derivatives. For example, Figure 11.4 shows Sage [1] calculating a couple of derivatives.

[1] Sage is free software that has both symbolic and numeric computational capabilities. See http://www.sagemath.org/

The KAIST Sage cell server will let you try Sage online, its address is http://aleph.sagemath.org/.

- Numeric derivatives are less accurate than analytic derivatives, and are usually more costly to evaluate. Both factors usually cause optimization programs to be less successful when numeric derivatives are used.

- One advantage of numeric derivatives is that you don't have to worry about having made an error in calculating the analytic derivative. When programming analytic derivatives it's a good idea to check that they are correct by using numeric derivatives. This is a lesson I learned the hard way when writing my thesis.

- Numeric second derivatives are much more accurate if the data are scaled so that the elements of the gradient are of the same order of magnitude. Example: if the model is $y_t = h(\alpha x_t + \beta z_t) + \varepsilon_t$, and estimation is by NLS, suppose that $D_\alpha s_n(\cdot) = 1000$ and $D_\beta s_n(\cdot) = 0.001$. One could define $\alpha^* = \alpha/1000$; $x_t^* = 1000 x_t$;$\beta^* = 1000\beta$; $z_t^* = z_t/1000$. In this case, the gradients $D_{\alpha^*} s_n(\cdot)$ and $D_\beta s_n(\cdot)$ will both be 1.

  In general, estimation programs always work better if data is scaled in this way, since roundoff errors are less likely to become important. *This is important in practice.*

- There are algorithms (such as BFGS and DFP) that use the sequential gradient evaluations to build up an approximation to the Hessian. The iterations are faster because the actual Hessian isn't calculated, but more iterations usually are required for convergence. Versions of BFGS are probably the most widely used optimizers in econometrics.

- Switching between algorithms during iterations is sometimes useful.

Figure 11.4: Using Sage to get analytic derivatives

test

```
f(x,y) = x*y + sin(x) + cos(y)
derivative(f,x)
```

$(x, y) \mapsto y + \cos(x)$

```
f(x,b) = log(1/(1+exp(-x*b)))
derivative(f,b)
```

$(x, b) \mapsto \frac{xe^{(-bx)}}{(e^{(-bx)}+1)}$

11.3 Simulated Annealing

Simulated annealing is an algorithm which can find an optimum in the presence of nonconcavities, discontinuities and multiple local minima/maxima. Basically, the algorithm randomly selects evaluation points, accepts all points that yield an increase in the objective function, but also accepts some points that decrease the objective function. This allows the algorithm to escape from local minima. As more and more points are tried, periodically the algorithm focuses on the best point so far, and reduces the range over which random points are generated. Also, the probability that a negative move is accepted reduces. The algorithm relies on many evaluations, as in the search method, but focuses in on promising areas, which reduces function evaluations with respect to the search method. It does not require derivatives to be evaluated. I have a program to do this if you're interested.

11.4 A practical example: Maximum likelihood estimation using count data: The MEPS data and the Poisson model

To show optimazation methods in practice, using real economic data, this section presents maximum likelihood estimation results for a particular model using real data. The focus at present is simply on numeric optimization. Later, after studying maximum likelihood estimation, this section can be read again.

Demand for health care is usually thought of a a derived demand: health care is an input to a home production function that produces health, and health is an argument of the utility function. Grossman (1972), for example, models health as a capital stock that is subject to depreciation (e.g., the effects

of ageing). Health care visits restore the stock. Under the home production framework, individuals decide when to make health care visits to maintain their health stock, or to deal with negative shocks to the stock in the form of accidents or illnesses. As such, individual demand will be a function of the parameters of the individuals' utility functions.

The MEPS health data file , `meps1996.data,` contains 4564 observations on six measures of health care usage. The data is from the 1996 Medical Expenditure Panel Survey (MEPS). You can get more information at `http://www.meps.ahrq.gov/`. The six measures of use are are office-based visits (OBDV), outpatient visits (OPV), inpatient visits (IPV), emergency room visits (ERV), dental visits (VDV), and number of prescription drugs taken (PRESCR). These form columns 1 - 6 of `meps1996.data`. The conditioning variables are public insurance (PUBLIC), private insurance (PRIV), sex (SEX), age (AGE), years of education (EDUC), and income (INCOME). These form columns 7 - 12 of the file, in the order given here. PRIV and PUBLIC are 0/1 binary variables, where a 1 indicates that the person has access to public or private insurance coverage. SEX is also 0/1, where 1 indicates that the person is female. This data will be used in examples fairly extensively in what follows.

The program ExploreMEPS.m shows how the data may be read in, and gives some descriptive information about variables, which follows:

All of the measures of use are count data, which means that they take on the values $0, 1, 2, \ldots$. It might be reasonable to try to use this information by specifying the density as a count data density. One of the simplest count data densities is the Poisson density, which is

$$f_Y(y) = \frac{\exp(-\lambda)\lambda^y}{y!}.$$

For this density, $E(Y) = V(Y) = \lambda$. The Poisson average log-likelihood function is

$$s_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(-\lambda_i + y_i \ln \lambda_i - \ln y_i!\right)$$

We will parameterize the model as

$$\begin{aligned} \lambda_i &= \exp(\mathbf{x}_i'\beta) \\ \mathbf{x}_i &= [1 \; PUBLIC \; PRIV \; SEX \; AGE \; EDUC \; INC]' \end{aligned} \tag{11.1}$$

This ensures that the mean is positive, as is required for the Poisson model, and now the mean (and the variance) depend upon explanatory variables. Note that for this parameterization

$$\frac{\partial \lambda}{\partial x_j} = \lambda \beta_j$$

so

$$\beta_j x_j = \frac{\partial \lambda}{\partial x_j}\frac{x_j}{\lambda} = \eta_{x_j}^{\lambda},$$

the elasticity of the conditional mean of y with respect to the j^{th} conditioning variable.

The program EstimatePoisson.m estimates a Poisson model using the full data set. The results of the estimation, using OBDV as the dependent variable are here:

```
MPITB extensions found

OBDV
```

```
******************************************************
Poisson model, MEPS 1996 full data set

MLE Estimation Results
BFGS convergence: Normal convergence

Average Log-L: -3.671090
Observations: 4564

                estimate     st. err      t-stat     p-value
constant          -0.791       0.149      -5.290       0.000
pub. ins.          0.848       0.076      11.093       0.000
priv. ins.         0.294       0.071       4.137       0.000
sex                0.487       0.055       8.797       0.000
age                0.024       0.002      11.471       0.000
edu                0.029       0.010       3.061       0.002
inc               -0.000       0.000      -0.978       0.328

Information Criteria
CAIC : 33575.6881      Avg. CAIC:   7.3566
 BIC : 33568.6881       Avg. BIC:   7.3551
```

```
 AIC : 33523.7064       Avg. AIC:   7.3452
*******************************************************
```

11.5 Numeric optimization: pitfalls

In this section we'll examine two common problems that can be encountered when doing numeric optimization of nonlinear models, and some solutions.

Poor scaling of the data

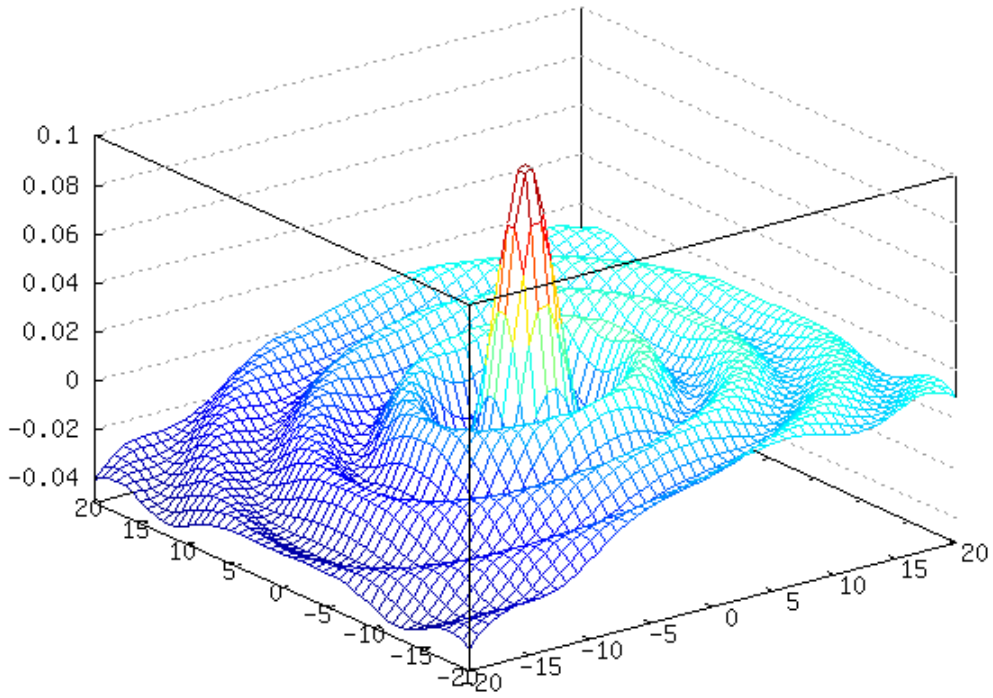
When the data is scaled so that the magnitudes of the first and second derivatives are of different orders, problems can easily result. If we uncomment the appropriate line in EstimatePoisson.m, the data will not be scaled, and the estimation program will have difficulty converging (it seems to take an infinite amount of time). With unscaled data, the elements of the score vector have very different magnitudes at the initial value of θ (all zeros). To see this run CheckScore.m. With unscaled data, one element of the gradient is very large, and the maximum and minimum elements are 5 orders of magnitude apart. This causes convergence problems due to serious numerical inaccuracy when doing inversions to calculate the BFGS direction of search. With scaled data, none of the elements of the gradient are very large, and the maximum difference in orders of magnitude is 3. Convergence is quick.

Figure 11.5: Mountains with low fog

Multiple optima

Multiple optima (one global, others local) can complicate life, since we have limited means of determining if there is a higher maximum than the one we're at. Think of climbing a mountain in an unknown range, in a very foggy place. A nice picture is Figure 11.5, but try to imagine the scene if the clouds were 2000m thicker. A representation is Figure 11.6). You can go up until there's nowhere else to go up, but since you're in the fog you don't know if the true summit is across the gap that's at your feet. Do you claim victory and go home, or do you trudge down the gap and explore the other side?

Figure 11.6: A foggy mountain



The best way to avoid stopping at a local maximum is to use many starting values, for example on a grid, or randomly generated. Or perhaps one might have priors about possible values for the parameters (*e.g.,* from previous studies of similar data).

Let's try to find the true minimizer of minus 1 times the foggy mountain function (since the algorithms are set up to minimize). From the picture, you can see it's close to $(0, 0)$, but let's pretend there is fog, and that we don't know that. The program FoggyMountain.m shows that poor start values can lead to problems. It uses SA, which finds the true global minimum, and it shows that BFGS using a battery of random start values can also find the global minimum help. The output of one run is here:

```
MPITB extensions found

=========================================================
BFGSMIN final results

Used numeric gradient

---------------------------------------------------------
STRONG CONVERGENCE
Function conv 1  Param conv 1  Gradient conv 1
---------------------------------------------------------
Objective function value -0.0130329
Stepsize 0.102833
43 iterations
```

```
------------------------------------------------

 param    gradient  change
 15.9999  -0.0000   0.0000
-28.8119   0.0000   0.0000
The result with poor start values
ans =

   16.000  -28.812


================================================
SAMIN final results
NORMAL CONVERGENCE

Func. tol. 1.000000e-10 Param. tol. 1.000000e-03
Obj. fn. value -0.100023

   parameter     search width
            0.037419            0.000018
           -0.000000            0.000051
================================================
Now try a battery of random start values and
```

```
a short BFGS on each, then iterate to convergence
The result using 20 randoms start values
ans =

   3.7417e-02   2.7628e-07

The true maximizer is near (0.037,0)
```

In that run, the single BFGS run with bad start values converged to a point far from the true minimizer, which simulated annealing and BFGS using a battery of random start values both found the true maximizer. Using a battery of random start values, we managed to find the global max. The moral of the story is to be cautious and don't publish your results too quickly.

11.6 Exercises

1. In octave, type "`help bfgsmin_example`", to find out the location of the file. Edit the file to examine it and learn how to call `bfgsmin`. Run it, and examine the output.

2. In octave, type "`help samin_example`", to find out the location of the file. Edit the file to examine it and learn how to call `samin`. Run it, and examine the output.

3. Numerically minimize the function $\sin(x) + 0.01\,(x - a)^2$, setting $a = 0$, using the software of your choice. Plot the function over the interval $(-2\pi, 2\pi)$. Does the software find the global minimum? Does this depend on the starting value you use? Outline a strategy that would allow you to find the minimum reliably, when a can take on any given value in the interval $(-\pi, \pi)$.

4. Numerically compute the OLS estimator of the Nerlove model by using an interative minimization algorithm to minimize the sum of squared residuals. Verify that the results coincide with those given in subsection 3.8. The important part of this problem is to learn how to minimize a function that depends on both parameters and data. Try to write your function so that it is easy to use it with an arbitrary data set.

Chapter 12

Asymptotic properties of extremum estimators

Readings: Hayashi (2000), Ch. 7; Gourieroux and Monfort (1995), Vol. 2, Ch. 24; Amemiya, Ch. 4 section 4.1; Davidson and MacKinnon, pp. 591-96; Gallant, Ch. 3; Newey and McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, Ch. 36.

12.1 Extremum estimators

We'll begin with study of *extremum estimators* in general. Let $\mathbf{Z}_n = \{z_1, z_2, ..., z_n\}$ be the available data, arranged in a $n \times p$ matrix, based on a sample of size n (there are p variables). Our paradigm is that data are generated as a draw from the joint density $f_{Z_n}(z)$. This density may not be known, but it exists in principle. The draw from the density may be thought of as the outcome of a random ex-

periment that is characterized by the probability space $\{\Omega, \mathcal{F}, P\}$. When the experiment is performed, $\omega \in \Omega$ is the result, and $\mathbf{Z}_n(\omega) = \{Z_1(\omega), Z_2(\omega), ..., Z_n(\omega)\} = \{z_1, z_2, ..., z_n\}$ is the realized data. The probability space is rich enough to allow us to consider events defined in terms of an infinite sequence of data $\mathbf{Z} = \{z_1, z_2, ..., \}$.

Definition 25. [Extremum estimator] An extremum estimator $\hat{\theta}$ is the optimizing element of an objective function $s_n(\mathbf{Z}_n, \theta)$ over a set $\overline{\Theta}$.

Because the data $\mathbf{Z}_n(\omega)$ depends on ω, we can emphasize this by writing $s_n(\omega, \theta)$. I'll be loose with notation and interchange when convenient.

Example 26. OLS. Let the d.g.p. be $y_t = \mathbf{x}_t'\theta^0 + \varepsilon_t$, $t = 1, 2, ..., n$, $\theta^0 \in \Theta$. Stacking observations vertically, $\mathbf{y}_n = \mathbf{X}_n\theta^0 + \varepsilon_n$, where $\mathbf{X}_n = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}'$. Let $\mathbf{Z}_n = [\mathbf{y}_n \, \mathbf{X}_n]$. The least squares estimator is defined as

$$\hat{\theta} \equiv \arg\min_{\Theta} s_n(\mathbf{Z}_n, \theta)$$

where

$$s_n(\mathbf{Z}_n, \theta) = 1/n \sum_{t=1}^{n} (y_t - \mathbf{x}_t'\theta)^2$$

As you already know, $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

.

Example 27. Maximum likelihood. Suppose that the continuous random variables $Y_t \sim IIN(\theta^0, \sigma_0^2)$, $t = 1, 2, ..., n$. If ϵ is a standard normal random variable, its density is

$$f_\epsilon(z; \theta) = (2\pi)^{-1/2} \exp\left(-\frac{z^2}{2}\right).$$

We have that $\epsilon_t = (Y_t - \theta_0)/\sigma_0$ is standard normal, and the Jacobian $|\partial\epsilon_t/\partial y_t| = 1/\sigma_0$. Thus, doing a change of variable, the density of a single observation on Y is

$$f_Y(y_t;\theta,\sigma) = (2\pi)^{-1/2}(1/\sigma)\exp\left(-\frac{1}{2}\left(\frac{y_t-\theta}{\sigma}\right)^2\right).$$

The maximum likelihood estimator is maximizes the joint density of the sample. Because the data are i.i.d., the joint density of the sample $\{y_1, y_2, ..., y_n\}$ is the product of the densities of each observation, and the ML estimator is

$$\hat{\theta} \equiv \arg\max_{\Theta}\ \mathcal{L}_n(\theta) = \prod_{t=1}^{n}(2\pi)^{-1/2}(1/\sigma)\exp\left(-\frac{(y_t-\theta)^2}{2}\right)$$

Because the natural logarithm is strictly increasing on $(0,\infty)$, maximization of the average logarithmic likelihood function is achieved at the same $\hat{\theta}$ as for the likelihood function. So, the ML estimator $\hat{\theta} \equiv \arg\max_{\Theta} s_n(\theta)$ where

$$s_n(\theta) = (1/n)\ln\mathcal{L}_n(\theta) = -\ln\sqrt{2\pi} - \log\sigma - (1/n)\sum_{t=1}^{n}\frac{(y_t-\theta)^2}{2}$$

Solution of the f.o.c. leads to the familiar result that $\hat{\theta} = \bar{\mathbf{y}}$. We'll come back to this in more detail later.

Example 28. Bayesian estimator

Bayesian point estimators such as the posterior mode, median or mean can be expressed as extremum estimators. For example, the posterior mean $E(\theta|Z_n)$ is the minimizer (with respect to ζ) of

the function

$$s_n(\zeta) = \int_\Theta (\theta - \zeta)^2 f(Z_n; \theta)\pi(\theta)/f(Z_n)d\theta$$

where $f(Z_n; \theta)$ is the likelihood function, $\pi(\theta)$ is a prior density, and $f(Z_n)$ is the marginal likelihood of the data. These concepts are explained later, for now the point is that Bayesian estimators can be thought of as extremum estimators, and the theory for extremum estimators will apply.

Note that the objective function $s_n(\mathbf{Z}_n, \theta)$ is a random function, because it depends on $\mathbf{Z}_n(\omega) = \{Z_1(\omega), Z_2(\omega), ..., Z_n(\omega)\} = \{z_1, z_2, ..., z_n\}$. We need to consider what happens as different outcomes $\omega \in \Omega$ occur. These different outcomes lead to different data being generated, and the different data causes the objective function to change. Note, however, that for a fixed $\omega \in \Omega$, the data $\mathbf{Z}_n(\omega) = \{Z_1(\omega), Z_2(\omega), ..., Z_n(\omega)\} = \{z_1, z_2, ..., z_n\}$ are a fixed realization, and the objective function $s_n(\mathbf{Z}_n, \theta)$ becomes a non-random function of θ. When actually computing an extremum estimator, we treat the data as fixed, and employ algorithms for optimization of nonstochastic functions. When analyzing the properties of an extremum estimator, we need to investigate what happens throughout Ω: we do not focus only on the ω that generated the observed data. This is because we would like to find estimators that work well on average for any data set that can result from $\omega \in \Omega$.

We'll often write the objective function suppressing the dependence on $\mathbf{Z}_n$, as $s_n(\omega, \theta)$ or simply $s_n(\theta)$, depending on context. The first of these emphasizes the fact that the objective function is random, and the second is more compact. However, the data is still in there, and because the data is randomly sampled, the objective function is random, too.

12.2 Existence

If $s_n(\theta)$ is continuous in θ and $\overline{\Theta}$ is compact, then a maximizer exists, by the Weierstrass maximum theorem (Debreu, 1959). In some cases of interest, $s_n(\theta)$ may not be continuous. Nevertheless, it may still converge to a continous function, in which case existence will not be a problem, at least asymptotically. Henceforth in this course, we assume that $s_n(\theta)$ is continuous.

12.3 Consistency

The following theorem is patterned on a proof in Gallant (1987) (the article, ref. later), which we'll see in its original form later in the course. It is interesting to compare the following proof with Amemiya's Theorem 4.1.1, which is done in terms of convergence in probability.

Theorem 29. [Consistency of e.e.] *Suppose that $\hat{\theta}_n$ is obtained by maximizing $s_n(\theta)$ over $\overline{\Theta}$.*

Assume

(a) Compactness: The parameter space Θ is an open bounded subset of Euclidean space $\Re^K$. So the closure of Θ, $\overline{\Theta}$, is compact.

(b) Uniform Convergence: There is a nonstochastic function $s_\infty(\theta)$ that is continuous in θ on $\overline{\Theta}$ such that

$$\lim_{n\to\infty} \sup_{\theta\in\overline{\Theta}} |s_n(\omega,\theta) - s_\infty(\theta)| = 0, \ a.s.$$

(c) Identification: $s_\infty(\cdot)$ has a unique global maximum at $\theta^0 \in \Theta$, i.e., $s_\infty(\theta^0) > s_\infty(\theta)$, $\forall \theta \neq \theta^0, \theta \in \overline{\Theta}$

Then $\hat{\theta}_n \overset{a.s.}{\to} \theta^0$.

Proof: Select a $\omega \in \Omega$ and hold it fixed. Then $\{s_n(\omega, \theta)\}$ is a fixed sequence of functions. Suppose that ω is such that $s_n(\omega, \theta)$ converges to $s_\infty(\theta)$. This happens with probability one by assumption (b). The sequence $\{\hat{\theta}_n\}$ lies in the compact set $\overline{\Theta}$, by assumption (a) and the fact that maximixation is over $\overline{\Theta}$. Since every sequence from a compact set has at least one limit point (Bolzano-Weierstrass), say that $\hat{\theta}$ is a limit point of $\{\hat{\theta}_n\}$. There is a subsequence $\{\hat{\theta}_{n_m}\}$ ($\{n_m\}$ is simply a sequence of increasing integers) with $\lim_{m\to\infty} \hat{\theta}_{n_m} = \hat{\theta}$. By uniform convergence and continuity,

$$\lim_{m\to\infty} s_{n_m}(\hat{\theta}_{n_m}) = s_\infty(\hat{\theta}).$$

To see this, first of all, select an element $\hat{\theta}_t$ from the sequence $\left\{\hat{\theta}_{n_m}\right\}$. Then uniform convergence implies

$$\lim_{m\to\infty} s_{n_m}(\hat{\theta}_t) = s_\infty(\hat{\theta}_t)$$

Continuity of $s_\infty(\cdot)$ implies that

$$\lim_{t\to\infty} s_\infty(\hat{\theta}_t) = s_\infty(\hat{\theta})$$

since the limit as $t \to \infty$ of $\left\{\hat{\theta}_t\right\}$ is $\hat{\theta}$. So the above claim is true.

Next, by maximization

$$s_{n_m}(\hat{\theta}_{n_m}) \geq s_{n_m}(\theta^0)$$

which holds in the limit, so

$$\lim_{m\to\infty} s_{n_m}(\hat{\theta}_{n_m}) \geq \lim_{m\to\infty} s_{n_m}(\theta^0).$$

However,

$$\lim_{m\to\infty} s_{n_m}(\hat{\theta}_{n_m}) = s_\infty(\hat{\theta}),$$

as seen above, and

$$\lim_{m \to \infty} s_{n_m}(\theta^0) = s_\infty(\theta^0)$$

by uniform convergence, so

$$s_\infty(\hat{\theta}) \geq s_\infty(\theta^0).$$

But by assumption (3), there is a unique global maximum of $s_\infty(\theta)$ at θ^0, so we must have $s_\infty(\hat{\theta}) = s_\infty(\theta^0)$, and $\hat{\theta} = \theta^0$ in the limit. Finally, all of the above limits hold almost surely, since so far we have held ω fixed, but now we need to consider all $\omega \in \Omega$. Therefore $\{\hat{\theta}_n\}$ has only one limit point, θ^0, except on a set $C \subset \Omega$ with $P(C) = 0$.

Discussion of the proof:

- This proof relies on the identification assumption of a unique global maximum at θ^0. An equivalent way to state this is

(c) Identification: Any point θ in $\overline{\Theta}$ with $s_\infty(\theta) \geq s_\infty(\theta^0)$ must be such that $\| \theta - \theta^0 \| = 0$, which matches the way we will write the assumption in the section on nonparametric inference.

- We assume that $\hat{\theta}_n$ is in fact a global maximum of $s_n(\theta)$. It is not required to be unique for n finite, though the identification assumption requires that the limiting objective function have a unique maximizing argument. The previous section on numeric optimization methods showed that actually finding the global maximum of $s_n(\theta)$ may be a non-trivial problem.

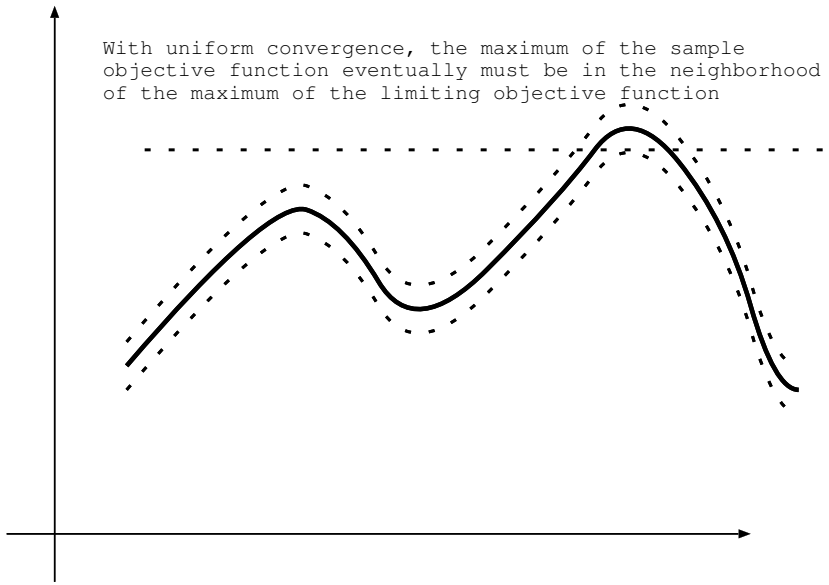
- See Amemiya's Example 4.1.4 for a case where discontinuity leads to breakdown of consistency.

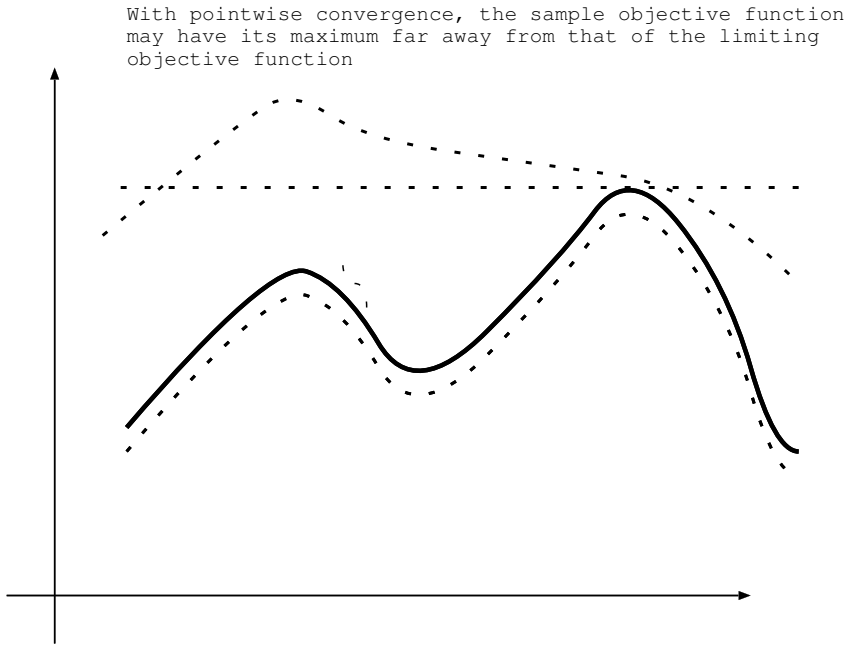
- The assumption that θ^0 is in the interior of $\overline{\Theta}$ (part of the identification assumption) has not been used to prove consistency, so we could directly assume that θ^0 is simply an element of a compact

set $\overline{\Theta}$. The reason that we assume it's in the interior here is that this is necessary for subsequent proof of asymptotic normality, and I'd like to maintain a minimal set of simple assumptions, for clarity. Parameters on the boundary of the parameter set cause theoretical difficulties that we will not deal with in this course. Just note that conventional hypothesis testing methods do not apply in this case.

- Note that $s_n(\theta)$ is not required to be continuous, though $s_\infty(\theta)$ is.

- The following figures illustrate why uniform convergence is important. In the second figure, if the function is not converging quickly enough around the lower of the two maxima. If the pointwise convergence in this region is slow enough, there is no guarantee that the maximizer will be in the neighborhood of the global maximizer of $s_\infty(\theta)$, even when n is very large. Uniform comvergence means that we are in the situation of the top graphic. As long as n is large enough, the maximum will be in the neighborhood of the global maximum of $s_\infty(\theta)$.


With uniform convergence, the maximum of the sample
objective function eventually must be in the neighborhood
of the maximum of the limiting objective function



Sufficient conditions for assumption (b)

We need a uniform strong law of large numbers in order to verify assumption (2) of Theorem 29. To verify the uniform convergence assumption, it is often feasible to employ the following set of stronger assumptions:

- the parameter space is compact, which is given by assumption (b)
- the objective function $s_n(\theta)$ is continuous and bounded with probability one on the entire parameter space

- a standard SLLN can be shown to apply to some point θ in the parameter space. That is, we can show that $s_n(\theta) \xrightarrow{a.s.} s_\infty(\theta)$ for some θ. Note that in most cases, the objective function will be an average of terms, such as

$$s_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} s_t(\theta)$$

  As long as the $s_t(\theta)$ are not too strongly dependent, and have finite variances, we can usually find a SLLN that will apply.

With these assumptions, it can be shown that pointwise convergence holds throughout the parameter space, so we obtain the needed uniform convergence.

These are reasonable conditions in many cases, and henceforth when dealing with specific estimators we'll simply assume that pointwise almost sure convergence can be extended to uniform almost sure convergence in this way.

More on the limiting objective function

The limiting objective function in assumption (b) is $s_\infty(\theta)$. What is the nature of this function and where does it come from?

- Remember our paradigm - data is presumed to be generated as a draw from $f_{Z_n}(z)$, and the objective function is $s_n(Z_n, \theta)$.
- Usually, $s_n(Z_n, \theta)$ is an average of terms.
- The limiting objective function is found by applying a strong (weak) law of large numbers to $s_n(Z_n, \theta)$.

- A strong (weak) LLN says that an average of terms converges almost surely (in probability) to the limit of the expectation of the average.

Supposing one holds,

$$s_\infty(\theta) = \lim_{n\to\infty} \mathcal{E} s_n(Z_n, \theta) = \lim_{n\to\infty} \int_{\mathcal{Z}_n} s_n(z, \theta) f_{Z_n}(z) dz$$

Now suppose that the density $f_{Z_n}(z)$ that characterizes the DGP is parametric: $f_{Z_n}(z; \rho)$, $\rho \in \varrho$, and the data is generated by $\rho^0 \in \varrho$. Now we have two parameters to worry about, θ and ρ. We are probably interested in learning about the true DGP, which means that ρ^0 is the item of interest. When the DGP is parametric, the limiting objective function is

$$s_\infty(\theta) = \lim_{n\to\infty} \mathcal{E} s_n(Z_n, \theta) = \lim_{n\to\infty} \int_{\mathcal{Z}_n} s_n(z, \theta) f_{Z_n}(z; \rho^0) dz$$

and we can write the limiting objective function as $s_\infty(\theta, \rho^0)$ to emphasize the dependence on the parameter of the DGP. From the theorem, we know that $\hat{\theta}_n \xrightarrow{a.s.} \theta^0$ What is the relationship between θ^0 and ρ^0?

- ρ and θ may have different dimensions. Often, the statistical model (with parameter θ) only partially describes the DGP. For example, the case of OLS with errors of unknown distribution. In some cases, the dimension of θ may be greater than that of ρ. For example, fitting a polynomial to an unknown nonlinear function.

- If knowledge of θ^0 is sufficient for knowledge of ρ^0, we have a correctly and fully specified model. θ^0 is referred to as the *true parameter value*.

- If knowledge of θ^0 is sufficient for knowledge of some but not all elements of ρ^0, we have a correctly specified semiparametric model. θ^0 is referred to as the *true parameter value*, understanding that not all parameters of the DGP are estimated.

- If knowledge of θ^0 is not sufficient for knowledge of any elements of ρ^0, or if it causes us to draw false conclusions regarding at least some of the elements of ρ^0, our model is misspecified. θ^0 is referred to as the *pseudo-true parameter value*.

Summary

The theorem for consistency is really quite intuitive. It says that with probability one, an extremum estimator converges to the value that maximizes the limit of the expectation of the objective function. Because the objective function may or may not make sense, depending on how good or poor is the model, we may or may not be estimating parameters of the DGP.

12.4 Example: Consistency of Least Squares

We suppose that data is generated by random sampling of (Y, X), where $y_t = \beta_0 x_t + \varepsilon_t$. (X, ε) has the common distribution function $F_Z = \mu_x \mu_\varepsilon$ (x and ε are independent) with support $\mathcal{Z} = \mathcal{X} \times \mathcal{E}$. Suppose that the variances σ_X^2 and σ_ε^2 are finite. The sample objective function for a sample size n is

$$\begin{aligned} s_n(\theta) &= 1/n \sum_{t=1}^{n} (y_t - \beta x_t)^2 = 1/n \sum_{i=1}^{n} (\beta_0 x_t + \varepsilon_t - \beta x_t)^2 \\ &= 1/n \sum_{t=1}^{n} (x_t (\beta_0 - \beta))^2 + 2/n \sum_{t=1}^{n} x_t (\beta_0 - \beta) \varepsilon_t + 1/n \sum_{t=1}^{n} \varepsilon_t^2 \end{aligned}$$

- Considering the last term, by the SLLN,

$$1/n \sum_{t=1}^{n} \varepsilon_t^2 \overset{a.s.}{\to} \int_{\mathcal{X}} \int_{\mathcal{E}} \varepsilon^2 d\mu_{\mathcal{X}} d\mu_{\mathcal{E}} = \sigma_\varepsilon^2.$$

- Considering the second term, since $E(\varepsilon) = 0$ and X and ε are independent, the SLLN implies that it converges to zero.

- Finally, for the first term, for a given β, we assume that a SLLN applies so that

$$\begin{aligned} 1/n \sum_{t=1}^{n} \left(x_t \left(\beta_0 - \beta\right)\right)^2 &\overset{a.s.}{\to} \int_{\mathcal{X}} \left(x \left(\beta_0 - \beta\right)\right)^2 d\mu_{\mathcal{X}} \\ &= \left(\beta^0 - \beta\right)^2 \int_{\mathcal{X}} x^2 d\mu_{\mathcal{X}} \\ &= \left(\beta^0 - \beta\right)^2 E\left(X^2\right) \end{aligned} \tag{12.1}$$

Finally, the objective function is clearly continuous, and the parameter space is assumed to be compact, so the convergence is also uniform. Thus,

$$s_\infty(\beta) = \left(\beta^0 - \beta\right)^2 E\left(X^2\right) + \sigma_\varepsilon^2$$

A minimizer of this is clearly $\beta = \beta^0$.

Exercise 30. Show that in order for the above solution to be unique it is necessary that $E(X^2) \neq 0$. Interpret this condition.

This example shows that Theorem 29 can be used to prove strong consistency of the OLS estimator. There are easier ways to show this, of course - this is only an example of application of the theorem.

12.5 Example: Inconsistency of Misspecified Least Squares

You already know that the OLS estimator is inconsistent when relevant variables are omitted. Let's verify this result in the context of extremum estimators. We suppose that data is generated by random sampling of (Y, X), where $y_t = \beta_0 x_t + \varepsilon_t$. (X, ε) has the common distribution function $F_Z = \mu_x \mu_\varepsilon$ (x and ε are independent) with support $\mathcal{Z} = \mathcal{X} \times \mathcal{E}$. Suppose that the variances σ_X^2 and σ_ε^2 are finite. However, the econometrician is unaware of the true DGP, and instead proposes the misspecified model $y_t = \gamma_0 w_t + \eta_t$. Suppose that $E(W\epsilon) = 0$ but that $E(WX) \neq 0$.

The sample objective function for a sample size n is

$$\begin{aligned} s_n(\gamma) &= 1/n \sum_{t=1}^{n} (y_t - \gamma w_t)^2 = 1/n \sum_{i=1}^{n} (\beta_0 x_t + \varepsilon_t - \gamma w_t)^2 \\ &= 1/n \sum_{t=1}^{n} (\beta_0 x_t)^2 + 1/n \sum_{t=1}^{n} (\gamma w_t)^2 + 1/n \sum_{t=1}^{n} \varepsilon_t^2 + 2/n \sum_{t=1}^{n} \beta_0 x_t \varepsilon_t - 2/n \sum_{t=1}^{n} \beta_0 \gamma x_t w_t - 2/n \sum_{t=1}^{n} \varepsilon_t x_t w_t \end{aligned}$$

Using arguments similar to above,

$$s_\infty(\gamma) = \gamma^2 E\left(W^2\right) - 2\beta_0 \gamma E(WX) + C$$

So, $\gamma_0 = \frac{\beta_0 E(WX)}{E(W^2)}$, which is the true parameter of the DGP, multiplied by the pseudo-true value of a regression of X on W. The OLS estimator is not consistent for the true parameter, β_0

12.6 Example: Linearization of a nonlinear model

Ref. Gourieroux and Monfort, section 8.3.4. White, *Intn'l Econ. Rev.* 1980 is an earlier reference.

Suppose we have a nonlinear model

$$y_i = h(x_i, \theta^0) + \varepsilon_i$$

where

$$\varepsilon_i \sim iid(0, \sigma^2)$$

The *nonlinear least squares* estimator solves

$$\hat{\theta}_n = \arg\min \frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i, \theta))^2$$

We'll study this more later, but for now it is clear that the foc for minimization will require solving a set of nonlinear equations. A common approach to the problem seeks to avoid this difficulty by *linearizing* the model. A first order Taylor's series expansion about the point x_0 with remainder gives

$$y_i = h(x^0, \theta^0) + (x_i - x_0)' \frac{\partial h(x_0, \theta^0)}{\partial x} + \nu_i$$

where ν_i encompasses both ε_i and the Taylor's series remainder. Note that ν_i is no longer a classical error - its mean is not zero. We should expect problems.

Define

$$\begin{aligned} \alpha^* &= h(x_0, \theta^0) - x_0' \frac{\partial h(x^0, \theta^0)}{\partial x} \\ \beta^* &= \frac{\partial h(x_0, \theta^0)}{\partial x} \end{aligned}$$

Given this, one might try to estimate α^* and β^* by applying OLS to

$$y_i = \alpha + \beta x_i + \nu_i$$

- Question, will $\hat{\alpha}$ and $\hat{\beta}$ be consistent for α^* and β^*?
- The answer is no, as one can see by interpreting $\hat{\alpha}$ and $\hat{\beta}$ as extremum estimators. Let $\gamma = (\alpha, \beta')'$.

$$\hat{\gamma} = \arg\min s_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)^2$$

The objective function converges to its expectation

$$s_n(\gamma) \overset{u.a.s.}{\to} s_\infty(\gamma) = \mathcal{E}_X\mathcal{E}_{Y|X}\left(y - \alpha - \beta x\right)^2$$

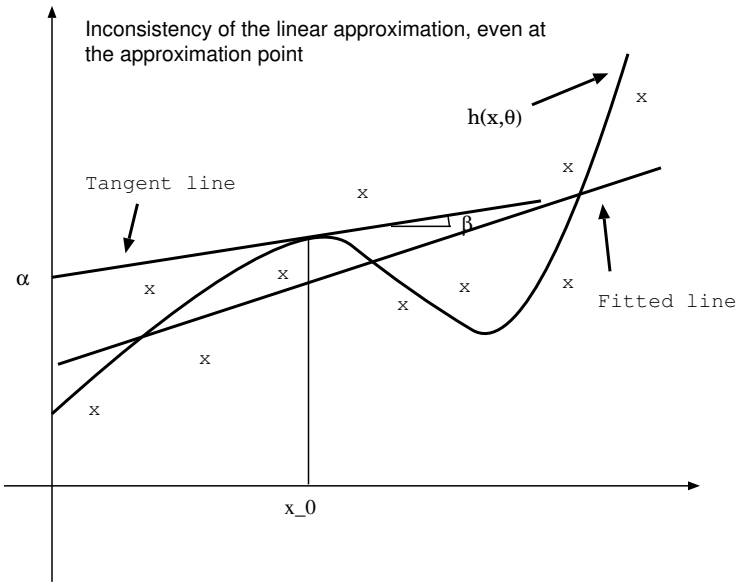
and $\hat{\gamma}$ converges $a.s.$ to the γ^0 that minimizes $s_\infty(\gamma)$:

$$\gamma^0 = \arg\min \mathcal{E}_X\mathcal{E}_{Y|X}\left(y - \alpha - \beta x\right)^2$$

Noting that

$$\begin{aligned}\mathcal{E}_X\mathcal{E}_{Y|X}\left(y - \alpha - x'\beta\right)^2 &= \mathcal{E}_X\mathcal{E}_{Y|X}\left(h(x,\theta^0) + \varepsilon - \alpha - \beta x\right)^2 \\ &= \sigma^2 + \mathcal{E}_X\left(h(x,\theta^0) - \alpha - \beta x\right)^2\end{aligned}$$

since cross products involving ε drop out. α^0 and β^0 correspond to the hyperplane that is closest to the true regression function $h(x,\theta^0)$ according to the mean squared error criterion. This depends on both the shape of $h(\cdot)$ and the density function of the conditioning variables.



- It is clear that the tangent line does not minimize MSE, since, for example, if $h(x, \theta^0)$ is concave, all errors between the tangent line and the true function are negative.

- Note that the true underlying parameter θ^0 is not estimated consistently, either (it may be of a different dimension than the dimension of the parameter of the approximating model, which is 2 in this example).

- Second order and higher-order approximations suffer from exactly the same problem, though to a less severe degree, of course. For this reason, translog, Generalized Leontiev and other "flexible functional forms" based upon second-order approximations in general suffer from bias and inconsistency. The bias may not be too important for analysis of conditional means, but it can be very important for analyzing first and second derivatives. In production and consumer

analysis, first and second derivatives (*e.g.,* elasticities of substitution) are often of interest, so in this case, one should be cautious of unthinking application of models that impose stong restrictions on second derivatives.

- This sort of linearization about a long run equilibrium is a common practice in dynamic macroeconomic models. It is justified for the purposes of theoretical analysis of a model *given* the model's parameters, but it is not justifiable for the estimation of the parameters of the model using data. The section on simulation-based methods offers a means of obtaining consistent estimators of the parameters of dynamic macro models that are too complex for standard methods of analysis.

12.7 Asymptotic Normality

A consistent estimator is oftentimes not very useful unless we know how fast it is likely to be converging to the true value, and the probability that it is far away from the true value. Establishment of asymptotic normality with a known scaling factor solves these two problems. The following theorem is similar to Amemiya's Theorem 4.1.3 (pg. 111).

Theorem 31. [Asymptotic normality of e.e.] *In addition to the assumptions of Theorem 29, assume*
(a) $\mathcal{J}_n(\theta) \equiv D_\theta^2 s_n(\theta)$ *exists and is continuous in an open, convex neighborhood of* θ^0.
(b) $\{\mathcal{J}_n(\theta_n)\} \xrightarrow{a.s.} \mathcal{J}_\infty(\theta^0)$, *a finite negative definite matrix, for any sequence* $\{\theta_n\}$ *that converges almost surely to* θ^0.
(c) $\sqrt{n} D_\theta s_n(\theta^0) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta^0)\right]$, *where* $\mathcal{I}_\infty(\theta^0) = \lim_{n\to\infty} Var\sqrt{n} D_\theta s_n(\theta^0)$
Then $\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right]$

Proof: By Taylor expansion:

$$D_\theta s_n(\hat{\theta}_n) = D_\theta s_n(\theta^0) + D_\theta^2 s_n(\theta^*)\left(\hat{\theta} - \theta^0\right)$$

where $\theta^* = \lambda\hat{\theta} + (1-\lambda)\theta^0$, $0 \leq \lambda \leq 1$.

- Note that $\hat{\theta}$ will be in the neighborhood where $D_\theta^2 s_n(\theta)$ exists with probability one as n becomes large, by consistency.
- Now the l.h.s. of this equation is zero, at least asymptotically, since $\hat{\theta}_n$ is a maximizer and the f.o.c. must hold exactly since the limiting objective function is strictly concave in a neighborhood of θ^0.
- Also, since θ^* is between $\hat{\theta}_n$ and θ^0, and since $\hat{\theta}_n \xrightarrow{a.s.} \theta^0$, assumption (b) gives

$$D_\theta^2 s_n(\theta^*) \xrightarrow{a.s.} \mathcal{J}_\infty(\theta^0)$$

So

$$0 = D_\theta s_n(\theta^0) + \left[\mathcal{J}_\infty(\theta^0) + o_s(1)\right]\left(\hat{\theta} - \theta^0\right)$$

And

$$0 = \sqrt{n} D_\theta s_n(\theta^0) + \left[\mathcal{J}_\infty(\theta^0) + o_s(1)\right]\sqrt{n}\left(\hat{\theta} - \theta^0\right)$$

Now $\sqrt{n} D_\theta s_n(\theta^0) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta^0)\right]$ by assumption c, so

$$-\left[\mathcal{J}_\infty(\theta^0) + o_s(1)\right]\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta^0)\right]$$

Also, $[\mathcal{J}_\infty(\theta^0) + o_s(1)] \xrightarrow{a.s.} \mathcal{J}(\theta^0)$, so

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right]$$

by the Slutsky Theorem (see Gallant, Theorem 4.6).

Figure

- **Skip this in lecture.** A note on the order of these matrices: Supposing that $s_n(\theta)$ is representable as an average of n terms, which is the case for all estimators we consider, $D_\theta^2 s_n(\theta)$ is also an average of n matrices, the elements of which are not centered (they do not have zero expectation). Supposing a SLLN applies, the almost sure limit of $D_\theta^2 s_n(\theta^0)$, $\mathcal{J}_\infty(\theta^0) = O(1)$, as we saw in Example 89. On the other hand, assumption (c):$\sqrt{n}D_\theta s_n(\theta^0) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta^0)\right]$ means that

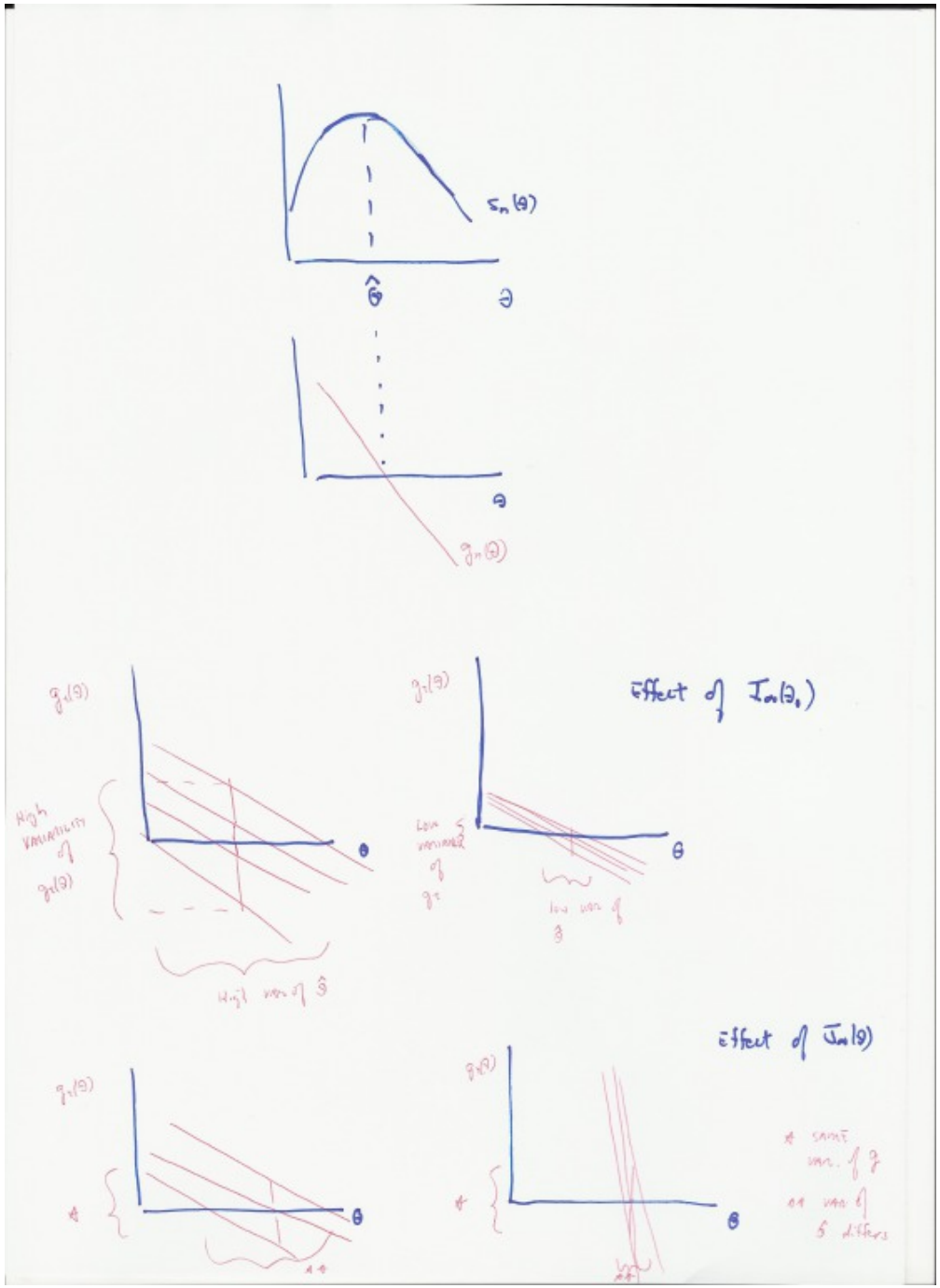
  $$\sqrt{n}D_\theta s_n(\theta^0) = O_p(1)$$

  where we use the result of Example 87. If we were to omit the $\sqrt{n}$, we'd have

  $$\begin{aligned} D_\theta s_n(\theta^0) &= n^{-\frac{1}{2}} O_p(1) \\ &= O_p\left(n^{-\frac{1}{2}}\right) \end{aligned}$$

  where we use the fact that $O_p(n^r)O_p(n^q) = O_p(n^{r+q})$. The sequence $D_\theta s_n(\theta^0)$ *is* centered, so we need to scale by $\sqrt{n}$ to avoid convergence to zero.

Figure 12.1: Effects of I_∞ and J_∞



12.8 Example: Classical linear model

Let's use the results to get the asymptotic distribution of the OLS estimator applied to the classical model, to verify that we obtain the results seen before. The OLS criterion is

$$\begin{aligned} s_n(\beta) &= \frac{1}{n}(y - X\beta)'(y - X\beta) \\ &= \frac{1}{n}\left(X\beta^0 + \epsilon - X\beta\right)'\left(X\beta^0 + \epsilon - X\beta\right) \\ &= \frac{1}{n}\left[\left(\beta^0 - \beta\right)'X'X\left(\beta^0 - \beta\right) - 2\epsilon'X\beta + \epsilon'\epsilon\right] \end{aligned}$$

The first derivative is

$$D_\beta s_n(\beta) = \frac{1}{n}\left[-2X'X\left(\beta^0 - \beta\right) - 2X'\epsilon\right]$$

so, evaluating at β^0,

$$D_\beta s_n(\beta^0) = -2\frac{X'\epsilon}{n}$$

This has expectation 0, so the variance is the expectation of the outer product:

$$\begin{aligned} Var\sqrt{n}D_\beta s_n(\beta^0) &= E\left[\left(-\sqrt{n}2\frac{X'\epsilon}{n}\right)\left(-\sqrt{n}2\frac{X'\epsilon}{n}\right)'\right] \\ &= E4\frac{X'\epsilon\epsilon'X}{n} \\ &= 4\sigma_\epsilon^2 E\left(\frac{X'X}{n}\right) \end{aligned}$$

(assuming regressors independent of errors). Therefore

$$\begin{aligned}\mathcal{I}_\infty(\beta^0) &= \lim_{n\to\infty} Var\sqrt{n}D_\beta s_n(\beta^0) \\ &= 4\sigma_\epsilon^2 Q_X\end{aligned}$$

where $Q_X = limE\left(\frac{X'X}{n}\right)$, a finite p.d. matrix, is obtained using a LLN.

The second derivative is

$$\mathcal{J}_n(\beta) = D_\beta^2 s_n(\beta^0) = \frac{1}{n}\left[2X'X\right].$$

A SLLN tells us that this converges almost surely to the limit of its expectation:

$$\mathcal{J}_\infty(\beta^0) = 2Q_X$$

There's no parameter in that last expression, so uniformity is not an issue.

The asymptotic normality theorem (31) tells us that

$$\sqrt{n}\left(\hat{\beta} - \beta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\beta^0)^{-1}\mathcal{I}_\infty(\beta^0)\mathcal{J}_\infty(\beta^0)^{-1}\right]$$

which is, given the above,

$$\sqrt{n}\left(\hat{\beta} - \beta^0\right) \xrightarrow{d} N\left[0, \left(\frac{Q_X^{-1}}{2}\right)\left(4\sigma_\epsilon^2 Q_X\right)\left(\frac{Q_X^{-1}}{2}\right)\right]$$

or

$$\sqrt{n}\left(\hat{\beta} - \beta^0\right) \xrightarrow{d} N\left[0, Q_X^{-1}\sigma_\epsilon^2\right].$$

This is the same thing we saw in equation 4.1, of course. So, the theory seems to work :-)

12.9 Exercises

1. Suppose that $x_i \sim$ uniform(0,1), and $y_i = 1 - x_i^2 + \varepsilon_i$, where ε_i is iid(0,σ^2). Suppose we estimate the misspecified model $y_i = \alpha + \beta x_i + \eta_i$ by OLS. Find the numeric values of α^0 and β^0 that are the probability limits of $\hat{\alpha}$ and $\hat{\beta}$

2. Verify your results using Octave by generating data that follows the above model, and calculating the OLS estimator. When the sample size is very large the estimator should be very close to the analytical results you obtained in question 1.

3. Use the asymptotic normality theorem to find the asymptotic distribution of the ML estimator of β^0 for the model $y = x\beta^0 + \varepsilon$, where $\varepsilon \sim N(0,1)$ and is independent of x. This means finding $\frac{\partial^2}{\partial\beta\partial\beta'} s_n(\beta)$, $\mathcal{J}(\beta^0)$, $\frac{\partial s_n(\beta)}{\partial\beta}\Big|$, and $\mathcal{I}(\beta^0)$. The expressions may involve the unspecified density of x.

Chapter 13

Maximum likelihood estimation

The maximum likelihood estimator is important because it uses all of the information in a fully specified statistical model. Its use of all of the information causes it to have a number of attractive properties, foremost of which is asymptotic efficiency. For this reason, the ML estimator can serve as a benchmark against which other estimators may be measured. The ML estimator requires that the statistical model be fully specified, which essentially means that there is enough information to draw data from the DGP, given the parameter. This is a fairly strong requirement, and for this reason we need to be concerned about the possible misspecification of the statistical model. If this is the case, the ML estimator will not have the nice properties that it has under correct specification.

13.1 The likelihood function

Suppose we have a sample of size n of the random vectors y and z. Suppose the joint density of $Y = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix}$ and $Z = \begin{pmatrix} z_1 & \dots & z_n \end{pmatrix}$ is characterized by a parameter vector ψ_0 :

$$f_{YZ}(Y, Z, \psi_0).$$

This is the joint density of the sample. This density can be factored as

$$f_{YZ}(Y, Z, \psi_0) = f_{Y|Z}(Y|Z, \theta_0) f_Z(Z, \rho_0)$$

The *likelihood function* is just this density evaluated at other values ψ

$$L(Y, Z, \psi) = f(Y, Z, \psi), \psi \in \Psi,$$

where Ψ is a *parameter space.*

The *maximum likelihood estimator* of ψ_0 is the value of ψ that maximizes the likelihood function.

Note that if θ_0 and ρ_0 share no elements, then the maximizer of the conditional likelihood function $f_{Y|Z}(Y|Z, \theta)$ with respect to θ is the same as the maximizer of the overall likelihood function $f_{YZ}(Y, Z, \psi) = f_{Y|Z}(Y|Z, \theta) f_Z(Z, \rho)$, for the elements of ψ that correspond to θ. In this case, the variables Z are said to be *exogenous* for estimation of θ, and we may more conveniently work with the conditional likelihood function $f_{Y|Z}(Y|Z, \theta)$ for the purposes of estimating θ_0.

When this is the case, the maximum likelihood estimator of $\theta_0 = \arg\max f_{Y|Z}(Y|Z, \theta)$. We'll suppose this framework in what follows.

- If the n observations are independent, the likelihood function can be written as

$$L(Y|Z,\theta) = \prod_{t=1}^{n} f(y_t|z_t,\theta)$$

- If this is not possible, we can always factor the likelihood into *contributions of observations,* by using the fact that a joint density can be factored into the product of a marginal and conditional (doing this iteratively)

$$L(Y,\theta) = f(y_1|z_1,\theta)f(y_2|y_1,z_2,\theta)f(y_3|y_1,y_2,z_3,\theta)\cdots f(y_n|y_1,y_2,\ldots y_{t-n},z_n,\theta)$$

To simplify notation, define

$$x_t \;=\; \{y_1, y_2, ..., y_{t-1}, z_t\}$$

so $x_1 = z_1$, $x_2 = \{y_1, z_2\}$, *etc.* - it contains exogenous and predetermined endogeous variables. Now the likelihood function can be written as

$$L(Y,\theta) = \prod_{t=1}^{n} f(y_t|x_t,\theta)$$

The criterion function can be defined as the average log-likelihood function:

$$s_n(\theta) = \frac{1}{n}\ln L(Y,\theta) = \frac{1}{n}\sum_{t=1}^{n}\ln f(y_t|x_t,\theta)$$

The maximum likelihood estimator may thus be defined equivalently as

$$\hat{\theta} = \arg\max s_n(\theta),$$

where the set maximized over is defined below. Since $\ln(\cdot)$ is a monotonic increasing function, $\ln L$ and L maximize at the same value of θ. Dividing by n has no effect on $\hat{\theta}$.

Example 32. Example: Bernoulli trial
Suppose that we are flipping a coin that may be biased, so that the probability of a heads may not be 0.5. Maybe we're interested in estimating the probability of a heads. Let $Y = 1(heads)$ be a binary variable that indicates whether or not a heads is observed. The outcome of a toss is a Bernoulli random variable:

$$\begin{aligned} f_Y(y, p_0) &= p_0^y (1 - p_0)^{1-y}, y \in \{0, 1\} \\ &= 0, y \notin \{0, 1\} \end{aligned}$$

So a representative term that enters the likelihood function is

$$f_Y(y, p) = p^y (1 - p)^{1-y}$$

and

$$\ln f_Y(y, p) = y \ln p + (1 - y) \ln (1 - p)$$

The derivative of this is

$$\begin{aligned}\frac{\partial \ln f_Y(y,p)}{\partial p} &= \frac{y}{p} - \frac{(1-y)}{(1-p)} \\ &= \frac{y-p}{p\,(1-p)}\end{aligned}$$

Averaging this over a sample of size n gives

$$\frac{\partial s_n(p)}{\partial p} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - p}{p\,(1-p)}$$

Setting to zero and solving gives

$$\hat{p} = \bar{y} \tag{13.1}$$

So it's easy to calculate the MLE of p_0 in this case. For future reference, note that $E(Y) = \sum_{Y=0}^{Y=1} y p_0^y \left(1-p_0\right)^{1-y} = p_0$ and $Var(Y) = E(Y^2) - \left[E(Y)\right]^2 = p_0 - p_0^2$.

For this example, $s_n(p) = \frac{1}{n}\sum_{t=1}^{n} y_t \ln p + (1-y_t)\ln(1-p)$.

- A LLN tells us that $s_n(p) \rightarrow^{a.s.} p_0 \ln p + (1-p_0)\ln(1-p)$.
- The parameter space is compact (p_0 lies between 0 and 1)
- the objective function is continous
- thus, the a.s. convergence is also uniform.

The consistency theorem for extremum estimators tells us that the ML estimator converges to the value that maximized the limiting objective function. Because $s_\infty(p) = p_0 \ln p + (1-p_0)\ln(1-p)$, we

can easily check that the maximizer is p_0. So, the ML estimator is consistent for the true probability.

In practice, we need to ensure that p stays between 0 and 1. To do this with an unconstrained optimization algorithm, we can use a *parameterization.* See subsection 13.8 for an example.

Now imagine that we had a bag full of bent coins, each bent around a sphere of a different radius (with the head pointing to the outside of the sphere). We might suspect that the probability of a heads could depend upon the radius. Suppose that $p_i \equiv p(x_i, \beta) = (1 + \exp(-x_i'\beta))^{-1}$ where $x_i = \begin{bmatrix} 1 & r_i \end{bmatrix}'$, so that β is a 2×1 vector. Now

$$\frac{\partial p_i(\beta)}{\partial \beta} = p_i \left(1 - p_i\right) x_i$$

so

$$\begin{aligned} \frac{\partial \ln f_Y(y, \beta)}{\partial \beta} &= \frac{y - p_i}{p_i \left(1 - p_i\right)} p_i \left(1 - p_i\right) x_i \\ &= \left(y_i - p(x_i, \beta)\right) x_i \end{aligned}$$

So the derivative of the average log lihelihood function is now

$$\frac{\partial s_n(\beta)}{\partial \beta} = \frac{\sum_{i=1}^n \left(y_i - p(x_i, \beta)\right) x_i}{n}$$

This is a set of 2 nonlinear equations in the two unknown elements in β. There is no explicit solution for the two elements that set the equations to zero. This is commonly the case with ML estimators: they are often nonlinear, and finding the value of the estimate often requires use of numeric methods to find solutions to the first order conditions. See Chapter 11 for more information on how to do this.

Example 33. Example: Likelihood function of classical linear regression model

The classical linear regression model with normality is outlined in Section 3.6. The likelihood function for this model is presented in Section 4.3. A Octave/Matlab example that shows how to compute the maximum likelihood estimator for data that follows the CLRM with normality is in NormalExample.m , which makes use of NormalLF.m .

13.2 Consistency of MLE

The MLE is an extremum estimator, given basic assumptions it is consistent for the value that maximizes the limiting objective function, following Theorem 29. The question is: what is the value that maximizes $s_\infty(\theta)$?

Remember that $s_n(\theta) = \frac{1}{n} \ln L(Y, \theta)$, and $L(Y, \theta_0)$ is the true density of the sample data. For any $\theta \neq \theta_0$

$$\mathcal{E}\left(\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right) \leq \ln\left(\mathcal{E}\left(\frac{L(\theta)}{L(\theta_0)}\right)\right)$$

by Jensen's inequality ($\ln(\cdot)$ is a concave function).

Now, the expectation on the RHS is

$$\mathcal{E}\left(\frac{L(\theta)}{L(\theta_0)}\right) = \int \frac{L(\theta)}{L(\theta_0)} L(\theta_0) dy = 1,$$

since $L(\theta_0)$ *is* the density function of the observations, and since the integral of any density is 1. Therefore, since $\ln(1) = 0$,

$$\mathcal{E}\left(\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right) \leq 0,$$

or

$$\mathcal{E}\left(s_n\left(\theta\right)\right) - \mathcal{E}\left(s_n\left(\theta_0\right)\right) \leq 0.$$

A SLLN tells us that $s_n(\theta) \overset{a.s.}{\to} s_\infty(\theta, \theta_0) = \lim \mathcal{E}\left(s_n\left(\theta\right)\right)$, and with continuity and a compact parameter space, this is uniform, so

$$s_\infty(\theta, \theta_0) - s_\infty(\theta_0, \theta_0) \leq 0$$

except on a set of zero probability. Note: the θ_0 appears because the expectation is taken with respect to the true density $L(\theta_0)$.

By the identification assumption there is a unique maximizer, so the inequality is strict if $\theta \neq \theta_0$:

$$s_\infty(\theta, \theta_0) - s_\infty(\theta_0, \theta_0) < 0, \forall \theta \neq \theta_0, \text{a.s.}$$

Therefore, θ_0 is the unique maximizer of $s_\infty(\theta, \theta_0)$, and thus, Theorem 29 tells us that

$$\lim_{n \to \infty} \hat{\theta} = \theta_0, \text{ a.s.}$$

So, the ML estimator is consistent for the true parameter value.

13.3 The score function

Assumption: (Differentiability) Assume that $s_n(\theta)$ is twice continuously differentiable in a neighborhood $N(\theta_0)$ of θ_0, at least when n is large enough.

To maximize the log-likelihood function, take derivatives:

$$\begin{aligned} g_n(Y,\theta) &= D_\theta s_n(\theta) \\ &= \frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f(y_t|x_x,\theta) \\ &\equiv \frac{1}{n}\sum_{t=1}^{n} g_t(\theta). \end{aligned}$$

This is the *score vector* (with dim $K \times 1$). Note that the score function has Y as an argument, which implies that it is a random function. Y (and any exogeneous variables) will often be suppressed for clarity, but one should not forget that they are still there.

The ML estimator $\hat{\theta}$ sets the derivatives to zero:

$$g_n(\hat{\theta}) = \frac{1}{n}\sum_{t=1}^{n} g_t(\hat{\theta}) \equiv 0.$$

We will show that $\mathcal{E}_\theta\left[g_t(\theta)\right] = 0$, $\forall t$. *This is the expectation taken with respect to the density* $f(\theta)$, not necessarily $f\left(\theta_0\right)$.

$$\begin{aligned} \mathcal{E}_\theta\left[g_t(\theta)\right] &= \int [D_\theta \ln f(y_t|x_t,\theta)] f(y_t|x,\theta) dy_t \\ &= \int \frac{1}{f(y_t|x_t,\theta)} \left[D_\theta f(y_t|x_t,\theta)\right] f(y_t|x_t,\theta) dy_t \\ &= \int D_\theta f(y_t|x_t,\theta) dy_t. \end{aligned}$$

Given some regularity conditions on boundedness of $D_\theta f$, we can switch the order of integration and

differentiation, by the dominated convergence theorem. This gives

$$\begin{aligned} \mathcal{E}_\theta \left[g_t(\theta) \right] &= D_\theta \int f(y_t|x_t, \theta) dy_t \\ &= D_\theta 1 \\ &= 0 \end{aligned} \tag{13.2}$$

where we use the fact that the integral of the density is 1.

- So $\mathcal{E}_\theta(g_t(\theta) = 0$: *the expectation of the score vector is zero.*
- This hold for all t, so it implies that $\mathcal{E}_\theta g_n(Y, \theta) = 0$.

13.4 Asymptotic normality of MLE

Recall that we assume that the log-likelihood function $s_n(\theta)$ is twice continuously differentiable. Take a first order Taylor's series expansion of $g(Y, \hat{\theta})$ about the true value θ_0 :

$$0 \equiv g(\hat{\theta}) = g(\theta_0) + (D_{\theta'} g(\theta^*)) \left(\hat{\theta} - \theta_0 \right)$$

or with appropriate definitions

$$\mathcal{J}(\theta^*) \left(\hat{\theta} - \theta_0 \right) = -g(\theta_0),$$

where $\theta^* = \lambda \hat{\theta} + (1 - \lambda)\theta_0, 0 < \lambda < 1$. Assume $\mathcal{J}(\theta^*)$ is invertible (we'll justify this in a minute). So

$$\sqrt{n} \left(\hat{\theta} - \theta_0 \right) = -\mathcal{J}(\theta^*)^{-1} \sqrt{n} g(\theta_0) \tag{13.3}$$

Now consider $\mathcal{J}(\theta^*)$, the matrix of second derivatives of the average log likelihood function. This is

$$\begin{aligned}\mathcal{J}(\theta^*) &= D_{\theta'} g(\theta^*) \\ &= D_\theta^2 s_n(\theta^*) \\ &= \frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f_t(\theta^*)\end{aligned}$$

where the notation

$$D_\theta^2 s_n(\theta) \equiv \frac{\partial^2 s_n(\theta)}{\partial\theta\partial\theta'}.$$

Given that this is an average of terms, it should usually be the case that this satisfies a strong law of large numbers (SLLN). *Regularity conditions* are a set of assumptions that guarantee that this will happen. There are different sets of assumptions that can be used to justify appeal to different SLLN's. For example, the $D_\theta^2 \ln f_t(\theta^*)$ must not be too strongly dependent over time, and their variances must not become infinite. We don't assume any particular set here, since the appropriate assumptions will depend upon the particularities of a given model. However, we assume that a SLLN applies.

Also, since we know that $\hat{\theta}$ is consistent, and since $\theta^* = \lambda\hat{\theta} + (1-\lambda)\theta_0$, we have that $\theta^* \overset{a.s.}{\to} \theta_0$. Also, by the above differentiability assumption, $\mathcal{J}(\theta)$ is continuous in θ. Given this, $\mathcal{J}(\theta^*)$ converges to the limit of it's expectation:

$$\mathcal{J}(\theta^*) \overset{a.s.}{\to} \lim_{n\to\infty} \mathcal{E}\left(D_\theta^2 s_n(\theta_0)\right) = \mathcal{J}_\infty(\theta_0) < \infty$$

This matrix converges to a finite limit.

Re-arranging orders of limits and differentiation, which is legitimate given certain regularity con-

ditions related to the boundedness of the log-likelihood function, we get

$$\begin{aligned}\mathcal{J}_\infty(\theta_0) &= D_\theta^2 \lim_{n\to\infty} \mathcal{E}\left(s_n(\theta_0)\right) \\ &= D_\theta^2 s_\infty(\theta_0, \theta_0)\end{aligned}$$

We've already seen that

$$s_\infty(\theta, \theta_0) < s_\infty(\theta_0, \theta_0)$$

i.e., θ_0 maximizes the limiting objective function. Since there is a unique maximizer, and by the assumption that $s_n(\theta)$ is twice continuously differentiable (which holds in the limit), then $\mathcal{J}_\infty(\theta_0)$ must be negative definite, and therefore of full rank. Therefore the previous inversion is justified, asymptotically, and we have

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\mathcal{J}(\theta^*)^{-1}\sqrt{n}g(\theta_0). \tag{13.4}$$

Now consider $\sqrt{n}g(\theta_0)$. This is

$$\begin{aligned}\sqrt{n}g_n(\theta_0) &= \sqrt{n}D_\theta s_n(\theta) \\ &= \frac{\sqrt{n}}{n}\sum_{t=1}^{n} D_\theta \ln f_t(y_t|x_t, \theta_0) \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n} g_t(\theta_0)\end{aligned}$$

We've already seen that $\mathcal{E}_\theta\left[g_t(\theta)\right] = 0$. As such, it is reasonable to assume that a CLT applies.

Note that $g_n(\theta_0) \xrightarrow{a.s.} 0$, by consistency. To avoid this collapse to a degenerate r.v. (a constant vector) we need to scale by $\sqrt{n}$. A generic CLT states that, for X_n a random vector that satisfies

certain conditions,

$$X_n - E(X_n) \xrightarrow{d} N(0, \lim V(X_n))$$

The "certain conditions" that X_n must satisfy depend on the case at hand. Usually, X_n will be of the form of an average, scaled by $\sqrt{n}$:

$$X_n = \sqrt{n}\frac{\sum_{t=1}^{n} X_t}{n}$$

This is the case for $\sqrt{n}g(\theta_0)$ for example. Then the properties of X_n depend on the properties of the X_t. For example, if the X_t have finite variances and are not too strongly dependent, then a CLT for dependent processes will apply. Supposing that a CLT applies, and noting that $E(\sqrt{n}g_n(\theta_0) = 0$, we get

$$\sqrt{n}g_n(\theta_0) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta_0)\right] \tag{13.5}$$

where

$$\begin{aligned} \mathcal{I}_\infty(\theta_0) &= \lim_{n\to\infty} \mathcal{E}_{\theta_0}\left(n\left[g_n(\theta_0)\right]\left[g_n(\theta_0)\right]'\right) \\ &= \lim_{n\to\infty} V_{\theta_0}\left(\sqrt{n}g_n(\theta_0)\right) \end{aligned}$$

This can also be written as

- $\mathcal{I}_\infty(\theta_0)$ is known as the *information matrix*.
- Combining [13.4] and [13.5], and noting that $\mathcal{J}(\theta^*) \xrightarrow{a.s.} \mathcal{J}_\infty(\theta_0)$, we get

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{a}{\sim} N\left[0, \mathcal{J}_\infty(\theta_0)^{-1}\mathcal{I}_\infty(\theta_0)\mathcal{J}_\infty(\theta_0)^{-1}\right].$$

The MLE estimator is asymptotically normally distributed.

Definition 34. Consistent and asymptotically normal (CAN). An estimator $\hat{\theta}$ of a parameter θ_0 is $\sqrt{n}$-consistent and asymptotically normally distributed if $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, V_\infty\right)$ where V_∞ is a finite positive definite matrix.

There do exist, in special cases, estimators that are consistent such that $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{p} 0$. These are known as *superconsistent* estimators, since in ordinary circumstances with stationary data, $\sqrt{n}$ is the highest factor that we can multiply by and still get convergence to a stable limiting distribution.

Definition 35. Asymptotically unbiased. An estimator $\hat{\theta}$ of a parameter θ_0 is asymptotically unbiased if

$\lim_{n\to\infty} \mathcal{E}_\theta(\hat{\theta}) = \theta$.

Estimators that are CAN are asymptotically unbiased, though not all consistent estimators are asymptotically unbiased. Such cases are unusual, though.

13.5 The information matrix equality

We will show that $\mathcal{J}_\infty(\theta) = -I_\infty(\theta)$. Let $f_t(\theta)$ be short for $f(y_t|x_t,\theta)$

$$\begin{aligned} 1 &= \int f_t(\theta)dy, \text{ so} \\ 0 &= \int D_\theta f_t(\theta)dy \\ &= \int \left(D_\theta \ln f_t(\theta)\right) f_t(\theta)dy \end{aligned}$$

Now differentiate again:

$$\begin{aligned} 0 &= \int \left[D_\theta^2 \ln f_t(\theta)\right] f_t(\theta) dy + \int \left[D_\theta \ln f_t(\theta)\right] D_{\theta'} f_t(\theta) dy \\ &= \mathcal{E}_\theta \left[D_\theta^2 \ln f_t(\theta)\right] + \int \left[D_\theta \ln f_t(\theta)\right] \left[D_{\theta'} \ln f_t(\theta)\right] f_t(\theta) dy \\ &= \mathcal{E}_\theta \left[D_\theta^2 \ln f_t(\theta)\right] + \mathcal{E}_\theta \left[D_\theta \ln f_t(\theta)\right] \left[D_{\theta'} \ln f_t(\theta)\right] \\ &= \mathcal{E}_\theta \left[\mathcal{J}_t(\theta)\right] + \mathcal{E}_\theta \left[g_t(\theta)\right] \left[g_t(\theta)\right]' \end{aligned} \tag{13.6}$$

Now sum over n and multiply by $\frac{1}{n}$

$$\mathcal{E}_\theta \frac{1}{n} \sum_{t=1}^{n} \left[\mathcal{J}_t(\theta)\right] = -\mathcal{E}_\theta \left[\frac{1}{n} \sum_{t=1}^{n} \left[g_t(\theta)\right] \left[g_t(\theta)\right]'\right] \tag{13.7}$$

The scores g_t and g_s are uncorrelated for $t \neq s$, since for $t > s$, $f_t(y_t|y_1, ..., y_{t-1}, \theta)$ has conditioned on prior information, so what was random in s is fixed in t. (This forms the basis for a specification test proposed by White: if the scores appear to be correlated one may question the specification of the model). This allows us to write

$$\mathcal{E}_\theta \left[\mathcal{J}_n(\theta)\right] = -\mathcal{E}_\theta \left(n \left[g(\theta)\right] \left[g(\theta)\right]'\right)$$

since all cross products between different periods expect to zero. Finally take limits, we get

$$\mathcal{J}_\infty(\theta) = -\mathcal{I}_\infty(\theta). \tag{13.8}$$

This holds for all θ, in particular, for θ_0. Using this,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \overset{a.s.}{\to} N\left[0, \mathcal{J}_\infty(\theta_0)^{-1}\mathcal{I}_\infty(\theta_0)\mathcal{J}_\infty(\theta_0)^{-1}\right]$$

simplifies to

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \overset{a.s.}{\to} N\left[0, \mathcal{I}_\infty(\theta_0)^{-1}\right] \tag{13.9}$$

or

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \overset{a.s.}{\to} N\left[0, -\mathcal{J}_\infty(\theta_0)^{-1}\right] \tag{13.10}$$

To estimate the asymptotic variance, we need estimators of $\mathcal{J}_\infty(\theta_0)$ and $\mathcal{I}_\infty(\theta_0)$. We can use

$$\begin{aligned}
\widehat{\mathcal{I}_\infty(\theta_0)} &= \frac{1}{n}\sum_{t=1}^{n} g_t(\hat{\theta})g_t(\hat{\theta})' \\
\widehat{\mathcal{J}_\infty(\theta_0)} &= \mathcal{J}_n(\hat{\theta}).
\end{aligned}$$

as is intuitive if one considers equation 13.7. Note, one can't use

$$\widehat{I_\infty(\theta_0)} = n\left[g_n(\hat{\theta})\right]\left[g_n(\hat{\theta})\right]'$$

to estimate the information matrix. Why not?

From this we see that there are alternative ways to estimate $V_\infty(\theta_0)$ that are all valid. These include

$$\begin{aligned}
\widehat{V_\infty(\theta_0)} &= -\widehat{\mathcal{J}_\infty(\theta_0)}^{-1} \\
\widehat{V_\infty(\theta_0)} &= \widehat{\mathcal{I}_\infty(\theta_0)}^{-1} \\
\widehat{V_\infty(\theta_0)} &= \widehat{\mathcal{J}_\infty(\theta_0)}^{-1}\widehat{\mathcal{I}_\infty(\theta_0)}\widehat{\mathcal{J}_\infty(\theta_0)}^{-1}
\end{aligned}$$

These are known as the *inverse Hessian, outer product of the gradient* (OPG) and *sandwich* estimators, respectively. The sandwich form is the most robust, since it coincides with the covariance estimator of the *quasi*-ML estimator.

With a litte more detail, the methods are:

- The sandwich version:

$$\widehat{V_\infty} = n \left\{ \begin{array}{c} \left\{ \sum_{t=1}^n D_\theta^2 \ln f(y_t|Y_{t-1}, \hat{\theta}) \right\} \times \\ \left\{ \sum_{t=1}^n \left[D_\theta \ln f(y_t|Y_{t-1}, \hat{\theta}) \right] \left[D_\theta \ln f(y_t|Y_{t-1}, \hat{\theta}) \right]' \right\}^{-1} \times \\ \left\{ \sum_{t=1}^n D_\theta^2 \ln f(y_t|Y_{t-1}, \hat{\theta}) \right\} \end{array} \right\}^{-1}$$

- or the inverse of the negative of the Hessian (since the middle and last term cancel, except for a minus sign):

$$\widehat{V_\infty} = \left[-1/n \sum_{t=1}^n D_\theta^2 \ln f(y_t|Y_{t-1}, \hat{\theta}) \right]^{-1},$$

- or the inverse of the outer product of the gradient (since the middle and last cancel except for a minus sign, and the first term converges to minus the inverse of the middle term, which is still inside the overall inverse)

$$\widehat{V_\infty} = \left\{ 1/n \sum_{t=1}^n \left[D_\theta \ln f(y_t|Y_{t-1}, \hat{\theta}) \right] \left[D_\theta \ln f(y_t|Y_{t-1}, \hat{\theta}) \right]' \right\}^{-1}.$$

This simplification is a special result for the MLE estimator - it doesn't apply to GMM estimators in general.

- Asymptotically, if the model is correctly specified, all of these forms converge to the same limit. In small samples they will differ. In particular, there is evidence that the outer product of the gradient formula does not perform very well in small samples (see Davidson and MacKinnon, pg. 477).
- White's *Information matrix test* (Econometrica, 1982) is based upon comparing the two ways to estimate the information matrix: outer product of gradient or negative of the Hessian. If they differ by too much, this is evidence of misspecification of the model.

Example, Coin flipping, again

In section 32 we saw that the MLE for the parameter of a Bernoulli trial, with i.i.d. data, is the sample mean: $\hat{p} = \bar{y}$ (equation 13.1). Now let's find the limiting variance of $\sqrt{n}(\hat{p} - p_0)$. We can do this in a simple way:

$$\begin{aligned}
\lim Var\sqrt{n}(\hat{p} - p_0) &= \lim nVar(\hat{p} - p_0) \\
&= \lim nVar(\hat{p}) \\
&= \lim nVar(\bar{y}) \\
&= \lim nVar\left(\frac{\sum y_t}{n}\right) \\
&= \lim \frac{1}{n}\sum Var(y_t) \text{ (by independence of obs.)} \\
&= \lim \frac{1}{n} nVar(y) \text{ (by identically distributed obs.)} \\
&= Var(y) \\
&= p_0(1 - p_0)
\end{aligned}$$

While that is simple, let's verify this using the methods of Chapter 12 give the same answer. The log-likelihood function is

$$s_n(p) = \frac{1}{n}\sum_{t=1}^{n}\{y_t \ln p + (1-y_t)\ln(1-p)\}$$

so

$$Es_n(p) = p^0 \ln p + \left(1-p^0\right)\ln(1-p)$$

by the fact that the observations are i.i.d. Thus, $s_\infty(p) = p^0 \ln p + (1-p^0)\ln(1-p)$. A bit of calculation shows that

$$D_\theta^2 s_n(p)\big|_{p=p^0} \equiv \mathcal{J}_n(\theta) = \frac{-1}{p^0(1-p^0)},$$

which doesn't depend upon n. By results we've seen on MLE, $\lim Var\sqrt{n}(\hat{p}-p^0) = -\mathcal{J}_\infty^{-1}(p^0)$. And in this case, $-\mathcal{J}_\infty^{-1}(p^0) = p^0(1-p^0)$. So, we get the same limiting variance using both methods.

Exercise 36. For this example, find $\mathcal{I}_\infty(p_0)$.

13.6 The Cramér-Rao lower bound

Theorem 37. [Cramer-Rao Lower Bound] *The limiting variance of a CAN estimator of θ_0, say $\tilde{\theta}$, minus the inverse of the information matrix is a positive semidefinite matrix.*

Proof: Since the estimator is CAN, it is asymptotically unbiased, so

$$\lim_{n\to\infty} \mathcal{E}_\theta(\tilde{\theta}-\theta) = 0$$

Differentiate wrt θ' :

$$\begin{aligned} D_{\theta'} \lim_{n\to\infty} \mathcal{E}_\theta(\tilde{\theta} - \theta) &= \lim_{n\to\infty} \int D_{\theta'} \left[f(Y,\theta) \left(\tilde{\theta} - \theta\right)\right] dy \\ &= 0 \text{ (this is a } K \times K \text{ matrix of zeros).} \end{aligned}$$

Noting that $D_{\theta'} f(Y,\theta) = f(\theta) D_{\theta'} \ln f(\theta)$, we can write

$$\lim_{n\to\infty} \int \left(\tilde{\theta} - \theta\right) f(\theta) D_{\theta'} \ln f(\theta) dy + \lim_{n\to\infty} \int f(Y,\theta) D_{\theta'} \left(\tilde{\theta} - \theta\right) dy = 0.$$

Now note that $D_{\theta'} \left(\tilde{\theta} - \theta\right) = -I_K$, and $\int f(Y,\theta)(-I_K) dy = -I_K$. With this we have

$$\lim_{n\to\infty} \int \left(\tilde{\theta} - \theta\right) f(\theta) D_{\theta'} \ln f(\theta) dy = I_K.$$

Playing with powers of n we get

$$\lim_{n\to\infty} \int \sqrt{n} \left(\tilde{\theta} - \theta\right) \sqrt{n} \underbrace{\frac{1}{n} \left[D_{\theta'} \ln f(\theta)\right]} f(\theta) dy = I_K$$

Note that the bracketed part is just the transpose of the score vector, $g(\theta)$, so we can write

$$\lim_{n\to\infty} \mathcal{E}_\theta \left[\sqrt{n} \left(\tilde{\theta} - \theta\right) \sqrt{n} g(\theta)'\right] = I_K$$

This means that the covariance of the score function with $\sqrt{n} \left(\tilde{\theta} - \theta\right)$, for $\tilde{\theta}$ any CAN estimator, is an

identity matrix. Using this, suppose the variance of $\sqrt{n}\left(\tilde{\theta}-\theta\right)$ tends to $V_\infty(\tilde{\theta})$. Therefore,

$$V_\infty\left[\begin{array}{c}\sqrt{n}\left(\tilde{\theta}-\theta\right)\\ \sqrt{n}g(\theta)\end{array}\right]=\left[\begin{array}{cc}V_\infty(\tilde{\theta}) & I_K\\ I_K & \mathcal{I}_\infty(\theta)\end{array}\right]. \tag{13.11}$$

Since this is a covariance matrix, it is positive semi-definite. Therefore, for any K -vector α,

$$\left[\begin{array}{cc}\alpha' & -\alpha'\mathcal{I}_\infty^{-1}(\theta)\end{array}\right]\left[\begin{array}{cc}V_\infty(\tilde{\theta}) & I_K\\ I_K & \mathcal{I}_\infty(\theta)\end{array}\right]\left[\begin{array}{c}\alpha\\ -\mathcal{I}_\infty(\theta)^{-1}\alpha\end{array}\right]\geq 0.$$

This simplifies to

$$\alpha'\left[V_\infty(\tilde{\theta})-\mathcal{I}_\infty^{-1}(\theta)\right]\alpha\geq 0.$$

Since α is arbitrary, $V_\infty(\tilde{\theta})-\mathcal{I}_\infty^{-1}(\theta)$ is positive semidefinite. This conludes the proof.

This means that $\mathcal{I}_\infty^{-1}(\theta)$ is a *lower bound* for the asymptotic variance of a CAN estimator.

(*Asymptotic efficiency*) Given two CAN estimators of a parameter θ_0, say $\tilde{\theta}$ and $\hat{\theta}$, $\hat{\theta}$ is asymptotically efficient with respect to $\tilde{\theta}$ if $V_\infty(\tilde{\theta})-V_\infty(\hat{\theta})$ is a positive semidefinite matrix.

A direct proof of asymptotic efficiency of an estimator is infeasible, but if one can show that the asymptotic variance is equal to the inverse of the information matrix, then the estimator is asymptotically efficient. In particular, *the MLE is asymptotically efficient with respect to any other CAN estimator.*

13.7 Likelihood ratio-type tests

Suppose we would like to test a set of q possibly nonlinear restrictions $r(\theta) = 0$, where the $q \times k$ matrix $D_{\theta'} r(\theta)$ has rank q. The Wald test can be calculated using the unrestricted model. The score test can be calculated using only the restricted model. The likelihood ratio test, on the other hand, uses both the restricted and the unrestricted estimators. The test statistic is

$$LR = 2\left(\ln L(\hat{\theta}) - \ln L(\tilde{\theta})\right)$$

where $\hat{\theta}$ is the unrestricted estimate and $\tilde{\theta}$ is the restricted estimate. To show that it is asymptotically χ^2, take a second order Taylor's series expansion of $\ln L(\tilde{\theta})$ about $\hat{\theta}$:

$$\ln L(\tilde{\theta}) \simeq \ln L(\hat{\theta}) + \frac{n}{2}\left(\tilde{\theta} - \hat{\theta}\right)' \mathcal{J}(\hat{\theta}) \left(\tilde{\theta} - \hat{\theta}\right)$$

(note, the first order term drops out since $D_\theta \ln L(\hat{\theta}) \equiv 0$ by the fonc and we need to multiply the second-order term by n since $\mathcal{J}(\theta)$ is defined in terms of $\frac{1}{n} \ln L(\theta)$) so

$$LR \simeq -n\left(\tilde{\theta} - \hat{\theta}\right)' \mathcal{J}(\hat{\theta}) \left(\tilde{\theta} - \hat{\theta}\right)$$

As $n \to \infty$, $\mathcal{J}(\hat{\theta}) \to \mathcal{J}_\infty(\theta_0) = -\mathcal{I}(\theta_0)$, by the information matrix equality. So

$$LR \stackrel{a}{=} n\left(\tilde{\theta} - \hat{\theta}\right)' \mathcal{I}_\infty(\theta_0) \left(\tilde{\theta} - \hat{\theta}\right) \tag{13.12}$$

We also have that, from the theory on the asymptotic normality of the MLE and the information matrix equality

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \stackrel{a}{=} \mathcal{I}_\infty(\theta_0)^{-1} n^{1/2} g(\theta_0).$$

An analogous result for the restricted estimator is (this is unproven here, to prove this set up the Lagrangean for MLE subject to $R\beta = r$, and manipulate the first order conditions) :

$$\sqrt{n}\left(\tilde{\theta}-\theta_0\right) \stackrel{a}{=} \mathcal{I}_\infty(\theta_0)^{-1}\left(I_n - R'\left(R\mathcal{I}_\infty(\theta_0)^{-1}R'\right)^{-1} R\mathcal{I}_\infty(\theta_0)^{-1}\right) n^{1/2} g(\theta_0).$$

Combining the last two equations

$$\sqrt{n}\left(\tilde{\theta}-\hat{\theta}\right) \stackrel{a}{=} -n^{1/2}\mathcal{I}_\infty(\theta_0)^{-1}R'\left(R\mathcal{I}_\infty(\theta_0)^{-1}R'\right)^{-1} R\mathcal{I}_\infty(\theta_0)^{-1} g(\theta_0)$$

so, substituting into [13.12]

$$LR \stackrel{a}{=} \left[n^{1/2} g(\theta_0)' \mathcal{I}_\infty(\theta_0)^{-1} R'\right]\left[R\mathcal{I}_\infty(\theta_0)^{-1}R'\right]^{-1}\left[R\mathcal{I}_\infty(\theta_0)^{-1} n^{1/2} g(\theta_0)\right]$$

But since

$$n^{1/2} g(\theta_0) \xrightarrow{d} N\left(0, \mathcal{I}_\infty(\theta_0)\right)$$

the linear function

$$R\mathcal{I}_\infty(\theta_0)^{-1} n^{1/2} g(\theta_0) \xrightarrow{d} N(0, R\mathcal{I}_\infty(\theta_0)^{-1}R').$$

We can see that LR is a quadratic form of this rv, with the inverse of its variance in the middle, so

$$LR \xrightarrow{d} \chi^2(q).$$

Summary of MLE

- Consistent
- Asymptotically normal (CAN)
- Asymptotically efficient
- Asymptotically unbiased
- LR test is available for testing hypothesis
- The presentation is for general MLE: we haven't specified the distribution or the linearity/non-linearity of the estimator

13.8 Examples

ML of Nerlove model, assuming normality

As we saw in Section 4.3, the ML and OLS estimators of β in the linear model $y = X\beta + \epsilon$ coincide when ϵ is assumed to be i.i.d. normally distributed. The Octave script NerloveMLE.m verifies this result, for the basic Nerlove model (eqn. 3.10). The output of the script follows:

```
**************************************************
check MLE with normality, compare to OLS

MLE Estimation Results
```

```
BFGS convergence: Normal convergence

Average Log-L: -0.465806
Observations: 145

estimate      st. err      t-stat     p-value
constant       -3.527       1.689      -2.088       0.037
output          0.720       0.032      22.491       0.000
labor           0.436       0.241       1.808       0.071
fuel            0.427       0.074       5.751       0.000
capital        -0.220       0.318      -0.691       0.490
sig             0.386       0.041       9.290       0.000

Information Criteria
CAIC : 170.9442       Avg. CAIC:   1.1789
BIC : 164.9442        Avg. BIC:   1.1375
AIC : 147.0838        Avg. AIC:   1.0144
*******************************************************
```

Compare the output to that of Nerlove.m , which does OLS. The script also provides a basic example of how to use the MLE estimation routing `mle_results.m`

Example: Binary response models: theory

This section extends the Bernoulli trial model to binary response models with conditioning variables, as such models arise in a variety of contexts.

Assume that

$$\begin{aligned} y^* &= x'\theta - \varepsilon \\ y &= 1(y^* > 0) \\ \varepsilon &\sim N(0,1) \end{aligned}$$

Here, y^* is an unobserved (latent) continuous variable, and y is a binary variable that indicates whether $y^*$is negative or positive. Then the *probit* model results, where $Pr(y = 1|x) = Pr(\varepsilon < x'\theta) = \Phi(x'\theta)$, where

$$\Phi(\bullet) = \int_{-\infty}^{x\beta} (2\pi)^{-1/2} \exp(-\frac{\varepsilon^2}{2}) d\varepsilon$$

is the standard normal distribution function.

The *logit* model results if the errors ϵ are not normal, but rather have a logistic distribution. This distribution is similar to the standard normal, but has fatter tails. The probability has the following parameterization

$$Pr(y = 1|x) = \Lambda(x'\theta) = (1 + \exp(-x'\theta))^{-1}.$$

In general, a binary response model will require that the choice probability be parameterized in some form which could be logit, probit, or something else. For a vector of explanatory variables x, the

response probability will be parameterized in some manner

$$Pr(y = 1|x) = p(x, \theta)$$

Again, if $p(x, \theta) = \Lambda(x'\theta)$, we have a logit model. If $p(x, \theta) = \Phi(x'\theta)$, where $\Phi(\cdot)$ is the standard normal distribution function, then we have a probit model.

Regardless of the parameterization, we are dealing with a Bernoulli density,

$$f_{Y_i}(y_i|x_i) = p(x_i, \theta)^{y_i}(1 - p(x, \theta))^{1-y_i}$$

so as long as the observations are independent, the maximum likelihood (ML) estimator, $\hat{\theta}$, is the maximizer of

$$\begin{aligned} s_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i \ln p(x_i, \theta) + (1 - y_i)\ln\left[1 - p(x_i, \theta)\right]\right) \\ &\equiv \frac{1}{n}\sum_{i=1}^{n} s(y_i, x_i, \theta). \end{aligned} \tag{13.13}$$

Following the above theoretical results, $\hat{\theta}$ tends in probability to the θ_0 that maximizes the uniform almost sure limit of $s_n(\theta)$. Noting that $\mathcal{E}y_i = p(x_i, \theta_0)$, and following a SLLN for i.i.d. processes, $s_n(\theta)$ converges almost surely to the expectation of a representative term $s(y, x, \theta)$. First one can take the expectation conditional on x to get

$$\mathcal{E}_{y|x}\left\{y \ln p(x, \theta) + (1 - y)\ln\left[1 - p(x, \theta)\right]\right\} = p(x, \theta_0)\ln p(x, \theta) + \left[1 - p(x, \theta_0)\right]\ln\left[1 - p(x, \theta)\right].$$

Next taking expectation over x we get the limiting objective function

$$s_\infty(\theta) = \int_{\mathcal{X}} \{p(x,\theta_0) \ln p(x,\theta) + [1 - p(x,\theta_0)] \ln [1 - p(x,\theta)]\} \mu(x) dx, \tag{13.14}$$

where $\mu(x)$ is the (joint - the integral is understood to be multiple, and $\mathcal{X}$ is the support of x) density function of the explanatory variables x. This is clearly continuous in θ, as long as $p(x,\theta)$ is continuous, and if the parameter space is compact we therefore have uniform almost sure convergence. Note that $p(x,\theta)$ is continous for the logit and probit models, for example. The maximizing element of $s_\infty(\theta)$, θ^*, solves the first order conditions

$$\int_{\mathcal{X}} \left\{ \frac{p(x,\theta_0)}{p(x,\theta^*)} \frac{\partial}{\partial \theta} p(x,\theta^*) - \frac{1 - p(x,\theta_0)}{1 - p(x,\theta^*)} \frac{\partial}{\partial \theta} p(x,\theta^*) \right\} \mu(x) dx = 0$$

This is clearly solved by $\theta^* = \theta_0$. Provided the solution is unique, $\hat{\theta}$ is consistent. Question: what's needed to ensure that the solution is unique?

The asymptotic normality theorem tells us that

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1} \mathcal{I}_\infty(\theta^0) \mathcal{J}_\infty(\theta^0)^{-1}\right].$$

In the case of i.i.d. observations $\mathcal{I}_\infty(\theta_0) = \lim_{n\to\infty} Var \sqrt{n} D_\theta s_n(\theta_0)$ is simply the expectation of a typical element of the outer product of the gradient.

- There's no need to subtract the mean, since it's zero, following the f.o.c. in the consistency proof above and the fact that observations are i.i.d.

- The terms in n also drop out by the same argument:

$$\begin{aligned}
\lim_{n\to\infty} Var\sqrt{n}D_\theta s_n(\theta_0) &= \lim_{n\to\infty} Var\sqrt{n}D_\theta \frac{1}{n}\sum_t s(\theta_0) \\
&= \lim_{n\to\infty} Var\frac{1}{\sqrt{n}}D_\theta \sum_t s(\theta_0) \\
&= \lim_{n\to\infty} \frac{1}{n}Var\sum_t D_\theta s(\theta_0) \\
&= \lim_{n\to\infty} VarD_\theta s(\theta_0) \\
&= VarD_\theta s(\theta_0)
\end{aligned}$$

So we get

$$\mathcal{I}_\infty(\theta_0) = \mathcal{E}\left\{\frac{\partial}{\partial\theta}s(y,x,\theta_0)\frac{\partial}{\partial\theta'}s(y,x,\theta_0)\right\}.$$

Likewise,

$$\mathcal{J}_\infty(\theta_0) = \mathcal{E}\frac{\partial^2}{\partial\theta\partial\theta'}s(y,x,\theta_0).$$

Expectations are jointly over y and x, or equivalently, first over y conditional on x, then over x. From above, a typical element of the objective function is

$$s(y,x,\theta_0) = y\ln p(x,\theta_0) + (1-y)\ln\left[1-p(x,\theta_0)\right].$$

Now suppose that we are dealing with a correctly specified logit model:

$$p(x,\theta) = \left(1+\exp(-\mathbf{x}'\theta)\right)^{-1}.$$

We can simplify the above results in this case. We have that

$$\begin{aligned}
\frac{\partial}{\partial\theta}p(x,\theta) &= \left(1+\exp(-\mathbf{x}'\theta)\right)^{-2}\exp(-\mathbf{x}'\theta)\mathbf{x} \\
&= \left(1+\exp(-\mathbf{x}'\theta)\right)^{-1}\frac{\exp(-\mathbf{x}'\theta)}{1+\exp(-\mathbf{x}'\theta)}\mathbf{x} \\
&= p(x,\theta)\left(1-p(x,\theta)\right)\mathbf{x} \\
&= \left(p(x,\theta)-p(x,\theta)^2\right)\mathbf{x}.
\end{aligned}$$

So

$$\begin{aligned}
\frac{\partial}{\partial\theta}s(y,x,\theta_0) &= \left[y-p(x,\theta_0)\right]\mathbf{x} \qquad (13.15)\\
\frac{\partial^2}{\partial\theta\partial\theta'}s(\theta_0) &= -\left[p(x,\theta_0)-p(x,\theta_0)^2\right]\mathbf{x}\mathbf{x}'.
\end{aligned}$$

Taking expectations over y then $\mathbf{x}$ gives

$$\begin{aligned}
\mathcal{I}_\infty(\theta_0) &= \int E_Y\left[y^2-2p(x,\theta_0)p(x,\theta_0)+p(x,\theta_0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx \qquad (13.16)\\
&= \int\left[p(x,\theta_0)-p(x,\theta_0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx. \qquad (13.17)
\end{aligned}$$

where we use the fact that $E_Y(y) = E_Y(y^2) = p(\mathbf{x},\theta_0)$. Likewise,

$$\mathcal{J}_\infty(\theta_0) = -\int\left[p(x,\theta_0)-p(x,\theta_0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx. \qquad (13.18)$$

Note that we arrive at the expected result: the information matrix equality holds (that is, $\mathcal{J}_\infty(\theta_0) =$

$-\mathcal{I}_\infty(\theta_0)$). With this,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta_0)^{-1}\mathcal{I}_\infty(\theta_0)\mathcal{J}_\infty(\theta_0)^{-1}\right]$$

simplifies to

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left[0, -\mathcal{J}_\infty(\theta_0)^{-1}\right]$$

which can also be expressed as

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta_0)^{-1}\right].$$

On a final note, the logit and standard normal CDF's are very similar - the logit distribution is a bit more fat-tailed. While coefficients will vary slightly between the two models, functions of interest such as estimated probabilities $p(x, \hat{\theta})$ will be virtually identical for the two models.

Estimation of the logit model

In this section we will consider maximum likelihood estimation of the logit model for binary 0/1 dependent variables. We will use the BFGS algorithm to find the MLE.

A binary response is a variable that takes on only two values, customarily 0 and 1, which can be thought of as codes for whether or not a condisiton is satisfied. For example, 0=drive to work, 1=take the bus. Often the observed binary variable, say y, is related to an unobserved (latent) continuous varable, say y^*. We would like to know the effect of covariates, x, on y. The model can be represented

as

$$\begin{aligned} y^* &= g(x) - \varepsilon \\ y &= 1(y^* > 0) \\ Pr(y = 1) &= F_\varepsilon[g(x)] \\ &\equiv p(x, \theta) \end{aligned}$$

The log-likelihood function is

$$s_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i \ln p(x_i, \theta) + (1 - y_i) \ln \left[1 - p(x_i, \theta) \right] \right)$$

For the logit model, the probability has the specific form

$$p(x, \theta) = \frac{1}{1 + \exp(-x'\theta)}$$

You should download and examine LogitDGP.m , which generates data according to the logit model, logit.m , which calculates the loglikelihood, and EstimateLogit.m , which sets things up and calls the estimation routine, which uses the BFGS algorithm.

Here are some estimation results with $n = 100$, and the true $\theta = (0, 1)'$.

```
*********************************************
Trial of MLE estimation of Logit model

MLE Estimation Results
BFGS convergence: Normal convergence
```

```
Average Log-L: 0.607063
Observations: 100

          estimate   st. err    t-stat   p-value
constant    0.5400    0.2229    2.4224    0.0154
slope       0.7566    0.2374    3.1863    0.0014

Information Criteria
CAIC : 132.6230
 BIC : 130.6230
 AIC : 125.4127
*********************************************
```

The estimation program is calling `mle_results()`, which in turn calls a number of other routines.

Duration data and the Weibull model

In some cases the dependent variable may be the time that passes between the occurence of two events. For example, it may be the duration of a strike, or the time needed to find a job once one is unemployed. Such variables take on values on the positive real line, and are referred to as duration data.

A *spell* is the period of time between the occurence of initial event and the concluding event. For example, the initial event could be the loss of a job, and the final event is the finding of a new job. The spell is the period of unemployment.

Let t_0 be the time the initial event occurs, and t_1 be the time the concluding event occurs. For simplicity, assume that time is measured in years. The random variable D is the duration of the spell, $D = t_1 - t_0$. Define the density function of D, $f_D(t)$, with distribution function $F_D(t) = \Pr(D < t)$.

Several questions may be of interest. For example, one might wish to know the expected time one has to wait to find a job given that one has already waited s years. The probability that a spell lasts more than s years is

$$\Pr(D > s) = 1 - \Pr(D \leq s) = 1 - F_D(s).$$

The density of D conditional on the spell being longer than s years is

$$f_D(t|D > s) = \frac{f_D(t)}{1 - F_D(s)}.$$

The expectanced additional time required for the spell to end given that is has already lasted s years is the expectation of D with respect to this density, minus s.

$$E = \mathcal{E}(D|D > s) - s = \left(\int_t^\infty z \frac{f_D(z)}{1 - F_D(s)} dz \right) - s$$

To estimate this function, one needs to specify the density $f_D(t)$ as a parametric density, then estimate by maximum likelihood. There are a number of possibilities including the exponential density, the lognormal, *etc.* A reasonably flexible model that is a generalization of the exponential density is the Weibull density

$$f_D(t|\theta) = e^{-(\lambda t)^\gamma} \lambda \gamma (\lambda t)^{\gamma - 1}.$$

According to this model, $\mathcal{E}(D) = \lambda^{-\gamma}$. The log-likelihood is just the product of the log densities.

To illustrate application of this model, 402 observations on the lifespan of dwarf mongooses (see Figure 13.1) in Serengeti National Park (Tanzania) were used to fit a Weibull model. The "spell" in this case is the lifetime of an individual mongoose. The parameter estimates and standard errors are $\hat{\lambda} = 0.559\,(0.034)$ and $\hat{\gamma} = 0.867\,(0.033)$ and the log-likelihood value is -659.3. Figure 13.2 presents fitted life expectancy (expected additional years of life) as a function of age, with 95% confidence bands. The plot is accompanied by a nonparametric Kaplan-Meier estimate of life-expectancy. This nonparametric estimator simply averages all spell lengths greater than age, and then subtracts age. This is consistent by the LLN.

In the figure one can see that the model doesn't fit the data well, in that it predicts life expectancy quite differently than does the nonparametric model. For ages 4-6, the nonparametric estimate is outside the confidence interval that results from the parametric model, which casts doubt upon the parametric model. Mongooses that are between 2-6 years old seem to have a lower life expectancy than is predicted by the Weibull model, whereas young mongooses that survive beyond infancy have a higher life expectancy, up to a bit beyond 2 years. Due to the dramatic change in the death rate as a function of t, one might specify $f_D(t)$ as a mixture of two Weibull densities,

$$f_D(t|\theta) = \delta\left(e^{-(\lambda_1 t)^{\gamma_1}}\lambda_1\gamma_1(\lambda_1 t)^{\gamma_1-1}\right) + (1-\delta)\left(e^{-(\lambda_2 t)^{\gamma_2}}\lambda_2\gamma_2(\lambda_2 t)^{\gamma_2-1}\right).$$

The parameters γ_i and $\lambda_i, i = 1, 2$ are the parameters of the two Weibull densities, and δ is the parameter that mixes the two.

With the same data, θ can be estimated using the mixed model. The results are a log-likelihood = -623.17. Note that a standard likelihood ratio test cannot be used to chose between the two models, since under the null that $\delta = 1$ (single density), the two parameters λ_2 and γ_2 are not identified. It is possible to take this into account, but this topic is out of the scope of this course. Nevertheless, the

Figure 13.1: Dwarf mongooses

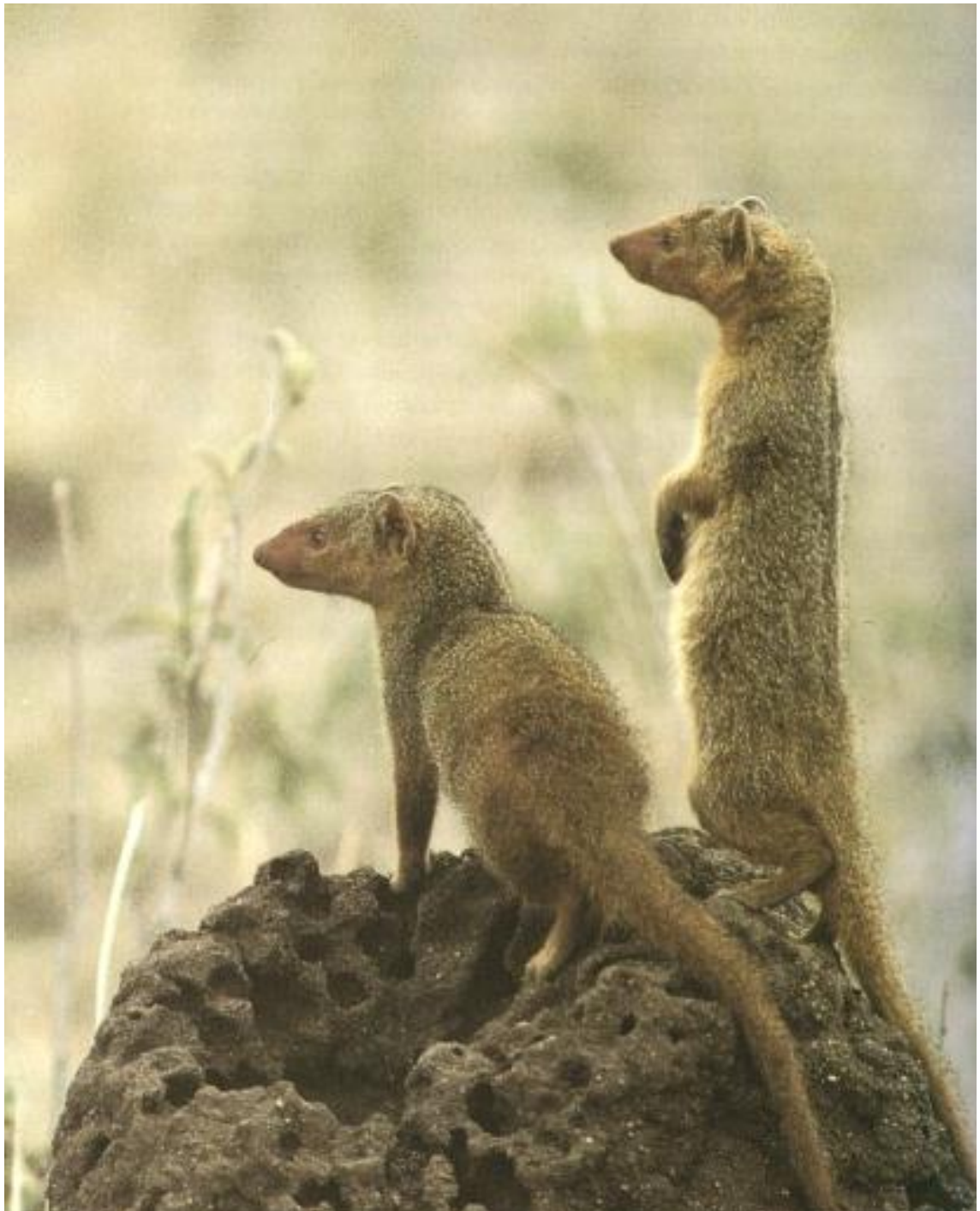
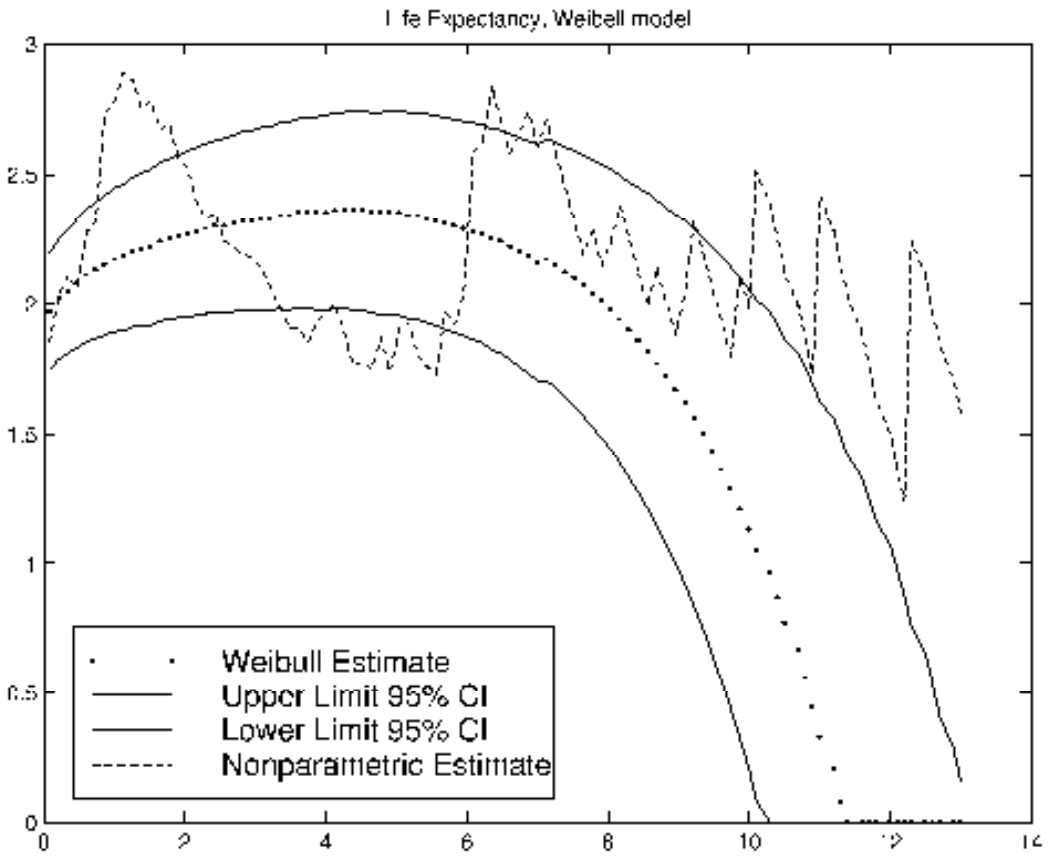


Figure 13.2: Life expectancy of mongooses, Weibull model



improvement in the likelihood function is considerable. The parameter estimates are

Parameter	Estimate	St. Error
λ_1	0.233	0.016
γ_1	1.722	0.166
λ_2	1.731	0.101
γ_2	1.522	0.096
δ	0.428	0.035

Note that the mixture parameter is highly significant. This model leads to the fit in Figure 13.3. Note that the parametric and nonparametric fits are quite close to one another, up to around 6 years. The disagreement after this point is not too important, since less than 5% of mongooses live more than 6 years, which implies that the Kaplan-Meier nonparametric estimate has a high variance (since it's an average of a small number of observations).

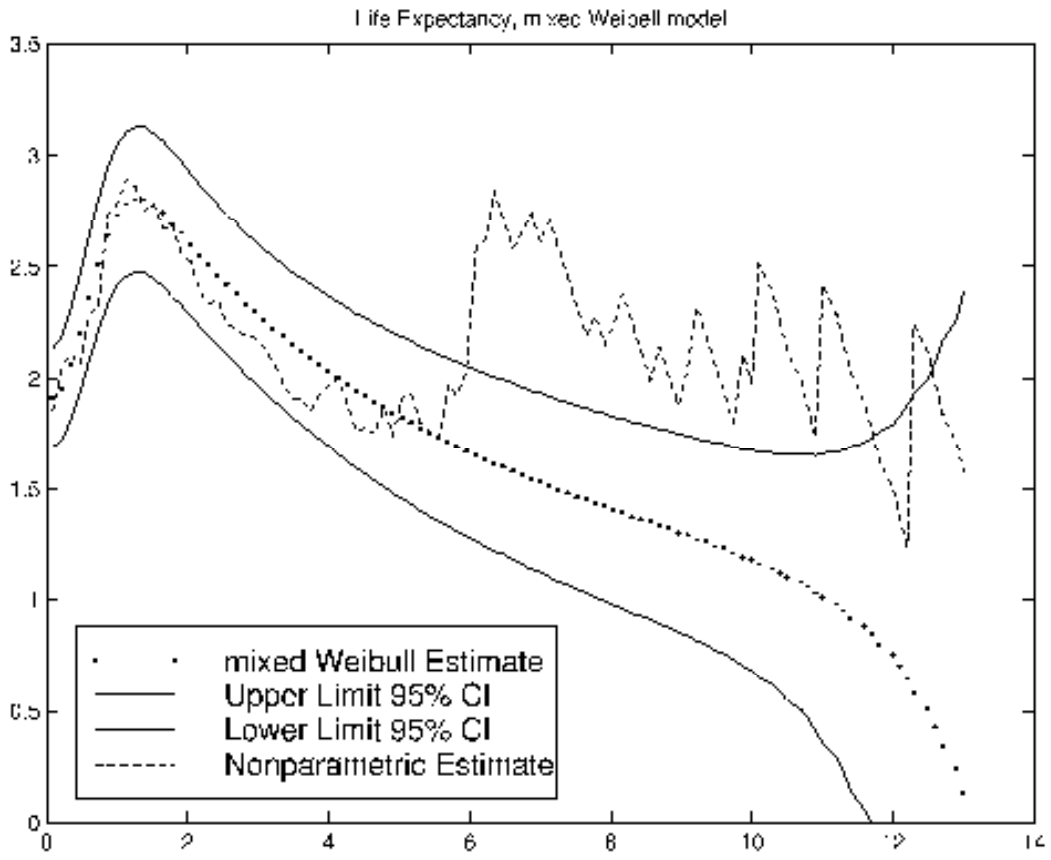
Mixture models are often an effective way to model complex responses, though they can suffer from overparameterization. Alternatives will be discussed later.

For examples of MLE using logit and Poisson model applied to data, see Section **??** in the chapter on Numerical Optimization. You should examine the scripts and run them to see how MLE is actually done.

Estimation of a simple DSGE model

Dynamic stochastic general equilibrium model are widely used tools in macroeconomics. These are models in which current decisions depend upon expectations of future events. An example is the simple real business cycle model discussed in the file rbc.pdf, by Fernández-Villaverde, which is available on

Figure 13.3: Life expectancy of mongooses, mixed Weibull model



the Dynare web page www.dynare.org. The file EstimateRBC_ML.mod shows how this model may be estimated, using maximum likelihood methods. The estimation process involves forming a linear approximation to the true model, which means that the estimator is not actually the true maximum likelihood estimator, it is actually a "quasi-ML" estimator. The quasi-likelihood is computed by putting the linearized model in state-space form, and then computing the likelihood iteratively using Kalman filtering, which relies on the assumption that shocks to the model are normally distributed. State space models and Kalman filtering are introduced in Section 15.5. Once the likelihood function is available, the methods studied in this Chapter may be applied. The intention at the moment is simple to show that ML is an estimation method that may be applied to complicated and more or less realistic economic models. If you play around with the estimation program, you will see that difficulties are encountered with estimating certain parameters. This may be due to an excessive information loss due to the linearization.

13.9 Exercises

1. Consider coin tossing with a single possibly biased coin. The density function for the random variable $y = 1(heads)$ is

$$\begin{aligned} f_Y(y, p_0) &= p_0^y (1 - p_0)^{1-y}, y \in \{0, 1\} \\ &= 0, y \notin \{0, 1\} \end{aligned}$$

Suppose that we have a sample of size n. We know from above that the ML estimator is $\widehat{p_0} = \bar{y}$. We also know from the theory above that

$$\sqrt{n}(\bar{y} - p_0) \overset{a}{\sim} N\left[0, \mathcal{J}_\infty(p_0)^{-1}\mathcal{I}_\infty(p_0)\mathcal{J}_\infty(p_0)^{-1}\right]$$

a) find the analytic expression for $g_t(\theta)$ and show that $\mathcal{E}_\theta[g_t(\theta)] = 0$
b) find the analytical expressions for $\mathcal{J}_\infty(p_0)$ and $\mathcal{I}_\infty(p_0)$ for this problem
c) verify that the result for $\lim Var\sqrt{n}(\hat{p} - p)$ found in section 13.5 is equal to $\mathcal{J}_\infty(p_0)^{-1}\mathcal{I}_\infty(p_0)\mathcal{J}_\infty(p_0)^{-1}$
d) Write an Octave program that does a Monte Carlo study that shows that $\sqrt{n}(\bar{y} - p_0)$ is approximately normally distributed when n is large. Please give me histograms that show the sampling frequency of $\sqrt{n}(\bar{y} - p_0)$ for several values of n.

2. The exponential density is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

Suppose we have an independently and identically distributed sample of size n, $\{x_i\}, i = 1, 2, ..., n$, where each x_i follows this exponential distribution.

(a) write the log likelihood function

(b) compute the maximum likelihood estimator of the parameter λ.

3. Suppose we have an i.i.d. sample of size n from the Poisson density. The Poisson density is $f_y(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$. Verify that the ML estimator is asymptotically distributed as $\sqrt{n}\left(\hat{\lambda} - \lambda_0\right) \xrightarrow{d} N(0, \lambda_0)$, where λ_0 is the true parameter value. Hint: compute the asymptotic variance using $-\mathcal{J}_\infty(\lambda_0)^{-1}$.

4. Consider the model $y_t = x_t'\beta + \alpha\epsilon_t$ where the errors follow the Cauchy (Student-t with 1 degree of freedom) density. So

$$f(\epsilon_t) = \frac{1}{\pi\left(1 + \epsilon_t^2\right)}, -\infty < \epsilon_t < \infty$$

The Cauchy density has a shape similar to a normal density, but with much thicker tails. Thus, extremely small and large errors occur much more frequently with this density than would happen if the errors were normally distributed. Find the score function $g_n(\theta)$ where $\theta = \begin{pmatrix} \beta' & \alpha \end{pmatrix}'$.

5. Consider the model classical linear regression model $y_t = x_t'\beta + \epsilon_t$ where $\epsilon_t \sim IIN(0, \sigma^2)$. Find the score function $g_n(\theta)$ where $\theta = \begin{pmatrix} \beta' & \sigma \end{pmatrix}'$.

6. Compare the first order conditions that define the ML estimators of problems 4 and 5 and interpret the differences. *Why* are the first order conditions that define an efficient estimator different in the two cases?

7. Assume a d.g.p. follows the logit model: $\Pr(y = 1|x) = (1 + exp(-\beta^0 x))^{-1}$.

(a) Assume that $x \sim$ uniform(-a,a). Find the asymptotic distribution of the ML estimator of β^0 (this is a scalar parameter).

(b) Now assume that $x \sim$ uniform(-2a,2a). Again find the asymptotic distribution of the ML estimator of β^0.

(c) Comment on the results

8. There is an ML estimation routine in the provided software that accompanies these notes. Edit (to see what it does) then run the script mle_example.m. Interpret the output.

9. Estimate the simple Nerlove model discussed in section 3.8 by ML, assuming that the errors are i.i.d. $N(0, \sigma^2)$ and compare to the results you get from running Nerlove.m .

10. Using logit.m and EstimateLogit.m as templates, write a function to calculate the probit log likelihood, and a script to estimate a probit model. Run it using data that actually follows a logit model (you can generate it in the same way that is done in the logit example).

11. Study `mle_results.m` to see what it does. Examine the functions that `mle_results.m` calls, and in turn the functions that those functions call. Write a complete description of how the whole chain works.

12. In Subsection 11.4 a model is presented for data on health care usage, along with some Octave scripts. Look at the Poisson estimation results for the OBDV measure of health care use and give an economic interpretation. Estimate Poisson models for the other 5 measures of health care usage, using the provided scripts.

Chapter 14

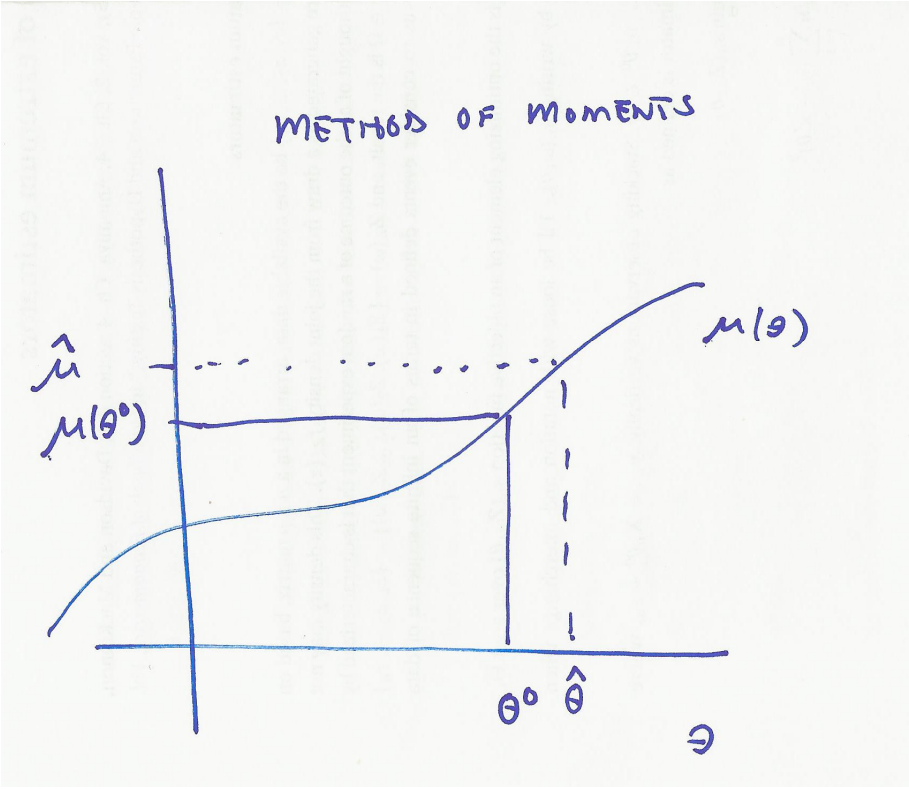
Generalized method of moments

Readings: Hamilton Ch. 14*; Davidson and MacKinnon, Ch. 17 (see pg. 587 for refs. to applications); Newey and McFadden (1994), "Large Sample Estimation and Hypothesis Testing", in *Handbook of Econometrics, Vol. 4*, Ch. 36.

14.1 Motivation

The principle of the method of moments is to set the population moment to the sample moment, then invert to solve for the estimator of the parameter. This is illustrated in Figure 14.1. The sample moment $\hat{\mu}$ will converge to the true moment $\mu(\theta^0)$, so the estimator will converge to the true parameter value. We need that the moment function be invertible.

Figure 14.1: Method of Moments



Sampling from $\chi^2(\theta^0)$

Example 38. (Method of moments, v1) Suppose we draw a random sample of y_t from the $\chi^2(\theta^0)$ distribution. Here, θ^0 is the parameter of interest. The first moment (expectation), μ_1, of a random variable will in general be a function of the parameters of the distribution: $\mu_1 = \mu_1(\theta^0)$.

In this example, if $Y \sim \chi^2(\theta^0)$, then $E(Y) = \theta^0$, so the relationship is the identity function $\mu_1(\theta^0) = \theta^0$, though in general the relationship may be more complicated. The sample first moment is

$$\widehat{\mu_1} = \bar{y} = \sum_{t=1}^{n} y_t/n.$$

Define a moment condition as

$$m_1(\theta) = \mu_1(\theta) - \widehat{\mu_1}.$$

The method of moments principle is to choose the estimator of the parameter to set the estimate of the population moment equal to the sample moment, or equivalently to make the moment condition equal to zero: $m_1(\hat{\theta}) \equiv 0$. Then the equation is solved for the estimator. In this case,

$$m_1(\hat{\theta}) = \hat{\theta} - \sum_{t=1}^{n} y_t/n = 0$$

is solved by $\hat{\theta} = \bar{y}$. Since $\bar{y} = \sum_{t=1}^{n} y_t/n \xrightarrow{p} \theta^0$ by the LLN, the estimator is consistent.

Example 39. (Method of moments, v2) The variance of a $\chi^2(\theta^0)$ r.v. is

$$V(y_t) = E\left(y_t - \theta^0\right)^2 = 2\theta^0.$$

The sample variance is $\hat{V}(y_t) = \frac{\sum_{t=1}^{n}(y_t-\bar{y})^2}{n}$. Define the an alternative moment condition as the population moment minus the sample moment:

$$\begin{aligned} m_2(\theta) &= V(y_t) - \hat{V}(y_t) \\ &= 2\theta - \frac{\sum_{t=1}^{n}(y_t-\bar{y})^2}{n} \end{aligned}$$

We can see that the average moment condition is the average of the contributions

$$m_{2t}(\theta) = V(y_t) - (y_t-\bar{y})^2$$

The MM estimator using the variance would set

$$m_2(\hat{\theta}) = 2\hat{\theta} - \frac{\sum_{t=1}^{n}(y_t-\bar{y})^2}{n} \equiv 0.$$

Again, by the LLN, the sample variance is consistent for the true variance, that is,

$$\frac{\sum_{t=1}^{n}(y_t-\bar{y})^2}{n} \xrightarrow{p} 2\theta^0.$$

So, the estimator is half the sample variance:

$$\hat{\theta} = \frac{1}{2}\frac{\sum_{t=1}^{n}(y_t-\bar{y})^2}{n},$$

This estimator is also consistent for θ^0.

Example 40. Try some MM estimation yourself: here's an Octave script that implements the two MM estimators discussed above: GMM/chi2mm.m

Note that when you run the script, the two estimators give different results. Each of the two estimators is consistent.

- With two moment-parameter equations and only one parameter, we have *overidentification,* which means that we have more information than is strictly necessary for consistent estimation of the parameter.
- The idea behind GMM is to combine information from the two moment-parameter equations to form a new estimator which will be *more efficient,* in general (proof of this below).

Sampling from $t(\theta^0)$

Here's another example based upon the t-distribution. The density function of a t-distributed r.v. Y_t is

$$f_{Y_t}(y_t, \theta^0) = \frac{\Gamma\left[\left(\theta^0 + 1\right)/2\right]}{\left(\pi\theta^0\right)^{1/2}\Gamma\left(\theta^0/2\right)}\left[1 + \left(y_t^2/\theta^0\right)\right]^{-\left(\theta^0+1\right)/2}$$

Given an iid sample of size n, one could estimate θ^0 by maximizing the log-likelihood function

$$\hat{\theta} \equiv \arg\max_{\Theta} \ln \mathcal{L}_n(\theta) = \sum_{t=1}^{n} \ln f_{Y_t}(y_t, \theta)$$

- This approach is attractive since ML estimators are asymptotically efficient. This is because the ML estimator uses all of the available information (e.g., the distribution is fully specified up

to a parameter). Recalling that a distribution is completely characterized by its moments, the ML estimator is interpretable as a GMM estimator that uses *all* of the moments. The method of moments estimator uses only K moments to estimate a $K-$dimensional parameter. Since information is discarded, in general, by the MM estimator, efficiency is lost relative to the ML estimator.

Example 41. (Method of moments). A t-distributed r.v. with density $f_{Y_t}(y_t, \theta^0)$ has mean zero and variance $V(y_t) = \theta^0/(\theta^0 - 2)$ (for $\theta^0 > 2$).

We can define a moment condition as the difference between the theoretical variance and the sample variance: $m_1(\theta) = \theta/(\theta - 2) - 1/n \sum_{t=1}^{n} y_t^2$. When evaluated at the true parameter value θ^0, both $\mathcal{E}_{\theta^0}[m_1(\theta^0)] = 0$.

Choosing $\hat{\theta}$ to set $m_1(\hat{\theta}) \equiv 0$ yields a MM estimator:

$$\hat{\theta} = \frac{2}{1 - \frac{n}{\sum_i y_i^2}} \tag{14.1}$$

This estimator is based on only one moment of the distribution - it uses less information than the ML estimator, so it is intuitively clear that the MM estimator will be inefficient relative to the ML estimator.

Example 42. (Method of moments). An alternative MM estimator could be based upon the fourth moment of the t-distribution. The fourth moment of a t-distributed r.v. is

$$\mu_4 \equiv E(y_t^4) = \frac{3(\theta^0)^2}{(\theta^0 - 2)(\theta^0 - 4)},$$

provided that $\theta^0 > 4$. We can define a second moment condition

$$m_2(\theta) = \frac{3\left(\theta\right)^2}{\left(\theta - 2\right)\left(\theta - 4\right)} - \frac{1}{n}\sum_{t=1}^{n} y_t^4$$

A second, different MM estimator chooses $\hat{\theta}$ to set $m_2(\hat{\theta}) \equiv 0$. If you solve this you'll see that the estimate is different from that in equation 14.1.

This estimator isn't efficient either, since it uses only one moment. A GMM estimator would use the two moment conditions together to estimate the single parameter. The GMM estimator is overidentified, which leads to an estimator which is efficient relative to the just identified MM estimators (more on efficiency later).

14.2 Definition of GMM estimator

For the purposes of this course, the following definition of the GMM estimator is sufficiently general:

Definition 43. The GMM estimator of the k -dimensional parameter vector θ^0, $\hat{\theta} \equiv \arg\min_\Theta s_n(\theta) \equiv m_n(\theta)' W_n m_n(\theta)$, where $m_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} m_t(\theta)$ is a g-vector, $g \geq k$, with $\mathcal{E}_\theta m(\theta) = 0$, and W_n converges almost surely to a finite $g \times g$ symmetric positive definite matrix W_∞.

What's the reason for using GMM if MLE is asymptotically efficient?

- Robustness: GMM is based upon a limited set of moment conditions. For consistency, only these moment conditions need to be correctly specified, whereas MLE in effect requires correct specification of *every conceivable* moment condition. GMM is *robust with respect to distributional*

misspecification. The price for robustness is usually a loss of efficiency with respect to the MLE estimator. Keep in mind that the true distribution is not known so if we erroneously specify a distribution and estimate by MLE, the estimator will be inconsistent in general (not always).

- Feasibility: in some cases the MLE estimator is not available, because we are not able to deduce or compute the likelihood function. More on this in the section on simulation-based estimation. The GMM estimator may still be feasible even though MLE is not available.

Example 44. The Octave script GMM/chi2gmm.m implements GMM using the same χ^2 data as was using in Example 40, above. The two moment conditions, based on the sample mean and sample variance are combined. The weight matrix is an identity matrix, I_2. In Octave, type "help gmm_estimate" to get more information on how the GMM estimation routine works.

14.3 Consistency

We simply assume that the assumptions of Theorem 29 hold, so the GMM estimator is strongly consistent. The main requirement is that the moment conditions have mean zero at the true paraeter value, θ^0. This will be the case if our moment conditions are correctly specified. With this, it is clear that the minimum of the limiting objective function occurs at the true parameter value. The only assumption that warrants additional comments is that of identification. In Theorem 29, the third assumption reads: (c) *Identification:* $s_\infty(\cdot)$ has a unique global maximum at θ^0, *i.e.*, $s_\infty(\theta^0) > s_\infty(\theta)$, $\forall \theta \neq \theta^0$. Taking the case of a quadratic objective function $s_n(\theta) = m_n(\theta)' W_n m_n(\theta)$, first consider $m_n(\theta)$.

- Applying a uniform law of large numbers, we get $m_n(\theta) \overset{a.s.}{\to} m_\infty(\theta)$.

- Since $\mathcal{E}_{\theta^0} m_n(\theta^0) = 0$ by assumption, $m_\infty(\theta^0) = 0$.
- Since $s_\infty(\theta^0) = m_\infty(\theta^0)' W_\infty m_\infty(\theta^0) = 0$, in order for asymptotic identification, we need that $m_\infty(\theta) \neq 0$ for $\theta \neq \theta^0$, for at least some element of the vector. There can be no other parameter value that sets the moment conditions to zero (at least, in the limit). This and the assumption that $W_n \xrightarrow{a.s.} W_\infty$, a finite positive $g \times g$ definite $g \times g$ matrix guarantee that θ^0 is asymptotically identified.
- Note that asymptotic identification does not rule out the possibility of lack of identification for a given data set - there may be multiple minimizing solutions in finite samples.

Example 45. Increase n in the Octave script GMM/chi2gmm.m to see evidence of the consistency of the GMM estimator.

14.4 Asymptotic normality

We also simply assume that the conditions of Theorem 31 hold, so we will have asymptotic normality. However, we do need to find the structure of the asymptotic variance-covariance matrix of the estimator. From Theorem 31, we have

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1} \mathcal{I}_\infty(\theta^0) \mathcal{J}_\infty(\theta^0)^{-1}\right]$$

where $\mathcal{J}_\infty(\theta^0)$ is the almost sure limit of $\frac{\partial^2}{\partial\theta\partial\theta'} s_n(\theta)$ when evaluated at θ^0 and

$$\mathcal{I}_\infty(\theta^0) = \lim_{n\to\infty} Var \sqrt{n} \frac{\partial}{\partial\theta} s_n(\theta^0).$$

We need to determine the form of these matrices given the objective function $s_n(\theta) = m_n(\theta)' W_n m_n(\theta)$.

Now using the product rule from the introduction,

$$\frac{\partial}{\partial \theta} s_n(\theta) = 2 \left[\frac{\partial}{\partial \theta} m_n^{'} (\theta) \right] W_n m_n (\theta)$$

(this is analogous to $\frac{\partial}{\partial \beta} \beta' X' X \beta = 2X'X\beta$ which appears when computing the first order conditions for the OLS estimator)

Define the $k \times g$ matrix

$$D_n(\theta) \equiv \frac{\partial}{\partial \theta} m_n' (\theta) ,$$

so:

$$\frac{\partial}{\partial \theta} s(\theta) = 2D(\theta) W m (\theta) . \tag{14.2}$$

(Note that $s_n(\theta)$, $D_n(\theta)$, W_n and $m_n(\theta)$ all depend on the sample size n, but it is omitted to unclutter the notation).

To take second derivatives, let D_i be the $i-$ th row of $D(\theta)$. This is a $1 \times G$ row vector. Using the product rule (25.1),

$$\begin{aligned} \frac{\partial^2}{\partial \theta' \partial \theta_i} s(\theta) &= \frac{\partial}{\partial \theta'} 2D_i(\theta) W m (\theta) \\ &= 2D_i W D' + 2m' W \left[\frac{\partial}{\partial \theta'} D_i' \right] \end{aligned}$$

Note that the first term contains a D', which appears due to $\frac{\partial}{\partial \theta'} m_n (\theta)$. When evaluating the second term:

$$2m(\theta)' W \left[\frac{\partial}{\partial \theta'} D(\theta)_i' \right]$$

(where the dependence of D upon θ is emphasized) at θ^0, assume that $\frac{\partial}{\partial \theta'} D(\theta)_i'$ satisfies a LLN (it is an average), so that it converges almost surely to a finite limit. In this case, we have

$$2m(\theta^0)'W \left[\frac{\partial}{\partial \theta'} D(\theta^0)_i' \right] \xrightarrow{a.s.} 0,$$

because $m(\theta^0) = o_p(1)$ and $W \xrightarrow{a.s.} W_\infty$.

Stacking these results over the k rows of D, we get

$$\lim \frac{\partial^2}{\partial \theta \partial \theta'} s_n(\theta^0) = \mathcal{J}_\infty(\theta^0) = 2D_\infty W_\infty D_\infty', a.s.,$$

where we define $\lim D = D_\infty$, $a.s.$, and $\lim W = W_\infty$, a.s. (we assume a LLN holds).

With regard to $\mathcal{I}_\infty(\theta^0)$, following equation 14.2, and noting that the scores have mean zero at θ^0 (since $\mathcal{E}m(\theta^0) = 0$ by assumption), we have

$$\begin{aligned}
\mathcal{I}_\infty(\theta^0) &= \lim_{n \to \infty} Var \sqrt{n} \frac{\partial}{\partial \theta} s_n(\theta^0) \\
&= \lim_{n \to \infty} \mathcal{E} 4nDWm(\theta^0)m(\theta^0)'WD' \\
&= \lim_{n \to \infty} \mathcal{E} 4DW \left\{ \sqrt{n} m(\theta^0) \right\} \left\{ \sqrt{n} m(\theta^0)' \right\} WD'
\end{aligned}$$

Now, given that $m(\theta^0)$ is an average of centered (mean-zero) quantities, it is reasonable to expect a CLT to apply, after multiplication by $\sqrt{n}$. Assuming this,

$$\sqrt{n} m(\theta^0) \xrightarrow{d} N(0, \Omega_\infty), \tag{14.3}$$

where

$$\Omega_\infty = \lim_{n\to\infty} \mathcal{E}\left[nm(\theta^0)m(\theta^0)'\right].$$

Using this, and the last equation, we get

$$\mathcal{I}_\infty(\theta^0) = 4D_\infty W_\infty \Omega_\infty W_\infty D_\infty'$$

Using these results, the asymptotic normality theorem (31) gives us

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D_\infty' \left(D_\infty W_\infty D_\infty'\right)^{-1}\right],$$

the asymptotic distribution of the GMM estimator for arbitrary weighting matrix W_n. Note that for J_∞ to be positive definite, D_∞ must have full row rank, $\rho(D_\infty) = k$. This is related to identification. If the rows of $m_n(\theta)$ were not linearly independent of one another, then neither D_n nor D_∞ would have full row rank. Identification plus two times differentiability of the objective function lead to J_∞ being positive definite.

There are two things that affect the asymptotic variance:

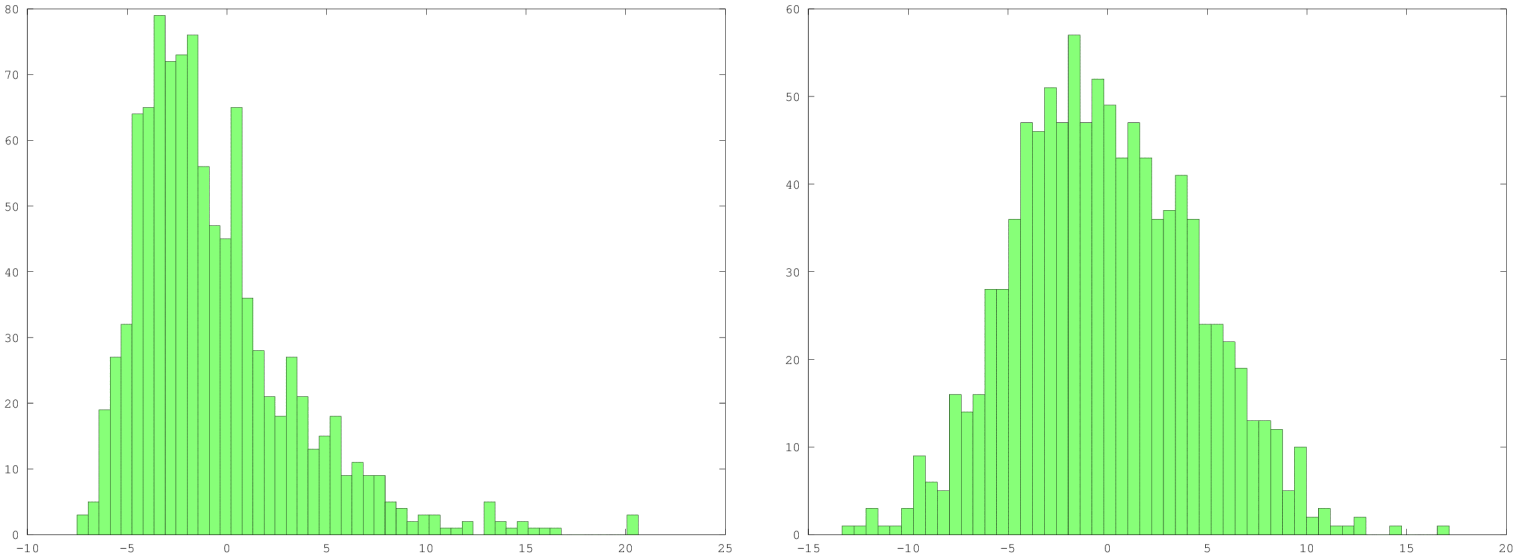
- the choice of the moment conditions, $m_n(\theta)$, which determines both D_∞ and Ω_∞
- the choice of the weight matrix W_n, which determines W_∞

We would probably like to know how to choose both $m_n(\theta)$ and W_n so that the asymptotic variance is a small as possible.

Example 46. The Octave script GMM/AsymptoticNormalityGMM.m does a Monte Carlo of the GMM estimator for the χ^2 data. Histograms for 1000 replications of $\sqrt{n}\left(\hat{\theta} - \theta^0\right)$ are given in Figure

Figure 14.2: Asymptotic Normality of GMM estimator, χ^2 example

(a) $n = 10$ (b) $n = 1000$



14.2. On the left are results for $n = 10$, on the right are results for $n = 1000$. Note that the two distributions are more or less centered at 0. The distribution for the small sample size is somewhat asymmetric, which shows that the small sample distribution may be poorly approximated by the asymptotic distribution. This has mostly disappeared for the larger sample size.

14.5 Choosing the weighting matrix

W is a *weighting matrix,* which determines the relative importance of violations of the individual moment conditions. For example, if we are much more sure of the first moment condition, which is

based upon the variance, than of the second, which is based upon the fourth moment, we could set

$$W = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

with a much larger than b. In this case, errors in the second moment condition have less weight in the objective function.

- Since moments are not independent, in general, we should expect that there be a correlation between the moment conditions, so it may not be desirable to set the off-diagonal elements to 0. W may be a random, data dependent matrix.
- We have already seen that the choice of W will influence the asymptotic distribution of the GMM estimator. Since the GMM estimator is already inefficient w.r.t. MLE, we might like to choose the W matrix to make the GMM estimator efficient *within the class of GMM estimators* defined by $m_n(\theta)$.
- To provide a little intuition, consider the linear model $y = \mathbf{x}'\beta + \varepsilon$, where $\varepsilon \sim N(0, \Omega)$. That is, he have heteroscedasticity and autocorrelation.
- Let P be the Cholesky factorization of Ω^{-1}, e.g, $P'P = \Omega^{-1}$.
- Then the model $Py = P\mathbf{X}\beta + P\varepsilon$ satisfies the classical assumptions of homoscedasticity and nonautocorrelation, because $V(P\varepsilon) = PV(\varepsilon)P' = P\Omega P' = P(P'P)^{-1}P' = PP^{-1}\left(P'\right)^{-1}P' = I_n$. (Note: we use $(AB)^{-1} = B^{-1}A^{-1}$ for A, B both nonsingular). This means that the transformed model is efficient.

- The OLS estimator of the model $Py = P\mathbf{X}\beta + P\varepsilon$ minimizes the objective function $(y - \mathbf{X}\beta)'\Omega^{-1}(y - \mathbf{X}\beta)$. Interpreting $(y - \mathbf{X}\beta) = \varepsilon(\beta)$ as moment conditions (note that they do have zero expectation when evaluated at β^0), the optimal weighting matrix is seen to be the inverse of the covariance matrix of the moment conditions. This result carries over to GMM estimation. (Note: this presentation of GLS is not a GMM estimator as defined above, because the number of moment conditions here is equal to the sample size, n. Later we'll see that GLS can be put into the GMM framework defined above).

Theorem 47. *If $\hat{\theta}$ is a GMM estimator that minimizes $m_n(\theta)'W_n m_n(\theta)$, the asymptotic variance of $\hat{\theta}$ will be minimized by choosing W_n so that $W_n \xrightarrow{a.s} W_\infty = \Omega_\infty^{-1}$, where $\Omega_\infty = \lim_{n\to\infty} \mathcal{E}\left[nm(\theta^0)m(\theta^0)'\right]$.*

Proof: *For $W_\infty = \Omega_\infty^{-1}$, the asymptotic variance*

$$\left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D_\infty' \left(D_\infty W_\infty D_\infty'\right)^{-1}$$

simplifies to $\left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}$. Now, let A be the difference between the general form and the simplified form:

$$A = \left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D_\infty' \left(D_\infty W_\infty D_\infty'\right)^{-1} \; - \; \left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}$$

Set $B = \left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty - \left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1} D_\infty \Omega_\infty^{-1}$. One can show that $A = B\Omega_\infty B'$. This is a quadratic form in a p.d. matrix, so it is p.s.d., which concludes the proof.

The result

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}\right] \tag{14.4}$$

allows us to treat

$$\hat{\theta} \approx N\left(\theta^0, \frac{\left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}}{n}\right),$$

where the $\approx$ means "approximately distributed as." To operationalize this we need estimators of D_∞ and Ω_∞.

- The obvious estimator of $\widehat{D_\infty}$ is simply $\frac{\partial}{\partial \theta} m_n'\left(\hat{\theta}\right)$, which is consistent by the consistency of $\hat{\theta}$, assuming that $\frac{\partial}{\partial \theta} m_n'$ is continuous in θ. Stochastic equicontinuity results can give us this result even if $\frac{\partial}{\partial \theta} m_n'$ is not continuous.

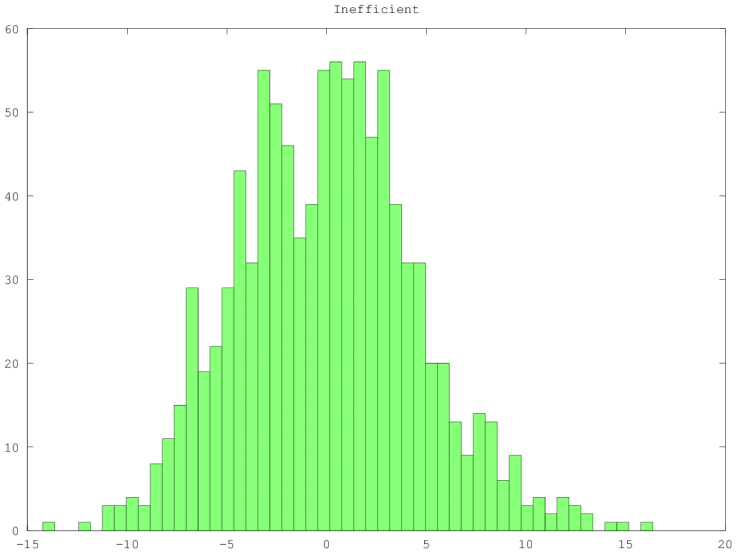
Example 48. To see the effect of using an efficient weight matrix, consider the Octave script GMM/EfficientGMM.m. This modifies the previous Monte Carlo for the χ^2 data. This new Monte Carlo computes the GMM estimator in two ways:
1) based on an identity weight matrix
2) using an estimated optimal weight matrix. The estimated efficient weight matrix is computed as the inverse of the estimated covariance of the moment conditions, using the inefficient estimator of the first step. See the next section for more on how to do this.
Figure 14.3 shows the results, plotting histograms for 1000 replications of $\sqrt{n}\left(\hat{\theta} - \theta^0\right)$. Note that the use of the estimated efficient weight matrix leads to much better results in this case. This is a simple case where it is possible to get a good estimate of the efficient weight matrix. This is not always so. See the next section.

14.6 Estimation of the variance-covariance matrix

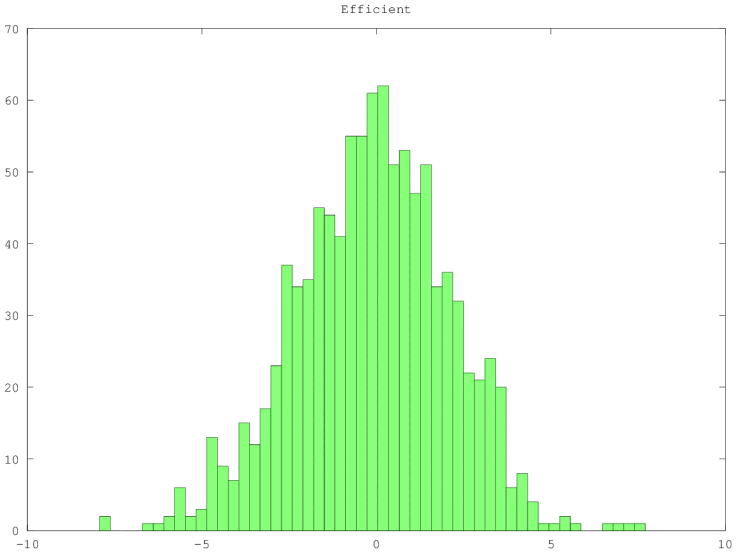
(See Hamilton Ch. 10, pp. 261-2 and 280-84)*.

Figure 14.3: Inefficient and Efficient GMM estimators, χ^2 data

(a) inefficient



(b) efficient



In the case that we wish to use the optimal weighting matrix, we need an estimate of Ω_∞, the limiting variance-covariance matrix of $\sqrt{n}m_n(\theta^0)$. While one could think of estimating Ω_∞ parametrically (along the lines of feasible GLE) we in general have little information upon which to base a parametric specification, so nonparametric estimation is normally used. In general, we expect that:

- m_t will be autocorrelated ($\Gamma_{ts} = \mathcal{E}(m_t m_{t-s}') \neq 0$). Note that this autocovariance will not depend on t if the moment conditions are covariance stationary.
- contemporaneously correlated, since the individual moment conditions will not in general be independent of one another ($\mathcal{E}(m_{it}m_{jt}) \neq 0$).
- and have different variances ($\mathcal{E}(m_{it}^2) = \sigma_{it}^2$).

Since we need to estimate so many components if we are to take the parametric approach, it is unlikely that we would arrive at a correct parametric specification. For this reason, research has focused on consistent nonparametric estimators of Ω_∞.

Henceforth we assume that m_t is covariance stationary (the covariance between m_t and m_{t-s} does not depend on t). Define the $v-th$ autocovariance of the moment conditions $\Gamma_v = \mathcal{E}(m_t m_{t-s}')$.

Exercise 49. Show that $\mathcal{E}(m_t m_{t+s}') = \Gamma_v'$.

Recall that m_t and m are functions of θ, so for now assume that we have some consistent estimator

of θ^0, so that $\hat{m}_t = m_t(\hat{\theta})$. Now

$$\begin{aligned}
\Omega_n &= \mathcal{E}\left[nm(\theta^0)m(\theta^0)'\right] = \mathcal{E}\left[n\left(1/n\sum_{t=1}^{n} m_t\right)\left(1/n\sum_{t=1}^{n} m_t'\right)\right] \\
&= \mathcal{E}\left[1/n\left(\sum_{t=1}^{n} m_t\right)\left(\sum_{t=1}^{n} m_t'\right)\right] \\
&= \Gamma_0 + \frac{n-1}{n}\left(\Gamma_1 + \Gamma_1'\right) + \frac{n-2}{n}\left(\Gamma_2 + \Gamma_2'\right)\cdots + \frac{1}{n}\left(\Gamma_{n-1} + \Gamma_{n-1}'\right)
\end{aligned}$$

A natural, consistent estimator of Γ_v is

$$\widehat{\Gamma_v} = 1/n \sum_{t=v+1}^{n} \hat{m}_t \hat{m}_{t-v}'.$$

(you might use $n - v$ in the denominator instead). So, a natural, but inconsistent, estimator of Ω_∞ would be

$$\begin{aligned}
\hat{\Omega} &= \widehat{\Gamma_0} + \frac{n-1}{n}\left(\widehat{\Gamma_1} + \widehat{\Gamma_1'}\right) + \frac{n-2}{n}\left(\widehat{\Gamma_2} + \widehat{\Gamma_2'}\right) + \cdots + \left(\widehat{\Gamma_{n-1}} + \widehat{\Gamma_{n-1}'}\right) \\
&= \widehat{\Gamma_0} + \sum_{v=1}^{n-1} \frac{n-v}{n}\left(\widehat{\Gamma_v} + \widehat{\Gamma_v'}\right).
\end{aligned}$$

This estimator is inconsistent in general, since the number of parameters to estimate is more than the number of observations, and increases more rapidly than n, so information does not build up as $n \to \infty$.

On the other hand, supposing that Γ_v tends to zero sufficiently rapidly as v tends to ∞, a modified

estimator

$$\hat{\Omega} = \widehat{\Gamma_0} + \sum_{v=1}^{q(n)} \left(\widehat{\Gamma_v} + \widehat{\Gamma_v'} \right),$$

where $q(n) \xrightarrow{p} \infty$ as $n \to \infty$ will be consistent, provided $q(n)$ grows sufficiently slowly. The term $\frac{n-v}{n}$ can be dropped because $q(n)$ must be $o_p(n)$. This allows information to accumulate at a rate that satisfies a LLN. A disadvantage of this estimator is that it may not be positive definite. This could cause one to calculate a negative χ^2 statistic, for example!

- Note: the formula for $\hat{\Omega}$ requires an estimate of $m(\theta^0)$, which in turn requires an estimate of θ, which is based upon an estimate of Ω! The solution to this circularity is to set the weighting matrix W arbitrarily (for example to an identity matrix), obtain a first consistent but inefficient estimate of θ^0, then use this estimate to form $\hat{\Omega}$, then re-estimate θ^0. The process can be iterated until neither $\hat{\Omega}$ nor $\hat{\theta}$ change appreciably between iterations.

Newey-West covariance estimator

The Newey-West estimator (*Econometrica*, 1987) solves the problem of possible nonpositive definiteness of the above estimator. Their estimator is

$$\hat{\Omega} = \widehat{\Gamma_0} + \sum_{v=1}^{q(n)} \left[1 - \frac{v}{q+1} \right] \left(\widehat{\Gamma_v} + \widehat{\Gamma_v'} \right).$$

This estimator is p.d. by construction. The condition for consistency is that $n^{-1/4} q \to 0$. Note that this is a very slow rate of growth for q. This estimator is nonparametric - we've placed no parametric restrictions on the form of Ω. It is an example of a *kernel* estimator.

In a more recent paper, Newey and West (*Review of Economic Studies*, 1994) use *pre-whitening* before applying the kernel estimator. The idea is to fit a VAR model to the moment conditions. It is expected that the residuals of the VAR model will be more nearly white noise, so that the Newey-West covariance estimator might perform better with short lag lengths..

The VAR model is

$$\hat{m}_t = \Theta_1 \hat{m}_{t-1} + \cdots + \Theta_p \hat{m}_{t-p} + u_t$$

This is estimated, giving the residuals $\hat{u}_t$. Then the Newey-West covariance estimator is applied to these pre-whitened residuals, and the covariance Ω is estimated combining the fitted VAR

$$\widehat{\hat{m}_t} = \widehat{\Theta_1} \hat{m}_{t-1} + \cdots + \widehat{\Theta_p} \hat{m}_{t-p}$$

with the kernel estimate of the covariance of the u_t. See Newey-West for details.

- I have a program that does this if you're interested.

14.7 Estimation using conditional moments

So far, the moment conditions have been presented as unconditional expectations. One common way of defining unconditional moment conditions is based upon conditional moment conditions.

Suppose that a random variable Y has zero expectation conditional on the random variable X

$$\mathcal{E}_{Y|X} Y = \int Y f(Y|X) dY = 0$$

Then the unconditional expectation of the product of Y and a function $g(X)$ of X is also zero. The

unconditional expectation is

$$\mathcal{E}Yg(X) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} Yg(X)f(Y,X)dY \right) dX.$$

This can be factored into a conditional expectation and an expectation w.r.t. the marginal density of X :

$$\mathcal{E}Yg(X) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} Yg(X)f(Y|X)dY \right) f(X)dX.$$

Since $g(X)$ doesn't depend on Y it can be pulled out of the integral

$$\mathcal{E}Yg(X) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} Yf(Y|X)dY \right) g(X)f(X)dX.$$

But the term in parentheses on the rhs is zero by assumption, so

$$\mathcal{E}Yg(X) = 0$$

as claimed.

This is important econometrically, since models often imply restrictions on conditional moments. Suppose a model tells us that the function $K(y_t, x_t)$ has expectation, conditional on the information set I_t, equal to $k(x_t, \theta)$,

$$\mathcal{E}_\theta K(y_t, x_t)|I_t = k(x_t, \theta).$$

- For example, in the context of the classical linear model $y_t = x_t'\beta + \varepsilon_t$, we can set $K(y_t, x_t) = y_t$ so that $k(x_t, \theta) = x_t'\beta$.

With this, the error function

$$\epsilon_t(\theta) = K(y_t, x_t) - k(x_t, \theta)$$

has conditional expectation equal to zero

$$\mathcal{E}_\theta \epsilon_t(\theta) | I_t = 0.$$

This is a scalar moment condition, which isn't sufficient to identify a K -dimensional parameter θ ($K > 1$). However, the above result allows us to form various unconditional expectations

$$m_t(\theta) = Z(w_t)\epsilon_t(\theta)$$

where $Z(w_t)$ is a $g \times 1$-vector valued function of w_t and w_t is a set of variables drawn from the information set I_t. The $Z(w_t)$ are *instrumental variables.* We now have g moment conditions, so as long as $g > K$ the necessary condition for identification holds.

One can form the $n \times g$ matrix

$$\begin{aligned} Z_n &= \begin{bmatrix} Z_1(w_1) & Z_2(w_1) & \cdots & Z_g(w_1) \\ Z_1(w_2) & Z_2(w_2) & & Z_g(w_2) \\ \vdots & & & \vdots \\ Z_1(w_n) & Z_2(w_n) & \cdots & Z_g(w_n) \end{bmatrix} \\ &= \begin{bmatrix} Z_1' \\ Z_2' \\ \\ Z_n' \end{bmatrix} \end{aligned}$$

With this we can form the g moment conditions

$$m_n(\theta) = \frac{1}{n} Z_n' \begin{bmatrix} \epsilon_1(\theta) \\ \epsilon_2(\theta) \\ \vdots \\ \epsilon_n(\theta) \end{bmatrix}$$

Define the vector of error functions

$$h_n(\theta) = \begin{bmatrix} \epsilon_1(\theta) \\ \epsilon_2(\theta) \\ \vdots \\ \epsilon_n(\theta) \end{bmatrix}$$

With this, we can write

$$\begin{aligned} m_n(\theta) &= \frac{1}{n} Z_n' h_n(\theta) \\ &= \frac{1}{n} \sum_{t=1}^{n} Z_t h_t(\theta) \\ &= \frac{1}{n} \sum_{t=1}^{n} m_t(\theta) \end{aligned}$$

where $Z_{(t,\cdot)}$ is the t^{th} row of Z_n. This fits the previous treatment.

14.8 A specification test

The first order conditions for minimization, using the an estimate of the optimal weighting matrix, are

$$\frac{\partial}{\partial\theta}s(\hat{\theta}) = 2\left[\frac{\partial}{\partial\theta}m_n^{'}\left(\hat{\theta}\right)\right]\hat{\Omega}^{-1}m_n\left(\hat{\theta}\right) \equiv 0$$

or

$$D(\hat{\theta})\hat{\Omega}^{-1}m_n(\hat{\theta}) \equiv 0$$

Consider a Taylor expansion of $m(\hat{\theta})$ about the true parameter value:

$$m(\hat{\theta}) = m_n(\theta^0) + D_n'(\theta^*)\left(\hat{\theta} - \theta^0\right) \tag{14.5}$$

where θ^* is between $\hat{\theta}$ and θ^0. Multiplying by $D(\hat{\theta})\hat{\Omega}^{-1}$ we obtain

$$D(\hat{\theta})\hat{\Omega}^{-1}m(\hat{\theta}) = D(\hat{\theta})\hat{\Omega}^{-1}m_n(\theta^0) + D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\left(\hat{\theta} - \theta^0\right)$$

The lhs is zero, so

$$D(\hat{\theta})\hat{\Omega}^{-1}m_n(\theta^0) = -\left[D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right]\left(\hat{\theta} - \theta^0\right)$$

or

$$\left(\hat{\theta} - \theta^0\right) = -\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1}D(\hat{\theta})\hat{\Omega}^{-1}m_n(\theta^0)$$

With this, and taking into account the original expansion (equation 14.5), we get

$$\sqrt{n}m(\hat{\theta}) = \sqrt{n}m_n(\theta^0) - \sqrt{n}D_n'(\theta^*)\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1}D(\hat{\theta})\hat{\Omega}^{-1}m_n(\theta^0).$$

With some factoring, this last can be written as

$$\sqrt{n}m(\hat{\theta}) = \left(\hat{\Omega}^{1/2} - D_n'(\theta^*)\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1}D(\hat{\theta})\hat{\Omega}^{-1/2}\right)\left(\sqrt{n}\hat{\Omega}^{-1/2}m_n(\theta^0)\right)$$

(verify it by multiplying out the last expression. Also, a note: the matrix square root of a matrix A is any matrix $A^{1/2}$ such that $A = A^{1/2}A^{1/2}$. Any positive definite matrix has an invertible matrix square root.)

Next, multiply by $\hat{\Omega}^{-1/2}$ to get

$$\sqrt{n}\hat{\Omega}^{-1/2}m(\hat{\theta}) = \left(I_g - \hat{\Omega}^{-1/2}D_n'(\theta^*)\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1}D(\hat{\theta})\hat{\Omega}^{-1/2}\right)\left(\sqrt{n}\hat{\Omega}^{-1/2}m_n(\theta^0)\right) \equiv PX$$

Now, from 14.3 we have

$$X \equiv \sqrt{n}\hat{\Omega}^{-1/2}m_n(\theta^0) \xrightarrow{d} N(0, I_g)$$

- the big matrix $P = I_g - \hat{\Omega}^{-1/2}D_n'(\theta^*)\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1}D(\hat{\theta})\hat{\Omega}^{-1/2}$ converges in probability to $P_\infty = I_g - \Omega_\infty^{-1/2}D_\infty'\left(D_\infty\Omega_\infty^{-1}D_\infty'\right)^{-1}D_\infty\Omega_\infty^{-1/2}$.
- One can easily verify that P_∞ is idempotent and has rank $g - K$, (recall that the rank of an idempotent matrix is equal to its trace).
- We know as a basic result from statistics that $X'PX \xrightarrow{d} \chi^2(d)$, because it is a quadratic form of

standard normal variables, weighted by an idempotent matrix.

- So, a quadratic form on the r.h.s. has an asymptotic chi-square distribution. The quadratic form made using the l.h.s. must also have the same distribution, so we finally get

$$\left(\sqrt{n}\hat{\Omega}^{-1/2}m(\hat{\theta})\right)'\left(\sqrt{n}\hat{\Omega}^{-1/2}m(\hat{\theta})\right) = nm(\hat{\theta})'\hat{\Omega}^{-1}m(\hat{\theta}) \xrightarrow{d} \chi^2(g-K)$$

  or

$$n \cdot s_n(\hat{\theta}) \xrightarrow{d} \chi^2(g-K)$$

  supposing the model is correctly specified.

This is a convenient test since we just multiply the optimized value of the objective function by n, and compare with a $\chi^2(g-K)$ critical value. The test is a general test of whether or not the moments used to estimate are correctly specified.

This won't work when the estimator is just identified. The f.o.c. are

$$D_\theta s_n(\theta) = D\hat{\Omega}^{-1}m(\hat{\theta}) \equiv 0.$$

But with exact identification, both D and $\hat{\Omega}$ are square and invertible (at least asymptotically, assuming that asymptotic normality hold), so

$$m(\hat{\theta}) \equiv 0.$$

So the moment conditions are zero *regardless* of the weighting matrix used. As such, we might as well use an identity matrix and save trouble. Also $s_n(\hat{\theta}) = 0$, so the test breaks down.

A note: this sort of test often over-rejects in finite samples. One should be cautious in rejecting a

model when this test rejects.

This test goes by several names: Hansen test, Sargan test, Hansen-Sargan test, J test. I call it the GMM criterion test. An old name for GMM estimation is "minimum chi-square" estimation. This makes sense: the criterion function at the estimate (which makes the criterion as small as possible), scaled by n, has a χ^2 distribution.

14.9 Example: Generalized instrumental variables estimator

The IV estimator may appear a bit unusual at first, but it will grow on you over time. We have in fact already seen the IV estimator above, in the discussion of conditional moments. That presentation allows the function $k(x_t, \theta)$ to be nonlinear. Let's look in more detail at the commonly encountered special case of a linear model with iid errors, but with correlation between regressors and errors:

$$\begin{aligned} y_t &= x_t'\theta + \varepsilon_t \\ \mathcal{E}(x_t'\varepsilon_t) &\neq 0 \end{aligned}$$

- Let's assume, just to keep things simple, that the errors are iid
- The model in matrix form is $y = X\theta + \epsilon$

Let $K = dim(x_t)$. Consider some vector z_t of dimension $G \times 1$, where $G \geq K$. Assume that $E(z_t \epsilon_t) = 0$. The variables z_t are *instrumental variables.* Consider the moment conditions

$$\begin{aligned} m_t(\theta) &= z_t \epsilon_t \\ &= z_t \left(y_t - x_t' \theta \right) \end{aligned}$$

We can arrange the instruments in the $n \times G$ matrix

$$Z = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_n' \end{bmatrix}$$

The average moment conditions are

$$\begin{aligned} m_n(\theta) &= \frac{1}{n} Z' \epsilon \\ &= \frac{1}{n} (Z' y - Z' X \theta) \end{aligned}$$

The *generalized instrumental variables* estimator is just the GMM estimator based upon these moment conditions. When $G = K$, we have exact identification, and it is referred to as the instrumental variables estimator.

The first order conditions for GMM are $D_n W_n m_n(\hat{\theta}) = 0$, which imply that

$$D_n W_n Z' X \hat{\theta}_{IV} = D_n W_n Z' y$$

Exercise 50. Verify that $D_n = -\frac{X'Z}{n}$. Remember that (assuming differentiability) identification of the GMM estimator requires that this matrix must converge to a matrix with full row rank. Can just any variable that is uncorrelated with the error be used as an instrument, or is there some other condition?

Exercise 51. Verify that the efficient weight matrix is $W_n = \left(\frac{Z'Z}{n}\right)^{-1}$ (up to a constant).

If we accept what is stated in these two exercises, then

$$\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1} Z'X\hat{\theta}_{IV} = \frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1} Z'y$$

Noting that the powers of n cancel, we get

$$X'Z\left(Z'Z\right)^{-1} Z'X\hat{\theta}_{IV} = X'Z\left(Z'Z\right)^{-1} Z'y$$

or

$$\hat{\theta}_{IV} = \left(X'Z\left(Z'Z\right)^{-1} Z'X\right)^{-1} X'Z\left(Z'Z\right)^{-1} Z'y \tag{14.6}$$

Another way of arriving to the same point is to define the projection matrix P_Z

$$P_Z = Z(Z'Z)^{-1}Z'$$

Anything that is projected onto the space spanned by Z will be uncorrelated with ε, by the definition of Z. Transforming the model with this projection matrix we get

$$P_Z y = P_Z X\beta + P_Z \varepsilon$$

or

$$y^* = X^*\theta + \varepsilon^*$$

Now we have that ε^* and X^* are uncorrelated, since this is simply

$$\begin{aligned} \mathcal{E}(X^{*\prime}\varepsilon^*) &= \mathcal{E}(X'P_Z'P_Z\varepsilon) \\ &= \mathcal{E}(X'P_Z\varepsilon) \end{aligned}$$

and

$$P_Z X = Z(Z'Z)^{-1}Z'X$$

is the fitted value from a regression of X on Z. This is a linear combination of the columns of Z, so it must be uncorrelated with ε. This implies that applying OLS to the model

$$y^* = X^*\theta + \varepsilon^*$$

will lead to a consistent estimator, given a few more assumptions.

Exercise 52. Verify algebraically that applying OLS to the above model gives the IV estimator of equation 14.6.

With the definition of P_Z, we can write

$$\hat{\theta}_{IV} = (X'P_Z X)^{-1} X'P_Z y$$

from which we obtain

$$\begin{aligned} \hat{\theta}_{IV} &= (X'P_Z X)^{-1} X'P_Z (X\theta^0 + \varepsilon) \\ &= \theta^0 + (X'P_Z X)^{-1} X'P_Z \varepsilon \end{aligned}$$

so

$$\begin{aligned} \hat{\theta}_{IV} - \theta^0 &= (X'P_Z X)^{-1} X'P_Z \varepsilon \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'\varepsilon \end{aligned}$$

Now we can introduce factors of n to get

$$\hat{\theta}_{IV} - \theta^0 = \left(\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}^{-1}\right)\left(\frac{Z'X}{n}\right)\right)^{-1}\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'\varepsilon}{n}\right)$$

Assuming that each of the terms with a n in the denominator satisfies a LLN, so that

- $\frac{Z'Z}{n} \xrightarrow{p} Q_{ZZ}$, a finite pd matrix
- $\frac{X'Z}{n} \xrightarrow{p} Q_{XZ}$, a finite matrix with rank K (= cols(X)). That is to say, the instruments must be correlated with the regressors. More precisely, each regressor must be correlated with at least one instrument. Otherwise, the row of Q_{XZ} corresponding to that regressor would be all zeros, and thus the rank of the matrix would be less than K.

- $\frac{Z'\varepsilon}{n} \xrightarrow{p} 0$

then the plim of the rhs is zero. This last term has plim 0 because we assume that Z and ε are uncorrelated, e.g.,

$$\mathcal{E}(z_t'\varepsilon_t) = 0,$$

Given these assumptions, the IV estimator is consistent

$$\hat{\theta}_{IV} \xrightarrow{p} \theta^0.$$

Furthermore, scaling by $\sqrt{n}$, we have

$$\sqrt{n}\left(\hat{\theta}_{IV} - \theta^0\right) = \left(\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'X}{n}\right)\right)^{-1}\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'\varepsilon}{\sqrt{n}}\right)$$

Assuming that the far right term satifies a CLT, so that

- $\frac{Z'\varepsilon}{\sqrt{n}} \xrightarrow{d} N(0, Q_{ZZ}\sigma^2)$

then we get

$$\sqrt{n}\left(\hat{\theta}_{IV} - \theta^0\right) \xrightarrow{d} N\left(0, (Q_{XZ}Q_{ZZ}^{-1}Q_{XZ}')^{-1}\sigma^2\right)$$

The estimators for Q_{XZ} and Q_{ZZ} are the obvious ones. An estimator for σ^2 is

$$\widehat{\sigma_{IV}^2} = \frac{1}{n}\left(y - X\hat{\theta}_{IV}\right)'\left(y - X\hat{\theta}_{IV}\right).$$

This estimator is consistent following the proof of consistency of the OLS estimator of σ^2, when the classical assumptions hold.

The formula used to estimate the variance of $\hat{\theta}_{IV}$ is

$$\hat{V}(\hat{\theta}_{IV}) = \left((X'Z) (Z'Z)^{-1} (Z'X) \right)^{-1} \widehat{\sigma^2_{IV}}$$

The GIV estimator is

1. Consistent
2. Asymptotically normally distributed
3. Biased in general, because even though $\mathcal{E}(X'P_Z\varepsilon) = 0$, $\mathcal{E}(X'P_ZX)^{-1}X'P_Z\varepsilon$ may not be zero, because $(X'P_ZX)^{-1}$ and $X'P_Z\varepsilon$ are not independent.

An important point is that the asymptotic distribution of $\hat{\beta}_{IV}$ depends upon Q_{XZ} and Q_{ZZ}, and these depend upon the choice of Z. *The choice of instruments influences the efficiency of the estimator.* This point was made above, when optimal instruments were discussed.

- When we have two sets of instruments, Z_1 and Z_2 such that $Z_1 \subset Z_2$, then the IV estimator using Z_2 is at least as efficiently asymptotically as the estimator that used Z_1. More instruments leads to more asymptotically efficient estimation, in general. The same holds for GMM in general: adding moment conditions cannot cause the asymptotic variance to become larger.
- The penalty for indiscriminant use of instruments is that the small sample bias of the IV estimator rises as the number of instruments increases. The reason for this is that P_ZX becomes closer and closer to X itself as the number of instruments increases.

Exercise 53. How would one adapt the GIV estimator presented here to deal with the case of heteroscedastic and/or autocorrelated errors?

Example 54. Recall Example 19 which deals with a dynamic model with measurement error. The model is

$$\begin{aligned} y_t^* &= \alpha + \rho y_{t-1}^* + \beta x_t + \epsilon_t \\ y_t &= y_t^* + \upsilon_t \end{aligned}$$

where ϵ_t and υ_t are independent Gaussian white noise errors. Suppose that y_t^* is not observed, and instead we observe y_t. If we estimate the equation

$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \nu_t$$

by OLS, we have seen in Example 19 that the estimator is biased an inconsistent. What about using the GIV estimator? Consider using as instruments $Z = [1\, x_t\, x_{t-1}\, x_{t-2}]$. The lags of x_t are correlated with y_{t-1} as long as ρ and β are different from zero, and by assumption x_t and its lags are uncorrelated with ϵ_t and υ_t (and thus they're also uncorrelated with ν_t). Thus, these are legitimate instruments. As we have 4 instruments and 3 parameters, this is an overidentified situation. The Octave script GMM/MeasurementErrorIV.m does a Monte Carlo study using 1000 replications, with a sample size of 100. The results are comparable with those in Example 19. Using the GIV estimator, descriptive statistics for 1000 replications are

```
octave:3> MeasurementErrorIV
rm: cannot remove ‘meas_error.out’: No such file or directory
      mean  st. dev.       min       max
```

```
    0.000     0.241    -1.250     1.541
   -0.016     0.149    -0.868     0.827
   -0.001     0.177    -0.757     0.876
octave:4>
```

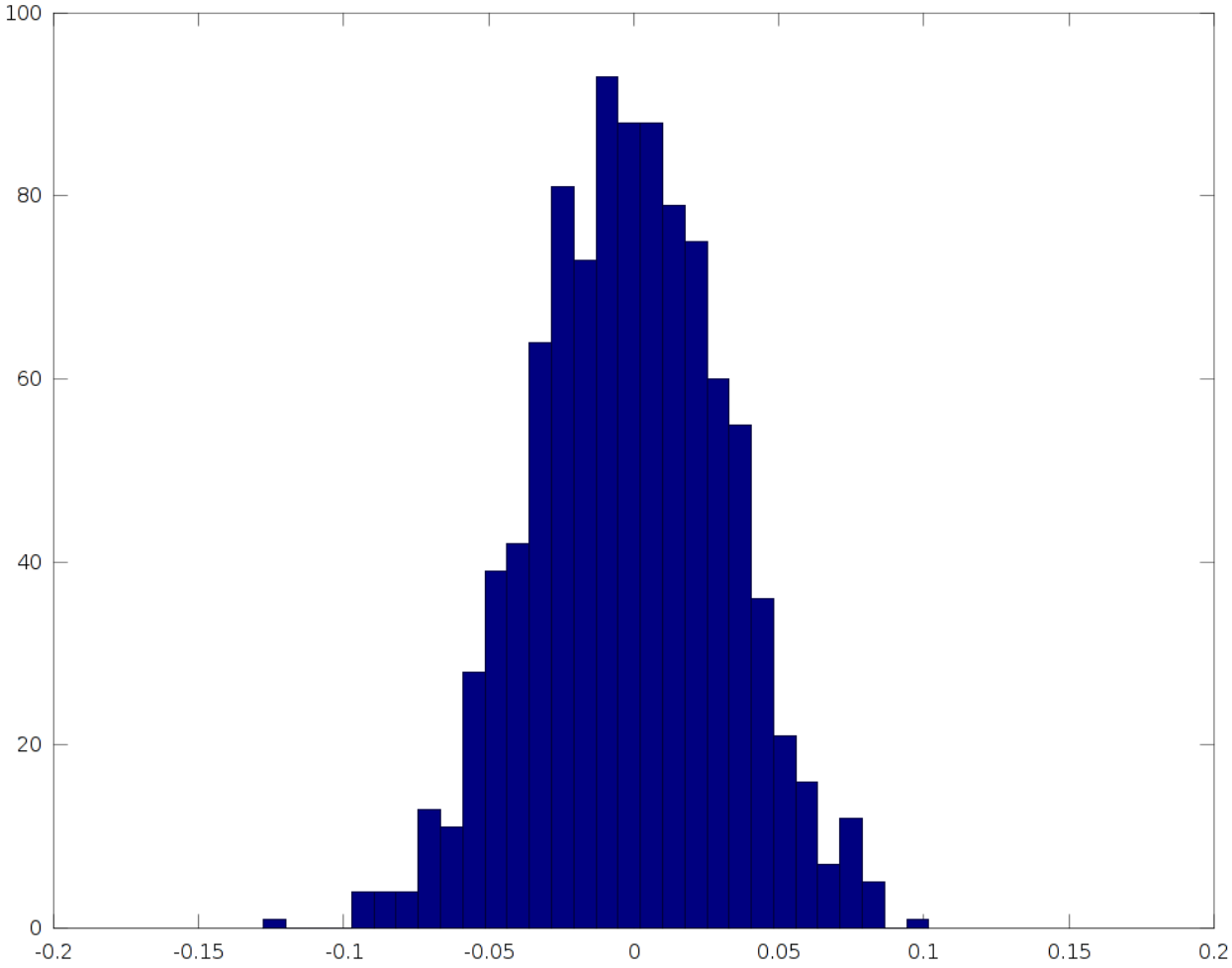
If you compare these with the results for the OLS estimator, you will see that the bias of the GIV estimator is much less for estimation of ρ. If you increase the sample size, you will see that the GIV estimator is consistent, but that the OLS estimator is not.

A histogram for $\hat{\rho} - \rho$ is in Figure 14.4. You can compare with the similar figure for the OLS estimator, Figure 7.5. As mentioned above, when the GMM estimator is overidentified and we use a consistent estimate of the efficient weight matrix, we have the criterion-based specification test $n \cdot s_n(\hat{\theta})$ available. The Octave script GMM/SpecTest.m does a Monte Carlo study of this test, for the dynamic model with measurement error, and shows that it over-rejects a correctly specified model in this case. For example, if the significance level is set to 10%, the test rejects about 16% of the time. This is a common result for this test.

2SLS

In the general discussion of GIV above, we haven't considered from where we get the instruments. Two stage least squares is an example of a particular GIV estimator, where the instruments are obtained in a particular way. Consider a single equation from a system of simultaneous equations. Refer back

Figure 14.4: GIV estimation results for $\hat{\rho} - \rho$, dynamic model with measurement error



to equation 10.3 for context. The model is

$$\begin{aligned} y &= Y_1\gamma_1 + X_1\beta_1 + \varepsilon \\ &= Z\delta + \varepsilon \end{aligned}$$

where $Z = \begin{bmatrix} Y_1 & X_1 \end{bmatrix}$ and $\delta = \begin{bmatrix} \gamma_1' & \beta_1' \end{bmatrix}'$ and Y_1 are current period endogenous variables that are correlated with the error term. X_1 are exogenous and predetermined variables that are assumed not to be correlated with the error term. Let X be all of the weakly exogenous variables (please refer back for context). The problem, recall, is that the variables in Y_1 are correlated with ϵ.

- Define $\hat{Z} = \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix}$ as the vector of predictions of Z when regressed upon X:

  $$\hat{Z} = X\left(X'X\right)^{-1}X'Z$$

  Remember that X are *all* of the exogenous variables from all equations. The fitted values of a regression of X_1 on X are just X_1, because X contains X_1. So, $\hat{Y}_1$ are the reduced form predictions of Y_1.

- Since $\hat{Z}$ is a linear combination of the weakly exogenous variables X, it must be uncorrelated with ε. This suggests the K-dimensional moment condition $m_t(\delta) = \hat{\mathbf{z}}_t\left(y_t - \mathbf{z}_t'\delta\right)$ and so

  $$m(\delta) = 1/n\sum_t \hat{\mathbf{z}}_t\left(y_t - \mathbf{z}_t'\delta\right).$$

- Since we have K parameters and K moment conditions, the GMM estimator will set m identically

equal to zero, regardless of W, so we have

$$\hat{\delta} = \left(\sum_t \hat{\mathbf{z}}_t \mathbf{z}_t' \right)^{-1} \sum_t \left(\hat{\mathbf{z}}_t y_t \right) = \left(\hat{\mathbf{Z}}' \mathbf{Z} \right)^{-1} \hat{\mathbf{Z}}' \mathbf{y}$$

This is the standard formula for 2SLS. We use the exogenous variables and the reduced form predictions of the endogenous variables as instruments, and apply IV estimation. See Hamilton pp. 420-21 for the varcov formula (which is the standard formula for 2SLS), and for how to deal with ε_t heterogeneous and dependent (basically, just use the Newey-West or some other consistent estimator of Ω, and apply the usual formula).

- Note that autocorrelation of ε_t causes lagged endogenous variables to loose their status as legitimate instruments. Some caution is warranted if this suspicion arises.
- An example of 2SLS estimation is given in Section 10.10.
- We can also estimate this same model using plain GMM estimation, this is done in Simeq/KleinGMM.m. This script shows the use of the Newey-West covariance estimator.

14.10 Nonlinear simultaneous equations

GMM provides a convenient way to estimate nonlinear systems of simultaneous equations. We have a system of equations of the form

$$\begin{aligned} y_{1t} &= f_1(\mathbf{z}_t, \theta_1^0) + \varepsilon_{1t} \\ y_{2t} &= f_2(\mathbf{z}_t, \theta_2^0) + \varepsilon_{2t} \\ &\vdots \\ y_{Gt} &= f_G(\mathbf{z}_t, \theta_G^0) + \varepsilon_{Gt}, \end{aligned}$$

or in compact notation

$$y_t = f(\mathbf{z}_t, \theta^0) + \varepsilon_t,$$

where $f(\cdot)$ is a G -vector valued function, and $\theta^0 = (\theta_1^{0\prime}, \theta_2^{0\prime}, \cdots, \theta_G^{0\prime})'$. We assume that $\mathbf{z}_t$ contains the current period endogenous variables, so we have a simultaneity problem.

We need to find an $A_i \times 1$ vector of instruments $\mathbf{x}_{it}$, for each equation, that are uncorrelated with ε_{it}. Typical instruments would be low order monomials in the exogenous variables in $\mathbf{z}_t$, with their lagged values. Then we can define the $\left(\sum_{i=1}^G A_i\right) \times 1$ orthogonality conditions

$$m_t(\theta) = \begin{bmatrix} (y_{1t} - f_1(\mathbf{z}_t, \theta_1))\, \mathbf{x}_{1t} \\ (y_{2t} - f_2(\mathbf{z}_t, \theta_2))\, \mathbf{x}_{2t} \\ \vdots \\ (y_{Gt} - f_G(\mathbf{z}_t, \theta_G))\, \mathbf{x}_{Gt} \end{bmatrix}.$$

- once we have gotten this far, we can just proceed with GMM estimation, one-step, two-step,

CUE, or whatever.

- A note on identification: selection of instruments that ensure identification is a non-trivial problem. Identification in nonlinear models is not as easy to check as it is with linear models, where counting zero restrictions works.

- A note on efficiency: the selected set of instruments has important effects on the efficiency of estimation. There are some papers that study this problem, but the results are fairly complicated and difficult to implement. I think it's safe to say that the great majority of applied work does not attempt to use optimal instruments.

14.11 Maximum likelihood

In the introduction we argued that ML will in general be more efficient than GMM since ML implicitly uses all of the moments of the distribution while GMM uses a limited number of moments. Actually, a distribution with P parameters can be uniquely characterized by P moment conditions. However, some sets of P moment conditions may contain more information than others, since the moment conditions could be highly correlated. A GMM estimator that chose an optimal set of P moment conditions would be fully efficient. Here we'll see that the optimal moment conditions are simply the scores of the ML estimator.

Let y_t be a G -vector of variables, and let $Y_t = (y_1', y_2', ..., y_t')'$. Then at time t, Y_{t-1} has been observed (refer to it as the information set, since we assume the conditioning variables have been selected to take advantage of all useful information). The likelihood function is the joint density of the sample:

$$\mathcal{L}(\theta) = f(y_1, y_2, ..., y_n, \theta)$$

which can be factored as

$$\mathcal{L}(\theta) = f(y_n|Y_{n-1}, \theta) \cdot f(Y_{n-1}, \theta)$$

and we can repeat this to get

$$\mathcal{L}(\theta) = f(y_n|Y_{n-1}, \theta) \cdot f(y_{n-1}|Y_{n-2}, \theta) \cdot ... \cdot f(y_1).$$

The log-likelihood function is therefore

$$\ln \mathcal{L}(\theta) = \sum_{t=1}^{n} \ln f(y_t|Y_{t-1}, \theta).$$

Define

$$m_t(Y_t, \theta) \equiv D_\theta \ln f(y_t|Y_{t-1}, \theta)$$

as the *score* of the t^{th} observation. It can be shown that, under the regularity conditions, that the scores have conditional mean zero when evaluated at θ^0 (see 13.2):

$$\mathcal{E}\left(m_t(Y_t, \theta^0)|Y_{t-1}\right) = 0$$

so one could interpret these as moment conditions to use to define a just-identified GMM estimator (if there are K parameters there are K score equations). The GMM estimator sets

$$1/n \sum_{t=1}^{n} m_t(Y_t, \hat{\theta}) = 1/n \sum_{t=1}^{n} D_\theta \ln f(y_t|Y_{t-1}, \hat{\theta}) = 0,$$

which are precisely the first order conditions of MLE. Therefore, MLE can be interpreted as a GMM estimator. The GMM varcov formula is $V_\infty = \left(D_\infty \Omega^{-1} D'_\infty\right)^{-1}$.

Consistent estimates of variance components are as follows

- D_∞

$$\widehat{D_\infty} = \frac{\partial}{\partial \theta'} m(Y_t, \hat{\theta}) = 1/n \sum_{t=1}^{n} D_\theta^2 \ln f(y_t | Y_{t-1}, \hat{\theta})$$

- Ω

  It is important to note that m_t and m_{t-s}, $s > 0$ are both conditionally and unconditionally uncorrelated. Conditional uncorrelation follows from the fact that m_{t-s} is a function of Y_{t-s}, which is in the information set at time t. Unconditional uncorrelation follows from the fact that conditional uncorrelation hold regardless of the realization of Y_{t-1}, so marginalizing with respect to Y_{t-1} preserves uncorrelation (see the section on ML estimation, above). The fact that the scores are serially uncorrelated implies that Ω can be estimated by the estimator of the 0^{th} autocovariance of the moment conditions:

$$\widehat{\Omega} = 1/n \sum_{t=1}^{n} m_t(Y_t, \hat{\theta}) m_t(Y_t, \hat{\theta})' = 1/n \sum_{t=1}^{n} \left[D_\theta \ln f(y_t | Y_{t-1}, \hat{\theta}) \right] \left[D_\theta \ln f(y_t | Y_{t-1}, \hat{\theta}) \right]'$$

  There is no need for a Newey-West style estimator, the heteroscedastic-consistent estimator of White is sufficient. Also, the fact that the scores of ML are uncorrelated suggests a means of testing the correct specification of the model: see if the fitted scores ($m_t(\hat{\theta})$ show evidence of serial correlation. If they do, the correctness of the specification of the model is subject to doubt.

14.12 Example: OLS as a GMM estimator - the Nerlove model again

The simple Nerlove model can be estimated using GMM. The Octave script NerloveGMM.m estimates the model by GMM and by OLS. It also illustrates that the weight matrix does not matter when the moments just identify the parameter. You are encouraged to examine the script and run it.

14.13 Example: The MEPS data

The MEPS data on health care usage discussed in section 11.4 estimated a Poisson model by "maximum likelihood" (probably misspecified). Perhaps the same latent factors (e.g., chronic illness) that induce one to make doctor visits also influence the decision of whether or not to purchase insurance. If this is the case, the PRIV variable could well be endogenous, in which case, the Poisson "ML" estimator would be inconsistent, even if the conditional mean were correctly specified. The Octave script meps.m estimates the parameters of the model presented in equation 11.1, using Poisson "ML" (better thought of as quasi-ML), and IV estimation[1]. Both estimation methods are implemented using a GMM form. Running that script gives the output

```
OBDV



**************************************************
IV
```

[1]The validity of the instruments used may be debatable, but real data sets often don't contain ideal instruments.

GMM Estimation Results
BFGS convergence: Normal convergence

Objective function value: 0.004273
Observations: 4564
No moment covariance supplied, assuming efficient weight matrix

	Value	df	p-value
X^2 test	19.502	3.000	0.000

	estimate	st. err	t-stat	p-value
constant	-0.441	0.213	-2.072	0.038
pub. ins.	-0.127	0.149	-0.851	0.395
priv. ins.	-1.429	0.254	-5.624	0.000
sex	0.537	0.053	10.133	0.000
age	0.031	0.002	13.431	0.000
edu	0.072	0.011	6.535	0.000
inc	0.000	0.000	4.500	0.000

****

****

```
Poisson QML

GMM Estimation Results
BFGS convergence: Normal convergence

Objective function value: 0.000000
Observations: 4564
No moment covariance supplied, assuming efficient weight matrix

Exactly identified, no spec. test

              estimate     st. err      t-stat     p-value
constant        -0.791       0.149      -5.289       0.000
pub. ins.        0.848       0.076      11.092       0.000
priv. ins.       0.294       0.071       4.136       0.000
sex              0.487       0.055       8.796       0.000
age              0.024       0.002      11.469       0.000
edu              0.029       0.010       3.060       0.002
inc             -0.000       0.000      -0.978       0.328
*******************************************************
```

Note how the Poisson QML results, estimated here using a GMM routine, are the same as were obtained using the ML estimation routine (see subsection 11.4). This is an example of how (Q)ML may

be represented as a GMM estimator. Also note that the IV and QML results are considerably different. Treating PRIV as potentially endogenous causes the sign of its coefficient to change. Perhaps it is logical that people who own private insurance make fewer visits, if they have to make a co-payment. Note that income becomes positive and significant when PRIV is treated as endogenous.

Perhaps the difference in the results depending upon whether or not PRIV is treated as endogenous can suggest a method for testing exogeneity....

14.14 Example: The Hausman Test

This section discusses the Hausman test, which was originally presented in Hausman, J.A. (1978), Specification tests in econometrics, *Econometrica*, **46**, 1251-71.

Consider the simple linear regression model $y_t = x_t'\beta + \epsilon_t$. We assume that the functional form and the choice of regressors is correct, but that the some of the regressors may be correlated with the error term, which as you know will produce inconsistency of $\hat{\beta}$. For example, this will be a problem if

- if some regressors are endogeneous
- some regressors are measured with error
- some relevant regressors are omitted (equivalent to imposing false restrictions)
- lagged values of the dependent variable are used as regressors and ϵ_t is autocorrelated.

To illustrate, the Octave program OLSvsIV.m performs a Monte Carlo experiment where errors are correlated with regressors, and estimation is by OLS and IV. The true value of the slope coefficient used to generate the data is $\beta = 2$. Figure 14.5 shows that the OLS estimator is quite biased, while

Figure 14.5: OLS

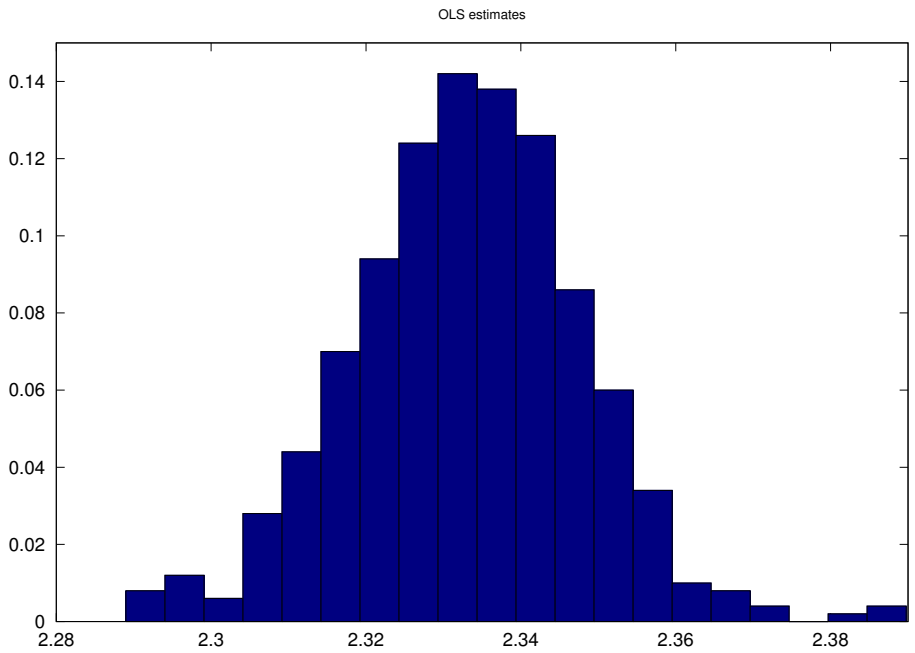
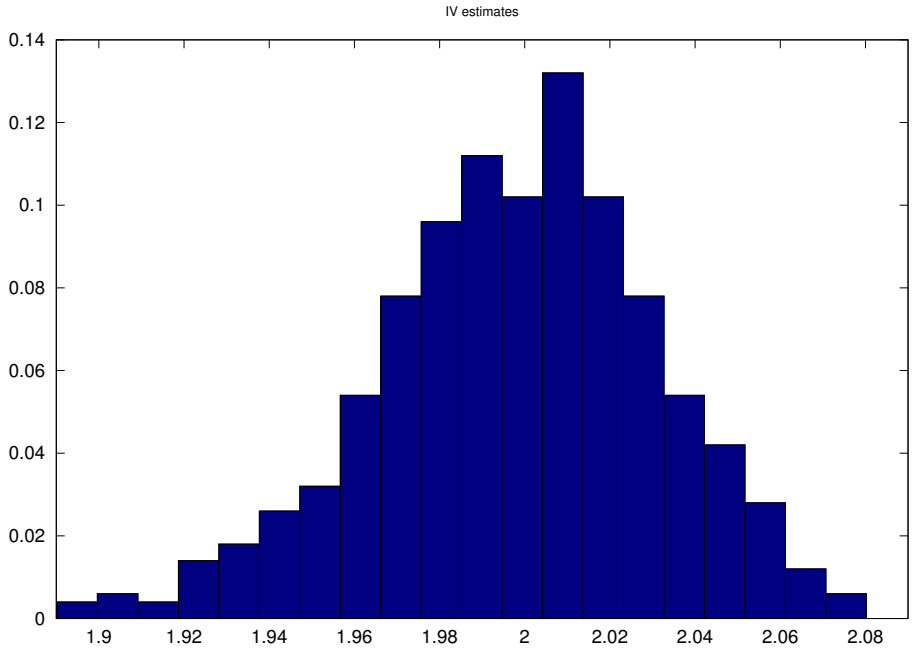


Figure 14.6 shows that the IV estimator is on average much closer to the true value. If you play with the program, increasing the sample size, you can see evidence that the OLS estimator is asymptotically biased, while the IV estimator is consistent. You can also play with the covariances of the instrument and regressor, and the covariance of the regressor and the error.

We have seen that inconsistent and the consistent estimators converge to different probability limits. This is the idea behind the Hausman test - a pair of consistent estimators converge to the same probability limit, while if one is consistent and the other is not they converge to different limits. If we accept that one is consistent (*e.g.*, the IV estimator), but we are doubting if the other is consistent (*e.g.,* the OLS estimator), we might try to check if the difference between the estimators is significantly

Figure 14.6: IV



different from zero.

- If we're doubting about the consistency of OLS (or QML, *etc.*), why should we be interested in testing - why not just use the IV estimator? Because the OLS estimator is *more efficient* when the regressors are exogenous and the other classical assumptions (including normality of the errors) hold.
- Play with the above script to convince yourself of this point: make exogeneity hold, and compare the variances of OLS and IV
- When we have a more efficient estimator that relies on stronger assumptions (such as exogeneity) than the IV estimator, we might prefer to use it, unless we have evidence that the assumptions are false.

So, let's consider the covariance between the MLE estimator $\hat{\theta}$ (or any other fully efficient estimator) and some other CAN estimator, say $\tilde{\theta}$. Now, let's recall some results from MLE. Equation 13.4 is:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{a.s.} -\mathcal{J}_\infty(\theta_0)^{-1}\sqrt{n}g(\theta_0).$$

Equation 13.8 is

$$\mathcal{J}_\infty(\theta) = -\mathcal{I}_\infty(\theta).$$

Combining these two equations, we get

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{a.s.} I_\infty(\theta_0)^{-1}\sqrt{n}g(\theta_0).$$

Also, equation 13.11 tells us that the asymptotic covariance between any CAN estimator and the MLE score vector is

$$V_\infty \left[\begin{array}{c} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n} g(\theta) \end{array} \right] = \left[\begin{array}{cc} V_\infty(\tilde{\theta}) & I_K \\ I_K & \mathcal{I}_\infty(\theta) \end{array} \right].$$

Now, consider

$$\begin{bmatrix} I_K & 0_K \\ 0_K & I_\infty(\theta)^{-1} \end{bmatrix} \left[\begin{array}{c} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n} g(\theta) \end{array} \right] \stackrel{a.s.}{\rightarrow} \left[\begin{array}{c} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n}\left(\hat{\theta} - \theta\right) \end{array} \right].$$

The asymptotic covariance of this is

$$\begin{aligned} V_\infty \left[\begin{array}{c} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n}\left(\hat{\theta} - \theta\right) \end{array} \right] &= \begin{bmatrix} I_K & 0_K \\ 0_K & I_\infty(\theta)^{-1} \end{bmatrix} \left[\begin{array}{cc} V_\infty(\tilde{\theta}) & I_K \\ I_K & \mathcal{I}_\infty(\theta) \end{array} \right] \begin{bmatrix} I_K & 0_K \\ 0_K & I_\infty(\theta)^{-1} \end{bmatrix} \\ &= \left[\begin{array}{cc} V_\infty(\tilde{\theta}) & I_\infty(\theta)^{-1} \\ I_\infty(\theta)^{-1} & I_\infty(\theta)^{-1} \end{array} \right], \end{aligned}$$

which, for clarity in what follows, we might write as (note to self for lectures: the 2,2 element has changed)

$$V_\infty \left[\begin{array}{c} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n}\left(\hat{\theta} - \theta\right) \end{array} \right] = \left[\begin{array}{cc} V_\infty(\tilde{\theta}) & I_\infty(\theta)^{-1} \\ I_\infty(\theta)^{-1} & V_\infty(\hat{\theta}) \end{array} \right].$$

So, the asymptotic covariance between the MLE and any other CAN estimator is equal to the MLE asymptotic variance (the inverse of the information matrix).

Now, suppose we wish to test whether the the two estimators are in fact both converging to θ_0, versus the alternative hypothesis that the "MLE" estimator is not in fact consistent (the consistency

of $\tilde{\theta}$ is a maintained hypothesis). Under the null hypothesis that they are, we have

$$\begin{bmatrix} I_K & -I_K \end{bmatrix} \begin{bmatrix} \sqrt{n}\left(\tilde{\theta} - \theta_0\right) \\ \sqrt{n}\left(\hat{\theta} - \theta_0\right) \end{bmatrix} = \sqrt{n}\left(\tilde{\theta} - \hat{\theta}\right),$$

will be asymptotically normally distributed as (work out on blackboard)

$$\sqrt{n}\left(\tilde{\theta} - \hat{\theta}\right) \xrightarrow{d} N\left(0, V_\infty(\tilde{\theta}) - V_\infty(\hat{\theta})\right).$$

So,

$$n\left(\tilde{\theta} - \hat{\theta}\right)' \left(V_\infty(\tilde{\theta}) - V_\infty(\hat{\theta})\right)^{-1} \left(\tilde{\theta} - \hat{\theta}\right) \xrightarrow{d} \chi^2(\rho),$$

where ρ is the rank of the difference of the asymptotic variances. A statistic that has the same asymptotic distribution is

$$\left(\tilde{\theta} - \hat{\theta}\right)' \left(\hat{V}(\tilde{\theta}) - \hat{V}(\hat{\theta})\right)^{-1} \left(\tilde{\theta} - \hat{\theta}\right) \xrightarrow{d} \chi^2(\rho).$$

This is the Hausman test statistic, in its original form. The reason that this test has power under the alternative hypothesis is that in that case the "MLE" estimator will not be consistent, and will converge to θ_A, say, where $\theta_A \neq \theta_0$. Then the mean of the asymptotic distribution of vector $\sqrt{n}\left(\tilde{\theta} - \hat{\theta}\right)$ will be $\theta_0 - \theta_A$, a non-zero vector, so the test statistic will eventually reject, regardless of how small a significance level is used.

- Note: if the test is based on a sub-vector of the entire parameter vector of the MLE, it is possible that the inconsistency of the MLE will not show up in the portion of the vector that has been used. If this is the case, the test may not have power to detect the inconsistency. This may occur,

for example, when the consistent but inefficient estimator is not identified for all the parameters of the model, so that we estimate only some of the parameters using the inefficient estimator, and the test does not include the others.

Some things to note:

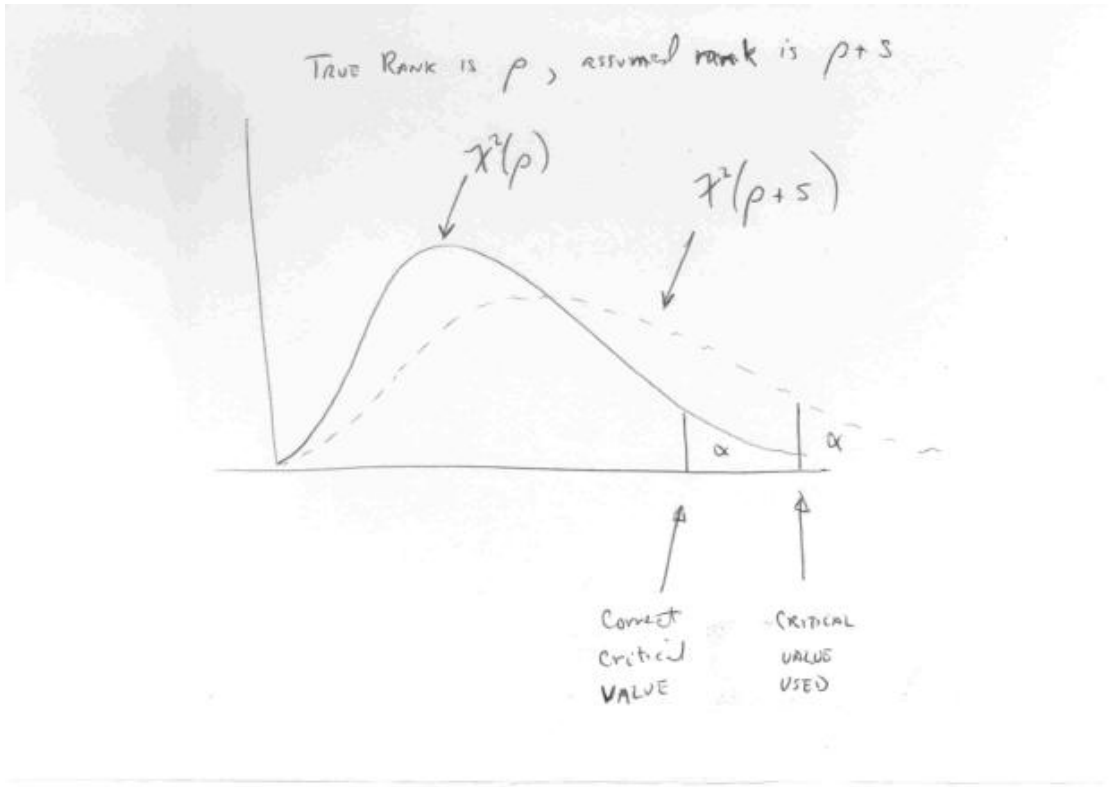
- The rank, ρ, of the difference of the asymptotic variances is often less than the dimension of the matrices, and it may be difficult to determine what the true rank is. This can occur when certain moment conditions are used to define both estimators, which introduces some linear dependence between the estimators. If the true rank is lower than what is taken to be true, the test will be biased against rejection of the null hypothesis. The null is that both estimators are consistent. Failure to reject when this hypothesis is false would cause us to use an inconsistent estimator: not a desirable outcome! The contrary holds if we underestimate the rank.

- A solution to this problem is to use a rank 1 test, by comparing only a single coefficient. For example, if a variable is suspected of possibly being endogenous, that variable's coefficients may be compared.

- This simple formula only holds when the estimator that is being tested for consistency is *fully* efficient under the null hypothesis. This means that it must be a ML estimator or a fully efficient estimator that has the same asymptotic distribution as the ML estimator. This is quite restrictive since modern estimators such as GMM, QML, or even OLS with heteroscedastic consistent standard errors are not in general fully efficient.

Following up on this last point, let's think of two not necessarily efficient estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, where one is assumed to be consistent, but the other may not be. We assume for expositional simplicity

Figure 14.7: Incorrect rank and the Hausman test



that both $\hat{\theta}_1$ and $\hat{\theta}_2$ belong to the same parameter space, and that each estimator can be expressed as generalized method of moments (GMM) estimator. The estimators are defined (suppressing the dependence upon data) by

$$\hat{\theta}_i \;=\; \arg\min_{\theta_i \in \Theta} \; m_{_i}(\theta_i)' \, W_i \, m_i(\theta_i)$$

where $m_i(\theta_i)$ is a $g_i \times 1$ vector of moment conditions, and W_i is a $g_i \times g_i$ positive definite weighting matrix, $i = 1, 2$. Consider the omnibus GMM estimator

$$\left(\hat{\theta}_1, \hat{\theta}_2\right) = \arg\min_{\Theta \times \Theta} \left[\begin{array}{cc} m_1(\theta_1)' & m_2(\theta_2)' \end{array} \right] \left[\begin{array}{cc} W_1 & \mathbf{0}_{(g_1 \times g_2)} \\ \mathbf{0}_{(g_2 \times g_1)} & W_2 \end{array} \right] \left[\begin{array}{c} m_1(\theta_1) \\ m_2(\theta_2) \end{array} \right]. \tag{14.7}$$

Suppose that the asymptotic covariance of the omnibus moment vector is

$$\begin{aligned} \Sigma \;&=\; \lim_{n \to \infty} Var \left\{ \sqrt{n} \left[\begin{array}{c} m_1(\theta_1) \\ m_2(\theta_2) \end{array} \right] \right\} \\ &\equiv \; \left(\begin{array}{cc} \Sigma_1 & \Sigma_{12} \\ \cdot & \Sigma_2 \end{array} \right). \end{aligned} \tag{14.8}$$

The standard Hausman test is equivalent to a Wald test of the equality of θ_1 and θ_2 (or subvectors of the two) applied to the omnibus GMM estimator, but with the covariance of the moment conditions estimated as

$$\widehat{\Sigma} = \left(\begin{array}{cc} \widehat{\Sigma_1} & \mathbf{0}_{(g_1 \times g_2)} \\ \mathbf{0}_{(g_2 \times g_1)} & \widehat{\Sigma_2} \end{array} \right).$$

While this is clearly an inconsistent estimator in general, the omitted Σ_{12} term cancels out of the test

statistic when one of the estimators is asymptotically efficient, as we have seen above, and thus it need not be estimated.

The general solution when neither of the estimators is efficient is clear: the entire Σ matrix must be estimated consistently, since the Σ_{12} term will not cancel out. Methods for consistently estimating the asymptotic covariance of a vector of moment conditions are well-known, *e.g.,* the Newey-West estimator discussed previously. The Hausman test using a proper estimator of the overall covariance matrix will now have an asymptotic χ^2 distribution when neither estimator is efficient.

However, the test suffers from a loss of power due to the fact that the omnibus GMM estimator of equation 14.7 is defined using an inefficient weight matrix. A new test can be defined by using an alternative omnibus GMM estimator

$$\left(\hat{\theta}_1, \hat{\theta}_2\right) = \arg\min_{\Theta\times\Theta} \left[\begin{array}{cc} m_1(\theta_1)' & m_2(\theta_2)' \end{array} \right] \left(\widetilde{\Sigma}\right)^{-1} \left[\begin{array}{c} m_1(\theta_1) \\ m_2(\theta_2) \end{array} \right], \tag{14.9}$$

where $\widetilde{\Sigma}$ is a consistent estimator of the overall covariance matrix Σ of equation 14.8. By standard arguments, this is a more efficient estimator than that defined by equation 14.7, so the Wald test using this alternative is more powerful. See my article in *Applied Economics*, 2004, for more details, including simulation results. The Octave script hausman.m calculates the Wald test corresponding to the efficient joint GMM estimator (the "H2" test in my paper), for a simple linear model.

14.15 Application: Nonlinear rational expectations

Readings: Hansen and Singleton, *Econometrics,* 1982; Tauchen, *Journal of Business and Economic Statistics*, 1986.

Though GMM estimation has many applications, application to rational expectations models is elegant, since theory directly suggests the moment conditions. Hansen and Singleton's 1982 paper is also a classic worth studying in itself. Though I strongly recommend reading the paper, I'll use a simplified model with similar notation to Hamilton's. The literature on estimation of these models has grown a lot since these early papers. After work like the cited papers, people moved to ML estimation of linearized models, using Kalman filtering. Current methods are usually Bayesian, and involve sophisticated filtering methods to compute the likelihood function for nonlinear models with non-normal shocks. There is a lot of interesting stuff that is beyond the scope of this course. I have done some work using simulation-based estimation methods applied to such models. The methods explained in this section are intended to provide an example of GMM estimation. They are not the state of the art for estimation of such models.

We assume a representative consumer maximizes expected discounted utility over an infinite horizon. Expectations are rational, and the agent has full information (is fully aware of the history of the world up to the current time period - how's that for an assumption!). Utility is temporally additive, and the expected utility hypothesis holds. The future consumption stream is the stochastic sequence $\{c_t\}_{t=0}^{\infty}$. The objective function at time t is the discounted expected utility

$$\sum_{s=0}^{\infty} \beta^s \mathcal{E}\left(u(c_{t+s})|I_t\right). \tag{14.10}$$

- The parameter β is between 0 and 1, and reflects discounting.
- I_t is the *information set* at time t, and includes the all realizations of all random variables indexed t and earlier.
- The choice variable is c_t - current consumption, which is constrained to be less than or equal to

current wealth w_t.

- Suppose the consumer can invest in a risky asset. A dollar invested in the asset yields a gross return

$$(1 + r_{t+1}) = \frac{p_{t+1} + d_{t+1}}{p_t}$$

where p_t is the price and d_t is the dividend in period t. Thus, r_{t+1} is the net return on a dollar invested in period t.

- The price of c_t is normalized to 1.

- Current wealth $w_t = (1 + r_t)i_{t-1}$, where i_{t-1} is investment in period $t - 1$. So the problem is to allocate current wealth between current consumption and investment to finance future consumption: $w_t = c_t + i_t$.

- Future net rates of return $r_{t+s}, s > 0$ are *not known* in period t: the asset is risky.

A partial set of necessary conditions for utility maximization have the form:

$$u'(c_t) = \beta \mathcal{E} \left\{ (1 + r_{t+1}) \, u'(c_{t+1}) | I_t \right\}. \tag{14.11}$$

To see that the condition is necessary, suppose that the lhs $<$ rhs. Then by reducing current consumption marginally would cause equation 14.10 to drop by $u'(c_t)$, since there is no discounting of the current period. At the same time, the marginal reduction in consumption finances investment, which has gross return $(1 + r_{t+1})$, which could finance consumption in period $t + 1$. This increase in consumption would cause the objective function to increase by $\beta \mathcal{E} \left\{ (1 + r_{t+1}) \, u'(c_{t+1}) | I_t \right\}$. Therefore, unless the condition holds, the expected discounted utility function is not maximized.

- To use this we need to choose the functional form of utility. A constant relative risk aversion (CRRA) form is

$$u(c_t) = \frac{c_t^{1-\gamma} - 1}{1-\gamma}$$

where γ is the coefficient of relative risk aversion. With this form,

$$u'(c_t) = c_t^{-\gamma}$$

so the foc are

$$c_t^{-\gamma} = \beta \mathcal{E}\left\{(1 + r_{t+1})\, c_{t+1}^{-\gamma} | I_t\right\}$$

While it is true that

$$\mathcal{E}\left(c_t^{-\gamma} - \beta\left\{(1 + r_{t+1})\, c_{t+1}^{-\gamma}\right\}\right) | I_t = 0$$

so that we could use this to define moment conditions, it is unlikely that c_t is stationary, even though it is in real terms, and our theory requires stationarity. To solve this, divide though by $c_t^{-\gamma}$

$$\mathcal{E}\left(1\text{-}\beta\left\{(1 + r_{t+1})\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right\}\right) | I_t = 0$$

(note that c_t can be passed though the conditional expectation since c_t is chosen based only upon information available in time t). That is to say, c_t is in the information set I_t.

Now

$$1\text{-}\beta\left\{(1 + r_{t+1})\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right\}$$

is analogous to $h_t(\theta)$ defined above: it's a scalar moment condition. To get a vector of moment condi-

tions we need some instruments. Suppose that $\mathbf{z}_t$ is a vector of variables drawn from the information set I_t. We can use the necessary conditions to form the expressions

$$\left[1 - \beta\left(1 + r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right]\mathbf{z}_t \equiv m_t(\theta)$$

- θ represents β and γ.
- Therefore, the above expression may be interpreted as a moment condition which can be used for GMM estimation of the parameters θ^0.

Note that at time t, m_{t-s} has been observed, and is therefore an element of the information set. By rational expectations, the autocovariances of the moment conditions other than Γ_0 should be zero. The optimal weighting matrix is therefore the inverse of the variance of the moment conditions:

$$\Omega_{\infty} = \lim E\left[nm(\theta^0)m(\theta^0)'\right]$$

which can be consistently estimated by

$$\hat{\Omega} = 1/n \sum_{t=1}^{n} m_t(\hat{\theta})m_t(\hat{\theta})'$$

As before, this estimate depends on an initial consistent estimate of θ, which can be obtained by setting the weighting matrix W arbitrarily (to an identity matrix, for example). After obtaining $\hat{\theta}$, we then minimize

$$s(\theta) = m(\theta)'\hat{\Omega}^{-1}m(\theta).$$

This process can be iterated, e.g., use the new estimate to re-estimate Ω, use this to estimate θ^0, and

repeat until the estimates don't change.

- In principle, we could use a very large number of moment conditions in estimation, since *any current or lagged variable* could be used in $\mathbf{x}_t$. Since use of more moment conditions will lead to a more (asymptotically) efficient estimator, one might be tempted to use many instrumental variables. We will do a computer lab that will show that this may not be a good idea with finite samples. This issue has been studied using Monte Carlos (Tauchen, *JBES,* 1986). The reason for poor performance when using many instruments is that the estimate of Ω becomes very imprecise.

- Empirical papers that use this approach often have serious problems in obtaining precise estimates of the parameters, and identification can be problematic. Note that we are basing everything on a single partial first order condition. Probably this f.o.c. is simply not informative enough.

14.16 Empirical example: a portfolio model

The Octave program portfolio.m performs GMM estimation of a portfolio model, using the data file tauchen.data. The columns of this data file are c, p, and d in that order. There are 95 observations (source: Tauchen, *JBES*, 1986). As instruments we use lags of c and r, as well as a constant. For a single lag the estimation results are

```
MPITB extensions found
```

```
******************************************************
Example of GMM estimation of rational expectations model

GMM Estimation Results
BFGS convergence: Normal convergence

Objective function value: 0.000014
Observations: 94

               Value          df     p-value
X^2 test       0.001       1.000       0.971

         estimate     st. err      t-stat     p-value
beta        0.915       0.009      97.271       0.000
gamma       0.569       0.319       1.783       0.075
******************************************************
```

For two lags the estimation results are

```
MPITB extensions found


******************************************************
Example of GMM estimation of rational expectations model
```

```
GMM Estimation Results
BFGS convergence: Normal convergence

Objective function value: 0.037882
Observations: 93

                Value          df     p-value
X^2 test        3.523       3.000       0.318

          estimate     st. err      t-stat     p-value
beta         0.857       0.024      35.636       0.000
gamma       -2.351       0.315      -7.462       0.000
************************************************************
```

Pretty clearly, the results are sensitive to the choice of instruments. Maybe there is some problem here: poor instruments, or possibly a conditional moment that is not very informative. Moment conditions formed from Euler conditions sometimes do not identify the parameter of a model. See Hansen, Heaton and Yarron, (1996) *JBES* V14, N3. I believe that this is the case here, though I haven't checked it carefully.

Aside on ML estimation of RBC model. A similar model is the RBC model discussed by Fernández-Villaverde: Fernández-Villaverde's RBC example. Files to estimate this model by maximum

likelihood are provided here. The main point for the purposes of this course is that methods other than GMM based on the Euler equation do exist, and work better. For those of you who go on to do empirical macro work, this example may be useful in the future.

14.17 Exercises

1. Do the exercises in section 14.9.

2. Show how the GIV estimator presented in section 14.9 can be adapted to account for an error term with HET and/or AUT.

3. For the GIV estimator presented in section 14.9, find the form of the expressions $\mathcal{I}_\infty(\theta^0)$ and $\mathcal{J}_\infty(\theta^0)$ that appear in the asymptotic distribution of the estimator, assuming that an efficient weight matrix is used.

4. The Octave script meps.m estimates a model for office-based doctpr visits (OBDV) using two different moment conditions, a Poisson QML approach and an IV approach. If all conditioning variables are exogenous, both approaches should be consistent. If the PRIV variable is endogenous, only the IV approach should be consistent. Neither of the two estimators is efficient in any case, since we already know that this data exhibits variability that exceeds what is implied by the Poisson model (e.g., negative binomial and other models fit much better). Test the exogeneity of the variable PRIV with a GMM-based Hausman-type test, using the Octave script hausman.m for hints about how to set up the test.

5. Using Octave, generate data from the logit dgp. The script EstimateLogit.m should prove quite helpful.

   (a) Recall that $E(y_t|\mathbf{x}_t) = \mathbf{p}(\mathbf{x}_t, \theta) = [1 + \exp(-\mathbf{x}_t\prime\theta)]^{-1}$. Consider the moment condtions (exactly identified) $m_t(\theta) = [y_t - p(\mathbf{x}_t, \theta)]\mathbf{x}_t$. Estimate by GMM (using `gmm_results`), using these moments.

(b) Estimate by ML (using `mle_results`).

(c) The two estimators should coincide. Prove analytically that the estimators coicide.

6. When working out the structure of Ω_n, show that $\mathcal{E}(m_t m'_{t+s}) = \Gamma'_v$.

7. Verify the missing steps needed to show that $n \cdot m(\hat{\theta})'\hat{\Omega}^{-1}m(\hat{\theta})$ has a $\chi^2(g-K)$ distribution. That is, show that the monster matrix is idempotent and has trace equal to $g-K$.

8. For the portfolio example, experiment with the program using lags of 3 and 4 periods to define instruments

   (a) Iterate the estimation of $\theta = (\beta, \gamma)$ and Ω to convergence.

   (b) Comment on the results. Are the results sensitive to the set of instruments used? Look at $\hat{\Omega}$ as well as $\hat{\theta}$. Are these good instruments? Are the instruments highly correlated with one another? Is there something analogous to collinearity going on here?

9. Run the Octave script GMM/chi2gmm.m with several sample sizes. Do the results you obtain seem to agree with the consistency of the GMM estimator? Explain.

10. The GMM estimator with an arbitrary weight matrix has the asymptotic distribution

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \left(D_\infty W_\infty D'_\infty\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D'_\infty \left(D_\infty W_\infty D'_\infty\right)^{-1}\right]$$

Supposing that you compute a GMM estimator using an arbitrary weight matrix, so that this result applies. Carefully explain how you could test the hypothesis $H_0 : R\theta^0 = r$ versus $H_A : R\theta^0 \neq r$, where R is a given $q \times k$ matrix, and r is a given $q \times 1$ vector. I suggest that you use

a Wald test. Explain exactly what is the test statistic, and how to compute every quantity that appears in the statistic.

11. (proof that the GMM optimal weight matrix is one such that $W_\infty = \Omega_\infty^{-1}$) Consider the difference of the asymptotic variance using an arbitrary weight matrix, minus the asymptotic variance using the optimal weight matrix:

$$A = (D_\infty W_\infty D_\infty')^{-1} D_\infty W_\infty \Omega_\infty W_\infty D_\infty' (D_\infty W_\infty D_\infty')^{-1} - \left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}$$

Set $B = (D_\infty W_\infty D_\infty')^{-1} D_\infty W_\infty - (D_\infty \Omega_\infty^{-1} D_\infty')^{-1} D_\infty \Omega_\infty^{-1}$. Verify that $A = B\Omega_\infty B'$. What is the implication of this? Explain.

12. Recall the dynamic model with measurement error that was discussed in class:

$$\begin{aligned} y_t^* &= \alpha + \rho y_{t-1}^* + \beta x_t + \epsilon_t \\ y_t &= y_t^* + \upsilon_t \end{aligned}$$

where ϵ_t and υ_t are independent Gaussian white noise errors. Suppose that y_t^* is not observed, and instead we observe y_t. We can estimate the equation

$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \nu_t$$

using GIV, as was done above. The Octave script GMM/SpecTest.m performs a Monte Carlo study of the performance of the GMM criterion test,

$$n \cdot s_n(\hat{\theta}) \xrightarrow{d} \chi^2(g - K)$$

Examine the script and describe what it does. Run this script to verify that the test over-rejects. Increase the sample size, to determine if the over-rejection problem becomes less severe. Discuss your findings.

Chapter 15

Models for time series data

Hamilton, *Time Series Analysis* is a good reference for this chapter.

Up to now we've considered the behavior of the dependent variable y_t as a function of other variables x_t. These variables can of course contain lagged dependent variables, e.g., $x_t = (w_t, y_{t-1}, ..., y_{t-j})$. Pure time series methods consider the behavior of y_t as a function only of its own lagged values, unconditional on other observable variables. One can think of this as modeling the behavior of y_t after marginalizing out all other variables. While it's not immediately clear why a model that has other explanatory variables should marginalize to a linear in the parameters time series model, most applied time series work is done with linear models, though nonlinear time series is also a large and growing field.

Basic concepts

Definition 55. [Stochastic process] A stochastic process is a sequence of random variables, indexed by time: $\{Y_t\}_{t=-\infty}^{\infty}$

Definition 56. [Time series] A time series is **one** observation of a stochastic process, over a specific interval: $\{y_t\}_{t=1}^{n}$.

So a time series is a sample of size n from a stochastic process. It's important to keep in mind that conceptually, one could draw another sample, and that the values would be different.

Definition 57. [Autocovariance] The j^{th} autocovariance of a stochastic process is $\gamma_{jt} = \mathcal{E}(y_t - \mu_t)(y_{t-j} - \mu_{t-j})$ where $\mu_t = \mathcal{E}(y_t)$.

Definition 58. [Covariance (weak) stationarity] A stochastic process is covariance stationary if it has time constant mean and autocovariances of all orders:

$$\begin{aligned} \mu_t &= \mu, \quad \forall t \\ \gamma_{jt} &= \gamma_j, \quad \forall t \end{aligned}$$

As we've seen, this implies that $\gamma_j = \gamma_{-j}$: the autocovariances depend only one the interval between observations, but not the time of the observations.

Definition 59. [Strong stationarity] A stochastic process is strongly stationary if the joint distribution of an arbitrary collection of the $\{Y_t\}$ doesn't depend on t.

Since moments are determined by the distribution, strong stationarity⇒weak stationarity.

What is the mean of Y_t? The time series is one sample from the stochastic process. One could think of M repeated samples from the stoch. proc., e.g., $\{y_{tm}\}$ By a LLN, we would expect that

$$\lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} y_{tm} \xrightarrow{p} \mathcal{E}(Y_t)$$

The problem is, we have only one sample to work with, since we can't go back in time and collect another. How can $\mathcal{E}(Y_t)$ be estimated then? It turns out that *ergodicity* is the needed property.

Definition 60. [Ergodicity]. A stationary stochastic process is ergodic (for the mean) if the time average converges to the mean

$$\frac{1}{n} \sum_{t=1}^{n} y_t \xrightarrow{p} \mu \tag{15.1}$$

A sufficient condition for ergodicity is that the autocovariances be absolutely summable:

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty$$

This implies that the autocovariances die off, so that the y_t are not so strongly dependent that they don't satisfy a LLN.

Definition 61. [Autocorrelation] The j^{th} autocorrelation, ρ_j is just the j^{th} autocovariance divided by the variance:

$$\rho_j = \frac{\gamma_j}{\gamma_0} \tag{15.2}$$

Definition 62. [White noise] White noise is just the time series literature term for a classical error. ϵ_t is white noise if i) $\mathcal{E}(\epsilon_t) = 0, \forall t$, ii) $V(\epsilon_t) = \sigma^2, \forall t$ and iii) ϵ_t and ϵ_s are independent, $t \neq s$. Gaussian white noise just adds a normality assumption.

15.1 ARMA models

With these concepts, we can discuss ARMA models. These are closely related to the AR and MA error processes that we've already discussed. The main difference is that the lhs variable is observed directly now.

MA(q) processes

A q^{th} order moving average (MA) process is

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where ε_t is white noise. The variance is

$$\begin{aligned} \gamma_0 &= \mathcal{E}\left(y_t - \mu\right)^2 \\ &= \mathcal{E}\left(\varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}\right)^2 \\ &= \sigma^2 \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right) \end{aligned}$$

Similarly, the autocovariances are

$$\begin{aligned}\gamma_j &= \mathcal{E}\left[(y_t - \mu)(y_{t-j} - \mu)\right] \\ &= \sigma^2(\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \cdots + \theta_q\theta_{q-j}), j \leq q \\ &= 0, j > q\end{aligned}$$

Therefore an MA(q) process is necessarily covariance stationary and ergodic, as long as σ^2 and all of the θ_j are finite.

For example, if we have an MA(1) model, then $E(y_t) = \mu$, $V(y_t) = \sigma^2(1 + \theta_1^2)$, and $\gamma_1 = \sigma^2\theta_1$. The higher order autocovariances are zero.

Thus, if the model is MA(1), the density of the vector of n observations, y, is

$$f_Y(y|\rho) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(y - \mu)' \Sigma^{-1} (y - \mu)\right)$$

where

$$\Sigma = \sigma^2 \begin{bmatrix} 1 + \theta_1^2 & \theta_1 & 0 & \cdots & 0 \\ \theta_1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \theta_1 \\ 0 & \cdots & 0 & \theta_1 & 1 + \theta_1^2 \end{bmatrix}.$$

With this, it is very easy to program the log-likelihood function. For higher order MA models, the only difference is the structure of Σ becomes more complicated. In this form, one needs a lot of computer memory. A more economical approach uses the Kalman filter, which we'll see in the discussion of state

space models.

- An issue to be aware of is that MA models are not identified, in that there exist multiple parameter values that give the same value of the likelihood function.
- For example, the MA(1) model with $\tilde{\sigma}^2 = \theta^2\sigma^2$ and $\tilde{\theta}_1 = \frac{1}{\theta_1}$ has identical first and second moments to the original model, so the likelihood function has the same value.
- Normally, the parameterization that leads to an *invertible* MA model is the one that is selected. An invertible MA model is one that has a representation as a AR(∞) model. For the MA(1) model, the invertible parameterization has $|\theta_1| < 1$.
- This implies that parameter restrictions will need to be imposed when estimating the MA model, to enforce selection of the invertible model.

AR(p) processes

An AR(p) process can be represented as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

This is just a linear regression model, and assuming stationarity, we can estimate the parameters by OLS. What is needed for stationarity?

The dynamic behavior of an AR(p) process can be studied by writing this p^{th} order difference

equation as a vector first order difference equation (this is known as the companion form):

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} \phi_1 & \phi_2 & \cdots & & \phi_p \\ 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \cdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

or

$$Y_t = C + FY_{t-1} + E_t$$

With this, we can recursively work forward in time:

$$\begin{aligned} Y_{t+1} &= C + FY_t + E_{t+1} \\ &= C + F\left(C + FY_{t-1} + E_t\right) + E_{t+1} \\ &= C + FC + F^2Y_{t-1} + FE_t + E_{t+1} \end{aligned}$$

and

$$\begin{aligned} Y_{t+2} &= C + FY_{t+1} + E_{t+2} \\ &= C + F\left(C + FC + F^2Y_{t-1} + FE_t + E_{t+1}\right) + E_{t+2} \\ &= C + FC + F^2C + F^3Y_{t-1} + F^2E_t + FE_{t+1} + E_{t+2} \end{aligned}$$

or in general

$$Y_{t+j} = C + FC + \cdots + F^j C + F^{j+1} Y_{t-1} + F^j E_t + F^{j-1} E_{t+1} + \cdots + F E_{t+j-1} + E_{t+j}$$

Consider the impact of a shock in period t on y_{t+j}. This is simply

$$\frac{\partial Y_{t+j}}{\partial E_t'}_{(1,1)} = F^j_{(1,1)}$$

If the system is to be stationary, then as we move forward in time this impact must die off. Otherwise a shock causes a permanent change in the mean of y_t. Therefore, stationarity requires that

$$\lim_{j \to \infty} F^j_{(1,1)} = 0$$

- Save this result, we'll need it in a minute.

Consider the eigenvalues of the matrix F. These are the λ such that

$$|F - \lambda I_P| = 0$$

The determinant here can be expressed as a polynomial. For example, for $p = 1$, the matrix F is simply

$$F = \phi_1$$

so

$$|\phi_1 - \lambda| = 0$$

can be written as

$$\phi_1 - \lambda = 0$$

When $p = 2$, the matrix F is

$$F = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix}$$

so

$$F - \lambda I_P = \begin{bmatrix} \phi_1 - \lambda & \phi_2 \\ 1 & -\lambda \end{bmatrix}$$

and

$$|F - \lambda I_P| = \lambda^2 - \lambda\phi_1 - \phi_2$$

So the eigenvalues are the roots of the polynomial

$$\lambda^2 - \lambda\phi_1 - \phi_2$$

which can be found using the quadratic equation. This generalizes. For a p^{th} order AR process, the eigenvalues are the roots of

$$\lambda^p - \lambda^{p-1}\phi_1 - \lambda^{p-2}\phi_2 - \cdots - \lambda\phi_{p-1} - \phi_p = 0$$

Supposing that all of the roots of this polynomial are distinct, then the matrix F can be factored as

$$F = T\Lambda T^{-1}$$

where T is the matrix which has as its columns the eigenvectors of F, and Λ is a diagonal matrix with

the eigenvalues on the main diagonal. Using this decomposition, we can write

$$F^j = \left(T\Lambda T^{-1}\right)\left(T\Lambda T^{-1}\right)\cdots\left(T\Lambda T^{-1}\right)$$

where $T\Lambda T^{-1}$ is repeated j times. This gives

$$F^j = T\Lambda^j T^{-1}$$

and

$$\Lambda^j = \begin{bmatrix} \lambda_1^j & 0 & & 0 \\ 0 & \lambda_2^j & & \\ & & \ddots & \\ 0 & & & \lambda_p^j \end{bmatrix}$$

Supposing that the λ_i $i = 1, 2, ..., p$ are all real valued, it is clear that

$$\lim_{j\to\infty} F^j_{(1,1)} = 0$$

requires that

$$|\lambda_i| < 1, i = 1, 2, ..., p$$

e.g., the eigenvalues must be less than one in absolute value.

- It may be the case that some eigenvalues are complex-valued. The previous result generalizes to the requirement that the eigenvalues be less than one in *modulus,* where the modulus of a complex number $a + bi$ is

$$mod(a + bi) = \sqrt{a^2 + b^2}$$

This leads to the famous statement that "stationarity requires the roots of the determinantal polynomial to lie inside the complex unit circle." *draw picture here.*

- When there are roots on the unit circle (unit roots) or outside the unit circle, we leave the world of stationary processes.
- Dynamic multipliers: $\partial y_{t+j}/\partial \varepsilon_t = F^j_{(1,1)}$ is a *dynamic multiplier* or an *impulse-response* function. Real eigenvalues lead to steady movements, whereas complex eigenvalues lead to ocillatory behavior. Of course, when there are multiple eigenvalues the overall effect can be a mixture. *pictures*

Moments of AR(p) process

The AR(p) process is

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Assuming stationarity, $\mathcal{E}(y_t) = \mu, \forall t$, so

$$\mu = c + \phi_1 \mu + \phi_2 \mu + ... + \phi_p \mu$$

so

$$\mu = \frac{c}{1 - \phi_1 - \phi_2 - ... - \phi_p}$$

and

$$c = \mu - \phi_1 \mu - ... - \phi_p \mu$$

so

$$\begin{aligned} y_t - \mu &= \mu - \phi_1\mu - ... - \phi_p\mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \mu \\ &= \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t \end{aligned}$$

With this, the second moments are easy to find: The variance is

$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + ... + \phi_p\gamma_p + \sigma^2$$

The autocovariances of orders $j \geq 1$ follow the rule

$$\begin{aligned} \gamma_j &= \mathcal{E}\left[(y_t - \mu)\left(y_{t-j} - \mu\right)\right] \\ &= \mathcal{E}\left[(\phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t)\left(y_{t-j} - \mu\right)\right] \\ &= \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + ... + \phi_p\gamma_{j-p} \end{aligned}$$

Using the fact that $\gamma_{-j} = \gamma_j$, one can take the $p+1$ equations for $j = 0, 1, ..., p$, which have $p+1$ unknowns (σ^2, $\gamma_0, \gamma_1, ..., \gamma_p$) and solve for the unknowns. With these, the γ_j for $j > p$ can be solved for recursively.

ARMA model

An ARMA(p, q) model is $(1 + \phi_1 L + \phi_2 L^2 + ... + \phi_p L^p)y_t = c + (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q)\epsilon_t$. These are popular in applied time series analysis. A high order AR process *may* be well approximated by a low order MA process, and a high order MA process *may* be well approximated by a low order AR

process. By combining low order AR and MA processes in the same model, one can hope to fit a wide variety of time series using a parsimonious number of parameters. There is much literature on how to choose p and q, which is outside the scope of this course. Estimation can be done using the Kalman filter, assuming that the errors are normally distributed.

15.2 VAR models

Consider the model

$$\begin{aligned} y_t &= C + A_1 y_{t-1} + \epsilon_t \\ E(\epsilon_t \epsilon_t') &= \Sigma \\ E(\epsilon_t \epsilon_s') &= 0, t \neq s \end{aligned} \tag{15.3}$$

where y_t and ϵ_t are $G \times 1$ vectors, C is a $G \times 1$ of constants, and $A_1$is a $G \times G$ matrix of parameters. The matrix Σ is a $G \times G$ covariance matrix. Assume that we have n observations. This is a *vector autoregressive* model, of order 1 - commonly referred to as a VAR(1) model. It is a collection of G AR(1) models, augmented to include lags of other endogenous variables, and the G equations are contemporaneously correlated. The extension to a VAR(p) model is quite obvious.

As shown in Section 10.3, it is efficient to estimate a VAR model using OLS equation by equation, there is no need to use GLS, in spite of the cross equation correlations.

A VAR model of this form can be thought of as the reduced form of a dynamic simultaneous equations system, with all of the variables treated as endogenous, and with lags of all of the endogenous

variables present. The simultaneous equations model is (see equation 10.1)

$$Y_t'\Gamma = X_t'B + E_t'$$

which can be written after transposing (and adapting notation to use small case, pulling the constant out of X_t and using v_t for the error) as $\Gamma' y_t = a + B' x_t + v_t$. Let $x_t = y_{t-1}$. Then we have $\Gamma' y_t = a + B' y_{t-1} + v_t$. Premultiplying by the inverse of Γ' gives

$$y_t = (\Gamma')^{-1} a + (\Gamma')^{-1} B' y_{t-1} + (\Gamma')^{-1} v_t.$$

Finally define $C = (\Gamma')^{-1} a$, $A_1 = (\Gamma')^{-1} B'$ and $\epsilon_t = (\Gamma')^{-1} v_t$, and we have the VAR(1) model of equation 15.3. C. Sims originally proposed reduced form VAR models as an alternative to structural simultaneous equatons models, which were perceived to require too many unrealistic assumptions for their identification. However, the search for structural interpretations of VAR models slowly crept back into the literature, leading to "structural VARs". A structural VAR model is really just a dynamic linear simultaneous equations model, with other imaginative and hopefully more realistic methods used for identification. The issue of identifying the structural parameters Γ and B is more or less the same problem that was studied in the context of simultaneous equatons. There, identification was obtained through zero restrictions. In the structural VAR literature, zero restrictions are often used, but other information may also be used, such as covariance matrix restrictions or sign restrictions. Interest often focuses on the impulse-response functions. Identification of the impact of structural shocks (how to estimate the impact-response functions) is complicated, with many alternative methodologies, and is often a topic of much disagreement among practitioners. The estimated impulse response functions are often sensitive to the identification strategy that is used. There is a large literature. Papers by C.

Sims are a good place to start, if one wants to learn more. He also offers a good deal of useful software on his web page.

An issue which arises when a VAR(p) model $y_t = C + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t$ is contemplated is that the number of parameters increases rapidly in p, which introduces severe collinearity problems. One can use Bayesian methods such as the "Minnesota prior" (Litterman), which is a prior that each variable separately follows a random walk (an AR(1) model with $\rho = 1$). The prior on A_1 is that it is an identity matrix, and the prior on the A_j, $j > 1$ is that they are zero matrices. This can be done using stochastic restrictions similar to what was in the discussion of collinearity and ridge regression. For example, a VAR(2) model in de-meaned variables, with G variables, can be written as

$$Y = \begin{bmatrix} Y_{-1} & Y_{-2} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} + \epsilon$$

We can impose the stochastic restriction that $A_1 = I_2 - v_1$ and that $A_2 = 0_2 - v_2$. Augmenting the data with these 4 "artificial observations", we get

$$\begin{bmatrix} Y \\ I_G \\ 0_G \end{bmatrix} = \begin{bmatrix} Y_{-1} & Y_{-2} \\ I_G & 0_G \\ 0_G & I_G \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} + \begin{bmatrix} \epsilon \\ v_1 \\ v_2 \end{bmatrix}$$

Then we can impose how important the restrictions are by weighting the stochastic restrictions, along

the lines of a GLS heteroscedasticity correction:

$$\begin{bmatrix} Y \\ k_1 I_G \\ 0_G \end{bmatrix} = \begin{bmatrix} Y_{-1} & Y_{-2} \\ k_1 I_G & 0_G \\ 0_G & k_2 I_G \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} + \begin{bmatrix} \epsilon \\ k_1 v_1 \\ k_2 v_2 \end{bmatrix}$$

Then we fit by OLS. When k_1 is small, the estimated A_1 will be forced to be close to an identity matrix. When k_2 is small, the second lag coefficients are all forced to zero. Jointly, these restrictions push the model in the direction of separate random walks for each variable. The degree to which the model is pushed depends on the ks. When the ks are large, the fit is close to the unrestricted OLS fit. An example is given in BVar.m

"Bayesian VARs" is a now a substantial body of literature. An introduction to more formal Bayesian methods is given in a chapter that follows. For highly parameterized models, Bayesian methods can help to impose structure.

15.3 ARCH, GARCH and Stochastic volatility

ARCH (autoregressive conditionally heteoscedastic) models appeared in the literature in 1982, in Engle, Robert F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation", Econometrica 50:987-1008. This paper stimulated a very large growth in the literature for a number of years afterward. The related GARCH (generalized ARCH) model is now one of the most widely used models for financial time series.

Financial time series often exhibit several type of behavior:

- volatility clustering: periods of low variation can be followed by periods of high variation

- fat tails, or excess kurtosis: the marginal density of a series is more strongly peaked and has fatter tails than does a normal distribution with the same mean and variance.
- other features, such as leverage (correlation between returns and volatility) and perhaps slight autocorrelation within the bounds allowed by arbitrage.

The data set "nysewk.gdt", which is provided with Gretl, provides an example. If we compute 100 times the growth rate of the series, using log differences, we can obtain the plots in Figure 15.1. In the first we clearly see volatility clusters, and in the second, we see excess kurtosis and tails fatter than the normal distribution. The skewness suggests that leverage may be present.

- regress returns on its own lag and on squared returns and lags: low predictability
- regress squared returns on its own lags and on returns: more predictable, evidence of leverage

The presence of volatility clusters indicates that the variance of the series is not constant over time, conditional on past events. Engle's ARCH paper was the first to model this feature.

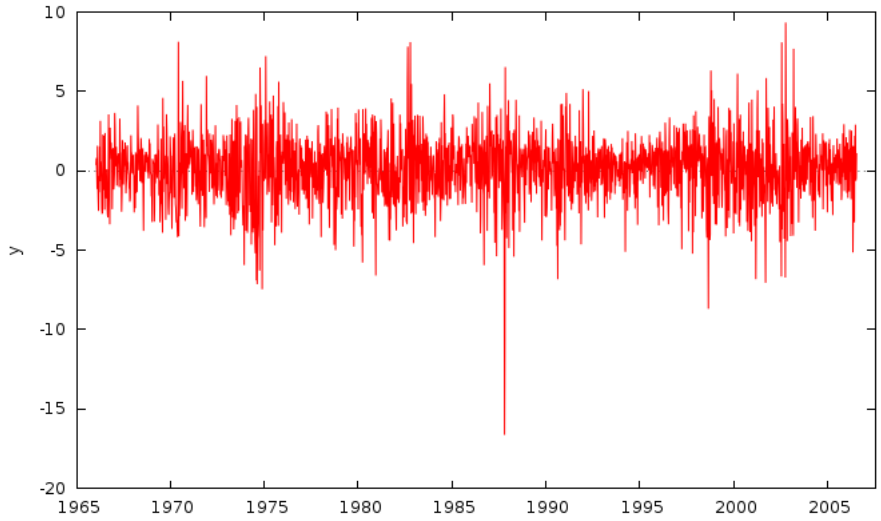
ARCH

A basic ARCH specification is

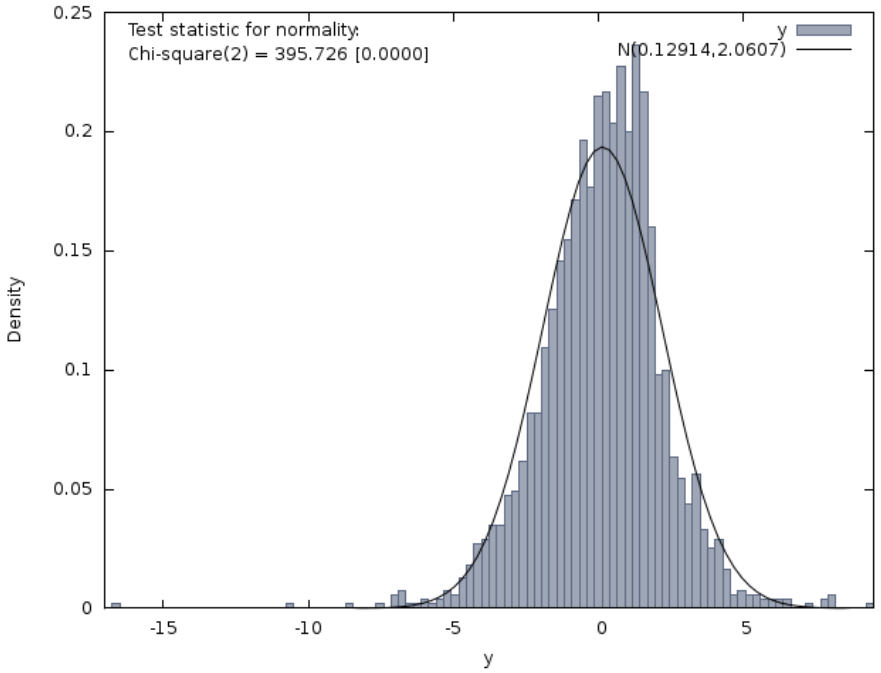
$$\begin{aligned} y_t &= \mu + \rho y_{t-1} + \epsilon_t \\ &\equiv g_t + \epsilon_t \\ \epsilon_t &= \sigma_t u_t \\ \sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 \end{aligned}$$

Figure 15.1: NYSE weekly close price, 100 ×log differences

(a) Time series plot



(b) Frequency distribution



where the u_t are Gaussian white noise shocks. The ARCH variance is a moving average process. Previous large shocks to the series cause the conditional variance of the series to increase. There is no leverage: negative shocks have the same impact on the future variance as do positive shocks..

- for σ_t^2 to be positive for all realizations of $\{\epsilon_t\}$, we need $\omega > 0$, $\alpha_i \geq 0$, $\forall i$.
- to ensure that the model is covariance stationary, we need $\sum_i \alpha_i < 1$. Otherwise, the variances will explode off to infinity.

Given that ϵ_t is normally distributed, to find the likelihood in terms of the observable y_t instead of the unobservable ϵ_t, first note that the series $u_t = (y_t - g_t)/\sigma_t = \frac{\epsilon_t}{\sigma_t}$ is iid Gaussian, so the likelihood is simply the product of standard normal densities.

$$\begin{aligned} u &\sim N(0, I), \text{ so} \\ f(u) &= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_t^2}{2}\right) \end{aligned}$$

The joint density for y can be constructed using a change of variables:

- We have $u_t = (y_t - \mu - \rho y_{t-1})/\sigma_t$, so $\frac{\partial u_t}{\partial y_t} = \frac{1}{\sigma_t}$ and $|\frac{\partial u}{\partial y'}| = \Pi_{t=1}^{n} \frac{1}{\sigma_t}$,
- doing a change of variables,

$$f(y;\theta) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_t} \exp\left(-\frac{1}{2}\left(\frac{y_t - \mu - \rho y_{t-1}}{\sigma_t}\right)^2\right)$$

  where θ includes the parameters in g_t and the alpha parameters of the ARCH specification.

Taking logs,

$$\ln L(\theta) = -n \ln \sqrt{2\pi} - \sum_{t=1}^{n} \ln \sigma_t - \frac{1}{2} \sum_{t=1}^{n} \left(\frac{y_t - \mu - \rho y_{t-1}}{\sigma_t} \right)^2 .$$

In principle, this is easy to maximize. Some complications can arise when the restrictions for positivity and stationarity are imposed. Consider a fairly short data series with low volatility in the initial part, and high volatility at the end. This data appears to have a nonstationary variance sequence. If one attempts to estimate and ARCH model with stationarity imposed, the data and the restrictions are saying two different things, which can make maximization of the likelihood function difficult.

The Octave script ArchExample.m illustrates estimation of an ARCH(1) model, using the NYSE closing price data.

GARCH

Note that an ARCH model specifies the variance process as a moving average. For the same reason that an ARMA model may be used to parsimoniously model a series instead of a high order AR or MA, one can do the same thing for the variance series. A basic GARCH(p,q) (Bollerslev, Tim (1986). "Generalized Autoregressive Conditional Heteroskedasticity", Journal of Econometrics, 31:307-327) specification is

$$\begin{aligned} y_t &= \mu + \rho y_{t-1} + \epsilon_t \\ \epsilon_t &= \sigma_t u_t \\ \sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \end{aligned}$$

The idea is that a GARCH model with low values of p and q may fit the data as well or better than an ARCH model with large q.

- the model also requires restrictions for positive variance and stationarity, which are:
  - $\omega > 0$
  - $\alpha_i \geq 0,\ i = 1, ..., q$
  - $\beta_i \geq 0,\ i = 1, ..., p$
  - $\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i < 1.$
- to estimate a GARCH model, you need to initialize σ_0^2 at some value. The sample unconditional variance is one possibility. Another choice could be the sample variance of the initial elements of the sequence. One can also "backcast" the conditional variance.

The GARCH model also requires restrictions on the parameters to ensure stationarity and positivity of the variance. A useful modification is the EGARCH model (exponential GARCH, Nelson, D. B. (1991). "Conditional heteroskedasticity in asset returns: A new approach", Econometrica 59: 347-370). This model treats the logarithm of the variance as an ARMA process, so the variance will be positive without restrictions on the parameters. It is also possible to introduce asymmetry (leverage) and non-normality.

The Octave script GarchExample.m illustrates estimation of a GARCH(1,1) model, using the NYSE closing price data. You can get the same results more quickly using Gretl, which takes advantage of C code for the model. If you play with the example, you can see that the results are sensitive to start values. The likelihood function does not appear to have a nice well-defined global maximum. Thus,

one needs to use care when estimating this sort of model, or rely on some software that is known to work well.

Note that the test of homoscedasticity against ARCH or GARCH involves parameters being on the boundary of the parameter space. Also, the reducton of GARCH to ARCH has the same problem. Testing needs to be done taking this into account. See Demos and Sentana (1998) *Journal of Econometrics*.

Stochastic volatility

In ARCH and GARCH models, the same shocks that affect the level also affect the variance. The stochastic volatility model allows the variance to have its own random component. A simple example is

$$y_t = \exp(h_t)\epsilon_t$$
$$h_t = \alpha + \rho h_{t-1} + \sigma\eta_t$$

In this model, the log of the standard error of the observed sequence follows an AR(1) model. Once can introduce leverage by allowing correlation between ϵ_t and h_t. Variants of this sort of model are widely used to model financial data, competing with the GARCH(1,1) model for being the most popular choice. Many estimation methods have been proposed.

15.4 Diffusion models

Financial data is often modeled using a continuous time specification. An example is the following model, taken from a paper of mine (with D. Kristensen).

A basic model is a simple continuous time stochastic volatility model with leverage. Log price $p_t = \log(P_t)$, solves the following pure diffusion model,

$$dp_t = (\mu_0 + \mu_1 \exp(h_t - \alpha))\, dt + \exp\left(\frac{h_t}{2}\right) dW_{1,t}$$

where the spot volatility (the instantaneous variance of returns), $\exp(h_t)$ is modeled using its logarithm:

$$dh_t = \kappa(\alpha - h_t)dt + \sigma dW_{2,t}.$$

Here, $W_{1,t}$ and $W_{2,t}$ are two standard Brownian motions with instantaneous correlation $\rho = Cov(dW_{1,t}, dW_{2,t})$. This model is the well-known log-Normal volatility model of Wiggins (1987); see also Chesney and Scott (1989). The parameters are interpreted as follows: μ_0 is the baseline drift of returns; μ_1 allows drift to depend upon spot volatility; α is the mean of log volatility; κ is the speed of mean reversion of log volatility, such that low values of κ imply high persistence of log volatility; σ is the so-called volatility of volatility; and ρ is a leverage parameter that affects the correlation between returns and log volatility. We collect the parameters in $\theta = (\mu_0, \mu_1, \alpha, \kappa, \sigma, \rho)$.

An extension is to add jumps to the above model. These occur with Poisson frequency, and are conditionally normally distributed. More specifically, log-price p_t solves the following continuous-time

jump-diffusion model,

$$dp_t = (\mu_0 + \mu_1 \exp(h_t/2))\, dt + \exp\left(\frac{h_t}{2}\right) dW_{1,t} + J_t dN_t.$$

The Poisson process N_t counts the number of jumps up to time t, and has jump intensity $\lambda_t = \lambda_0 + \lambda_1 \exp(h_t - \alpha)$ that varies with the volatility, while jump sizes, conditional on the occurence of a jump, are independent and conditionally normally distributed: $J_t | \mathcal{F}_{t-} > 0 \sim N(\mu_J, \sigma_J^2)$, where $\mathcal{F}_t$ is the standard filtration. The inclusion of the jump component adds four parameters to θ as defined above, μ_J, σ_J^2 and λ_0, and λ_1. This jump model was considered in, for example, Andersen, Benzoni and Lund (2002).

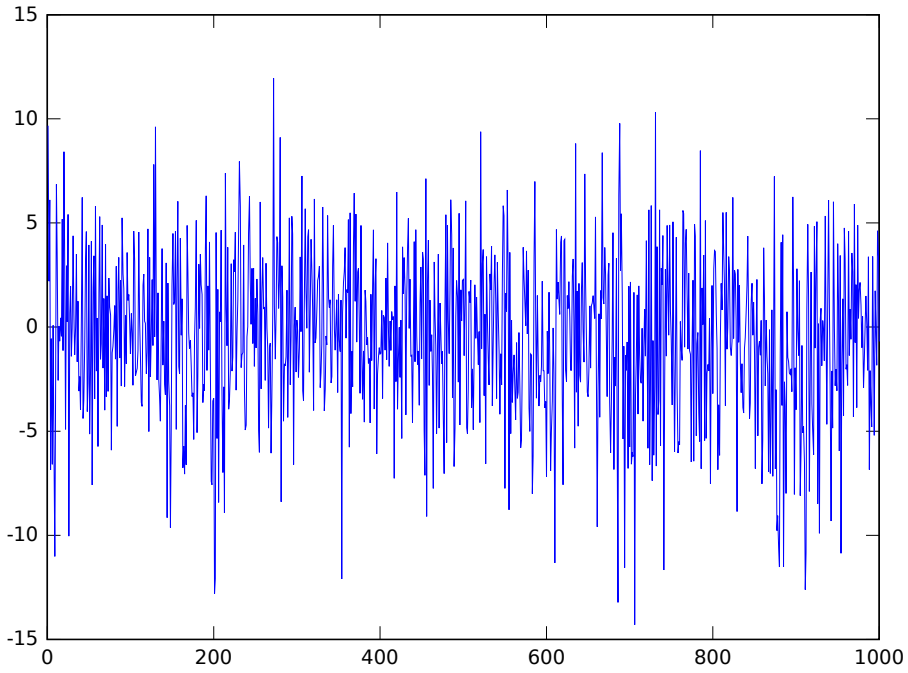
An example of how returns, $r_t = 100(p_t - p_{t-1})$, generated by such a model might look is given in Figure 15.2. The spot volatility is plotted in Figure 15.3. Returns are observable, but spot volatility is not.

One might want to try to infer the parameters of the model, and also the latent spot volatility, using the observable data. Estimation of the parameters of such models is complicated by the fact that data is available in discrete time: $p_1, p_2, ...p_n$, but the model is in continuous time. One can "discretize" the model, to obtain something like the discrete time SV model of the previous section, but the discrete time transition density implied by the approximating model is not the same as the true transition density

$$p_t \sim f_p(p_t | p_{t-1}, h_{t-1}; \theta),$$

induced by the continuous time model. This true density is unknown, however, so using it for ML estimation is not possible. If one estimates the discrete time version treating it as the actual density, there is an approximation misspecification that causes the estimates to be inconsistent: we're not

Figure 15.2: Returns from jump-diffusion model

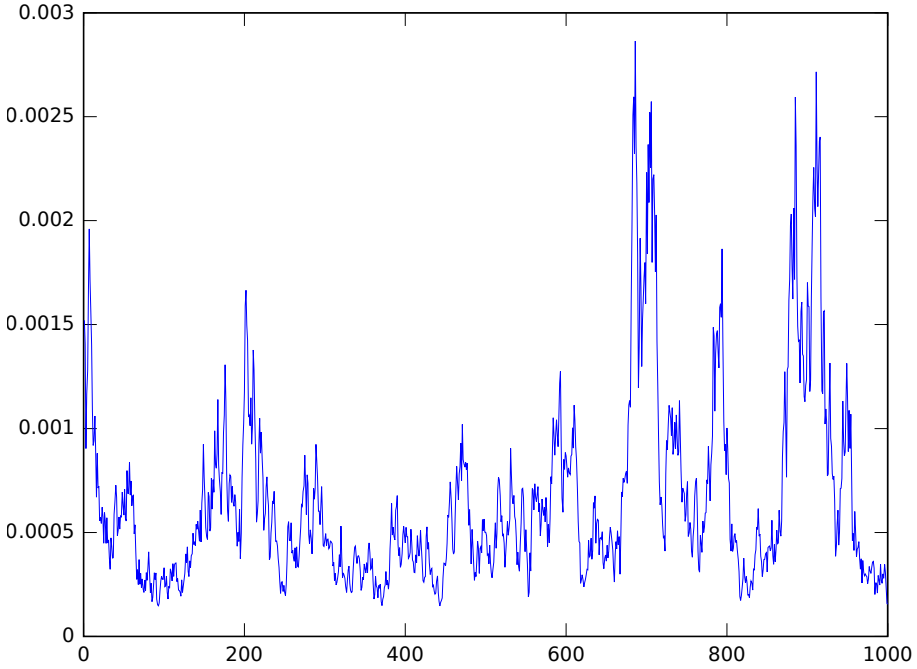


doing ML, we're doing quasi-ML, which is in general an inconsistent estimator. Consistent estimation of parameters is discussed in Section 22.1, in the Chapter on simulation-based estimation. A means of learning about spot volatility, h_t, is discussed in the chapter on nonparametric inference, in Section 20.5.

15.5 State space models

For linear time series models with Gaussian shocks, it is often useful to put the model in state space form, as in this form, the Kalman filter provides a convenient way to compute the likelihood function.

Figure 15.3: Spot volatility, jump-diffusion model



For example, with an MA model, we can compute the likelihood function using the joint density of the whole sample, $y \sim N(0, \Sigma)$ where Σ is an $n \times n$ matrix that depends on σ^2 and ϕ, The log likelihood is $f(y|\sigma^2, \phi)$

See Fernández-Villaverde's notes Fernández-Villaverde's Kalman filter notes and Mikusheva's MIT OpenCourseWare notes, lectures 20 and 21: Mikusheva's Kalman filter notes. I will follow Mikusheva's notes in class.

For nonlinear state space models, or non-Gaussian state space models, the basic Kalman filter cannot be used, and the particle filter is becoming a widely-used means of computing the likelihood. This is a fairly new, computationally demanding technique, and is currently (this was written in 2013) an active area of research. Papers by Fernández-Villaverde and Rubio-Ramírez provide interesting and reasonably accessible applications in the context of estimating macroeconomic (DSGE) models.

15.6 Nonstationarity and cointegration

I'm going to follow Karl Whelan's notes, which are available at Whelan notes.

15.7 Exercises

1. Use Matlab to estimate the same GARCH(1,1) model as in the GarchExample.m script provided above. Also, estimate an ARCH(4) model for the same data. If unconstrained estimation does not satisfy stationarity restrictions, then do contrained estimation. Compare likelihood values. Which of the two models do you prefer? But do the models have the same number of parameters? Find out what is the "consistent Akaike information criterion" or the "Bayes

information criterion" and what they are used for. Compute one or the other, or both, and discuss what they tell you about selecting between the two models.

Chapter 16

Bayesian methods

References I have used to prepare these notes: Cameron and Trivedi, *Microeconometrics: Methods and Applications*, Chapter 13; Chernozhukov and Hong (2003), "An MCMC approach to classical estimation", *Journal of Econometrics;* Gallant and Tauchen, "EMM: A program for efficient method of moments estimation"; Hoogerheide, van Dijk and van Oest (2007) "Simulation Based Bayesian Econometric Inference: Principles and Some Recent Computational Advances". You might also like to read See Mikusheva's MIT OpenCourseWare notes, lectures 23, 24 and 25: Bayesian notes.

This chaper provides a brief introduction to Bayesian methods, which form a large part of econometric research, especially in the last two decades. Advances in computational methods (e.g., MCMC, particle filtering), combined with practical advantages of Bayesian methods (e.g., no need for minimization and improved identification coming from the prior) have contributed to the popularity of this approach.

16.1 Definitions

The Bayesian approach treats the parameter of a model as a random vector. The parameter has a density, $\pi(\theta)$, which is known as the *prior*. It is assumed that the econometrician can provide this density, which reflects current beliefs about the parameter.

We also have sample information, $y=\{y_1, y_2, ...y_n\}$. We're already familiar with the *likelihood function,* $f(y|\theta)$, which is the density of the sample given a parameter value.

Given these two pieces, we can write the joint density of the sample and the parameter:

$$f(y, \theta) = f(y|\theta)\pi(\theta)$$

We can get the *marginal likelihood* by integrating out the parameter, integrating over its support Θ:

$$f(y) = \int_{\Theta} f(y, \theta) d\theta$$

The last step is to get the *posterior* of the parameter. This is simply the density of the parameter conditional on the sample, and we get it in the normal way we get a conditional density, using Bayes' theorem

$$f(\theta|y) = \frac{f(y, \theta)}{f(y)} = \frac{f(y|\theta)\pi(\theta)}{f(y)}$$

The posterior reflects the learning that occurs about the parameter when one receives the sample information. The sources of information used to make the posterior are the prior and the likelihood function. Once we have the posterior, one can provide a complete probabilistic description about our updated beliefs about the parameter, using quantiles or moments of the posterior. The posterior mean or median provide the Bayesian analogue of the frequentist point estimator in the form of the ML

estimator. We can define regions analogous to confidence intervals by using quantiles of the posterior, or the marginal posterior.

So far, this is pretty straightforward. The complications are often computational. To illustrate, the posterior mean is

$$E(\theta|y) = \int_\Theta \theta f(\theta|y) d\theta = \frac{\int_\Theta \theta f(y|\theta)\pi(\theta) d\theta}{\int_\Theta f(y,\theta) d\theta}$$

One can see that a means of integrating will be needed. Only in very special cases will the integrals have analytic solutions. Otherwise, computational methods will be needed.

16.2 Philosophy, etc.

So, the classical paradigm views the data as generated by a data generating process, which is a perhaps unknown model characterized by a parameter vector, and the data is generated from the model at a particular value of the parameter vector. Bayesians view data as given, and update beliefs about a random parameter using the information about the parameter contained in the data.

Bayesians and frequentists have a long tradition of arguing about the meaning and interpretation of their respective procedures. Here's my take on the debate. Fundamentally, I take the frequentist view: I find it pleasing to think about a model with a fixed non-random parameter about which we would like to learn. I like the idea of a point estimator that gives a best guess about the true parameter. However, we shouldn't reinvent the wheel each time we get a new sample: previous samples have information about the parameter, and we should use all of the available information. A pure frequentist approach would require writing the joint likelihood of all samples, which would almost certainly constitute an impossible task. The Bayesian approach concentrates all of the information coming from previous work in the form of a prior. A fairly simple, easy to use prior may not exactly capture all previous

information, but it could offer a handy and reasonably accurate summary. So, the idea of a prior as a summary of what we have learned may simply be viewed as a practical solution to the problem of using all the available information. Given that it's a summary, one may as well use a convenient form, as long as it's plausible and the results don't depend too exaggerately on the prior.

About the likelihood function, fortunately, Bayesians and frequentists are in agreement, so there's no need for further comment.

When we get to how to generate and interpret results, there is some divergence. Frequentists maximize the likelihood function, and compute standard errors, etc., using the methods already explained in these notes. A frequentist could test the hypothesis that $\theta_0 = \theta^*$ by seeing if the data are sufficiently likely conditional on the parameter value θ^*. A Bayesian would check if θ^* is a plausible value conditional on the observed data.

I have criticized the frequentist practice of using only the current sample, ignoring what previous work has told us about the parameter, simply because it's too hard to write the overall joint likelihood for all samples. So to be fair, here's a criticism of the Bayesian approach. If we're doing Bayesian learning, what is it we're learning about? If it's not a fixed parameter value then what is it? What is the process that generated the sample data? If the parameter is random, was the sample generated at a single realization, or at many realizations? If we had an infinite sample, then the Bayesian estimators (e.g., posterior mean or median) converge to a point. What is that point if it's not the same true parameter value that the frequentists are trying to estimate? Why would one use noninformative priors for one's whole career - don't we believe what we learned from the last paper we wrote? These questions often receive no answer, or obscure answers.

It turns out that one can analyze Bayesian estimators from a classical (frequentist) perspective. It also turns out that Bayesian estimators may be easier to compute reliably than analogous classical

estimators. These computational advantages, combined with the ability to use information from previous work in an intelligent way, make the study of Bayesian methods attractive for frequentists. If a Bayesian takes the view that there is a fixed data generating process, and Bayesian learning leads in the limit to the same fixed true value that frequentists posit, then the study of frequentist theory will be useful to a Bayesian practitioner.

For the rest of this, I will adopt the classical, frequentist perspective, and study the behavior of Bayesian estimators in this context.

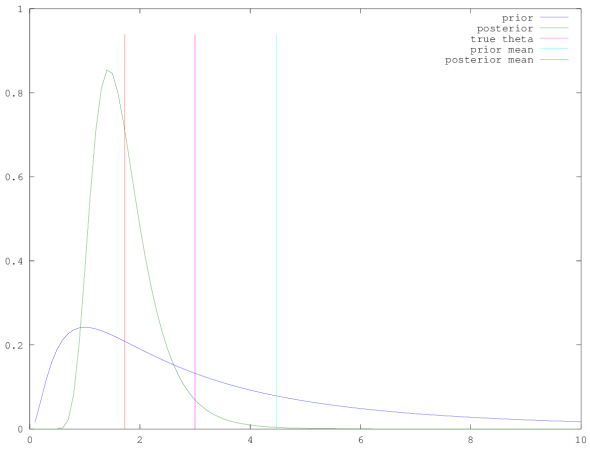
16.3 Example

Suppose data is generated by i.i.d. sampling from an exponential distribution with mean θ. An exponential random variable takes values on the positive real numbers. Waiting times are often modeled using the exponential distribution.

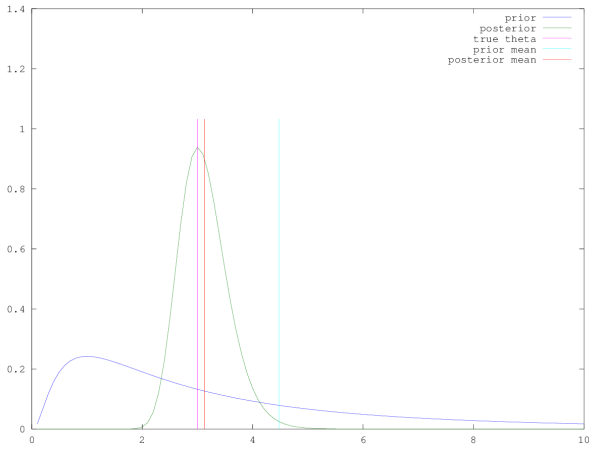
- The density of a typical sample element is $f(y|\theta) = \frac{1}{\theta}e^{-y/\theta}$. The likelihood is simply the product of the sample contributions.
- Suppose the prior for θ is $\theta \sim$lognormal(1,1). This means that the logarithm of θ is standard normal. We use a lognormal prior because it enforces the requirement that the parameter of the exponential density be positive.
- The Octave script BayesExample1.m implements Bayesian estimation for this setup.
- with a sample of 10 observations, we obtain the results in panel (a) of Figure 16.1, while with a sample of size 50 we obtain the results in panel (b). Note how the posterior is more concentrated

Figure 16.1: Bayesian estimation, exponential likelihood, lognormal prior

(a) N=10



(b) N=50



around the true parameter value in panel (b). Also note how the posterior mean is closer to the prior mean when the sample is small. When the sample is small, the likelihood function has less weight, and more of the information comes from the prior. When the sample is larger, the likelihood function will have more weight, and its effect will dominate the prior's.

16.4 Theory

Chernozhukov and Hong (2003) "An MCMC Approach to Classical Estimation" http://www.sciencedirect.com/science/article/pii/S0304407603001003 is a very interesting article that shows how Bayesian methods may be used with criterion functions that are associated with classical estimation techniques. For example, it is possible to compute a posterior mean version of a GMM estimator. Chernozhukov and Hong provide their Theorem 2, which proves consistency and asymptotic normality for a general

Figure 16.2: Chernozhukov and Hong, Theorem 2

Theorem 2 (LTE in large samples). *Under Assumptions* 1–4,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \xi_{J_n(\theta_0)} + U_n + o_p(1), \quad \Omega_n^{-1/2}(\theta_0) J_n(\theta_0) U_n \to_d \mathcal{N}(0, I).$$

Hence

$$\Omega_n^{-1/2}(\theta_0) J_n(\theta_0)(\sqrt{n}(\hat{\theta} - \theta_0) - \xi_{J_n(\theta_0)}) \to_d \mathcal{N}(0, I).$$

If the loss function ρ_n is symmetric, i.e. $\rho_n(h) = \rho_n(-h)$ for all h, $\xi_{J_n(\theta_0)} = 0$ for each n.

class of such estimators. When the criterion function $L_n(\theta)$ in their paper is set to the log-likelihood function, the pseudo-prior $\pi(\theta)$ is a real Bayesian prior, and the penalty function ρ_n is the squared loss function (see the paper), then the class of estimators discussed by CH reduces to the ordinary Bayesian posterior mean. As such, their Theorem 2, in Figure 16.2 tells us that this estimator is consistent and asymptotically normally distributed. In particular, the Bayesian posterior mean has the same asymptotic distribution as does the ordinary maximum likelihood estimator.

- the intuition is clear: as the amount of information coming from the sample increases, the likelihood function brings an increasing amount of information relative to the prior. Eventually, the prior is no longer important for determining the shape of the posterior.

- when the sample is large, the shape of the posterior depends on the likelihood function. The likelihood function collapses around θ_0 when the sample is generated at θ_0. The same is true of the posterior, it narrows around θ_0. This causes the posterior mean to converge to the true parameter value. In fact, all quantiles of the posterior converge to θ_0. Chernozhukov and Hong

discuss estimators defined using quantiles.

- For an econometrician coming from the frequentist perspective, this is attractive. The Bayesian estimator has the same asymptotic behavior as the MLE. There may be computational advantages to using the Bayesian approach, because there is no need for optimization. If the objective function that defines the classical estimator is irregular (multiple local optima, nondifferentiabilities, noncontinuities...), then optimization may be very difficult. However, Bayesian methods that use integration may be more tractible. Tthis is the main motivation of CH's paper. Additional advantages include the benefits if an informative prior is available. When this is the case, the Bayesian estimator can have better small sample performance than the maximum likelihood estimator.

16.5 Computational methods

- To compute the posterior mean, we need to evaluate $E(\theta|y) = \int_\Theta \theta f(\theta|y)d\theta = \int_\Theta \theta f(y|\theta)\pi(\theta)d\theta / \int_\Theta f(y,\theta)d\theta$.
- Note that both of the integrals are multiple integrals, with the dimension given by that of the parameter, θ.
- Under some special circumstances, the integrals may have analytic solutions: e.g., Gaussian likelihood with a Gaussian prior leads to a Gaussian posterior.
- When the dimension of the parameter is low, quadrature methods may be used. What was done in as was done in BayesExample1.m is an unsophisticated example of this. More sophisticated

methods use an intelligently chosen grid to reduce the number of function evaluations. Still, these methods only work for dimensions up to 3 or so.

- Otherwise, some form of simulation-based "Monte Carlo" integration must be used. The basic idea is that $E(\theta|y)$ can be approximated by $(1/S)\sum_{s=1}^{S}\theta^s$, where θ^s is a random draw from the posterior distribution $f(\theta|y)$. The *trick is how to make draws from the posterior* when in general we can't compute the posterior.

  - the law of large numbers tells us that this average will converge to the desired expectation as S gets large
  - convergence will be more rapid if the random draws are independent of one another, but insisting on independence may have computational drawbacks.

- Monte Carlo methods include importance sampling, Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC, also known as particle filtering). The great expansion of these methods over the years has caused Bayesian econometrics to become much more widely used than it was in the not so distant (for some of us) past. There is much literature - here we will only look at a basic example that captures the main ideas.

MCMC

Variants of Markov chain Monte Carlo have become a very widely used means of computing Bayesian estimates. See Tierney (1994) "Markov Chains for Exploring Posterior Distributions" *Annals of Statistics* and Chib and Greenberg (1995) "Understanding the Metropolis-Hastings algorithm" *The American Statistician.*

Let's consider the basic Metropolis-Hastings MCMC algorithm. We will generate a long realization of a Markov chain process for θ, as follows:

The prior density is $\pi(\theta)$, as above. Let $g(\theta^*; \theta^s)$ be a proposal density, which generates a new trial parameter value θ^* given the most recently accepted parameter value θ^s. A proposal will be accepted if

$$\frac{f(\theta^*|y)}{f(\theta^s|y)} \frac{g(\theta^s; \theta^*)}{g(\theta^*; \theta^s)} > \alpha$$

where α is a $U(0, 1)$ random variate.

There are two parts to the numerator and denominator: the posterior, and the proposal density.

- Focusing on the numerator, when the trial value of the proposal has a higher posterior, acceptance is favored.

- The other factor is the density associated with returning to θ^s when starting at θ^*, which has to do with the reversability of the Markov chain. If this is too low, acceptance is not favored. We don't want to jump to a new region if we will never get back, as we need to sample from the entire support of the posterior.

- The two together mean that we will jump to a new area only if we are able to eventually jump back with a reasonably high probability. The probability of jumping is higher when the new area has a higher posterior density, but lower if it's hard to get back.

- The idea is to sample from all regions of the posterior, those with high and low density, sampling more heavily from regions of high density. We want to go occasionally to regions of low density, but it is important not to get stuck there.

- Consider a bimodal density: we want to explore the area around both modes. To be able to do that, it is important that the proposal density allows us to be able to jump between modes. Understanding in detail why this makes sense is the tricky and elegant part of the theory, see the references for more information.

- Note that the ratio of posteriors is equal to the ratio of likelihoods times the ratio of priors:

$$\frac{f(\theta^*|y)}{f(\theta^s|y)} = \frac{f(y|\theta^*)}{f(y|\theta^s)}\frac{\pi(\theta^*)}{\pi(\theta^s)}$$

  because the marginal likelihood $f(y)$ is the same in both cases. We don't need to compute that integral! We don't need to know the posterior, either. The acceptance criterion can be written as: accept if

$$\frac{f(y|\theta^*)}{f(y|\theta^s)}\frac{\pi(\theta^*)}{\pi(\theta^s)}\frac{g(\theta^s;\theta^*)}{g(\theta^*;\theta^s)} > \alpha$$

  otherwise, reject

- From this, we see that the information needed to determine if a proposal is accepted or rejected is the prior, the proposal density, and the likelihood function $f(y|\theta)$.

  - in principle, the prior is non-negotiable. In practice, people often chose priors with convenience in mind
  - the likelihood function is what it is
  - the place where artistry comes to bear is the choice of the proposal density

- the steps are:

1. the algorithm is initialized at some θ^1
2. for $s = 2, ..., S$,
   (a) draw θ^* from $g(\theta^*; \theta^s)$
   (b) according to the acceptance/rejection criterion, if the result is acceptance, set $\theta^{s+1} = \theta^*$, otherwise set $\theta^{s+1} = \theta^s$
   (c) iterate

- Once the chain is considered to have stabilized, say at iteration r, the values of θ^s for $s > r$ are taken to be draws from the posterior. The posterior mean is computed as the simple average of the value. Quantiles, etc., can be computed in the appropriate fashion.
- the art of applying these methods consists of providing a good proposal density so that the acceptance rate is reasonably high. Otherwise, the chain will be highly autocorrelated, with long intervals where the same value of θ appears (many proposals rejected). There is a vast literature on this, and the vastness of the literature should serve as a warning that getting this to work in practice is not necessarily a simple matter. If it were, there would be fewer papers on the topic.
  - too high acceptance rate: this is usually due to a proposal density that gives proposals very close to the current value, e.g, a random walk with very low variance. This means that the posterior is being explored inefficiently, we travel around through the support at a very low rate, which means the chain will have to run for a long time to do a thorough exploration.
  - too low acceptance rate: this means that the steps are too large, and we attempt to move to low posterior density regions too frequently. The chain will become highly autocorrelated,

so long periods convey little additional information relative to a subset of the values in the interval

- look at the example mh.m

16.6 Examples

MCMC for the simple example

The simple exponential example with lognormal prior can be implemented using MH MCMC, and this is done in the Octave script BayesExample2.m . Play around with the sample size and the tuning parameter, and note the effects on the computed posterior mean and on the acceptance rate. An example of output is given in Figure 16.3. In that Figure, the chain shows relatively long periods of rejection, meaning that the tuning parameter needs to be lowered, to cause the random walk to be a little less random.

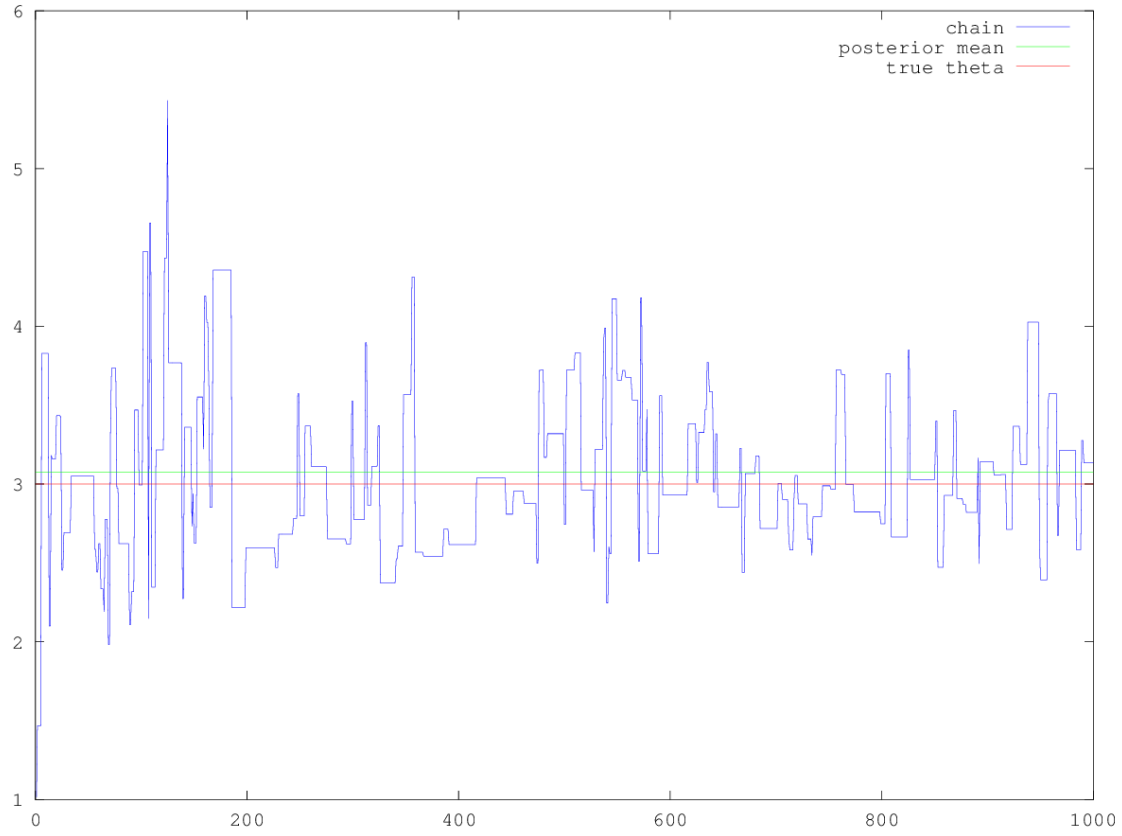
Bayesian VAR with Minnesota priors

Consider a VAR(p) model, where the data have been de-meaned:

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t$$

This follows an SUR structure, so OLS estimation is appropriate, even though we expect that $V(\epsilon_t) = \Sigma$ is a full $G \times G$ matrix (heteroscedasticity and autocorrelation, which would normally lead on to think of GLS estimation). As was previously noted, a problem with the estimation of this model is that

Figure 16.3: Metropolis-Hastings MCMC, exponential likelihood, lognormal prior



the number of parameters increases rapidly in the number of lags, p. One can use Bayesian methods such as the "Minnesota prior" (Doan, T., Litterman, R., Sims, C. (1984). "Forecasting and conditional projection using realistic prior distributions". Econometric Reviews 3: 1–100), which is a prior that each variable separately follows a random walk (an AR(1) model with $\rho = 1$). The prior on A_1 is that it is an identity matrix, and the prior on the $A_j, j > 1$ is that they are zero matrices. This can be done using stochastic restrictions similar to what was in the discussion of collinearity and ridge regression. To be specific, note that the model can be written as

$$Y = Y_{-1}A_1' + Y_{-2}A_2' + \cdots + Y_{-p}A_p' + E$$

where

$$Y = \begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{bmatrix}$$

is the $n \times G$ matrix of all the data, and the right hand side $Y's$ are this matrix, lagged the indicated number of times. The initial data with missing lags has been dropped, and n refers to the number of complete observations, including all needed lags.

Exercise 63. Convince yourself that this matrix representation is the same as $y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t$, just writing all observations at once, and transposing.

Now, consider the prior that each variable separately follows a random walk. If this were exactly

true, then $A_1 = I_G$, and all the $A_s = 0_G$, a $G \times G$ matrix of zeros, for $s = 2, 3, ..., p$. Consider the prior

$$\begin{aligned} A_1 &\sim N(I_G, \sigma_1^2 I_G) \\ A_2 &\sim N(0_G, \sigma_2^2 I_G) \\ &\vdots \\ A_p &\sim N(0_G, \sigma_p^2 I_G) \end{aligned}$$

and all of the matrices of parameters are independent of one another. In the same way we formulated the ridge regression estimator in Section 7.1, we can write the model and priors as

$$\begin{bmatrix} Y \\ I_G \\ 0_G \\ \vdots \\ 0_G \end{bmatrix} = \begin{bmatrix} Y_{-1} & Y_{-2} & \cdots & Y_{-p} \\ I_G & 0_G & \cdots & 0_G \\ 0_G & I_G & \cdots & 0_G \\ \vdots & & \ddots & \\ 0_G & & & I_G \end{bmatrix} \begin{bmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \end{bmatrix} + \begin{bmatrix} E \\ v_1 \\ v_2 \\ \vdots \\ v_p \end{bmatrix}$$

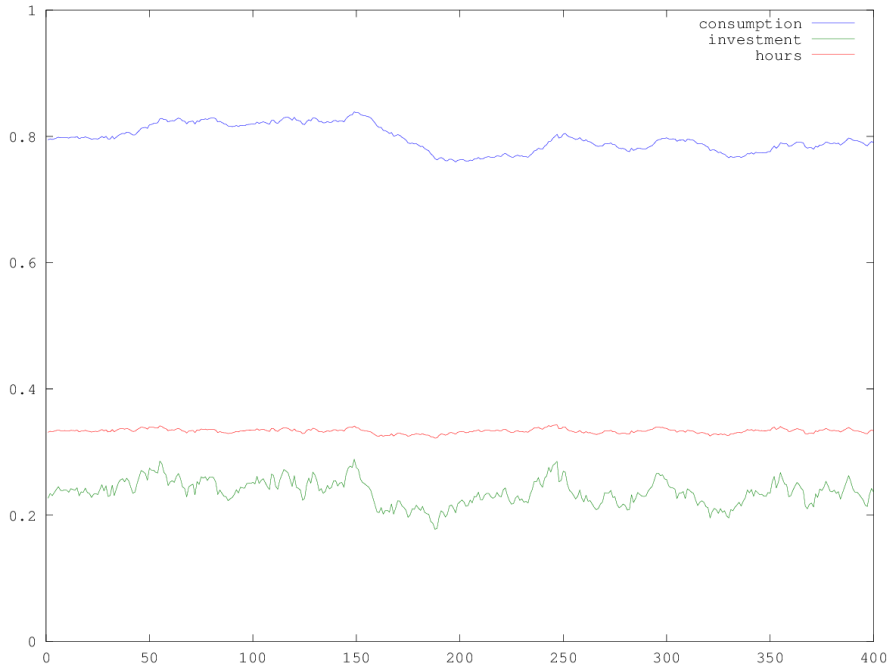
The final blocks may be multiplied by a prior precision, to enforce the prior to the desired degree, and then estimation may be done using OLS, just as we did when introducing ordinary ridge regression. This is a simple example of a Bayesian VAR: the VAR(p) model, combined with a certain prior (random walk, and Gaussian prior), implemented using mixed estimation.

We have previously seen a simple RBC model, in Section 13.8. If you run rbc.mod using Dynare, it will generate simulated data from this model. The data file rbcdata.m contains 400 observations on consumption, investment and hours worked, generated by this model. The data are plotted in Figure 16.4. Hours worked is quite stable around the steady state value of 1/3, but consumption

and investment fluctuate a little more. Let's estimate a Bayesian VAR, using this data. The script EstimateBVAR.m gives the results

```
octave:1> EstimateBVAR
plain OLS results
A1
-0.463821 -9.234590 -10.554874
0.065112 1.650430 0.913775
0.300228 1.465854 2.509376
A2
1.47732 9.31190 10.57178
-0.18877 -1.20802 -1.30911
-0.10263 -0.64005 -0.78350
r-squares OLS fit:  0.98860 0.82362 0.79340
##################################
Minnesota prior results
A1
1.004195 0.037051 0.026717
-0.046089 0.706809 -0.273725
0.058231 0.160412 1.121387
A2
0.0066800 0.0739334 0.0674895
-0.0358829 -0.1631035 -0.1307285
0.0357671 0.1514610 0.1249041
```

Figure 16.4: Data from RBC model

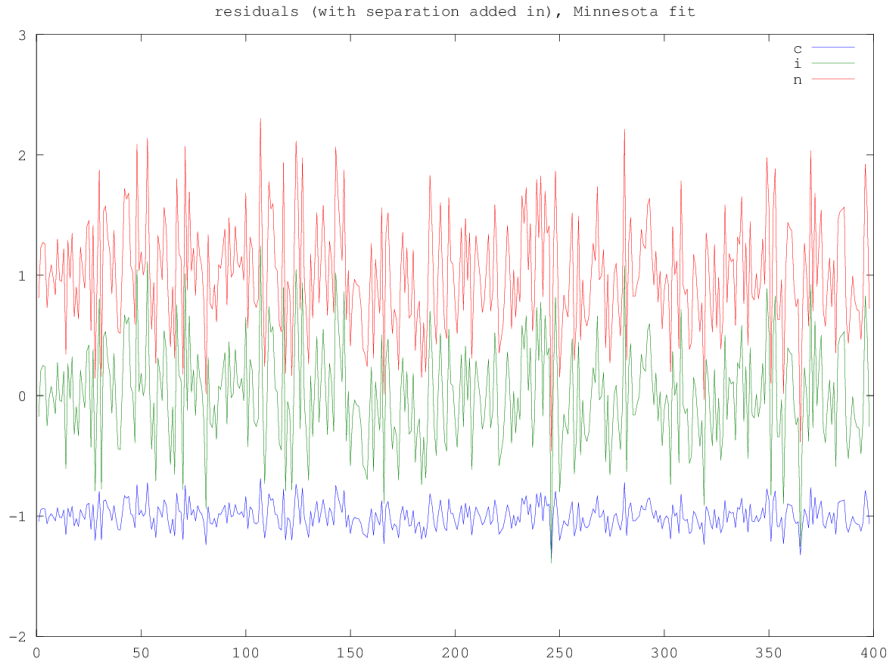



```
r-squares Minnesota fit:  0.98859 0.82341 0.79320
```

Note how the $R^2 s$ hardly change, but the estimated coefficients are much more similar to AR1 fits. The prior seems to be imposing discipline on the coefficients, without affecting goodness of fit in any serious way. Having a look at the residuals, see Figure 16.5. Note that the residuals for investment and hours are obviously very highly correlated. This is because the model that generated the data contains only one shock (a technology shock), so the stochastic behavior of the variables is necessarily fairly tightly linked.

16.5

Figure 16.5: BVAR residuals, with separation



Bayesian estimation of DSGE model

In Section 13.8, a simple DSGE model was estimated by ML. EstimateRBC_Bayesian.mod is a Dynare .mod file that lets you do the same thing using Bayesian methods, with MCMC.

- ML is not able to successfully estimate all parameters
- The Bayesian method does manage to estimate all parameters: the prior is helping
- note that the posterior is substantially different than the prior: we learn a lot from the sample
- both ML and Bayesian are using the same likelihood function, calculated using Kalman filtering on a linearized model, assuming Gaussian errors.
- if the model is solved using a higher order solution, the Kalman filter cannot be used, and Dynare uses particle filtering instead. This is very time consuming, as you can check.

Another example of Bayesian estimation of a DSGE model is given in Section 22.6.

16.7 Exercises

1. Experiment with the examples to learn about tuning, etc.

Chapter 17

Introduction to panel data

Reference: Cameron and Trivedi, 2005, *Microeconometrics: Methods and Applications*, Part V, Chapters 21 and 22 (plus 23 if you have special interest in the topic). The GRETL manual, chapters 16 and 17 is also a nice reference.

In this chapter we'll look at panel data. Panel data is an important area in applied econometrics, simply because much of the available data has this structure. Also, it provides an example where things we've already studied (GLS, endogeneity, GMM, Hausman test) come into play. There has been much work in this area, and the intention is not to give a complete overview, but rather to highlight the issues and see how the tools we have studied can be applied.

17.1 Generalities

Panel data combines cross sectional and time series data: we have a time series for each of the agents observed in a cross section. The addition of temporal information can in principle allow us to

investigate issues such as persistence, habit formation, and dynamics. Starting from the perspective of a single time series, the addition of cross-sectional information allows investigation of heterogeneity. In both cases, if parameters are common across units or over time, the additional data allows for more precise estimation.

The basic idea is to allow variables to have two indices, $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$. The simple linear model

$$y_i = \alpha + x_i\beta + \epsilon_i$$

becomes

$$y_{it} = \alpha + x_{it}\beta + \epsilon_{it}$$

We could think of allowing the parameters to change over time and over cross sectional units. This would give

$$y_{it} = \alpha_{it} + x_{it}\beta_{it} + \epsilon_{it}$$

The problem here is that there are more parameters than observations, so the model is not identified. We need some restraint! The proper restrictions to use of course depend on the problem at hand, and a single model is unlikely to be appropriate for all situations. For example, one could have time and cross-sectional dummies, and slopes that vary by time:

$$y_{it} = \alpha_i + \alpha_t + x_{it}\beta_t + \epsilon_{it}$$

There is a lot of room for playing around here. We also need to consider whether or not n and T are fixed or growing. We'll need at least one of them to be growing in order to do asymptotics.

To provide some focus, we'll consider common slope parameters, but agent-specific intercepts,

which:

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it} \tag{17.1}$$

I will refer to this as the "simple linear panel model". This is the model most often encountered in the applied literature. It is like the original cross-sectional model, in that the $\beta's$ are constant over time for all i. However we're now allowing for the constant to vary across i (some individual heterogeneity). The $\beta's$ are fixed over time, which is a testable restriction, of course. We can consider what happens as $n \to \infty$ but T is fixed. This would be relevant for microeconometric panels, (e.g., the PSID data) where a survey of a large number of individuals may be done for a limited number of time periods. Macroeconometric applications might look at longer time series for a small number of cross-sectional units (e.g., 40 years of quarterly data for 15 European countries). For that case, we could keep n fixed (seems appropriate when dealing with the EU countries), and do asymptotics as T increases, as is normal for time series. The asymptotic results depend on how we do this, of course.

Why bother using panel data, what are the benefits? The model

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$$

is a restricted version of

$$y_{it} = \alpha_i + x_{it}\beta_i + \epsilon_{it}$$

which could be estimated for each i in turn. Why use the panel approach?

- Because the restrictions that $\beta_i = \beta_j = ... = \beta$, if true, lead to more efficient estimation. Estimation for each i in turn will be very uninformative if T is small.

- Another reason is that panel data allows us to estimate parameters that are not identified by

cross sectional (time series) data. For example, if the model is

$$y_{it} = \alpha_i + \alpha_t + x_{it}\beta_t + \epsilon_{it}$$

and we have only cross sectional data, we cannot estimate the α_i. If we have only time series data on a single cross sectional unit $i = 1$, we cannot estimate the α_t. Cross-sectional variation allows us to estimate parameters indexed by time, and time series variation allows us to estimate parameters indexed by cross-sectional unit. Parameters indexed by both i and t will require other forms of restrictions in order to be estimable.

The main issues are:

- can β be estimated consistently? This is almost always a goal.
- can the α_i be estimated consistently? This is often of secondary interest.
- sometimes, we're interested in estimating the distribution of α_i across i.
- are the α_i correlated with x_{it}?
- does the presence of α_i complicate estimation of β?
- what about the covariance stucture? We're likely to have HET and AUT, so GLS issue will probably be relevant. Potential for efficiency gains.

17.2 Static models and correlations between variables

To begin with, assume that:

- the x_{it} are weakly exogenous variables (uncorrelated with ϵ_{it})
- the model is static: x_{it} does not contain lags of y_{it}.
- then the basic problem we have in the panel data model $y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$ is the presence of the α_i. These are individual-specific parameters. Or, possibly more accurately, they can be thought of as individual-specific variables that are not observed (latent variables). The reason for thinking of them as variables is because the agent may choose their values following some process.

Define $\alpha = E(\alpha_i)$, so $E(\alpha_i - \alpha) = 0$. Our model $y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$ may be written

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}\beta + \epsilon_{it} \\ &= \alpha + x_{it}\beta + (\alpha_i - \alpha + \epsilon_{it}) \\ &= \alpha + x_{it}\beta + \eta_{it} \end{aligned}$$

Note that $E(\eta_{it}) = 0$. A way of thinking about the data generating process is this:

- First, α_i is drawn, either in turn from the set of n fixed values, or randomly
- then x is drawn from $f_X(z|\alpha_i)$.
- In either case, the important point is that the distribution of x may vary depending on the realization of α_i.
  - Thus, there may be correlation between α_i and x_{it}, which means that $E(x_{it}\eta_{it}) \neq 0$ in the above equation.

- This means that OLS estimation of the model would lead to biased and inconsistent estimates.
- However, it is possible (but unlikely for economic data) that x_{it} and η_{it} are independent or at least uncorrelated, if the distribution of x_{it} is constant with respect to the realization of α_i. In this case OLS estimation would be consistent.

Fixed effects: when $E(x_{it}\eta_{it}) \neq 0$, the model is called the "fixed effects model"

Random effects: when $E(x_{it}\eta_{it}) = 0$, the model is called the "random effects model".

I find this to be pretty poor nomenclature, because the issue is not whether "effects" are fixed or random (they are always random, unconditional on i). The issue is whether or not the "effects" are correlated with the other regressors. In economics, it seems likely that the unobserved variable α is probably correlated with the observed regressors, x (this is simply the presence of collinearity between observed and unobserved variables, and collinearity is usually the rule rather than the exception). So, we expect that the "fixed effects" model is probably the relevant one unless special circumstances mean that the α_i are uncorrelated with the x_{it}.

17.3 Estimation of the simple linear panel model

"Fixed effects": The "within" estimator

How can we estimate the parameters of the simple linear panel model (equation 17.1) and what properties do the estimators have? First, we assume that the α_i are correlated with the x_{it} ("fixed effects" model). The model can be written as $y_{it} = \alpha + x_{it}\beta + \eta_{it}$, and we have that $E(x_{it}\eta_{it}) \neq 0$. As such, OLS estimation of this model will give biased an inconsistent estimated of the parameters α and

β. The "within" estimator is a solution - this involves subtracting the time series average from each cross sectional unit.

$$\begin{aligned}
\overline{x}_i &= \frac{1}{T}\sum_{t=1}^{T} x_{it} \\
\overline{\epsilon_i} &= \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} \\
\overline{y}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} = \alpha_i + \frac{1}{T}\sum_{t=1}^{T} x_{it}\beta + \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} \\
\overline{y}_i &= \alpha_i + \overline{x}_i\beta + \overline{\epsilon_i} \qquad (17.2)
\end{aligned}$$

The transformed model is

$$\begin{aligned}
y_{it} - \overline{y}_i &= \alpha_i + x_{it}\beta + \epsilon_{it} - \alpha_i - \overline{x}_i\beta - \overline{\epsilon_i} \qquad (17.3) \\
y^*_{it} &= x^*_{it}\beta + \epsilon^*_{it}
\end{aligned}$$

where $x^*_{it} = x_{it} - \overline{x}_i$ and $\epsilon^*_{it} = \epsilon_{it} - \overline{\epsilon}_i$. In this model, it is clear that x^*_{it} and ϵ^*_{it} are uncorrelated, as long as the original regressors x_{it} are *strongly* exogenous with respect to the original error ϵ_{it} ($E(x_{it}\epsilon_{is}) = 0$, $\forall t, s$). In this case, OLS will give consistent estimates of the parameters of this model, β.

What about the α_i? Can they be consistently estimated? An estimator is

$$\hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}\left(y_{it} - x_{it}\hat{\beta}\right)$$

It's fairly obvious that this is a consistent estimator *if* $T \to \infty$. For a short panel with fixed T, this

estimator is not consistent. Nevertheless, the variation in the $\hat{\alpha}_i$ can be fairly informative about the heterogeneity. A couple of notes:

- an equivalent approach is to estimate the model

$$y_{it} = \sum_{j=1}^{n} d_{j,it}\alpha_i + x_{it}\beta + \epsilon_{it}$$

  by OLS. The d_j, $j = 1, 2, ..., n$ are n dummy variables that take on the value 1 if $j = i$, zero otherwise. They are indicators of the cross sectional unit of the observation. (Write out form of regressor matrix on blackboard). Estimating this model by OLS gives numerically exactly the same results as the "within" estimator, and you get the $\hat{\alpha}_i$ automatically. See Cameron and Trivedi, section 21.6.4 for details. An interesting and important result known as the Frisch-Waugh-Lovell Theorem can be used to show that the two means of estimation give identical results.

- This last expression makes it clear why the "within" estimator cannot estimate slope coefficients corresponding to variables that have no time variation. Such variables are perfectly collinear with the cross sectional dummies d_j. The corresponding coefficients are not identified.

- OLS estimation of the "within" model is consistent, but probably not efficient, because it is highly probable that the ϵ_{it} are not iid. There is very likely heteroscedasticity across the i and autocorrelation between the T observations corresponding to a given i. *One needs to estimate the covariance matrix of the parameter estimates taking this into account.* It is possible to use GLS corrections if you make assumptions regarding the het. and autocor. Quasi-GLS, using a possibly misspecified model of the error covariance, can lead to more efficient estimates

than simple OLS. One can then combine it with subsequent panel-robust covariance estimation to deal with the misspecification of the error covariance, which would invalidate inferences if ignored. The White heteroscedasticity consistent covariance estimator is easily extended to panel data with independence across i, but with heteroscedasticity and autocorrelation within i, and heteroscedasticity between i. See Cameron and Trivedi, Section 21.2.3.

Estimation with random effects

The original model is

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$$

This can be written as

$$\begin{aligned} y_{it} &= \alpha + x_{it}\beta + (\alpha_i - \alpha + \epsilon_{it}) \\ y_{it} &= \alpha + x_{it}\beta + \eta_{it} \end{aligned} \tag{17.4}$$

where $E(\eta_{it}) = 0$, and $E(x_{it}\eta_{it}) = 0$. As such, the OLS estimator of this model is consistent. We can recover estimates of the α_i as discussed above. It is to be noted that the error η_{it} is almost certainly heteroscedastic and autocorrelated, so OLS will not be efficient, and inferences based on OLS need to be done taking this into account. One could attempt to use GLS, or panel-robust covariance matrix estimation, or both, as above.

There are other estimators when we have random effects, a well-known example being the "between" estimator, which operates on the time averages of the cross sectional units. There is no advantage to doing this, as the overall estimator is already consistent, and averaging looses information (efficiency loss). One would still need to deal with cross sectional heteroscedasticity when using the between

estimator, so there is no gain in simplicity, either.

It is to be emphasized that "random effects" is not a plausible assumption with most economic data, so use of this estimator is discouraged, even if your statistical package offers it as an option. Think carefully about whether the assumption is warranted before trusting the results of this estimator.

Hausman test

Suppose you're doubting about whether fixed or random effects are present.

- If we have correlation between x_{it} and α_i (fixed effects), then the "within" estimator will be consistent, but the random effects estimator of the previous section will not.
- Evidence that the two estimators are converging to different limits is evidence in favor of fixed effects, not random effects.
- A Hausman test statistic can be computed, using the difference between the two estimators.
  - The null hypothesis is that the effects are uncorrelated with the regressors in x_{it} ("random effects") so that both estimators are consistent under the null.
  - When the test rejects, we conclude that fixed effects are present, so the "within" estimator should be used.
  - Now, what happens if the test does not reject? One could optimistically turn to the random effects model, but it's probably more realistic to conclude that the test may have low power. Failure to reject does not mean that the null hypothesis is true. After all, estimation of the covariance matrices needed to compute the Hausman test is a non-trivial issue, and is a source of considerable noise in the test statistic (noise=low power).

– Finally, the simple version of the Hausman test requires that the estimator under the null be fully efficient. Achieving this goal is probably a utopian prospect. A conservative approach would acknowledge that neither estimator is likely to be efficient, and to operate accordingly. I have a little paper on this topic, Creel, *Applied Economics,* 2004. See also Cameron and Trivedi, section 21.4.3.

In class, do the first part of the example, below.

17.4 Dynamic panel data

When we have panel data, we have information on both y_{it} as well as $y_{i,t-1}$. One may naturally think of including $y_{i,t-1}$ as a regressor, to capture dynamic effects that can't be analyed with only cross-sectional data. Excluding dynamic effects is often the reason for detection of spurious AUT of the errors. With dynamics, there is likely to be less of a problem of autocorrelation, but one should still be concerned that some might still be present. The model, using a single lag of the dependent variable, becomes

$$\begin{aligned} y_{it} &= \alpha_i + \gamma y_{i,t-1} + x_{it}\beta + \epsilon_{it} \\ y_{it} &= \alpha + \gamma y_{i,t-1} + x_{it}\beta + (\alpha_i - \alpha + \epsilon_{it}) \\ y_{it} &= \alpha + \gamma y_{i,t-1} + x_{it}\beta + \eta_{it} \end{aligned}$$

We assume that the x_{it} are uncorrelated with ϵ_{it}.

- Note that α_i is a component that determines both y_{it} and its lag, $y_{i,t-1}$. Thus, α_i and $y_{i,t-1}$ are correlated, even if the α_i are pure random effects (uncorrelated with x_{it}).

- So, $y_{i,t-1}$ is correlated with η_{it}.
- For this reason, OLS estimation is inconsistent even for the random effects model, and it's also of course still inconsistent for the fixed effects model.
- When regressors are correlated with the errors, the natural thing to do is start thinking of instrumental variables estimation, or GMM.

To illustrate, consider a simple linear dynamic panel model

$$y_{it} = \alpha_i + \phi_0 y_{it-1} + \epsilon_{it} \tag{17.5}$$

where $\epsilon_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\phi_0 = 0, 0.3, 0.6, 0.9$ and α_i and ϵ_i are independently distributed. Tables 17.1 and 17.2 present bias and RMSE for the "within" estimator (labeled as ML) and some simulation-based estimators. Note that the "within" estimator is very biased, and has a large RMSE. The overidentified SBIL estimator (see Creel and Kristensen, "Indirect Likelihood Inference") has the lowest RMSE. Simulation-based estimators are discussed in a later Chapter. Perhaps these results will stimulate your interest.

Table 17.1: Dynamic panel data model. Bias. Source for ML and II is Gouriéroux, Phillips and Yu, 2010, Table 2. SBIL, SMIL and II are exactly identified, using the ML auxiliary statistic. SBIL(OI) and SMIL(OI) are overidentified, using both the naive and ML auxiliary statistics.

T	N	ϕ	ML	II	SBIL	SBIL(OI)
5	100	0.0	-0.199	0.001	0.004	-0.000
5	100	0.3	-0.274	-0.001	0.003	-0.001
5	100	0.6	-0.362	0.000	0.004	-0.001
5	100	0.9	-0.464	0.000	-0.022	-0.000
5	200	0.0	-0.200	0.000	0.001	0.000
5	200	0.3	-0.275	-0.010	0.001	-0.001
5	200	0.6	-0.363	-0.000	0.001	-0.001
5	200	0.9	-0.465	-0.003	-0.010	0.001

Table 17.2: Dynamic panel data model. RMSE. Source for ML and II is Gouriéroux, Phillips and Yu, 2010, Table 2. SBIL, SMIL and II are exactly identified, using the ML auxiliary statistic. SBIL(OI) and SMIL(OI) are overidentified, using both the naive and ML auxiliary statistics.

T	N	ϕ	ML	II	SBIL	SBIL(OI)
5	100	0.0	0.204	0.057	0.059	0.044
5	100	0.3	0.278	0.081	0.065	0.041
5	100	0.6	0.365	0.070	0.071	0.036
5	100	0.9	0.467	0.076	0.059	0.033
5	200	0.0	0.203	0.041	0.041	0.031
5	200	0.3	0.277	0.074	0.046	0.029
5	200	0.6	0.365	0.050	0.050	0.025
5	200	0.9	0.467	0.054	0.046	0.027

Arellano-Bond estimator

The first thing is to realize that the α_i that are a component of the error are correlated with all regressors in the general case of fixed effects. Getting rid of the α_i is a step in the direction of solving the problem. We could subtract the time averages, as above for the "within" estimator, but this would give us problems later when we need to define instruments. Instead, consider the model in first differences

$$\begin{aligned} y_{it} - y_{i,t-1} &= (\alpha_i + \gamma y_{i,t-1} + x_{it}\beta + \epsilon_{it}) - (\alpha_i + \gamma y_{i,t-2} + x_{i,t-1}\beta + \epsilon_{i,t-1}) \\ &= \gamma (y_{i,t-1} - y_{i,t-2}) + (x_{it} - x_{i,t-1}) \beta + \epsilon_{it} - \epsilon_{i,t-1} \end{aligned}$$

or

$$\Delta y_{it} = \gamma \Delta y_{i,t-1} + \Delta x_{it}\beta + \Delta \epsilon_{it}$$

- Now the pesky α_i are no longer in the picture.
- Note that we loose one observation when doing first differencing.
- OLS estimation of this model will still be inconsistent, because $y_{i,t-1}$is clearly correlated with $\epsilon_{i,t-1}$.
- Note also that the error $\Delta\epsilon_{it}$ is serially correlated even if the ϵ_{it} are not.
- There is no problem of correlation between Δx_{it} and $\Delta\epsilon_{it}$. Thus, to do GMM, we need to find instruments for $\Delta y_{i,t-1}$, but the variables in Δx_{it} can serve as their own instruments.

How about using $y_{i.t-2}$ as an instrument?

- It is clearly correlated with $\Delta y_{i,t-1} = (y_{i,t-1} - y_{i,t-2})$
- as long as the ϵ_{it} are not serially correlated, then $y_{i.t-2}$ is not correlated with $\Delta\epsilon_{it} = \epsilon_{it} - \epsilon_{i,t-1}$.
- We can also use additional lags $y_{i.t-s}$, $s \geq 2$ to increase efficiency, because GMM with additional instruments is asymptotically more efficient than with less instruments (but small sample bias may become a serious problem).

This sort of estimator is widely known in the literature as an Arellano-Bond estimator, due to the influential 1991 paper of Arellano and Bond (1991).

- Note that this sort of estimators requires $T = 3$ at a minimum.
- For $t = 1$ and $t = 2$, we cannot compute the moment conditions.
  - for $t = 1$, we do not have $y_{i,t-1} = y_{i,0}$, so we can't compute dependent variable.
  - for $t = 2$, we can compute the dependent variable Δy_{i2}, but not the regressor $\Delta y_{i,2-1} = y_{i,1} - y_{i,0}$.
- for $t = 3$, we can compute the dep. var. $\Delta y_{i,3}$, the regressor $\Delta y_{i,2} = y_{i,2} - y_{i,1}$, and we have $y_{i,1,}$ to serve as an instrument for $\Delta y_{i,2}$
- If $T > 3$, then when $t = 4$, we can use both $y_{i,1}$ and $y_{i,2}$ as instruments. This sort of unbalancedness in the instruments requires a bit of care when programming. Also, additional instruments increase asymptotic efficiency but can lead to increased small sample bias, so one should be a little careful with using too many instruments. Some robustness checks, looking at the stability of the estimates are a way to proceed.

One should note that serial correlation of the ϵ_{it} will cause this estimator to be inconsistent. Serial correlation of the errors *may* be due to dynamic misspecification, and this can be solved by including additional lags of the dependent variable. However, serial correlation may also be due to factors not captured in lags of the dependent variable. If this is a possibility, then the validity of the Arellano-Bond type instruments is in question.

- A final note is that the error $\Delta\epsilon_{it}$ is serially correlated even when the ϵ_{it} are not, and very likely heteroscedastic across i. One needs to take this into account when computing the covariance of the GMM estimator. One can also attempt to use GLS style weighting to improve efficiency. There are many possibilities.

- there is a "system" version of this sort of estimator that adds additional moment conditions, to improve efficiency

17.5 Example

Use the GRETL data set abdata.gdt to illustrate fixed effects, random effects, and DPD estimation For FE and RE, use the model

$$n_{it} = \alpha_i + \beta_t + \gamma w_{it} + \delta k_{it} + \phi ys_{it} + \epsilon_{it}$$

- do residuals appear to be normally distributed?

- is there evidence of serial correlation of residuals? (save them, and run an AR1 on them)

- Hausman test: rejects RE: we should favor FE. However, if errors are not normal, or if there is serial correlation, the test is not valid. Nevertheless, FE is probably favored on strictly theoretical grounds.

Given that the residuals seem to be serially correlated, we need to introduce dynamic structure. For DPD, use the model

$$n_{it} = \alpha_i + \beta_t + \rho_1 n_{i,t-1} + \gamma w_{it} + \delta k_{it} + \phi y s_{it} + \epsilon_{it}$$

- the estimate of ρ_1 is economically and statistically significant
- note the important differences in the other coefficients compared to the FE model
- check the serial correlation of the residuals: if it exists, the instruments are not valid. Possible solution is to augment the AR order, but the sample size gets smaller with every additional lag.

17.6 Exercises

1. In the context of a dynamic model with fixed effects, why is the differencing used in the "within" estimation approach (equation 17.3) problematic? That is, why does the Arellano-Bond estimator operate on the model in first differences instead of using the within approach?

2. Consider the simple linear panel data model with random effects (equation 17.4). Suppose that the ϵ_{it} are independent across cross sectional units, so that $E(\epsilon_{it}\epsilon_{js}) = 0$, $i \neq j$, $\forall t, s$. With a cross sectional unit, the errors are independently and identically distributed, so $E(\epsilon_{it}^2) = \sigma_i^2$, but $E(\epsilon_{it}\epsilon_{is}) = 0$, $t \neq s$. More compactly, let $\epsilon_i = \begin{bmatrix} \epsilon_{i1} & \epsilon_{i2} & \cdots & \epsilon_{iT} \end{bmatrix}'$. Then the assumptions are that $E(\epsilon_i \epsilon_i') = \sigma_i^2 I_T$, and $E(\epsilon_i \epsilon_j') = 0$, $i \neq j$.

(a) write out the form of the entire covariance matrix ($nT \times nT$) of all errors, $\Sigma = E(\epsilon\epsilon')$, where $\epsilon = \left[\begin{array}{cccc} \epsilon_1' & \epsilon_2' & \cdots & \epsilon_T' \end{array} \right]'$ is the column vector of nT errors.

(b) suppose that n is fixed, and consider asymptotics as T grows. Is it possible to estimate the Σ_i consistently? If so, how?

(c) suppose that T is fixed, and consider asymptotics an n grows. Is it possible to estimate the Σ_i consistently? If so, how?

(d) For one of the two preceeding parts (b) and (c), consistent estimation is possible. For that case, outline how to do "within" estimation using a GLS correction.

Chapter 18

Quasi-ML

Quasi-ML is the estimator one obtains when a misspecified probability model is used to calculate an "ML" estimator.

Given a sample of size n of a random vector $\mathbf{y}$ and a vector of conditioning variables $\mathbf{x}$, suppose the joint density of $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_n \end{pmatrix}$ conditional on $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_n \end{pmatrix}$ is a member of the parametric family $p_{\mathcal{Y}}(\mathbf{Y}|\mathbf{X}, \rho)$, $\rho \in \Xi$. The true joint density is associated with the vector ρ^0 :

$$p_{\mathcal{Y}}(\mathbf{Y}|\mathbf{X}, \rho^0).$$

As long as the marginal density of $\mathbf{X}$ doesn't depend on ρ^0, this conditional density fully characterizes the random characteristics of samples: i.e., it fully describes the probabilistically important features

of the d.g.p. The *likelihood function* is just this density evaluated at other values ρ

$$L(\mathbf{Y}|\mathbf{X},\rho) = p_{\mathcal{Y}}(\mathbf{Y}|\mathbf{X},\rho), \rho \in \Xi.$$

- Let $\mathbf{Y}_{t-1} = \begin{pmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_{t-1} \end{pmatrix}$, $\mathbf{Y}_0 = 0$, and let $\mathbf{X}_t = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_t \end{pmatrix}$ The likelihood function, taking into account possible dependence of observations, can be written as

$$\begin{aligned} L(\mathbf{Y}|\mathbf{X},\rho) &= \prod_{t=1}^{n} p_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\mathbf{X}_t,\rho) \\ &\equiv \prod_{t=1}^{n} p_t(\rho) \end{aligned}$$

- The average log-likelihood function is:

$$s_n(\rho) = \frac{1}{n}\ln L(\mathbf{Y}|\mathbf{X},\rho) = \frac{1}{n}\sum_{t=1}^{n}\ln p_t(\rho)$$

- Suppose that we do not have knowledge of the family of densities $p_t(\rho)$. Mistakenly, we may assume that the conditional density of $\mathbf{y}_t$ is a member of the family $f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\mathbf{X}_t,\theta)$, $\theta \in \Theta$, where there is no θ^0 such that $f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\mathbf{X}_t,\theta^0) = p_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\mathbf{X}_t,\rho^0), \forall t$ (this is what we mean by "misspecified").

- This setup allows for heterogeneous time series data, with dynamic misspecification.

The QML estimator is the argument that maximizes the **misspecified** average log likelihood, which

we refer to as the quasi-log likelihood function. This objective function is

$$\begin{aligned} s_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n} \ln f_t(\mathbf{y}_t|\mathbf{Y}_{t-1}, \mathbf{X}_t, \theta^0) \\ &\equiv \frac{1}{n}\sum_{t=1}^{n} \ln f_t(\theta) \end{aligned}$$

and the QML is

$$\hat{\theta}_n = \arg\max_{\Theta} s_n(\theta)$$

A SLLN for dependent sequences applies (we assume), so that

$$s_n(\theta) \overset{a.s.}{\to} \lim_{n\to\infty} \mathcal{E}\frac{1}{n}\sum_{t=1}^{n} \ln f_t(\theta) \equiv s_\infty(\theta)$$

We assume that this can be strengthened to uniform convergence, a.s., following the previous arguments. The "pseudo-true" value of θ is the value that maximizes $\bar{s}(\theta)$:

$$\theta^0 = \arg\max_{\Theta} s_\infty(\theta)$$

Given assumptions so that theorem 29 is applicable, we obtain

$$\lim_{n\to\infty} \hat{\theta}_n = \theta^0, \text{a.s.}$$

- Applying the asymptotic normality theorem,

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \overset{d}{\to} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right]$$

where

$$\mathcal{J}_\infty(\theta^0) = \lim_{n\to\infty} \mathcal{E} D_\theta^2 s_n(\theta^0)$$

and

$$\mathcal{I}_\infty(\theta^0) = \lim_{n\to\infty} Var\sqrt{n} D_\theta s_n(\theta^0).$$

- Note that asymptotic normality only requires that the additional assumptions regarding $\mathcal{J}$ and $\mathcal{I}$ hold in a neighborhood of θ^0 for $\mathcal{J}$ and at θ^0, for $\mathcal{I}$, not throughout Θ. In this sense, asymptotic normality is a local property.

18.1 Consistent Estimation of Variance Components

Consistent estimation of $\mathcal{J}_\infty(\theta^0)$ is straightforward. Assumption (b) of Theorem 31 implies that

$$\mathcal{J}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f_t(\hat{\theta}_n) \overset{a.s.}{\to} \lim_{n\to\infty} \mathcal{E}\frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f_t(\theta^0) = \mathcal{J}_\infty(\theta^0).$$

That is, just calculate the Hessian using the estimate $\hat{\theta}_n$ in place of θ^0.

Consistent estimation of $\mathcal{I}_\infty(\theta^0)$ is more difficult, and may be impossible.

- **Notation**: Let $g_t \equiv D_\theta f_t(\theta^0)$

We need to estimate

$$\begin{aligned}
\mathcal{I}_\infty(\theta^0) &= \lim_{n\to\infty} Var\sqrt{n}D_\theta s_n(\theta^0) \\
&= \lim_{n\to\infty} Var\sqrt{n}\frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f_t(\theta^0) \\
&= \lim_{n\to\infty} \frac{1}{n}Var\sum_{t=1}^{n} g_t \\
&= \lim_{n\to\infty} \frac{1}{n}\mathcal{E}\left\{\left(\sum_{t=1}^{n}(g_t - \mathcal{E}g_t)\right)\left(\sum_{t=1}^{n}(g_t - \mathcal{E}g_t)\right)'\right\}
\end{aligned}$$

This is going to contain a term

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}(\mathcal{E}g_t)(\mathcal{E}g_t)'$$

which will not tend to zero, in general. This term is not consistently estimable in general, since it requires calculating an expectation using the true density under the d.g.p., which is unknown.

- There are important cases where $\mathcal{I}_\infty(\theta^0)$ *is* consistently estimable. For example, suppose that the data come from a random sample (*i.e.,* they are iid). This would be the case with cross sectional data, for example. (Note: under i.i.d. sampling, the joint distribution of (y_t, x_t) is identical. This does not imply that the conditional density $f(y_t|x_t)$ is identical).

- With random sampling, the limiting objective function is simply

  $$s_\infty(\theta^0) = \mathcal{E}_X\mathcal{E}_0 \ln f(y|x, \theta^0)$$

  where $\mathcal{E}_0$ means expectation of $y|x$ and $\mathcal{E}_X$ means expectation respect to the marginal density of x.

- By the requirement that the limiting objective function be maximized at θ^0 we have

$$D_\theta \mathcal{E}_X \mathcal{E}_0 \ln f(y|x, \theta^0) = D_\theta s_\infty(\theta^0) = 0$$

- The dominated convergence theorem allows switching the order of expectation and differentiation, so

$$D_\theta \mathcal{E}_X \mathcal{E}_0 \ln f(y|x, \theta^0) = \mathcal{E}_X \mathcal{E}_0 D_\theta \ln f(y|x, \theta^0) = 0$$

  The CLT implies that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} D_\theta \ln f(y|x, \theta^0) \xrightarrow{d} N(0, \mathcal{I}_\infty(\theta^0)).$$

  That is, it's not necessary to subtract the individual means, since they are zero. Given this, and due to independent observations, a consistent estimator is

$$\widehat{\mathcal{I}} = \frac{1}{n} \sum_{t=1}^{n} D_\theta \ln f_t(\hat{\theta}) D_{\theta'} \ln f_t(\hat{\theta})$$

This is an important case where consistent estimation of the covariance matrix is possible. Other cases exist, even for dynamically misspecified time series models.

18.2 Example: the MEPS Data

To check the plausibility of the Poisson model for the MEPS data, we can compare the sample unconditional variance with the estimated unconditional variance according to the Poisson model: $\widehat{V(y)} = \frac{\sum_{t=1}^{n} \hat{\lambda}_t}{n}$. Using the program PoissonVariance.m, for OBDV and ERV, we get the results in

Table 18.1. We see that even after conditioning, the overdispersion is not captured in either case.

Table 18.1: Marginal Variances, Sample and Estimated (Poisson)

	OBDV	ERV
Sample	38.09	0.151
Estimated	3.28	0.086

There is huge problem with OBDV, and a significant problem with ERV. In both cases the Poisson model does not appear to be plausible. You can check this for the other use measures if you like.

Infinite mixture models: the negative binomial model

Reference: Cameron and Trivedi (1998) *Regression analysis of count data,* chapter 4.

The two measures seem to exhibit extra-Poisson variation. To capture unobserved heterogeneity, a possibility is the *random parameters* approach. Consider the possibility that the constant term in a Poisson model were random:

$$\begin{aligned} f_Y(y|\mathbf{x},\varepsilon) &= \frac{\exp(-\theta)\theta^y}{y!} \\ \theta &= \exp(\mathbf{x}\prime\beta + \varepsilon) \\ &= \exp(\mathbf{x}\prime\beta)\exp(\varepsilon) \\ &= \lambda\nu \end{aligned}$$

where $\lambda = \exp(\mathbf{x}\prime\beta)$ and $\nu = \exp(\varepsilon)$. Now ν captures the randomness in the constant. The problem

is that we don't observe ν, so we will need to marginalize it to get a usable density

$$f_Y(y|\mathbf{x}) = \int_{-\infty}^{\infty} \frac{\exp[-\theta]\theta^y}{y!} f_v(z)dz$$

This density *can* be used directly, perhaps using numerical integration to evaluate the likelihood function. In some cases, though, the integral will have an analytic solution. For example, if ν follows a certain one parameter gamma density, then

$$f_Y(y|\mathbf{x}, \phi) = \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^{\psi} \left(\frac{\lambda}{\psi+\lambda}\right)^{y} \tag{18.1}$$

where $\phi = (\lambda, \psi)$. ψ appears since it is the parameter of the gamma density.

- For this density, $E(y|\mathbf{x}) = \lambda$, which we have parameterized $\lambda = \exp(\mathbf{x}'\beta)$
- The variance depends upon how ψ is parameterized.
  - If $\psi = \lambda/\alpha$, where $\alpha > 0$, then $V(y|\mathbf{x}) = \lambda + \alpha\lambda$. Note that λ is a function of $\mathbf{x}$, so that the variance is too. This is referred to as the NB-I model.
  - If $\psi = 1/\alpha$, where $\alpha > 0$, then $V(y|\mathbf{x}) = \lambda + \alpha\lambda^2$. This is referred to as the NB-II model.

So both forms of the NB model allow for overdispersion, with the NB-II model allowing for a more radical form.

Testing reduction of a NB model to a Poisson model cannot be done by testing $\alpha = 0$ using standard Wald or LR procedures. The critical values need to be adjusted to account for the fact that $\alpha = 0$ is on the boundary of the parameter space. Without getting into details, suppose that the data were in

fact Poisson, so there is equidispersion and the true $\alpha = 0$. Then about half the time the sample data will be underdispersed, and about half the time overdispersed. When the data is underdispersed, the MLE of α will be $\hat{\alpha} = 0$. Thus, under the null, there will be a probability spike in the asymptotic distribution of $\sqrt{n}(\hat{\alpha} - \alpha) = \sqrt{n}\hat{\alpha}$ at 0, so standard testing methods will not be valid.

This program will do estimation using the NB model. Note how modelargs is used to select a NB-I or NB-II density. Here are NB-I estimation results for OBDV:

```
MPITB extensions found

OBDV

============================================================
BFGSMIN final results

Used analytic gradient

------------------------------------------------------------
STRONG CONVERGENCE
Function conv 1  Param conv 1  Gradient conv 1
------------------------------------------------------------
Objective function value 2.18573
Stepsize 0.0007
17 iterations
------------------------------------------------------------

 param    gradient  change
  1.0965   0.0000  -0.0000
```

```
  0.2551  -0.0000   0.0000
  0.2024  -0.0000   0.0000
  0.2289   0.0000  -0.0000
  0.1969   0.0000  -0.0000
  0.0769   0.0000  -0.0000
  0.0000  -0.0000   0.0000
  1.7146  -0.0000   0.0000


******************************************************
Negative Binomial model, MEPS 1996 full data set

MLE Estimation Results
BFGS convergence: Normal convergence

Average Log-L: -2.185730
Observations: 4564

              estimate     st. err      t-stat     p-value
constant        -0.523       0.104      -5.005       0.000
pub. ins.        0.765       0.054      14.198       0.000
priv. ins.       0.451       0.049       9.196       0.000
sex              0.458       0.034      13.512       0.000
age              0.016       0.001      11.869       0.000
edu              0.027       0.007       3.979       0.000
inc              0.000       0.000       0.000       1.000
alpha            5.555       0.296      18.752       0.000

Information Criteria
```

```
CAIC : 20026.7513      Avg. CAIC:   4.3880
 BIC : 20018.7513       Avg. BIC:   4.3862
 AIC : 19967.3437       Avg. AIC:   4.3750
*******************************************************
```

Note that the parameter values of the last BFGS iteration are different that those reported in the final results. This reflects two things - first, the data were scaled before doing the BFGS minimization, but the `mle_results` script takes this into account and reports the results using the original scaling. But also, the parameterization $\alpha = \exp(\alpha^*)$ is used to enforce the restriction that $\alpha > 0$. The unrestricted parameter $\alpha^* = \log \alpha$ is used to define the log-likelihood function, since the BFGS minimization algorithm does not do contrained minimization. To get the standard error and t-statistic of the estimate of α, we need to use the delta method. This is done inside `mle_results`, making use of the function parameterize.m .

Likewise, here are NB-II results:

```
MPITB extensions found

OBDV

=======================================================
BFGSMIN final results

Used analytic gradient

-------------------------------------------------------
```

```
STRONG CONVERGENCE
Function conv 1  Param conv 1  Gradient conv 1
--------------------------------------------------
Objective function value 2.18496
Stepsize 0.0104394
13 iterations
--------------------------------------------------

 param    gradient  change
  1.0375   0.0000  -0.0000
  0.3673  -0.0000   0.0000
  0.2136   0.0000  -0.0000
  0.2816   0.0000  -0.0000
  0.3027   0.0000   0.0000
  0.0843  -0.0000   0.0000
 -0.0048   0.0000  -0.0000
  0.4780  -0.0000   0.0000


**************************************************
Negative Binomial model, MEPS 1996 full data set

MLE Estimation Results
BFGS convergence: Normal convergence

Average Log-L: -2.184962
Observations: 4564

              estimate     st. err      t-stat     p-value
```

```
constant        -1.068       0.161      -6.622       0.000
pub. ins.        1.101       0.095      11.611       0.000
priv. ins.       0.476       0.081       5.880       0.000
sex              0.564       0.050      11.166       0.000
age              0.025       0.002      12.240       0.000
edu              0.029       0.009       3.106       0.002
inc             -0.000       0.000      -0.176       0.861
alpha            1.613       0.055      29.099       0.000

Information Criteria
CAIC : 20019.7439      Avg. CAIC:   4.3864
 BIC : 20011.7439       Avg. BIC:   4.3847
 AIC : 19960.3362       Avg. AIC:   4.3734
**************************************************
```

- For the OBDV usage measurel, the NB-II model does a slightly better job than the NB-I model, in terms of the average log-likelihood and the information criteria (more on this last in a moment).
- Note that both versions of the NB model fit much better than does the Poisson model (see 11.4).
- The estimated α is highly significant.

To check the plausibility of the NB-II model, we can compare the sample unconditional variance with the estimated unconditional variance according to the NB-II model: $\widehat{V(y)} = \frac{\sum_{t=1}^{n} \hat{\lambda}_t + \hat{\alpha}\left(\hat{\lambda}_t\right)^2}{n}$. For OBDV and ERV (estimation results not reported), we get For OBDV, the overdispersion problem is significantly better than in the Poisson case, but there is still some that is not captured. For ERV, the negative binomial model seems to capture the overdispersion adequately.

Table 18.2: Marginal Variances, Sample and Estimated (NB-II)

	OBDV	ERV
Sample	38.09	0.151
Estimated	30.58	0.182

Finite mixture models: the mixed negative binomial model

The finite mixture approach to fitting health care demand was introduced by Deb and Trivedi (1997). The mixture approach has the intuitive appeal of allowing for subgroups of the population with different health status. If individuals are classified as healthy or unhealthy then two subgroups are defined. A finer classification scheme would lead to more subgroups. Many studies have incorporated objective and/or subjective indicators of health status in an effort to capture this heterogeneity. The available objective measures, such as limitations on activity, are not necessarily very informative about a person's overall health status. Subjective, self-reported measures may suffer from the same problem, and may also not be exogenous

Finite mixture models are conceptually simple. The density is

$$f_Y(y, \phi_1, ..., \phi_p, \pi_1, ..., \pi_{p-1}) = \sum_{i=1}^{p-1} \pi_i f_Y^{(i)}(y, \phi_i) + \pi_p f_Y^p(y, \phi_p),$$

where $\pi_i > 0, i = 1, 2, ..., p$, $\pi_p = 1 - \sum_{i=1}^{p-1} \pi_i$, and $\sum_{i=1}^{p} \pi_i = 1$. Identification requires that the π_i are ordered in some way, for example, $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_p$ and $\phi_i \neq \phi_j, i \neq j$. This is simple to accomplish post-estimation by rearrangement and possible elimination of redundant component densities.

- The properties of the mixture density follow in a straightforward way from those of the components. In particular, the moment generating function is the same mixture of the moment

generating functions of the component densities, so, for example, $E(Y|x) = \sum_{i=1}^{p} \pi_i \mu_i(x)$, where $\mu_i(x)$ is the mean of the i^{th} component density.

- Mixture densities may suffer from overparameterization, since the total number of parameters grows rapidly with the number of component densities. It is possible to constrained parameters across the mixtures.

- Testing for the number of component densities is a tricky issue. For example, testing for $p = 1$ (a single component, which is to say, no mixture) versus $p = 2$ (a mixture of two components) involves the restriction $\pi_1 = 1$, which is on the boundary of the parameter space. Not that when $\pi_1 = 1$, the parameters of the second component can take on any value without affecting the density. Usual methods such as the likelihood ratio test are not applicable when parameters are on the boundary under the null hypothesis. Information criteria means of choosing the model (see below) are valid.

The following results are for a mixture of 2 NB-II models, for the OBDV data, which you can replicate using this program .

```
OBDV


**********************************************************
Mixed Negative Binomial model, MEPS 1996 full data set

MLE Estimation Results
BFGS convergence: Normal convergence
```

Average Log-L: -2.164783
Observations: 4564

	estimate	st. err	t-stat	p-value
constant	0.127	0.512	0.247	0.805
pub. ins.	0.861	0.174	4.962	0.000
priv. ins.	0.146	0.193	0.755	0.450
sex	0.346	0.115	3.017	0.003
age	0.024	0.004	6.117	0.000
edu	0.025	0.016	1.590	0.112
inc	-0.000	0.000	-0.214	0.831
alpha	1.351	0.168	8.061	0.000
constant	0.525	0.196	2.678	0.007
pub. ins.	0.422	0.048	8.752	0.000
priv. ins.	0.377	0.087	4.349	0.000
sex	0.400	0.059	6.773	0.000
age	0.296	0.036	8.178	0.000
edu	0.111	0.042	2.634	0.008
inc	0.014	0.051	0.274	0.784
alpha	1.034	0.187	5.518	0.000
Mix	0.257	0.162	1.582	0.114

Information Criteria

CAIC :	19920.3807	Avg. CAIC:	4.3647
BIC :	19903.3807	Avg. BIC:	4.3610
AIC :	19794.1395	Avg. AIC:	4.3370

****

It is worth noting that the mixture parameter is not significantly different from zero, but also not that the coefficients of public insurance and age, for example, differ quite a bit between the two latent classes.

Information criteria

As seen above, a Poisson model can't be tested (using standard methods) as a restriction of a negative binomial model. But it seems, based upon the values of the likelihood functions and the fact that the NB model fits the variance much better, that the NB model is more appropriate. How can we determine which of a set of competing models is the best?

The information criteria approach is one possibility. Information criteria are functions of the log-likelihood, with a penalty for the number of parameters used. Three popular information criteria are the Akaike (AIC), Bayes (BIC) and consistent Akaike (CAIC). The formulae are

$$\begin{aligned} CAIC &= -2\ln L(\hat{\theta}) + k(\ln n + 1) \\ BIC &= -2\ln L(\hat{\theta}) + k\ln n \\ AIC &= -2\ln L(\hat{\theta}) + 2k \end{aligned}$$

It can be shown that the CAIC and BIC will select the correctly specified model from a group of models, asymptotically. This doesn't mean, of course, that the correct model is necesarily in the group. The AIC is not consistent, and will asymptotically favor an over-parameterized model over the correctly specified model. Here are information criteria values for the models we've seen, for OBDV. Pretty clearly, the NB models are better than the Poisson. The one additional parameter gives a very significant improvement in the likelihood function value. Between the NB-I and NB-II models, the

Table 18.3: Information Criteria, OBDV

Model	AIC	BIC	CAIC
Poisson	7.345	7.355	7.357
NB-I	4.375	4.386	4.388
NB-II	4.373	4.385	4.386
MNB-II	4.337	4.361	4.365

NB-II is slightly favored. But one should remember that information criteria values are statistics, with variances. With another sample, it may well be that the NB-I model would be favored, since the differences are so small. The MNB-II model is favored over the others, by all 3 information criteria.

Why is all of this in the chapter on QML? Let's suppose that the correct model for OBDV is in fact the NB-II model. It turns out in this case that the Poisson model will give consistent estimates of the slope parameters (if a model is a member of the linear-exponential family and the conditional mean is correctly specified, then the parameters of the conditional mean will be consistently estimated). So the Poisson estimator would be a QML estimator that is consistent for some parameters of the true model. The ordinary OPG or inverse Hessian "ML" covariance estimators are however biased and inconsistent, since the information matrix equality does not hold for QML estimators. But for i.i.d. data (which is the case for the MEPS data) the QML asymptotic covariance can be consistently estimated, as discussed above, using the sandwich form for the ML estimator. `mle_results` in fact reports sandwich results, so the Poisson estimation results would be reliable for inference even if the true model is the NB-I or NB-II. Not that they are in fact similar to the results for the NB models.

However, if we assume that the correct model is the MNB-II model, as is favored by the information criteria, then both the Poisson and NB-x models will have misspecified mean functions, so the parameters that influence the means would be estimated with bias and inconsistently.

18.3 Exercises

1. Considering the MEPS data (the description is in Section 11.4), for the OBDV (y) measure, let η be a latent index of health status that has expectation equal to unity.[1] We suspect that η and $PRIV$ may be correlated, but we assume that η is uncorrelated with the other regressors. We assume that

$$\begin{aligned} &E(y|PUB, PRIV, AGE, EDUC, INC, \eta) \\ &= \exp(\beta_1 + \beta_2 PUB + \beta_3 PRIV + \beta_4 AGE + \beta_5 EDUC + \beta_6 INC)\eta. \end{aligned}$$

We use the Poisson QML estimator of the model

$$\begin{aligned} y &\sim \text{Poisson}(\lambda) \\ \lambda &= \exp(\beta_1 + \beta_2 PUB + \beta_3 PRIV + \\ &\quad \beta_4 AGE + \beta_5 EDUC + \beta_6 INC). \end{aligned} \tag{18.2}$$

Since much previous evidence indicates that health care services usage is overdispersed[2], this is almost certainly not an ML estimator, and thus is not efficient. However, when η and $PRIV$ are uncorrelated, this estimator is consistent for the β_i parameters, since the conditional mean is correctly specified in that case. When η and $PRIV$ are correlated, Mullahy's (1997) NLIV

[1] A restriction of this sort is necessary for identification.

[2] Overdispersion exists when the conditional variance is greater than the conditional mean. If this is the case, the Poisson specification is not correct.

estimator that uses the residual function

$$\varepsilon = \frac{y}{\lambda} - 1,$$

where λ is defined in equation 18.2, with appropriate instruments, is consistent. As instruments we use all the exogenous regressors, as well as the cross products of PUB with the variables in $Z = \{AGE, EDUC, INC\}$. That is, the full set of instruments is

$$W = \{1 \quad PUB \quad Z \quad PUB \times Z\}.$$

(a) Calculate the Poisson QML estimates.

(b) Calculate the generalized IV estimates (do it using a GMM formulation - see the portfolio example for hints how to do this).

(c) Calculate the Hausman test statistic to test the exogeneity of PRIV.

(d) comment on the results

Chapter 19

Nonlinear least squares (NLS)

Readings: Davidson and MacKinnon, Ch. 2* and 5*; Gallant, Ch. 1

19.1 Introduction and definition

Nonlinear least squares (NLS) is a means of estimating the parameter of the model

$$y_t = f(\mathbf{x}_t, \theta^0) + \varepsilon_t.$$

- In general, ε_t will be heteroscedastic and autocorrelated, and possibly nonnormally distributed. However, dealing with this is exactly as in the case of linear models, so we'll just treat the iid case here,

$$\varepsilon_t \sim iid(0, \sigma^2)$$

If we stack the observations vertically, defining

$$\mathbf{y} = (y_1, y_2, ..., y_n)'$$

$$\mathbf{f} = (f(x_1, \theta), f(x_1, \theta), ..., f(x_1, \theta))'$$

and

$$\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$$

we can write the n observations as

$$\mathbf{y} = \mathbf{f}(\theta) + \varepsilon$$

Using this notation, the NLS estimator can be defined as

$$\hat{\theta} \equiv \arg\min_{\Theta} s_n(\theta) = \frac{1}{n} \left[\mathbf{y} - \mathbf{f}(\theta)\right]' \left[\mathbf{y} - \mathbf{f}(\theta)\right] = \frac{1}{n} \parallel \mathbf{y} - \mathbf{f}(\theta) \parallel^2$$

- The estimator minimizes the weighted sum of squared errors, which is the same as minimizing the Euclidean distance between $\mathbf{y}$ and $\mathbf{f}(\theta)$.

The objective function can be written as

$$s_n(\theta) = \frac{1}{n} \left[\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{f}(\theta) + \mathbf{f}(\theta)'\mathbf{f}(\theta)\right],$$

which gives the first order conditions

$$-\left[\frac{\partial}{\partial \theta}\mathbf{f}(\hat{\theta})'\right]\mathbf{y} + \left[\frac{\partial}{\partial \theta}\mathbf{f}(\hat{\theta})'\right]\mathbf{f}(\hat{\theta}) \equiv 0.$$

Define the $n \times K$ matrix

$$\mathbf{F}(\hat{\theta}) \equiv D_{\theta'}\mathbf{f}(\hat{\theta}). \tag{19.1}$$

In shorthand, use $\hat{\mathbf{F}}$ in place of $\mathbf{F}(\hat{\theta})$. Using this, the first order conditions can be written as

$$-\hat{\mathbf{F}}'\mathbf{y} + \hat{\mathbf{F}}'\mathbf{f}(\hat{\theta}) \equiv 0,$$

or

$$\hat{\mathbf{F}}'\left[\mathbf{y} - \mathbf{f}(\hat{\theta})\right] \equiv 0. \tag{19.2}$$

This bears a good deal of similarity to the f.o.c. for the linear model - the derivative of the prediction is orthogonal to the prediction error. If $\mathbf{f}(\theta) = \mathbf{X}\theta$, then $\hat{\mathbf{F}}$ is simply $\mathbf{X}$, so the f.o.c. (with spherical errors) simplify to

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta = 0,$$

the usual 0LS f.o.c.

We can interpret this geometrically: *INSERT drawings of geometrical depiction of OLS and NLS (see Davidson and MacKinnon, pgs. 8,13 and 46).*

- Note that the nonlinearity of the manifold leads to potential multiple local maxima, minima and saddlepoints: the objective function $s_n(\theta)$ is not necessarily well-behaved and may be difficult to minimize.

19.2 Identification

As before, identification can be considered conditional on the sample, and asymptotically. The condition for asymptotic identification is that $s_n(\theta)$ tend to a limiting function $s_\infty(\theta)$ such that $s_\infty(\theta^0) < s_\infty(\theta)$, $\forall \theta \neq \theta^0$. This will be the case if $s_\infty(\theta^0)$ is strictly convex at θ^0, which requires that $D^2_\theta s_\infty(\theta^0)$ be positive definite. Consider the objective function:

$$\begin{aligned} s_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n}\left[y_t - f(\mathbf{x}_t,\theta)\right]^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}\left[f(\mathbf{x}_t,\theta^0) + \varepsilon_t - f_t(\mathbf{x}_t,\theta)\right]^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}\left[f_t(\theta^0) - f_t(\theta)\right]^2 + \frac{1}{n}\sum_{t=1}^{n}(\varepsilon_t)^2 \\ &- \frac{2}{n}\sum_{t=1}^{n}\left[f_t(\theta^0) - f_t(\theta)\right]\varepsilon_t \end{aligned}$$

- As in example 12.4, which illustrated the consistency of extremum estimators using OLS, we conclude that the second term will converge to a constant which does not depend upon θ.
- A LLN can be applied to the third term to conclude that it converges pointwise to 0, as long as $\mathbf{f}(\theta)$ and ε are uncorrelated.
- Next, pointwise convergence needs to be stregnthened to uniform almost sure convergence. There are a number of possible assumptions one could use. Here, we'll just assume it holds.

- Turning to the first term, we'll assume a pointwise law of large numbers applies, so

$$\frac{1}{n}\sum_{t=1}^{n}\left[f_t(\theta^0)-f_t(\theta)\right]^2 \overset{a.s.}{\to} \int\left[f(z,\theta^0)-f(z,\theta)\right]^2 d\mu(z), \tag{19.3}$$

  where $\mu(x)$ is the distribution function of x. In many cases, $f(x,\theta)$ *will* be bounded and continuous, for all $\theta \in \Theta$, so strengthening to uniform almost sure convergence is immediate. For example if $f(x,\theta) = [1+\exp(-x\theta)]^{-1}$, $f : \Re^K \to (0,1)$, a bounded range, and the function is continuous in θ.

Given these results, it is clear that a minimizer is θ^0. When considering identification (asymptotic), the question is whether or not there may be some other minimizer. A local condition for identification is that

$$\frac{\partial^2}{\partial\theta\partial\theta'}s_\infty(\theta) = \frac{\partial^2}{\partial\theta\partial\theta'}\int\left[f(x,\theta^0)-f(x,\theta)\right]^2 d\mu(x)$$

be positive definite at θ^0. Evaluating this derivative, we obtain (after a little work)

$$\frac{\partial^2}{\partial\theta\partial\theta'}\int\left[f(x,\theta^0)-f(x,\theta)\right]^2 d\mu(x)\Big|_{\theta^0} = 2\int\left[D_\theta f(z,\theta^0)'\right]\left[D_{\theta'}f(z,\theta^0)\right]' d\mu(z)$$

the expectation of the outer product of the gradient of the regression function evaluated at θ^0. (Note: the uniform boundedness we have already assumed allows passing the derivative through the integral, by the dominated convergence theorem.) This matrix will be positive definite (wp1) as long as the gradient vector is of full rank (wp1). The tangent space to the regression manifold must span a K -dimensional space if we are to consistently estimate a K -dimensional parameter vector. This is analogous to the requirement that there be no perfect colinearity in a linear model. This is a necessary

condition for identification. Note that the LLN implies that the above expectation is equal to

$$\mathcal{J}_\infty(\theta^0) = 2 \lim \mathcal{E} \frac{\mathbf{F}'\mathbf{F}}{n}$$

19.3 Consistency

We simply assume that the conditions of Theorem 29 hold, so the estimator is consistent. Given that the strong stochastic equicontinuity conditions hold, as discussed above, and given the above identification conditions an a compact estimation space (the closure of the parameter space Θ), the consistency proof's assumptions are satisfied.

19.4 Asymptotic normality

As in the case of GMM, we also simply assume that the conditions for asymptotic normality as in Theorem 31 hold. The only remaining problem is to determine the form of the asymptotic variance-covariance matrix. Recall that the result of the asymptotic normality theorem is

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right],$$

where $\mathcal{J}_\infty(\theta^0)$ is the almost sure limit of $\frac{\partial^2}{\partial\theta\partial\theta'}s_n(\theta)$ evaluated at θ^0, and

$$\mathcal{I}_\infty(\theta^0) = \lim Var\sqrt{n}D_\theta s_n(\theta^0)$$

The objective function is

$$s_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}\left[y_t - f(\mathbf{x}_t, \theta)\right]^2$$

So

$$D_\theta s_n(\theta) = -\frac{2}{n}\sum_{t=1}^{n}\left[y_t - f(\mathbf{x}_t, \theta)\right] D_\theta f(\mathbf{x}_t, \theta).$$

Evaluating at θ^0,

$$D_\theta s_n(\theta^0) = -\frac{2}{n}\sum_{t=1}^{n}\varepsilon_t D_\theta f(\mathbf{x}_t, \theta^0).$$

Note that the expectation of this is zero, since ϵ_t and $\mathbf{x}_t$ are assumed to be uncorrelated. So to calculate the variance, we can simply calculate the second moment about zero. Also note that

$$\begin{aligned}\sum_{t=1}^{n}\varepsilon_t D_\theta f(\mathbf{x}_t, \theta^0) &= \frac{\partial}{\partial\theta}\left[\mathbf{f}(\theta^0)\right]'\varepsilon \\ &= \mathbf{F}'\varepsilon\end{aligned}$$

With this we obtain

$$\begin{aligned}\mathcal{I}_\infty(\theta^0) &= \lim Var\sqrt{n}D_\theta s_n(\theta^0) \\ &= \lim n\mathcal{E}\frac{4}{n^2}\mathbf{F}'\varepsilon\varepsilon^{'}\mathbf{F} \\ &= 4\sigma^2\lim\mathcal{E}\frac{\mathbf{F}'\mathbf{F}}{n}\end{aligned}$$

We've already seen that

$$\mathcal{J}_\infty(\theta^0) = 2\lim\mathcal{E}\frac{\mathbf{F}'\mathbf{F}}{n},$$

where the expectation is with respect to the joint density of x and ε. Combining these expressions for $\mathcal{J}_\infty(\theta^0)$ and $\mathcal{I}_\infty(\theta^0)$, and the result of the asymptotic normality theorem, we get

$$\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left(0, \left(\lim \mathcal{E}\frac{\mathbf{F}'\mathbf{F}}{n}\right)^{-1}\sigma^2\right).$$

We can consistently estimate the variance covariance matrix using

$$\left(\frac{\hat{\mathbf{F}}'\hat{\mathbf{F}}}{n}\right)^{-1}\hat{\sigma}^2, \tag{19.4}$$

where $\hat{\mathbf{F}}$ is defined as in equation 19.1 and

$$\hat{\sigma}^2 = \frac{\left[\mathbf{y}-\mathbf{f}(\hat{\theta})\right]'\left[\mathbf{y}-\mathbf{f}(\hat{\theta})\right]}{n},$$

the obvious estimator. Note the close correspondence to the results for the linear model.

19.5 Example: The Poisson model for count data

Suppose that y_t conditional on $\mathbf{x}_t$ is independently distributed Poisson. A Poisson random variable is a *count data* variable, which means it can take the values {0,1,2,...}. This sort of model has been used to study visits to doctors per year, number of patents registered by businesses per year, *etc.*

The Poisson density is

$$f(y_t) = \frac{\exp(-\lambda_t)\lambda_t^{y_t}}{y_t!}, y_t \in \{0,1,2,...\}.$$

The mean of y_t is λ_t, as is the variance. Note that λ_t must be positive. Suppose that the true mean is

$$\lambda_t^0 = \exp(\mathbf{x}_t'\beta^0),$$

which enforces the positivity of λ_t. Suppose we estimate β^0 by nonlinear least squares:

$$\hat{\beta} = \arg\min s_n(\beta) = \frac{1}{T}\sum_{t=1}^{n}\left(y_t - \exp(\mathbf{x}_t'\beta)\right)^2$$

We can write

$$\begin{aligned} s_n(\beta) &= \frac{1}{T}\sum_{t=1}^{n}\left(\exp(\mathbf{x}_t'\beta^0 + \varepsilon_t - \exp(\mathbf{x}_t'\beta)\right)^2 \\ &= \frac{1}{T}\sum_{t=1}^{n}\left(\exp(\mathbf{x}_t'\beta^0 - \exp(\mathbf{x}_t'\beta)\right)^2 + \frac{1}{T}\sum_{t=1}^{n}\varepsilon_t^2 + 2\frac{1}{T}\sum_{t=1}^{n}\varepsilon_t\left(\exp(\mathbf{x}_t'\beta^0 - \exp(\mathbf{x}_t'\beta)\right) \end{aligned}$$

The last term has expectation zero since the assumption that $\mathcal{E}(y_t|\mathbf{x}_t) = \exp(\mathbf{x}_t'\beta^0)$ implies that $\mathcal{E}\left(\varepsilon_t|\mathbf{x}_t\right) = 0$, which in turn implies that functions of $\mathbf{x}_t$ are uncorrelated with ε_t. Applying a strong LLN, and noting that the objective function is continuous on a compact parameter space, we get

$$s_\infty(\beta) = \mathcal{E}_{\mathbf{x}}\left(\exp(\mathbf{x}'\beta^0 - \exp(\mathbf{x}'\beta)\right)^2 + \mathcal{E}_{\mathbf{x}}\exp(\mathbf{x}'\beta^0)$$

where the last term comes from the fact that the conditional variance of ε is the same as the variance of y. This function is clearly minimized at $\beta = \beta^0$, so the NLS estimator is consistent as long as identification holds.

Exercise 64. Determine the limiting distribution of $\sqrt{n}\left(\hat{\beta} - \beta^0\right)$. This means finding the the specific

forms of $\frac{\partial^2}{\partial\beta\partial\beta'}s_n(\beta)$, $\mathcal{J}(\beta^0)$, $\frac{\partial s_n(\beta)}{\partial\beta}\Big|$, and $\mathcal{I}(\beta^0)$. Again, use a CLT as needed, no need to verify that it can be applied.

19.6 The Gauss-Newton algorithm

Readings: Davidson and MacKinnon, Chapter 6, pgs. 201-207*.

The Gauss-Newton optimization technique is specifically designed for nonlinear least squares. The idea is to linearize the nonlinear model, rather than the objective function. The model is

$$\mathbf{y} = \mathbf{f}(\theta^0) + \varepsilon.$$

At some θ in the parameter space, not equal to θ^0, we have

$$\mathbf{y} = \mathbf{f}(\theta) + \nu$$

where ν is a combination of the fundamental error term ε and the error due to evaluating the regression function at θ rather than the true value θ^0. Take a first order Taylor's series approximation around a point θ^1 :

$$\mathbf{y} = \mathbf{f}(\theta^1) + \left[D_{\theta'}\mathbf{f}\left(\theta^1\right)\right]\left(\theta - \theta^1\right) + \nu + \text{approximation error.}$$

Define $\mathbf{z} \equiv \mathbf{y} - \mathbf{f}(\theta^1)$ and $b \equiv (\theta - \theta^1)$. Then the last equation can be written as

$$\mathbf{z} = \mathbf{F}(\theta^1)b + \omega,$$

where, as above, $\mathbf{F}(\theta^1) \equiv D_{\theta'}\mathbf{f}(\theta^1)$ is the $n \times K$ matrix of derivatives of the regression function,

evaluated at θ^1, and ω is ν plus approximation error from the truncated Taylor's series.

- Note that $\mathbf{F}$ is known, given θ^1.
- Note that one could estimate b simply by performing OLS on the above equation.
- Given $\hat{b}$, we calculate a new round estimate of θ^0 as $\theta^2 = \hat{b} + \theta^1$. With this, take a new Taylor's series expansion around θ^2 and repeat the process. Stop when $\hat{b} = 0$ (to within a specified tolerance).

To see why this might work, consider the above approximation, but evaluated at the NLS estimator:

$$\mathbf{y} = \mathbf{f}(\hat{\theta}) + \mathbf{F}(\hat{\theta})\left(\theta - \hat{\theta}\right) + \omega$$

The OLS estimate of $b \equiv \theta - \hat{\theta}$ is

$$\hat{b} = \left(\hat{\mathbf{F}}'\hat{\mathbf{F}}\right)^{-1}\hat{\mathbf{F}}'\left[\mathbf{y} - \mathbf{f}(\hat{\theta})\right].$$

This must be zero, since

$$\hat{\mathbf{F}}'\left(\hat{\theta}\right)\left[\mathbf{y} - \mathbf{f}(\hat{\theta})\right] \equiv 0$$

by definition of the NLS estimator (these are the normal equations as in equation 19.2, Since $\hat{b} \equiv 0$ when we evaluate at $\hat{\theta}$, updating would stop.

- The Gauss-Newton method doesn't require second derivatives, as does the Newton-Raphson method, so it's faster.

- The varcov estimator, as in equation 19.4 is simple to calculate, since we have $\hat{\mathbf{F}}$ as a by-product of the estimation process (*i.e.,* it's just the last round "regressor matrix"). In fact, a normal OLS program will give the NLS varcov estimator directly, since it's just the OLS varcov estimator from the last iteration.

- The method can suffer from convergence problems since $\mathbf{F}(\theta)'\mathbf{F}(\theta)$, may be very nearly singular, even with an asymptotically identified model, especially if θ is very far from $\hat{\theta}$. Consider the example

$$y = \beta_1 + \beta_2 x_t \beta^3 + \varepsilon_t$$

  When evaluated at $\beta_2 \approx 0$, β_3 has virtually no effect on the NLS objective function, so $\mathbf{F}$ will have rank that is "essentially" 2, rather than 3. In this case, $\mathbf{F}'\mathbf{F}$ will be nearly singular, so $(\mathbf{F}'\mathbf{F})^{-1}$ will be subject to large roundoff errors.

19.7 Application: Limited dependent variables and sample selection

Readings: Davidson and MacKinnon, Ch. 15* (a quick reading is sufficient), J. Heckman, "Sample Selection Bias as a Specification Error", *Econometrica*, 1979 (This is a classic article, not required for reading, and which is a bit out-dated. Nevertheless it's a good place to start if you encounter sample selection problems in your research).

Sample selection is a common problem in applied research. The problem occurs when observations used in estimation are sampled non-randomly, according to some selection scheme.

Example: Labor Supply

Labor supply of a person is a positive number of hours per unit time supposing the offer wage is higher than the reservation wage, which is the wage at which the person prefers not to work. The model (very simple, with t subscripts suppressed):

- Characteristics of individual: $\mathbf{x}$
- Latent labor supply: $s^* = \mathbf{x}'\beta + \omega$
- Offer wage: $w^o = \mathbf{z}'\gamma + \nu$
- Reservation wage: $w^r = \mathbf{q}'\delta + \eta$

Write the wage differential as

$$\begin{aligned} w^* &= (\mathbf{z}'\gamma + \nu) - (\mathbf{q}'\delta + \eta) \\ &\equiv \mathbf{r}'\theta + \varepsilon \end{aligned}$$

We have the set of equations

$$\begin{aligned} s^* &= \mathbf{x}'\beta + \omega \\ w^* &= \mathbf{r}'\theta + \varepsilon. \end{aligned}$$

Assume that

$$\begin{bmatrix} \omega \\ \varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \right).$$

We assume that the offer wage and the reservation wage, as well as the latent variable s^* are unobservable. What is observed is

$$\begin{aligned} w &= 1\left[w^* > 0\right] \\ s &= ws^*. \end{aligned}$$

In other words, we observe whether or not a person is working. If the person is working, we observe labor supply, which is equal to latent labor supply, s^*. Otherwise, $s = 0 \neq s^*$. Note that we are using a simplifying assumption that individuals can freely choose their weekly hours of work.

Suppose we estimated the model

$$s^* = \mathbf{x}'\beta + \text{residual}$$

using only observations for which $s > 0$. The problem is that these observations are those for which $w^* > 0$, or equivalently, $-\varepsilon < \mathbf{r}'\theta$ and

$$\mathcal{E}\left[\omega | -\varepsilon < \mathbf{r}'\theta\right] \neq 0,$$

since ε and ω are dependent. Furthermore, this expectation will in general depend on $\mathbf{x}$ since elements of $\mathbf{x}$ can enter in $\mathbf{r}$. Because of these two facts, least squares estimation is biased and inconsistent.

Consider more carefully $\mathcal{E}\left[\omega | -\varepsilon < \mathbf{r}'\theta\right]$. Given the joint normality of ω and ε, we can write (see for example Spanos *Statistical Foundations of Econometric Modelling,* pg. 122)

$$\omega = \rho\sigma\varepsilon + \eta,$$

where η has mean zero and is independent of ε. With this we can write

$$s^* = \mathbf{x}'\beta + \rho\sigma\varepsilon + \eta.$$

If we condition this equation on $-\varepsilon < \mathbf{r}'\theta$ we get

$$s = \mathbf{x}'\beta + \rho\sigma\mathcal{E}(\varepsilon| - \varepsilon < \mathbf{r}'\theta) + \eta$$

which may be written as

$$s = \mathbf{x}'\beta + \rho\sigma\mathcal{E}(\varepsilon|\varepsilon > -\mathbf{r}'\theta) + \eta$$

- A useful result is that for

$$z \sim N(0,1)$$

$$E(z|z > z^*) = \frac{\phi(z^*)}{\Phi(-z^*)},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution function, respectively. The quantity on the RHS above is known as the *inverse Mill's ratio:*

$$IMR(\mathbf{z}^*) = \frac{\phi(z^*)}{\Phi(-z^*)}$$

With this we can write (making use of the fact that the standard normal density is symmetric about zero, so that $\phi(-a) = \phi(a)$):

$$s = \mathbf{x}'\beta + \rho\sigma\frac{\phi\left(\mathbf{r}'\theta\right)}{\Phi\left(\mathbf{r}'\theta\right)} + \eta \tag{19.5}$$

$$\equiv \begin{bmatrix} \mathbf{x}' & \frac{\phi(\mathbf{r}'\theta)}{\Phi(\mathbf{r}'\theta)} \end{bmatrix} \begin{bmatrix} \beta \\ \zeta \end{bmatrix} + \eta. \tag{19.6}$$

where $\zeta = \rho\sigma$. The error term η has conditional mean zero, and is uncorrelated with the regressors $\begin{bmatrix} \mathbf{x}' & \frac{\phi(\mathbf{r}'\theta)}{\Phi(\mathbf{r}'\theta)} \end{bmatrix}$. At this point, we can estimate the equation by NLS.

- Heckman showed how one can estimate this in a two step procedure where first θ is estimated, then equation 19.6 is estimated by least squares using the estimated value of θ to form the regressors. This is inefficient and estimation of the covariance is a tricky issue. It is probably easier (and more efficient) just to do MLE.

- The model presented above depends strongly on joint normality. There exist many alternative models which weaken the maintained assumptions. It is possible to estimate consistently without distributional assumptions. See Ahn and Powell, *Journal of Econometrics*, 1994.

Chapter 20

Nonparametric inference

A good reference is Li and Racine (2007) *Nonparametric Econometrics: Theory and Practice.*

20.1 Possible pitfalls of parametric inference: estimation

Readings: H. White (1980) "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, pp. 149-70.

In this section we consider a simple example, which illustrates both why nonparametric methods may in some cases be preferred to parametric methods.

We suppose that data is generated by random sampling of (y, x), where $y = f(x) + \varepsilon$, x is uniformly distributed on $(0, 2\pi)$, and ε is a classical error with variance equal to 1. Suppose that

$$f(x) = 1 + \frac{3x}{2\pi} - \left(\frac{x}{2\pi}\right)^2$$

The problem of interest is to estimate the elasticity of $f(x)$ with respect to x, throughout the range of x.

In general, the functional form of $f(x)$ is unknown. One idea is to take a Taylor's series approximation to $f(x)$ about some point x_0. Flexible functional forms such as the transcendental logarithmic (usually known as the translog) can be interpreted as second order Taylor's series approximations. We'll work with a first order approximation, for simplicity. Approximating about x_0:

$$h(x) = f(x_0) + D_x f(x_0)\,(x - x_0)$$

If the approximation point is $x_0 = 0$, we can write

$$h(x) = a + bx$$

The coefficient a is the value of the function at $x = 0$, and the slope is the value of the derivative at $x = 0$. These are of course not known. One might try estimation by ordinary least squares. The objective function is

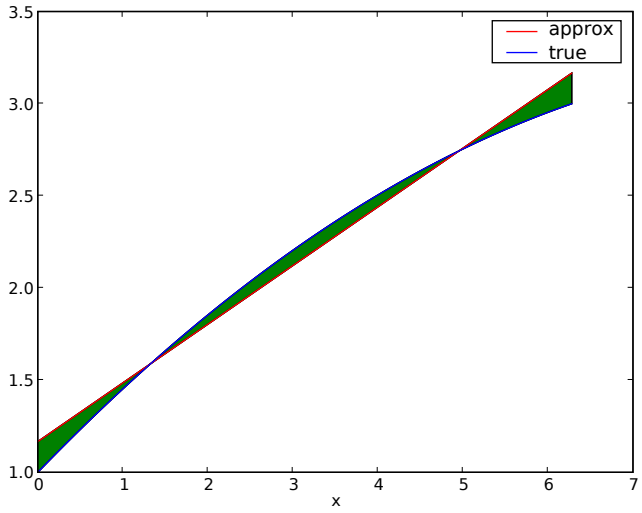
$$s(a, b) = 1/n \sum_{t=1}^{n} (y_t - h(x_t))^2 .$$

The limiting objective function, following the argument we used to get equations 12.1 and 19.3 is

$$s_\infty(a, b) = \int_0^{2\pi} (f(x) - h(x))^2 \, dx.$$

The theorem regarding the consistency of extremum estimators (Theorem 29) tells us that $\hat{a}$ and $\hat{b}$ will converge almost surely to the values that minimize the limiting objective function. Solving the

Figure 20.1: True and simple approximating functions



first order conditions[1] reveals that $s_\infty(a,b)$ obtains its minimum at $\left\{a^0 = \frac{7}{6}, b^0 = \frac{1}{\pi}\right\}$. The estimated approximating function $\hat{h}(x)$ therefore tends almost surely to

$$h_\infty(x) = 7/6 + x/\pi$$

In Figure 20.1 we see the true function and the limit of the approximation to see the asymptotic bias as a function of x.

(The approximating model is the straight line, the true model has curvature.) Note that the approximating model is in general inconsistent, even at the approximation point. This shows that "flexible functional forms" based upon Taylor's series approximations do not in general lead to consis-

[1]The following results were obtained using the free computer algebra system (CAS) Maxima. Unfortunately, I have lost the source code to get the results. It's not hard to do, though.

tent estimation of functions.

The approximating model seems to fit the true model fairly well, asymptotically. However, we are interested in the elasticity of the function. Recall that an elasticity is the marginal function divided by the average function:

$$\varepsilon(x) = \frac{\phi'(x)}{\phi(x)/x}$$

Good approximation of the elasticity over the range of x will require a good approximation of both $f(x)$ and $f'(x)$ over the range of x. The approximating elasticity is

$$\eta(x) = \frac{h'(x)}{h(x)/x}$$

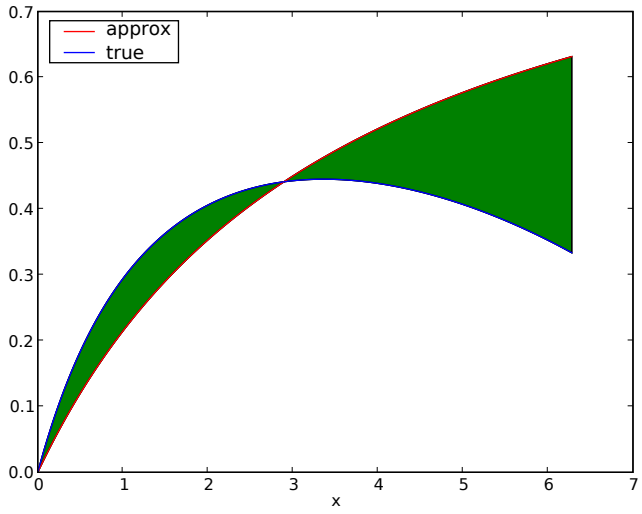
In Figure 20.2 we see the true elasticity and the elasticity obtained from the limiting approximating model.

The true elasticity is the line that has negative slope for large x. Visually we see that the elasticity is not approximated so well. Root mean squared error in the approximation of the elasticity is

$$\left(\int_0^{2\pi} (\varepsilon(x) - \eta(x))^2 \, dx\right)^{1/2} = .31546$$

Now suppose we use the leading terms of a trigonometric series as the approximating model. The reason for using a trigonometric series as an approximating model is motivated by the asymptotic properties of the Fourier flexible functional form (Gallant, 1981, 1982), which we will study in more detail below. Normally with this type of model the number of basis functions is an increasing function of the sample size. Here we hold the set of basis function fixed. We will consider the asymptotic behavior of a fixed model, which we interpret as an approximation to the estimator's behavior in finite

Figure 20.2: True and approximating elasticities



samples. Consider the set of basis functions:

$$Z(x) = \begin{bmatrix} 1 & x & \cos(x) & \sin(x) & \cos(2x) & \sin(2x) \end{bmatrix}.$$

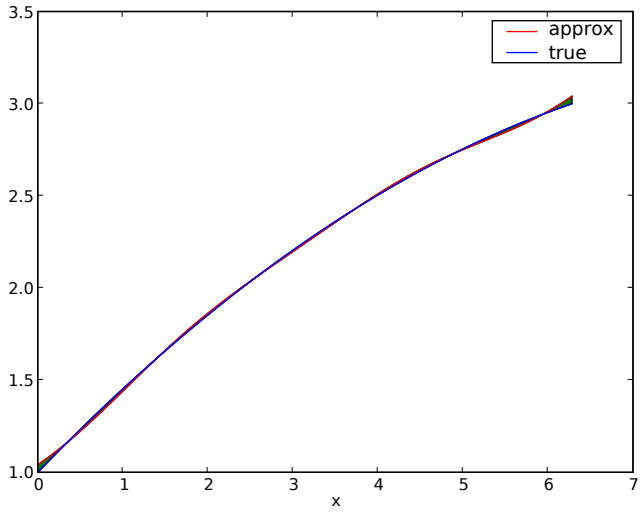
The approximating model is

$$g_K(x) = Z(x)\alpha.$$

Maintaining these basis functions as the sample size increases, we find that the limiting objective function is minimized at

$$\left\{a_1 = \frac{7}{6}, a_2 = \frac{1}{\pi}, a_3 = -\frac{1}{\pi^2}, a_4 = 0, a_5 = -\frac{1}{4\pi^2}, a_6 = 0\right\}.$$

Figure 20.3: True function and more flexible approximation

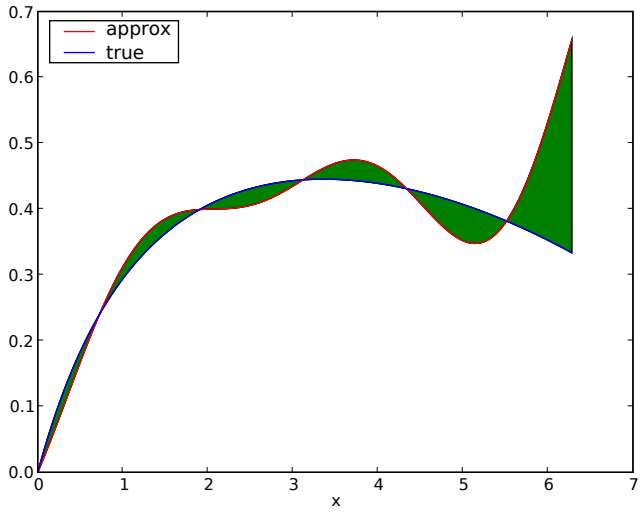


Substituting these values into $g_K(x)$ we obtain the almost sure limit of the approximation

$$g_\infty(x) = 7/6 + x/\pi + (\cos x)\left(-\frac{1}{\pi^2}\right) + (\sin x)\,0 + (\cos 2x)\left(-\frac{1}{4\pi^2}\right) + (\sin 2x)\,0 \tag{20.1}$$

In Figure 20.3 we have the approximation and the true function: Clearly the truncated trigonometric series model offers a better approximation, asymptotically, than does the linear model. In Figure 20.4 we have the more flexible approximation's elasticity and that of the true function: On average, the fit is better, though there is some implausible wavyness in the estimate. Root mean squared error in the

Figure 20.4: True elasticity and more flexible approximation



approximation of the elasticity is

$$\left(\int_0^{2\pi} \left(\varepsilon(x) - \frac{g'_\infty(x)x}{g_\infty(x)} \right)^2 dx \right)^{1/2} = .16213,$$

about half that of the RMSE when the first order approximation is used. If the trigonometric series contained infinite terms, this error measure would be driven to zero, as we shall see.

20.2 Possible pitfalls of parametric inference: hypothesis testing

What do we mean by the term "nonparametric inference"? Simply, this means inferences that are possible without restricting the functions of interest to belong to a parametric family.

- Consider means of testing for the hypothesis that consumers maximize utility. A consequence of utility maximization is that the Slutsky matrix $D_p^2 h(p, U)$, where $h(p, U)$ are the a set of compensated demand functions, must be negative semi-definite. One approach to testing for utility maximization would estimate a set of normal demand functions $x(p, m)$.

- Estimation of these functions by normal parametric methods requires specification of the functional form of demand, for example

$$x(p, m) = x(p, m, \theta^0) + \varepsilon, \theta^0 \in \Theta,$$

  where $x(p, m, \theta^0)$ is a function of known form and θ^0 is a finite dimensional parameter.

- After estimation, we could use $\hat{x} = x(p, m, \hat{\theta})$ to calculate (by solving the integrability problem, which is non-trivial) $\widehat{D}_p^2 h(p, U)$. If we can statistically reject that the matrix is negative semi-definite, we might conclude that consumers don't maximize utility.

- The problem with this is that the reason for rejection of the theoretical proposition may be that our choice of functional form is incorrect. In the introductory section we saw that functional form misspecification leads to inconsistent estimation of the function and its derivatives.

- Testing using parametric models always means we are testing a compound hypothesis. The hypothesis that is tested is 1) the economic proposition we wish to test, and 2) the model is correctly specified. Failure of either 1) or 2) can lead to rejection (as can a Type-I error, even when 2) holds). This is known as the "model-induced augmenting hypothesis."

- Varian's WARP allows one to test for utility maximization without specifying the form of the demand functions. The only assumptions used in the test are those directly implied by theory (well, maybe that's too strong: we also assume that the data are observed without measurement error), so rejection of the hypothesis calls into question the theory (unless there's measurement error in the data).

- Nonparametric inference also allows direct testing of economic propositions, avoiding the "model-induced augmenting hypothesis". The cost of nonparametric methods is usually an increase in complexity, and a loss of power, compared to what one would get using a well-specified parametric model. The benefit is robustness against possible misspecification.

20.3 Estimation of regression functions

The Fourier functional form

Readings: Gallant, 1987, "Identification and consistency in semi-nonparametric regression," in *Advances in Econometrics, Fifth World Congress,* V. 1, Truman Bewley, ed., Cambridge.

Suppose we have a multivariate model

$$y = f(\mathbf{x}) + \varepsilon,$$

where $f(x)$ is of unknown form and x is a $P-$dimensional vector. For simplicity, assume that ε is a classical error. Let us take the estimation of the vector of elasticities with typical element

$$\xi_{x_i} = \frac{\mathbf{x}_i}{f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial x_i f(x)},$$

at an arbitrary point $\mathbf{x}_i$.

The Fourier form, following Gallant (1982), but with a somewhat different parameterization, may be written as

$$g_K(\mathbf{x} \mid \theta_K) = \alpha + \mathbf{x}'\beta + 1/2\mathbf{x}'\mathbf{C}\mathbf{x} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left(u_{j\alpha} \cos(j\mathbf{k}'_\alpha \mathbf{x}) - v_{j\alpha} \sin(j\mathbf{k}'_\alpha \mathbf{x})\right). \tag{20.2}$$

where the K-dimensional parameter vector

$$\theta_K = \{\alpha, \beta', vec^*(C)', u_{11}, v_{11}, \ldots, u_{JA}, v_{JA}\}'. \tag{20.3}$$

- We assume that the conditioning variables $\mathbf{x}$ have each been transformed to lie in an interval that is shorter than 2π. This is required to avoid periodic behavior of the approximation, which is desirable since economic functions aren't periodic. For example, subtract sample means, divide by the maxima of the conditioning variables, and multiply by $2\pi - eps$, where eps is some positive number less than 2π in value.

- The k_α are "elementary multi-indices" which are simply $P-$ vectors formed of integers (negative, positive and zero). The k_α, $\alpha = 1, 2, ..., A$ are required to be linearly independent, and we follow

the convention that the first non-zero element be positive. For example

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 1 \end{bmatrix}'$$

is a potential multi-index to be used, but

$$\begin{bmatrix} 0 & -1 & -1 & 0 & 1 \end{bmatrix}'$$

is not since its first nonzero element is negative. Nor is

$$\begin{bmatrix} 0 & 2 & -2 & 0 & 2 \end{bmatrix}'$$

a multi-index we would use, since it is a scalar multiple of the original multi-index.

- We parameterize the matrix C differently than does Gallant because it simplifies things in practice. The cost of this is that we are no longer able to test a quadratic specification using nested testing.

The vector of first partial derivatives is

$$D_x g_K(\mathbf{x} \mid \theta_K) = \beta + \mathbf{C}\mathbf{x} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left[(-u_{j\alpha} \sin(j\mathbf{k}'_\alpha \mathbf{x}) - v_{j\alpha} \cos(j\mathbf{k}'_\alpha \mathbf{x}))\, j\mathbf{k}_\alpha \right] \tag{20.4}$$

and the matrix of second partial derivatives is

$$D_x^2 g_K(\mathbf{x}|\theta_K) = \mathbf{C} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left[(-u_{j\alpha} \cos(j\mathbf{k}'_\alpha \mathbf{x}) + v_{j\alpha} \sin(j\mathbf{k}'_\alpha \mathbf{x}))\, j^2 \mathbf{k}_\alpha \mathbf{k}'_\alpha \right] \tag{20.5}$$

To define a compact notation for partial derivatives, let λ be an N-dimensional multi-index with no negative elements. Define $\mid \lambda \mid^*$ as the sum of the elements of λ. If we have N arguments $\mathbf{x}$ of the (arbitrary) function $h(\mathbf{x})$, use $D^\lambda h(\mathbf{x})$ to indicate a certain partial derivative:

$$D^\lambda h(\mathbf{x}) \equiv \frac{\partial^{|\lambda|^*}}{\partial x_1^{\lambda_1} \partial x_2^{\lambda_2} \cdots \partial x_N^{\lambda_N}} h(\mathbf{x})$$

When λ is the zero vector, $D^\lambda h(\mathbf{x}) \equiv h(\mathbf{x})$. Taking this definition and the last few equations into account, we see that it is possible to define $(1 \times K)$ vector $Z^\lambda(\mathbf{x})$ so that

$$D^\lambda g_K(\mathbf{x}|\theta_K) = \mathbf{z}^\lambda(\mathbf{x})'\theta_K. \tag{20.6}$$

- Both the approximating model and the derivatives of the approximating model are linear in the parameters.
- For the approximating model to the function (not derivatives), write $g_K(\mathbf{x}|\theta_K) = \mathbf{z}'\theta_K$ for simplicity.

The following theorem can be used to prove the consistency of the Fourier form.

Theorem 65. *[Gallant and Nychka, 1987] Suppose that $\hat{h}_n$ is obtained by maximizing a sample objective function $s_n(h)$ over $\mathcal{H}_{K_n}$ where $\mathcal{H}_K$ is a subset of some function space $\mathcal{H}$ on which is defined a norm $\parallel h \parallel$. Consider the following conditions:*

(a) Compactness: The closure of $\mathcal{H}$ with respect to $\parallel h \parallel$ is compact in the relative topology defined by $\parallel h \parallel$.

(b) Denseness: $\cup_K \mathcal{H}_K$, $K = 1, 2, 3, ...$ *is a dense subset of the closure of* $\mathcal{H}$ *with respect to* $\| h \|$ *and* $\mathcal{H}_K \subset \mathcal{H}_{K+1}$.

(c) Uniform convergence: There is a point h^* *in* $\mathcal{H}$ *and there is a function* $s_\infty(h, h^*)$ *that is continuous in* h *with respect to* $\| h \|$ *such that*

$$\lim_{n\to\infty} \sup_{\mathcal{H}} \mid s_n(h) - s_\infty(h, h^*) \mid = 0$$

almost surely.

(d) Identification: Any point h *in the closure of* $\mathcal{H}$ *with* $s_\infty(h, h^*) \geq s_\infty(h^*, h^*)$ *must have* $\| h - h^* \| = 0$.

Under these conditions $\lim_{n\to\infty} \| h^* - \hat{h}_n \| = 0$ *almost surely, provided that* $\lim_{n\to\infty} K_n = \infty$ *almost surely.*

The modification of the original statement of the theorem that has been made is to set the parameter space Θ in Gallant and Nychka's (1987) Theorem 0 to a single point and to state the theorem in terms of maximization rather than minimization.

This theorem is very similar in form to Theorem 29. The main differences are:

1. A generic norm $\| h \|$ is used in place of the Euclidean norm. This norm may be stronger than the Euclidean norm, so that convergence with respect to $\| h \|$ implies convergence w.r.t the Euclidean norm. Typically we will want to make sure that the norm is strong enough to imply convergence of all functions of interest.

2. The "estimation space" $\mathcal{H}$ is a function space. It plays the role of the parameter space Θ in our discussion of parametric estimators. There is no restriction to a parametric family, only a

restriction to a space of functions that satisfy certain conditions. This formulation is much less restrictive than the restriction to a parametric family.

3. There is a denseness assumption that was not present in the other theorem.

We will not prove this theorem (the proof is quite similar to the proof of theorem [29], see Gallant, 1987) but we will discuss its assumptions, in relation to the Fourier form as the approximating model.

Sobolev norm Since all of the assumptions involve the norm $\| h \|$, we need to make explicit what norm we wish to use. We need a norm that guarantees that the errors in approximation of the functions we are interested in are accounted for. Since we are interested in first-order elasticities in the present case, we need close approximation of both the function $f(x)$ and its first derivative $f'(x)$, throughout the range of x. Let $\mathcal{X}$ be an open set that contains all values of x that we're interested in. The Sobolev norm is appropriate in this case. It is defined, making use of our notation for partial derivatives, as:

$$\| h \|_{m,\mathcal{X}} = \max_{|\lambda^*| \leq m} \sup_{\mathcal{X}} \left| D^{\lambda} h(x) \right|$$

To see whether or not the function $f(x)$ is well approximated by an approximating model $g_K(x \mid \theta_K)$, we would evaluate

$$\| f(\mathbf{x}) - g_K(\mathbf{x} \mid \theta_K) \|_{m,\mathcal{X}} .$$

We see that this norm takes into account errors in approximating the function and partial derivatives up to order m. If we want to estimate first order elasticities, as is the case in this example, the relevant m would be $m = 1$. Furthermore, since we examine the sup over $\mathcal{X}$, convergence w.r.t. the Sobolev means *uniform* convergence, so that we obtain consistent estimates for all values of x.

Compactness Verifying compactness with respect to this norm is quite technical and unenlightening. It is proven by Elbadawi, Gallant and Souza, *Econometrica*, 1983. The basic requirement is that if we need consistency w.r.t. $\| h \|_{m,\mathcal{X}}$, then the functions of interest must belong to a Sobolev space which takes into account derivatives of order $m+1$. A Sobolev space is the set of functions

$$\mathcal{W}_{m,\mathcal{X}}(D) = \{h(\mathbf{x}) :\| h(\mathbf{x}) \|_{m,\mathcal{X}} < D\},$$

where D is a finite constant. In plain words, the functions must have bounded partial derivatives of one order higher than the derivatives we seek to estimate.

The estimation space and the estimation subspace Since in our case we're interested in consistent estimation of first-order elasticities, we'll define the estimation space as follows:

Definition 66. [Estimation space] The estimation space $\mathcal{H} = \mathcal{W}_{2,\mathcal{X}}(D)$. The estimation space is an open set, and we presume that $h^* \in \mathcal{H}$.

So we are assuming that the function to be estimated has bounded second derivatives throughout $\mathcal{X}$.

With seminonparametric estimators, we don't actually optimize over the estimation space. Rather, we optimize over a subspace, $\mathcal{H}_{K_n}$, defined as:

Definition 67. [Estimation subspace] The estimation subspace $\mathcal{H}_K$ is defined as

$$\mathcal{H}_K = \{g_K(\mathbf{x}|\theta_K) : g_K(\mathbf{x}|\theta_K) \in \mathcal{W}_{2,\mathcal{Z}}(D), \theta_K \in \Re^K\},$$

where $g_K(\mathbf{x}, \theta_K)$ is the Fourier form approximation as defined in Equation 20.2.

Denseness The important point here is that $\mathcal{H}_K$ is a space of functions that is indexed by a finite dimensional parameter (θ_K has K elements, as in equation 20.3). With n observations, $n > K$, this parameter is estimable. Note that the true function h^* is not necessarily an element of $\mathcal{H}_K$, so optimization over $\mathcal{H}_K$ may not lead to a consistent estimator. In order for optimization over $\mathcal{H}_K$ to be equivalent to optimization over $\mathcal{H}$, at least asymptotically, we need that:

1. The dimension of the parameter vector, $\dim \theta_{K_n} \to \infty$ as $n \to \infty$. This is achieved by making A and J in equation 20.2 increasing functions of n, the sample size. It is clear that K will have to grow more slowly than n. The second requirement is:

2. We need that the $\mathcal{H}_K$ be dense subsets of $\mathcal{H}$.

The estimation subspace $\mathcal{H}_K$, defined above, is a subset of the closure of the estimation space, $\overline{\mathcal{H}}$. A set of subsets $\mathcal{A}_a$ of a set $\mathcal{A}$ is "dense" if the closure of the countable union of the subsets is equal to the closure of $\mathcal{A}$:

$$\overline{\cup_{a=1}^{\infty} \mathcal{A}_a} = \overline{\mathcal{A}}$$

Use a picture here. The rest of the discussion of denseness is provided just for completeness: there's no need to study it in detail. To show that $\mathcal{H}_K$ is a dense subset of $\overline{\mathcal{H}}$ with respect to $\| h \|_{1,\mathcal{X}}$, it is useful to apply Theorem 1 of Gallant (1982), who in turn cites Edmunds and Moscatelli (1977). We reproduce the theorem as presented by Gallant, with minor notational changes, for convenience of reference:

Theorem 68. *[Edmunds and Moscatelli, 1977] Let the real-valued function $h^*(\mathbf{x})$ be continuously differentiable up to order m on an open set containing the closure of $\mathcal{X}$. Then it is possible to choose a triangular array of coefficients $\theta_1, \theta_2, \ldots \theta_K, \ldots,$ such that for every q with $0 \leq q < m$, and every $\varepsilon > 0$, $\| h^*(\mathbf{x}) - h_K(\mathbf{x}|\theta_K) \|_{q,\mathcal{X}} = o(K^{-m+q+\varepsilon})$ as $K \to \infty$.*

In the present application, $q = 1$, and $m = 2$. By definition of the estimation space, the elements of $\mathcal{H}$ are once continuously differentiable on $\mathcal{X}$, which is open and contains the closure of $\mathcal{X}$, so the theorem is applicable. Closely following Gallant and Nychka (1987), $\cup_\infty \mathcal{H}_K$ is the countable union of the $\mathcal{H}_K$. The implication of Theorem 68 is that there is a sequence of $\{h_K\}$ from $\cup_\infty \mathcal{H}_K$ such that

$$\lim_{K\to\infty} \| h^* - h_K \|_{1,\mathcal{X}} = 0,$$

for all $h^* \in \mathcal{H}$. Therefore,

$$\mathcal{H} \subset \overline{\cup_\infty \mathcal{H}_K}.$$

However,

$$\cup_\infty \mathcal{H}_K \subset \mathcal{H},$$

so

$$\overline{\cup_\infty \mathcal{H}_K} \subset \overline{\mathcal{H}}.$$

Therefore

$$\overline{\mathcal{H}} = \overline{\cup_\infty \mathcal{H}_K},$$

so $\cup_\infty \mathcal{H}_K$ is a dense subset of $\mathcal{H}$, with respect to the norm $\| h \|_{1,\mathcal{X}}$.

Uniform convergence We now turn to the limiting objective function. We estimate by OLS. The sample objective function stated in terms of maximization is

$$s_n(\theta_K) = -\frac{1}{n} \sum_{t=1}^{n} (y_t - g_K(\mathbf{x}_t \mid \theta_K))^2$$

With random sampling, as in the case of Equations 12.1 and 19.3, the limiting objective function is

$$s_\infty(g, f) = -\int_{\mathcal{X}} (f(\mathbf{x}) - g(\mathbf{x}))^2 \, d\mu x - \sigma_\varepsilon^2. \tag{20.7}$$

where the true function $f(x)$ takes the place of the generic function h^* in the presentation of the theorem. Both $g(x)$ and $f(x)$ are elements of $\overline{\cup_\infty \mathcal{H}_K}$.

The pointwise convergence of the objective function needs to be strengthened to uniform convergence. We will simply assume that this holds, since the way to verify this depends upon the specific application. We also have continuity of the objective function in g, with respect to the norm $\| h \|_{1,\mathcal{X}}$ since

$$\begin{aligned} & \lim_{\|g^1 - g^0\|_{1,\mathcal{X}} \to 0} \left\{ s_\infty\left(g^1, f\right) - s_\infty\left(g^0, f\right) \right\} \\ = & \lim_{\|g^1 - g^0\|_{1,\mathcal{X}} \to 0} \int_{\mathcal{X}} \left[\left(g^1(\mathbf{x}) - f(\mathbf{x})\right)^2 - \left(g^0(\mathbf{x}) - f(\mathbf{x})\right)^2 \right] d\mu x. \end{aligned}$$

By the dominated convergence theorem (which applies since the finite bound D used to define $\mathcal{W}_{2,\mathcal{Z}}(D)$ is dominated by an integrable function), the limit and the integral can be interchanged, so by inspection, the limit is zero.

Identification The identification condition requires that for any point (g, f) in $\overline{\mathcal{H}} \times \overline{\mathcal{H}}$, $s_\infty(g, f) \geq s_\infty(f, f) \Rightarrow \| g - f \|_{1,\mathcal{X}} = 0$. This condition is clearly satisfied given that g and f are once continuously differentiable (by the assumption that defines the estimation space).

Review of concepts For the example of estimation of first-order elasticities, the relevant concepts are:

- Estimation space $\mathcal{H} = \mathcal{W}_{2,\mathcal{X}}(D)$: the function space in the closure of which the true function must lie.
- Consistency norm $\| h \|_{1,\mathcal{X}}$. The closure of $\mathcal{H}$ is compact with respect to this norm.
- Estimation subspace $\mathcal{H}_K$. The estimation subspace is the subset of $\mathcal{H}$ that is representable by a Fourier form with parameter θ_K. These are dense subsets of $\mathcal{H}$.
- Sample objective function $s_n(\theta_K)$, the negative of the sum of squares. By standard arguments this converges uniformly to the
- Limiting objective function $s_\infty(\ g, f)$, which is continuous in g and has a global maximum in its first argument, over the closure of the infinite union of the estimation subpaces, at $g = f$.
- As a result of this, first order elasticities

$$\frac{\mathbf{x}_i}{f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial x_i f(x)}$$

  are consistently estimated for all $\mathbf{x} \in \mathcal{X}$.

Discussion Consistency requires that the number of parameters used in the expansion increase with the sample size, tending to infinity. If parameters are added at a high rate, the bias tends relatively rapidly to zero. A basic problem is that a high rate of inclusion of additional parameters causes the variance to tend more slowly to zero. The issue of how to chose the rate at which parameters are added and which to add first is fairly complex. A problem is that the allowable rates for asymptotic

normality to obtain (Andrews 1991; Gallant and Souza, 1991) are very strict. Supposing we stick to these rates, our approximating model is:

$$g_K(\mathbf{x}|\theta_K) = \mathbf{z}'\theta_K.$$

- Define $\mathbf{Z}_K$ as the $n \times K$ matrix of regressors obtained by stacking observations. The LS estimator is

  $$\hat{\theta}_K = \left(\mathbf{Z}_K'\mathbf{Z}_K\right)^+ \mathbf{Z}_K' y,$$

  where $(\cdot)^+$ is the Moore-Penrose generalized inverse.

  - This is used since $\mathbf{Z}_K'\mathbf{Z}_K$ may be singular, as would be the case for $K(n)$ large enough when some dummy variables are included.

- . The prediction, $\mathbf{z}'\hat{\theta}_K$, of the unknown function $f(\mathbf{x})$ is asymptotically normally distributed:

  $$\sqrt{n}\left(\mathbf{z}'\hat{\theta}_K - f(x)\right) \xrightarrow{d} N(0, AV),$$

  where

  $$AV = \lim_{n\to\infty} E\left[\mathbf{z}'\left(\frac{\mathbf{Z}_K'\mathbf{Z}_K}{n}\right)^+ \mathbf{z}\hat{\sigma}^2\right].$$

  Formally, this is exactly the same as if we were dealing with a parametric linear model. I emphasize, though, that this is only valid if K grows very slowly as n grows. If we can't stick to acceptable rates, we should probably use some other method of approximating the small sample distribution. Bootstrapping is a possibility. We'll discuss this in the section on simulation.

Kernel regression estimators

Readings: Bierens, 1987, "Kernel estimators of regression functions," in *Advances in Econometrics, Fifth World Congress,* V. 1, Truman Bewley, ed., Cambridge.

An alternative method to the semi-nonparametric method is a fully nonparametric method of estimation. Kernel regression estimation is an example (others are splines, nearest neighbor, etc.). We'll consider the Nadaraya-Watson kernel regression estimator in a simple case.

- Suppose we have an iid sample from the joint density $f(x, y)$, where x is k -dimensional. The model is

$$y_t = g(x_t) + \varepsilon_t,$$

where

$$E(\varepsilon_t | x_t) = 0.$$

- The conditional expectation of y given x is $g(x)$. By definition of the conditional expectation, we have

$$\begin{aligned} g(x) &= \int y \frac{f(x, y)}{h(x)} dy \\ &= \frac{1}{h(x)} \int y f(x, y) dy, \end{aligned}$$

where $h(x)$ is the marginal density of x :

$$h(x) = \int f(x, y) dy.$$

- This suggests that we could estimate $g(x)$ by estimating $h(x)$ and $\int yf(x,y)dy$.

Estimation of the denominator

A kernel estimator for $h(x)$ has the form

$$\hat{h}(x) = \frac{1}{n}\sum_{t=1}^{n} \frac{K\left[\left(x - x_t\right)/\gamma_n\right]}{\gamma_n^k},$$

where n is the sample size and k is the dimension of x.

- The function $K(\cdot)$ (the kernel) is absolutely integrable:

$$\int |K(x)|dx < \infty,$$

  and $K(\cdot)$ integrates to 1 :

$$\int K(x)dx = 1.$$

  In this respect, $K(\cdot)$ is like a density function, but we do not necessarily restrict $K(\cdot)$ to be nonnegative.

- The *window width* parameter, γ_n is a sequence of positive numbers that satisfies

$$\begin{aligned} \lim_{n\to\infty} \gamma_n &= 0 \\ \lim_{n\to\infty} n\gamma_n^k &= \infty \end{aligned}$$

  So, the window width must tend to zero, but not too quickly.

- To show pointwise consistency of $\hat{h}(x)$ for $h(x)$, first consider the expectation of the estimator (because the estimator is an average of iid terms, we only need to consider the expectation of a representative term):

$$E\left[\hat{h}(x)\right] = \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right] h(z) dz.$$

Change variables as $z^* = (x-z)/\gamma_n$, so $z = x - \gamma_n z^*$ and $|\frac{dz}{dz^{*\prime}}| = \gamma_n^k$, we obtain

$$\begin{aligned} E\left[\hat{h}(x)\right] &= \int \gamma_n^{-k} K\left(z^*\right) h(x - \gamma_n z^*) \gamma_n^k dz^* \\ &= \int K\left(z^*\right) h(x - \gamma_n z^*) dz^*. \end{aligned}$$

Now, asymptotically,

$$\begin{aligned} \lim_{n\to\infty} E\left[\hat{h}(x)\right] &= \lim_{n\to\infty} \int K\left(z^*\right) h(x - \gamma_n z^*) dz^* \\ &= \int \lim_{n\to\infty} K\left(z^*\right) h(x - \gamma_n z^*) dz^* \\ &= \int K\left(z^*\right) h(x) dz^* \\ &= h(x) \int K\left(z^*\right) dz^* \\ &= h(x), \end{aligned}$$

since $\gamma_n \to 0$ and $\int K\left(z^*\right) dz^* = 1$ by assumption. (Note: that we can pass the limit through the integral is a result of the dominated convergence theorem. For this to hold we need that $h(\cdot)$ be dominated by an absolutely integrable function.)

- Next, considering the variance of $\hat{h}(x)$, we have, due to the iid assumption

$$\begin{aligned} n\gamma_n^k V\left[\hat{h}(x)\right] &= n\gamma_n^k \frac{1}{n^2}\sum_{t=1}^{n} V\left\{\frac{K\left[(x-x_t)/\gamma_n\right]}{\gamma_n^k}\right\} \\ &= \gamma_n^{-k}\frac{1}{n}\sum_{t=1}^{n} V\left\{K\left[(x-x_t)/\gamma_n\right]\right\} \end{aligned}$$

- By the representative term argument, this is

$$n\gamma_n^k V\left[\hat{h}(x)\right] = \gamma_n^{-k} V\left\{K\left[(x-z)/\gamma_n\right]\right\}$$

- Also, since $V(x) = E(x^2) - E(x)^2$ we have

$$\begin{aligned} n\gamma_n^k V\left[\hat{h}(x)\right] &= \gamma_n^{-k} E\left\{\left(K\left[(x-z)/\gamma_n\right]\right)^2\right\} - \gamma_n^{-k}\left\{E\left(K\left[(x-z)/\gamma_n\right]\right)\right\}^2 \\ &= \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right]^2 h(z)dz - \gamma_n^k\left\{\int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right] h(z)dz\right\}^2 \\ &= \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right]^2 h(z)dz - \gamma_n^k E\left[\widehat{h}(x)\right]^2 \end{aligned}$$

  The second term converges to zero:

$$\gamma_n^k E\left[\widehat{h}(x)\right]^2 \to 0,$$

  by the previous result regarding the expectation and the fact that $\gamma_n \to 0$. Therefore,

$$\lim_{n\to\infty} n\gamma_n^k V\left[\hat{h}(x)\right] = \lim_{n\to\infty} \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right]^2 h(z)dz.$$

Using exactly the same change of variables as before, this can be shown to be

$$\lim_{n\to\infty} n\gamma_n^k V\left[\hat{h}(x)\right] = h(x)\int \left[K(z^*)\right]^2 dz^*.$$

Since both $\int \left[K(z^*)\right]^2 dz^*$ and $h(x)$ are bounded, this is bounded, and since $n\gamma_n^k \to \infty$ by assumption, we have that

$$V\left[\hat{h}(x)\right] \to 0.$$

- Since the bias and the variance both go to zero, we have pointwise consistency (convergence in quadratic mean implies convergence in probability).

Estimation of the numerator

To estimate $\int yf(x,y)dy$, we need an estimator of $f(x,y)$. The estimator has the same form as the estimator for $h(x)$, only with one dimension more:

$$\hat{f}(x,y) = \frac{1}{n}\sum_{t=1}^{n} \frac{K_*\left[\left(y-y_t\right)/\gamma_n, \left(x-x_t\right)/\gamma_n\right]}{\gamma_n^{k+1}}$$

The kernel $K_*(\cdot)$ is required to have mean zero:

$$\int yK_*(y,x)\,dy = 0$$

and to marginalize to the previous kernel for $h(x)$:

$$\int K_*(y,x)\,dy = K(x).$$

With this kernel, we have

$$\int y\hat{f}(y,x)dy = \frac{1}{n}\sum_{t=1}^{n} y_t \frac{K\left[(x-x_t)/\gamma_n\right]}{\gamma_n^k}$$

by marginalization of the kernel, so we obtain

$$\begin{aligned}
\hat{g}(x) &= \frac{1}{\hat{h}(x)}\int y\hat{f}(y,x)dy \\
&= \frac{\frac{1}{n}\sum_{t=1}^{n} y_t \frac{K[(x-x_t)/\gamma_n]}{\gamma_n^k}}{\frac{1}{n}\sum_{t=1}^{n} \frac{K[(x-x_t)/\gamma_n]}{\gamma_n^k}} \\
&= \frac{\sum_{t=1}^{n} y_t K\left[(x-x_t)/\gamma_n\right]}{\sum_{t=1}^{n} K\left[(x-x_t)/\gamma_n\right]}.
\end{aligned}$$

This is the Nadaraya-Watson kernel regression estimator.

Discussion

- The kernel regression estimator for $g(x_t)$ is a weighted average of the y_j, $j = 1, 2, ..., n$, where higher weights are associated with points that are closer to x_t. The weights sum to 1. See this link for a graphic interpretation.
- The window width parameter γ_n imposes smoothness. The estimator is increasingly flat as $\gamma_n \to \infty$, since in this case each weight tends to $1/n$.
- A large window width reduces the variance (strong imposition of flatness), but increases the bias.
- A small window width reduces the bias, but makes very little use of information except points that are in a small neighborhood of x_t. Since relatively little information is used, the variance is

large when the window width is small.

- The standard normal density is a popular choice for $K(.)$ and $K_*(y, x)$, though there are possibly better alternatives.

Choice of the window width: Cross-validation

The selection of an appropriate window width is important. One popular method is cross validation. This consists of splitting the sample into two parts (e.g., 50%-50%). The first part is the "in sample" data, which is used for estimation, and the second part is the "out of sample" data, used for evaluation of the fit though RMSE or some other criterion. The steps are:

1. Split the data. The out of sample data is y^{out} and x^{out}.
2. Choose a window width γ.
3. With the in sample data, fit $\hat{y}_t^{out}$ corresponding to each x_t^{out}. This fitted value is a function of the in sample data, as well as the evaluation point x_t^{out}, but it does not involve y_t^{out}.
4. Repeat for all out of sample points.
5. Calculate RMSE(γ)
6. Go to step 2, or to the next step if enough window widths have been tried.
7. Select the γ that minimizes RMSE(γ) (Verify that a minimum has been found, for example by plotting RMSE as a function of γ).

8. Re-estimate using the best γ and all of the data.

This same principle can be used to choose A and J in a Fourier form model.

20.4 Density function estimation

Kernel density estimation

The previous discussion suggests that a kernel density estimator may easily be constructed. We have already seen how joint densities may be estimated. If were interested in a conditional density, for example of y conditional on x, then the kernel estimate of the conditional density is simply

$$\begin{aligned}
\widehat{f}_{y|x} &= \frac{\hat{f}(x,y)}{\hat{h}(x)} \\
&= \frac{\frac{1}{n}\sum_{t=1}^{n}\frac{K_*[(y-y_t)/\gamma_n,(x-x_t)/\gamma_n]}{\gamma_n^{k+1}}}{\frac{1}{n}\sum_{t=1}^{n}\frac{K[(x-x_t)/\gamma_n]}{\gamma_n^{k}}} \\
&= \frac{1}{\gamma_n}\frac{\sum_{t=1}^{n}K_*\left[\left(y-y_t\right)/\gamma_n,\left(x-x_t\right)/\gamma_n\right]}{\sum_{t=1}^{n}K\left[\left(x-x_t\right)/\gamma_n\right]}
\end{aligned}$$

where we obtain the expressions for the joint and marginal densities from the section on kernel regression.

Semi-nonparametric maximum likelihood

Readings: Gallant and Nychka, *Econometrica*, 1987. For a Fortran program to do this and a useful discussion in the user's guide, see this link. See also Cameron and Johansson, *Journal of Applied Econometrics*, V. 12, 1997.

MLE is the estimation method of choice when we are confident about specifying the density. Is is possible to obtain the benefits of MLE when we're not so confident about the specification? In part, yes.

Suppose we're interested in the density of y conditional on x (both may be vectors). Suppose that the density $f(y|x,\phi)$ is a reasonable starting approximation to the true density. This density can be reshaped by multiplying it by a squared polynomial. The new density is

$$g_p(y|x,\phi,\gamma) = \frac{h_p^2(y|\gamma)f(y|x,\phi)}{\eta_p(x,\phi,\gamma)}$$

where

$$h_p(y|\gamma) = \sum_{k=0}^{p} \gamma_k y^k$$

and $\eta_p(x,\phi,\gamma)$ is a normalizing factor to make the density integrate (sum) to one. Because $h_p^2(y|\gamma)/\eta_p(x,\phi,\gamma)$ is a homogenous function of θ it is necessary to impose a normalization to identify the parameters: γ_0

is set to 1. The normalization factor $\eta_p(\phi, \gamma)$ is calculated (following Cameron and Johansson) using

$$\begin{aligned} E(Y^r) &= \sum_{y=0}^{\infty} y^r f_Y(y|\phi,\gamma) \\ &= \sum_{y=0}^{\infty} y^r \frac{\left[h_p\left(y|\gamma\right)\right]^2}{\eta_p(\phi,\gamma)} f_Y(y|\phi) \\ &= \sum_{y=0}^{\infty} \sum_{k=0}^{p} \sum_{l=0}^{p} y^r f_Y(y|\phi)\gamma_k\gamma_l y^k y^l / \eta_p(\phi,\gamma) \\ &= \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k\gamma_l \left\{ \sum_{y=0}^{\infty} y^{r+k+l} f_Y(y|\phi) \right\} / \eta_p(\phi,\gamma) \\ &= \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k\gamma_l m_{k+l+r} / \eta_p(\phi,\gamma). \end{aligned}$$

By setting $r = 0$ we get that the normalizing factor is

20.8

$$\eta_p(\phi,\gamma) = \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k\gamma_l m_{k+l} \tag{20.8}$$

Recall that γ_0 is set to 1 to achieve identification. The m_r in equation 20.8 are the raw moments of the baseline density. Gallant and Nychka (1987) give conditions under which such a density may be treated as correctly specified, asymptotically. Basically, the order of the polynomial must increase as the sample size increases. However, there are technicalities.

Similarly to Cameron and Johannson (1997), we may develop a negative binomial polynomial (NBP) density for count data. The negative binomial baseline density may be written (see equation 18.1) as

$$f_Y(y|\phi) = \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^{\psi} \left(\frac{\lambda}{\psi+\lambda}\right)^{y}$$

where $\phi = \{\lambda, \psi\}$, $\lambda > 0$ and $\psi > 0$. The usual means of incorporating conditioning variables $\mathbf{x}$ is the parameterization $\lambda = e^{\mathbf{x}'\beta}$. When $\psi = \lambda/\alpha$ we have the negative binomial-I model (NB-I). When $\psi = 1/\alpha$ we have the negative binomial-II (NP-II) model. For the NB-I density, $V(Y) = \lambda + \alpha\lambda$. In the case of the NB-II model, we have $V(Y) = \lambda + \alpha\lambda^2$. For both forms, $E(Y) = \lambda$.

The reshaped density, with normalization to sum to one, is

$$f_Y(y|\phi,\gamma) = \frac{\left[h_p\left(y|\gamma\right)\right]^2}{\eta_p(\phi,\gamma)} \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^{\psi} \left(\frac{\lambda}{\psi+\lambda}\right)^{y}. \tag{20.9}$$

To get the normalization factor, we need the moment generating function:

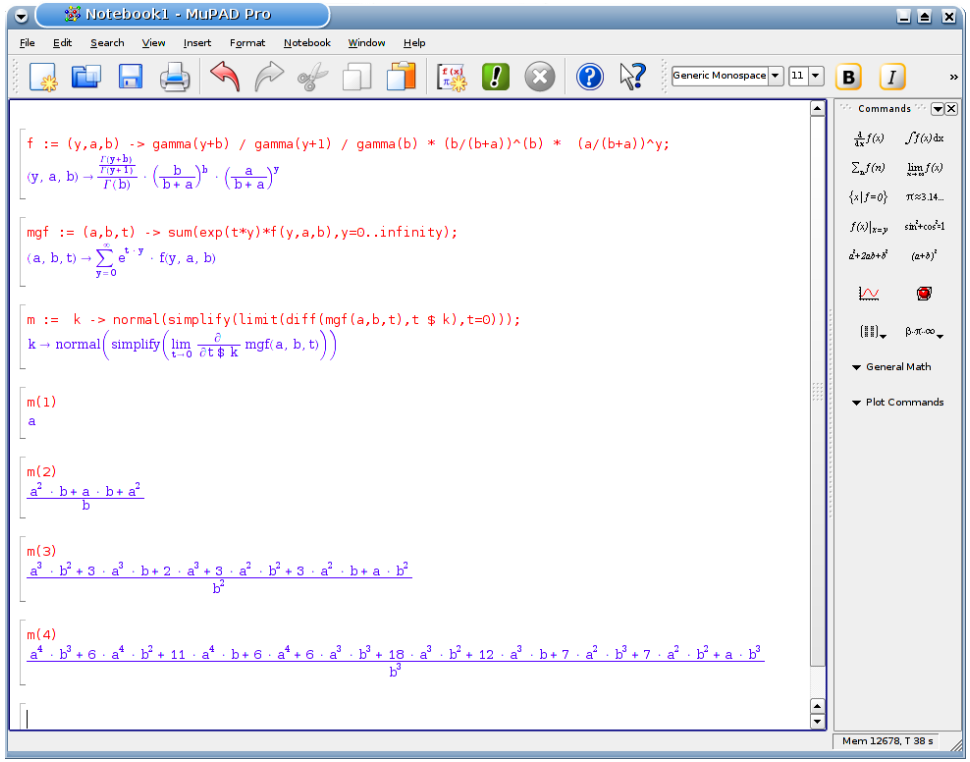
$$M_Y(t) = \psi^{\psi}\left(\lambda - e^t\lambda + \psi\right)^{-\psi}. \tag{20.10}$$

To illustrate, Figure 20.5 shows calculation of the first four raw moments of the NB density, calculated using MuPAD, which is a Computer Algebra System that (used to be?) free for personal use. These are the moments you would need to use a second order polynomial ($p = 2$). MuPAD will output these results in the form of C code, which is relatively easy to edit to write the likelihood function for the model. This has been done in NegBinSNP.cc, which is a C++ version of this model that can be compiled to use with octave using the `mkoctfile` command. Note the impressive length of the expressions when the degree of the expansion is 4 or 5! This is an example of a model that would be difficult to formulate without the help of a program like *MuPAD.*

It is possible that there is conditional heterogeneity such that the appropriate reshaping should be more local. This can be accomodated by allowing the γ_k parameters to depend upon the conditioning variables, for example using polynomials.

Gallant and Nychka, *Econometrica*, 1987 prove that this sort of density can approximate a wide

Figure 20.5: Negative binomial raw moments




```
f := (y,a,b) -> gamma(y+b) / gamma(y+1) / gamma(b) * (b/(b+a))^(b) *  (a/(b+a))^y;
```

$$(y, a, b) \to \frac{\Gamma(y+b)}{\Gamma(y+1)\,\Gamma(b)} \cdot \left(\frac{b}{b+a}\right)^{b} \cdot \left(\frac{a}{b+a}\right)^{y}$$

```
mgf := (a,b,t) -> sum(exp(t*y)*f(y,a,b),y=0..infinity);
```

$$(a, b, t) \to \sum_{y=0}^{\infty} e^{t\cdot y} \cdot f(y, a, b)$$

```
m :=  k -> normal(simplify(limit(diff(mgf(a,b,t),t $ k),t=0)));
```

$$k \to \mathrm{normal}\left(\mathrm{simplify}\left(\lim_{t\to 0} \frac{\partial^{k}}{\partial t^{k}}\, \mathrm{mgf}(a, b, t)\right)\right)$$

```
m(1)
```

$$a$$

```
m(2)
```

$$\frac{a^2 \cdot b + a \cdot b + a^2}{b}$$

```
m(3)
```

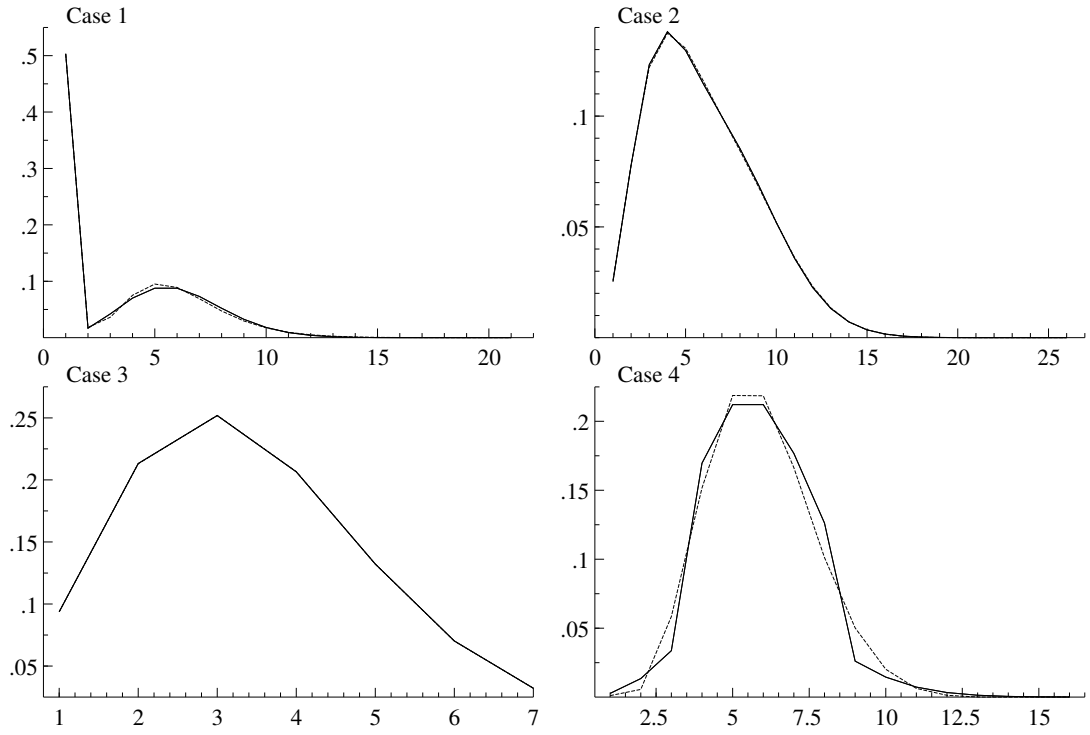
$$\frac{a^3 \cdot b^2 + 3 \cdot a^3 \cdot b + 2 \cdot a^3 + 3 \cdot a^2 \cdot b^2 + 3 \cdot a^2 \cdot b + a \cdot b^2}{b^2}$$

```
m(4)
```

$$\frac{a^4 \cdot b^3 + 6 \cdot a^4 \cdot b^2 + 11 \cdot a^4 \cdot b + 6 \cdot a^4 + 6 \cdot a^3 \cdot b^3 + 18 \cdot a^3 \cdot b^2 + 12 \cdot a^3 \cdot b + 7 \cdot a^2 \cdot b^3 + 7 \cdot a^2 \cdot b^2 + a \cdot b^3}{b^3}$$

variety of densities arbitrarily well as the degree of the polynomial increases with the sample size. This approach is not without its drawbacks: the sample objective function can have an *extremely* large number of local maxima that can lead to numeric difficulties. If someone could figure out how to do in a way such that the sample objective function was nice and smooth, they would probably get the paper published in a good journal. Any ideas?

Here's a plot of true and the limiting SNP approximations (with the order of the polynomial fixed) to four different count data densities, which variously exhibit over and underdispersion, as well as excess zeros. The baseline model is a negative binomial density.



20.5 Examples

MEPS health care usage data

We'll use the MEPS OBDV data to illustrate kernel regression and semi-nonparametric maximum likelihood.

Kernel regression estimation

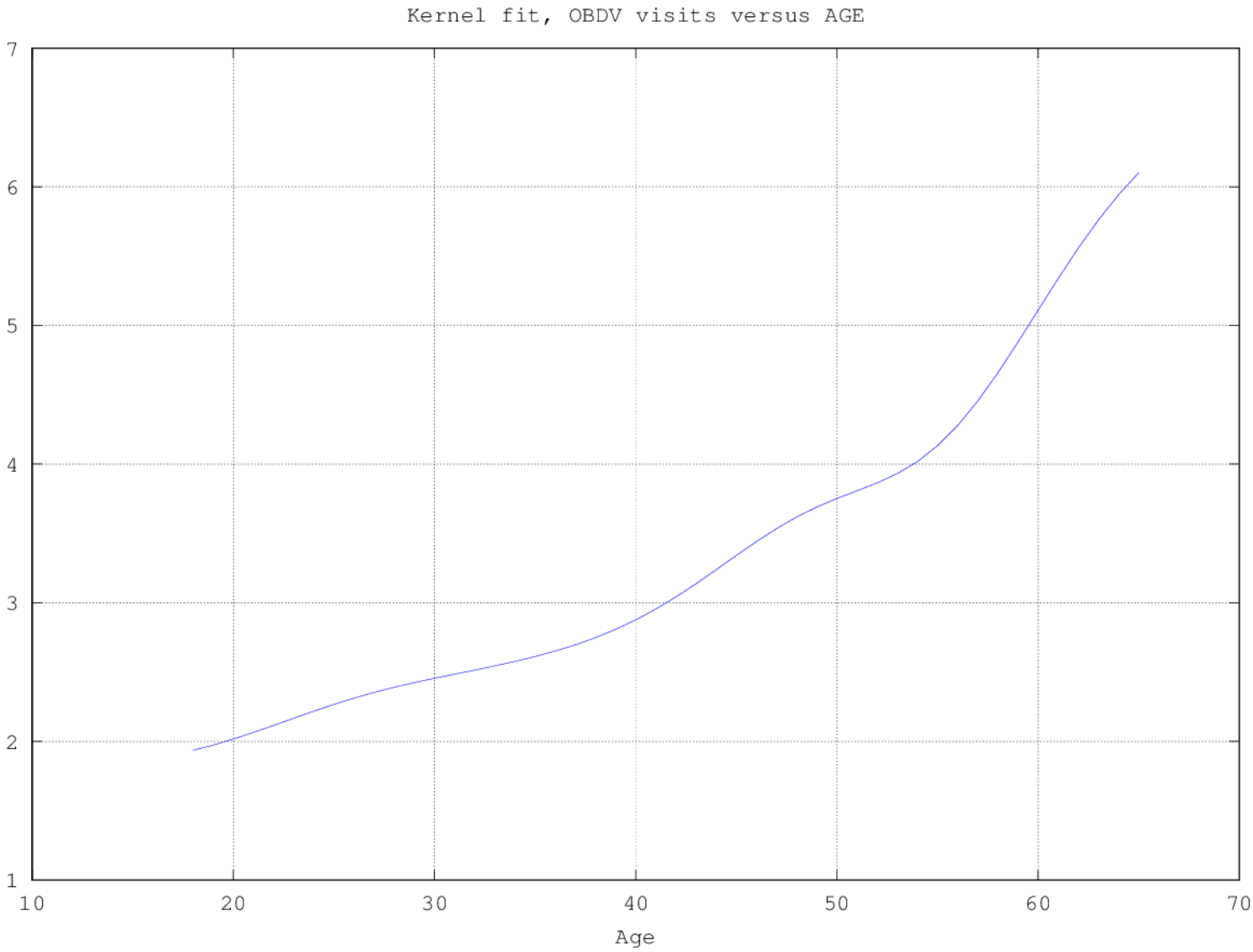
Let's try a kernel regression fit for the OBDV data. The program OBDVkernel.m loads the MEPS OBDV data, scans over a range of window widths and calculates leave-one-out CV scores, and plots the fitted OBDV usage versus AGE, using the best window width. The plot is in Figure 20.6. Note that usage increases with age, just as we've seen with the parametric models. Once could use bootstrapping to generate a confidence interval to the fit.

Seminonparametric ML estimation and the MEPS data

Now let's estimate a seminonparametric density for the OBDV data. We'll reshape a negative binomial density, as discussed above. The program EstimateNBSNP.m loads the MEPS OBDV data and estimates the model, using a NB-I baseline density and a 2nd order polynomial expansion. The output is:

```
OBDV

==========================================================
```

Figure 20.6: Kernel fitted OBDV usage versus AGE

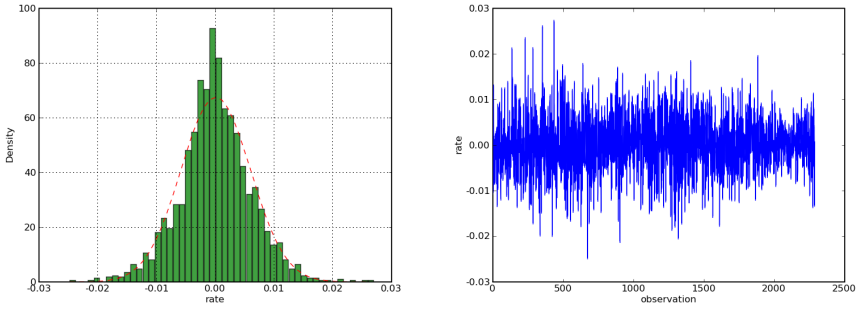


```
BFGSMIN final results

Used numeric gradient

---------------------------------------------------
STRONG CONVERGENCE
Function conv 1  Param conv 1  Gradient conv 1
---------------------------------------------------
Objective function value 2.17061
Stepsize 0.0065
24 iterations
---------------------------------------------------

 param    gradient  change
  1.3826   0.0000  -0.0000
  0.2317  -0.0000   0.0000
  0.1839   0.0000   0.0000
  0.2214   0.0000  -0.0000
  0.1898   0.0000  -0.0000
  0.0722   0.0000  -0.0000
 -0.0002   0.0000  -0.0000
  1.7853  -0.0000  -0.0000
 -0.4358   0.0000  -0.0000
  0.1129   0.0000   0.0000


***************************************************
NegBin SNP model, MEPS full data set
```

```
MLE Estimation Results
BFGS convergence: Normal convergence

Average Log-L: -2.170614
Observations: 4564

              estimate     st. err      t-stat     p-value
constant        -0.147       0.126      -1.173       0.241
pub. ins.        0.695       0.050      13.936       0.000
priv. ins.       0.409       0.046       8.833       0.000
sex              0.443       0.034      13.148       0.000
age              0.016       0.001      11.880       0.000
edu              0.025       0.006       3.903       0.000
inc             -0.000       0.000      -0.011       0.991
gam1             1.785       0.141      12.629       0.000
gam2            -0.436       0.029     -14.786       0.000
lnalpha          0.113       0.027       4.166       0.000

Information Criteria
CAIC : 19907.6244      Avg. CAIC:   4.3619
 BIC : 19897.6244       Avg. BIC:   4.3597
 AIC : 19833.3649       Avg. AIC:   4.3456
**************************************************
```

Note that the CAIC and BIC are lower for this model than for the models presented in Table 18.3. This model fits well, still being parsimonious. You can play around trying other use measures, using a NP-II baseline density, and using other orders of expansions. Density functions formed in this way may have **MANY** local maxima, so you need to be careful before accepting the results of a casual run.

Figure 20.7: Dollar-Euro



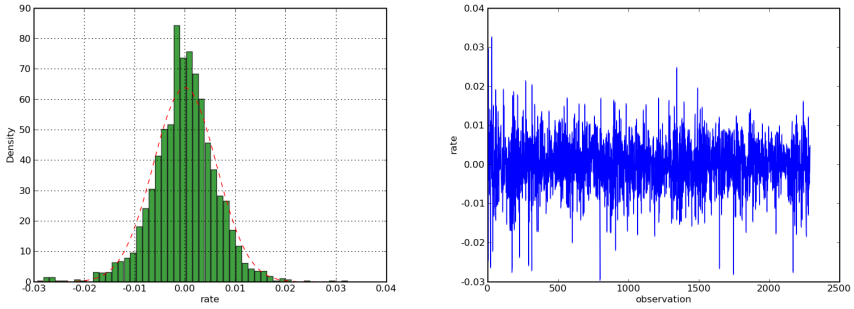
To guard against having converged to a local maximum, one can try using multiple starting values, or one could try simulated annealing as an optimization method. If you uncomment the relevant lines in the program, you can use SA to do the minimization. This will take a *lot* of time, compared to the default BFGS minimization. The chapter on parallel computations might be interesting to read before trying this.

Financial data and volatility

The data set rates contains the growth rate (100×log difference) of the daily spot $/euro and $/yen exchange rates at New York, noon, from January 04, 1999 to February 12, 2008. There are 2291 observations. See the README file for details. Figures 20.7 and 20.8 show the data and their histograms.

- at the center of the histograms, the bars extend above the normal density that best fits the data, and the tails are fatter than those of the best fit normal density. This feature of the data is known as *leptokurtosis*.

Figure 20.8: Dollar-Yen



- in the series plots, we can see that the variance of the growth rates is not constant over time. Volatility clusters are apparent, alternating between periods of stability and periods of more wild swings. This is known as *conditional heteroscedasticity*. ARCH and GARCH well-known models that are often applied to this sort of data.

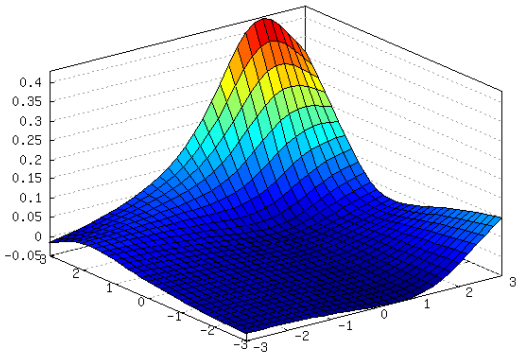
- Many structural economic models often cannot generate data that exhibits conditional heteroscedasticity without directly assuming shocks that are conditionally heteroscedastic. It would be nice to have an economic explanation for how conditional heteroscedasticity, leptokurtosis, and other (leverage, etc.) features of financial data result from the behavior of economic agents, rather than from a black box that provides shocks.

The Octave script kernelfit.m performs kernel regression to fit $E(y_t^2|y_{t-1}^2,y_{t-2}^2)$, and generates the plots in Figure 20.9.

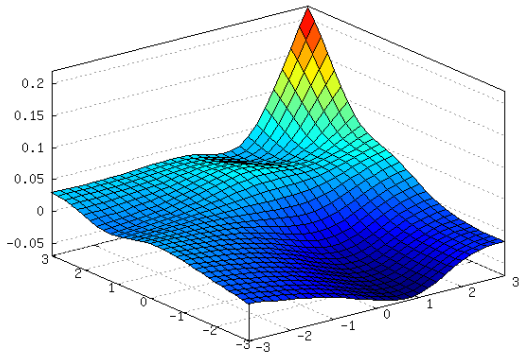
- From the point of view of learning the practical aspects of kernel regression, note how the data is compactified in the example script.

Figure 20.9: Kernel regression fitted conditional second moments, Yen/Dollar and Euro/Dollar

(a) Yen/Dollar



(b) Euro/Dollar



- In the Figure, note how current volatility depends on lags of the squared return rate - it is high when both of the lags are high, but drops off quickly when either of the lags is low.
- The fact that the plots are not flat suggests that this conditional moment contain information about the process that generates the data. Perhaps attempting to match this moment might be a means of estimating the parameters of the dgp. We'll come back to this later.

Limited information nonparametric filtering

Add discussion from JEF paper.

20.6 Exercises

1. In Octave, type "`edit kernel_example`".

   (a) Look this script over, and describe in words what it does.

   (b) Run the script and interpret the output.

   (c) Experiment with different bandwidths, and comment on the effects of choosing small and large values.

2. In Octave, type "`help kernel_regression`".

   (a) How can a kernel fit be done without supplying a bandwidth?

   (b) How is the bandwidth chosen if a value is not provided?

   (c) What is the default kernel used?

3. Using the Octave script OBDVkernel.m as a model, plot kernel regression fits for OBDV visits as a function of income and education.

Chapter 21

Quantile regression

References: Cameron and Trivedi, Chapter 4, and Chernozhukov's MIT OpenCourseWare notes, lecture 8 Chernozhukov's quantile reg notes.

This chapter gives a brief outline of quantile regression. The intention is to learn what quantile regression is, and its potential uses, but without going into the topic in depth.

21.1 Quantiles of the linear regression model

The classical linear regression model $y_t = x_t'\beta + \epsilon_t$ with normal errors implies that the distribution of y_t conditional on x_t is

$$y_t \sim N(x_t'\beta, \sigma^2)$$

The α quantile of Y, conditional on $X = x$ (notation: $Y_{\alpha|X=x}$) is the smallest value z such that $Pr(Y \leq z|X = x) = \alpha$. If $F_{Y|X=x}$ is the conditional CDF of Y, then the α-conditional quantile is

$Y_{\alpha|X=x} = \inf y : \alpha \leq F_{Y|X=x}(y)$. When $\alpha = 0.5$, we are talking about the conditional median $Y_{0.5|X=x}$, but we could be interested in other quantiles, too.

Note that $Pr(Y < x'\beta|X = x) = 0.5$ when the model follows the classical assumptions with normal errors, because the normal distribution is symmetric about the mean, so $Y_{0.5|X=x} = x'\beta$. One can estimate the conditional median just by using the fitted conditional mean, because the mean and median are the same given normality.

How about other quantiles? We have $y = x'\beta + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$. Conditional on x, $x'\beta$ is given, and the distribution of ϵ does not depend on x. Note that ϵ/σ is standard normal, and the α quantile of ϵ/σ is simply the inverse of the standard normal CDF evaulated at α, $\Phi^{-1}(\alpha)$, where Φ is the standard normal CDF function. The probit function $\Phi^{-1}(\alpha)$ is tabulated (or can be found in Octave using the norminv function). It is plotted in Figure 21.1.

The α quantile of ϵ is $\sigma\Phi^{-1}(\alpha)$. Thus, the α conditional quantile of y is $Y_{\alpha|X=x} = x'\beta + \sigma\Phi^{-1}(\alpha)$. Some quantiles are pictured in Figure 21.2. These give confidence intervals for the the fitted value, $x'\beta$.

- The conditional quantiles for the classical model are linear functions of x
- all have the same slope: the only thing that changes with α is the intercept $\sigma\Phi^{-1}(\alpha)$.
- If the error is heteroscedastic, so that $\sigma = \sigma(x)$, quantiles can have different slopes. *Draw a picture.*

Figure 21.1: Inverse CDF for N(0,1)

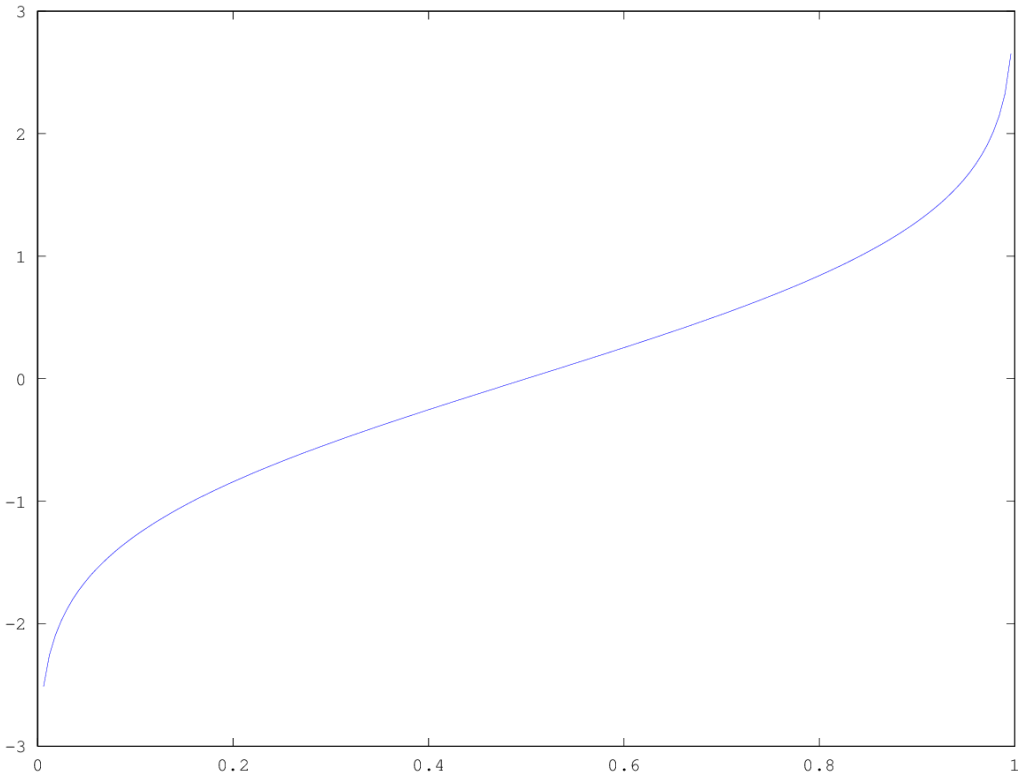
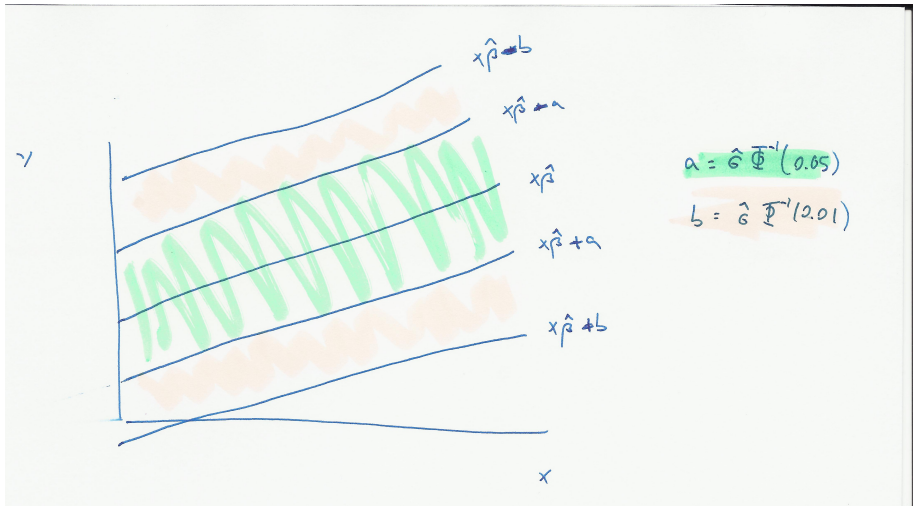


Figure 21.2: Quantiles of classical linear regression model



21.2 Fully nonparametric conditional quantiles

To compute conditional quantiles for the classical linear model, we used the assumption of normality. Can we estimate conditional quantiles without making distributional assumptions? Yes, we can! (nod to Obama). You can do fully nonparametric conditional density estimation, as in Chapter 20, and use the fitted conditional density to compute quantiles.

- Note that estimating quantiles where α is close to 0 or 1 is difficult, because you have few observations that lie in the neighborhood of the quantile, so you should expect a large variance if you go the nonparametric route. For more central quantiles, like the median, this will be less of a problem.

- For this reason, we may go the semi-parametric route, which imposes more structure. When people talk about quantile regression, they usually mean the semi-parametric approach.

21.3 Quantile regression as a semi-parametric estimator

The most widely used method does not take either of the extreme positions, it is not fully parametric, like the linear regression model with known distribution of errors, but some parametric restrictions are made, to improve efficiency compared to the fully nonparametric approach.

- The assumption is that the α-conditional quantile of the dependent variable Y is a linear function of the conditioning variables X: $Y_{\alpha|X=x} = x'\beta_\alpha$.
- This is a generalization of what we get from the classical model with normality, where the slopes of the quantiles with respect to the regressors are constant for all α.
  - For the classical model with normality, $\frac{\partial}{\partial x} Y_{\alpha|X=x} = \beta$.
  - With the assumption of linear quantiles without distributional assumptions, $\frac{\partial}{\partial x} Y_{\alpha|X=x} = \beta_\alpha$, so the slopes are allowed to change with α.
- This is a step in the direction of flexibility, but it also means we need to estimate many parameters if we're interested in many quantiles: there may be an efficiency loss due to using many parameters to avoid distributional assumptions.
- The question is how to estimate β_α when we don't make distributional assumptions.

It turns out that the problem can be expressed as an extremum estimator, $\widehat{\beta_\alpha} = \arg\min s_n(\beta)$ where

$$s_n(\beta) = \sum_{i=1}^{n} \left[1(y_i \geq x_i'\beta_\alpha)\alpha + 1(y_i < x_i'\beta_\alpha)(1-\alpha)\right] |y_i - x_i'\beta_\alpha|$$

First, suppose that $\alpha = 0.5$, so we are estimating the median. Then the objective simplifies to minimizing the absolute deviations:

$$s_n(\beta) = \sum_{i=1}^{n} |y_i - x_i'\beta_\alpha|$$

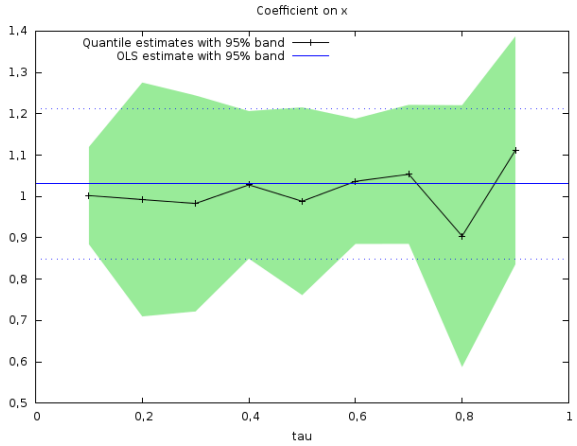
The presence of the weights in the general version accounts for the fact that if we're estimating the $\alpha = 0.1$ quantile, we expect 90% of the y_i to be greater than $x_i'\beta_\alpha$, and only 10% to be smaller. We need to downweight the likely events and upweight the unlikely events so that the objective function minimizes at the appropriate place.

- One note is that median regression may be a useful means of dealing with data that satisfies the classical assumptions, except for contamination by outliers. *In class, use Gretl to show this.*
- Note that the quantile regression objective function is discontinuous. Minimization can be done quickly using linear programming. BFGS won't work.
- the asymptotic distribution is normal, with the sandwich form typical of extremum estimators. Estimation of the terms is not completely straightforward, so methods like bootstrapping may be preferable.
- the asymptotic variance depends upon which quantile we're estimating. When α is close to 0 or 1, the asymptotic variance becomes large, and the asymptotic appoximation is unreliable for the small sample distribution. Extreme quantiles are hard to estimate with precision, because the data is sparse in those regions.

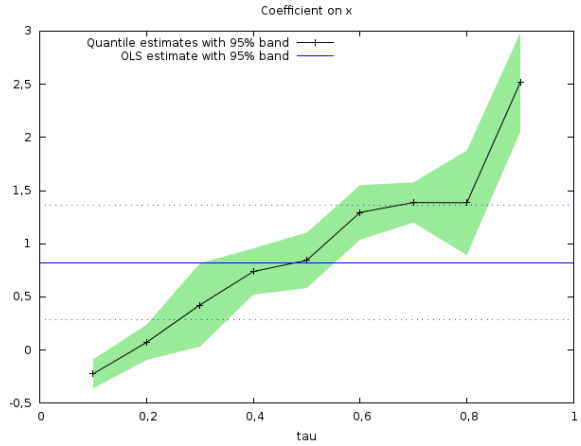
The artificial data set quantile.gdt allows you to explore quantile regression with GRETL, and to see how median regression can help to deal with data contamination.

Figure 21.3: Quantile regression results

(a) homoscedastic data



(b) heteroscedastic data



- If you do quantile regression of the variable y versus x, we are in a situation where the assumptions of the classical model hold. Quantiles all have approximately the same slope (the true value is 1).
- With heteroscedastic data, the quantiles have different slopes.
- see Figure 21.3

Chapter 22

Simulation-based methods for estimation and inference

Readings: Gourieroux and Monfort (1996) *Simulation-Based Econometric Methods* (Oxford University Press). There are many articles. Some of the seminal papers are Gallant and Tauchen (1996), "Which Moments to Match?", ECONOMETRIC THEORY, Vol. 12, 1996, pages 657-681; Gourieroux, Monfort and Renault (1993), "Indirect Inference," *J. Apl. Econometrics;* Pakes and Pollard (1989) *Econometrica*; McFadden (1989) *Econometrica.*

Simulation-based methods use computer power as a major input to do econometrics. Of course, computer power has always been used, but when intensive use of computer power is contemplated, it is possible to do things that are otherwise infeasible. Examples include obtaining more accurate results that what asymptotic theory gives us, using methods like bootstrapping, or to perform estimation using simulation, when analytic expressions for objective functions that define estimators are not available.

22.1 Motivation

Simulation methods are of interest when the DGP is fully characterized by a parameter vector, so that simulated data can be generated, but the likelihood function and moments of the observable varables are not calculable, so that MLE or GMM estimation is not possible. Many moderately complex models result in intractible likelihoods or moments, as we will see. Simulation-based estimation methods open up the possibility to estimate truly complex models. The desirability introducing a great deal of complexity may be an issue[1], but it least it becomes a possibility.

Example: Multinomial and/or dynamic discrete response models

(following McFadden, 1989)

Let y_i^* be a latent random vector of dimension m. Suppose that

$$y_i^* = X_i\beta + \varepsilon_i$$

where X_i is $m \times K$. Suppose that

$$\varepsilon_i \sim N(0, \Omega) \tag{22.1}$$

Henceforth drop the i subscript when it is not needed for clarity.

- y^* is not observed. Rather, we observe a many-to-one mapping

$$y = \tau(y^*)$$

[1] Remember that a model is an abstraction from reality, and abstraction helps us to isolate the important features of a phenomenon.

This mapping is such that each element of y is either zero or one (in some cases only one element will be one).

- Define

$$A_i = A(y_i) = \{y^* | y_i = \tau(y^*)\}$$

Suppose random sampling of (y_i, X_i). In this case the elements of y_i may not be independent of one another (and clearly are not if Ω is not diagonal). However, y_i is independent of y_j, $i \neq j$.

- Let $\theta = (\beta', (vec^*\Omega)')'$ be the vector of parameters of the model. The contribution of the i^{th} observation to the likelihood function is

$$p_i(\theta) = \int_{A_i} n(y_i^* - X_i\beta, \Omega) dy_i^*$$

where

$$n(\varepsilon, \Omega) = (2\pi)^{-M/2} |\Omega|^{-1/2} \exp\left[\frac{-\varepsilon'\Omega^{-1}\varepsilon}{2}\right]$$

is the multivariate normal density of an M -dimensional random vector. The log-likelihood function is

$$\ln \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln p_i(\theta)$$

and the MLE $\hat{\theta}$ solves the score equations

$$\frac{1}{n} \sum_{i=1}^{n} g_i(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{D_\theta p_i(\hat{\theta})}{p_i(\hat{\theta})} \equiv 0.$$

- The problem is that evaluation of $\mathcal{L}_i(\theta)$ and its derivative w.r.t. θ by standard methods of

numeric integration such as quadrature is computationally infeasible when m (the dimension of y) is higher than 3 or 4 (as long as there are no restrictions on Ω).

- The mapping $\tau(y^*)$ has not been made specific so far. This setup is quite general: for different choices of $\tau(y^*)$ it nests the case of dynamic binary discrete choice models as well as the case of multinomial discrete choice (the choice of one out of a finite set of alternatives).

  - Multinomial discrete choice is illustrated by a (very simple) job search model. We have cross sectional data on individuals' matching to a set of m jobs that are available (one of which is unemployment). The utility of alternative j is

    $$u_j = X_j\beta + \varepsilon_j$$

    Utilities of jobs, stacked in the vector u_i are not observed. Rather, we observe the vector formed of elements

    $$y_j = 1\left[u_j > u_k, \forall k \in m, k \neq j\right]$$

    Only one of these elements is different than zero.

  - Dynamic discrete choice is illustrated by repeated choices over time between two alternatives. Let alternative j have utility

    $$\begin{aligned} u_{jt} &= W_{jt}\beta - \varepsilon_{jt}, \\ j &\in \{1,2\} \\ t &\in \{1,2,...,m\} \end{aligned}$$

Then

$$\begin{aligned} y^* &= u_2 - u_1 \\ &= (W_2 - W_1)\beta + \varepsilon_2 - \varepsilon_1 \\ &\equiv X\beta + \varepsilon \end{aligned}$$

Now the mapping is (element-by-element)

$$y = 1\left[y^* > 0\right],$$

that is $y_{it} = 1$ if individual i chooses the second alternative in period t, zero otherwise.

Example: Marginalization of latent variables

Economic data often presents substantial heterogeneity that may be difficult to model. A possibility is to introduce latent random variables. This can cause the problem that there may be no known closed form for the distribution of observable variables after marginalizing out the unobservable latent variables. For example, count data (that takes values $0, 1, 2, 3, ...$) is often modeled using the Poisson distribution

$$\Pr(y = i) = \frac{\exp(-\lambda)\lambda^i}{i!}$$

The mean and variance of the Poisson distribution are both equal to λ :

$$\mathcal{E}(y) = V(y) = \lambda.$$

Often, one parameterizes the conditional mean as

$$\lambda_i = \exp(X_i\beta).$$

This ensures that the mean is positive (as it must be). Estimation by ML is straightforward.

Often, count data exhibits "overdispersion" which simply means that

$$V(y) > \mathcal{E}(y).$$

If this is the case, a solution is to use the negative binomial distribution rather than the Poisson. An alternative is to introduce a latent variable that reflects heterogeneity into the specification:

$$\lambda_i = \exp(X_i\beta + \eta_i)$$

where η_i has some specified density with support S (this density may depend on additional parameters). Let $d\mu(\eta_i)$ be the density of η_i. In some cases, the marginal density of y

$$\Pr(y = y_i|X_i) = \int_S \frac{\exp\left[-\exp(X_i\beta + \eta_i)\right]\left[\exp(X_i\beta + \eta_i)\right]^{y_i}}{y_i!} d\mu(\eta_i)$$

will have a closed-form solution (one can derive the negative binomial distribution in this way if η has an exponential distribution - see equation 18.1), but often this will not be possible. In this case, simulation is a means of calculating $\Pr(y = i|X_i)$, which is then used to do ML estimation. This would be an example of the Simulated Maximum Likelihood (SML) estimation.

- In this case, since there is only one latent variable, quadrature is probably a better choice.

However, a more flexible model with heterogeneity would allow all parameters (not just the constant) to vary. For example

$$\Pr(y = y_i) = \int_S \frac{\exp\left[-\exp(X_i\beta_i)\right]\left[\exp(X_i\beta_i)\right]^{y_i}}{y_i!} d\mu(\beta_i)$$

entails a $K = \dim \beta_i$-dimensional integral, which will not be evaluable by quadrature when K gets large.

Estimation of models specified in terms of stochastic differential equations

It is often convenient to formulate models in terms of continuous time using differential equations. An example was the jump-diffusion model discussed in Section 15.4. A realistic model should account for exogenous shocks to the system, which can be done by assuming a random component. This leads to a model that is expressed as a system of stochastic differential equations. Consider the process

$$dy_t = g(\theta, y_t)dt + h(\theta, y_t)dW_t$$

which is assumed to be stationary. $\{W_t\}$ is a standard Brownian motion (Weiner process), such that

$$W(T) = \int_0^T dW_t \sim N(0, T)$$

Brownian motion is a continuous-time stochastic process such that

- $W(0) = 0$
- $[W(s) - W(t)] \sim N(0, s - t)$

- $[W(s) - W(t)]$ and $[W(j) - W(k)]$ are independent for $s > t > j > k$. That is, non-overlapping segments are independent.

One can think of Brownian motion the accumulation over time of independent normally distributed shocks, each with an infinitesimal variance.

- The function $g(\theta, y_t)$ is the deterministic part.
- $h(\theta, y_t)$ determines the variance of the shocks.

To estimate a model of this sort, we typically have data that are assumed to be observations of y_t in discrete points y_1, y_2, ...y_T. That is, though y_t is a continuous process it is observed in discrete time.

To perform inference on θ, direct ML or GMM estimation is not usually feasible, because one cannot, in general, deduce the transition density $f(y_t|y_{t-1}, \theta)$. This density is necessary to evaluate the likelihood function or to evaluate moment conditions (which are based upon expectations with respect to this density).

- A typical solution is to "discretize" the model, by which we mean to find a discrete time approximation to the model. The discretized version of the model is

$$\begin{aligned} y_t - y_{t-\Delta} &= \tilde{g}(\phi, y_{t-1})\Delta + \sqrt{\Delta}\tilde{h}(\phi, y_{t-1})\varepsilon_t \\ \varepsilon_t &\sim N(0,1) \end{aligned}$$

  where Δ is a discrete time interval

  - I have changed the parameter from θ to ϕ to emphasize that this is an approximation, which will be more or less good. As such "ML" estimation of ϕ is actually quasi-maximum

likelihood estimation. When actual data is available on a daily, say, basis, then you could set $\Delta = 1$, and use the discretized model to do QML estimation. However, the time interval Δ may be too large to give an accurate approximation to the model, and if this is the case, the QML estimator could suffer from a large bias for estimation of the original parameter, θ.

  - Nevertheless, the approximation shouldn't be too bad, especially if Δ is small. For example, one could simulate the model at a frequency of 1 minute, saving every 1440th point on the path ($60 \times 24 = 1440$), which would give a good approximation of the evolution of the daily observations. The "Euler approximation" method for simulating such models is based upon this fact. Simulation-based inference allows for direct inference on θ, which is what we would like to do.

- The important point about these three examples is that computational difficulties prevent direct application of ML, GMM, etc. Nevertheless the model is fully specified in probabilistic terms up to a parameter vector. This means that the model is simulable, conditional on the parameter vector.

22.2 Simulated maximum likelihood (SML)

For simplicity, consider cross-sectional data. An ML estimator solves

$$\hat{\theta}_{ML} = \arg\max s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \ln p(y_t | X_t, \theta)$$

where $p(y_t|X_t,\theta)$ is the density function of the t^{th} observation. When $p(y_t|X_t,\theta)$ does not have a known closed form, $\hat{\theta}_{ML}$ is an infeasible estimator. However, it may be possible to define a random function such that

$$\mathcal{E}_\nu f(\nu, y_t, X_t, \theta) = p(y_t|X_t,\theta)$$

where the density of ν is known. If this is the case, the simulator

$$\tilde{p}\left(y_t, X_t, \theta\right) = \frac{1}{H}\sum_{s=1}^{H} f(\nu_{ts}, y_t, X_t, \theta)$$

is unbiased for $p(y_t|X_t,\theta)$.

- The SML simply substitutes $\tilde{p}\left(y_t, X_t, \theta\right)$ in place of $p(y_t|X_t,\theta)$ in the log-likelihood function, that is

$$\hat{\theta}_{SML} = \arg\max s_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln \tilde{p}\left(y_t, X_t, \theta\right)$$

Example: multinomial probit

Recall that the utility of alternative j is

$$u_j = X_j\beta + \varepsilon_j$$

and the vector y is formed of elements

$$y_j = 1\left[u_j > u_k, k \in m, k \neq j\right]$$

The problem is that $\Pr(y_j = 1|\theta)$ can't be calculated when m is larger than 4 or 5. However, it is easy to simulate this probability.

- Draw $\tilde{\varepsilon}_i$ from the distribution $N(0, \Omega)$
- Calculate $\tilde{u}_i = X_i\beta + \tilde{\varepsilon}_i$ (where X_i is the matrix formed by stacking the X_{ij})
- Define $\tilde{y}_{ij} = 1\left[u_{ij} > u_{ik}, \forall k \in m, k \neq j\right]$
- Repeat this H times and define
$$\tilde{\pi}_{ij} = \frac{\sum_{h=1}^{H} \tilde{y}_{ijh}}{H}$$
- Define $\tilde{\pi}_i$ as the m-vector formed of the $\tilde{\pi}_{ij}$. Each element of $\tilde{\pi}_i$ is between 0 and 1, and the elements sum to one.
- Now $\tilde{p}\left(y_i, X_i, \theta\right) = y_i'\tilde{\pi}_i$
- The SML multinomial probit log-likelihood function is
$$\ln \mathcal{L}(\beta, \Omega) = \frac{1}{n}\sum_{i=1}^{n} y_i' \ln \tilde{p}\left(y_i, X_i, \theta\right)$$

This is to be maximized w.r.t. β and Ω.

Notes:

- The H draws of $\tilde{\varepsilon}_i$ are draw *only once* and are used repeatedly during the iterations used to find $\hat{\beta}$ and $\hat{\Omega}$. The draws are different for each i. If the $\tilde{\varepsilon}_i$ are re-drawn at every iteration the estimator will not converge.

- The log-likelihood function with this simulator is a discontinuous function of β and Ω. This does not cause problems from a theoretical point of view since it can be shown that $\ln \mathcal{L}(\beta, \Omega)$ is stochastically equicontinuous. However, it does cause problems if one attempts to use a gradient-based optimization method such as Newton-Raphson.

- It may be the case, particularly if few simulations, H, are used, that some elements of $\tilde{\pi}_i$ are zero. If the corresponding element of y_i is equal to 1, there will be a log(0) problem.

- Solutions to discontinuity:

  - 1) use an estimation method that doesn't require a continuous and differentiable objective function, for example, simulated annealing. This is computationally costly.
  - 2) Smooth the simulated probabilities so that they are continuous functions of the parameters. For example, apply a kernel transformation such as

$$\tilde{y}_{ij} = \Phi\left(A \times \left[u_{ij} - \max_{k=1}^{m} u_{ik}\right]\right) + .5 \times 1\left[u_{ij} = \max_{k=1}^{m} u_{ik}\right]$$

    where A is a large positive number. This approximates a step function such that $\tilde{y}_{ij}$ is very close to zero if u_{ij} is not the maximum, and $\tilde{y}_{ij}$ is very close to 1 if u_{ij} is the maximum. This makes $\tilde{y}_{ij}$ a continuous function of β and Ω, so that $\tilde{p}_{ij}$ and therefore $\ln \mathcal{L}(\beta, \Omega)$ will be continuous and differentiable. Consistency requires that $A(n) \xrightarrow{p} \infty$, so that the approximation to a step function becomes arbitrarily close as the sample size increases. There are alternative methods (e.g., Gibbs sampling) that may work better, but this is too technical to discuss here.

- To solve to log(0) problem, one possibility is to search the web for the slog function. Also, increase H if this is a serious problem.

Properties

The properties of the SML estimator depend on how H is set. The following is taken from Lee (1995) "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models," *Econometric Theory,* **11,** pp. 437-83.

Theorem 69. *[Lee] 1) if* $\lim_{n\to\infty} n^{1/2}/H = 0$, *then*

$$\sqrt{n}\left(\hat{\theta}_{SML} - \theta^0\right) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\theta^0))$$

2) if $\lim_{n\to\infty} n^{1/2}/H = \lambda$, λ *a finite constant, then*

$$\sqrt{n}\left(\hat{\theta}_{SML} - \theta^0\right) \xrightarrow{d} N(B, \mathcal{I}^{-1}(\theta^0))$$

where B is a finite vector of constants.

- This means that the SML estimator is asymptotically biased if H doesn't grow faster than $n^{1/2}$.
- The varcov is the typical inverse of the information matrix, so that as long as H grows fast enough the estimator is consistent and fully asymptotically efficient.

22.3 Method of simulated moments (MSM)

Suppose we have a DGP$(y|x,\theta)$ which is simulable given θ, but is such that the density of y is not calculable.

Once could, in principle, base a GMM estimator upon the moment conditions

$$m_t(\theta) = [K(y_t, x_t) - k(x_t, \theta)]\, z_t$$

where

$$k(x_t, \theta) = \int K(y_t, x_t) p(y|x_t, \theta) dy,$$

z_t is a vector of instruments in the information set and $p(y|x_t, \theta)$ is the density of y conditional on x_t. The problem is that this density is not available.

- However $k(x_t, \theta)$ is readily simulated using

$$\tilde{k}\,(x_t, \theta) = \frac{1}{H} \sum_{h=1}^{H} K(\tilde{y}_t^h, x_t)$$

- By the law of large numbers, $\tilde{k}\,(x_t, \theta) \xrightarrow{a.s.} k\,(x_t, \theta)$, as $H \to \infty$, which provides a clear intuitive basis for the estimator, though in fact we obtain consistency even for H finite, since a law of large numbers is also operating across the n observations of real data, so errors introduced by simulation cancel themselves out.

- This allows us to form the moment conditions

$$\widetilde{m_t}(\theta) = \left[K(y_t, x_t) - \tilde{k}\,(x_t, \theta)\right] z_t \tag{22.2}$$

where z_t is drawn from the information set. As before, form

$$\begin{aligned} \widetilde{m}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{m_t}(\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[K(y_t,x_t) - \frac{1}{H}\sum_{h=1}^{H}k(\tilde{y}_t^h,x_t)\right]z_t \end{aligned} \tag{22.3}$$

with which we form the GMM criterion and estimate as usual. Note that the unbiased simulator $k(\tilde{y}_t^h, x_t)$ appears linearly within the sums.

Properties

Suppose that the optimal weighting matrix is used. McFadden (ref. above) and Pakes and Pollard (refs. above) show that the asymptotic distribution of the MSM estimator is very similar to that of the infeasible GMM estimator. In particular, assuming that the optimal weighting matrix is used, and for H finite,

$$\sqrt{n}\left(\hat{\theta}_{MSM} - \theta^0\right) \xrightarrow{d} N\left[0, \left(1+\frac{1}{H}\right)\left(D_\infty \Omega^{-1} D_\infty'\right)^{-1}\right] \tag{22.4}$$

where $\left(D_\infty \Omega^{-1} D_\infty'\right)^{-1}$ is the asymptotic variance of the infeasible GMM estimator.

- That is, the asymptotic variance is inflated by a factor $1 + 1/H$. For this reason the MSM estimator is not fully asymptotically efficient relative to the infeasible GMM estimator, for H finite, but the efficiency loss is small and controllable, by setting H reasonably large.
- The estimator is asymptotically unbiased even for $H = 1$. This is an advantage relative to SML.
- If one doesn't use the optimal weighting matrix, the asymptotic varcov is just the ordinary GMM

varcov, inflated by $1 + 1/H$.

- The above presentation is in terms of a specific moment condition based upon the conditional mean. The MSM can be applied to moment conditions of other forms, too.

  - A leading example is Indirect Inference, where we set $m_n(\theta) = \hat{\phi}\text{-}\frac{1}{S}\sum\tilde{\phi}^s(\theta)$, and then we just do ordinary GMM. Here, $\hat{\phi}$ is an extremum estimator corresponding to some auxiliary model. The $\tilde{\phi}^s(\theta)$ are the same extremum estimator, applied to simulated data generated from the model. The logic is that $\hat{\phi}$ will converge to a pseudo-true value, and $\tilde{\phi}^s(\theta)$ will converge to another pseudo-true value, depending on the value of θ that generated the data. When $\theta = \theta^0$, the two pseudo-true values will be the same. Trying to make the average of the simulated estimators as close as possible to the estimator generated by the real data will cause the MSM estimator to be consistent, given identification.
  - For such an estimator to have good efficiency, we need the auxiliary model to fit well: it should pick up the relevant features of the data.
  - a drawback of the II estimator is that the auxiliary model must be estimated many times. This is not a problem if it's a simple linear model, but it could be a problem if it's more complicated. For efficiency, we need a good fit, and a simple linear model may not provide this. The EMM estimator discussed below is asymptotically equivalent to II, and it requires the auxiliary model to be estimated only once.

Comments

Why is SML inconsistent if H is finite, while MSM is? The reason is that SML is based upon an average of **logarithms** of an unbiased simulator (the densities of the observations). To use the multinomial probit model as an example, the log-likelihood function is

$$\ln \mathcal{L}(\beta, \Omega) = \frac{1}{n} \sum_{i=1}^{n} y_i' \ln p_i(\beta, \Omega)$$

The SML version is

$$\ln \mathcal{L}(\beta, \Omega) = \frac{1}{n} \sum_{i=1}^{n} y_i' \ln \tilde{p}_i(\beta, \Omega)$$

The problem is that

$$E \ln(\tilde{p}_i(\beta, \Omega)) \neq \ln(\mathcal{E}\tilde{p}_i(\beta, \Omega))$$

in spite of the fact that

$$\mathcal{E}\tilde{p}_i(\beta, \Omega) = p_i(\beta, \Omega)$$

due to the fact that $\ln(\cdot)$ is a nonlinear transformation. The only way for the two to be equal (in the limit) is if H tends to infinite so that $\tilde{p}(\cdot)$ tends to $p(\cdot)$.

The reason that MSM does not suffer from this problem is that in this case the unbiased simulator appears *linearly* within every sum of terms, and it appears within a sum over n (see equation [22.3]). Therefore the SLLN applies to cancel out simulation errors, from which we get consistency. That is,

using simple notation for the random sampling case, the moment conditions

$$\tilde{m}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[K(y_t, x_t) - \frac{1}{H}\sum_{h=1}^{H} k(\tilde{y}_t^h, x_t)\right] z_t \tag{22.5}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[k(x_t, \theta^0) + \varepsilon_t - \frac{1}{H}\sum_{h=1}^{H}\left[k(x_t, \theta) + \tilde{\varepsilon}_{ht}\right]\right] z_t \tag{22.6}$$

converge almost surely to

$$\tilde{m}_\infty(\theta) = \int \left[k(x, \theta^0) - k(x, \theta)\right] z(x) d\mu(x).$$

(note: z_t is assume to be made up of functions of x_t). The objective function converges to

$$s_\infty(\theta) = \tilde{m}_\infty(\theta)' \Omega_\infty^{-1} \tilde{m}_\infty(\theta)$$

which obviously has a minimum at θ^0, henceforth consistency.

- If you look at equation 22.6 a bit, you will see why the variance inflation factor is $(1 + \frac{1}{H})$.

22.4 Efficient method of moments (EMM)

The choice of which moments upon which to base a GMM estimator can have very pronounced effects upon the efficiency of the estimator.

- A poor choice of moment conditions may lead to very inefficient estimators, and can even cause identification problems (as we've seen with the GMM problem set).
- The drawback of the above approach MSM is that the moment conditions used in estimation are

selected arbitrarily. The asymptotic efficiency of the estimator may be low.

- The asymptotically optimal choice of moments would be the score vector of the likelihood function,

$$m_t(\theta) = D_\theta \ln p_t(\theta \mid I_t)$$

  As before, this choice is unavailable.

The efficient method of moments (EMM) (see Gallant and Tauchen (1996), "Which Moments to Match?", ECONOMETRIC THEORY, Vol. 12, 1996, pages 657-681) seeks to provide moment conditions that closely mimic the score vector. If the approximation is very good, the resulting estimator will be very nearly fully efficient.

The DGP is characterized by random sampling from the density

$$p(y_t|x_t, \theta^0) \equiv p_t(\theta^0)$$

We can define an auxiliary model, called the "score generator", which simply provides a (misspecified) parametric density

$$f(y|x_t, \lambda) \equiv f_t(\lambda)$$

- This density is known up to a parameter λ. We assume that this density function *is* calculable. Therefore quasi-ML estimation is possible. Specifically,

$$\hat{\lambda} = \arg\max_{\Lambda} s_n(\lambda) = \frac{1}{n}\sum_{t=1}^{n} \ln f_t(\lambda).$$

- After determining $\hat{\lambda}$ we can calculate the score functions $D_\lambda \ln f(y_t|x_t, \hat{\lambda})$.

- The important point is that even if the density is misspecified, there is a pseudo-true λ^0 for which the true expectation, taken with respect to the true but unknown density of y, $p(y|x_t, \theta^0)$, and then marginalized over x is zero:

$$\exists \lambda^0 : \mathcal{E}_X \mathcal{E}_{Y|X} \left[D_\lambda \ln f(y|x, \lambda^0) \right] = \int_X \int_{Y|X} D_\lambda \ln f(y|x, \lambda^0) p(y|x, \theta^0) dy d\mu(x) = 0$$

- We have seen in the section on QML that $\hat{\lambda} \xrightarrow{p} \lambda^0$; this suggests using the moment conditions

$$m_n(\theta, \hat{\lambda}) = \frac{1}{n} \sum_{t=1}^{n} \int D_\lambda \ln f_t(\hat{\lambda}) p_t(\theta) dy \tag{22.7}$$

- These moment conditions are not calculable, since $p_t(\theta)$ is not available, but they are simulable using

$$\widetilde{m_n}(\theta, \hat{\lambda}) = \frac{1}{n} \sum_{t=1}^{n} \frac{1}{H} \sum_{h=1}^{H} D_\lambda \ln f(\tilde{y}_t^h | x_t, \hat{\lambda})$$

  where $\tilde{y}_t^h$ is a draw from $DGP(\theta)$, holding x_t fixed. By the LLN and the fact that $\hat{\lambda}$ converges to λ^0,

$$\widetilde{m}_\infty(\theta^0, \lambda^0) = 0.$$

  This is not the case for other values of θ, assuming that λ^0 is identified.

- The advantage of this procedure is that if $f(y_t|x_t, \lambda)$ closely approximates $p(y|x_t, \theta)$, then $\widetilde{m}_n(\theta, \hat{\lambda})$ will closely approximate the optimal moment conditions which characterize maximum likelihood estimation, which is fully efficient.

- If one has prior information that a certain density approximates the data well, it would be a

good choice for $f(\cdot)$.

- If one has no density in mind, there exist good ways of approximating unknown distributions parametrically: Philips' ERA's (*Econometrica*, 1983) and Gallant and Nychka's (*Econometrica, 1987)* SNP density estimator which we saw before. Since the SNP density is consistent, the efficiency of the indirect estimator is the same as the infeasible ML estimator.

Optimal weighting matrix

I will present the theory for H finite, and possibly small. This is done because it is sometimes impractical to estimate with H very large. Gallant and Tauchen give the theory for the case of H so large that it may be treated as infinite (the difference being irrelevant given the numerical precision of a computer). The theory for the case of H infinite follows directly from the results presented here.

The moment condition $\widetilde{m}(\theta, \hat{\lambda})$ depends on the pseudo-ML estimate $\hat{\lambda}$. We can apply Theorem 31 to conclude that

$$\sqrt{n}\left(\hat{\lambda} - \lambda^0\right) \xrightarrow{d} N\left[0, \mathcal{J}(\lambda^0)^{-1}\mathcal{I}(\lambda^0)\mathcal{J}(\lambda^0)^{-1}\right] \tag{22.8}$$

If the density $f(y_t|x_t, \hat{\lambda})$ were in fact the true density $p(y|x_t, \theta)$, then $\hat{\lambda}$ would be the maximum likelihood estimator, and $\mathcal{J}(\lambda^0)^{-1}\mathcal{I}(\lambda^0)$ would be an identity matrix, due to the information matrix equality. However, in the present case we assume that $f(y_t|x_t, \hat{\lambda})$ is only an approximation to $p(y|x_t, \theta)$, so there is no cancellation.

Recall that $\mathcal{J}(\lambda^0) \equiv p\lim\left(\frac{\partial^2}{\partial\lambda\partial\lambda'}s_n(\lambda^0)\right)$. Comparing the definition of $s_n(\lambda)$ with the definition of the moment condition in Equation 22.7, we see that

$$\mathcal{J}(\lambda^0) = D_{\lambda'}m(\theta^0, \lambda^0).$$

As in Theorem 31,

$$\mathcal{I}(\lambda^0) = \lim_{n\to\infty} \mathcal{E}\left[n \left.\frac{\partial s_n(\lambda)}{\partial \lambda}\right|_{\lambda^0} \left.\frac{\partial s_n(\lambda)}{\partial \lambda'}\right|_{\lambda^0} \right].$$

In this case, this is simply the asymptotic variance covariance matrix of the moment conditions, Ω. Now take a first order Taylor's series approximation to $\sqrt{n}m_n(\theta^0, \hat{\lambda})$ about λ^0 :

$$\sqrt{n}\tilde{m}_n(\theta^0, \hat{\lambda}) = \sqrt{n}\tilde{m}_n(\theta^0, \lambda^0) + \sqrt{n}D_{\lambda'}\tilde{m}(\theta^0, \lambda^0)\left(\hat{\lambda} - \lambda^0\right) + o_p(1)$$

First consider $\sqrt{n}\tilde{m}_n(\theta^0, \lambda^0)$. It is straightforward but somewhat tedious to show that the asymptotic variance of this term is $\frac{1}{H}I_\infty(\lambda^0)$.

Next consider the second term $\sqrt{n}D_{\lambda'}\tilde{m}(\theta^0, \lambda^0)\left(\hat{\lambda} - \lambda^0\right)$. Note that $D_{\lambda'}\tilde{m}_n(\theta^0, \lambda^0) \xrightarrow{a.s.} \mathcal{J}(\lambda^0)$, so we have

$$\sqrt{n}D_{\lambda'}\tilde{m}(\theta^0, \lambda^0)\left(\hat{\lambda} - \lambda^0\right) = \sqrt{n}\mathcal{J}(\lambda^0)\left(\hat{\lambda} - \lambda^0\right), a.s.$$

But noting equation 22.8

$$\sqrt{n}\mathcal{J}(\lambda^0)\left(\hat{\lambda} - \lambda^0\right) \overset{a}{\sim} N\left[0, \mathcal{I}(\lambda^0)\right]$$

Now, combining the results for the first and second terms,

$$\sqrt{n}\tilde{m}_n(\theta^0, \hat{\lambda}) \overset{a}{\sim} N\left[0, \left(1 + \frac{1}{H}\right)\mathcal{I}(\lambda^0)\right]$$

Suppose that $\widehat{\mathcal{I}(\lambda^0)}$ is a consistent estimator of the asymptotic variance-covariance matrix of the moment conditions. This may be complicated if the score generator is a poor approximator, since the individual score contributions may not have mean zero in this case (see the section on QML) . Even if this is the case, the individuals means can be calculated by simulation, so it is always possible to

consistently estimate $\mathcal{I}(\lambda^0)$ when the model is simulable. On the other hand, if the score generator is taken to be correctly specified, the ordinary estimator of the information matrix is consistent. Combining this with the result on the efficient GMM weighting matrix in Theorem 47, we see that defining $\hat{\theta}$ as

$$\hat{\theta} = \arg\min_{\Theta} m_n(\theta, \hat{\lambda})' \left[\left(1 + \frac{1}{H}\right) \widehat{\mathcal{I}(\lambda^0)} \right]^{-1} m_n(\theta, \hat{\lambda})$$

is the GMM estimator with the efficient choice of weighting matrix.

- If one has used the Gallant-Nychka ML estimator as the auxiliary model, the appropriate weighting matrix is simply the information matrix of the auxiliary model, since the scores are uncorrelated. (e.g., it really is ML estimation asymptotically, since the score generator can approximate the unknown density arbitrarily well).

Asymptotic distribution

Since we use the optimal weighting matrix, the asymptotic distribution is as in Equation 14.4, so we have (using the result in Equation 22.8):

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \left(D_\infty \left[\left(1 + \frac{1}{H}\right) \mathcal{I}(\lambda^0)\right]^{-1} D_\infty'\right)^{-1}\right],$$

where

$$D_\infty = \lim_{n\to\infty} \mathcal{E}\left[D_\theta m_n'(\theta^0, \lambda^0)\right].$$

This can be consistently estimated using

$$\hat{D} = D_\theta m_n'(\hat{\theta}, \hat{\lambda})$$

Diagnostic testing

The fact that

$$\sqrt{n} m_n(\theta^0, \hat{\lambda}) \overset{a}{\sim} N\left[0, \left(1 + \frac{1}{H}\right) \mathcal{I}(\lambda^0)\right]$$

implies that

$$n m_n(\hat{\theta}, \hat{\lambda})' \left[\left(1 + \frac{1}{H}\right) \mathcal{I}(\hat{\lambda})\right]^{-1} m_n(\hat{\theta}, \hat{\lambda}) \overset{a}{\sim} \chi^2(q)$$

where q is $\dim(\lambda) - \dim(\theta)$, since without $\dim(\theta)$ moment conditions the model is not identified, so testing is impossible. One test of the model is simply based on this statistic: if it exceeds the $\chi^2(q)$ critical point, something may be wrong (the small sample performance of this sort of test would be a topic worth investigating).

- Information about what is wrong can be gotten from the pseudo-t-statistics:

$$\left(\text{diag}\left[\left(1 + \frac{1}{H}\right) \mathcal{I}(\hat{\lambda})\right]^{1/2}\right)^{-1} \sqrt{n} m_n(\hat{\theta}, \hat{\lambda})$$

  can be used to test which moments are not well modeled. Since these moments are related to parameters of the score generator, which are usually related to certain features of the model, this information can be used to revise the model. These aren't actually distributed as $N(0, 1)$, since $\sqrt{n} m_n(\theta^0, \hat{\lambda})$ and $\sqrt{n} m_n(\hat{\theta}, \hat{\lambda})$ have different distributions (that of $\sqrt{n} m_n(\hat{\theta}, \hat{\lambda})$ is somewhat

more complicated). It can be shown that the pseudo-t statistics are biased toward nonrejection. See Gourieroux *et. al.* or Gallant and Long, 1995, for more details.

22.5 Indirect likelihood inference

This method is something I've been working on for the last few years with Dennis Kristensen. The main reference is Creel and Kristensen, 2013. The method is related to "Approximate Bayesian Computing". Our paper adds some formal results that relates the idea to GMM, and which gives the first applications in economics. It is a very useful method, in my opinion. We have used it to estimate complicated models such as DSGE models and continuous time jump diffusions, with good success. It combines simulation based estimation, nonparametric fitting, and Bayesian methods. The following is a brief description, and an example.

Suppose we have a fully specified model indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^k$. Given a sample $Y_n = (y_1, ..., y_n)$ generated at the unknown true parameter value θ_0, the generalized method of moments estimator is based on a vector of statistics $Z_n = Z_n(Y_n)$, that lead to moment conditions $m_n(\theta) = Z_n - E_\theta(Z_n)$, where E_θ indicates expectations under the model. In a similar vein, Creel and Kristensen (CK13) propose a Bayesian indirect likelihood (BIL) estimator

$$\hat{\theta}_{BIL} = E(\theta|Z_n) = \int_\Theta \theta f_n(\theta|Z_n)\, d\theta, \tag{22.9}$$

where, for some prior density $\pi(\theta)$ on the parameter space Θ, $f_n(\theta|Z_n)$ is the posterior distribution given by

$$f_n(\theta|Z_n) = \frac{f_n(Z_n, \theta)}{f_n(Z_n)} = \frac{f_n(Z_n|\theta)\pi(\theta)}{\int_\Theta f_n(Z_n|\theta)\pi(\theta)\, d\theta}.$$

This is very much like the widely used Bayesian posterior mean, except that the likelihood is formulated in terms of the density of the statistic, $f_n(Z_n|\theta)$, rather than the full sample. Advantages of the BIL estimator over GMM are the avoidance of optimization, avoidance of the need to compute the efficient weight matrix, and higher order efficiency relative to the GMM estimator that uses the optimal weight matrix (CK13).

Computation of the BIL estimator requires knowledge of $f_n(Z_n|\theta)$, which is normally not known. Just as the simulated method of moments may be required when GMM is not feasible, simulation and nonparametric regression may be used to compute a simulated BIL (SBIL) estimator. The method explored in this paper is $k-$nearest neighbors (KNN) nonparametric regression. This is implemented as follows: Make i.i.d. draws θ^s, $s = 1, ..., S$, from the pseudo-prior density $\pi(\theta)$, for each draw generate a sample $Y_n(\theta^s)$ from the model at this parameter value, and then compute the corresponding statistic $Z_n^s = Z(Y_n(\theta^s))$, $s = 1, ..., S$. Let $\mathcal{Z}_S = \{Z_n^s\}$, $s = 1, 2, ..., S$ be the set of S draws of the statistic. Given the i.i.d. draws (θ^s, Z_n^s), $s = 1, ...S$, we can obtain the SBIL estimator using

$$\hat{\theta}_{SBIL} = \hat{E}_S\left[\theta|Z_n\right] = \frac{\sum_{s=1}^S \theta^s K_h\left(Z_n^s - Z_n\right)}{\sum_{s=1}^S K_h\left(Z_n^s - Z_n\right)}. \tag{22.10}$$

where $K_h(z) \geq 0$ is a kernel function that depends on a bandwidth parameter h. This is most obviously a kernel regression estimator, but it can also be a nearest neighbor estimator if the bandwidth is adaptive, which is what we do.

In the literature on kernel regression, it is well-known that the specific kernel function chosen is of less importance than is choosing the bandwidth appropriately, given the chosen kernel (REF). For this reason, we use a truncated Gaussian kernel, and focus on choosing the bandwidth well. The specific

kernel function we use is

$$K_h\left(Z_n^s - Z_n\right) = \begin{array}{ll} \phi\left(\frac{a\left\|\Sigma^{-1/2}\left(Z_n^s - Z_n\right)\right\|}{h}\right) & \text{if } \left\|\Sigma^{-1/2}\left(Z_n^s - Z_n\right)\right\| \leq h \\ 0 & \text{if } \left\|\Sigma^{-1/2}\left(Z_n^s - Z_n\right)\right\| > h \end{array}$$

where $\phi(\cdot)$ is the multivariate standard normal density function. The matrix Σ is the diagonal matrix containing the sample variances of the S replications of Z_n^s. This matrix plays the important role of putting the elements of the auxiliary statistic vector on the same scale, so that statistics with larger variances do not dominate the distance measure. The bandwidth h is adaptive, it is the k^{th} order statistic of the S distances

$$d_s = \left\|\Sigma^{-1/2}\left(Z_n^s - Z_n\right)\right\|, \; s = 1, 2, ..., S. \tag{22.11}$$

Define d_s^k as the kth order statistic of these distances, so the kernel bandwidth is $h = d_s^k$. Thus, this adaptive kernel regression estimator is a KNN estimator, where only the closest k neighbors affect the fit. The scalar tuning parameter a influences how rapidly the weights decline as the distance between the simulated statistic and the observed statistic increases. We set $a = 2$. With these choices, the SBIL estimator is a weighted average of the θ^s such that the corresponding simulated (scaled) Z_n^s is among the k nearest neighbors to Z_n, and the weights are declining as the distance from Z_n increases. The problem of choosing the bandwidth becomes one of choosing the number of neighbors to use.

In the i.i.d. sample context that applies to the simulated pairs (θ^s, Z_n^s), the KNN estimator is consistent for the true posterior mean $E(\theta|Z_n)$ as S increases, as long as the chosen number of neighbors, k, grows slowly with S (Li and Racine, 2007, Ch. 14). Because S, the number of simulations can be made as large as needed, consistency of KNN regression means that the SBIL estimator can be made arbitrarily close to the infeasible BIL estimator. Nevertheless, methods for choosing k as a function of

S to obtain good performance of the SBIL estimator without requiring the number of simulations to be extremely large are desirable, to limit the computational demand. Another factor that obviously affects the performance of the SBIL estimator is the choice of the vector of statistics that form Z_n. These issues have been addressed, but they are beyond the score of these notes.

This discussion makes clear the nature of the estimator. The infeasible BIL estimator is a posterior mean, conditional on a statistic, rather than the full sample. It turns out that the BIL estimator is first order asymptotically equivalent to the optimal GMM estimator that uses the same statistic (CK13). Thus, the relationship between the BIL estimator and the ordinary posterior mean $E(\theta|Y_n)$ based on the full sample is essentially the same as the relationship between the GMM estimator and the maximum likelihood estimator: the first is in general not fully efficient, and the issue of the choice of statistics arises. The relationship between the SBIL estimator and the infeasible BIL estimator is like that between an ordinary Bayesian posterior mean computed using Markov chain Monte Carlo or some other computational technique, and the desired true posterior mean: the first is a numeric approximation of the second, which can be made as precise as needed by means of additional computional resources. Our argument for using the SBIL estimator is one of convenience and performance. In terms of convenience, the SBIL estimator can be reliably computed using simple means that are very amenable to parallel computing techniques. Regarding performance, we show by example that Z_n can be found which lead to good estimation results, even for complicated models (e.g., DSGE models) that traditionally have required sophisticated estimation techniques.

A Simple DSGE Model

CK13 shows that SBIL estimation is tractible and gives reliable results for estimating the parameters of a simple DSGE model. SBIL estimation can be done quickly and easily enough so that it is possible

to show its good performance via Monte Carlo, and example software has been provided that allows confirmation of these results in little time. Here we use the same model to illustrate the methods we propose. The model is as follows: A single good can be consumed or used for investment, and a single competitive firm maximizes profits. The variables are: y output; c consumption; k capital; i investment, n labor; w real wages; r return to capital. The household maximizes expected discounted utility

$$E_t \sum_{s=0}^{\infty} \beta^s \left(\frac{c_{t+s}^{1-\gamma}}{1-\gamma} + (1 - n_{t+s})\eta_t \psi \right)$$

subject to the budget constraint $c_t + i_t = r_t k_t + w_t n_t$ and the accumulation of capital $k_{t+1} = i_t + (1-\delta k_t)$. There is a preference shock, η_t, that affects the desirability of leisure. The shock evolves according to $\ln \eta_t = \rho_\eta \ln \eta_{t-1} + \sigma_\eta \epsilon_t$. The competitive firm produces the good y_t using the technology $y_t = k_t^\alpha n_t^{1-\alpha} z_t$. Technology shocks z_t also follow an AR(1) process in logarithms: $\ln z_t = \rho_z \ln z_{t-1} + \sigma_z u_t$. The innovations to the preference and technology shocks, ϵ_t and u_t, are mutually independent i.i.d. standard normally distributed. The good y_t can be allocated by the consumer to consumption or investment: $y_t = c_t + i_t$. The consumer provides capital and labor to the firm, and is paid at the rates r_t and w_t, respectively. The unknown parameters are collected in $\theta = (\alpha, \beta, \delta, \gamma, \psi, \rho_z, \rho_\eta, \sigma_z, \sigma_\eta)$. In total, we have seven variables and two shocks.

In the estimation, we treat capital stock k as unobserved, while the remaing variables are observed. The true parameter values are given in Table 22.1. Following Ruge-Murcia (2012), rather than set a true value and prior for ψ and estimate this parameter directly, we instead treat steady state hours $\bar{n}$ as a parameter to estimate, along with the other parameters, excepting ψ. Following Ruge-Murcia (2012), the true value for ψ was found using the other true parameter values, along with the restriction that true steady state hours $\bar{n} = 1/3$. The true value is $\psi = \bar{c}^{-\gamma} (1-\alpha) \bar{k}^\alpha \bar{n}^{-\alpha} = 3.417$, where overbars

indicate steady state values of the variables.

The solution method is third-order perturbation using Dynare (Adjemian et al., 2011). The pseudo-prior $\pi(\theta)$ is a uniform distribution over the hypercube defined by the bounds of the parameter space, which are found in columns 3 and 4 of Table 22.1. The chosen limits cause the pseudo-prior means to be biased for the true parameter values (see column 5 of the Table), and they are intended to be broad, in comparison to the fairly strongly informative priors that are often used when estimating DSGE models (see column 6 of the Table). To generate simulations, a parameter value θ^s is drawn from the prior, then the model is solved at this parameter value, and a simulated sample is drawn. The sample size is $n = 80$, which mimics 20 years of quarterly data. With a simulated sample, we can generate a realization of the vector of statistics, $Z_n^s(\theta^s)$.

Auxiliary statistics should capture the essential features of the data. In this case, it is natural to use the parameter estimates of a vector autogressive (VAR) model as an auxiliary statistic. In this example, I use a Bayesian VAR model of order 1 for the 5 observable variables consumption, investment, hours, wages and interest rate (the capital stock is taken to be unobservable), subject to the Minnesota priors that the variables individually follow a random walk process (Doan et al., 1984, see also Section 15.2). This is done after de-meaning and standardizing the variables. This gives 40 statistics: 25 regression coefficients, and 15 estimated covariance matrix elements. Additional auxiliary statistics are the sample means and standard deviations of the variables (10 additional statistics), and certain statistics suggested by consideration of the specific DSGE model. For example, the model implies that $w = \psi\eta c^\gamma$, so $\log w = \log\psi + \gamma \log c + \log\eta$, where $\log\eta$ follows an AR(1) process. Because w and c are observable, this equation can be estimated. We use a generalized instrumental variables estimator (GIV), using the lags of the logarithms of the observable variables as instruments. The GIV estimation results give a statistic $\hat{\gamma}$ which is likely to be informative about the parameter

γ. The residuals from the GIV estimation can be used to fit the regression $\widehat{\log \eta_t} = \rho_\eta \widehat{\log \eta_{t-1}} + \epsilon_t$, which leads to statistics that should be informative for the parameters ρ_η and σ_η. In total, the vector of auxiliary statistics has 57 elements.

Table 22.1 gives the true parameter values and the limits of the uniform priors, along with information about the informativeness of the prior. Table 22.2 give results for a Monte Carlo study of the performance of the SBIL estimator. If you compare RMSE in the two tables, you'll see that the SBIL estimator achieves a considerable reduction of RMSE relative to that of the prior. Also, the SBIL estimator (with bootstrap-based bias correction) is essentially unbiased for all of the model's parameters.

A simple script that does SBIL estimation of the same DSGE model as discussed above is at DSGE by SBIL. This implements a less sophisticated verion of the estimator than was used to make the tables presented here (less careful choice of auxiliary statistics, no bias correction, etc), but it conveys the main ideas.

These results show that simulation-based estimation enables estimation of the parameters of a nonlinear structural model, with good precision.

- The GMM estimation of the portfolio model, in Section 14.15, gave quite unreliable results. That estimation method used moment conditions derived from the Euler equation, crossed with instruments chosen by an ad hoc procedure.

- We have also seen, in Section 13.8, that "maximum likelihood" estimation of such models, after linearization (which means we're not really doing ML estimation), is in many cases not very successful, at least when all parameters are to be estimated.

- There are alternative methods to actually do ML estimation, using *particle filtering*, which is a

nonlinear filter, which can replace the Kalman filter when the model is nonlinear. This method is computationally quite intensive.

- IL estimation is perfectly feasible. It is somewhat computationally intensive, but the fact that it is possible to investigate its properties by Monte Carlo illustrates that it is not excessively computationally intensive.

Table 22.1: True parameter values and bound of priors

		Prior bounds			
Parameter	True value	Lower	Upper	Prior Bias	Prior RMSE
α	0.330	0.2	0.4	-0.030	0.065
β	0.990	0.97	0.999	-0.006	0.010
δ	0.025	0.005	0.04	0.000	0.009
γ	2.000	0.0	5.0	0.500	1.527
ρ_z	0.900	0.5	0.999	-0.150	0.208
σ_z	0.010	0.001	0.1	0.041	0.049
ρ_η	0.700	0.5	0.999	0.049	0.152
σ_η	0.005	0.001	0.1	0.046	0.054
$\bar{n}$	7/24	8/24	9/24	0.000	0.024

Table 22.2: Monte Carlo results, bias corrected estimators

		Mean		Bias		RMSE	
Parameter	True value	1st round	2nd round	1st round	2nd round	1st round	2nd round
α	0.330	0.329	0.330	-0.001	-0.000	0.002	0.002
β	0.990	0.990	0.990	0.000	-0.000	0.001	0.001
δ	0.025	0.025	0.025	-0.000	-0.000	0.001	0.001
γ	2.000	2.065	2.027	0.065	0.027	0.290	0.292
ρ_z	0.900	0.891	0.899	-0.009	-0.001	0.052	0.052
σ_z	0.010	0.010	0.010	0.000	-0.000	0.002	0.002
ρ_η	0.700	0.707	0.701	0.007	0.001	0.071	0.076
σ_η	0.005	0.005	0.005	0.000	-0.000	0.001	0.002
$\bar{n}$	1/3	0.333	0.333	-0.000	-0.000	0.004	0.004

A jump-diffusion model

Another example of IL estimation is the jump-diffusion model that was introduced in section 15.4. To estimate the model, we need an auxiliary statistic that is informative about the parameters of the model. The stylized facts are:

- contemporaneous correlation between returns and volatility (leverage)
- volatility clusters: serial correlation of volatility
- fat tails
- slight autocorrelation of returns

We need statistics that can pick up all of this. Ideally, the statistics should be reasonably fast and easy to compute. This suggests using

- an EGARCH-type model for returns. Can capture all 4 stylized facts. Requires iterative maximization, but is stable and reliable.
- HAR-J type models where high frequency realized volatility measures are explained with their own lags, and with measures of jump activity that rely on jump-robust measures of volatility. This picks up volatility clusters, and the effect of jumps, which can generate fat tails. Estimation is by OLS.
- an autoregressive model for returns. This picks up a slight autocorrelation below the level that would trigger arbitrage. It can also pick up the effect of measurement error in returns. If this exists, then returns would contain an MA(1) error, which would also leads to a small correlation in returns. OLS
- ordinary descriptive statistics for returns and realized volatility measures.

IL estimation was done for

22.6 Examples

SML of a Poisson model with latent heterogeneity

We have seen (see equation **??**) that a Poisson model with latent heterogeneity that follows an exponential distribution leads to the negative binomial model. To illustrate SML, we can integrate out

the latent heterogeneity using Monte Carlo, rather than the analytical approach which leads to the negative binomial model. In actual practice, one would not want to use SML in this case, but it is a nice example since it allows us to compare SML to the actual ML estimator. The Octave function defined by PoissonLatentHet.m calculates the simulated log likelihood for a Poisson model where $\lambda = \exp x_t'\beta + \sigma\eta)$, where $\eta \sim N(0,1)$. This model is similar to the negative binomial model, except that the latent variable is normally distributed rather than gamma distributed. The Octave script EstimatePoissonLatentHet.m estimates this model using the MEPS OBDV data that has already been discussed. Note that simulated annealing is used to maximize the log likelihood function. Attempting to use BFGS leads to trouble. I suspect that the log likelihood is approximately non-differentiable in places, around which it is very flat, though I have not checked if this is true. If you run this script, you will see that it takes a long time to get the estimation results, which are:

```
*********************************************************
Poisson Latent Heterogeneity model, SML estimation, MEPS 1996 full data set

MLE Estimation Results
BFGS convergence: Max. iters. exceeded

Average Log-L: -2.171826
Observations: 4564

                estimate     st. err    t-stat        p-value
constant        -1.592       0.146     -10.892       0.000
pub. ins.        1.189       0.068      17.425       0.000
```

```
priv. ins.       0.655       0.065      10.124       0.000
sex              0.615       0.044      13.888       0.000
age              0.018       0.002      10.865       0.000
edu              0.024       0.010       2.523       0.012
inc             -0.000       0.000      -0.531       0.596
lnalpha          0.203       0.014      14.036       0.000

Information Criteria
CAIC : 19899.8396      Avg. CAIC:   4.3602
BIC : 19891.8396       Avg. BIC:   4.3584
AIC : 19840.4320       Avg. AIC:   4.3472
*******************************************************
octave:3>
```

If you compare these results to the results for the negative binomial model, given in subsection (18.2), you can see that the present model fits better according to the CAIC criterion. The present model is considerably less convenient to work with, however, due to the computational requirements. The chapter on parallel computing is relevant if you wish to use models of this sort.

MSM

An example of estimation using the MSM is given in the script file MSM_Example.m. The first order moving average (MA(1)) model has been widely used to investigate the performance of the indirect

inference estimator, and a *pth*-order autoregressive model is often used to generate the auxiliary statistic (see, for example, Gouriéroux, Monfort and Renault, 1993; Chumacero, 2001). In this section we estimate the MA(1) model

$$\begin{aligned} y_t &= \epsilon_t + \psi\epsilon_{t-1} \\ \epsilon_t &\sim i.i.d.\, N(0, \sigma^2) \end{aligned}$$

The parameter vector is $\theta = (\psi, \sigma)$. We set the parameter space for the initial simulated annealing stage (to get good start values for the gradient-based algorithm) to $\Theta = (-1, 1) \times (0, 2)$, which imposes invertibility, which is needed for the parameter to be identified. The statistic Z_n is the vector of estimated parameters $(\rho_0, \rho_1, ..., \rho_P, \sigma_v^2)$ of an AR(P) model $y_t = \rho_0 + \sum_{p=1}^{P} \rho_p y_{t-p} + \upsilon_t$, fit to the data using ordinary least squares.

We estimate θ using MSM implemented as II, using continuously updated GMM (Hanson, Heaton and Yaron, 1996). The moment conditions that define the continuously updated indirect inference (CU-II) estimator are $m_n(\theta) = Z_n - \bar{Z}_{S,n}(\theta)$ where $\bar{Z}_{S,n}(\theta) = \frac{1}{S}\sum_{s=1}^{S} Z_n^s(\theta)$, and the weight matrix at each iteration is the inverse of $\Omega_n^S(\theta) = \frac{1}{S}\sum_{s=1}^{S}\left[Z_n^s(\theta) - \bar{Z}_{S,n}(\theta)\right]\left[Z_n^s(\theta) - \bar{Z}_{S,n}(\theta)\right]'$, where $S = 100$.

Example: EMM estimation of a discrete choice model

In this section consider EMM estimation. There is a sophisticated package by Gallant and Tauchen for this, but here we'll look at some simple, but hopefully didactic code. The file probitdgp.m generates data that follows the probit model. The file emm_moments.m defines EMM moment conditions, where the DGP and score generator can be passed as arguments. Thus, it is a general purpose moment condition for EMM estimation. This file is interesting enough to warrant some discussion.

A listing appears in Listing 19.1. Line 3 defines the DGP, and the arguments needed to evaluate it are defined in line 4. The score generator is defined in line 5, and its arguments are defined in line 6. The QML estimate of the parameter of the score generator is read in line 7. Note in line 10 how the random draws needed to simulate data are passed with the data, and are thus fixed during estimation, to avoid "chattering". The simulated data is generated in line 16, and the derivative of the score generator using the simulated data is calculated in line 18. In line 20 we average the scores of the score generator, which are the moment conditions that the function returns.

```
1  function scores = emm_moments(theta, data, momentargs)
2    k = momentargs{1};
3    dgp = momentargs{2}; # the data generating process (DGP)
4    dgpargs = momentargs{3}; # its arguments (cell array)
5    sg = momentargs{4}; # the score generator (SG)
6    sgargs = momentargs{5}; # SG arguments (cell array)
7    phi = momentargs{6}; # QML estimate of SG parameter
8    y = data(:,1);
9    x = data(:,2:k+1);
10   rand_draws = data(:,k+2:columns(data)); # passed with data to ensure fixed across iterations
11   n = rows(y);
12   scores = zeros(n,rows(phi)); # container for moment contributions
13   reps = columns(rand_draws); # how many simulations?
14   for i = 1:reps
15     e = rand_draws(:,i);
16     y = feval(dgp, theta, x, e, dgpargs); # simulated data
17     sgdata = [y x]; # simulated data for SG
18     scores = scores + numgradient(sg, {phi, sgdata, sgargs}); # gradient of SG
19   endfor
20   scores = scores / reps; # average over number of simulations
```

```
21 endfunction
```

Listing 22.1: emm_moments.m

The file emm_example.m performs EMM estimation of the probit model, using a logit model as the score generator. The results we obtain are

```
Score generator results:
================================================
BFGSMIN final results

Used analytic gradient

------------------------------------------------
STRONG CONVERGENCE
Function conv 1  Param conv 1  Gradient conv 1
------------------------------------------------
Objective function value 0.281571
Stepsize 0.0279
15 iterations
------------------------------------------------

 param    gradient  change
  1.8979   0.0000   0.0000
  1.6648  -0.0000   0.0000
  1.9125  -0.0000   0.0000
  1.8875  -0.0000   0.0000
  1.7433  -0.0000   0.0000
```

```
======================================================

Model results:
******************************************************
EMM example

GMM Estimation Results
BFGS convergence: Normal convergence

Objective function value: 0.000000
Observations: 1000

Exactly identified, no spec. test

      estimate     st. err      t-stat     p-value
p1       1.069       0.022      47.618       0.000
p2       0.935       0.022      42.240       0.000
p3       1.085       0.022      49.630       0.000
p4       1.080       0.022      49.047       0.000
p5       0.978       0.023      41.643       0.000
******************************************************
```

It might be interesting to compare the standard errors with those obtained from ML estimation, to check efficiency of the EMM estimator. One could even do a Monte Carlo study.

Indirect likelihood inference

A simple script that does SBIL estimation of the same DSGE model as discussed above is at DSGE by SBIL. This implements a less sophisticated verion of the estimator than was used to make the tables presented here (less careful choice of auxiliary statistics, no bias correction, etc), but it conveys the main ideas.

22.7 Exercises

1. (basic) Examine the Octave script and function discussed in subsection 22.6 and describe what they do.

2. (basic) Examine the Octave scripts and functions discussed in subsection 22.6 and describe what they do.

3. (advanced, but even if you don't do this you should be able to describe what needs to be done) Write Octave code to do SML estimation of the probit model. Do an estimation using data generated by a probit model (probitdgp.m might be helpful). Compare the SML estimates to ML estimates.

4. (more advanced) Do a little Monte Carlo study to compare ML, SML and EMM estimation of the probit model. Investigate how the number of simulations affect the two simulation-based estimators.

Chapter 23

Parallel programming for econometrics

The following borrows heavily from Creel (2005).

Parallel computing can offer an important reduction in the time to complete computations. This is well-known, but it bears emphasis since it is the main reason that parallel computing may be attractive to users. To illustrate, the Intel Pentium IV (Willamette) processor, running at 1.5GHz, was introduced in November of 2000. The Pentium IV (Northwood-HT) processor, running at 3.06GHz, was introduced in November of 2002. An approximate doubling of the performance of a commodity CPU took place in two years. Extrapolating this admittedly rough snapshot of the evolution of the performance of commodity processors, one would need to wait more than 6.6 years and then purchase a new computer to obtain a 10-fold improvement in computational performance. The examples in this chapter show that a 10-fold improvement in performance can be achieved immediately, using

distributed parallel computing on available computers.

Recent (this is written in 2005) developments that may make parallel computing attractive to a broader spectrum of researchers who do computations. The first is the fact that setting up a cluster of computers for distributed parallel computing is not difficult. If you are using the ParallelKnoppix bootable CD that accompanies these notes, you are less than 10 minutes away from creating a cluster, supposing you have a second computer at hand and a crossover ethernet cable. See the ParallelKnoppix tutorial. A second development is the existence of extensions to some of the high-level matrix programming (HLMP) languages[1] that allow the incorporation of parallelism into programs written in these languages. A third is the spread of dual and quad-core CPUs, so that an ordinary desktop or laptop computer can be made into a mini-cluster. Those cores won't work together on a single problem unless they are told how to.

Following are examples of parallel implementations of several mainstream problems in econometrics. A focus of the examples is on the possibility of hiding parallelization from end users of programs. If programs that run in parallel have an interface that is nearly identical to the interface of equivalent serial versions, end users will find it easy to take advantage of parallel computing's performance. We continue to use Octave, taking advantage of the MPI Toolbox (MPITB) for Octave, by by Fernández Baldomero *et al.* (2004). There are also parallel packages for Ox, R, and Python which may be of interest to econometricians, but as of this writing, the following examples are the most accessible introduction to parallel programming for econometricians.

[1]By "high-level matrix programming language" I mean languages such as MATLAB (TM the Mathworks, Inc.), Ox (TM OxMetrics Technologies, Ltd.), and GNU Octave (`www.octave.org`), for example.

23.1 Example problems

This section introduces example problems from econometrics, and shows how they can be parallelized in a natural way.

Monte Carlo

A Monte Carlo study involves repeating a random experiment many times under identical conditions. Several authors have noted that Monte Carlo studies are obvious candidates for parallelization (Doornik *et al.* 2002; Bruche, 2003) since blocks of replications can be done independently on different computers. To illustrate the parallelization of a Monte Carlo study, we use same trace test example as do Doornik, *et. al.* (2002). tracetest.m is a function that calculates the trace test statistic for the lack of cointegration of integrated time series. This function is illustrative of the format that we adopt for Monte Carlo simulation of a function: it receives a single argument of cell type, and it returns a row vector that holds the results of one random simulation. The single argument in this case is a cell array that holds the length of the series in its first position, and the number of series in the second position. It generates a random result though a process that is internal to the function, and it reports some output in a row vector (in this case the result is a scalar).

mc_example1.m is an Octave script that executes a Monte Carlo study of the trace test by repeatedly evaluating the `tracetest.m` function. The main thing to notice about this script is that lines 7 and 10 call the function `montecarlo.m`. When called with 3 arguments, as in line 7, `montecarlo.m` executes serially on the computer it is called from. In line 10, there is a fourth argument. When called with four arguments, the last argument is the number of slave hosts to use. We see that running the Monte Carlo study on one or more processors is transparent to the user - he or she must only indicate

the number of slave computers to be used.

ML

For a sample $\{(y_t, x_t)\}_n$ of n observations of a set of dependent and explanatory variables, the maximum likelihood estimator of the parameter θ can be defined as

$$\hat{\theta} = \arg\max s_n(\theta)$$

where

$$s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \ln f(y_t|x_t, \theta)$$

Here, y_t may be a vector of random variables, and the model may be dynamic since x_t may contain lags of y_t. As Swann (2002) points out, this can be broken into sums over blocks of observations, for example two blocks:

$$s_n(\theta) = \frac{1}{n} \left\{ \left(\sum_{t=1}^{n_1} \ln f(y_t|x_t, \theta) \right) + \left(\sum_{t=n_1+1}^{n} \ln f(y_t|x_t, \theta) \right) \right\}$$

Analogously, we can define up to n blocks. Again following Swann, parallelization can be done by calculating each block on separate computers.

mle_example1.m is an Octave script that calculates the maximum likelihood estimator of the parameter vector of a model that assumes that the dependent variable is distributed as a Poisson random variable, conditional on some explanatory variables. In lines 1-3 the data is read, the name of the density function is provided in the variable `model`, and the initial value of the parameter vector is set. In line 5, the function `mle_estimate` performs ordinary serial calculation of the ML estimator,

while in line 7 the same function is called with 6 arguments. The fourth and fifth arguments are empty placeholders where options to `mle_estimate` may be set, while the sixth argument is the number of slave computers to use for parallel execution, 1 in this case. A person who runs the program sees no parallel programming code - the parallelization is transparent to the end user, beyond having to select the number of slave computers. When executed, this script prints out the estimates `theta_s` and `theta_p`, which are identical.

It is worth noting that a different likelihood function may be used by making the `model` variable point to a different function. The likelihood function itself is an ordinary Octave function that is not parallelized. The `mle_estimate` function is a generic function that can call any likelihood function that has the appropriate input/output syntax for evaluation either serially or in parallel. Users need only learn how to write the likelihood function using the Octave language.

GMM

For a sample as above, the GMM estimator of the parameter θ can be defined as

$$\hat{\theta} \equiv \arg\min_{\Theta} s_n(\theta)$$

where

$$s_n(\theta) = m_n(\theta)' W_n m_n(\theta)$$

and

$$m_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} m_t(y_t | x_t, \theta)$$

Since $m_n(\theta)$ is an average, it can obviously be computed blockwise, using for example 2 blocks:

$$m_n(\theta) = \frac{1}{n}\left\{\left(\sum_{t=1}^{n_1} m_t(y_t|x_t,\theta)\right) + \left(\sum_{t=n_1+1}^{n} m_t(y_t|x_t,\theta)\right)\right\} \tag{23.1}$$

Likewise, we may define up to n blocks, each of which could potentially be computed on a different machine.

gmm_example1.m is a script that illustrates how GMM estimation may be done serially or in parallel. When this is run, `theta_s` and `theta_p` are identical up to the tolerance for convergence of the minimization routine. The point to notice here is that an end user can perform the estimation in parallel in virtually the same way as it is done serially. Again, `gmm_estimate`, used in lines 8 and 10, is a generic function that will estimate any model specified by the `moments` variable - a different model can be estimated by changing the value of the `moments` variable. The function that `moments` points to is an ordinary Octave function that uses no parallel programming, so users can write their models using the simple and intuitive HLMP syntax of Octave. Whether estimation is done in parallel or serially depends only the seventh argument to `gmm_estimate` - when it is missing or zero, estimation is by default done serially with one processor. When it is positive, it specifies the number of slave nodes to use.

Kernel regression

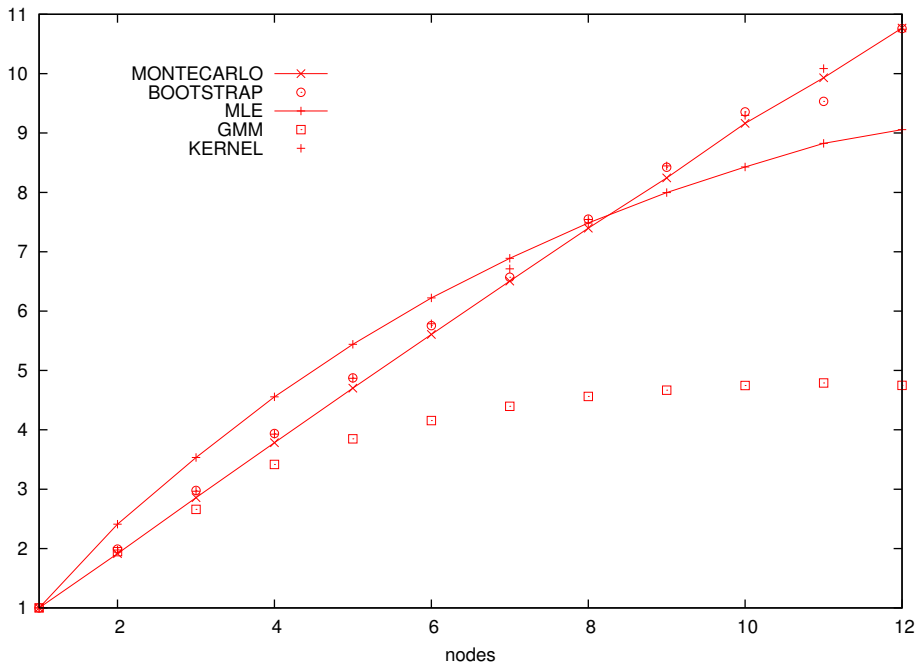
The Nadaraya-Watson kernel regression estimator of a function $g(x)$ at a point x is

$$\begin{aligned}\hat{g}(x) &= \frac{\sum_{t=1}^{n} y_t K\left[\left(x - x_t\right)/\gamma_n\right]}{\sum_{t=1}^{n} K\left[\left(x - x_t\right)/\gamma_n\right]} \\ &\equiv \sum_{t=1}^{n} w_t y_y\end{aligned}$$

We see that the weight depends upon every data point in the sample. To calculate the fit at every point in a sample of size n, on the order of n^2k calculations must be done, where k is the dimension of the vector of explanatory variables, x. Racine (2002) demonstrates that MPI parallelization can be used to speed up calculation of the kernel regression estimator by calculating the fits for portions of the sample on different computers. We follow this implementation here. kernel_example1.m is a script for serial and parallel kernel regression. Serial execution is obtained by setting the number of slaves equal to zero, in line 15. In line 17, a single slave is specified, so execution is in parallel on the master and slave nodes.

The example programs show that parallelization may be mostly hidden from end users. Users can benefit from parallelization without having to write or understand parallel code. The speedups one can obtain are highly dependent upon the specific problem at hand, as well as the size of the cluster, the efficiency of the network, *etc.* Some examples of speedups are presented in Creel (2005). Figure 23.1 reproduces speedups for some econometric problems on a cluster of 12 desktop computers. The speedup for k nodes is the time to finish the problem on a single node divided by the time to finish the problem on k nodes. Note that you can get 10X speedups, as claimed in the introduction. It's pretty obvious that much greater speedups could be obtained using a larger cluster, for the "embarrassingly

Figure 23.1: Speedups from parallelization



parallel" problems.

Bibliography

[1] Bruche, M. (2003) A note on embarassingly parallel computation using OpenMosix and Ox, working paper, Financial Markets Group, London School of Economics.

[2] Creel, M. (2005) User-friendly parallel computations with econometric examples, *Computational Economics,* V. 26, pp. 107-128.

[3] Doornik, J.A., D.F. Hendry and N. Shephard (2002) Computationally-intensive econometrics using a distributed matrix-programming language, *Philosophical Transactions of the Royal Society of London, Series A*, 360, 1245-1266.

[4] Fernández Baldomero, J. (2004) LAM/MPI parallel computing under GNU Octave, atc.ugr.es/javier-bin/mpitb.

[5] Racine, Jeff (2002) Parallel distributed kernel estimation, *Computational Statistics & Data Analysis*, **40**, 293-302.

[6] Swann, C.A. (2002) Maximum likelihood estimation using parallel computing: an introduction to MPI, *Computational Economics*, **19**, 145-178.

Chapter 24

Introduction to Octave

Why is Octave being used here, since it's not that well-known by econometricians? Well, because it is a high quality environment that is easily extensible, uses well-tested and high performance numerical libraries, it is licensed under the GNU GPL, so you can get it for free and modify it if you like, and it runs on both GNU/Linux, Mac OSX and Windows systems. It's also quite easy to learn.

24.1 Getting started

Get the ParallelKnoppix CD, as was described in Section 1.5. Then burn the image, and boot your computer with it. This will give you this same PDF file, but with all of the example programs ready to run. The editor is configure with a macro to execute the programs using Octave, which is of course installed. From this point, I assume you are running the CD (or sitting in the computer room across the hall from my office), or that you have configured your computer to be able to run the `*.m` files mentioned below.

24.2 A short introduction

The objective of this introduction is to learn just the basics of Octave. There are other ways to use Octave, which I encourage you to explore. These are just some rudiments. After this, you can look at the example programs scattered throughout the document (and edit them, and run them) to learn more about how Octave can be used to do econometrics. Students of mine: your problem sets will include exercises that can be done by modifying the example programs in relatively minor ways. So study the examples!

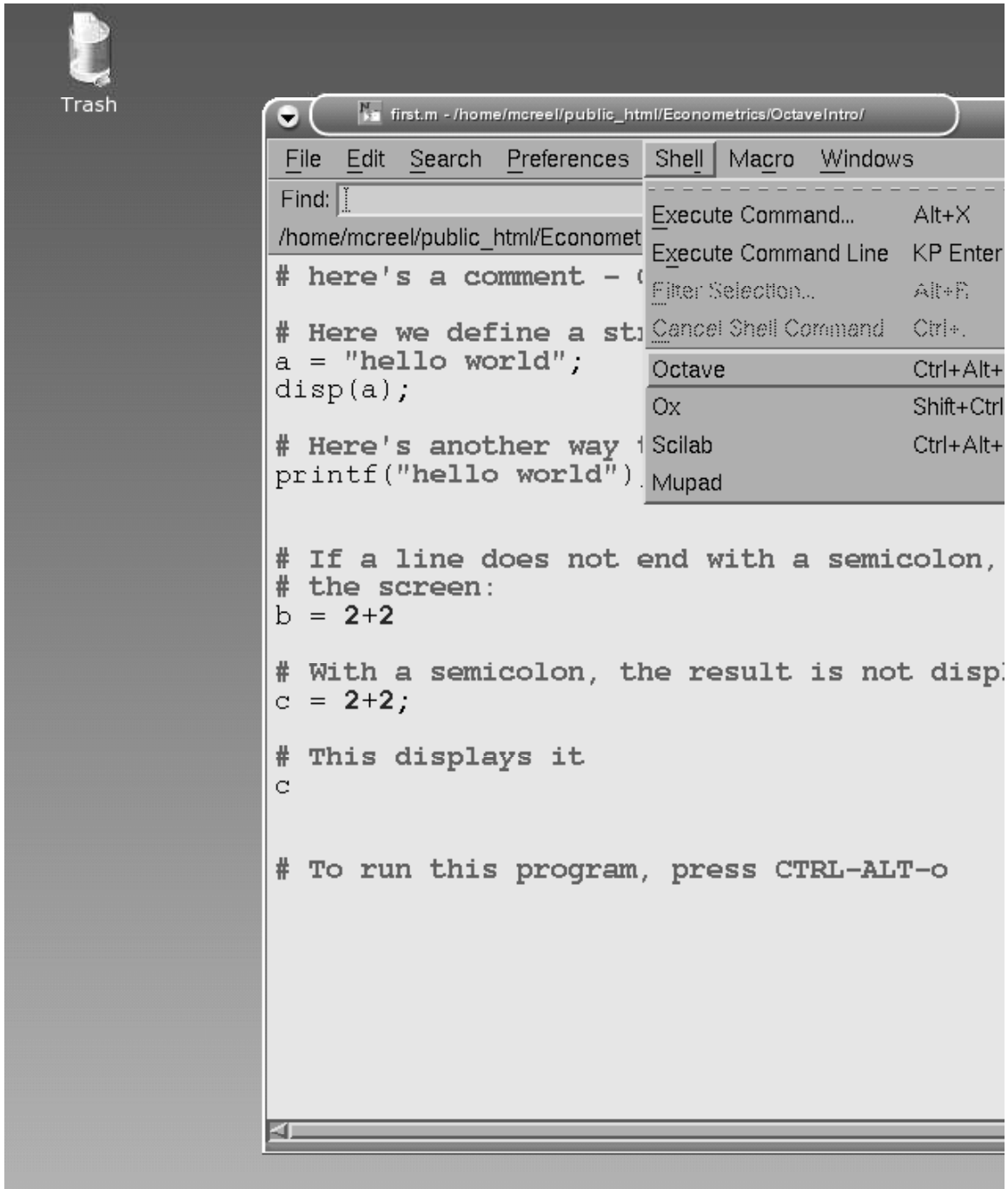
Octave can be used interactively, or it can be used to run programs that are written using a text editor. We'll use this second method, preparing programs with NEdit, and calling Octave from within the editor. The program first.m gets us started. To run this, open it up with NEdit (by finding the correct file inside the `/home/knoppix/Desktop/Econometrics` folder and clicking on the icon) and then type CTRL-ALT-o, or use the Octave item in the Shell menu (see Figure 24.1).

Note that the output is not formatted in a pleasing way. That's because `printf()` doesn't automatically start a new line. Edit `first.m` so that the 8th line reads "`printf("hello world\n");`" and re-run the program.

We need to know how to load and save data. The program second.m shows how. Once you have run this, you will find the file "`x`" in the directory `Econometrics/Examples/OctaveIntro/` You might have a look at it with NEdit to see Octave's default format for saving data. Basically, if you have data in an ASCII text file, named for example "`myfile.data`", formed of numbers separated by spaces, just use the command "`load myfile.data`". After having done so, the matrix "`myfile`" (without extension) will contain the data.

Please have a look at CommonOperations.m for examples of how to do some basic things in Octave. Now that we're done with the basics, have a look at the Octave programs that are included as examples.

Figure 24.1: Running an Octave program



If you are looking at the browsable PDF version of this document, then you should be able to click on links to open them. If not, the example programs are available here and the support files needed to run these are available here. Those pages will allow you to examine individual files, out of context. To actually use these files (edit and run them), you should go to the home page of this document, since you will probably want to download the pdf version together with all the support files and examples. Or get the bootable CD.

There are some other resources for doing econometrics with Octave. You might like to check the article Econometrics with Octave and the Econometrics Toolbox , which is for Matlab, but much of which could be easily used with Octave.

24.3 If you're running a Linux installation...

Then to get the same behavior as found on the CD, you need to:

- Get the collection of support programs and the examples, from the document home page.
- Put them somewhere, and tell Octave how to find them, e.g., by putting a link to the MyOctaveFiles directory in `/usr/local/share/octave/site-m`
- Make sure nedit is installed and configured to run Octave and use syntax highlighting. Copy the file `/home/econometrics/.nedit` from the CD to do this. Or, get the file NeditConfiguration and save it in your $HOME directory with the name "`.nedit`". Not to put too fine a point on it, please note that there is a period in that name.
- Associate `*.m` files with NEdit so that they open up in the editor when you click on them. That should do it.

Chapter 25

Notation and Review

- All vectors will be column vectors, unless they have a transpose symbol (or I forget to apply this rule - your help catching typos and er0rors is much appreciated). For example, if x_t is a $p \times 1$ vector, x_t' is a $1 \times p$ vector. When I refer to a p-vector, I mean a column vector.

25.1 Notation for differentiation of vectors and matrices

[3, Chapter 1]

Let $s(\cdot) : \Re^p \to \Re$ be a real valued function of the p-vector θ. Then $\frac{\partial s(\theta)}{\partial \theta}$ is organized as a p-vector,

$$\frac{\partial s(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial s(\theta)}{\partial \theta_1} \\ \frac{\partial s(\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial s(\theta)}{\partial \theta_p} \end{bmatrix}$$

Following this convention, $\frac{\partial s(\theta)}{\partial \theta'}$ is a $1 \times p$ vector, and $\frac{\partial^2 s(\theta)}{\partial \theta \partial \theta'}$ is a $p \times p$ matrix. Also,

$$\frac{\partial^2 s(\theta)}{\partial \theta \partial \theta'} = \frac{\partial}{\partial \theta}\left(\frac{\partial s(\theta)}{\partial \theta'}\right) = \frac{\partial}{\partial \theta'}\left(\frac{\partial s(\theta)}{\partial \theta}\right).$$

Exercise 70. For a and x both p-vectors, show that $\frac{\partial a'x}{\partial x} = a$.

Let $f(\theta)$:$\Re^p \to \Re^n$ be a n-vector valued function of the p-vector θ. Let $f(\theta)'$ be the $1 \times n$ valued transpose of f . Then $\left(\frac{\partial}{\partial \theta} f(\theta)'\right)' = \frac{\partial}{\partial \theta'} f(\theta)$.

Definition. Product rule. Let $f(\theta)$:$\Re^p \to \Re^n$ and $h(\theta)$:$\Re^p \to \Re^n$ be n-vector valued functions of the p-vector θ. Then

$$\frac{\partial}{\partial \theta'} h(\theta)' f(\theta) = h'\left(\frac{\partial}{\partial \theta'} f\right) + f'\left(\frac{\partial}{\partial \theta'} h\right)$$

has dimension $1 \times p$. Applying the transposition rule we get

$$\frac{\partial}{\partial \theta} h(\theta)' f(\theta) = \left(\frac{\partial}{\partial \theta} f'\right) h + \left(\frac{\partial}{\partial \theta} h'\right) f$$

which has dimension $p \times 1$.

Exercise 71. For A a $p \times p$ matrix and x a $p \times 1$ vector, show that $\frac{\partial x'Ax}{\partial x} = A + A'$.

Definition 72. Chain rule. Let $f(\cdot)$:$\Re^p \to \Re^n$ a n-vector valued function of a p-vector argument, and let $g()$:$\Re^r \to \Re^p$ be a p-vector valued function of an r-vector valued argument ρ. Then

$$\frac{\partial}{\partial \rho'} f\left[g\left(\rho\right)\right] = \left.\frac{\partial}{\partial \theta'} f(\theta)\right|_{\theta = g(\rho)} \frac{\partial}{\partial \rho'} g(\rho)$$

has dimension $n \times r$.

Exercise 73. For x and β both $p \times 1$ vectors, show that $\frac{\partial \exp(x'\beta)}{\partial \beta} = \exp(x'\beta)x$.

25.2 Convergenge modes

Readings: [1, Chapter 4];[4, Chapter 4].

We will consider several modes of convergence. The first three modes discussed are simply for background. The stochastic modes are those which will be used later in the course.

Definition 74. A sequence is a mapping from the natural numbers $\{1, 2, ...\} = \{n\}_{n=1}^{\infty} = \{n\}$ to some other set, so that the set is ordered according to the natural numbers associated with its elements.

Real-valued sequences:

Definition 75. *[Convergence]* A real-valued sequence of vectors $\{a_n\}$ *converges* to the vector a if for any $\varepsilon > 0$ there exists an integer N_ε such that for all $n > N_\varepsilon, \| a_n - a \| < \varepsilon$. a is the *limit* of a_n, written $a_n \to a$.

Deterministic real-valued functions

Consider a sequence of functions $\{f_n(\omega)\}$ where

$$f_n : \Omega \to T \subseteq \Re.$$

Ω may be an arbitrary set.

Definition 76. *[Pointwise convergence]* A sequence of functions $\{f_n(\omega)\}$ *converges pointwise* on Ω to the function $f(\omega)$ if for all $\varepsilon > 0$ and $\omega \in \Omega$ there exists an integer $N_{\varepsilon\omega}$ such that

$$|f_n(\omega) - f(\omega)| < \varepsilon, \forall n > N_{\varepsilon\omega}.$$

It's important to note that $N_{\varepsilon\omega}$ depends upon ω, so that converge may be much more rapid for certain ω than for others. Uniform convergence requires a similar rate of convergence throughout Ω.

Definition 77. *[Uniform convergence]* A sequence of functions $\{f_n(\omega)\}$ *converges uniformly* on Ω to the function $f(\omega)$ if for any $\varepsilon > 0$ there exists an integer N such that

$$\sup_{\omega\in\Omega} |f_n(\omega) - f(\omega)| < \varepsilon, \forall n > N.$$

(insert a diagram here showing the envelope around $f(\omega)$ in which $f_n(\omega)$ must lie).

Stochastic sequences

In econometrics, we typically deal with stochastic sequences. Given a probability space $(\Omega, \mathcal{F}, P)$, recall that a random variable maps the sample space to the real line, i.e., $X(\omega) : \Omega \to \Re$. A sequence of random variables $\{X_n(\omega)\}$ is a collection of such mappings, *i.e.,* each $X_n(\omega)$ is a random variable with respect to the probability space $(\Omega, \mathcal{F}, P)$. For example, given the model $Y = X\beta^0 + \varepsilon$, the OLS estimator $\hat{\beta}_n = (X'X)^{-1} X'Y$, where n is the sample size, can be used to form a sequence of random vectors $\{\hat{\beta}_n\}$. A number of modes of convergence are in use when dealing with sequences of random variables. Several such modes of convergence should already be familiar:

Definition 78. *[Convergence in probability]* Let $X_n(\omega)$ be a sequence of random variables, and let

$X(\omega)$ be a random variable. Let $\mathcal{A}_n = \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}$. Then $\{X_n(\omega)\}$ converges in probability to $X(\omega)$ if

$$\lim_{n\to\infty} P(\mathcal{A}_n) = 0, \forall \varepsilon > 0.$$

Convergence in probability is written as $X_n \xrightarrow{p} X$, or plim $X_n = X$.

Definition 79. *[Almost sure convergence]* Let $X_n(\omega)$ be a sequence of random variables, and let $X(\omega)$ be a random variable. Let $\mathcal{A} = \{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$. Then $\{X_n(\omega)\}$ converges almost surely to $X(\omega)$ if

$$P(\mathcal{A}) = 1.$$

In other words, $X_n(\omega) \to X(\omega)$ (ordinary convergence of the two functions) except on a set $C = \Omega - \mathcal{A}$ such that $P(C) = 0$. Almost sure convergence is written as $X_n \xrightarrow{a.s.} X$, or $X_n \to X, a.s.$ One can show that

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X.$$

Definition 80. *[Convergence in distribution]* Let the r.v. X_n have distribution function F_n and the r.v. X_n have distribution function F. If $F_n \to F$ at every continuity point of F, then X_n converges in distribution to X.

Convergence in distribution is written as $X_n \xrightarrow{d} X$. It can be shown that convergence in probability implies convergence in distribution.

Stochastic functions

Simple laws of large numbers (LLN's) allow us to directly conclude that $\hat{\beta}_n \xrightarrow{a.s.} \beta^0$ in the OLS example, since

$$\hat{\beta}_n = \beta^0 + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right),$$

and $\frac{X'\varepsilon}{n} \xrightarrow{a.s.} 0$ by a SLLN. Note that this term is not a function of the parameter β. This easy proof is a result of the linearity of the model, which allows us to express the estimator in a way that separates parameters from random functions. In general, this is not possible. We often deal with the more complicated situation where the stochastic sequence depends on parameters in a manner that is not reducible to a simple sequence of random variables. In this case, we have a sequence of random functions that depend on θ: $\{X_n(\omega, \theta)\}$, where each $X_n(\omega, \theta)$ is a random variable with respect to a probability space $(\Omega, \mathcal{F}, P)$ and the parameter θ belongs to a parameter space $\theta \in \Theta$.

Definition 81. *[Uniform almost sure convergence]* $\{X_n(\omega, \theta)\}$ converges uniformly almost surely in Θ to $X(\omega, \theta)$ if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} |X_n(\omega, \theta) - X(\omega, \theta)| = 0, \text{(a.s.)}$$

Implicit is the assumption that all $X_n(\omega, \theta)$ and $X(\omega, \theta)$ are random variables w.r.t. $(\Omega, \mathcal{F}, P)$ for all $\theta \in \Theta$. We'll indicate uniform almost sure convergence by $\xrightarrow{u.a.s.}$ and uniform convergence in probability by $\xrightarrow{u.p.}$.

- An equivalent definition, based on the fact that "almost sure" means "with probability one" is

$$\Pr\left(\lim_{n\to\infty} \sup_{\theta\in\Theta} |X_n(\omega, \theta) - X(\omega, \theta)| = 0\right) = 1$$

This has a form similar to that of the definition of a.s. convergence - the essential difference is the addition of the sup.

25.3 Rates of convergence and asymptotic equality

It's often useful to have notation for the relative magnitudes of quantities. Quantities that are small relative to others can often be ignored, which simplifies analysis.

Definition 82. *[Little-o]* Let $f(n)$ and $g(n)$ be two real-valued functions. The notation $f(n) = o(g(n))$ means $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

Definition 83. *[Big-O]* Let $f(n)$ and $g(n)$ be two real-valued functions. The notation $f(n) = O(g(n))$ means there exists some N such that for $n > N$, $\left|\frac{f(n)}{g(n)}\right| < K$, where K is a finite constant.

This definition doesn't require that $\frac{f(n)}{g(n)}$ have a limit (it may fluctuate boundedly).
If $\{f_n\}$ and $\{g_n\}$ are sequences of random variables analogous definitions are

Definition 84. The notation $f(n) = o_p(g(n))$ means $\frac{f(n)}{g(n)} \xrightarrow{p} 0$.

Example 85. The least squares estimator $\hat{\theta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\theta^0 + \varepsilon) = \theta^0 + (X'X)^{-1}X'\varepsilon$. Since $\operatorname{plim}\frac{(X'X)^{-1}X'\varepsilon}{1} = 0$, we can write $(X'X)^{-1}X'\varepsilon = o_p(1)$ and $\hat{\theta} = \theta^0 + o_p(1)$. Asymptotically, the term $o_p(1)$ is negligible. This is just a way of indicating that the LS estimator is consistent.

Definition 86. The notation $f(n) = O_p(g(n))$ means there exists some N_ε such that for $\varepsilon > 0$ and all $n > N_\varepsilon$,

$$P\left(\left|\frac{f(n)}{g(n)}\right| < K_\varepsilon\right) > 1 - \varepsilon,$$

where K_ε is a finite constant.

Example 87. If $X_n \sim N(0,1)$ then $X_n = O_p(1)$, since, given ε, there is always some K_ε such that $P(|X_n| < K_\varepsilon) > 1 - \varepsilon$.

Useful rules:

- $O_p(n^p)O_p(n^q) = O_p(n^{p+q})$
- $o_p(n^p)o_p(n^q) = o_p(n^{p+q})$

Example 88. Consider a random sample of iid r.v.'s with mean 0 and variance σ^2. The estimator of the mean $\hat{\theta} = 1/n \sum_{i=1}^n x_i$ is asymptotically normally distributed, e.g., $n^{1/2}\hat{\theta} \overset{A}{\sim} N(0, \sigma^2)$. So $n^{1/2}\hat{\theta} = O_p(1)$, so $\hat{\theta} = O_p(n^{-1/2})$. Before we had $\hat{\theta} = o_p(1)$, now we have have the stronger result that relates the rate of convergence to the sample size.

Example 89. Now consider a random sample of iid r.v.'s with mean μ and variance σ^2. The estimator of the mean $\hat{\theta} = 1/n \sum_{i=1}^n x_i$ is asymptotically normally distributed, e.g., $n^{1/2}\left(\hat{\theta} - \mu\right) \overset{A}{\sim} N(0, \sigma^2)$. So $n^{1/2}\left(\hat{\theta} - \mu\right) = O_p(1)$, so $\hat{\theta} - \mu = O_p(n^{-1/2})$, so $\hat{\theta} = O_p(1)$.

These two examples show that averages of centered (mean zero) quantities typically have plim 0, while averages of uncentered quantities have finite nonzero plims. Note that the definition of O_p does not mean that $f(n)$ and $g(n)$ are of the same order. Asymptotic equality ensures that this is the case.

Definition 90. Two sequences of random variables $\{f_n\}$ and $\{g_n\}$ are asymptotically equal (written $f_n \stackrel{a}{=} g_n$) if

$$plim\left(\frac{f(n)}{g(n)}\right) = 1$$

Finally, analogous almost sure versions of o_p and O_p are defined in the obvious way.

For a and x both $p \times 1$ vectors, show that $D_x a'x = a$.

For A a $p \times p$ matrix and x a $p \times 1$ vector, show that $D_x^2 x'Ax = A + A'$.

For x and β both $p \times 1$ vectors, show that $D_\beta \exp x'\beta = \exp(x'\beta)x$.

For x and β both $p \times 1$ vectors, find the analytic expression for $D_\beta^2 \exp x'\beta$.

Write an Octave program that verifies each of the previous results by taking numeric derivatives. For a hint, type `help numgradient` and `help numhessian` inside octave.

Chapter 26

Licenses

This document and the associated examples and materials are copyright Michael Creel, under the terms of the GNU General Public License, ver. 2., or at your option, under the Creative Commons Attribution-Share Alike License, Version 2.5. The licenses follow.

26.1 The GPL

```
GNU GENERAL PUBLIC LICENSE
Version 2, June 1991

Copyright (C) 1989, 1991 Free Software Foundation, Inc.
59 Temple Place, Suite 330, Boston, MA 02111-1307 USA
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.
```

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you

distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any

patent must be licensed for everyone’s free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

GNU GENERAL PUBLIC LICENSE
TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".

Activities other than copying, distribution and modification are not

covered by this License; they are outside its scope. The act of
running the Program is not restricted, and the output from the Program
is covered only if its contents constitute a work based on the
Program (independent of having been made by running the Program).
Whether that is true depends on what the Program does.

 1. You may copy and distribute verbatim copies of the Program's
source code as you receive it, in any medium, provided that you
conspicuously and appropriately publish on each copy an appropriate
copyright notice and disclaimer of warranty; keep intact all the
notices that refer to this License and to the absence of any warranty;
and give any other recipients of the Program a copy of this License
along with the Program.

You may charge a fee for the physical act of transferring a copy, and
you may at your option offer warranty protection in exchange for a fee.

 2. You may modify your copy or copies of the Program or any portion
of it, thus forming a work based on the Program, and copy and
distribute such modifications or work under the terms of Section 1
above, provided that you also meet all of these conditions:

 a) You must cause the modified files to carry prominent notices

stating that you changed the files and the date of any change.

b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.

c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If

identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include

anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this

License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions

either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING,

REPAIR OR CORRECTION.

12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

<one line to give the program's name and a brief idea of what it does.>
Copyright (C) <year> <name of author>

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA

Also add information on how to contact you by electronic and paper mail.

If the program is interactive, make it output a short notice like this
when it starts in an interactive mode:

```
 Gnomovision version 69, Copyright (C) year name of author
 Gnomovision comes with ABSOLUTELY NO WARRANTY; for details type ‘show w’.
 This is free software, and you are welcome to redistribute it
 under certain conditions; type ‘show c’ for details.
```

The hypothetical commands ‘show w’ and ‘show c’ should show the appropriate
parts of the General Public License. Of course, the commands you use may
be called something other than ‘show w’ and ‘show c’; they could even be
mouse-clicks or menu items--whatever suits your program.

You should also get your employer (if you work as a programmer) or your
school, if any, to sign a "copyright disclaimer" for the program, if
necessary. Here is a sample; alter the names:

```
 Yoyodyne, Inc., hereby disclaims all copyright interest in the program
 ‘Gnomovision’ (which makes passes at compilers) written by James Hacker.
```

```
<signature of Ty Coon>, 1 April 1989
Ty Coon, President of Vice
```

```
This General Public License does not permit incorporating your program into
proprietary programs. If your program is a subroutine library, you may
consider it more useful to permit linking proprietary applications with the
library. If this is what you want to do, use the GNU Library General
Public License instead of this License.
```

26.2 Creative Commons

Legal Code

Attribution-ShareAlike 2.5

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER

THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

1. "Collective Work" means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.

2. "Derivative Work" means a work based upon the Work or upon the Work and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work will not be considered a Derivative Work for the purpose of this License. For the avoidance of doubt, where the Work is a musical composition or sound recording, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered a Derivative Work for the purpose of this License.

3. "Licensor" means the individual or entity that offers the Work under the terms of this License.

4. "Original Author" means the individual or entity who created the Work.

5. "Work" means the copyrightable work of authorship offered under the terms of this License.

6. "You" means an individual or entity exercising rights under this License who has not previously

violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

7. "License Elements" means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

2. Fair Use Rights. Nothing in this license is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

1. to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

2. to create and reproduce Derivative Works;

3. to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works;

4. to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission Derivative Works.

5.

For the avoidance of doubt, where the work is a musical composition:

1. Performance Royalties Under Blanket Licenses. Licensor waives the exclusive right to collect, whether individually or via a performance rights society (e.g. ASCAP, BMI, SESAC), royalties for the public performance or public digital performance (e.g. webcast) of the Work.

2. Mechanical Rights and Statutory Royalties. Licensor waives the exclusive right to collect,

whether individually or via a music rights society or designated agent (e.g. Harry Fox Agency), royalties for any phonorecord You create from the Work ("cover version") and distribute, subject to the compulsory license created by 17 USC Section 115 of the US Copyright Act (or the equivalent in other jurisdictions).

6. Webcasting Rights and Statutory Royalties. For the avoidance of doubt, where the Work is a sound recording, Licensor waives the exclusive right to collect, whether individually or via a performance-rights society (e.g. SoundExchange), royalties for the public digital performance (e.g. webcast) of the Work, subject to the compulsory license created by 17 USC Section 114 of the US Copyright Act (or the equivalent in other jurisdictions).

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4. Restrictions.The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

1. You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this License, and You must include a copy of, or the Uniform Resource Identifier for, this License with every copy or phonorecord of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this License or the recipients' exercise of the rights granted hereunder. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in

a manner inconsistent with the terms of this License Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this License. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any credit as required by clause 4(c), as requested. If You create a Derivative Work, upon notice from any Licensor You must, to the extent practicable, remove from the Derivative Work any credit as required by clause 4(c), as requested.

2. You may distribute, publicly display, publicly perform, or publicly digitally perform a Derivative Work only under the terms of this License, a later version of this License with the same License Elements as this License, or a Creative Commons iCommons license that contains the same License Elements as this License (e.g. Attribution-ShareAlike 2.5 Japan). You must include a copy of, or the Uniform Resource Identifier for, this License or other license specified in the previous sentence with every copy or phonorecord of each Derivative Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Derivative Works that alter or restrict the terms of this License or the recipients' exercise of the rights granted hereunder, and You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Derivative Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this License Agreement. The above applies to the Derivative Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Derivative Work itself to be made subject to the terms of this License.

3. If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Derivative Works or Collective Works, You must keep intact all copyright notices for the Work

and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or (ii) if the Original Author and/or Licensor designate another party or parties (e.g. a sponsor institute, publishing entity, journal) for attribution in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; the title of the Work if supplied; to the extent reasonably practicable, the Uniform Resource Identifier, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and in the case of a Derivative Work, a credit identifying the use of the Work in the Derivative Work (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Derivative Work or Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE MATERIALS, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY

SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

1. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Derivative Works or Collective Works from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

2. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

1. Each time You distribute or publicly digitally perform the Work or a Collective Work, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

2. Each time You distribute or publicly digitally perform a Derivative Work, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

3. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect

the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

4. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

5. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, neither party will use the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time.

Creative Commons may be contacted at http://creativecommons.org/.

Chapter 27

The attic

This holds material that is not really ready to be incorporated into the main body, but that I don't want to lose. Basically, ignore it, unless you'd like to help get it ready for inclusion.

Invertibility of AR process

To begin with, define the lag operator L

$$Ly_t = y_{t-1}$$

The lag operator is defined to behave just as an algebraic quantity, e.g.,

$$\begin{aligned} L^2 y_t &= L(Ly_t) \\ &= Ly_{t-1} \\ &= y_{t-2} \end{aligned}$$

or

$$\begin{aligned}(1-L)(1+L)y_t &= 1 - Ly_t + Ly_t - L^2 y_t \\ &= 1 - y_{t-2}\end{aligned}$$

A mean-zero AR(p) process can be written as

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} = \varepsilon_t$$

or

$$y_t(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) = \varepsilon_t$$

Factor this polynomial as

$$1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p = (1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_p L)$$

For the moment, just assume that the λ_i are coefficients to be determined. Since L is defined to operate as an algebraic quantitiy, determination of the λ_i is the same as determination of the λ_i such that the following two expressions are the same for all z :

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = (1 - \lambda_1 z)(1 - \lambda_2 z) \cdots (1 - \lambda_p z)$$

Multiply both sides by z^{-p}

$$z^{-p} - \phi_1 z^{1-p} - \phi_2 z^{2-p} - \cdots \phi_{p-1} z^{-1} - \phi_p = (z^{-1} - \lambda_1)(z^{-1} - \lambda_2) \cdots (z^{-1} - \lambda_p)$$

and now define $\lambda = z^{-1}$ so we get

$$\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_{p-1}\lambda - \phi_p = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_p)$$

The LHS is precisely the determinantal polynomial that gives the eigenvalues of F. Therefore, the λ_i that are the coefficients of the factorization are simply the eigenvalues of the matrix F.

Now consider a different stationary process

$$(1 - \phi L)y_t = \varepsilon_t$$

- Stationarity, as above, implies that $|\phi| < 1$.

Multiply both sides by $1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j$ to get

$$\left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right)(1 - \phi L)y_t = \left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right)\varepsilon_t$$

or, multiplying the polynomials on the LHS, we get

$$\left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j - \phi L - \phi^2 L^2 - ... - \phi^j L^j - \phi^{j+1} L^{j+1}\right) y_t = \left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right)\varepsilon_t$$

and with cancellations we have

$$\left(1 - \phi^{j+1} L^{j+1}\right) y_t = \left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right)\varepsilon_t$$

so

$$y_t = \phi^{j+1} L^{j+1} y_t + \left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right)\varepsilon_t$$

Now as $j \to \infty$, $\phi^{j+1} L^{j+1} y_t \to 0$, since $|\phi| < 1$, so

$$y_t \cong \left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right) \varepsilon_t$$

and the approximation becomes better and better as j increases. However, we started with

$$(1 - \phi L) y_t = \varepsilon_t$$

Substituting this into the above equation we have

$$y_t \cong \left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right) (1 - \phi L) y_t$$

so

$$\left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right) (1 - \phi L) \cong 1$$

and the approximation becomes arbitrarily good as j increases arbitrarily. Therefore, for $|\phi| < 1$, define

$$(1 - \phi L)^{-1} = \sum_{j=0}^{\infty} \phi^j L^j$$

Recall that our mean zero AR(p) process

$$y_t (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) = \varepsilon_t$$

can be written using the factorization

$$y_t (1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_p L) = \varepsilon_t$$

where the λ are the eigenvalues of F, and given stationarity, all the $|\lambda_i| < 1$. Therefore, we can invert each first order polynomial on the LHS to get

$$y_t = \left(\sum_{j=0}^{\infty} \lambda_1^j L^j\right)\left(\sum_{j=0}^{\infty} \lambda_2^j L^j\right)\cdots\left(\sum_{j=0}^{\infty} \lambda_p^j L^j\right)\varepsilon_t$$

The RHS is a product of infinite-order polynomials in L, which can be represented as

$$y_t = (1 + \psi_1 L + \psi_2 L^2 + \cdots)\varepsilon_t$$

where the ψ_i are real-valued and absolutely summable.

- The ψ_i are formed of products of powers of the λ_i, which are in turn functions of the ϕ_i.
- The ψ_i are real-valued because any complex-valued λ_i always occur in conjugate pairs. This means that if $a + bi$ is an eigenvalue of F, then so is $a - bi$. In multiplication

$$\begin{aligned}(a + bi)(a - bi) &= a^2 - abi + abi - b^2 i^2 \\ &= a^2 + b^2\end{aligned}$$

  which is real-valued.

- This shows that an AR(p) process is representable as an infinite-order MA(q) process.
- Recall before that by recursive substitution, an AR(p) process can be written as

$$Y_{t+j} = C + FC + \cdots + F^j C + F^{j+1} Y_{t-1} + F^j E_t + F^{j-1} E_{t+1} + \cdots + F E_{t+j-1} + E_{t+j}$$

If the process is mean zero, then everything with a C drops out. Take this and lag it by j periods to get

$$Y_t = F^{j+1}Y_{t-j-1} + F^j E_{t-j} + F^{j-1}E_{t-j+1} + \cdots + FE_{t-1} + E_t$$

As $j \to \infty$, the lagged Y on the RHS drops out. The E_{t-s} are vectors of zeros except for their first element, so we see that the first equation here, in the limit, is just

$$y_t = \sum_{j=0}^{\infty} \left(F^j\right)_{1,1} \varepsilon_{t-j}$$

which makes explicit the relationship between the ψ_i and the ϕ_i (and the λ_i as well, recalling the previous factorization of F^j).

Invertibility of MA(q) process

An MA(q) can be written as

$$y_t - \mu = (1 + \theta_1 L + ... + \theta_q L^q)\varepsilon_t$$

As before, the polynomial on the RHS can be factored as

$$(1 + \theta_1 L + ... + \theta_q L^q) = (1 - \eta_1 L)(1 - \eta_2 L)...(1 - \eta_q L)$$

and each of the $(1 - \eta_i L)$ can be inverted as long as each of the $|\eta_i| < 1$. If this is the case, then we can write

$$(1 + \theta_1 L + ... + \theta_q L^q)^{-1}(y_t - \mu) = \varepsilon_t$$

where

$$(1 + \theta_1 L + ... + \theta_q L^q)^{-1}$$

will be an infinite-order polynomial in L, so we get

$$\sum_{j=0}^{\infty} -\delta_j L^j (y_{t-j} - \mu) = \varepsilon_t$$

with $\delta_0 = -1$, or

$$(y_t - \mu) - \delta_1 (y_{t-1} - \mu) - \delta_2 (y_{t-2} - \mu) + ... = \varepsilon_t$$

or

$$y_t = c + \delta_1 y_{t-1} + \delta_2 y_{t-2} + ... + \varepsilon_t$$

where

$$c = \mu + \delta_1 \mu + \delta_2 \mu + ...$$

So we see that an MA(q) has an infinite AR representation, as long as the $|\eta_i| < 1$, $i = 1, 2, ..., q$.

- It turns out that one can always manipulate the parameters of an MA(q) process to find an invertible representation. For example, the two MA(1) processes

  $$y_t - \mu = (1 - \theta L)\varepsilon_t$$

  and

  $$y_t^* - \mu = (1 - \theta^{-1} L)\varepsilon_t^*$$

have exactly the same moments if

$$\sigma^2_{\varepsilon^*} = \sigma^2_{\varepsilon}\theta^2$$

For example, we've seen that

$$\gamma_0 = \sigma^2(1+\theta^2).$$

Given the above relationships amongst the parameters,

$$\gamma_0^* = \sigma^2_{\varepsilon}\theta^2(1+\theta^{-2}) = \sigma^2(1+\theta^2)$$

so the variances are the same. It turns out that *all* the autocovariances will be the same, as is easily checked. This means that the two MA processes are *observationally equivalent*. As before, it's impossible to distinguish between observationally equivalent processes on the basis of data.

- For a given MA(q) process, it's always possible to manipulate the parameters to find an invertible representation (which is unique).
- It's important to find an invertible representation, since it's the only representation that allows one to represent ε_t as a function of past $y's$. The other representations express ϵ_t as a function of future $y's$
- Why is invertibility important? The most important reason is that it provides a justification for the use of parsimonious models. Since an AR(1) process has an MA(∞) representation, one can reverse the argument and note that at least some MA(∞) processes have an AR(1) representation. Likewise, some AR(∞) processes have an MA(1) representation. At the time of estimation, it's a lot easier to estimate the single AR(1) or MA(1) coefficient rather than the infinite number of coefficients associated with the MA(∞) or AR(∞) representation.

- This is the reason that ARMA models are popular. Combining low-order AR and MA models can usually offer a satisfactory representation of univariate time series data using a reasonable number of parameters.

- Stationarity and invertibility of ARMA models is similar to what we've seen - we won't go into the details. Likewise, calculating moments is similar.

Exercise 91. Calculate the autocovariances of an ARMA(1,1) model:$(1 + \phi L)y_t = c + (1 + \theta L)\epsilon_t$

Optimal instruments for GMM

PLEASE IGNORE THE REST OF THIS SECTION: there is a flaw in the argument that needs correction. In particular, it may be the case that $E(Z_t\epsilon_t) \neq 0$ if instruments are chosen in the way suggested here.

An interesting question that arises is how one should choose the instrumental variables $Z(w_t)$ to achieve maximum efficiency.

Note that with this choice of moment conditions, we have that $D_n \equiv \frac{\partial}{\partial\theta}m'(\theta)$ (a $K \times g$ matrix) is

$$\begin{aligned} D_n(\theta) &= \frac{\partial}{\partial\theta}\frac{1}{n}\left(Z_n'h_n(\theta)\right)' \\ &= \frac{1}{n}\left(\frac{\partial}{\partial\theta}h_n'(\theta)\right)Z_n \end{aligned}$$

which we can define to be

$$D_n(\theta) = \frac{1}{n}H_nZ_n.$$

where H_n is a $K \times n$ matrix that has the derivatives of the individual moment conditions as its columns.

Likewise, define the var-cov. of the moment conditions

$$\begin{aligned}\Omega_n &= \mathcal{E}\left[n m_n(\theta^0) m_n(\theta^0)'\right] \\ &= \mathcal{E}\left[\frac{1}{n} Z_n' h_n(\theta^0) h_n(\theta^0)' Z_n\right] \\ &= Z_n' \mathcal{E}\left(\frac{1}{n} h_n(\theta^0) h_n(\theta^0)'\right) Z_n \\ &\equiv Z_n' \frac{\Phi_n}{n} Z_n\end{aligned}$$

where we have defined $\Phi_n = V\left(h_n(\theta^0)\right)$. Note that the dimension of this matrix is growing with the sample size, so it is not consistently estimable without additional assumptions.

The asymptotic normality theorem above says that the GMM estimator using the optimal weighting matrix is distributed as

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N(0, V_\infty)$$

where

$$V_\infty = \lim_{n\to\infty} \left(\left(\frac{H_n Z_n}{n}\right) \left(\frac{Z_n' \Phi_n Z_n}{n}\right)^{-1} \left(\frac{Z_n' H_n'}{n}\right) \right)^{-1}. \tag{27.1}$$

Using an argument similar to that used to prove that Ω_∞^{-1} is the efficient weighting matrix, we can show that putting

$$Z_n = \Phi_n^{-1} H_n'$$

causes the above var-cov matrix to simplify to

$$V_\infty = \lim_{n\to\infty} \left(\frac{H_n \Phi_n^{-1} H_n'}{n}\right)^{-1}. \tag{27.2}$$

and furthermore, this matrix is smaller that the limiting var-cov for any other choice of instrumental variables. (To prove this, examine the difference of the inverses of the var-cov matrices with the optimal intruments and with non-optimal instruments. As above, you can show that the difference is positive semi-definite).

- Note that both H_n, which we should write more properly as $H_n(\theta^0)$, since it depends on θ^0, and Φ must be consistently estimated to apply this.
- Usually, estimation of H_n is straightforward - one just uses

$$\widehat{H} = \frac{\partial}{\partial \theta} h_n' \left(\tilde{\theta} \right),$$

  where $\tilde{\theta}$ is some initial consistent estimator based on non-optimal instruments.
- Estimation of Φ_n may not be possible. It is an $n \times n$ matrix, so it has more unique elements than n, the sample size, so without restrictions on the parameters it can't be estimated consistently. Basically, you need to provide a parametric specification of the covariances of the $h_t(\theta)$ in order to be able to use optimal instruments. A solution is to approximate this matrix parametrically to define the instruments. Note that the simplified var-cov matrix in equation 27.2 will not apply if approximately optimal instruments are used - it will be necessary to use an estimator based upon equation 27.1, where the term $n^{-1} Z_n' \Phi_n Z_n$ must be estimated consistently apart, for example by the Newey-West procedure.

27.1 Hurdle models

Returning to the Poisson model, lets look at actual and fitted count probabilities. Actual relative frequencies are $f(y = j) = \sum_i 1(y_i = j)/n$ and fitted frequencies are $\hat{f}(y = j) = \sum_{i=1}^{n} f_Y(j|x_i, \hat{\theta})/n$ We

Table 27.1: Actual and Poisson fitted frequencies

Count	OBDV		ERV	
Count	Actual	Fitted	Actual	Fitted
0	0.32	0.06	0.86	0.83
1	0.18	0.15	0.10	0.14
2	0.11	0.19	0.02	0.02
3	0.10	0.18	0.004	0.002
4	0.052	0.15	0.002	0.0002
5	0.032	0.10	0	2.4e-5

see that for the OBDV measure, there are many more actual zeros than predicted. For ERV, there are somewhat more actual zeros than fitted, but the difference is not too important.

Why might OBDV not fit the zeros well? What if people made the decision to contact the doctor for a first visit, they are sick, then the *doctor* decides on whether or not follow-up visits are needed. This is a principal/agent type situation, where the total number of visits depends upon the decision of both the patient and the doctor. Since different parameters may govern the two decision-makers choices, we might expect that different parameters govern the probability of zeros versus the other counts. Let λ_p be the parameters of the patient's demand for visits, and let λ_d be the paramter of the doctor's "demand" for visits. The patient will initiate visits according to a discrete choice model, for example, a logit model:

$$\begin{aligned}\Pr(Y=0) &= f_Y(0,\lambda_p) = 1 - 1/\left[1+\exp(-\lambda_p)\right] \\ \Pr(Y>0) &= 1/\left[1+\exp(-\lambda_p)\right],\end{aligned}$$

The above probabilities are used to estimate the binary 0/1 hurdle process. Then, for the observations where visits are positive, a truncated Poisson density is estimated. This density is

$$\begin{aligned}f_Y(y,\lambda_d|y>0) &= \frac{f_Y(y,\lambda_d)}{\Pr(y>0)} \\ &= \frac{f_Y(y,\lambda_d)}{1-\exp(-\lambda_d)}\end{aligned}$$

since according to the Poisson model with the doctor's paramaters,

$$\Pr(y=0) = \frac{\exp(-\lambda_d)\lambda_d^0}{0!}.$$

Since the hurdle and truncated components of the overall density for Y share no parameters, they may be estimated separately, which is computationally more efficient than estimating the overall model. (Recall that the BFGS algorithm, for example, will have to invert the approximated Hessian. The computational overhead is of order K^2 where K is the number of parameters to be estimated) . The expectation of Y is

$$\begin{aligned}E(Y|x) &= \Pr(Y>0|x)E(Y|Y>0,x) \\ &= \left(\frac{1}{1+\exp(-\lambda_p)}\right)\left(\frac{\lambda_d}{1-\exp(-\lambda_d)}\right)\end{aligned}$$

Here are hurdle Poisson estimation results for OBDV, obtained from this estimation program

```
**************************************************************************
MEPS data, OBDV
logit results
Strong convergence
Observations = 500
Function value      -0.58939
t-Stats
                     params       t(OPG)     t(Sand.)       t(Hess)
constant            -1.5502      -2.5709      -2.5269      -2.5560
pub_ins              1.0519       3.0520       3.0027       3.0384
priv_ins            0.45867       1.7289       1.6924       1.7166
sex                 0.63570       3.0873       3.1677       3.1366
age                0.018614       2.1547       2.1969       2.1807
educ               0.039606       1.0467      0.98710       1.0222
inc                0.077446       1.7655       2.1672       1.9601
Information Criteria
Consistent Akaike
       639.89
Schwartz
       632.89
Hannan-Quinn
       614.96
```

Akaike

603.39

****

The results for the truncated part:

```
**************************************************************************
MEPS data, OBDV
tpoisson results
Strong convergence
Observations = 500
Function value       -2.7042
t-Stats
                    params       t(OPG)     t(Sand.)       t(Hess)
constant           0.54254       7.4291       1.1747       3.2323
pub_ins            0.31001       6.5708       1.7573       3.7183
priv_ins          0.014382      0.29433      0.10438      0.18112
sex                0.19075       10.293       1.1890       3.6942
age               0.016683       16.148       3.5262       7.9814
educ              0.016286       4.2144      0.56547       1.6353
inc             -0.0079016      -2.3186     -0.35309     -0.96078
Information Criteria
Consistent Akaike
       2754.7
Schwartz
       2747.7
Hannan-Quinn
       2729.8
```

Akaike

2718.2

****

Fitted and actual probabilites (NB-II fits are provided as well) are:

Table 27.2: Actual and Hurdle Poisson fitted frequencies

Count	OBDV			ERV		
Count	Actual	Fitted HP	Fitted NB-II	Actual	Fitted HP	Fitted NB-II
0	0.32	0.32	0.34	0.86	0.86	0.86
1	0.18	0.035	0.16	0.10	0.10	0.10
2	0.11	0.071	0.11	0.02	0.02	0.02
3	0.10	0.10	0.08	0.004	0.006	0.006
4	0.052	0.11	0.06	0.002	0.002	0.002
5	0.032	0.10	0.05	0	0.0005	0.001

For the Hurdle Poisson models, the ERV fit is very accurate. The OBDV fit is not so good. Zeros are exact, but 1's and 2's are underestimated, and higher counts are overestimated. For the NB-II fits, performance is at least as good as the hurdle Poisson model, and one should recall that many fewer parameters are used. Hurdle version of the negative binomial model are also widely used.

Finite mixture models

The following are results for a mixture of 2 negative binomial (NB-I) models, for the OBDV data, which you can replicate using this estimation program

```
*******************************************************************************
MEPS data, OBDV
mixnegbin results
Strong convergence
Observations = 500
Function value      -2.2312
t-Stats
                    params       t(OPG)     t(Sand.)       t(Hess)
constant           0.64852       1.3851       1.3226       1.4358
pub_ins          -0.062139     -0.23188     -0.13802     -0.18729
priv_ins          0.093396      0.46948      0.33046      0.40854
sex                0.39785       2.6121       2.2148       2.4882
age               0.015969       2.5173       2.5475       2.7151
educ             -0.049175      -1.8013      -1.7061      -1.8036
inc               0.015880      0.58386      0.76782      0.73281
ln_alpha           0.69961       2.3456       2.0396       2.4029
constant           -3.6130      -1.6126      -1.7365      -1.8411
pub_ins             2.3456       1.7527       3.7677       2.6519
priv_ins           0.77431      0.73854       1.1366      0.97338
sex                0.34886      0.80035      0.74016      0.81892
age               0.021425       1.1354       1.3032       1.3387
educ               0.22461       2.0922       1.7826       2.1470
inc               0.019227      0.20453      0.40854      0.36313
```

```
ln_alpha            2.8419       6.2497       6.8702       7.6182
logit_inv_mix      0.85186       1.7096       1.4827       1.7883
Information Criteria
Consistent Akaike
       2353.8
Schwartz
       2336.8
Hannan-Quinn
       2293.3
Akaike
       2265.2
******************************************************************************
Delta method for mix parameter st.  err.
          mix       se_mix
      0.70096      0.12043
```

- The 95% confidence interval for the mix parameter is perilously close to 1, which suggests that there may really be only one component density, rather than a mixture. Again, this is *not* the way to test this - it is merely suggestive.

- Education is interesting. For the subpopulation that is "healthy", i.e., that makes relatively few visits, education seems to have a positive effect on visits. For the "unhealthy" group, education has a negative effect on visits. The other results are more mixed. A larger sample could help clarify things.

The following are results for a 2 component constrained mixture negative binomial model where all the slope parameters in $\lambda_j = e^{\mathbf{x}\beta_j}$ are the same across the two components. The constants and the overdispersion parameters α_j are allowed to differ for the two components.

```
*************************************************************************
MEPS data, OBDV
cmixnegbin results
Strong convergence
Observations = 500
Function value      -2.2441
t-Stats
                     params       t(OPG)     t(Sand.)       t(Hess)
constant          -0.34153     -0.94203     -0.91456     -0.97943
pub_ins            0.45320       2.6206       2.5088       2.7067
priv_ins           0.20663       1.4258       1.3105       1.3895
sex                0.37714       3.1948       3.4929       3.5319
age               0.015822       3.1212       3.7806       3.7042
educ              0.011784      0.65887      0.50362      0.58331
inc               0.014088      0.69088      0.96831      0.83408
ln_alpha            1.1798       4.6140       7.2462       6.4293
const_2             1.2621      0.47525       2.5219       1.5060
lnalpha_2           2.7769       1.5539       6.4918       4.2243
logit_inv_mix       2.4888      0.60073       3.7224       1.9693

Information Criteria
Consistent Akaike
       2323.5
```

```
Schwartz
       2312.5
Hannan-Quinn
       2284.3
Akaike
       2266.1
**************************************************************************
Delta method for mix parameter st.  err.
          mix       se_mix
      0.92335     0.047318
```

- Now the mixture parameter is even closer to 1.
- The slope parameter estimates are pretty close to what we got with the NB-I model.

Bibliography

[1] Davidson, R. and J.G. MacKinnon (1993) *Estimation and Inference in Econometrics*, Oxford Univ. Press.

[2] Davidson, R. and J.G. MacKinnon (2004) *Econometric Theory and Methods*, Oxford Univ. Press.

[3] Gallant, A.R. (1985) *Nonlinear Statistical Models*, Wiley.

[4] Gallant, A.R. (1997) *An Introduction to Econometric Theory*, Princeton Univ. Press.

[5] Hamilton, J. (1994) *Time Series Analysis*, Princeton Univ. Press

[6] Hayashi, F. (2000) *Econometrics*, Princeton Univ. Press.

[7] Wooldridge (2003), *Introductory Econometrics,* Thomson. (undergraduate level, for supplementary use only).

Index