Graduate Econometrics Lecture Notes

Michael Creel[*]

Version 0.2, Copyright (C) Jan 28, 2002 by Michael Creel

Contents

[*]Dept. of Economics and Economic History, Universitat Autònoma de Barcelona. michael.creel@uab.es

1 License, availability and use

1.1 License

These lecture notes are copyrighted by Michael Creel with the date that appears above. The are provided under the terms of the GNU General Public License, which forms Section 25 of the notes. The main thing you need to know is that you are free to modify and distribute these notes in any way you like, as long as you do so under the terms of the GPL. In particular, you must provide the source code for your version of the notes.

1.2 Obtaining the notes

These notes are part of the OMEGA (Open-source Materials for Econometrics, GPL Archive) project at http://pareto.uab.es/omega. They were prepared using L$_Y$X (http://www.lyx.org). L$_Y$X is an open source "what you see is what you mean" word processor. It can export your work in TEX, HTML, PDF and several other forms. It will run on Unix, Windows, and MacOS systems. The source code is the L$_Y$X file notes.lyx, which is available at http://pareto.uab.es/omega/Project_001/. There you will find the L$_Y$X source file, as well as PDF, HTML, TEX and zipped HTML versions of the notes.

1.3 Use

You are free to use the notes as you like, for study, preparing a course, etc. I find that a hard copy is of most use for lecturing or study, while the html version is useful for quick reference or answering students' questions in office hours. I would greatly appreciate that you inform me of any errors you find. I'd also welcome contributions in any area, especially in the areas of time series and nonstationary data.

2 Economic and econometric models

A model from economic theory:

$$x_i = x_i(p_i, m_i, z_i)$$

- x_i is $G \times 1$ vector of quantities demanded

- p_i is $G \times 1$ vector of prices

- m_i is income

- z_i is a vector of individual characteristics related to preferences

Suppose a sample of one observation of n individuals' demands at time period t (this is a *cross section*). The model is not estimable as it stands.

- The form of the demand function is different for all i.

- Some components of z_i are subject to fluctuations that are not observable to outside modeler (people don't eat the same lunch every day). Break z_i into the observable components w_i and an unobservable component ε_i.

An estimable (*e.g.,* econometric) model is

$$x_i = \beta_0 + p_i'\beta_p + m_i\beta_m + w_i'\beta_w + \varepsilon_i$$

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

- The functions $x_i(\cdot)$ which may differ for all i have been restricted to all belong to the same parametric family.

- Of all parametric families of functions, we have restricted the model to the class of linear in the variables functions.

These are **very strong restrictions**, compared to the theoretical model. Furthermore, **these restrictions have no theoretical basis**. The validity of any results we obtain using this model will be contingent on these restrictions being 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. In the next sections we will obtain results supposing that the econometric model is correctly specified. Later we will examine the consequences of misspecification and see some methods for determining if a model is correctly specified.

3 Ordinary Least Squares

3.1 The classical linear model

The classical linear model is based upon several assumptions.

1. **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,

$$y = \mathbf{X}\beta_0 + \varepsilon,$$

   where y is $n \times 1$, $\mathbf{X} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}'$, where x_t is $K \times 1$, and β_0 and ε are conformable. The subscript "0" in β_0 means this is the true value of the unknown parameter. It will be suppressed when it's not necessary for clarity. Linear models are more general than they might first appear, since one can employ nonlinear transformations of the variables:

$$\varphi_0(z_t) = \begin{bmatrix} \varphi_1(w_t) & \varphi_2(w_t) & \cdots & \varphi_p(w_t) \end{bmatrix} \beta_0 + \varepsilon_t$$

   (The $\phi_i()$ are known functions). Defining $y_t = \varphi_0(z_t)$, $x_{t1} = \varphi_1(w_t)$, etc. leads to a model in the form of equation (**??**). For example, the Cobb-Douglas model

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

   can be transformed logarithmically to obtain

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

2. **IID mean zero errors**:

$$\mathcal{E}(\varepsilon) = 0$$

$$Var(\varepsilon) = \mathcal{E}(\varepsilon\varepsilon') = \sigma_0^2 I_n$$

3. **Nonstochastic, linearly independent regressors**

   (a) X has rank K

   (b) X is nonstochastic

   (c) $\lim_{n\to\infty} \frac{1}{n} X'X = Q_X$, a finite positive definite matrix.

4. **Normality (Optional)**: ε is normally distributed

3.2 Estimation by least squares

The objective is to gain information about the unknown parameters β_0 and σ_0^2.

$$\hat{\beta} = \arg\min s(\beta) = \sum_{t=1}^{n} \left(y_t - x_t'\beta\right)^2$$
$$s(\beta) = (y - X\beta)'(y - X\beta)$$
$$= y'y - 2y'X\beta + \beta'X'X\beta$$
$$= \| y - X\beta \|^2$$

This last expression makes it clear how the OLS estimator chooses $\hat{\beta}$: it minimizes the Euclidean distance between y and $X\beta$.

- To minimize the criterion $s(\beta)$, take the f.o.n.c. and set them to zero:

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

so

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

- To verify that this is a minimum, check the s.o.s.c.:

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

  Since $\rho(X) = K$, 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 in the vector $\hat{y} = X\hat{\beta}$.

- The *residuals* are in the vector $\hat{\varepsilon} = y - X\hat{\beta}$

- Note that

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

3.3 Estimating the error variance

The OLS estimator of σ_0^2 is

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

3.4 Geometric interpretation of least squares estimation

3.4.1 In X, Y Space

Do a plot with the true line, observations and the estimated line. Note the impact of outliers.

3.4.2　In Observation Space

If we want to plot in observation space, we'll need to use only two or three observations. Let's use two. With only two observations, we can't have $K > 1$. *Draw a picture with two observations and one regressor.*

- 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.

3.4.3　Projection Matrices

- We have that $X\hat{\beta}$ is the projection of y on 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 *off* the space spanned by X (that is onto the space that is

16

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 off 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

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

- 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.

17

3.5   Influential observations and outliers

The OLS estimator of the ith element of the vector β_0 is simply

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

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 detemined by the observations on the regressors, some observations may have more influence than others. Define

$$
\begin{aligned}
h_t &= (P_X)_{tt} \\
&= e_t' P_X e_t \\
&= \| P_X e_t \|^2 \leq \| e_t \|^2 = 1
\end{aligned}
$$

h_t is the tth element on the main diagonal of P_X (e_t is a n vector of zeros with a 1 in the tth position). So $0 \leq h_t \leq 1$, and

$$
TrP_X = K \Rightarrow \bar{h} = K/n.
$$

A better method is as follows. Consider estimation of β without using the tth observation (designate this estimator as $\hat{\beta}^{(t)}$). One can show (see Davidson and MacKinnon, pp. 32-5 for proof) that

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

18

so the change in the tth observations fitted value is

$$X_t\hat{\beta} - X_t\hat{\beta}^{(t)} = \left(\frac{h_t}{1 - h_t}\right)\hat{\varepsilon}_t$$

While and 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 $\left(\frac{h_t}{1-h_t}\right)\hat{\varepsilon}_t$ as a function of t.

After influential observations are detected, one needs to determine why they are influential. A common cause is a data entry error, which can easily be corrected. If the data is correct then there may be some special economic factors that affect some observations. These would need to be identified and incorporated in the model. Another possibility is that pure randomness caused us to sample a low-probability observation.

There exist *robust* estimation methods that downweight outliers.

3.6   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

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

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 *(show with the one regressor, two observation example).*

- The uncentered R^2 changes if we add a constant to y, since this changes ϕ. Another, more common definition measures the contribution of the variables, other than the constant term, to explaining the variation in y.

- 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.

The *centered* R_c^2 is defined as

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

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

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

20

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

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

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

3.7 Small sample properties of the least squares estimator

3.7.1 Unbiasedness

For $\hat{\beta}$ we have

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

For $\hat{\sigma}^2$ we have

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

$$= \frac{1}{n-K}\varepsilon'M\varepsilon$$

$$\mathcal{E}(\widehat{\sigma_0^2}) = \frac{1}{n-K}\mathcal{E}(Tr\varepsilon'M\varepsilon)$$

$$= \frac{1}{n-K}\mathcal{E}(TrM\varepsilon\varepsilon')$$

$$= \frac{1}{n-K}Tr\mathcal{E}(M\varepsilon\varepsilon')$$

$$= \frac{1}{n-K}\sigma_0^2 TrM$$

$$= \frac{1}{n-K}\sigma_0^2\left(n - TrX(X'X)^{-1}X'\right)$$

$$= \frac{1}{n-K}\sigma_0^2\left(n - Tr(X'X)^{-1}X'X\right)$$

$$= \sigma_0^2$$

3.7.2 Normality

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

This is a linear function of ε, which is normally distributed. Therefore

$$\hat{\beta} \sim N\left(\beta_0, (X'X)^{-1}\sigma_0^2\right)$$

3.7.3 Efficiency (Gauss-Markov theorem)

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

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

$$= Cy$$

It is also *unbiased*, as we proved above. One could consider other weights W in place of the OLS weights. We'll still insist upon unbiasedness. Consider $\tilde{\beta} = Wy$. If the estimator is unbiased

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

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

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

This is a proof of the Gauss-Markov Theorem.

Theorem 1 (Gauss-Markov) *Under the classical assumptions, the variance of any linear unbiased estimator minus the variance of the OLS estimator is a positive semidefinite matrix.*

- It is worth noting 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.

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.

4 Maximum likelihood estimation

4.1 The likelihood function

Suppose a sample of size n of a random vector y. Suppose the joint density of $Y = \begin{pmatrix} y_1 & \ldots & y_n \end{pmatrix}$ is characterized by a parameter vector θ_0 :

$$f_{\mathcal{Y}}(Y, \theta_0).$$

This will often be referred to using the simplified notation $f(\theta_0)$.

The *likelihood function* is just this density evaluated at other values θ

$$L(Y, \theta) = f_{\mathcal{Y}}(Y, \theta), \theta \in \Theta,$$

where Θ is a *parameter space.*

- If the n observations are independent, the likelihood function can be written as

$$L(Y, \theta) = \prod_{t=1}^{n} f(y_t, \theta)$$

  where the f_t are possibly of different form.

- Even 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)

$$L(Y, \theta) = f(y_1, \theta) f(y_2 | y_1, \theta) f(y_3 | y_1, y_2, \theta) \cdots f(y_n | y_1, y_2, \ldots y_{t-n}, \theta)$$

25

To simplify notation, define

$$\begin{aligned} x_t &= \{y_1, y_2, ..., y_{t-1}\}, t \geq 2 \\ &= \mathcal{S}, t = 1 \end{aligned}$$

where $\mathcal{S}$ is the sample space of Y. (With this, conditioning on x_1 has no effect and gives a marginal probability). 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 is defined 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 $\hat{\theta}$.

Note that one can easily modify this to include exogenous conditioning variables in x_t in addition to the y_t that are already there. This changes nothing in what follows, and therefore it is suppressed to clarify the notation.

4.2 Consistency of MLE

To show consistency of the MLE, we need to make explicit some assumptions.

Compact parameter space $\theta \in \Theta$, a open bounded subset of $\mathfrak{R}^K$. Maximixation is

over $\overline{\Theta}$, which is compact.

This implies that θ is an interior point of the *parameter space* $\overline{\Theta}$.

Uniform convergence

$$s_n(\theta) \overset{u.a.s}{\to} \lim_{n \to \infty} \mathcal{E}_{\theta_0} s_n(\theta) \equiv s_\infty(\theta, \theta_0), \forall \theta \in \overline{\Theta}.$$

We have suppressed Y here for simplicity. This requires that almost sure convergence holds for all possible parameter values.

Continuity $s_n(\theta)$ is continuous in $\theta, \theta \in \overline{\Theta}$. This implies that $s_\infty(\theta, \theta_0)$ is continuous in θ.

Identification $s_\infty(\theta, \theta_0)$ has a unique maximum in its first argument.

We will use these assumptions to show that $\hat{\theta} \overset{a.s.}{\to} \theta_0$.

First, $\hat{\theta}$ certainly exists, since a continuous function has a maximum on a compact set.

Second, 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. Therefore, since $\ln(1) = 0$,

$$\mathcal{E}\left(\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right) \leq 0,$$

or

$$\mathcal{E}\left(s_n\left(\theta\right)\right) - \mathcal{E}\left(s_n\left(\theta_0\right)\right) \leq 0.$$

Taking limits, this is

$$s_\infty(\theta, \theta_0) - s_\infty(\theta_0, \theta_0) \leq 0$$

except on a set of zero probability (by the uniform convergence assumption).

By the identification assumption there is a unique maximizer, so the inequality is strict if $\theta \neq \theta_0$:

$$s_\infty(\theta, \theta_0) - s_\infty(\theta_0, \theta_0) < 0, \forall \theta \neq \theta_0,$$

However, since $\hat{\theta}$ is a maximizer, independent of n, we must have

$$s_\infty(\hat{\theta}, \theta_0) - s_\infty(\theta_0, \theta_0) \geq 0.$$

These last two inequalities imply that

$$\lim_{n \to \infty} \hat{\theta} = \theta_0, \text{ a.s.}$$

This completes the proof of strong consistency of the MLE. One can use weaker assumptions to prove weak consistency (convergence in probability to θ_0) of the MLE. This is omitted here. Note that almost sure convergence implies convergence in probability.

4.3 The score function

Differentiability Assume that $s_n(\theta)$ is twice continuously differentiable in $N(\theta_0)$, 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 will often be suppressed for clarity, but one should not forget that it is 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 $\mathcal{E}_\theta[g_t(\theta)] = 0$, $\forall t$. *This is the expectation taken with respect to the density $f(\theta)$, not necessarily $f(\theta_0)$.*

$$
\begin{aligned}
\mathcal{E}_\theta[g_t(\theta)] &= \int [D_\theta \ln f(y_t|x,\theta)] f(y_t|x,\theta) dy_t \\
&= \int \frac{1}{f(y_t|x_t,\theta)} [D_\theta f(y_t|x_t,\theta)] f(y_t|x_t,\theta) dy_t \\
&= \int D_\theta f(y_t|x_t,\theta) dy_t.
\end{aligned}
$$

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

$$
\begin{aligned}
\mathcal{E}_\theta[g_t(\theta)] &= D_\theta \int ft(y_t|x_t,\theta) dy_t \\
&= D_\theta 1 \\
&= 0.
\end{aligned}
$$

29

- 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 $\mathcal{E}_\theta g_n(Y, \theta) = 0$.

4.4   Asymptotic normality of MLE

Recall that we assume that $s_n(\theta)$ 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 appropirate definitions

$$H(\theta^*) \left(\hat{\theta} - \theta_0\right) = -g(\theta_0),$$

where $\theta^* = \lambda \hat{\theta} + (1 - \lambda)\theta_0, 0 < \lambda < 1$. Assume $H(\theta^*)$ is invertible (we'll justify this in a minute). So

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -H(\theta^*)^{-1}\sqrt{n}g(\theta_0)$$

Now consider $H(\theta^*)$. This is

$$
\begin{aligned}
H(\theta^*) \;&=\; D_{\theta'} g(\theta^*) \\
&=\; D_\theta^2 s_n(\theta^*) \\
&=\; \frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f t(\theta^*)
\end{aligned}
$$

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 $\theta^* \overset{a.s.}{\to} \theta_0$. Given this, $H(\theta^*)$ converges to the limit of it's expectation:

$$H(\theta^*) \overset{a.s.}{\to} \lim_{n\to\infty} \mathcal{E}\left(D_\theta^2 s_n(\theta_0)\right) = H_\infty(\theta_0) < \infty$$

This matrix converges to a finite limit.

Re-arranging orders of limits and differentiation, which is legitimate given regularity conditions, we get

$$
\begin{aligned}
H_\infty(\theta_0) &= D_\theta^2 \lim_{n\to\infty} \mathcal{E}\left(s_n(\theta_0)\right) \\
&= D_\theta^2 s_\infty(\theta_0, \theta_0)
\end{aligned}
$$

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 $H_\infty(\theta_0)$ must be negative definite, and therefore of full rank. Therefore

the previous inversion is justified, asymptotically, and we have

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{a.s.}{\to} -H_{\infty}(\theta_0)^{-1}\sqrt{n}g(\theta_0). \tag{1}$$

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 $g_n(\theta_0) \overset{a.s.}{\to} 0$, 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,

$$V(X_n)^{-1/2}\left(X_n - E(X_n)\right) \overset{d}{\to} N(0,I)$$

where $V(X_n)^{1/2}$ is any matrix such that

$$\left(V(X_n)^{1/2}\right)\left(V(X_n)^{1/2}\right)' = V(X_n).$$

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

$$I_\infty(\theta_0)^{-1/2}\sqrt{n}g_n(\theta_0) \xrightarrow{d} N[0, I_K]$$

where

$$
\begin{aligned}
I_\infty(\theta_0) &= \lim_{n\to\infty} \mathcal{E}_{\theta_0}\left(n\left[g_n(\theta_0)\right]\left[g_n(\theta_0)\right]'\right) \\
&= \lim_{n\to\infty} V_{\theta_0}\left(\sqrt{n}g_n(\theta_0)\right)
\end{aligned}
$$

This can also be written as

$$\sqrt{n}g_n(\theta_0) \xrightarrow{d} N[0, I_\infty(\theta_0)] \tag{2}$$

- $I_\infty(\theta_0)$ is known as the *information matrix*.

- Combining [1] and [2], we get

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = N\left[0, H_\infty(\theta_0)^{-1} I_\infty(\theta_0) H_\infty(\theta_0)^{-1}\right].$$

The MLE estimator is asymptotically normally distributed.

Definition 2 (CAN) *An estimator $\hat{\theta}$ of a parameter θ_0 is $\sqrt{n}$-consistent and asymptotically normally distributed if*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(0, V_\infty) \tag{3}$$

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 normally, $\sqrt{n}$ is the highest factor that we can multiply by an still get convergence to a stable limiting distribution.

Definition 3 (Asymptotic unbiasedness) *An estimator $\hat{\theta}$ of a parameter θ_0 is asymptotically unbiased if*

$$\lim_{n \to \infty} \mathcal{E}_\theta(\hat{\theta}) = \theta. \tag{4}$$

Estimators that are CAN are asymptotically unbiased, though not all consistent estimators are asymptotically unbiased. Such cases are unusual, though. An example is:

Exercise 4 *Consider an estimator $\hat{\theta}$ with distribution*

$$\hat{\theta} = \left\{ \theta_0 \text{ with prob. } 1 - \frac{1}{n} n \text{ with prob. } \frac{1}{n} \right. . \tag{5}$$

Show that this estimator is consistent but asymptotically biased.

4.5 The information matrix equality

We will show that $H_\infty(\theta) = -I_\infty(\theta)$. 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:

$$
\begin{aligned}
0 &= \int \left[D_\theta^2 \ln f_t(\theta) \right] f_t(\theta) dy + \int \left[D_\theta \ln f_t(\theta) \right] D_{\theta'} f_t(\theta) dy \\
&= \mathcal{E}_\theta \left[D_\theta^2 \ln f_t(\theta) \right] + \int \left[D_\theta \ln f_t(\theta) \right] \left[D_\theta \ln f_t(\theta) \right]' f_t(\theta) dy \\
&= \mathcal{E}_\theta \left[D_\theta^2 \ln f_t(\theta) \right] + \mathcal{E}_\theta \left[D_\theta \ln f_t(\theta) \right] \left[D_\theta \ln f_t(\theta)' \right] \\
&= \mathcal{E}_\theta \left[H_t(\theta) \right] + \mathcal{E}_\theta \left[g_t(\theta) \right] \left[g_t(\theta) \right]'
\end{aligned}
$$

Now sum over n and multiply by $\frac{1}{n}$

$$
\mathcal{E}_\theta \frac{1}{n} \sum_{t=1}^n \left[H_t(\theta) \right] = - \mathcal{E}_\theta \left[\frac{1}{n} \sum_{t=1}^n \left[g_t(\theta) \right] \left[g_t(\theta) \right]' \right]
$$

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

$$
\mathcal{E}_\theta \left[H(\theta) \right] = - \mathcal{E}_\theta \left(n \left[g(\theta) \right] \left[g(\theta) \right]' \right)
$$

since all cross products between different periods expect to zero. Finally take limits, we get

$$
H_\infty(\theta) = - I_\infty(\theta).
$$

This holds for all θ, in particular, for θ_0. Using this,

$$
\sqrt{n} \left(\hat{\theta} - \theta_0 \right) \stackrel{a.s.}{\to} N \left[0, H_\infty(\theta_0)^{-1} I_\infty(\theta_0) H_\infty(\theta_0)^{-1} \right]
$$

simplifies to

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{a.s.}{\to} N\left[0, I_\infty(\theta_0)^{-1}\right]$$

To estimate the asymptotic variance, we need estimators of $H_\infty(\theta_0)$ and $I_\infty(\theta_0)$. We can use

$$\widehat{I_\infty(\theta_0)} = n\sum_{t=1}^{n} g_t(\hat{\theta})g_t(\hat{\theta})'$$
$$\widehat{H_\infty(\theta_0)} = H(\hat{\theta}).$$

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

$$\widehat{V_\infty(\theta_0)} = -\widehat{H_\infty(\theta_0)}^{-1}$$
$$\widehat{V_\infty(\theta_0)} = \widehat{I_\infty(\theta_0)}^{-1}$$
$$\widehat{V_\infty(\theta_0)} = \widehat{H_\infty(\theta_0)}^{-1}\widehat{I_\infty(\theta_0)}\widehat{H_\infty(\theta_0)}^{-1}$$

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.

4.6 The Cramér-Rao lower bound

Theorem 5 (Cramer-Rao Lower Bound) *The limiting variance of a CAN estimator, $\tilde{\theta}$, of θ_0 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

$$\lim_{n\to\infty} \int \sqrt{n}\left(\tilde\theta - \theta\right)\sqrt{n}\frac{1}{n}\left[D_{\theta'}\ln f(\theta)\right]f(\theta) dy = I_K$$

But $\frac{1}{n}D_{\theta'}\ln f(\theta)$ 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,

$$V_\infty \begin{bmatrix} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n}g(\theta) \end{bmatrix} = \begin{bmatrix} V_\infty(\tilde{\theta}) & I_K \\ I_K & I_\infty(\theta) \end{bmatrix}.$$

Since this is a covariance matrix, it is positive semi-definite. Therefore, for any K-vector α,

$$\begin{bmatrix} \alpha' & -\alpha' I_\infty^{-1}(\theta) \end{bmatrix} \begin{bmatrix} V_\infty(\tilde{\theta}) & I_K \\ I_K & I_\infty(\theta) \end{bmatrix} \begin{bmatrix} \alpha \\ -I_\infty(\theta)^{-1}\alpha \end{bmatrix} \geq 0.$$

This simplifies to

$$\alpha'\left(V_\infty(\tilde{\theta}) - I_\infty^{-1}(\theta)\right)\alpha \geq 0.$$

Since α is arbitrary, $V_\infty(\tilde{\theta}) - I_\infty(\theta)$ is positive semidefinite. This conludes the proof.

This means that $I_\infty^{-1}(\theta)$ is a *lower bound* for the asymptotic variance of a CAN estimator.

Definition 6 (Asymptotic Efficiency) *An estimator is $\hat{\theta}$ of a parameter θ_0 is asymptotically efficient if it is CAN and $V_\infty(\tilde{\theta}) - V_\infty(\hat{\theta})$ is positive semidefinite for $\tilde{\theta}$ any other CAN estimator of θ_0.*

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*.

Summary of MLE

- Consistent

- Asymptotically normal (CAN)

- Asymptotically efficient

- Asymptotically unbiased

- This is for general MLE: we haven't specified the distribution or the linearity/nonlinearity of the estimator

5 Asymptotic properties of the least squares estimator

5.1 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 $\frac{X'\varepsilon}{n}$,

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

$$V(x_t\varepsilon_t) = x_t x_t' \sigma_0^2,$$

and

$$\mathcal{E}(x_t\varepsilon_t\varepsilon_s x_s') = 0, t \neq s.$$

So the sum is a sum of independent, nonidentically distributed random variables, each with mean zero. Supposing that $V(x_t\varepsilon_t) < \infty, \forall t$, the KLLN implies

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

This implies that

$$\hat{\beta} \overset{a.s.}{\to} \beta_0.$$

This is the property of *strong consistency:* the estimator converges almost surely to the true value. If we has used a weak LLN (defined in terms of convergence in probability), we would have *(simple, weak) consistency.*

- The consistency proof does not use the normality assumption.

5.2 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:

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

- 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

$$
\begin{aligned}
\lim_{n\to\infty} V\left(\frac{X'\varepsilon}{\sqrt{n}}\right) &= \sigma_0^2 \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \\
&= \sigma_0^2 Q_X
\end{aligned}
$$

since cross-terms expect to zero by the assumption of uncorrelated errors.

- The mean is of course zero. This term is a sum of nonidentically, uncorrelated but possibly dependent terms, each with mean zero, weighted by $\sqrt{n}$. Apply-

ing the Lindeberg-Feller CLT for nonidentically but independently distributed random vectors:

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

Therefore,

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

- 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.

5.3 Asymptotic efficiency

The least squares objective function is

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

Supposing that ε is normally distributed, the model is

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

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

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,

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

It's clear that the fonc for the MLE of β_0 are the same as the fonc for OLS (up to multiplication by a constant), so *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.*

As we'll see later, it will be possible to use 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.

6 Restrictions and hypothesis tests

6.1 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

$$k^0 \ln q = \beta_0 + \beta_1 \ln k p_1 + \beta_2 \ln k p_2 + \beta_3 \ln k m + \varepsilon,$$

so

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

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.

6.1.1 Imposition

The general formulation of linear equality restrictions is the model

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

$$R\beta = r$$

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 thing less messy. The fonc are

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

$$D_{\lambda} s(\hat{\beta}, \hat{\lambda}) = R\hat{\beta}_R - r \equiv 0,$$

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

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

Aside: Stepwise Inversion

Note that

$$
\begin{bmatrix} (X'X)^{-1} & 0 \\ -R(X'X)^{-1} & I_Q \end{bmatrix} \begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \equiv AB
$$

$$
= \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -R(X'X)^{-1}R' \end{bmatrix}
$$

$$
\equiv \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -P \end{bmatrix}
$$

$$
\equiv C,
$$

and

$$
\begin{bmatrix} I_K & (X'X)^{-1}R'P^{-1} \\ 0 & -P^{-1} \end{bmatrix} \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -P \end{bmatrix} \equiv DC
$$

$$
= I_{K+Q},
$$

so

$$DAB = I_{K+Q}$$

$$DA = B^{-1}$$

$$B^{-1} = \begin{bmatrix} I_K & (X'X)^{-1}R'P^{-1} \\ 0 & -P^{-1} \end{bmatrix} \begin{bmatrix} (X'X)^{-1} & 0 \\ -R(X'X)^{-1} & I_Q \end{bmatrix}$$

$$= \begin{bmatrix} (X'X)^{-1} - (X'X)^{-1}R'P^{-1}R(X'X)^{-1} & (X'X)^{-1}R'P^{-1} \\ P^{-1}R(X'X)^{-1} & -P^{-1} \end{bmatrix},$$

so

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

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

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

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, $Var(Ax+b) = AVar(x)A'$.

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

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

$$\begin{bmatrix} R_1 & R_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = r$$

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

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

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

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

6.1.2 Properties of the restricted estimator

We have that

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

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

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

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.

6.2 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.

6.2.1 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 7

$$\frac{N(0,1)}{\sqrt{\frac{\chi^2(q)}{q}}} \sim t(q) \tag{6}$$

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

We need a few results on the ^{2} distribution.

Proposition 8 *If $x \sim N(\mu, I_n)$ is a vector of n independent r.v.'s, then*

$$x'x \sim \chi^2(n, \lambda) \tag{7}$$

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 9 *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 PP' (this is the Cholesky factorization). Then consider $y = P'x$. We have

$$y \sim N(0, P'VP)$$

but

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

so $PVP' = I_n$.

$$
\begin{aligned}
y &\sim N(0, I_n) \\
y'y &= x'PP'x \\
&= xV^{-1}x \\
&\sim \chi^2(n)
\end{aligned}
$$

A more general proposition which implies this result is

Proposition 10 *If the n dimensional random vector $x \sim N(0,V)$, then*

$$x'Bx \sim \chi^2(\rho(B)) \tag{8}$$

if and only if BV is idempotent.

An immediate consequence is

Proposition 11 *If the random vector (of dimension n) $x \sim N(0,I)$, and B is idempotent with rank r, then*

$$x'Bx \sim \chi^2(r). \tag{9}$$

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 12 *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.

6.2.2 F test

The F test allows testing multiple restrictions jointly.

Proposition 13 *If $x \sim \chi^2(r)$ and $y \sim \chi^2(s)$, then*

$$\frac{x/r}{y/s} \sim F(r,s) \tag{10}$$

provided that x and y are independent.

Proposition 14 *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 ^{2} distribution, it is simple to show that the following statistic has the F distribution:

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

A numerically equivalent expression is

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

- **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.

6.2.3 Wald-type tests

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 [9]

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

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.

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

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

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

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,

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

56

or

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

since the n's cancel and inserting the limit of a matrix of constants changes nothing.

However,

$$
\begin{aligned}
\lim nP &= \lim nR(X'X)^{-1}R' \\
&= \lim R\left(\frac{X'X}{n}\right)^{-1}R' \\
&= RQ_X^{-1}R'
\end{aligned}
$$

So there is a cancellation and we get

$$\sqrt{n}\hat{\lambda} \xrightarrow{d} N\left(0, \sigma_0^2 \lim nP^{-1}\right)$$

In this case,

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

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.

6.2.5 Likelihood ratio-type tests

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 $\hat{\theta}$:

$$\ln L(\tilde{\theta}) \simeq \ln L(\hat{\theta}) + \frac{n}{2}\left(\tilde{\theta} - \hat{\theta}\right)' H(\hat{\theta})\left(\tilde{\theta} - \hat{\theta}\right)$$

(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 $H(\theta)$ is defined in terms of $\frac{1}{n}\ln L(\theta)$) so

$$LR \simeq -n\left(\tilde{\theta} - \hat{\theta}\right)' H(\hat{\theta})\left(\tilde{\theta} - \hat{\theta}\right)$$

As $n \to \infty, H(\hat{\theta}) \to H_\infty(\theta_0) = -I(\theta_0)$, by the information matrix equality. So

$$LR \overset{a}{=} n\left(\tilde{\theta} - \hat{\theta}\right)' I_\infty(\theta_0)\left(\tilde{\theta} - \hat{\theta}\right)$$

We also have that, from [**??**] that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{a}{=} I_\infty(\theta_0)^{-1} n^{1/2} g(\theta_0).$$

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) :

$$\sqrt{n}\left(\tilde{\theta}-\theta_0\right) \overset{a}{=} I_\infty(\theta_0)^{-1}\left(I_n - R'\left(RI_\infty(\theta_0)^{-1}R'\right)^{-1}RI_\infty(\theta_0)^{-1}\right)n^{1/2}g(\theta_0).$$

Combining the last two equations

$$\sqrt{n}\left(\tilde{\theta}-\hat{\theta}\right) \overset{a}{=} -n^{1/2}I_\infty(\theta_0)^{-1}R'\left(RI_\infty(\theta_0)^{-1}R'\right)^{-1}RI_\infty(\theta_0)^{-1}g(\theta_0)$$

so, substituting into [**??**]

$$LR \overset{a}{=} \left[n^{1/2}g(\theta_0)'I_\infty(\theta_0)^{-1}R'\right]\left[RI_\infty(\theta_0)^{-1}R'\right]^{-1}\left[RI_\infty(\theta_0)^{-1}n^{1/2}g(\theta_0)\right]$$

But since

$$n^{1/2}g(\theta_0) \overset{d}{\to} N\left(0, I_\infty(\theta_0)\right)$$

the linear function

$$RI_\infty(\theta_0)^{-1}n^{1/2}g(\theta_0) \overset{d}{\to} N(0, RI_\infty(\theta_0)^{-1}R').$$

We can see that LR is a quadratic form of this rv, with the inverse of its variance in the middle, so

$$LR \overset{d}{\to} \chi^2(q).$$

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

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

60

equivalent to

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

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 [**??**] 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.

61

Now consider the likelihood ratio statistic

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

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:

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

so

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

62

Substituting these last expressions into [**??**], we get

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

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

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

- 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 samples. The numeric values of the tests also depend upon how σ^2 is esti-

mated, 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.

6.4 Interpretation of test statistics

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

6.5 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. *Draw a picture here.*

- 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:

65

– Rejection of hypotheses individually does not imply that the joint test will reject.

– Joint rejection does not imply individal tests will reject.

6.6 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

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

$$\varepsilon \quad \sim \quad IID(0, \sigma_0^2)$$

$$X \text{ is nonstochastic}$$

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 (y, x) 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.

67

- 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.

6.7 Testing nonlinear restrictions

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

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

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

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

under the null hypothesis. Substituting consistent estimators for β_0, Q_X and σ_0^2, the resulting statistic is

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

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 non-linear 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

$$\mathcal{E}(x) = \frac{\partial f(x)}{\partial x}\frac{x}{f(x)}$$

where I'm using element-by-element multiplication and division. Suppose we estimate a linear function

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

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

$$\eta_i(x) = \frac{\beta_i}{x'\beta}x_i$$

The estimator of the ith elasticity is

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

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

$$R_i(\beta) = \frac{\partial \eta_i(x)}{\partial \beta'}$$

$$= \frac{[\,0\ \ 0\ \ 0\ \ x_i\ \ 0\,]x'\beta - x(x_i\beta_i)}{(x'\beta)^2}$$

to obtain the i th row of $R(\beta)$, and apply the above formula. Note that the elasticity and the standard error are functions of x.

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

$$0 \ \leq s_i(p,m) \leq 1, \forall i$$

$$\sum_{i=1}^{G} s_i(p,m) \qquad = 1$$

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

71

specification.

7 Generalized least squares

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. The model is

$$
\begin{aligned}
y &= X\beta + \varepsilon \\
\mathcal{E}(\varepsilon) &= 0 \\
V(\varepsilon) &= \Sigma \\
\mathcal{E}(X'\varepsilon) &= 0
\end{aligned}
$$

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*.

- 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*.

- 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.

7.1 Effects of nonspherical disturbances on the OLS estimator

The least square estimator is

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

- Conditional on X, or supposing that X is independent of ε, we have unbiasedness, as before.

- The variance of $\hat{\beta}$, supposing X is nonstochastic, is

$$
\begin{aligned}
\mathcal{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] &= \mathcal{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \\
&= (X'X)^{-1}X'\Sigma X(X'X)^{-1}
\end{aligned}
$$

Due to this, any test statistic that is based upon $\widehat{\sigma^2}$ or the probability limit $\widehat{\sigma^2}$ of is invalid. 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, conditional on X

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

The problem is that Σ is unknown in general, so this distribution won't be useful

for testing hypotheses.

- Without normality, and unconditional on X 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

$$
\lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon'X}{n}\right) = \Omega
$$

so we obtain $\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right)$

Summary: OLS with heteroscedasticity and/or autocorrelation is:

- unbiased in the same circumstances in which the estimator is unbiased with iid errors

- 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.

7.2 The GLS estimator

Suppose Σ were known. Then one could form the Cholesky decomposition

$$
PP' = \Sigma^{-1}
$$

We have

$$PP'\Sigma = I_n$$

so

$$P'\left(P\Sigma P'\right) = P',$$

which implies that

$$P'\Sigma P = I_n$$

Consider the model

$$P'y = P'X\beta + P'\varepsilon,$$

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

$$
\begin{aligned}
y^* &= X^*\beta + \varepsilon^* \\
\mathcal{E}(\varepsilon^*) &= 0 \\
V(\varepsilon^*) &= I_n \\
\mathcal{E}(X^{*\prime}\varepsilon^*) &= 0
\end{aligned}
$$

satisfies the classical assumptions (with modifications to allow stochastic regressors

and nonnormality of ε). The GLS estimator is simply OLS applied to the transformed model:

$$
\begin{aligned}
\hat{\beta}_{GLS} &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \\
&= (X'PP'X)^{-1}X'PP'y \\
&= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y
\end{aligned}
$$

The GLS estimator is unbiased in the same circumstances under which the OLS estimator is unbiased. For example, assuming X is nonstochastic

$$
\begin{aligned}
\mathcal{E}(\hat{\beta}_{GLS}) &= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y\right\} \\
&= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(X\beta+\varepsilon)\right\} \\
&= \beta.
\end{aligned}
$$

The variance of the estimator, conditional on X can be calculated using

$$
\begin{aligned}
\hat{\beta}_{GLS} &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \\
&= (X^{*\prime}X^*)^{-1}X^{*\prime}(X^*\beta+\varepsilon^*) \\
&= \beta + (X^{*\prime}X^*)^{-1}X^{*\prime}\varepsilon^*
\end{aligned}
$$

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.

- 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, not that

$$Var(\hat{\beta}) - Var(\hat{\beta}_{GLS}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1} - (X'\Sigma^{-1}X)^{-1}$$
$$=$$

- As one can verify by calculating fonc, 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.

7.3 Feasible GLS

The problem is that Σ isn't known usually, so this estimator isn't available.

- Consider the dimension of Σ: it's an $n \times n$ matrix with $\left(n^2 - n\right)/2 + n = \left(n^2 + n\right)/2$ unique elements.

- 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 $\Sigma(X, \theta)$ is a continuous function 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.

7.4   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 ith 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.

7.4.1 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

$$n\left(R\hat{\beta} - r\right)'\left(R\left(\frac{X'X}{n}\right)^{-1}\hat{\Omega}\left(\frac{X'X}{n}\right)^{-1}R'\right)^{-1}\left(R\hat{\beta} - r\right) \overset{a}{\sim} \chi^2(q)$$

7.4.2 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

$$\frac{\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^1}{\sigma^2} = \frac{\varepsilon^{1\prime}M^1\varepsilon^1}{\sigma^2} \overset{d}{\to} \chi^2(n_1 - K)$$

and

$$\frac{\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^3}{\sigma^2} = \frac{\varepsilon^{3\prime}M^3\varepsilon^3}{\sigma^2} \overset{d}{\to} \chi^2(n_3 - K)$$

so

$$\frac{\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^1/(n_1 - K)}{\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^3/(n_3 - K)} \overset{d}{\to} F(n_1 - K, n_3 - K).$$

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 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

$$qF = \frac{P(ESS_R - ESS_U)/P}{ESS_U/(n - P - 1)}$$

83

Note that $ESS_R = TSS_U$, so dividing both numerator and denominator by this we get

$$qF = (n - P - 1) \frac{R^2}{1 - R^2}$$

Note that this is the R^2 or 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.

7.4.3 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.

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

$$
\varepsilon_t^2 = \left(z_t'\gamma\right)^\delta + v_t
$$

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

$$
\hat{\sigma}_t^2 = \left(z_t'\hat{\gamma}\right)^{\hat{\delta}} \overset{p}{\to} \sigma_t^2 \; .
$$

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

$$y_{it} = x_{it}'\beta + \varepsilon_{it}$$
$$\mathcal{E}(\varepsilon_{it}^2) = \sigma_i^2, \forall t$$

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.

7.5 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.

7.5.1 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

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

Draw a picture here.

7.5.2 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

$$
\begin{aligned}
\varepsilon_t \quad &= \rho\varepsilon_{t-1} + u_t \\
&= \rho\left(\rho\varepsilon_{t-2} + u_{t-1}\right) + u_t \\
&= \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t \\
&= \rho^2\left(\rho\varepsilon_{t-3} + u_{t-2}\right) + \rho u_{t-1} + u_t
\end{aligned}
$$

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

$$
\begin{aligned}
\mathcal{E}(\varepsilon_t^2) \quad &= \sigma_u^2 \sum_{m=0}^{\infty} \rho^{2m} \\
&= \frac{\sigma_u^2}{1-\rho^2}
\end{aligned}
$$

- If we had directly assumed that ε_t were covariance stationary, we could obtain this using

$$
\begin{aligned}
V(\varepsilon_t) \quad &= \rho^2 \mathcal{E}(\varepsilon_{t-1}^2) + 2\rho\mathcal{E}(\varepsilon_{t-1}u_t) + \mathcal{E}(u_t^2) \\
&= \rho^2 V(\varepsilon_t) + \sigma_u^2,
\end{aligned}
$$

so

$$
V(\varepsilon_t) = \frac{\sigma_u^2}{1-\rho^2}
$$

- The variance is the 0th order autocovariance: $\gamma_0 = V(\varepsilon_t)$

- Note that the variance does not depend on t

Likewise, the first order autocovariance γ_1 is

$$
\begin{aligned}
Cov(\varepsilon_t, \varepsilon_{t-1}) \quad &= \gamma_s = E((\rho\varepsilon_{t-1} + u_t)\varepsilon_{t-1}) \\
&= \rho V(\varepsilon_t) \\
&= \frac{\rho\sigma_u^2}{1-\rho^2}
\end{aligned}
$$

- 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. This is consistent as long as $\frac{1}{n}X'\Sigma X$ converges to a finite limiting matrix. It turns out that this requires that the regressors X satisfy the previous stationarity conditions and that $|\rho| < 1$, which we have assumed.

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

92

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

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

- 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

$$
\begin{aligned}
y_1^* &= \sqrt{1 - \hat{\rho}^2}y_1 \\
x_1^* &= \sqrt{1 - \hat{\rho}^2}x_1
\end{aligned}
$$

This somewhat odd result is related to the Cholesky factorization of Σ^{-1}. See Davidson and MacKinnon, pg. 348-49 for more discussion. Note that the vari-

ance of y_1^* is σ_u^2, asymptotically, 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.

7.5.3 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

$$\gamma_2 = [(u_t + \phi u_{t-1})(u_{t-2} + \phi u_{t-3})]$$

$$= 0$$

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

$$\rho_1 = \frac{\phi \sigma_u^2}{\sigma_u^2(1+\phi^2)} = \frac{\gamma_1}{\gamma_0}$$

$$= \frac{\phi}{(1+\phi^2)}$$

- 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:

$$\widehat{\sigma_\varepsilon^2} = \widehat{\sigma_u^2(1 + \phi^2)} = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^2$$

- 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.

- 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:

$$\hat{\phi}\widehat{\sigma_u^2} = \frac{1}{n}\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}$$

- 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.

7.5.4 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

$$
\begin{aligned}
X'\varepsilon &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
&= \sum_{t=1}^{n} x_t \varepsilon_t \\
&= \sum_{t=1}^{n} m_t
\end{aligned}
$$

97

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

$$\Omega_n = \mathcal{E}\frac{1}{n}\left[\left(\sum_{t=1}^{n} m_t\right)\left(\sum_{t=1}^{n} m_t'\right)\right]$$

98

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, $\hat{\Omega}_n \overset{p}{\to} \Omega$. 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.

7.5.5 Testing for autocorrelation

Durbin-Watson test   The Durbin-Watson test statistic is

$$\begin{aligned} DW &= \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2} \\ &= \frac{\sum_{t=2}^n \left(\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t \hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2 \right)}{\sum_{t=1}^n \hat{\varepsilon}_t^2} \end{aligned}$$

- 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

100

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 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. *Picture here.* There are means of determining exact critical values conditional on X.

- Note that DW can be used to test for nonlinearity (add discussion).

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 $\chi^2(P)$.

101

- 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.

- 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.

7.5.6 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. A simple example is the case of a single lag of the dependent variable with AR(1) errors. The model is

$$
\begin{aligned}
y_t &= x_t'\beta + y_{t-1}\gamma + \varepsilon_t \\
\varepsilon_t &= \rho\varepsilon_{t-1} + u_t
\end{aligned}
$$

Now we can write

$$
\begin{aligned}
\mathcal{E}(y_{t-1}\varepsilon_t) &= \mathcal{E}\left\{(x_{t-1}'\beta + y_{t-2}\gamma + \varepsilon_{t-1})(\rho\varepsilon_{t-1} + u_t)\right\} \\
&\neq 0
\end{aligned}
$$

since one of the terms is $\mathcal{E}(\rho\varepsilon_{t-1}^2)$ which is clearly nonzero. In this case $\mathcal{E}(X'\varepsilon) \neq 0$, 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.

8 Stochastic regressors

Up until now we've assumed that the regressors are nonstochastic. This is highly unrealistic in the case of economic data.

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 with is

1. **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.

(a) **IID mean zero errors**:

$$\mathcal{E}(\varepsilon) = 0$$
$$\mathcal{E}(\varepsilon\varepsilon') = \sigma_0^2 I_n$$

(b) **Stochastic, linearly independent regressors**

- X has rank K with probability 1

- X is stochastic

- X is uncorrelated with ε : $\mathcal{E}(X'\varepsilon) = 0$.

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

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

(c) **Normality (Optional)**: ε is normally distributed

8.1 Case 1

Normality of ε, X independent of ε

In this case,

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

Due to of independence of X and ε

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

Conditional on X,

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

- 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 and uncorrelated with ε 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.

8.2 Case 2

ε *nonnormally distributed, independent of* X

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 stochastic regressors that are independent of ε, with ε normal or nonnormal, $\hat{\beta}$ has the properties:

  1. Unbiasedness

  2. Consistency

  3. Gauss-Markov theorem holds, since it holds in the previous case and doesn't

107

depend on normality.

4. Asymptotic normality

5. Tests are asymptotically valid, but are not valid in small samples.

8.3   Case 3

Lagged dependent variables (dynamic models).

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

$$
\begin{aligned}
y_t &= z_t'\alpha + \sum_{s=1}^{p} \gamma_s y_{t-s} + \varepsilon_t \\
&= x_t'\beta + \varepsilon_t \\
\varepsilon &\sim iid(0, \sigma_0^2 I_n)
\end{aligned}
$$

where now x_t contains lagged dependent variables. Clearly X and ε aren't independent anymore, so one can't show unbiasedness. For example, consider

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

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.

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

108

8.4　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

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

  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*):

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

One can show that the variance of x_t depends upon t in this case.

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.*

For application of central limit theorems, a useful concept is that of a *martingale difference sequence.* This is a sequence $\{z_t\}$ such that

$$\mathcal{E}(z_t|\Omega_{t-1}) = 0,$$

where Ω_t is the *information set in period t*. At a minimum, Ω_t includes all z_s for $s = 1, 2, ..., t$. Note that

$$x_t' \varepsilon_t$$

is a martingale difference sequence. Hamilton, Proposition 7.8 (pg. 193) gives a central limit theorem for covariance stationary martingale difference sequences.

- 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 econometrics of nonstationary processes has been an active area of research in the last two decades. The standard asymptotics don't apply in this case. This isn't in the scope of this course.

9 Data problems

In this section well consider problems associated with the regressor matrix: collinearity, missing observation and measurement error.

9.1 Collinearity

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 ith 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

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t$$

$$x_{2t} = \alpha_1 + \alpha_2 x_{3t}$$

then we can write

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

- 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 fonc). 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 ith 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.

9.1.1 A brief aside on dummy variables

Introduce a brief discussion of dummy variables here.

9.1.2 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. This is reflected in imprecise estimates, *i.e.*, estimates with high variances. *With economic data, collinearity is commonly encountered, and is often a severe problem.*

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

$$X = \left[\begin{array}{cc} \mathbf{x} & W \end{array} \right]$$

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 assumptions, is

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

Using the partition,

$$X'X = \left[\begin{array}{cc} \mathbf{x}'\mathbf{x} & \mathbf{x}'W \\ W'\mathbf{x} & W'W \end{array} \right]$$

and following a rule for partitioned inversion,

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

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

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

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.

9.1.3 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.

9.1.4 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 $\hat{\beta}$:

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

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

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

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 al 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 solution to the problem.

9.2 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?

9.2.1 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^* 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.

9.2.2 Error of measurement of the regressors

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

$$y_t = x_t^{*\prime}\beta + \varepsilon_t$$

$$x_t = x_t^* + v_t$$

$$v_t \sim iid(0, \Sigma_v)$$

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

$$y_t = (x_t - v_t)'\beta + \varepsilon_t$$

$$= x_t'\beta - v_t'\beta + \varepsilon_t$$

$$= x_t'\beta + \omega_t$$

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

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

$$= -\Sigma_v \beta$$

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

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

121

We have that

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

and

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

since X^* and V are independent, and

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

Likewise,

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

so

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

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.

9.3 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.

9.3.1 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

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

$$= [X_1'X_1 + X_2'X_2]^{-1} [X_1'y_1 + X_2'\hat{y}_2]$$

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, and 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

$$\hat{\beta} = [X_1'X_1 + X_2'X_2]^{-1} \left[X_1'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 \right]$$

$$= [X_1'X_1 + X_2'X_2]^{-1} X_1'X_1\hat{\beta}_1 + [X_1'X_1 + X_2'X_2]^{-1} X_2'X_2\hat{\beta}_2$$

$$\equiv A\hat{\beta}_1 + (I_K - A)\hat{\beta}_2$$

where

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

and we use

$$\left[X_1'X_1 + X_2'X_2\right]^{-1} X_2'X_2 \;=\; [X_1'X_1 + X_2'X_2]^{-1} \left[(X_1'X_1 + X_2'X_2) - X_1'X_1\right]$$

$$= I_K - [X_1'X_1 + X_2'X_2]^{-1} X_1'X_1$$

$$= I_K - A.$$

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 the this 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 $\hat{y}_2 = \bar{y}$ does not satisfy the condition and therefore leads to a biased estimator.

Exercise 15 *Formally prove this last statement.*

- One possibility that has been suggested (see Greene, page 275) is to estimate β using a first round estimation using only the complete observations

$$\hat{\beta}_1 = (X_1'X_1)^{-1} X_1'y_1$$

125

then use this estimate, $\hat{\beta}_1$, to predict y_2 :

$$\hat{y}_2 \quad = X_2\hat{\beta}_1$$
$$= X_2(X_1'X_1)^{-1}X_1'y_1$$

Now, the overall estimate is a weighted average of $\hat{\beta}_1$ and $\hat{\beta}_2$, just as above, but we have

$$\hat{\beta}_2 \quad = (X_2'X_2)^{-1}X_2'\hat{y}_2$$
$$= (X_2'X_2)^{-1}X_2'X_2\hat{\beta}_1$$
$$= \hat{\beta}_1$$

This shows that this suggestion is completely empty of content: the final estimator is the same as the OLS estimator using only the complete observations.

9.3.2 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 $V(\varepsilon) = 25$. The figure sampsel.ps (on web site) illustrates this.

9.3.3 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 that the constant, subtution of $\bar{x}$ for the missing x *does not lead to bias*. This is a special case that doesn't hold for $K > 2$.

127

Exercise 16 *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.

10   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

$$c = A w_1^{\beta_1} w_2^{\beta_2} q^{\beta_q} e^{\varepsilon}$$

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 transform, 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:

$$x_t \quad = x(z_t)$$
$$y_t \quad = x_t'\beta + \varepsilon_t$$

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.

10.1 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

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

so the regression model to fit is

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

- 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.

131

- 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.

10.1.1 The translog form

In spite of the fact that FFF's aren't really as flexible as they were originally claimed to be, 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 of 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

$$y \quad = \ln(c)$$
$$x \quad = \ln(\tfrac{z}{\bar{z}})$$
$$= \ln(z) - \ln(\bar{z})$$
$$y \ = \alpha + x'\beta + 1/2 x' \Gamma x + \varepsilon$$

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

$$\frac{\partial y}{\partial x} = \beta + \Gamma x$$

$$= \frac{\partial \ln(c)}{\partial \ln(z)} \text{(the other part of } x \text{ is constant)}$$

$$= \frac{\partial c}{\partial z} \frac{z}{c}$$

which is the elasticity of w 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

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

133

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

$$\ln(c) = \alpha + x'\beta + z'\delta + 1/2 \begin{bmatrix} x' & z \end{bmatrix} \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma'_{12} & \Gamma_{22} \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}$$

$$= \alpha + x'\beta + z'\delta + 1/2x'\Gamma_{11}x + x'\Gamma_{12}z + 1/2z^2\gamma_{22}$$

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

$$\Gamma_{11} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{12} & \gamma_{22} \end{bmatrix}$$

$$\Gamma_{12} = \begin{bmatrix} \gamma_{13} \\ \gamma_{23} \end{bmatrix}$$

$$\Gamma_{22} = \gamma_{33}.$$

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:

$$s_1 = \beta_1 + \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}z$$

$$s_2 = \beta_2 + \gamma_{12}x_1 + \gamma_{22}x_2 + \gamma_{13}z$$

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

135

(adding in error terms)

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

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

136

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 corre-
lated 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 exten-
sion 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.

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

where the symbol $\otimes$ indicates the *Kronecker product*. The Kronecker product of two

matrices A and B is

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & \ddots & & \vdots \\ \vdots & & & \\ a_{pq}B & \cdots & & a_{pq}B \end{bmatrix}.$$

Personally, I can never keep straight the roles of A and B.

10.1.2 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

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

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:

$$
y^* = X^* \theta + \varepsilon^*
$$

139

Considering the error covariance, we can define

$$\Sigma_0^* = Var \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

$$\Sigma^* = I_n \otimes \Sigma_0^*$$

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

$$\hat{P}_0 = Chol \left(\hat{\Sigma}_0^* \right)^{-1}$$

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

$$\hat{P}y^* = \hat{P}X^*\theta + \hat{P}\varepsilon^*$$

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:

$$\hat{P}_0 y_y^* = \hat{P}_0 X_t^* \theta + \hat{P} \varepsilon^*$$

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

$$\beta_1 + \beta_2 \qquad = 1$$
$$\sum_{i=1}^{3} \gamma_{ij} \quad = 0, j = 1, 2, 3.$$

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.

10.2 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

$$y_t = \alpha + x_t'\beta + 1/2 x_t'\Gamma x_t + \varepsilon$$

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

$$M_1 \quad : y = X\beta + \varepsilon$$
$$\varepsilon_t \quad \sim iid(0, \sigma_\varepsilon^2)$$
$$M_2 \quad : y = Z\gamma + \eta$$
$$\eta \quad \sim iid(0, \sigma_\eta^2)$$

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

$$M_1 \quad : y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t$$

$$M_2 \quad : y_t = \gamma_1 + \gamma_2 x_{2t} + \gamma_3 x_{4t} + \eta_t$$

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.

11 Exogeneity 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 and measurement error in the regressors. Another important case is that of simultaneous equations. The cause is different, but the effect is the same.

11.1 Simultaneous equations

Up until now our model is

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

where, for purposes of estimation we can treat X as fixed. This means that when estimating β we *condition* on X. When analyzing dynamic models, we're not interested in conditioning on X, as we saw in the section on stochastic regressors. Nevertheless, the OLS estimator obtained by treating X as fixed continues to have desirable asymptotic properties even in that case.

Simultaneous equations is a different prospect. An example of a simultaneous equation system is a simple supply-demand system:

$$\text{Demand: } q_t = \alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t}$$

$$\text{Supply: } q_t = \beta_1 + \beta_2 p_t + \varepsilon_{2t}$$

$$\mathcal{E}\left(\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} & \varepsilon_{2t} \end{bmatrix} \right) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \cdot & \sigma_{22} \end{bmatrix}$$

$$\equiv \Sigma, \forall t$$

The presumption is that q_t and p_t are jointly determined at the same time by the in-

146

tersection 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 :

$$\alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t} = \beta_1 + \beta_2 p_t + \varepsilon_{2t}$$

$$\beta_2 p_t - \alpha_2 p_t = \alpha_1 - \beta_1 + \alpha_3 y_t + \varepsilon_{1t} - \varepsilon_{2t}$$

$$p_t = \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}$$

Now consider whether p_t is uncorrelated with ε_{1t} :

$$\mathcal{E}(p_t \varepsilon_{1t}) = \mathcal{E}\left\{ \left(\frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2} \right) \varepsilon_{1t} \right\}$$

$$= \frac{\sigma_{11} - \sigma_{12}}{\beta_2 - \alpha_2}$$

Because of this correlation, 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

$$Y_t'\Gamma = X_t'B + E_t'$$

$$E_t \sim N(0,\Sigma), \forall t$$

$$\mathcal{E}(E_t E_s') = 0, t \neq s$$

147

We can stack all n observations and write the model as

$$Y\Gamma = XB + E$$

$$\mathcal{E}(X'E) = 0_{(K \times G)}$$

$$vec(E) \sim N(0, \Psi)$$

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

$$\Psi = \begin{bmatrix} \sigma_{11}I_n & \sigma_{12}I_n & \cdots & \sigma_{1G}I_n \\ & \sigma_{22}I_n & & \vdots \\ & & \ddots & \vdots \\ \cdot & & & \sigma_{GG}I_n \end{bmatrix}$$

$$= I_n \otimes \Sigma$$

- X may contain lagged endogenous and exogenous variables. These variables are *predetermined.*

148

- We need to define what is meant by "endogenous" and "exogenous" when classifying the current period variables.

11.2 Exogeneity

The model defines a *data generating process*. The model involves two sets of variables, Y_t and X_t, as well as a parameter vector

$$\theta = \begin{bmatrix} vec(\Gamma)' & vec(B)' & vec^*(\Sigma)' \end{bmatrix}'$$

- In general, without additional restrictions, θ is a $G^2 + GK + (G^2 - G)/2 + G$ dimensional vector. This is the parameter vector that were interested in estimating.

- In principle, there exists a joint density function for Y_t and X_t, which depends on a parameter vector ϕ. Write this density as

$$f_t(Y_t, X_t | \phi, I_t)$$

where I_t is the information set in period t. This includes lagged $Y_t's$ and lagged X_t 's of course. This can be factored into the density of Y_t conditional on X_t times the marginal density of X_t :

$$f_t(Y_t, X_t | \phi, I_t) = f_t(Y_t | X_t, \phi, I_t) f_t(X_t | \phi, I_t)$$

This is a general factorization, but is may very well be the case that not all parameters in ϕ affect both factors. So use ϕ_1 to indicate elements of ϕ that enter into the conditional density and write ϕ_2 for parameters that enter into the

149

marginal. In general, ϕ_1 and ϕ_2 may share elements, of course. We have

$$f_t(Y_t, X_t | \phi, I_t) = f_t(Y_t | X_t, \phi_1, I_t) f_t(X_t | \phi_2, I_t)$$

- Recall that the model is

$$
\begin{aligned}
Y_t' \Gamma &= X_t' B + E_t' \\
E_t &\sim N(0, \Sigma), \forall t \\
\mathcal{E}(E_t E_s') &= 0, t \neq s
\end{aligned}
$$

Normality and lack of correlation over time imply that the observations are independent of one another, so we can write the log-likelihood function as the sum of likelihood contributions of each observation:

$$
\begin{aligned}
\ln L(Y | \theta, I_t) &= \sum_{t=1}^n \ln f_t(Y_t, X_t | \phi, I_t) \\
&= \sum_{t=1}^n \ln \left(f_t(Y_t | X_t, \phi_1, I_t) f_t(X_t | \phi_2, I_t) \right) \\
&= \sum_{t=1}^n \ln f_t(Y_t | X_t, \phi_1, I_t) + \sum_{t=1}^n \ln f_t(X_t | \phi_2, I_t) =
\end{aligned}
$$

Definition 17 (Weak Exogeneity) X_t *is weakly exogeneous for θ (the original parameter vector) if there is a mapping from ϕ to θ that is invariant to ϕ_2. More formally, for an arbitrary (ϕ_1, ϕ_2), $\theta(\phi) = \theta(\phi_1)$.*

This implies that ϕ_1 and ϕ_2 cannot share elements if X_t is weakly exogenous, since ϕ_1 would change as ϕ_2 changes, which prevents consideration of arbitrary combinations of (ϕ_1, ϕ_2).

Supposing that X_t is weakly exogenous, then the MLE of ϕ_1 using the joint density is the same as the MLE using only the conditional density

$$\ln L(Y|X, \theta, I_t) = \sum_{t=1}^{n} \ln f_t(Y_t|X_t, \phi_1, I_t)$$

since the conditional likelihood doesn't depend on ϕ_2. In other words, the joint and conditional log-likelihoods maximize at the same value of ϕ_1.

- With weak exogeneity, knowledge of the DGP of X_t is irrelevant for inference on ϕ_1, and knowledge of ϕ_1 is sufficient to recover the parameter of interest, θ. Since the DGP of X_t is irrelevant, we can treat X_t as fixed in inference.

- By the invariance property of MLE, the MLE of θ is $\theta(\hat{\phi}_1)$, and this mapping is assumed to exist in the definition of weak exogeneity.

- Of course, we'll need to figure out just what this mapping is to recover $\hat{\theta}$ from $\hat{\phi}_1$. This is the famous *identification problem.*

- With lack of weak exogeneity, the joint and conditional likelihood functions maximize in different places. For this reason, we can't treat X_t as fixed in inference. The joint MLE is valid, but the conditional MLE is not.

- In resume, we require the variables in X_t to be weakly exogenous if we are to be able to treat them as fixed in estimation. Lagged Y_t satisfy the definition, since they are in the conditioning information set, e.g., $Y_{t-1} \in I_t$. Lagged Y_t aren't exogenous in the normal usage of the word, since their values *are* determined within the model, just earlier on. *Weakly exogenous* variables include *exogenous* (in the normal sense) variables as well as all *predetermined* variables.

11.3 Reduced form

Recall that the model is

$$Y_t'\Gamma = X_t'B + E_t'$$

$$V(E_t) = \Sigma$$

This is the model in *structural form.*

Definition 18 (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

$$Y_t' = X_t'B\Gamma^{-1} + E_t'\Gamma^{-1}$$
$$= X_t'\Pi + V_t' =$$

Now only one current period endog appears in each equation. This is the *reduced form.*

Definition 19 (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:

$$q_t \ = \ \alpha_1 + \alpha_2 \left(\frac{q_t - \beta_1 - \varepsilon_{2t}}{\beta_2} \right) + \alpha_3 y_t + \varepsilon_{1t}$$

$$\beta_2 q_t - \alpha_2 q_t \ = \ \beta_2 \alpha_1 - \alpha_2 (\beta_1 + \varepsilon_{2t}) + \beta_2 \alpha_3 y_t + \beta_2 \varepsilon_{1t}$$

$$q_t \ = \ \frac{\beta_2 \alpha_1 - \alpha_2 \beta_1}{\beta_2 - \alpha_2} + \frac{\beta_2 \alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\beta_2 \varepsilon_{1t} - \alpha_2 \varepsilon_{2t}}{\beta_2 - \alpha_2}$$

$$= \ \pi_{11} + \pi_{21} y_t + V_{1t}$$

Similarly, the rf for price is

$$\beta_1 + \beta_2 p_t + \varepsilon_{2t} \ = \ \alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t}$$

$$\beta_2 p_t - \alpha_2 p_t \ = \ \alpha_1 - \beta_1 + \alpha_3 y_t + \varepsilon_{1t} - \varepsilon_{2t}$$

$$p_t \ = \ \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}$$

$$= \ \pi_{12} + \pi_{22} y_t + V_{2t}$$

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

$$V(V_{1t}) \ = \ \mathcal{E} \left[\left(\frac{\beta_2 \varepsilon_{1t} - \alpha_2 \varepsilon_{2t}}{\beta_2 - \alpha_2} \right) \left(\frac{\beta_2 \varepsilon_{1t} - \alpha_2 \varepsilon_{2t}}{\beta_2 - \alpha_2} \right) \right]$$

$$= \ \frac{\beta_2^2 \sigma_{11} - 2 \beta_2 \alpha_2 \sigma_{12} + \alpha_2 \sigma_{22}}{(\beta_2 - \alpha_2)^2}$$

- 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).

11.4 IV estimation

The simultaneous equations model is

$$Y\Gamma = XB + E$$

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

$$
\Gamma = \begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \end{bmatrix}
$$

$$
B = \begin{bmatrix} \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix}
$$

With this, the first equation can be written as

$$
\begin{aligned}
y &= Y_1\gamma_1 + X_1\beta_1 + \varepsilon \\
&= Z\delta + \varepsilon
\end{aligned}
$$

The problem, as we've seen is that Z is correlated with ε, since Y_1 is formed of endogs.

Let's change notation to our standard linear model, but with correlation between regressors and ther error term:

$$
\begin{aligned}
y &= X\beta + \varepsilon \\
\varepsilon &\sim iid(0, \sigma^2) \\
\mathcal{E}(X'\varepsilon) &\neq 0.
\end{aligned}
$$

Consider some matrix W which is formed of variables uncorrelated with ε. This matrix defines a projection matrix

$$
P_W = W(W'W)^{-1}W'
$$

so that anything that is projected onto the space spanned by W will be uncorrelated with ε, by the definition of W. Transforming the model with this projection matrix we

get

$$P_W y = P_W X + P_W \varepsilon$$

or

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

Now we have that ε^* and X^* are uncorrelated, since this is simply

$$
\begin{aligned}
\mathcal{E}(X^{*\prime} \varepsilon^*) &= \mathcal{E}(X' P'_W P_W \varepsilon) \\
&= \mathcal{E}(X' P_W \varepsilon)
\end{aligned}
$$

and

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

is the fitted value from a regression of X on W. This is a linear combination of the columns of W, so it must be uncorrelated with ε. This implies that applying OLS to the model

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

will lead to a consistent estimator, given a few more assumptions. This is the *generalized instrumental variables estimator*. W is known as the matrix of instruments. The estimator is

$$\hat{\beta}_{IV} = (X' P_W X)^{-1} X' P_W y$$

from which we obtain

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

so

$$\hat{\beta}_{IV} - \beta = (X'P_WX)^{-1}X'P_W\varepsilon$$
$$= (X'W(W'W)^{-1}W'X)^{-1}X'W(W'W)^{-1}W'\varepsilon$$

Now we can introduce factors of n to get

$$\hat{\beta}_{IV} - \beta = \left(\left(\frac{X'W}{n}\right)\left(\frac{W'W}{n}^{-1}\right)\left(\frac{W'X}{n}\right)\right)^{-1}\left(\frac{X'W}{n}\right)\left(\frac{W'W}{n}\right)^{-1}\left(\frac{W'\varepsilon}{n}\right)$$

Assuming that each of the terms with a n in the denominator satisfies a LLN, so that

- $\frac{W'W}{n} \overset{p}{\to} Q_{WW}$, a finite pd matrix

- $\frac{X'W}{n} \overset{p}{\to} Q_{XW}$, a finite matrix with rank K ($= \text{cols}(X)$)

- $\frac{W'\varepsilon}{n} \overset{p}{\to} 0$

then the plim of the rhs is zero. This last term has plim 0 since we assume that W and ε are uncorrelated, e.g.,

$$\mathcal{E}(W_t'\varepsilon) = 0,$$

Given these assumtions the IV estimator is consistent

$$\hat{\beta}_{IV} \overset{p}{\to} \beta.$$

Furthermore, scaling by $\sqrt{n}$, we have

$$\sqrt{n}\left(\hat{\beta}_{IV} - \beta\right) = \left(\left(\frac{X'W}{n}\right)\left(\frac{W'W}{n}\right)^{-1}\left(\frac{W'X}{n}\right)\right)^{-1}\left(\frac{X'W}{n}\right)\left(\frac{W'W}{n}\right)^{-1}\left(\frac{W'\varepsilon}{\sqrt{n}}\right)$$

Assuming that the far right term satifies a CLT, so that

- $\frac{W'\varepsilon}{\sqrt{n}} \xrightarrow{d} N(0, Q_{WW}\sigma^2)$

then we get

$$\sqrt{n}\left(\hat{\beta}_{IV} - \beta\right) \xrightarrow{d} N\left(0, (Q_{XW}Q_{WW}^{-1}Q'_{XW})^{-1}\sigma^2\right)$$

The estimators for Q_{XW} and Q_{WW} are the obvious ones. An estimator for σ^2 is

$$\widehat{\sigma^2_{IV}} = \frac{1}{n}\left(y - X\hat{\beta}_{IV}\right)'\left(y - X\hat{\beta}_{IV}\right).$$

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 $\hat{\beta}_{IV}$ is

$$\hat{V}(\hat{\beta}_{IV}) = \left((X'W)(W'W)^{-1}(W'X)\right)^{-1}\widehat{\sigma^2_{IV}}$$

The IV estimator is

1. Consistent

2. Asymptotically normally distributed

3. Biased in general, since even though $\mathcal{E}(X'P_W\varepsilon) = 0$, $\mathcal{E}(X'P_WX)^{-1}X'P_W\varepsilon$ may not be zero, since $(X'P_WX)^{-1}$ and $X'P_W\varepsilon$ are not independent.

An important point is that the asymptotic distribution of $\hat{\beta}_{IV}$ depends upon Q_{XW} and Q_{WW}, and these depend upon the choice of W. *The choice of instruments influences the efficiency of the estimator*.

- When we have two sets of instruments, W_1 and W_2 such that $W_1 \subset W_2$, then the IV estimator using W_2 is at least as efficiently asymptotically as the estimator that

used W_1. More instruments leads to more asymptotically efficient estimation, in general.

- There are special cases where there is no gain (simultaneous equations is an example of this, as we'll see).

- 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 $P_W X$ becomes closer and closer to X itself as the number of instruments increases.

- IV estimation can clearly be used in the case of simultaneous equations. The only issue is which instruments to use.

11.5   Identification by exclusion restrictions

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.

11.5.1 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

$$Z = \left[\begin{array}{cc} Y_1 & X_1 \end{array} \right]$$

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 exxluded 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

$$Z = \begin{bmatrix} Y_1 & X_1 \end{bmatrix}$$

Recall that we've partitioned the model

$$Y\Gamma = XB + E$$

as

$$Y = \begin{bmatrix} y & Y_1 & Y_2 \end{bmatrix}$$

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

Given the reduced form

$$Y = X\Pi + V$$

we can write the reduced form using the same partition

$$
\begin{bmatrix} y & Y_1 & Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \pi_{11} & \Pi_{12} & \Pi_{13} \\ \pi_{21} & \Pi_{22} & \Pi_{23} \end{bmatrix} + v \; V_1 \; V_2
$$

so we have

$$
Y_1 = X_1\Pi_{12} + X_2\Pi_{22} + V_1
$$

so

$$
\frac{1}{n}W'Z = \frac{1}{n}W' \begin{bmatrix} X_1\Pi_{12} + X_2\Pi_{22} + V_1 & X_1 \end{bmatrix}
$$

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

$$
plim\frac{1}{n}W'Z = plim\frac{1}{n}W' \begin{bmatrix} X_1\Pi_{12} + X_2\Pi_{22} & X_1 \end{bmatrix}
$$

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.

11.5.2 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

$$
\begin{aligned}
Y_t'\Gamma F &= X_t'BF + E_t F \\
V(E_t F) &= F'\Sigma F
\end{aligned}
$$

164

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 coeffi-

165

cients 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.,

$$\rho\left(\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix}\right) = cols\left(\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix}\right) = G - 1$$

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

$$\begin{bmatrix} 1 & \Gamma_{12} \end{bmatrix}\begin{bmatrix} f_{11} \\ F_2 \end{bmatrix} = 1 \Rightarrow f_{11} = 1$$

Therefore, as long as

$$\rho\left(\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix}\right) = G - 1$$

then

$$\begin{bmatrix} f_{11} \\ F_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0_{G-1} \end{bmatrix}$$

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

$$K^{**} \geq G^* - 1$$

167

which is the previously derived necessary condition.

- 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.

To give an example of determining identification status, consider the following macro

model (this is the widely known Klein's Model 1)

$$
\begin{aligned}
\text{Consumption: } C_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^p + W_t^g) + \varepsilon_{1t} \\
\text{Investment: } I_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} \\
\text{Private Wages: } W_t^p &= \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t} \\
\text{Output: } X_t &= C_t + I_t + G_t \\
\text{Profits: } P_t &= X_t - T_t - W_t^p \\
\text{Capital Stock: } K_t &= K_{t-1} + I_t
\end{aligned}
$$

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 written as $Y\Gamma = XB + E$ gives

$$
\Gamma = \begin{bmatrix}
1 & 0 & 0 & -1 & 0 & 0 \\
0 & 1 & 0 & -1 & 0 & -1 \\
-\alpha_3 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & -\gamma_1 & 1 & -1 & 0 \\
-\alpha_1 & -\beta_1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

170

$$
B = \begin{bmatrix}
\alpha_0 & \beta_0 & \gamma_0 & 0 & 0 & 0 \\
\alpha_3 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & \gamma_3 & 0 & 0 & 0 \\
\alpha_2 & \beta_2 & 0 & 0 & 0 & 0 \\
0 & \beta_3 & 0 & 0 & 0 & 1 \\
0 & 0 & \gamma_2 & 0 & 0 & 0
\end{bmatrix}
$$

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

171

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

$$K^{**} - L = G^* - 1$$
$$5 - L = 3 - 1$$
$$L = 3$$

- 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.

11.6 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.

172

- 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: 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 should be 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

$$y = \hat{Y}\gamma_1 + X_1\beta_1 + \varepsilon.$$

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

$$
\begin{aligned}
y &= P_X Y_1 \gamma_1 + P_X X_1 \beta_1 + \varepsilon \\
&= P_X Z \delta + \varepsilon
\end{aligned}
$$

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

174

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

$$
\hat{V}(\hat{\delta}) = \left(Z'\hat{Z} \right)^{-1} \left(\hat{Z}'\hat{Z} \right) \left(\hat{Z}'Z \right)^{-1} \hat{\sigma}_{IV}^2
$$

However, looking at the last term in brackets

$$
\hat{Z}'Z = \begin{bmatrix} \hat{Y} & X \end{bmatrix}' \begin{bmatrix} Y & X \end{bmatrix} = \begin{bmatrix} Y'(P_X)Y & Y'(P_X)X \\ X'Y & X'X \end{bmatrix}
$$

but since P_X is idempotent and since $P_X X = X$, we can write

$$
\begin{aligned}
\begin{bmatrix} \hat{Y}_i & X_i \end{bmatrix}' \begin{bmatrix} Y_i & X_i \end{bmatrix} &= \begin{bmatrix} Y_i' P_X P_X Y_i & Y_i' P_X X_i \\ X_i' P_X Y_i & X_i' X_i \end{bmatrix} \\
&= \begin{bmatrix} \hat{Y}_i & X_i \end{bmatrix}' \begin{bmatrix} \hat{Y}_i & X_i \end{bmatrix} \\
&= \hat{Z}'\hat{Z}
\end{aligned}
$$

Therefore, the second and last term in the variance formula cancel, so the 2SLS varcov estimator simplifies to

$$
\hat{V}(\hat{\delta}) = \left(Z'\hat{Z} \right)^{-1} \hat{\sigma}_{IV}^2
$$

which, following some algebra similar to the above, can also be written as

$$\hat{V}(\hat{\delta}) = \left(\hat{Z}'\hat{Z}\right)^{-1}\hat{\sigma}^2_{IV}$$

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).

11.7 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 maximixed value is

$$s(\hat{\beta}_{IV}) = \left(y - X\hat{\beta}_{IV}\right)' P_W \left(y - X\hat{\beta}_{IV}\right),$$

but

$$\begin{aligned}
\hat{\varepsilon}_{IV} &= y - X\hat{\beta}_{IV} \\
&= y - X(X'P_W X)^{-1}X'P_W y \\
&= \left(I - X(X'P_W X)^{-1}X'P_W\right) y \\
&= \left(I - X(X'P_W X)^{-1}X'P_W\right)(X\beta + \varepsilon) \\
&= A(X\beta + \varepsilon)
\end{aligned}$$

where

$$A \equiv I - X(X'P_W X)^{-1}X'P_W$$

so

$$s(\hat{\beta}_{IV}) = \left(\varepsilon' + \beta'X'\right) A'P_W A (X\beta + \varepsilon)$$

Moreover, $A'P_W A$ is idempotent, as can be verified by multiplication:

$$\begin{aligned}
A'P_W A &= \left(I - P_W X(X'P_W X)^{-1}X'\right) P_W \left(I - X(X'P_W X)^{-1}X'P_W\right) \\
&= \left(P_W - P_W X(X'P_W X)^{-1}X'P_W\right)\left(P_W - P_W X(X'P_W X)^{-1}X'P_W\right) \\
&= \left(I - P_W X(X'P_W X)^{-1}X'\right) P_W.
\end{aligned}$$

Furthermore, A is orthogonal to X

$$\begin{aligned}
AX &= \left(I - X(X'P_W X)^{-1}X'P_W\right) X \\
&= X - X \\
&= 0
\end{aligned}$$

so

$$s(\hat{\beta}_{IV}) = \varepsilon' A' P_W A \varepsilon$$

Supposing the ε are normally distributed, with variance σ^2, then the random variable

$$\frac{s(\hat{\beta}_{IV})}{\sigma^2} = \frac{\varepsilon' A' P_W A \varepsilon}{\sigma^2}$$

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,

$$\frac{s(\hat{\beta}_{IV})}{\widehat{\sigma^2}} \stackrel{a}{\sim} \chi^2(\rho(A' P_W A))$$

- 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

$$
\begin{aligned}
\rho(A' P_W A) &= Tr\left(P_W - P_W X (X' P_W X)^{-1} X' P_W\right) \\
&= Tr P_W - Tr X' P_W P_W X (X' P_W X)^{-1} \\
&= Tr W (W' W)^{-1} W' - K_X \\
&= Tr W' W (W' W)^{-1} - K_X \\
&= K_W - K_X
\end{aligned}
$$

178

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.

- Note that since

$$\hat{\varepsilon}_{IV} = A\varepsilon$$

and

$$s(\hat{\beta}_{IV}) = \varepsilon' A' P_W A \varepsilon$$

we can write

$$
\begin{aligned}
\frac{s(\hat{\beta}_{IV})}{\widehat{\sigma^2}} &= \frac{\left(\hat{\varepsilon}' W (W'W)^{-1} W'\right)\left(W(W'W)^{-1}W'\hat{\varepsilon}\right)}{\hat{\varepsilon}'\hat{\varepsilon}/n} \\
&= n(RSS_{\hat{\varepsilon}_{IV}|W}/TSS_{\hat{\varepsilon}_{IV}}) \\
&= nR_u^2
\end{aligned}
$$

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

179

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

and W is the matrix of instruments. If we have exact identification then $cols(W) = cols(X)$. 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.

11.8   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

$$
\Psi = \begin{bmatrix}
\sigma_{11}I_n & \sigma_{12}I_n & \cdots & \sigma_{1G}I_n \\
 & \sigma_{22}I_n & & \vdots \\
 & & \ddots & \vdots \\
\cdot & & & \sigma_{GG}I_n
\end{bmatrix}
$$
$$
= \Sigma \otimes I_n
$$

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).

11.8.1 3SLS

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

$$\mathcal{E}(\varepsilon\varepsilon') = \Psi$$
$$= \Sigma \otimes I_n$$

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 prediciitons 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. More on this later.

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 (**??**), 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.

The 3SLS estimator is based upon the rf parameter estimator $\hat{\Pi}$, calculated equation by equation using OLS:

$$\hat{\Pi} = (X'X)^{-1}X'Y$$

which is simply

$$\hat{\Pi} = (X'X)^{-1}X'\begin{bmatrix} y_1 & y_2 & \cdots & y_G \end{bmatrix}$$

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:

$$\begin{aligned} Y_t' &= X_t'B\Gamma^{-1} + E_t'\Gamma^{-1} \\ &= X_t'\Pi + V_t' \end{aligned}$$

and

$$V_t = \left(\Gamma^{-1}\right)' E_t \sim N\left(0, \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1}\right), \forall t$$

Let this var-cov matrix be indicated by

$$\Xi = \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1}$$

OLS equation by equation to get the rf is equivalent to

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix} = \begin{bmatrix} X & 0 & \cdots & 0 \\ 0 & X & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & X \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_G \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_G \end{bmatrix}
$$

where y_i is the $n \times 1$ vector of observations of the ith endog, X is the entire $n \times K$ matrix of exogs, π_i is the ith column of Π, and v_i is the ith 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 equations are contemporanously correlated, however. The general case would have a different X_i for each equation.

- Note that each equation of the system individually satisfies the classical assump-

187

tions.

- However, pooled estimation using the GLS correction is more efficient, since equation-by-equation estimation is equivalent to pooled estimation, since **X** is block diagonal, but ignoring the covariance information.

- The model is estimated by GLS, where Ξ is estimated using the OLS residuals from equation-by-equation estimation, which are consistent.

- In the special case that all the X_i are the same, which 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

$$\begin{aligned}
\hat{\pi}_{SUR} &= \left((I_n \otimes X)' (\Xi \otimes I_n)^{-1} (I_n \otimes X) \right)^{-1} (I_n \otimes X)' (\Xi \otimes I_n)^{-1} y \\
&= \left((\Xi^{-1} \otimes X') (I_n \otimes X) \right)^{-1} (\Xi^{-1} \otimes X') y \\
&= \left(\Xi \otimes (X'X)^{-1} \right) (\Xi^{-1} \otimes X') y \\
&= \left[I_G \otimes (X'X)^{-1} X' \right] y \\
&= \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \vdots \\ \hat{\pi}_G \end{bmatrix}
\end{aligned}$$

- 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. See two sections ahead.

11.8.2 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.

12 Limited dependent variables

Up until now we've considered models where the lhs variable typically is assumed to take on values on the real line. For example, if the model is $y_t = x_t'\beta + \varepsilon_t$ and ε_t is assumed to be normally distributed, then y_t will also be normally distributed, conditional on x_t, and therefore will take on values on $\mathfrak{R}$. This is unreasonable in many cases. For example, economic variables are often nonnegative (for example prices and quatities), or the variables may be restricted to integers (for example, the number of visits to the doctor a person makes in a year). In this section we'll see a few examples of models for these sorts of data.

12.1 Choice between two objects: the probit model

Suppose that an individual has to choose between two mutually exclusive possibilities, for example, between one of two job offers. Let indirect utility in the two states be $v_j(\mathbf{p}, m, \mathbf{z}) + \varepsilon_j, j = 0, 1$, where $\mathbf{p}$ is a price vector, m is income, and $\mathbf{z}$ is a vector of other variables related to the person's preferences or characteristics of the object. The first object ($j = 1$) is chosen if

$$\varepsilon_0 + v_0(m, \mathbf{p}, \mathbf{z}) < v_1(m, \mathbf{p}, \mathbf{z}) + \varepsilon_1$$

or if

$$\varepsilon_0 - \varepsilon_1 < v_1(m, \mathbf{p}, \mathbf{z}) - v_0(m, \mathbf{p}, \mathbf{z})$$

Define $\varepsilon = \varepsilon_0 - \varepsilon_1$, let $\mathbf{x}$ collect m, $\mathbf{p}$ and $\mathbf{z}$, and let $\Delta v(\mathbf{x}) = v_1(\mathbf{x}) - v_0(\mathbf{x})$. The first object is chosen if

$$\varepsilon < \Delta v(\mathbf{x}).$$

Define $y = 1$ if the consumer chooses object $j = 1$, $y = 0$ otherwise. The probability the first object is chosen is

$$\Pr(y = 1) = F_\varepsilon \left[\Delta v(\mathbf{x}) \right]$$
$$\equiv p(\mathbf{x}, \theta),$$

where θ are the parameters of the utility functions and the distribution function of ε.

A fairly simple version of this model is the standard *probit* model. Suppose that

$$v_0(m, \mathbf{p}, \mathbf{z}) = \alpha_0 + \beta_0 m + \mathbf{p}' \gamma_0$$
$$v_1(m, \mathbf{p}, \mathbf{z}) = \alpha_1 + \beta_1 m + \mathbf{p}' \gamma_1$$

and

$$
\begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \cdot & \sigma_{22} \end{bmatrix} \right).
$$

If we make the restrictions $\sigma_{11} = 0.5, \sigma_{12} = 0, \sigma_{22} = 0.5$ then

$$\varepsilon = \varepsilon_0 - \varepsilon_1 \sim N(0, 1).$$

Also,

$$\Delta v(\mathbf{w}) = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0) m + \mathbf{p}' (\gamma_1 - \gamma_0)$$
$$= \delta + \phi m + \mathbf{p}' \psi$$

193

and

$$\Pr(y \ = 1) = \Phi(\delta + \phi m + \mathbf{p}'\psi)$$

$$\equiv \Phi(\mathbf{x}'\theta).$$

where $\Phi(\cdot)$ is the standard normal distribution function and θ is the vector formed of the parameters δ, ϕ and ψ, which are in turn functions of the parameters α_i, β_i and γ_i, $i = 1, 2$.

Each observation can be thought of as a Bernoulli trial with probability of success equal to $\Phi(\mathbf{x}'\theta)$. The density function for a single Bernoulli trial is

$$\Pr(y|\mathbf{x}) = \Phi(\mathbf{x}'\theta)^y (1 - \Phi(\mathbf{x}'\theta))^{(1-y)}, y = 0, 1.$$

With n i.i.d. observations indexed by t, the likelihood function is

$$\ln \mathcal{L}(\theta) = \prod_{t=1}^{n} \Phi(\mathbf{x}_t'\theta)^{y_t} (1 - \Phi(\mathbf{x}_t'\theta))^{(1-y_t)}$$

and the average log-likelihood function is

$$s_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_t \ln \Phi(\mathbf{x}_t'\theta) + (1 - y_t) \ln \left[1 - \Phi(\mathbf{x}_t'\theta) \right] \right)$$

This is a nonlinear in the parameters function. We'll discuss how it can be maximized later. Note that Gauss has a function to calculate $\Phi(\cdot)$, it is cdfn$(\cdot)$. With this it's not hard to program the likelihood function.

A few comments:

- The parameters in θ are consistently estimated. On the other hand this has re-quired making assumptions regarding the parameters σ_{11}, σ_{12} and σ_{22}. Without

194

these restrictions the distribution function of ε is not identified, so the other parameters aren't identified. Also, knowledge of θ does not allow recovery of the α_i, β_i and γ_i, since there are twice as many unknowns as equations.

- The particular restrictions used to get that $\varepsilon \sim N(0,1)$ are not unique. We could have just as well assumed that $\varepsilon_0 \sim N(0,1)$ and $\varepsilon_1 = 0$. This would give the same distribution for ε.

- Binary response models of this sort are never identified without these sorts of restrictions.

The *logit* model is very similar to the probit model. Under the logit model the $\varepsilon_j, j = 12$ are assumed to be iid *extreme value* random variables. This leads to

$$F_\varepsilon(z) = \frac{1}{(1 + \exp(-z))}$$

so

$$\Pr(y = 1) = \frac{1}{(1 + \exp(-\mathbf{x}'\theta))}$$

It turns out that the probit and logit models give very similar estimates for $\Pr(y = 1)$ and the marginal effects $D_{\mathbf{x}} \Pr(y = 1)$. These functions and functionals of them are usually of most interest. Therefore the choice between logit and probit models is not very important, in the binary choice case. The coefficients are different (there is a scaling factor that related the coefficients). However, the coefficient themselves aren't usually of much interest and they are difficult to interpret.

12.2   Count data

Another situation where a continuous normally distributed dependent variable is unreasonable is the case where it represent the number of times some event occurs. For

example, the dependent variable could be the number of auto accidents in a week-end, or the number of political leaders that make fools of themselve in a week. Such variables are termed *count data* dependent variables.

The Poisson model is one of the simplest models for count data. The Poisson density is

$$f_Y(y) = \frac{\exp(-\lambda)\lambda^y}{y!}, y = 0, 1, 2, \ldots$$
$$\lambda > 0.$$

To allow for conditioning variables $\mathbf{x}$, make λ a function of $\mathbf{x}$. We need to ensure that λ is positive for all $\mathbf{x}$. The most popular parameterization is

$$\lambda = \exp(\mathbf{x}'\theta).$$

The log-likelihood for an individual observation is

$$s_t = -\exp(\mathbf{x}_t'\theta) + y_t \exp(\mathbf{x}_t'\theta) - \ln(y_t!)$$

and the average likelihood function is just the sum of this divided by the sample size:

$$s_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(-\exp(\mathbf{x}_t'\theta) + y_t \exp(\mathbf{x}_t'\theta) - \ln(y_t!) \right).$$

With this, θ can be estimated by ML.

The Poisson model exhibits a restriction that may not be desirable. This is that $\mathcal{E}(y) = V(y) = \lambda$. Usually there is no reason why the mean should be equal to the variance. There are generalizations of the Poisson model that relax this restriction. The way this is done is to make λ a function of another random variable, then integrate

196

this variable out. That is

$$\lambda = \exp(\mathbf{x}'\theta + \eta)$$

$$\eta \sim f_\eta(z, \phi)$$

so the joint density of y and η is the product of the conditional density of y given η, and the marginal of η:

$$f_{\mathcal{Y}}(y, \eta | \theta, \phi) = \frac{\exp(-\exp(\mathbf{x}'\theta + \eta))\exp(\mathbf{x}'\theta + \eta)^y}{y!} f_\eta(\eta, \phi)$$

and the marginal density of y is obtained by integrating out η:

$$f_{\mathcal{Y}}(y | \theta, \phi) = \int_{\mathcal{Z}} \frac{\exp(-\exp(\mathbf{x}'\theta + z))\exp(\mathbf{x}'\theta + z)^y}{y!} f_\eta(z, \phi) dz$$

This effectively introduces other parameters ϕ into the densisty which relax the Poisson mean-variance restriction.

12.3   Duration data

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 s years is

$$\Pr(D > s) = 1 - \Pr(D \leq s) = 1 - F_D(s).$$

The density of D conditional on the spell already having lasted 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.

$$E = \mathcal{E}(D|D > s) - s = \left(\int_t^\infty z \frac{f_D(z)}{1 - F_D(s)} ds \right) - s$$

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 length (in months) of strikes in the industrial sector were used to fit a Weibull model. The parameter

estimates are

& 0.559 & 0.034 \\\(\gamma\)& 0.867 & 0.033\end{tabular}\end{equation} and the log-likelihood value is -659.3

A plot of E, with 95% confidence bands follows. The plot is accompanied by a nonparametric Kaplan-Meier estimate of life-expectancy. This nonparametric estimator of E simply averages all spell lengths greater than t, and subtracts t. This is consistent by the LLN.

In the figure one can seel that the model doesn't fit the data well, in that it predicts E quite differently than does the nonparametric model. It seems that many strikes end quickly, since E is relatively low initially, but that if a strike lasts a month then it is likely to last considerably longer. Due to the dramatic change in the rate that spells end 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. 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

This model leads to a fit for E in the figure

Note that the parametric and nonparametric fits are quite close to one another, up to around 6 months. The disagreement after this point is not too important, since less that 5% of strikes last more than 6 months, which implies that the Kaplan-Meier nonparametric estimate has a high variance (since it's an average of a small number of observations).

12.4 The Newton method

The Newton-Raphson 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, e.g, 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

$$\theta^{k+1} = \theta^k - H(\theta^k)^{-1} g(\theta^k)$$

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 may have 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. This has the benefit that improvement in the objective function is guaranteed.

- 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.

Example of Newton iterations   Consider the funcion

$$f(x) = \ln x - 1.0 \frac{x}{1.0 + e^{-1.0x}}$$

This has a maximum at the point $x = 1.058416$ (approximately) as we can see by

$$f'(1.058416) = 2.206 \times 10^{-7}$$

Consider applying Newton-Raphson. The initial point is $z = 0.5$ The second order approximation is

$$g(x) = f(z) + f'(z)(x - z) + \frac{1}{2}f''(z)(x - z)^2$$

Plotting the true function and the approximation:

The next round expansion point is obtained by maximizing the approximation:

$g(x)$ Candidate(s) for extrema: $\{-.82563\}$, at $\{\{x = .78371\}\}$

Now set the expansion point to the new value, and re-plot:

$z_2 = .78371$

$g_2(x) = f(z_2) + f'(z_2)(x - z_2) + \frac{1}{2}f''(z_2)(x - z_2)^2$

$g_2(x)$ Candidate(s) for extrema: $\{-.73735\}$, at $\{\{x = .99458\}\}$

Another round:

$z_3 = .99458$

$g_3(x) = f(z_3) + f'(z_3)(x - z_3) + \frac{1}{2}f''(z_3)(x - z_3)^2$

$g_3(x)$ Candidate(s) for extrema: $\{-.72907\}$, at $\{\{x = 1.055\}\}$. So after two NR iterations we're already pretty close to the maximum and the approximation is quite close to the function, up to second order.

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, and in fact, more than about 6-10 decimals of precision is usually infeasible. 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:

$$|g_j(\theta^k) - g_j(\theta^{k-1})| < \varepsilon_4, \forall j$$

- Or, even better, check all of these.

- Also, if we're maximizing, it's good to check that the last round Hessian is negative definite.

203

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.**

13 Models for time series data

Hamilton, *Time Series Analysis* is a good reference for this section. This is very incomplete and contributions would be very welcome.

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 time series work is done with linear models, though nonlinear time series is also a large and growing field. We'll stick with linear time series models.

13.1 Basic concepts

Definition 20 (Stochastic process) *A stochastic process is a sequence of random variables, indexed by time:*

$$\{Y_t\}_{t=-\infty}^{\infty} \tag{11}$$

Definition 21 (Time series) *A time series is **one** observation of a stochastic process, over a specific interval:*

$$\{y_t\}_{t=1}^{n} \tag{12}$$

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 22 (Autocovariance) *The jth autocovariance of a stochastic process is*

$$\gamma_{jt} = \mathcal{E}(y_t - \mu_t)(y_{t-j} - \mu_{t-j}) \tag{13}$$

where $\mu_t = \mathcal{E}(y_t)$.

Definition 23 (Covariance (weak) stationarity) *A stochastic process is covariance stationary if it has time constant mean and autocovariances of all orders:*

$$\mu_t \quad = \mu, \forall t$$

$$\gamma_{jt} \quad = \gamma_j, \forall t$$

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 24 (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$\Rightarrow$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., $\{y_t^m\}$ 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 25 (Ergodicity) *A stationary stochastic process is ergodic (for the mean) if the time average converges to the mean*

$$\frac{1}{n}\sum_{t=1}^{n} y_t \xrightarrow{p} \mu \tag{14}$$

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 26 (Autocorrelation) *The jth autocorrelation, ρ_j is just the jth autocovariance divided by the variance:*

$$\rho_j = \frac{\gamma_j}{\gamma_0} \tag{15}$$

Definition 27 (White noise) *White noise is just the time series literature term for a classical error. ε_t is white noise if i) $\mathcal{E}(\varepsilon_t) = 0, \forall t$, ii) $V(\varepsilon_t) = \sigma^2, \forall t$, and iii) ε_t and ε_s are independent, $t \neq s$. Gaussian white noise just adds a normality assumption.*

13.2 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.

13.2.1 MA(q) processes

A qth 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

$$
\begin{aligned}
\gamma_0 \quad &= \mathcal{E}\left(y_t - \mu\right)^2 \\
&= \mathcal{E}\left(\varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}\right)^2 \\
&= \sigma^2 \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)
\end{aligned}
$$

Similarly, the autocovariances are

$$
\begin{aligned}
\gamma_j \quad &= \quad \theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \cdots + \theta_q \theta_{q-j}, j \le q \\
&= \quad 0, j > q
\end{aligned}
$$

Therefore an MA(q) process is necessarily covariance stationary and ergodic, as long as σ^2 and all of the θ_j are finite.

13.2.2 AR(p) processes

An AR(p) process can be represented as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

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:

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} \phi_1 & \phi_2 & \cdots & & \phi_p \\ 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \cdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

or

$$Y_t = C + F Y_{t-1} + E_t$$

With this, we can recursively work forward in time:

$$\begin{aligned} Y_{t+1} &= C + F Y_t + E_{t+1} \\ &= C + F \left(C + F Y_{t-1} + E_t \right) + E_{t+1} \\ &= C + FC + F^2 Y_{t-1} + F E_t + E_{t+1} \end{aligned}$$

and

$$\begin{aligned} Y_{t+2} &= C + F Y_{t+1} + E_{t+2} \\ &= C + F \left(C + FC + F^2 Y_{t-1} + F E_t + E_{t+1} \right) + E_{t+2} \\ &= C + FC + F^2 C + F^3 Y_{t-1} + F^2 E_t + F E_{t+1} + E_{t+2} \end{aligned}$$

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 for 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 pth 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

$$\lim_{j \to \infty} F_{(1,1)}^j = 0$$

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: $\partial y_{t+j} / \partial \varepsilon_t = F_{(1,1)}^j$ is a *dynamic multiplier* or an *impulse-response* function. Real eigenvalues lead to steady movements, whereas comlpex

eigenvalue lead to ocillatory behavior. Of course, when there are multiple eigen-

values the overall effect can be a mixture. *pictures*

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 $1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j$ to get

$$\left(1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j\right)(1 - \phi L)y_t = \left(1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j\right)\varepsilon_t$$

or, multiplying the polynomials on th LHS, we get

$$\left(1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j - \phi L - \phi^2 L^2 - \dots - \phi^j L^j - \phi^{j+1} L^{j+1}\right)y_t$$
$$== \left(1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j\right)\varepsilon_t$$

and with cancellations we have

$$\left(1 - \phi^{j+1} L^{j+1}\right)y_t = \left(1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j\right)\varepsilon_t$$

so

$$y_t = \phi^{j+1} L^{j+1} y_t + \left(1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j\right)\varepsilon_t$$

Now as $j \to \infty$, $\phi^{j+1} L^{j+1} y_t \to 0$, since $|\phi| < 1$, so

$$y_t \cong \left(1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j\right)\varepsilon_t$$

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

$$y_t \cong \left(1 + \phi L + \phi^2 L^2 + \dots + \phi^j L^j\right)(1 - \phi L)y_t$$

so

$$\left(1 + \phi L + \phi^2 L^2 + \ldots + \phi^j L^j\right)(1 - \phi L) \cong 1$$

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 conju-

216

gate 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

$$
Y_t = F^{j+1} Y_{t-j-1} + F^j E_{t-j} + F^{j-1} E_{t-j+1} + \cdots + F E_{t-1} + E_t
$$

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).

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

$$\mu = c + \phi_1 \mu + \phi_2 \mu + \ldots + \phi_p \mu$$

so

$$\mu = \frac{c}{1 - \phi_1 - \phi_2 - \ldots - \phi_p}$$

and

$$c = \mu - \phi_1 \mu - \ldots - \phi_p \mu$$

so

$$
\begin{aligned}
y_t - \mu \ &= \mu - \phi_1 \mu - \ldots - \phi_p \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \mu \\
&= \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \ldots + \phi_p(y_{t-p} - \mu) + \varepsilon_t
\end{aligned}
$$

With this, the second moments are easy to find: The variance is

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \ldots + \phi_p \gamma_p + \sigma^2$$

The autocovariances of orders $j \geq 1$ follow the rule

$$
\begin{aligned}
\gamma_j \qquad\qquad &= \mathcal{E}\left[(y_t - \mu)(y_{t-j} - \mu))\right] \\
&= \mathcal{E}\left[(\phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \ldots + \phi_p(y_{t-p} - \mu) + \varepsilon_t)(y_{t-j} - \mu)\right] \\
&= \phi_1 \gamma_{j-1} + \phi_2 \gamma_{j-2} + \ldots + \phi_p \gamma_{j-p}
\end{aligned}
$$

218

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 ($\sigma^2, \gamma_0, \gamma_1, ..., \gamma_p$) and solve for the unknowns. With these, the γ_j for $j > p$ can be solved for recursively.

13.2.3   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 $|\eta_i| < 1$. 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

$$y_t = c + \delta_1 y_{t-1} + \delta_2 y_{t-2} + \ldots + \varepsilon_t$$

where

$$c = \mu + \delta_1 \mu + \delta_2 \mu + \ldots$$

So we see that an MA(q) has an infinite AR representation, as long as the $|\eta_i| < 1$, $i = 1, 2, \ldots, q$.

- 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

$$\sigma_{\varepsilon^*}^2 = \sigma_\varepsilon^2 \theta^2$$

For example, we've seen that

$$\gamma_0 = \sigma^2 (1 + \theta^2).$$

Given the above relationships amongst the parameters,

$$\gamma_0^* = \sigma_\varepsilon^2 \theta^2 (1 + \theta^{-2}) = \sigma^2 (1 + \theta^2)$$

so the variances are the same. It turns out that *all* the autocovariances will be the

220

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

- 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. At the time of estimation, it's a lot easier to estimate the single AR(1) coefficient rather than the infinite number of coefficients associated with the MA 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 with 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 28 *Calculate the autocovariances of an ARMA(1,1) model:* $(1 + \phi L)y_t = c + (1 + \theta L)\varepsilon_t$

14 Introduction to the second half

We'll begin with study of *extremum estimators* in general. Let $\mathbf{Z}_n$ be the available data, based on a sample of size n.

Definition 29 *[Extremum estimator] An extremum estimator $\hat{\theta}$ is the optimizing element of an objective function $s_n(\mathbf{Z}_n, \theta)$ over a set Θ.*

We'll write the objective function suppressing the dependence on $\mathbf{Z}_n$.

Example 30 *Least squares, linear model*

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 $\mathbf{X}_n = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}'$. The least squares estimator is defined as

$$\hat{\theta} \equiv \arg\min_{\Theta} s_n(\theta) = 1/n\left[\mathbf{y}_n - \mathbf{X}_n\theta\right]'\left[\mathbf{y}_n - \mathbf{X}_n\theta\right]$$

We readily find that $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Example 31 *Maximum likelihood*

Suppose that the continuous random variable $y_t \sim IIN(\theta^0, 1)$. The maximum likelihood estimator is defined as

$$\hat{\theta} \equiv \arg\max_{\Theta} \ \mathcal{L}_n(\theta) = \prod_{t=1}^{n}(2\pi)^{-1/2}\exp\left(-\frac{(y_t - \theta)^2}{2}\right)$$

Because the logarithmic function is strictly increasing on $(0, \infty)$, maximization of the average logarithm of the likelihood function is achieved at the same $\hat{\theta}$ as for the likelihood function:

$$\hat{\theta} \equiv \arg\max_{\Theta} s_n(\theta) = 1/n \ln \mathcal{L}_n(\theta) = -1/2\ln 2\pi - 1/n\sum_{t=1}^{n}\frac{(y_t - \theta)^2}{2}$$

Solution of the f.o.c. leads to the familiar result that $\hat{\theta} = \bar{\mathbf{y}}$.

- MLE estimators are asymptotically efficient (Cramér-Rao theorem), *supposing the strong distributional assumptions upon which they are based are true.*

- One can investigate the properties of an "ML" estimator supposing that the distributional assumptions are incorrect. This gives a *quasi-ML estimator*, which we'll study later.

- The strong distributional assumptions of MLE may be questionable in many cases. It is possible to estimate using weaker distributional assumptions based only on some of the moments of a random variable(s).

Example 32 *Method of moments*

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, *i.e.,* $\mu_1(\theta^0)$.

- $\mu_1 = \mu_1(\theta^0)$ is a *moment-parameter equation.*

- In this example, 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

$$\widehat{\mu_1} = \sum_{t=1}^{n} y_t / n.$$

- Define

$$m_1(\theta) = \mu_1(\theta) - \widehat{\mu_1}$$

- 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, *i.e.,* $m_1(\hat{\theta}) \equiv 0$.

In this case,

$$m_1(\hat{\theta}) = \hat{\theta} - \sum_{t=1}^{n} y_t / n = 0.$$

Since $\sum_{t=1}^{n} y_t / n \xrightarrow{p} \theta^0$ by the WLLN, the estimator is consistent.

Example 33 *Method of moments, continued.*

Continuing with the above example, the variance of a $\chi^2(\theta^0)$ r.v. is

$$V(y_t) = E\left(y_t - \theta^0\right)^2 = 2\theta^0.$$

- Define

$$m_2(\theta) = 2\theta - \frac{\sum_{t=1}^{n} (y_t - \bar{y})^2}{n}$$

- The MM estimator would set

$$m_2(\hat{\theta}) = 2\hat{\theta} - \frac{\sum_{t=1}^{n} (y_t - \bar{y})^2}{n} \equiv 0,$$

  so,

$$\hat{\theta} = \frac{\sum_{t=1}^{n} (y_t - \bar{y})^2}{2n}.$$

  Again, since, by the WLLN, the sample variance is consistent for the true variance, that is,

$$\frac{\sum_{t=1}^{n} (y_t - \bar{y})^2}{n} \xrightarrow{p} 2\theta^0$$

  the MM estimator is consistent.

Example 34 *Generalized method of moments (GMM)*

The previous two examples give two estimators of θ^0 which are both consistent. With a given sample, the estimators will be different in general.

224

- With two moment-parameter equations and only one parameter, we have *overi-dentification,* which means that we have more information than is strictly necessary for consistent estimation of the parameter.

- The GMM combines information from the two moment-parameter equations to form a new estimator which will be *more efficient,* in general (proof of this below).

From the first example, define $m_{1t}(\theta) = \theta - y_t$. We already have that $m_1(\theta)$ is the sample average of $m_{1t}(\theta)$, *i.e.,*

$$
\begin{aligned}
m_1(\theta) &= 1/n \sum_{t=1}^{n} m_{1t}(\theta) \\
&= \theta - \sum_{t=1}^{n} y_t / n.
\end{aligned}
$$

Clearly, when evaluated at the true parameter value θ^0, both $E\left[m_{1t}(\theta^0)\right] = 0$ and $E\left[m_1(\theta^0)\right] = 0$.

From the second example we define additional moment conditions

$$
m_{2t}(\theta) = 2\theta - (y_t - \bar{y})^2
$$

and

$$
m_2(\theta) = 2\theta - \frac{\sum_{t=1}^{n}(y_t - \bar{y})^2}{n}.
$$

Again, it is clear from the SLLN that $m_2(\theta^0) \overset{a.s.}{\to} 0$. The MM estimator would chose $\hat{\theta}$ to set either $m_1(\hat{\theta}) = 0$ or $m_2(\hat{\theta}) = 0$. In general, no single value of θ will solve the two equations simultaneously.

- The GMM estimator is based on defining a measure of distance $d(m(\theta))$, where

225

$m(\theta) = (m_1(\theta), m_2(\theta))'$, and choosing

$$\hat{\theta} = \arg\min_{\Theta} s_n(\theta) = d\left(m(\theta)\right).$$

An example would be to choose $d(m) = m'Am$, where A is a positive definite matrix. While it's clear that the MM gives consistent estimates if there is a one-to-one relationship between parameters and moments, it's not immediately obvious that the GMM estimator is consistent. (We'll see later that it is.)

These examples show that these widely used estimators may all be interpreted as the solution of an optimization problem. For this reason, the study of extremum estimators is useful for its generality. We will see that the general results extend smoothly to the more specialized results available for specific estimators. After studying extremum estimators in general, we will study the GMM estimator, then QML and NLS. The reason we study GMM first is that LS, IV, NLS, MLE, QML and other well-known parametric estimators may all be interpreted as special cases of the GMM estimator, so the general results on GMM can simplify and unify the treatment of these other estimators. Nevertheless, there are some special results on QML and NLS, and both are important in empirical research, which makes focus on them useful.

One of the focal points of the course will be nonlinear models. This is not to suggest that linear models aren't useful. Linear models are more general than they might first appear, since one can employ nonlinear transformations of the variables:

$$\varphi_0(y_t) = \left[\begin{array}{cccc} \varphi_1(x_t) & \varphi_2(x_t) & \cdots & \varphi_p(x_t) \end{array}\right]\theta^0 + \varepsilon_t$$

For example,

$$\ln y_t = \alpha + \beta x_{1t} + \gamma x_{1t}^2 + \delta x_{1t} x_{2t} + \varepsilon_t$$

fits this form.

- The important point is that $\varphi_0(y_t)$ is *linear in the parameters* but not necessarily *linear in the variables.*

In spite of this generality, situations often arise which simply can not be convincingly represented by linear in the parameters models.

Example 35 *Expenditure shares*

Roy's Identity states that the quantity demanded of the ith of G goods is

$$x_i = \frac{-\partial v(p,y)/\partial p_i}{\partial v(p,y)/\partial y}.$$

An expenditure share is

$$s_i \equiv p_i x_i / y,$$

so necessarily $s_i \in [0,1]$, and $\sum_{i=1}^{G} s_i = 1$.

No linear in the parameters model for x_i or s_i with a parameter space that is defined independent of the data can guarantee that either of these conditions holds. These constraints will often be violated by estimated linear models, which calls into question their appropriateness in cases of this sort.

Example 36 *Binary limited dependent variable*

Suppose there is a latent process

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

but that y^* is not observed. Rather we observe

$$y = 1\left[\mathbf{x}'\beta^0 < \varepsilon\right]$$

227

so that y is either 0 or 1. In this case, we can write

$$y = F_\varepsilon(\mathbf{x}'\beta^0) + \eta$$

$$\mathcal{E}(\eta) = 0.$$

One could estimate this by (nonlinear) least squares

$$\hat{\beta} = \arg\min \frac{1}{n} \sum_t \left(y - F_\varepsilon(\mathbf{x}'\beta) \right)^2$$

The main point is that it is impossible that $F_\varepsilon(x'\beta^0)$ can be written as a linear in the parameters model, in the sense that there are no $\theta, \varphi(\mathbf{x})$ such that

$$F_\varepsilon(\mathbf{x}'\beta^0) = \varphi(\mathbf{x})'\theta, \forall \mathbf{x}$$

where $\varphi(\mathbf{x})$ is a p-vector valued function of the vector $\mathbf{x}$. This is because for any $\mathbf{x}$, we can always find a θ that is such that $\varphi(\mathbf{x})'\theta$ will be negative or greater than 1, which is illogical.

Since this sort of problem occurs often in empirical work, it is useful to study NLS and other nonlinear models.

After discussing these estimation methods for parametric models we'll briefly introduce *nonparametric estimation methods*. These methods allow one, for example, to estimate $f(x_t)$ consistently when we are not willing to assume that a model of the form

$$y_t = f(x_t) + \varepsilon_t$$

can be restricted to a parametric form

$$
\begin{aligned}
y_t &= f(x_t, \theta) + \varepsilon_t \\
\Pr(\varepsilon_t < z) &= F_\varepsilon(z | \phi, x_t) \\
\theta &\in \Theta, \phi \in \Phi
\end{aligned}
$$

where $f(\cdot)$ and perhaps $F_\varepsilon(z | \phi, x_t)$ are of known functional form. This is important since economic theory gives us general information about functions and the signs of their derivatives, but not about their specific form.

The final section deals with simulation methods in econometrics. These methods allow us to substitute computer power for mental power. Since computer power is becoming relatively cheap compared to mental effort, any econometrician who lives by the principles of economic theory should be interested in these techniques.

15 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.

15.1 Notation for differentiation of vectors and matrices

Readings: Gallant, Ch. 1, pp. 8-16.

Let $s(\cdot) : \Re^p \to \Re$ be a real valued function of the $p \times 1$ vector θ. Then $\frac{\partial s(\theta)}{\partial \theta}$ is organized as a $p \times 1$ 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. Note that

$$
\frac{\partial^2 s(\theta)}{\partial \theta \partial \theta'} = \frac{\partial}{\partial \theta} \left(\frac{\partial s(\theta)}{\partial \theta'} \right) = \frac{\partial}{\partial \theta'} \left(\frac{\partial s(\theta)}{\partial \theta} \right).
$$

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 $\left(\frac{\partial}{\partial \theta} f(\theta)' \right)' = \frac{\partial}{\partial \theta'} f(\theta)$.

- *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

230

$$\frac{\partial}{\partial \theta} h(\theta)' f(\theta) = \left(\frac{\partial}{\partial \theta} f' \right) h + \left(\frac{\partial}{\partial \theta} h' \right) f$$

which has dimension $p \times 1$.

- *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

$$\frac{\partial}{\partial \rho'} f[g(\rho)] = \left. \frac{\partial}{\partial \theta'} f(\theta) \right|_{\theta = g(\rho)} \frac{\partial}{\partial \rho'} g(\rho)$$

has dimension $n \times r$.

15.2 Convergenge modes

Readings: Davidson and MacKinnon, Ch. 4*; Amemiya Ch. 3; Hamilton Ch. 7; Davidson (1994) is a good advanced reference.

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 37 *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 38 [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 $\{f_n(w)\}$ where

$$f_n : \Omega \to T \subseteq \mathfrak{R}.$$

Ω may be an arbitrary set.

Definition 39 [Pointwise convergence] *A sequence of functions $\{f_n(w)\}$ 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*

$$|f_n(w) - f(\omega)| < \varepsilon, \forall n > N_{\varepsilon\omega}.$$

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 40 [Uniform convergence] *A sequence of functions $\{f_n(w)\}$ converges uniformly on Ω to the function $f(\omega)$ if for any $\varepsilon > 0$ there exists an integer N such that*

$$\sup_{\omega \in \Omega} |f_n(w) - f(\omega)| < \varepsilon, \forall n > N.$$

(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., $X(\omega) : \Omega \to \mathfrak{R}$. A sequence of random variables $\{X_n(\omega)\}$ is a collection

232

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 41 [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 $X_n \overset{p}{\to} X$, or plim $X_n = X$.

Definition 42 [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 $X_n \overset{a.s.}{\to} X$, or $X_n \to X, a.s.$ One can show that

$$X_n \overset{a.s.}{\to} X \Rightarrow X_n \overset{p}{\to} X.$$

Definition 43 [Convergence in distribution] *Let the r.v. X_n have distribution function*

233

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 44 [Uniform almost sure convergence] $\{X_n(\omega, \theta)\}$ *converges uniformly almost surely in Θ to $X(\omega, \theta)$ if*

$$\lim_{n \to \infty} \sup_{\theta \in \Theta} |X_n(\omega, \theta) - X(\omega, \theta)| = 0, (a.s.)$$

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

234

uniform convergence in probability by $\stackrel{u.p.}{\rightarrow}$.

- 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.

15.3   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 45 *[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 46 *[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 47 *The notation $f(n)=o_p(g(n))$ means $\frac{f(n)}{g(n)}\stackrel{P}{\rightarrow}0$.*

Example 48 *The least squares estimator $\hat{\theta}=(X'X)^{-1}X'Y=(X'X)^{-1}X'\left(X\theta^0+\varepsilon\right)=\theta^0+(X'X)^{-1}X'\varepsilon$. Since $plim\frac{(X'X)^{-1}X'\varepsilon}{1}=0$, 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 49 *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 50 *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 51 *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 52 *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., $n^{1/2}(\hat{\theta} - \mu) \overset{A}{\sim} N(0,\sigma^2)$. So $n^{1/2}(\hat{\theta} - \mu) = O_p(1)$, 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 53 *Two sequences of random variables $\{f_n\}$ and $\{g_n\}$ are asymptotically equal (written $f_n \overset{a}{=} g_n$) if*

$$plim\left(\frac{f(n)}{g(n)}\right) = 1$$

236

Finally, analogous almost sure versions of o_p and O_p are defined in the obvious way.

16 Asymptotic properties of extremum estimators

Readings: 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.

16.1 Extremum estimators

In Definition 29 we defined an extremum estimator $\hat{\theta}$ as the optimizing element of an objective function $s_n(\theta)$ over a set Θ. Let the objective function $s_n(\mathbf{Z}_n, \theta)$ depend upon a $n \times p$ random matrix $Z_n = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix}'$ where the z_t are p-vectors and p is finite.

Example 54 *Given the model $y_i = x_i'\theta + \varepsilon_i$, with n observations, define $z_i = (y_i, x_i')'$. The OLS estimator minimizes*

$$
\begin{aligned}
s_n(Z_n, \theta) &= 1/n \sum_{i=1}^{n} \left(y_i - x_i'\theta \right)^2 \\
&= 1/n \parallel Y - X\theta \parallel^2
\end{aligned}
$$

where Y and X are defined similarly to Z.

16.2 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 55 [Consistency of e.e.] *Suppose that $\hat{\theta}_n$ is obtained by maximizing $s_n(\theta)$ over $\overline{\Theta}$.*

Assume

1. *Compactness: The parameter space Θ is an open subset of Euclidean space $\Re^K$. The closure of Θ, $\overline{\Theta}$ is compact.*

2. *Uniform Convergence: There is a nonstochastic function $s_\infty(\theta)$ that is continuous in θ on $\overline{\Theta}$ such that*

$$\lim_{n \to \infty} \sup_{\theta \in \overline{\Theta}} |s_n(\theta) - s_\infty(\theta)| = 0, a.s.$$

3. *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 $s_n(\theta)$ converges uniformly 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 (1) and the fact that maximixation is over $\overline{\Theta}$. Since every sequence from a compact set has at least one limit point (Davidson, Thm. 2.12), 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 $\lim_{m \to \infty} \hat{\theta}_{n_m} = \hat{\theta}$ (for example, set each element of the sequence to $\hat{\theta}$). By uniform convergence and continuity

$$\lim_{m \to \infty} s_{n_m}(\hat{\theta}_{n_m}) = s_\infty(\hat{\theta}), a.s.$$

239

To see this, first of all, select an element $\hat{\theta}_t$ from the sequence $\{\hat{\theta}_{n_m}\}$. Then uniform convergence implies

$$\lim_{m \to \infty} s_{n_m}(\hat{\theta}_t) = s_\infty(\hat{\theta}_t), \text{a.s.}$$

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 $\{\hat{\theta}_t\}$ 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

$$s_\infty(\hat{\theta}) \geq s_\infty(\theta^0), \text{a.s.}$$

But by assumption (3), there is a unique global maximum of $s_\infty(\theta)$ at θ^0, so we must have $s_\infty(\hat{\theta}) = s_\infty(\theta^0)$, a.s. , and $\hat{\theta} = \theta^0$, a.s. 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 have $\| \theta - \theta^0 \| = 0$, 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 next section on numeric optimization methods will show 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.

241

With uniform convergence, the maximum of the sample
objective function eventually must be in the neighborhood
of the maximum of the limiting objective function

With pointwise convergence, the sample objective function
may have its maximum far away from that of the limiting
objective function
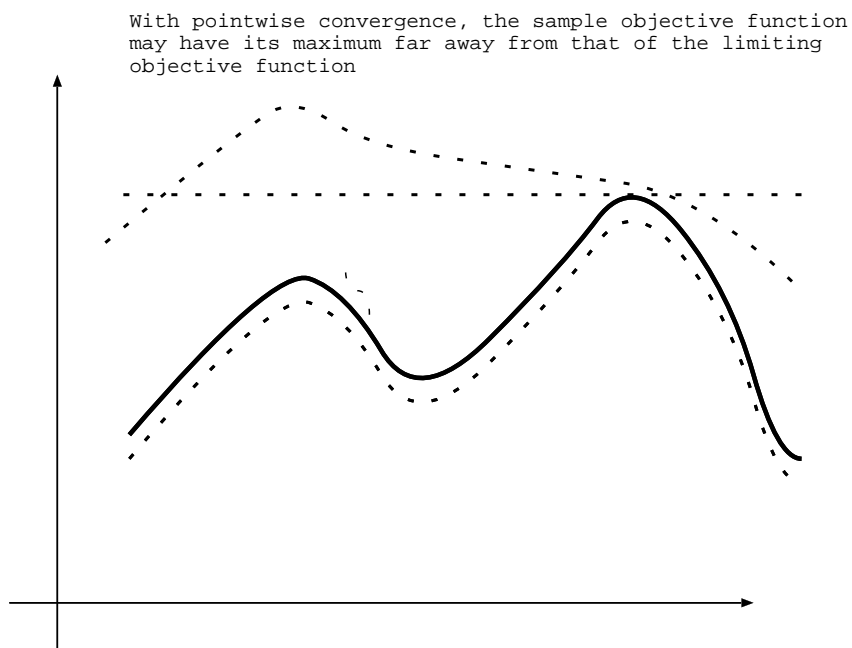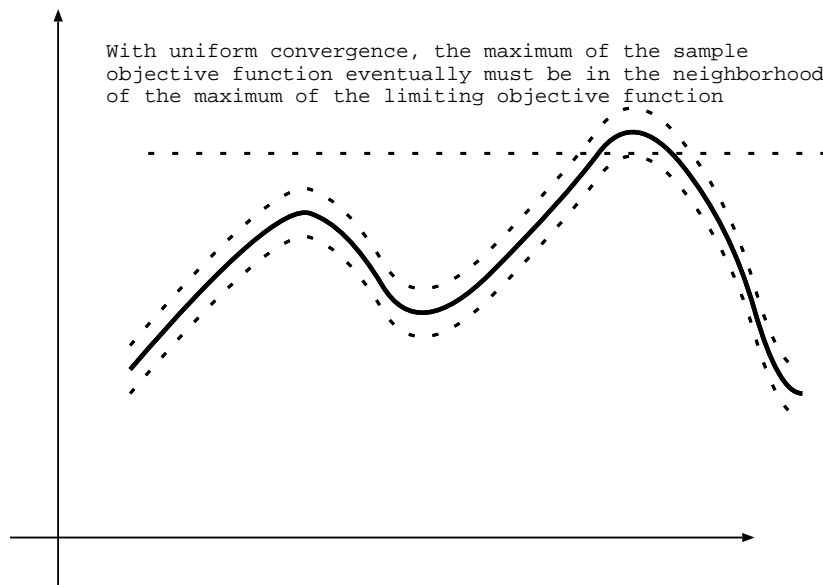
16.3   Example: Consistency of Least Squares

We suppose that data is generated by random sampling of (y, w), where $y_t = \alpha^0 + \beta^0 w_t$

$+\varepsilon_t$. (w_t, ε_t) has the common distribution function $\mu_w \mu_\varepsilon$ (w and ε are independent) with

support $\mathcal{W} \times \mathcal{E}$. Suppose that the variances σ_w^2 and σ_ε^2 are finite. Let $\theta^0 = (\alpha^0, \beta^0)' \in$ Θ, for which $\overline{\Theta}$ is compact. Let $x_t = (1, w_t)'$, so we can write $y_t = x_t'\theta^0 + \varepsilon_t$. The sample objective function for a sample size n is

$$
\begin{aligned}
s_n(\theta) &= 1/n \sum_{t=1}^{n} \left(y_t - x_t'\theta\right)^2 = 1/n \sum_{i=1}^{n} \left(x_t'\theta^0 + \varepsilon_t - x_t'\theta\right)^2 \\
&= 1/n \sum_{t=1}^{n} \left(x_t'\left(\theta^0 - \theta\right)\right)^2 + 2/n \sum_{t=1}^{n} x_t'\left(\theta^0 - \theta\right)\varepsilon_t + 1/n \sum_{t=1}^{n} \varepsilon_t^2
\end{aligned}
$$

Considering the last term, by the SLLN,

$$
1/n \sum_{t=1}^{n} \varepsilon_t^2 \xrightarrow{a.s.} \int_{\mathcal{W}} \int_{\mathcal{E}} \varepsilon^2 d\mu_{\mathcal{W}} d\mu_{\mathcal{E}} = \sigma_\varepsilon^2.
$$

This is completely unaffected by θ, so the pointwise almost sure convergence is also uniform. The same argument holds for the second term since $E(\varepsilon) = 0$ and w and ε are independent. Finally, for the first term, for a given θ

$$
\begin{aligned}
1/n \sum_{t=1}^{n} \left(x_t'\left(\theta^0 - \theta\right)\right)^2 &\xrightarrow{a.s.} \int_{\mathcal{W}} \left(x'\left(\theta^0 - \theta\right)\right)^2 d\mu_{\mathcal{W}} \qquad (16) \\
&= \left(\alpha^0 - \alpha\right)^2 + 2\left(\alpha^0 - \alpha\right)\left(\beta^0 - \beta\right)\int_{\mathcal{W}} w d\mu_{\mathcal{W}} + \left(\beta^0 - \beta\right)^2 \int_{\mathcal{W}} w^2 d\mu_{\mathcal{W}} \\
&= \left(\alpha^0 - \alpha\right)^2 + 2\left(\alpha^0 - \alpha\right)\left(\beta^0 - \beta\right)E(w) + \left(\beta^0 - \beta\right)^2 E\left(w^2\right)
\end{aligned}
$$

This convergence is also uniform, by the previous argument (that is, the expectations are not functions of parameters). So

$$
s_\infty(\theta) = \left(\alpha^0 - \alpha\right)^2 + 2\left(\alpha^0 - \alpha\right)\left(\beta^0 - \beta\right)E(w) + \left(\beta^0 - \beta\right)^2 E\left(w^2\right) + \sigma_\varepsilon^2
$$

A minimizer of this is clearly $\alpha = \alpha^0, \beta = \beta^0$.

Exercise 56 *Show that in order for the above solution to be unique it is necessary*

that $E(w^2) \neq 0$. Discuss the relationship between this condition and the problem of colinearity of regressors.

This example shows that Theorem 55 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. Also, the way we moved from $\overset{a.s.}{\to}$ to $\overset{u.a.s.}{\to}$ is a special case that relies on being able to neatly separate parameters and random variables. This won't always work. For this reason, we need a uniform strong law of large numbers in order to verify assumption (2) of Theorem 55. The following theorem is from Davidson, pg. 337.

Theorem 57 [Uniform Strong LLN] *Let $\{G_n(\theta)\}$ be a sequence of stochastic real-valued functions on a totally-bounded metric space (Θ, ρ). Then*

$$\sup_{\theta \in \Theta} |G_n(\theta)| \overset{a.s.}{\to} 0$$

if and only if

(a) $G_n(\theta) \overset{a.s.}{\to} 0$ for each $\theta \in \Theta_0$, where Θ_0 is a dense subset of Θ and

(b) $\{G_n(\theta)\}$ is strongly stochastically equicontinuous..

- Assumption (a) is simply pointwise almost sure convergence.

- For present purposes, just take $\Theta_0 = \Theta$, so don't worry about "dense."

- The metric space we are interested in now is simply $\Theta \subset \mathfrak{R}^K$, using the Euclidean norm.

- What is required is pointwise almost sure convergence and strong stochastic equicontinuity. Pointwise almost sure convergence comes from one of the usual SLLN's.

244

Strong stochastic equicontinuity requires that for $\forall \varepsilon > 0, \exists \delta > 0$ such that

$$\Pr \left(\limsup_{n \to \infty} \sup_{\theta \in \Theta} \sup_{\theta' \in S(\theta, \delta)} \left| G_n(\theta) - G_n(\theta') \right| > \varepsilon \right) = 0$$

Here, $S(\theta, \delta)$ is a $\delta-$ neighborhood of θ, *i.e.*, $S(\theta, \delta) = \{\theta^* : \rho(\theta^*, \theta) < \delta\}$.

This definition is basically requiring uniform continuity throughout Θ, with probability one as $n \to \infty$.

- A stronger condition that implies this one is: $G_n(\theta)$ is uniformly continuous in θ for all n, and also bounded for all n (w.p.1).

- Strong stochastic equicontinuity is basically a probabilistic, asymptotic version of uniform continuity.

- Note: a function that is continuous on a compact set is uniformly continuous.

- Taken together, these results imply that with a compact parameter space and a continuous, bounded objective function, pointwise almost sure convergence implies uniform almost sure 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.

- The limiting objective function can be continuous in θ even if $s_n(\theta)$ is discontinuous, since discontinuities can be smoothed out as we take expectations over the data. We'll see an example in the section on estimation by simulation methods. **ADD AN EXAMPLE HERE, ref. to ANDREWS**

16.4 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 58 [Asymptotic normality of e.e.] *In addition to the assumptions of Theorem 55, 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) $\{\mathcal{J}_n(\theta_n)\} \overset{a.s.}{\to} \mathcal{J}_\infty(\theta^0)$, a finite negative definite matrix, for any sequence $\{\theta_n\}$ that converges almost surely to θ^0.

(c) $\sqrt{n} D_\theta s_n(\theta^0) \overset{d}{\to} N\left[0, I_\infty(\theta^0)\right]$, where $I_\infty(\theta^0) = \lim_{n\to\infty} Var \sqrt{n} D_\theta s_n(\theta^0)$

Then $\sqrt{n}\left(\hat{\theta} - \theta^0\right) \overset{d}{\to} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1} I_\infty(\theta^0) \mathcal{J}_\infty(\theta^0)^{-1}\right]$

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 $\theta^* = \lambda\hat{\theta} + (1-\lambda)\theta^0$, $0 \le \lambda \le 1$.

- 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.

246

- Also, since θ^* is between $\hat{\theta}_n$ and θ^0, and since $\hat{\theta}_n \overset{a.s.}{\to} \theta^0$, assumption (b) gives

$$D_\theta^2 s_n(\theta^*) \overset{a.s.}{\to} \mathcal{J}_\infty(\theta^0)$$

So

$$0 = D_\theta s_n(\theta^0) + \left[\mathcal{J}_\infty(\theta^0) + o_p(1) \right] \left(\hat{\theta} - \theta^0 \right)$$

And

$$0 = \sqrt{n} D_\theta s_n(\theta^0) + \left[\mathcal{J}_\infty(\theta^0) + o_p(1) \right] \sqrt{n} \left(\hat{\theta} - \theta^0 \right)$$

Now $\mathcal{J}_\infty(\theta^0)$ is a finite negative definite matrix, so the $o_p(1)$ term is asymptotically irrelevant next to $\mathcal{J}_\infty(\theta^0)$, so we can write

$$0 \overset{a}{=} \sqrt{n} D_\theta s_n(\theta^0) + \mathcal{J}_\infty(\theta^0) \sqrt{n} \left(\hat{\theta} - \theta^0 \right)$$

$$\sqrt{n} \left(\hat{\theta} - \theta^0 \right) \overset{a}{=} -\mathcal{J}_\infty(\theta^0)^{-1} \sqrt{n} D_\theta s_n(\theta^0)$$

Because of assumption (c), and the formula for the variance of a linear combination of r.v.'s,

$$\sqrt{n} \left(\hat{\theta} - \theta^0 \right) \overset{d}{\to} N \left[0, \mathcal{J}_\infty(\theta^0)^{-1} I_\infty(\theta^0) \mathcal{J}_\infty(\theta^0)^{-1} \right]$$

- Assumption (b) is not implied by the Slutsky theorem. The Slutsky theorem says that $g(x_n) \overset{a.s.}{\to} g(x)$ if $x_n \to x$ and $g(\cdot)$ is continuous at x. However, the function $g(\cdot)$ can't depend on n to use this theorem. In our case $\mathcal{J}_n(\theta_n)$ *is a function of n.* A theorem which applies (Amemiya, Ch. 4) is

Theorem 59 *If $g_n(\theta)$ converges uniformly almost surely to a nonstochastic function $g_\infty(\theta)$ uniformly on an open neighborhood of θ^0, then $g_n(\hat{\theta}) \overset{a.s.}{\to} g_\infty(\theta^0)$ if $g_\infty(\theta^0)$ is continuous at θ^0 and $\hat{\theta} \overset{a.s.}{\to} \theta^0$.*

- To apply this to the second derivatives, sufficient conditions would be that the second derivatives be strongly stochastically equicontinuous on a neighborhood of θ^0, and that an ordinary LLN applies to the derivatives when evaluated at $\theta \in N(\theta^0)$.

- Stronger conditions that imply this are as above: continuous and bounded second derivatives in a neighborhood of θ^0.

- **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, $D^2_\theta s_n(\theta)$ 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 $D^2_\theta s_n(\theta^0)$, $\mathcal{J}_\infty(\theta^0) = O(1)$, as we saw in Example 52. On the other hand, assumption (c): $\sqrt{n} D_\theta s_n(\theta^0) \xrightarrow{d} N\left[0, I_\infty(\theta^0)\right]$ means that

$$\sqrt{n} D_\theta s_n(\theta^0) = O_p(1),$$

where we use the result of Example 50. If we were to omit the $\sqrt{n}$, we'd have

$$D_\theta s_n(\theta^0) = n^{-\frac{1}{2}} O_p(1)$$
$$= O_p\left(n^{-\frac{1}{2}}\right)$$

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.

16.5 Example: Binary response models.

Binary response models arise in a variety of contexts. The referendum contingent valuation (CV) method of infering the social value of a project provides a simple example. This example is a special case of more general discrete choice (or binary response) models. Individuals are asked if the would pay an amount A for provision of a project. Indirect utility in the base case (no project) is $v^0(m, \mathbf{z}) + \varepsilon^0$, where m is income and $\mathbf{z}$ is a vector of other variables such as prices, personal characteristics, *etc.* After provision, utility is $v^1(m, \mathbf{z}) + \varepsilon^1$. The random terms $\varepsilon^i, i = 1, 2$, reflect variations of preferences in the population. With this, an individual agrees[1] to pay A if

$$\varepsilon^0 - \varepsilon^1 < v^1(m - A, \mathbf{z}) - v^0(m, \mathbf{z})$$

Define $\varepsilon = \varepsilon^0 - \varepsilon^1$, let $\mathbf{w}$ collect m and $\mathbf{z}$, and let $\Delta v(\mathbf{w}, A) = v^1(m - A, \mathbf{z}) - v^0(m, \mathbf{z})$. Define $y = 1$ if the consumer agrees to pay A for the change, $y = 0$ otherwise. The probability of agreement is

$$\Pr(y = 1) = F_{\varepsilon}\left[\Delta v(\mathbf{w}, A)\right].$$

To simplify notation, define $p(\mathbf{w}, A) \equiv F_{\varepsilon}\left[\Delta v(\mathbf{w}, A)\right]$. To make the example specific, suppose that

$$v^1(m, \mathbf{z}) = \alpha - \beta m$$
$$v^0(m, \mathbf{z}) = -\beta m$$

---

[1]We assume here that responses are truthful, that is there is no strategic behavior and that individuals are able to order their preferences in this hypothetical situation.

and ε^0 and ε^1 are i.i.d. extreme value random variables. That is, utility depends only on income, preferences in both states are homothetic, and a specific distributional assumption is made on the distribution of preferences in the population. With these assumptions (the details are unimportant here, see articles by D. McFadden for details) it can be shown that

$$p(A, \theta) = \Lambda(\alpha + \beta A),$$

where $\Lambda(z)$ is the logistic distribution function

$$\Lambda(z) = (1 + \exp(-z))^{-1}.$$

This is the simple logit model: the choice probability is the logit function of a linear in parameters function.

Another simple example is a probit threshold-crossing model. Assume that

$$y^* = x\prime\beta - \varepsilon$$
$$y = 1(y^* > 0)$$
$$\varepsilon \sim N(0, 1)$$

Here, y^* is an unobserved (latent) continuous variable, and y is a binary variable that indicates whether y^* is negative or positive. Then $Pr(y = 1) = Pr(\varepsilon < x\beta) = \Phi(x\beta)$, where

$$\Phi(\bullet) = \int_{-\infty}^{x\beta} (2\pi)^{-1/2} \exp(-\frac{\varepsilon^2}{2}) d\varepsilon$$

is the standard normal distribution function.

In general, a binary response model will require that the choice probability be

parameterized in some form. For a vector of explanatory variables x, the response probability will be parameterized in some manner

$$Pr(y = 1|x) = p(x, \theta)$$

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

$$
\begin{aligned}
s_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i \ln p(x_i, \theta) + (1 - y_i) \ln[1 - p(x_i, \theta)]) \\
&\equiv \frac{1}{n}\sum_{i=1}^{n} s(y_i, x_i, \theta).
\end{aligned}
\tag{17}
$$

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

$$\mathcal{E}_{y|x}\{y \ln p(x, \theta) + (1 - y) \ln[1 - p(x, \theta)]\} = p(x, \theta^0) \ln p(x, \theta) + [1 - p(x, \theta^0)] \ln[1 - p(x, \theta)].$$

Next taking expectation over x we get the limiting objective function

$$s_\infty(\theta) = \int_{\mathcal{X}}\{p(x, \theta^0) \ln p(x, \theta) + [1 - p(x, \theta^0)] \ln[1 - p(x, \theta)]\}\mu(x)dx, \tag{18}$$

251

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

$$\sqrt{n} \left(\hat{\theta} - \theta^0 \right) \xrightarrow{d} N \left[0, \mathcal{J}_\infty(\theta^0)^{-1} I_\infty(\theta^0) \mathcal{J}_\infty(\theta^0)^{-1} \right].$$

In the case of i.i.d. observations $I_\infty(\theta^0) = \lim_{n \to \infty} Var \sqrt{n} D_\theta s_n(\theta^0)$ 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:

$$\lim_{n\to\infty} Var\sqrt{n}D_\theta s_n(\theta^0) = \lim_{n\to\infty} Var\sqrt{n}D_\theta \frac{1}{n}\sum_t s(\theta^0)$$

$$= \lim_{n\to\infty} Var\frac{1}{\sqrt{n}}D_\theta \sum_t s(\theta^0)$$

$$= \lim_{n\to\infty} \frac{1}{n}Var\sum_t D_\theta s(\theta^0)$$

$$= \lim_{n\to\infty} VarD_\theta s(\theta^0)$$

$$= VarD_\theta s(\theta^0)$$

So we get

$$I_\infty(\theta^0) = \mathcal{E}\left\{\frac{\partial}{\partial\theta}s(y,x,\theta^0)\frac{\partial}{\partial\theta'}s(y,x,\theta^0)\right\}.$$

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

253

$$\frac{\partial}{\partial\theta}p(x,\theta) \;=\; \left(1+\exp(-\mathbf{x}'\theta)\right)^{-2}\exp(-\mathbf{x}'\theta)\mathbf{x}$$

$$=\; \left(1+\exp(-\mathbf{x}'\theta)\right)^{-1}\frac{\exp(-\mathbf{x}'\theta)}{1+\exp(-\mathbf{x}'\theta)}\mathbf{x}$$

$$=\; p(x,\theta)\left(1-p(x,\theta)\right)\mathbf{x}$$

$$=\; \left(p(x,\theta)-p(x,\theta)^2\right)\mathbf{x}.$$

So

$$\frac{\partial}{\partial\theta}s(y,x,\theta^0) \;=\; \left[y-p(x,\theta^0)\right]\mathbf{x} \tag{19}$$

$$\frac{\partial^2}{\partial\theta\partial\theta'}s(\theta^0) \;=\; -\left[p(x,\theta^0)-p(x,\theta^0)^2\right]\mathbf{x}\mathbf{x}'. $$

Taking expectations over y then $\mathbf{x}$ gives

$$I_\infty(\theta^0) \;=\; \int\left[y^2-2p(x,\theta^0)p(x,\theta^0)+p(x,\theta^0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx \tag{20}$$

$$=\; \int\left[p(x,\theta^0)-p(x,\theta^0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx. \tag{21}$$

where we use the fact that $E(y)=E(y^2)=p(\mathbf{x},\theta^0)$. Likewise,

$$\mathcal{J}_\infty(\theta^0) = -\int\left[p(x,\theta^0)-p(x,\theta^0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx. \tag{22}$$

Note that we arrive at the expected result: the information matrix equality holds (that is, $\mathcal{J}_\infty(\theta^0) = -I_\infty(\theta^0)$). With this,

$$\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left[0,\mathcal{J}_\infty(\theta^0)^{-1}I_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right]$$

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

$$\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left[0, I_\infty(\theta^0)^{-1}\right].$$

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.

16.6   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

$$y_i = h(x^0, \theta^0) + (x_i - x_0)' \frac{\partial h(x_0, \theta^0)}{\partial x} + v_i$$

Define

$$\alpha^* = h(x_0, \theta^0) - x_0' \frac{\partial h(x^0, \theta^0)}{\partial x}$$

$$\beta^* = \frac{\partial h(x_0, \theta^0)}{\partial x}$$

Given this, one might try to estimate α^* and β^* by applying OLS to

$$y_i = \alpha + \beta x_i + v_i$$

- 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 $\gamma = (\alpha, \beta')'$.

$$\hat{\gamma} = \arg\min s_n(\gamma) = \frac{1}{n} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

The objective function converges to its expectation

$$s_n(\gamma) \overset{u.a.s.}{\to} s_\infty(\gamma) = \mathcal{E}_X \mathcal{E}_{Y|X} (y - \alpha - \beta x)^2$$

and $\hat{\gamma}$ converges *a.s.* to the γ^0 that minimizes $s_\infty(\gamma)$:

$$\gamma^0 = \arg\min \mathcal{E}_X \mathcal{E}_{Y|X} (y - \alpha - \beta x)^2$$

Noting that

$$\mathcal{E}_X \mathcal{E}_{Y|X} \left(y - \alpha - x'\beta \right)^2 = \mathcal{E}_X \mathcal{E}_{Y|X} \left(h(x,\theta^0) + \varepsilon - \alpha - \beta x \right)^2$$

$$= \sigma^2 + \mathcal{E}_X \left(h(x,\theta^0) - \alpha - \beta x \right)^2$$

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.
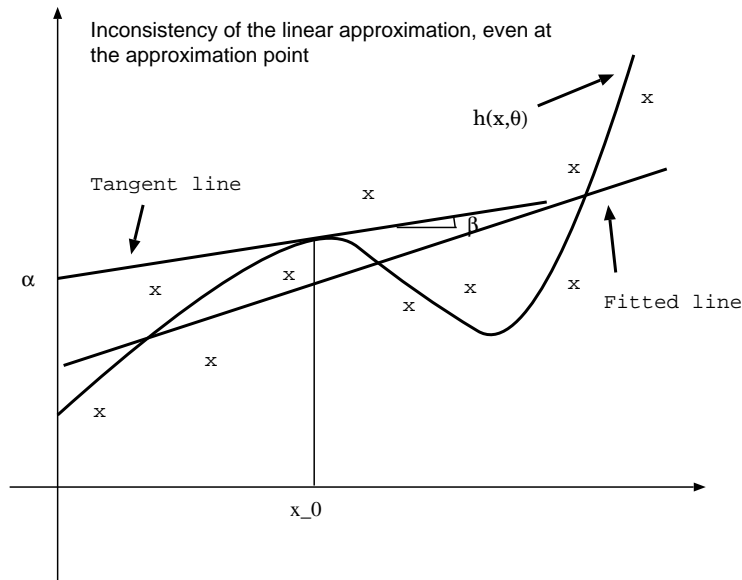
- It is clear that the tangent line does not minimize MSE, since, for example, if $h(x,\theta^0)$ is concave, all errors between the tangent line and the true function are negative.

- 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).

257

- 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.

17 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).

There 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 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.

17.1 Search

See Hamilton. Note, 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 $100^{10} = 1\,00000\,00000\,00000\,00000$ 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 the age of the earth. The search method is a very reasonable choice if K is small, but it quickly becomes infeasible if K is moderate or large.

Search in two dimensions with refinement

The maximizing point
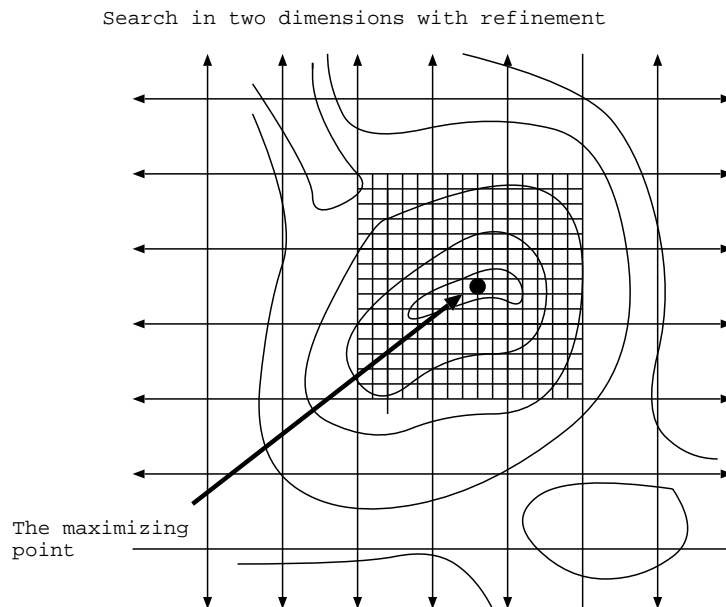
17.2 Derivative-based methods

17.2.1 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

$$\exists a : \frac{\partial s(\theta + ad)}{\partial a} > 0$$

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 de-

grees.)

- 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.

- Note also that this gives no guarantees to find a global maximum.

17.2.2 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....Draw banana function.

17.2.3 Newton-Raphson

The Newton-Raphson 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, e.g, 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

$$\theta^{k+1} = \theta^k - H(\theta^k)^{-1} g(\theta^k)$$

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 ma-

263

trix 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.

- 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.

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, and in fact, more than about 6-10 decimals of precision is usually infeasible. 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:

$$|g_j(\theta^k) - g_j(\theta^{k-1})| < \varepsilon_4, \forall j$$

- 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.**

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 Mathematica or *Scientific WorkPlace* to calculate analytic derivatives. Example: *Scientific WorkPlace* can be used to find that

$$\frac{\partial}{\partial \theta} \arctan \theta = \frac{1}{1 + \theta^2}$$

which I certainly didn't know before writing this example. Hal Varian has a book that discusses the use of Mathematica in this context in detail. Analytic derivatives usually lead to a much faster program, and are more accurate than numeric derivatives.

- Numeric derivatives lead to much slower estimation than analytic derivatives.

- Numeric 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 Davidon-Fletcher-Powell, see GAUSS OPTMUM) that use the sequential gradient evaluations to build up an approximation to the Hessian. The iterations are faster for this reason since the actual Hessian isn't calculated, but more iterations usually are required for convergence.

- Switching between algorithms during iterations is sometimes useful.

17.3 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.

18 Generalized method of moments (GMM)

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.

18.1 Definition

We've already seen one example of GMM in the introduction, based upon the χ^2 distribution. Consider the following example based upon the t-distribution. The density function of a t-distributed r.v. Y_t is

$$f_{Y_t}(y_t, \theta^0) = \frac{\Gamma\left[\left(\theta^0 + 1\right)/2\right]}{(\pi\theta^0)^{1/2}\,\Gamma\left(\theta^0/2\right)} \left[1 + \left(y_t^2/\theta^0\right)\right]^{-\left(\theta^0+1\right)/2}$$

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.

- Continuing with the example, 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 $\theta^0 > 2$).

- Using the notation introduced previously, define a moment condition $m_{1t}(\theta) = \theta/(\theta-2) - y_t^2$ and $m_1(\theta) = 1/n \sum_{t=1}^{n} m_{1t}(\theta) = \theta/(\theta-2) - 1/n \sum_{t=1}^{n} y_t^2$. As before, when evaluated at the true parameter value θ^0, both $\mathcal{E}_{\theta^0}\left[m_{1t}(\theta^0)\right] = 0$ and $\mathcal{E}_{\theta^0}\left[m_1(\theta^0)\right] = 0$.

- Choosing $\hat{\theta}$ to set $m_1(\hat{\theta}) \equiv 0$ yields a MM estimator:

$$\hat{\theta} = \frac{2}{1 - \frac{n}{\sum_i y_i^2}} \tag{23}$$

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.

- 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

$$\mu_4 \equiv E(y_t^4) = \frac{3\left(\theta^0\right)^2}{(\theta^0 - 2)(\theta^0 - 4)},$$

provided $\theta^0 > 4$. We can define a second moment condition

$$m_2(\theta) = \frac{3\left(\theta\right)^2}{(\theta-2)(\theta-4)} - \frac{1}{n}\sum_{t=1}^{n} y_t^4$$

- 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 23.

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).

- As before, set $m_n(\theta) = (m_1(\theta), m_2(\theta))'$. The n subscript is used to indicate the sample size. Note that $m(\theta^0) = O_p(n^{-1/2})$, since it is an average of centered random variables, whereas $m(\theta) = O_p(1)$, $\theta \neq \theta^0$, where expectations are taken using the true distribution with parameter θ^0. This is the fundamental reason that GMM is consistent.

- A GMM estimator requires defining a measure of distance, $d(m(\theta))$. A popular choice (for reasons noted below) is to set $d(m(\theta)) = m'W_n m$, and we minimize $s_n(\theta) = m(\theta)'W_n m(\theta)$. We assume W_n converges to a finite positive definite matrix.

- In general, assume we have g moment conditions, so $m(\theta)$ is a g-vector and W is a $g \times g$ matrix.

For the purposes of this course, the following definition of the GMM estimator is sufficiently general:

Definition 60 *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 $m_n(\theta) = \sum_{t=1}^n m_t(\theta)$ is a g-vector, $g \geq K$, with $\mathcal{E}m(\theta^0) = 0$, 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?* The answer is simple - 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.*

270

The price for robustness is loss of efficiency with respect to the MLE estimator. Keep in mind that the true distribution is <u>not known</u> so if we erroneously specify a distribution and estimate by MLE, the estimator will be inconsistent in general (not always).

18.2 Identification

In the consistency proof (Theorem 55) 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 $\mathcal{E}m_n(\theta^0) = 0$ 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. This and the assumption that $W_n \overset{a.s.}{\to} W_\infty$, 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.

18.3 Consistency

We simply assume that the assumptions of Theorem 55 hold, so the GMM estimator is strongly consistent.

18.4 Asymptotic normality

We also simply assume that the conditions of Theorem 58 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 58, we have

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1} I_\infty(\theta^0) \mathcal{J}_\infty(\theta^0)^{-1}\right]$$

where $\mathcal{J}_\infty(\theta^0)$ is the almost sure limit of $\frac{\partial^2}{\partial\theta\partial\theta'} s_n(\theta)$ and $I_\infty(\theta^0) = \lim_{n\to\infty} Var\sqrt{n}\frac{\partial}{\partial\theta} s_n(\theta^0)$. 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,

$$\frac{\partial}{\partial\theta} s(\theta) = 2\left[\frac{\partial}{\partial\theta} m'_n(\theta)\right] W_n m_n(\theta)$$

Define the $K \times g$ matrix

$$D_n(\theta) \equiv \frac{\partial}{\partial\theta} m'_n(\theta),$$

so

$$\frac{\partial}{\partial\theta} s(\theta) = 2D(\theta)W m(\theta).$$

(Note that $D_n(\theta)$, W_n and $m_n(\theta)$ all depend on the sample size n, but we will often simplify the notation to D, W, and m).

To take second derivatives, let D_i be the $i-$ th row of $D(\theta)$. Using the product rule,

$$
\begin{aligned}
\frac{\partial}{\partial\theta'\partial\theta_i} s(\theta) &= \frac{\partial}{\partial\theta'} 2D_i(\theta)W_n m(\theta)\\
&= 2D_i W D' + 2m'W\left[\frac{\partial}{\partial\theta'} D'_i\right]
\end{aligned}
$$

272

When evaluating the term

$$2m(\theta)'W\left[\frac{\partial}{\partial\theta'}D(\theta)'_i\right]$$

at θ^0, assume that $\frac{\partial}{\partial\theta'}D(\theta)'_i$ satisfies a LLN, so that it converges almost surely to a finite limit. In this case, we have

$$2m(\theta^0)'W\left[\frac{\partial}{\partial\theta'}D(\theta^0)'_i\right]\overset{a.s.}{\to}0,$$

since $m(\theta^0)=o_p(1)$, $W\overset{a.s.}{\to}W_\infty$.

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 $I_\infty(\theta^0)$,

$$
\begin{aligned}
I_\infty(\theta^0) &= \lim_{n\to\infty}Var\sqrt{n}\frac{\partial}{\partial\theta}s_n(\theta^0)\\
&= \lim_{n\to\infty}\mathcal{E}4nD_nW_nm(\theta^0)m(\theta)'\,W_nD'_n\\
&= \lim_{n\to\infty}\mathcal{E}4D_nW_n\left\{\sqrt{n}m(\theta^0)\right\}\left\{\sqrt{n}m(\theta)'\right\}\,W_nD'_n
\end{aligned}
$$

since $\mathcal{E}m(\theta^0)=0$ by assumption (that is, the moment conditions are correctly specified, by assumption, so there is no need to subtract the mean) . 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,

$$\sqrt{n}m(\theta^0)\overset{d}{\to}N(0,\Omega_\infty),$$

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

$$I_\infty(\theta^0) = 4D_\infty W_\infty \Omega_\infty W_\infty D_\infty'$$

Using these results, the asymptotic normality theorem 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, $\rho(D_\infty)=k$.

18.5 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*.

- 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, since $V(P\varepsilon) = PV(\varepsilon)P' = P\Omega P' = P(P'P)^{-1}P' = PP^{-1}(P')^{-1}P' = I_n$. (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, 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 61 *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 $W_n \overset{a.s}{\to} W_\infty = \Omega_\infty^{-1}$, 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, for any choice such that $W_\infty \neq \Omega_\infty^{-1}$, consider the difference of the inverses of the variances when $W = \Omega^{-1}$ versus when W is some arbitrary positive definite matrix:

$$
\begin{aligned}
&\left(D_\infty \Omega_\infty^{-1} D_\infty'\right) - \left(D_\infty W_\infty D_\infty'\right) \left[D_\infty W_\infty \Omega_\infty W_\infty D_\infty'\right]^{-1} \left(D_\infty W_\infty D_\infty'\right) \\
&= D_\infty \Omega_\infty^{-1/2} \left[I - \Omega_\infty^{1/2} \left(W_\infty D_\infty'\right) \left[D_\infty W_\infty \Omega_\infty W_\infty D_\infty'\right]^{-1} D_\infty W_\infty \Omega_\infty^{1/2}\right] \Omega_\infty^{-1/2} D_\infty'
\end{aligned}
$$

as can be verified by multiplication. The term in brackets is idempotent, which is also easy to check by multiplication, and is therefore positive semidefinite. A quadratic form in a positive semidefinite matrix is also positive semidefinite. The difference of the inverses of the variances is positive semidefinite, which implies that the difference of the variances is negative semidefinite, which proves the theorem.

The result

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}\right] \tag{24}$$

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. We now turn to estimation of Ω_∞.

18.6 Estimation of the variance-covariance matrix

(See Hamilton Ch. 10, pp. 261-2 and 280-84)[*].

In the case that we with to use the optimal weighting matrix, we need an estimate of Ω_∞, the variance-covariance matrix of the moment conditions $m = \sum_{t=1}^n m_t$. We assume that m_t is covariance stationary (the covariance between m_t and m_{t-s} does not depend on t). In general, we expect that:

- m_t will be autocorrelated ($\Gamma_s = \mathcal{E}(m_t m'_{t-s}) \neq 0$) Note that this autocovariance does not depend on t.

- contemporaneously correlated ($\mathcal{E}(m_{it} m_{jt}) \neq 0$)

- and heteroscedastic ($\mathcal{E}(m_{it}^2) = \sigma_i^2$, which depends upon i).

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_∞. *These estimators should work well asymptotically, but there's no guarantee that they will work well in small samples, since Ω_∞ may be estimated very imprecisely*. This is analogous to the fact that feasible GLS does not always work better than OLS with small samples. An interesting topic for research (actually, I'm pretty sure this has already been done) would be to compare the performance of GMM using the estimators discussed below with choices of W that may be inconsistent, but precisely estimable with small samples.

Define the $v - th$ autocovariance of the moment conditions $\Gamma_v = \mathcal{E}(m_t m'_{t-s})$. Note 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 $n \rightarrow \infty$.

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 $n \rightarrow \infty$ 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 is

278

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.

18.6.1 Newey-West covariance estimator

The Newey-West estimator (*Econometrica*, 1987) solves the problem of possible non-positive 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 $n^{-1/4}q \rightarrow 0$. 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.

18.7   Estimation using conditional moments

If the above VAR model *does* succeed in removing unmodeled heteroscedasticity and autocorrelation, might this imply that this information is not being used efficiently in estimation? In other words, since the performance of GMM depends on *which* moment conditions are used, if the set of selected moments exhibits heteroscedasticity and autocorrelation, can't we use this information, *a la* GLS, to guide us in selecting a better set of moment conditions to improve efficiency? The answer to this may not be so clear when moments are defined unconditionally, but it can be analyzed more carefully when the moments used in estimation are derived from 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

$$\mathcal{E}_{Y|X}Y = \int Yf(Y|X)dY = 0$$

Then the unconditional expectation of the product of Y and a function $g(X)$ of X is also zero. The unconditional expectation is

$$\mathcal{E}Yg(X) = \int_X \left(\int_{\mathcal{Y}} Yg(X)f(Y,X)dY \right) dX.$$

This can be factored into a conditional expectation and an expectation w.r.t. the marginal density of X :

$$\mathcal{E}Yg(X) = \int_X \left(\int_{\mathcal{Y}} Yg(X)f(Y|X)dY \right) f(X)dX.$$

Since $g(X)$ doesn't depend on Y it can be pulled out of the integral

$$\mathcal{E}Yg(X) = \int_X \left(\int_{\mathcal{Y}} Yf(Y|X)dY \right) g(X)f(X)dX.$$

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 $k(x_t,\theta)$,

$$\mathcal{E}_\theta K(y_t,x_t)|I_t = k(x_t,\theta).$$

Then the function

$$h_t(\theta) = K(y_t, x_t) - k(x_t, \theta)$$

has conditional expectation equal to zero

$$\mathcal{E}_\theta h_t(\theta) | I_t = 0.$$

This is a scalar moment condition, which wouldn't be sufficient to identify a K $(K > 1)$ dimensional parameter θ. However, the above result allows us to form various unconditional expectations

$$m_t(\theta) = Z(w_t) h_t(\theta)$$

where $Z(w_t)$ is a $gx1$-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

$$Z_n = \begin{bmatrix} Z_1(w_1) & Z_2(w_1) & \cdots & Z_g(w_1) \\ Z_1(w_2) & Z_2(w_2) & & Z_g(w_2) \\ \vdots & & & \vdots \\ Z_1(w_n) & Z_2(w_n) & \cdots & Z_g(w_n) \end{bmatrix}$$

$$= \begin{bmatrix} Z_1' \\ Z_2' \\ \\ Z_n' \end{bmatrix}$$

With this we can form the g moment conditions

$$
\begin{aligned}
m_n(\theta) &= \frac{1}{n} Z_n' \begin{bmatrix} h_1(\theta) \\ h_2(\theta) \\ \vdots \\ h_n(\theta) \end{bmatrix} \\
&= \frac{1}{n} Z_n' h_n(\theta) \\
&= \frac{1}{n} \sum_{t=1}^{n} Z_t h_t(\theta) \\
&= \frac{1}{n} \sum_{t=1}^{n} m_t(\theta)
\end{aligned}
$$

where $Z_{(t,\cdot)}$ is the tth row of Z_n. This fits the previous treatment. 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

$$
D_n(\theta) = \frac{1}{n} H_n Z_n.
$$

where H_n is a $K \times n$ matrix that has the derivatives of the individual moment conditions

283

as its columns. Likewise, define the var-cov. of the moment conditions

$$
\begin{aligned}
\Omega_n &= \mathcal{E}\left[nm_n(\theta^0)m_n(\theta^0)'\right] \\
&= \mathcal{E}\left[\frac{1}{n}Z_n'h_n(\theta)h_n(\theta)'Z_n\right] \\
&= Z_n'\mathcal{E}\left(\frac{1}{n}h_n(\theta)h_n(\theta)'\right)Z_n \\
&\equiv Z_n'\frac{\Phi_n}{n}Z_n
\end{aligned}
$$

where we have defined $\Phi_n = Varh_n(\theta)$. Note that matrix is growing with the sample size and 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
$$
V_\infty = \lim_{n\to\infty}\left(\left(\frac{H_nZ_n}{n}\right)\left(\frac{Z_n'\Phi_nZ_n}{n}\right)^{-1}\left(\frac{Z_n'H_n'}{n}\right)\right)^{-1}. \tag{25}
$$

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

$$
V_\infty = \lim_{n\to\infty}\left(\frac{H_n\Phi_n^{-1}H_n'}{n}\right)^{-1}. \tag{26}
$$

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 26 will not apply if approximately optimal instruments are used - it will be necessary to use an estimator based upon equation 25, where the term $\frac{Z'_n \Phi_n Z_n}{n}$ must be estimated consistently apart, for example by the Newey-West procedure.

18.8 Estimation using dynamic moment conditions

Note that dynamic moment conditions simplify the var-cov matrix, but are often harder to formulate. The will be added in future editions. For now, the Hansen application below is enough.

18.9 A specification test

The first order conditions for minimization, using the an estimate of the optimal weighting matrix, are

$$\frac{\partial}{\partial \theta} s(\hat{\theta}) = 2 \left[\frac{\partial}{\partial \theta} m_n' (\hat{\theta}) \right] \hat{\Omega}^{-1} m_n (\hat{\theta}) \equiv 0$$

or

$$D(\hat{\theta})\hat{\Omega}^{-1} m_n(\hat{\theta}) \equiv 0$$

Consider a Taylor expansion of $m(\hat{\theta})$:

$$m(\hat{\theta}) = m_n(\theta^0) + D_n'(\theta^0) \left(\hat{\theta} - \theta^0 \right) + o_p(1)$$

Multiplying by $D(\hat{\theta})\hat{\Omega}^{-1}$ we obtain

$$D(\hat{\theta})\hat{\Omega}^{-1} m(\hat{\theta}) = D(\hat{\theta})\hat{\Omega}^{-1} m_n(\theta^0) + D(\hat{\theta})\hat{\Omega}^{-1} D(\theta^0)' \left(\hat{\theta} - \theta^0 \right) + o_p(1)$$

The lhs is zero, and since $\hat{\theta}$ tends to θ^0 and $\hat{\Omega}$ tends to Ω_∞, we can write

$$D_\infty \Omega_\infty^{-1} m_n(\theta^0) \overset{a}{=} -D_\infty \Omega_\infty^{-1} D_\infty' \left(\hat{\theta} - \theta^0 \right)$$

or

$$\sqrt{n} \left(\hat{\theta} - \theta^0 \right) \overset{a}{=} -\sqrt{n} \left(D_\infty \Omega_\infty^{-1} D_\infty' \right)^{-1} D_\infty \Omega_\infty^{-1} m_n(\theta^0)$$

With this we can write

$$\sqrt{n} m(\hat{\theta}) \overset{a}{=} \sqrt{n} m_n(\theta^0) - \sqrt{n} D_\infty' \left(D_\infty \Omega_\infty^{-1} D_\infty' \right)^{-1} D_\infty \Omega_\infty^{-1} m_n(\theta^0)$$

This last can be written as

$$\sqrt{n}m(\hat{\theta}) \stackrel{a}{=} \sqrt{n}\left(\Omega_\infty^{1/2} - D_\infty'\left(D_\infty\Omega_\infty^{-1}D_\infty'\right)^{-1}D_\infty\Omega_\infty^{-1/2}\right)\Omega_\infty^{-1/2}m_n(\theta^0)$$

Or

$$\sqrt{n}\Omega_\infty^{-1/2}m(\hat{\theta}) \stackrel{a}{=} \sqrt{n}\left(I_g - \Omega_\infty^{-1/2}D_\infty'\left(D_\infty\Omega_\infty^{-1}D_\infty'\right)^{-1}D_\infty\Omega_\infty^{-1/2}\right)\Omega_\infty^{-1/2}m_n(\theta^0)$$

Now

$$\sqrt{n}\Omega_\infty^{-1/2}m_n(\theta^0) \stackrel{d}{\to} N(0, I_g)$$

and one can easily verify that

$$P = \left(I_g - \Omega_\infty^{-1/2}D_\infty'\left(D_\infty\Omega_\infty^{-1}D_\infty'\right)^{-1}D_\infty\Omega_\infty^{-1/2}\right)$$

is idempotent of rank $g - K$, (recall that the rank of an idempotent matrix is equal to its trace) so

$$\left(\sqrt{n}\Omega_\infty^{-1/2}m(\hat{\theta})\right)'\left(\sqrt{n}\Omega_\infty^{-1/2}m(\hat{\theta})\right) = nm(\hat{\theta})'\Omega_\infty^{-1}m(\hat{\theta}) \stackrel{a}{\sim} \chi^2(g - K)$$

Since $\hat{\Omega}$ converges to Ω_∞, we also have

$$nm(\hat{\theta})'\hat{\Omega}^{-1}m(\hat{\theta}) \stackrel{a}{\sim} \chi^2(g - K)$$

or

$$n \cdot s_n(\hat{\theta}) \stackrel{a}{\sim} \chi^2(g - K)$$

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

$$D_\theta s_n(\theta) = D\hat{\Omega}m(\hat{\theta}) \equiv 0.$$

  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. If the sample size is small, it might be better to use bootstrap critical values. That is, draw artificial samples of size n by sampling from the data with replacement. For R bootstrap samples, optimize and calculate the test statistic $n \cdot s(\hat{\theta}^j)$, $j = 1, 2, ..., R$. Define the bootstrap critical value C_b such that $\alpha \cdot 100$ percent of the $s(\hat{\theta}^j)$ exceed the value. Of course, R must be a very large number if $g - K$ is large, in order to determine the critical value with precision. This sort of test has been found to have quite good small sample properties.

18.10   Other estimators interpreted as GMM estimators

18.10.1   OLS with heteroscedasticity of unknown form

Example 62 *White's heteroscedastic consistent varcov estimator for OLS.*

Suppose $\mathbf{y} = \mathbf{X}\beta^0 + \varepsilon$, where $\varepsilon \sim N(0, \Sigma)$, Σ a diagonal matrix.

- The typical approach is to parameterize $\Sigma = \Sigma(\sigma)$, where σ is a finite dimensional parameter vector, and to estimate β and σ jointly (feasible GLS). This will work well if the parameterization of Σ is correct.

- If we're not confident about parameterizing Σ, we can still estimate β consistently by OLS. However, the typical covariance estimator $V(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \hat{\sigma}^2$ will be biased and inconsistent, and will lead to invalid inferences.

By exogeneity of the regressors x_t (a $K \times 1$ column vector) we have $E(x_t \varepsilon_t) = 0$, which suggests the moment condition

$$m_t(\beta) = x_t \left(y_t - \mathbf{x}_t' \beta \right).$$

In this case, we have exact identification (K parameters and K moment conditions). We have

$$m(\beta) = 1/n \sum_t m_t = 1/n \sum_t \mathbf{x}_t y_t - 1/n \sum_t \mathbf{x}_t \mathbf{x}_t' \beta.$$

For any choice of W, $m(\beta)$ will be identically zero at the minimum, due to exact identification. That is, since the number of moment conditions is identical to the number of parameters, the foc imply that $m(\hat{\beta}) \equiv 0$ regardless of W. There is no need to use the "optimal" weighting matrix in this case, an identity matrix works just as well for the

purpose of estimation. Therefore

$$\hat{\beta} = \left(\sum_t \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \sum_t \mathbf{x}_t y_t = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

which is the usual OLS estimator.

The GMM estimator of the asymptotic varcov matrix is $\left(\widehat{D_\infty} \widehat{\Omega}^{-1} \widehat{D_\infty}' \right)^{-1}$. Recall that $\widehat{D_\infty}$ is simply $\frac{\partial}{\partial \theta} m'\left(\hat{\theta}\right)$. In this case

$$\widehat{D_\infty} = -1/n \sum_t \mathbf{x}_t \mathbf{x}_t' = -\mathbf{X}'\mathbf{X}/n.$$

Recall that a possible estimator of Ω is

$$\hat{\Omega} = \widehat{\Gamma_0} + \sum_{v=1}^{n-1} \left(\widehat{\Gamma_v} + \widehat{\Gamma_v'} \right).$$

This is in general inconsistent, but in the present case of nonautocorrelation, it simplifies to

$$\hat{\Omega} = \widehat{\Gamma_0}$$

which has a constant number of elements to estimate, so information *will* accumulate, and consistency obtains. In the present case

$$
\begin{aligned}
\hat{\Omega} &= \widehat{\Gamma_0} = 1/n \left(\sum_{t=1}^{n} \hat{m}_t \hat{m}_t' \right) \\
&= 1/n \left[\sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t' \left(y_t - \mathbf{x}_t'\hat{\beta} \right)^2 \right] \\
&= 1/n \left[\sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t' \hat{\varepsilon}_t^2 \right] \\
&= \frac{\mathbf{X}'\hat{\mathbf{E}}\mathbf{X}}{n}
\end{aligned}
$$

290

where $\hat{\mathbf{E}}$ is an $n \times n$ diagonal matrix with $\hat{\varepsilon}_t^2$ in the position t,t (see the GAUSS command **diagrv** to achieve this).

Therefore, the GMM varcov. estimator, which is consistent, is

$$
\begin{aligned}
\hat{V}\left(\sqrt{n}\left(\hat{\beta}-\beta\right)\right) &= \left\{\left(-\frac{\mathbf{X}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{X}'\hat{\mathbf{E}}\mathbf{X}}{n}^{-1}\right)\left(-\frac{\mathbf{X}'\mathbf{X}}{n}\right)\right\}^{-1} \\
&= \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\hat{\mathbf{E}}\mathbf{X}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}
\end{aligned}
$$

This is the varcov estimator that White (1980) arrived at in an influential article. This estimator is consistent under heteroscedasticity of an unknown form. If there is autocorrelation, the Newey-West estimator can be used to estimate Ω - the rest is the same.

18.10.2 Weighted Least Squares

Consider the previous example of a linear model with heteroscedasticity of unknown form:

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

where Σ is a diagonal matrix.

Now, suppose that the form of Σ is known, so that $\Sigma(\theta^0)$ is a correct parametric specification (which may also depend upon $\mathbf{X}$). In this case, the GLS estimator is

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

This estimator can be interpreted as the solution to the K moment conditions

$$m(\tilde{\beta}) = 1/n \sum_t \frac{\mathbf{x}_t y_t}{\sigma_t(\theta^0)} - 1/n \sum_t \frac{\mathbf{x}_t \mathbf{x}_t'}{\sigma_t(\theta^0)} \tilde{\beta} \equiv 0.$$

That is, the GLS estimator in this case has an obvious representation as a GMM estimator. With autocorrelation, the representation exists but it is a little more complicated. Nevertheless, the idea is the same. There are a few points:

- The (feasible) GLS estimator is known to be asymptotically efficient in the class of linear asymptotically unbiased estimators (Gauss-Markov).

- This means that it is more efficient than the above example of OLS with White's heteroscedastic consistent covariance, which is an alternative GMM estimator.

- This means that the choice of the moment conditions is important to achieve efficiency.

18.10.3   2SLS

Consider the linear model

$$y_t = z_t'\beta + \varepsilon_t,$$

or

$$\mathbf{y} = \mathbf{Z}\beta + \varepsilon$$

using the usual construction, where β is $K \times 1$ and ε_t is i.i.d. Suppose that this equation is one of a system of simultaneous equations, so that z_t contains both endogenous and exogenous variables. Suppose that $\mathbf{x}_t$ is the vector of all exogenous and predetermined variables that are uncorrelated with ε_t (suppose that $\mathbf{x}_t$ is $r \times 1$).

- Define $\hat{\mathbf{Z}}$ as the vector of predictions of $\mathbf{Z}$ when regressed upon $\mathbf{X}$, e.g., $\hat{\mathbf{Z}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$

$$\hat{\mathbf{Z}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$$

- Since $\hat{\mathbf{Z}}$ is a linear combination of the exogenous variables $\mathbf{x}$, $\hat{\mathbf{z}}_t$ must be uncorrelated with ε. This suggests the K-dimensional moment condition $m_t(\beta) = \hat{\mathbf{z}}_t(y_t - \mathbf{z}_t'\beta)$ and so

$$m(\beta) = 1/n\sum_t \hat{\mathbf{z}}_t\left(y_t - \mathbf{z}_t'\beta\right).$$

- 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

$$\hat{\beta} = \left(\sum_t \hat{\mathbf{z}}_t\mathbf{z}_t'\right)^{-1}\sum_t(\hat{\mathbf{z}}_t y_t) = \left(\hat{\mathbf{Z}}'\mathbf{Z}\right)^{-1}\hat{\mathbf{Z}}'\mathbf{y}$$

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 ε_t dependent causes lagged endogenous variables to loose their status as legitimate instruments.

18.10.4 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 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

$$
m_t(\theta) = \begin{bmatrix}
(y_{1t} - f_1(\mathbf{z}_t, \theta_1)) \mathbf{x}_{1t} \\
(y_{2t} - f_2(\mathbf{z}_t, \theta_2)) \mathbf{x}_{2t} \\
\vdots \\
(y_{Gt} - f_G(\mathbf{z}_t, \theta_G)) \mathbf{x}_{Gt}
\end{bmatrix}.
$$

- A note on identification: selection of instruments that ensure identification is a non-trivial problem.

- A note on efficiency: the selected set of instruments has important effects on the efficiency of estimation. Unfortunately there is little theory offering guidance on

what is the optimal set. More on this later.

18.10.5 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

$$\ln L(\theta) = \sum_{t=1}^{n} \ln f(y_t | Y_{t-1}, \theta).$$

Define

$$m_t(Y_t, \theta) \equiv D_\theta \ln f(y_t | Y_{t-1}, \theta)$$

as the *score* of the tth observation. It can be shown that, under the regularity conditions, that the scores have conditional mean zero when evaluated at θ^0 (see notes to Introduction to Econometrics):

$$\mathcal{E}\{m_t(Y_t, \theta^0) | Y_{t-1}\} = 0$$

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 $AV(\hat{\theta}) = \left(D_\infty \Omega^{-1} D_\infty'\right)^{-1}$ (note, AV means asymptotic variance, by which I mean $\lim V\left(\sqrt{n}\left(\hat{\theta} - \theta\right)\right)$.

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 Davidson and MacKinnon, pg. 262-3 for more detail). The fact that the scores are serially uncorrelated implies that Ω can be estimated by the estimator of the 0th autocovariance of the moment conditions:

$$\widehat{\Omega} = 1/n \sum_{t=1}^n m_t(Y_t, \hat{\theta}) m_t(Y_t, \hat{\theta})' = 1/n \sum_{t=1}^n \left[D_\theta \ln f(y_t | Y_{t-1}, \hat{\theta}) \right] \left[D_\theta \ln f(y_t | Y_{t-1}, \hat{\theta}) \right]'$$

Recall from study of ML estimation that the information matrix equality states that

$$E \left\{ \left[D_\theta \ln f(y_t | Y_{t-1}, \theta^0) \right] \left[D_\theta \ln f(y_t | Y_{t-1}, \theta^0) \right]' \right\} = -E \left\{ D_\theta^2 \ln f(y_t | Y_{t-1}, \theta^0) \right\}$$

(*i.e.,* the expectation of the outer product of the gradient is equal to the negative of the expectation of the Hessian. This is a version of the information matrix inequality applied to the individual contributions to the log-likelihood function. It implies the usual form of the information matrix equality, see Davidson and MacKinnon, pg. 264).

This result implies that we can estimate $AV(\hat{\theta})$ in any of three ways:

- The full GMM version:

$$\widehat{AV(\hat{\theta})} = n \left\{ \begin{array}{c} \left\{ \sum_{t=1}^n D_\theta^2 \ln f(y_t | Y_{t-1}, \hat{\theta}) \right\} \cdot \\ \left\{ \sum_{t=1}^n \left[D_\theta \ln f(y_t | Y_{t-1}, \hat{\theta}) \right] \left[D_\theta \ln f(y_t | Y_{t-1}, \hat{\theta}) \right]' \right\}^{-1} \cdot \\ \left\{ \sum_{t=1}^n D_\theta^2 \ln f(y_t | Y_{t-1}, \hat{\theta}) \right\} \end{array} \right\}^{-1}$$

- or the inverse of the negative of the Hessian (since the middle and last term

cancel, except for a minus sign):

$$\widehat{AV(\hat{\theta})} = \left[-1/n \sum_{t=1}^{n} D_{\theta}^2 \ln f(y_t|Y_{t-1}, \hat{\theta}) \right]^{-1},$$

- 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)

$$\widehat{AV(\hat{\theta})} = \left\{ 1/n \sum_{t=1}^{n} \left[D_{\theta} \ln f(y_t|Y_{t-1}, \hat{\theta}) \right] \left[D_{\theta} \ln f(y_t|Y_{t-1}, \hat{\theta}) \right]' \right\}^{-1}$$

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.

18.11 Application: Nonlinear rational expectations

Readings: Hansen and Singleton, 1982[*]; Tauchen, 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.

- We assume a representative consumer maximizes expected discounted utility over an infinite horizon. Utility is temporally additive, and the expected utility

298

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

$$\sum_{s=0}^{\infty} \beta^s \mathcal{E}\left(u(c_{t+s})|I_t\right). \tag{27}$$

- 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 random variables indexed t and earlier.

- Suppose the consumer can invest in an assets. 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. The price of c_t is normalized to 1.

- Net rates of return r_{t+1} are *not known* in period t.

- Investment at time t may be worthwhile since it will lead to the possibility of higher consumption in later periods. However, current investment reduces current consumption.

A partial set of necessary conditions for utility maximization have the form:

$$u'(c_t) = \beta \mathcal{E}\left\{(1 + r_{t+1})u'(c_{t+1})|I_t\right\}. \tag{28}$$

To see that the condition is necessary, suppose that the lhs < rhs. Then by reducing current consumption marginally would cause equation 27 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 utility function is not maximized.

- To use this we need to choose the functional form of utility. A constant relative risk aversion form is

$$u(c_t) = \frac{c_t^\gamma}{\gamma}$$

where $1-\gamma$ is the coefficient of relative risk aversion ($\gamma < 1$). With this form,

$$u'(c_t) = c_t^{\gamma-1}$$

so the foc are

$$c_t^{\gamma-1} = \beta \mathcal{E}\left\{(1+r_{t+1})c_{t+1}^{\gamma-1}|I_t\right\}$$

While it is true that

$$\mathcal{E}\left(c_t^{\gamma-1} - \beta\left\{(1+r_{t+1})c_{t+1}^{\gamma-1}|I_t\right\}\right) = 0$$

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 $c_t^{\gamma-1}$

$$1 - \beta \mathcal{E}\left\{(1+r_{t+1})\left(\frac{c_{t+1}}{c_t}\right)^{\gamma-1}|I_t\right\} = 0$$

(note that c_t can be passed though the conditional expectation since c_t is chosen based only upon information available in time t).

- Suppose that $\mathbf{x}_t$ is a vector of variables drawn from the information set I_t. We

300

can use the necessary conditions to form the expressions

$$\left[1 - \beta\left(1 + r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{\gamma - 1}\right] \mathbf{x}_t \equiv m_t(\theta)$$

- 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.

- In principle, we could use a very large number of moment conditions in estimation, since *any current or lagged variable* could be used in $\mathbf{x}_t$.

- Note that at time t, m_{t-s} has been observed, and is therefore an element of the information set. By rat. exp., 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:

$$\Omega = E\left[m(\theta^0)m(\theta^0)'\right]$$

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.

- This whole approach relies on the very strong assumption that equation 28 holds without error. Supposing agents were heterogeneous, this wouldn't be reasonable. If there were an error term here, it could potentially be autocorrelated, which would no longer allow any variable in the information set to be used as an instrument..

- Supposing that the representative agent approach is ok, one might think that a large number of instruments should be used to increase the number of moment conditions. This is in fact not the case, as has been seen in Monte Carlo studies (Tauchen, *JBES,* 1986). The reason for poor performance when using many instruments is that the estimate of Ω becomes very imprecise.

18.12   Problems

1. Perform GMM estimation of the rational expectations model described above using the data in the file **gmmdata,** located , on the volcano server. The columns of this data file are c, p, and d, in that order. There are 95 observations (source: Tauchen, *JBES*, 1986). Use as instruments lags of c and $1 + r$.

    - Use lags of orders 1, 2, 3 and 4.

    - Iterate the estimation of $\theta = \beta, \gamma$ and Ω to convergence.

    - 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?

19 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 $\mathbf{x}$, the suppose the joint density of $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \ldots & \mathbf{y}_n \end{pmatrix}$ conditional on $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{pmatrix}$ 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: e.g., 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 $\mathbf{Y}_{t-1} = \begin{pmatrix} \mathbf{y}_1 & \ldots & \mathbf{y}_{t-1} \end{pmatrix}$, $\mathbf{Y}_0 = 0$, and let $\mathbf{X}_t = \begin{pmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_t \end{pmatrix}$ 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

$$s_n(\theta) \overset{a.s.}{\to} \lim_{n \to \infty} \mathcal{E} \frac{1}{n} \sum_{t=1}^{n} \ln f_t(\theta) \equiv \bar{s}(\theta)$$

We assume that this can be strengthened to uniform convergence, a.s., following the

304

previous arguments. The "pseudo-true" value of θ is the value that maximizes $\bar{s}(\theta)$:

$$\theta^0 = \arg\max_{\Theta} \bar{s}(\theta)$$

Given assumptions so that theorem 55 is applicable, we obtain

$$\lim_{n\to\infty} \hat{\theta}_n = \theta^0, \text{a.s.}$$

An example of sufficient conditions for consistency are

- Θ is compact

  - $s_n(\theta)$ is continuous and converges pointwise almost surely to $\bar{s}(\theta)$ (this means that $\bar{s}(\theta)$ will be continuous, and this combined with compactness of Θ means $\bar{s}(\theta)$ is uniformly continuous).

  - θ^0 is a unique global maximizer. A stronger version of this assumption that allows for asymptotic normality is that $D^2_\theta \bar{s}(\theta)$ exists and is negative definite in a neighborhood of θ^0.

- Applying the asymptotic normality theorem,

  $$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1} I_\infty(\theta^0) \mathcal{J}_\infty(\theta^0)^{-1}\right]$$

  where

  $$\mathcal{J}_\infty(\theta^0) = \lim_{n\to\infty} \mathcal{E} D^2_\theta s_n(\theta^0)$$

  and

  $$I_\infty(\theta^0) = \lim_{n\to\infty} Var\sqrt{n} D_\theta s_n(\theta^0).$$

- Note that asymptotic normality only requires that the additional assumptions regarding $\mathcal{J}$ and I hold in a neighborhood of θ^0 for $\mathcal{J}$ and at θ^0, for I, not throughout Θ. In this sense, asymptotic normality is a local property.

19.0.1 Consistent Estimation of Variance Components

Consistent estimation of $\mathcal{J}_\infty(\theta^0)$ is straightforward. Assumption (b) of Theorem 58 implies that

$$\mathcal{J}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f_t(\hat{\theta}_n) \overset{a.s.}{\rightarrow} \lim_{n\to\infty} \mathcal{E}\frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f_t(\theta^0) = \mathcal{J}_\infty(\theta^0).$$

That is, just calculate the Hessian using the estimate $\hat{\theta}_n$ in place of θ^0.

Consistent estimation of $I_\infty(\theta^0)$ is more difficult, and may be impossible.

- **Notation**: Let $g_t \equiv D_\theta f_t(\theta^0)$

We need to estimate

$$
\begin{aligned}
I_\infty(\theta^0) &= \lim_{n\to\infty} Var\sqrt{n}D_\theta s_n(\theta^0) \\
&= \lim_{n\to\infty} Var\sqrt{n}\frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f_t(\theta^0) \\
&= \lim_{n\to\infty} \frac{1}{n}Var\sum_{t=1}^{n} g_t \\
&= \lim_{n\to\infty} \frac{1}{n}\mathcal{E}\left\{\left(\sum_{t=1}^{n}(g_t - \mathcal{E}g_t)\right)\left(\sum_{t=1}^{n}(g_t - \mathcal{E}g_t)\right)'\right\}
\end{aligned}
$$

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 $I_\infty(\theta^0)$ *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: we have that 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

$$\bar{s}(\theta) = \mathcal{E}_X \, \mathcal{E}_0 \ln f(y|x, \theta^0)$$

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

$$D_\theta \mathcal{E}_X \, \mathcal{E}_0 \ln f(y|x, \theta^0) = D_\theta \, \bar{s}(\theta) = 0$$

- 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

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n D_\theta \ln f(y|x, \theta^0) \xrightarrow{d} N(0, I_\infty(\theta^0)).$$

307

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

$$\widehat{I} = \frac{1}{n} \sum_{t=1}^{n} D_\theta \ln f_t(\hat{\theta}) D_{\theta'} \ln f_t(\hat{\theta})$$

This is an important case where consistent estimation of the covariance matrix is possible. Other cases exist, even for dynamically misspecified time series models.

20 Nonlinear least squares (NLS)

Readings: Davidson and MacKinnon, Ch. 2* and 5*; Gallant, Ch. 1

20.1 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

$$\mathbf{F}(\hat{\theta}) \equiv D_{\theta'}\mathbf{f}(\hat{\theta}). \tag{29}$$

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

$$\hat{\mathbf{F}}'\left[\mathbf{y} - \mathbf{f}(\hat{\theta})\right] \equiv 0. \tag{30}$$

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 maxima, minima and saddlepoints: the objective function $s_n(\theta)$ is not necessarily well-behaved and may be difficult to minimize.

20.2 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 $D_\theta^2 s_\infty(\theta^0)$ be positive definite. Consider the objective function:

$$
\begin{aligned}
s_n(\theta) &= \frac{1}{n} \sum_{t=1}^{n} \left[y_t - f(\mathbf{x}_t, \theta) \right]^2 \\
&= \frac{1}{n} \sum_{t=1}^{n} \left[f(\mathbf{x}_t, \theta^0) + \varepsilon_t - f_t(\mathbf{x}_t, \theta) \right]^2 \\
&= \frac{1}{n} \sum_{t=1}^{n} \left[f_t(\theta^0) - f_t(\theta) \right]^2 + \frac{1}{n} \sum_{t=1}^{n} (\varepsilon_t)^2 \\
&\quad - \frac{2}{n} \sum_{t=1}^{n} \left[f_t(\theta^0) - f_t(\theta) \right] \varepsilon_t
\end{aligned}
$$

- As in example 16.3, 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.

311

- 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

$$\frac{1}{n} \sum_{t=1}^{n} \left[f_t(\theta^0) - f_t(\theta) \right]^2 \overset{a.s.}{\rightarrow} \int \left[f(z, \theta^0) - f(z, \theta) \right]^2 d\mu(z), \tag{31}$$

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 $f(x, \theta) = [1 + \exp(-x\theta)]^{-1}$, $f : \mathfrak{R}^K \rightarrow (0, 1)$, 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)

$$\frac{\partial^2}{\partial\theta\partial\theta'} \int \left[f(x, \theta^0) - f(x, \theta) \right]^2 d\mu(x) \bigg|_{\theta^0} = 2 \int \left[D_\theta f(z, \theta^0)' \right] \left[D_{\theta'} f(z, \theta^0) \right]' d\mu(z)$$

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

312

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

$$\mathcal{J}_\infty(\theta^0) = 2\lim \mathcal{E}\frac{\mathbf{F}'\mathbf{F}}{n}$$

20.3 Consistency

We simply assume that the conditions of Theorem 55 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..

20.4 Asymptotic normality

As in the case of GMM, we also simply assume that the conditions for asymptotic normality as in Theorem 58 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

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1} I_\infty(\theta^0) \mathcal{J}_\infty(\theta^0)^{-1}\right],$$

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

$$n\left[D_\theta s_n(\theta^0)\right]\left[D_\theta s_n(\theta^0)\right]' \xrightarrow{a.s.} I_\infty(\theta^0),$$

313

The objective function is

$$s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} [y_t - f(\mathbf{x}_t, \theta)]^2$$

So

$$D_\theta s_n(\theta) = -\frac{2}{n} \sum_{t=1}^{n} [y_t - f(\mathbf{x}_t, \theta)] D_\theta f(\mathbf{x}_t, \theta).$$

Evaluating at θ^0,

$$D_\theta s_n(\theta^0) = -\frac{2}{n} \sum_{t=1}^{n} \varepsilon_t D_\theta f(\mathbf{x}_t, \theta^0).$$

With this we obtain

$$n \left[D_\theta s_n(\theta^0) \right] \left[D_\theta s_n(\theta^0) \right]' = \frac{4}{n} \left[\sum_{t=1}^{n} \varepsilon_t D_\theta f(\mathbf{x}_t, \theta^0) \right] \left[\sum_{t=1}^{n} \varepsilon_t D_\theta f(\mathbf{x}_t, \theta^0) \right]'$$

Noting that

$$\sum_{t=1}^{n} \varepsilon_t D_\theta f(\mathbf{x}_t, \theta^0) = \frac{\partial}{\partial \theta} \left[\mathbf{f}(\theta^0) \right]' \varepsilon$$
$$= \mathbf{F}' \varepsilon$$

we can write the above as

$$n \left[D_\theta s_n(\theta^0) \right] \left[D_\theta s_n(\theta^0) \right]' = \frac{4}{n} \mathbf{F}' \varepsilon \varepsilon' \mathbf{F}$$

This converges almost surely to its expectation, following a LLN

$$I_\infty(\theta^0) = 4\sigma^2 \lim \mathcal{E} \frac{\mathbf{F}' \mathbf{F}}{n}$$

314

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 $I_\infty(\theta^0)$, 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

$$\left(\frac{\hat{\mathbf{F}}'\hat{\mathbf{F}}}{n}\right)^{-1}\hat{\sigma}^2, \tag{32}$$

where $\hat{\mathbf{F}}$ is defined as in equation 29 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.

20.5   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

$$\lambda_t^0 = \exp(\mathbf{x}_t'\boldsymbol{\beta}^0),$$

which enforces the positivity of λ_t. Suppose we estimate $\boldsymbol{\beta}^0$ by nonlinear least squares:

$$\hat{\boldsymbol{\beta}} = \arg\min s_n(\boldsymbol{\beta}) = \frac{1}{T}\sum_{t=1}^{n}\left(y_t - \exp(\mathbf{x}_t'\boldsymbol{\beta})\right)^2$$

We can write

$$
\begin{aligned}
s_n(\boldsymbol{\beta}) &= \frac{1}{T}\sum_{t=1}^{n}\left(\exp(\mathbf{x}_t'\boldsymbol{\beta}^0) + \varepsilon_t - \exp(\mathbf{x}_t'\boldsymbol{\beta})\right)^2 \\
&= \frac{1}{T}\sum_{t=1}^{n}\left(\exp(\mathbf{x}_t'\boldsymbol{\beta}^0) - \exp(\mathbf{x}_t'\boldsymbol{\beta})\right)^2 + \frac{1}{T}\sum_{t=1}^{n}\varepsilon_t^2 + 2\frac{1}{T}\sum_{t=1}^{n}\varepsilon_t\left(\exp(\mathbf{x}_t'\boldsymbol{\beta}^0) - \exp(\mathbf{x}_t'\boldsymbol{\beta})\right)
\end{aligned}
$$

The last term has expectation zero since the assumption that $\mathcal{E}(y_t|\mathbf{x}_t) = \exp(\mathbf{x}_t'\boldsymbol{\beta}^0)$ implies that $\mathcal{E}(\varepsilon_t|\mathbf{x}_t) = 0$, which in turn implies that functions of $\mathbf{x}_t$ are uncorrelated with ε_t. Applying a strong LLN, and noting that the obsective function is continuous on a compact parameter space, we get

$$s_\infty(\boldsymbol{\beta}) = \mathcal{E}_{\mathbf{x}}\left(\exp(\mathbf{x}'\boldsymbol{\beta}^0) - \exp(\mathbf{x}'\boldsymbol{\beta})\right)^2 + \mathcal{E}_{\mathbf{x}}\exp(\mathbf{x}'\boldsymbol{\beta}^0)$$

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 $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, so the NLS estimator is consistent as long as identification holds.

Exercise 63 *Determine the limiting distribution of $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)$. This means finding the the specific forms of $\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}s_n(\boldsymbol{\beta})$, $\mathcal{J}(\boldsymbol{\beta}^0)$, $\frac{\partial s_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\Big|$, and $I(\boldsymbol{\beta}^0)$. Again, use a CLT as needed, no need to verify that it can be applied.*

20.6   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

$$\mathbf{y} = \mathbf{f}(\theta) + \mathbf{v}$$

where $\mathbf{v}$ 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 :

$$\mathbf{y} = \mathbf{f}(\theta^1) + \left[D_{\theta'}\mathbf{f}\left(\theta^1\right)\right]\left(\theta - \theta^1\right) + \mathbf{v} + \text{approximationerror}.$$

This 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 $\mathbf{v}$ plus approximation error from the truncated Taylor's series.

- Note that $\mathbf{F}$ is known, given θ^1.

- Similarly, $\mathbf{z} \equiv \mathbf{y} - \mathbf{f}(\theta^1)$, which is also known.

- The other new element here is $b \equiv (\theta - \theta^1)$. 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

$$\hat{\mathbf{F}}'\left[\mathbf{y} - \mathbf{f}(\hat{\theta})\right] \equiv 0$$

by definition of the NLS estimator (these are the normal equations as in equation 30, 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 32 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

318

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, $\mathbf{F'F}$ will be nearly singular, so $(\mathbf{F'F})^{-1}$ will be subject to large roundoff errors.

20.7 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.

20.7.1 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

$$
s^* = \mathbf{x}'\beta + \text{residual}
$$

320

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

$$\mathcal{E}\left[\omega\right| -\varepsilon < \mathbf{r}'\theta] \neq 0,$$

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 $\mathcal{E}\left[\omega\right| -\varepsilon < \mathbf{r}'\theta]$. 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

$$s = \mathbf{x}'\beta + \rho\sigma\mathcal{E}(\varepsilon| -\varepsilon < \mathbf{r}'\theta) + \eta.$$

- 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

$$
\begin{aligned}
s &= \mathbf{x}'\beta + \rho\sigma\frac{\phi(\mathbf{r}'\theta)}{\Phi(\mathbf{r}'\theta)} + \eta & (33)\\[2ex]
&\equiv \begin{bmatrix} \mathbf{x}' & \frac{\phi(\mathbf{r}'\theta)}{\Phi(\mathbf{r}'\theta)} \end{bmatrix} \begin{bmatrix} \beta \\ \zeta \end{bmatrix} + \eta. & (34)
\end{aligned}
$$

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 34 is estimated by least squares using the estimated value of θ to form the regressors. This is inefficient and estimation of the co-variance 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.

21 Examples: demand for health care

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.

21.1 The MEPS data

The file health.mat (on the class web page) contains 500 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). The conditioning variables are private insurance (PRIV), public insurance (PUBLIC), age (AGE), sex (SEX), income (INCOME) and years of education (EDUC). 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.

Here are descriptive statistics for the measures of usage:

	mean	variance	mean/var	max	% zeros
OBDV	3.4120	37.446	0.091117	68.000	0.32000
OPV	0.20400	1.0944	0.18641	20.000	0.88800
ERV	0.18400	0.30614	0.60102	6.0000	0.86400
IPV	0.076000	0.14222	0.53437	5.0000	0.94600
DV	1.0360	3.1107	0.33304	16.000	0.55800
PRESCR	8.0500	214.39	0.037549	107.00	0.29000

Since health care visits are count data, a simple approach to modeling demand could be based upon the Poisson model. Recall that the Poisson model is

$$
\begin{aligned}
f_Y(y) &= \frac{exp(-\lambda)\lambda^y}{y!} \\
\lambda &= exp(\mathbf{x}\prime\beta)
\end{aligned}
$$

Here, we'll let the **x** vector be

$$
\mathbf{x} = \begin{bmatrix} 1 & PUBLIC & PRIV & SEX & AGE & EDUC & INC \end{bmatrix}\prime
$$

Recall that the Poisson model imposes that the conditional mean equals the conditional variance (*equidispersion*). We see from the above descriptive statistics that the data are all unconditionally overdispersed, since the unconditional variance is greater than the unconditional mean. To achieve conditional equidispersion, the model would have to fit quite well.

Here are results for OBDV:

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

MEPS data, OBDV

poisson results

Strong convergence

Observations = 500

Function value      -3.8679

                params      t(OPG)      t(Sand.)      t(Hess)

constant      -0.51541      -8.5242      -1.0992      -3.2325

pub_ins        0.61054      16.999        3.0582       7.6966

priv_ins       0.18459       5.1354       1.1697       2.4819

sex            0.35452      21.396        2.1007       7.0053

age            0.022112     24.396        4.3966      10.795

educ           0.027979      8.6896       0.93269      2.9554

inc            0.0070852     2.2891       0.30328      0.87485

Information Criteria

Consistent Akaike

     3918.4

Schwartz

     3911.4

Hannan-Quinn

     3893.5

Akaike

     3881.9

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

- The insurance variables have the expected sign, but PRIV is not significant.

Women and older people make more visits. Income appears not to affect demand for office based visits.

- Note that the t-stats differ quite a bit according to the covariance matrix estimator. This big difference is an indicator of possible misspecification. If there is misspecification, then only the sandwich form is valid (since we have a QML estimator in this case, and the information matrix equality doesn't hold). The information matrix test is based on this principle.

Here are results for ERV.

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

MEPS data, ERV

poisson results

Strong convergence

Observations = 500

Function value      -0.49978

                  params      t(OPG)     t(Sand.)      t(Hess)

constant         -1.1669     -2.0607     -1.6099      -1.8912

pub_ins          0.65307      2.3722      1.7257       2.3114

priv_ins        -0.26764    -0.93634    -0.83555     -0.90040

sex             -0.57001     -2.7777     -2.0050      -2.6389

age            0.0037963     0.60114     0.32714      0.45393

educ           0.0010258    0.026424    0.024977     0.026173

inc             -0.12531     -2.2085     -2.2781      -2.3102


Information Criteria

Consistent Akaike

     550.29

Schwartz

     543.29

Hannan-Quinn

     525.36

Akaike

     513.78

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

Table 1: Marginal Variances, Sample and Estimated (Poisson)

	OBDV	ERV
Sample	37.446	0.30614
Estimated	3.4540	0.19060

- In this case, private insurance has a negative impact.

- Women are less likely to make emergency room visits compared to men.

- Richer people make fewer visits, and the effect seems to be significant. Perhaps poor people do not have good insurance coverage and use emergency visits as a substitute for preventive care?

- There is less difference between the three forms of the t-statistics. Is this an indication that the Poisson model might work better for ERV than for OBDV?

To check the plausibility of the Poisson model, we can compare the sample unconditional variance with the estimated unconditional variance according to the Poisson model: $\widehat{V(y)} = \frac{\sum_{t=1}^{n} \hat{\lambda}_t}{n}$. For OBDV and ERV, we get We see that even after conditioning, the overdispersion is not captured in either case. There is huge problem with OBDV, and a significant problem with ERV. In both cases the Poisson model does not appear to be plausible.

21.2   Infinite mixture models

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 possibil-

ity that the constant term in a Poisson model were random:

$$f_Y(y, \varepsilon | \mathbf{x}) = \frac{\exp(-\lambda)\lambda^y}{y!}$$

$$\lambda = \exp(\mathbf{x'}\beta + \varepsilon)$$

$$= \exp(\mathbf{x'}\beta)\exp(\varepsilon)$$

$$= \theta\nu$$

where $\theta = \exp(\mathbf{x'}\beta)$ 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

$$f_Y(y|\mathbf{x}) = \int_{-\infty}^{\infty} \frac{\exp[-\lambda]\lambda^y}{y!} f_\lambda(z) dz$$

This density *can* be used directly, perhaps using numberical 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

$$f_Y(y|\phi) = \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^\psi \left(\frac{\lambda}{\psi+\lambda}\right)^y \tag{35}$$

where $\phi = (\lambda, \psi)$. ψ appear since it is the parameter of the gamma density.

- For this density, $E(y|x) = \lambda$. We again parameterize $\lambda = \exp(x'\beta)$

- The variance depends upon how ψ is parameterized.

  - If $\psi = \lambda/\alpha$, where $\alpha > 0$, then $V(y|x) = \lambda + \alpha\lambda$. Note that λ is a function of x, so that the variance is too. This is referred to as the NB-I model.

  - If $\psi = 1/\alpha$, where $\alpha > 0$, then $V(y|x) = \lambda + \alpha\lambda^2$. This is referred to as the NB-II model.

329

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.

Here are NB-I estimation results for OBDV

```
MEPS data, OBDV

negbin results

Strong convergence

Observations = 500

Function value      -2.2656

t-Stats
```

	params	t(OPG)	t(Sand.)	t(Hess)
constant	-0.055766	-0.16793	-0.17418	-0.17215
pub_ins	0.47936	2.9406	2.8296	2.9122
priv_ins	0.20673	1.3847	1.4201	1.4086
sex	0.34916	3.2466	3.4148	3.3434
age	0.015116	3.3569	3.8055	3.5974
educ	0.014637	0.78661	0.67910	0.73757
inc	0.012581	0.60022	0.93782	0.76330
ln_alpha	1.7389	23.669	11.295	16.660

```
Information Criteria

Consistent Akaike

     2323.3

Schwartz

     2315.3

Hannan-Quinn

     2294.8

Akaike

     2281.6
```

Here are NB-II results for OBDV

****

MEPS data, OBDV

negbin results

Strong convergence

Observations = 500

Function value   -2.2616

t-Stats

	params	t(OPG)	t(Sand.)	t(Hess)
constant	-0.65981	-1.8913	-1.4717	-1.6977
pub_ins	0.68928	2.9991	3.1825	3.1436
priv_ins	0.22171	1.1515	1.2057	1.1917
sex	0.44610	3.8752	2.9768	3.5164
age	0.024221	3.8193	4.5236	4.3239
educ	0.020608	0.94844	0.74627	0.86004
inc	0.020040	0.87374	0.72569	0.86579
ln_alpha	0.47421	5.6622	4.6278	5.6281

Information Criteria

Consistent Akaike

    2319.3

Schwartz

    2311.3

Hannan-Quinn

    2290.8

Akaike

    2277.6

Table 2: Marginal Variances, Sample and Estimated (NB-II)

	OBDV	ERV
Sample	37.446	0.30614
Estimated	26.962	0.27620

****

- For the OBDV model, the NB-II model does a better job, in terms of the average log-likelihood and the information criteria.

- Note that both versions of the NB model fit much better than does the Poisson model.

- The t-statistics are now similar for all three ways of calculating them, which might indicate that the serious specification problems of the Poisson model for the OBDV data are partially solved by moving to the NB model.

- The estimated $\ln \alpha$ 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: $\widehat{V(y)} = \frac{\sum_{t=1}^{n} \hat{\lambda}_t + \hat{\alpha}(\hat{\lambda}_t)^2}{n}$. For OBDV and ERV (estimation results not reported), we get The overdispersion problem is significantly better than in the Poisson case, but there is still some overdispersion that is not captured, for both OBDV and ERV.

21.3 Hurdle models

Returning to the Poisson model, lets look at actual and fitted count probabilities. Actual 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 see that for the OBDV measure, there are many more actual ze-

Table 3: 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

ros 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:

$$\Pr(Y = 0) \quad = f_Y(0, \lambda_p) = \quad 1 - 1/\left[1 + \exp(-\lambda_p)\right]$$

$$\Pr(Y > 0) \quad\quad\quad = \quad\quad 1/\left[1 + \exp(-\lambda_p)\right],$$

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

$$f_Y(y, \lambda_d | y > 0) \;=\; \frac{f_Y(y, \lambda_d)}{1 - \exp(-\lambda_d)}$$

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

$$\begin{aligned}
E(Y|x) \;&=\; \Pr(Y > 0|x)E(Y|Y > 0, x) \\
&=\; \frac{1}{1 + \exp(-\lambda_p)} \frac{\lambda_d}{1 - \exp(-\lambda_d)}
\end{aligned}$$

Here are hurdle Poisson estimation results for OBDV:

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

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

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

Table 4: 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

Fitted and actual probabilites (NB-II fits are provided as well) are:

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.

21.4 Finite mixture models

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 ith 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 are results for a mixture of 2 negative binomial (NB-I) models, for the

339

OBDV data.

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

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 suggetive.

- 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.

21.5   Comparing models using information criteria

A Poisson model can't be tested (using standard methods) as a restriction of a negative binomial model. Testing for collapse of a finite mixture to a mixture of fewer components has the same problem. How can we determine which 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

$$
\begin{aligned}
CAIC &= -2\ln L(\hat{\theta}) + k(\ln n + 1) \\
BIC &= -2\ln L(\hat{\theta}) + k\ln n \\
AIC &= -2\ln L(\hat{\theta}) + 2k
\end{aligned}
$$

344

Table 5: Information Criteria, OBDV

Model	AIC	BIC	CAIC
Poisson	3822	3911	3918
NB-I	2282	2315	2323
Hurdle Poisson	3333	3381	3395
MNB-I	2265	2337	2354
CMNB-I	2266	2312	2323

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. According to the AIC, the best is the MNB-I, which has relatively many parameters. The best according to the BIC is CMNB-I, and according to CAIC, the best is NB-I. The Poisson-based models do not do well.

22 Nonparametric inference

22.1 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. Suppose that

$$f(x) = 1 + \frac{3x}{2\pi} - \left(\frac{x}{2\pi}\right)^2$$

345

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 know 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 16 and 31 is

$$s_\infty(a,b) = \int_0^{2\pi} (f(x) - h(x))^2 \, dx.$$

The theorem regarding the consistency of extremum estimators (Theorem 55) tells us that $\hat{a}$ and $\hat{b}$ will converge almost surely to the values that minimize the limiting objective function. Solving the first order conditions[2] reveals that $s_\infty(a,b)$ obtains its

---

[2] All calculations were done using Scientific Workplace.

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

$$h_\infty(x) = 7/6 + x/\pi$$

We may plot 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 allow consistent estimation. The mathematical properties of the Taylor's series do not carry over when coefficients are estimated.

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:

$$\varepsilon(x) = x\phi'(x)/\phi(x)$$

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

$$\eta(x) = xh'(x)/h(x)$$

Plotting 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 approx-

imation 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 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

$$\left\{ a_1 = \frac{7}{6}, a_2 = \frac{1}{\pi}, a_3 = -\frac{1}{\pi^2}, a_4 = 0, a_5 = -\frac{1}{4\pi^2}, a_6 = 0 \right\}.$$

Substituting these values into $g_K(x)$ we obtain the almost sure limit of the approximation

$$g_\infty(x) = 7/6 + x/\pi + (\cos x)\left(-\frac{1}{\pi^2}\right) + (\sin x)0 + (\cos 2x)\left(-\frac{1}{4\pi^2}\right) + (\sin 2x)0 \quad (36)$$

Plotting the approximation and the true function:

Clearly the truncated trigonometric series model offers a better approximation, asymptotically, than does the linear model. Plotting elasticities: On average, the fit is better, though there is some implausible wavyness in the estimate.

Root mean squared error in the 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.

22.2 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 semidefinite. 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

$$x(p,m) = x(p,m,\theta^0) + \varepsilon, \theta^0 \in \Theta^0,$$

349

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. 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, so rejection of the hypothesis calls into question the theory.

- Nonparametric inference allows direct testing of economic propositions, without the "model-induced augmenting hypothesis".

22.3   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

$$g_K(\mathbf{x} \mid \theta_K) = \alpha + \mathbf{x}'\beta + 1/2\mathbf{x}'\mathbf{C}\mathbf{x} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left(u_{j\alpha} \cos(j\mathbf{k}_\alpha'\mathbf{x}) - v_{j\alpha} \sin(j\mathbf{k}_\alpha'\mathbf{x}) \right). \quad (37)$$

where the K-dimensional parameter vector

$$\theta_K = \{\alpha, \beta', vec^*(C)', u_{11}, v_{11}, \ldots, u_{JA}, v_{JA}\}'. \quad (38)$$

- 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

351

variables, and multiply by $2\pi - eps$, where eps is some positive number less than 2π in value.

- The k_α are "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

$$D_x g_K(\mathbf{x} \mid \theta_K) = \beta + \mathbf{C}\mathbf{x} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left[\left(-u_{j\alpha} \sin(j\mathbf{k}'_\alpha \mathbf{x}) - v_{j\alpha} \cos(j\mathbf{k}'_\alpha \mathbf{x}) \right) j\mathbf{k}_\alpha \right] \qquad (39)$$

and the matrix of second partial derivatives is

$$D_x^2 g_K(\mathbf{x}|\theta_K) = \mathbf{C} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left[\left(-u_{j\alpha} \cos(j\mathbf{k}'_\alpha \mathbf{x}) + v_{j\alpha} \sin(j\mathbf{k}'_\alpha \mathbf{x}) \right) j^2 \mathbf{k}_\alpha \mathbf{k}'_\alpha \right] \qquad (40)$$

To define a compact notation for partial derivatives, let λ be an N-dimensional multi-index with no negative elements. Define $|\lambda|^*$ 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

$$D^\lambda g_K(\mathbf{x}|\theta_K) = \mathbf{z}^\lambda(\mathbf{x})' \theta_K. \qquad (41)$$

- 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 64 *[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 $\| h \|$. Consider the following conditions:*

(a) Compactness: The closure of $\mathcal{H}$ with respect to $\| h \|$ is compact in the relative topology defined by $\| h \|$.

(b) Denseness: $\cup_K \mathcal{H}_K$, $K = 1, 2, 3, \ldots$ 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

$$\lim_{n \to \infty} \sup_{\overline{\mathcal{H}}} | s_n(h) - s_\infty(h, h^*) | = 0$$

almost surely.

(d) Identification: Any point h in the closure of $\mathcal{H}$ with $\bar{s}(h, \hbar) \geq s_\infty(h^*, h^*)$ 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 55. 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

354

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 [55], see Gallant, 1987) but we will discuss its assumptions, in relation to the Fourier form as the approximating model.

22.3.1   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 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:

$$\| h \|_{m,X} = \max_{|\lambda^*| \leq m} \sup_{X} \left| D^{\lambda} h(x) \right|$$

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

$$\| f(\mathbf{x}) - g_K(\mathbf{x} \mid \theta_K) \|_{m,X} \, .$$

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

355

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.

22.3.2   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.

22.3.3   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 65 *[Estimation space] The estimation space $\mathcal{H} = \mathcal{W}_{2,\mathcal{X}}(D)$. The estimation space is an open set, and we presume that $h^* \in \mathcal{H}$.*

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 66 *[Estimation subspace] The estimation subspace $\mathcal{H}_K$ is defined as*

$$\mathcal{H}_K = \{g_K(\mathbf{x}|\theta_K) : g_K(\mathbf{x}|\theta_K) \in \mathcal{W}_{2,Z}(D), \theta_K \in \Re^K\},$$

356

where $g_K(\mathbf{x}, \theta_K)$ is the Fourier form approximation as defined in Equation 37.

22.3.4 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 **??**). 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 37 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 67 *[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 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 $\varepsilon > 0$, $\| h^*(\mathbf{x}) - h_K(\mathbf{x}|\theta_K) \|_{q,X} = o(K^{-m+q+\varepsilon})$ 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 X, which is open and contains the closure of 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 67 is that there is a sequence of $\{h_K\}$ from $\cup_\infty \mathcal{H}_K$ such that

$$\lim_{K \to \infty} \| h^* - h_K \|_{1,X} = 0,$$

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 $\| h \|_{1,X}$.

22.3.5 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 16 and 31, the limiting objective function is

$$s_\infty(g, f) = -\int_X (f(\mathbf{x}) - g(\mathbf{x}))^2 \, d\mu x. \tag{42}$$

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 strong stochastic equicontinuity applies, so that we have uniform almost sure convergence. We also have continuity of the objective function in g, with respect to the norm $\| h \|_{1,X}$ since

$$\lim_{\|g^1 - g^0\|_{1,X} \to 0} \left\{ s_\infty\left(g^1, f\right) - s_\infty\left(g^0, f\right) \right\}$$
$$= \lim_{\|g^1 - g^0\|_{1,X} \to 0} \int_X \left[\left(g^1(\mathbf{x}) - f(\mathbf{x})\right)^2 - \left(g^0(\mathbf{x}) - f(\mathbf{x})\right)^2 \right] d\mu x.$$

By the dominated convergence theorem (which applies since the finite bound D used to define $\mathcal{W}_{2,Z}(D)$ is dominated by an integrable function), the limit and the integral can be interchanged, so by inspection, the limit is zero.

22.3.6 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 assumption).

22.3.7 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 $\mathbf{x} \in \mathcal{X}$.

22.3.8 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 (Gauss command pinv($\mathbf{X}$)).

  - 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

$$AV = \lim_{n\to\infty} E\left[\mathbf{z}'\left(\frac{\mathbf{Z}_K'\mathbf{Z}_K}{n}\right)^+ \mathbf{z}\hat{\sigma}^2\right].$$

361

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.

22.4   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

$$
\begin{aligned}
g(x) &= \int y \frac{f(x,y)}{h(x)} dy \\
&= \frac{1}{h(x)} \int y f(x,y) dy,
\end{aligned}
$$

362

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 $\int y f(x,y)dy$.

22.4.1 Estimation of the denominator

A kernel estimator for $h(x)$ has the form

$$\hat{h}(x) = \frac{1}{n}\sum_{t=1}^{n} \frac{K\left[(x-x_t)/\gamma_n\right]}{\gamma_n^k},$$

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

$$\lim_{n\to\infty} \gamma_n = 0$$
$$\lim_{n\to\infty} n\gamma_n^k = \infty$$

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 (since 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 $\left|\frac{dz}{dz^{*\prime}}\right| = \gamma_n^k$, 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

$$
\begin{aligned}
n\gamma_n^k V\left[\hat{h}(x)\right] &= \gamma_n^{-k} E\left\{(K\left[(x-z)/\gamma_n\right])^2\right\} - \gamma_n^{-k}\left\{E\left(K\left[(x-z)/\gamma_n\right]\right)\right\}^2 \\
&= \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right]^2 h(z)dz - \gamma_n^k\left\{\int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right] h(z)dz\right\}^2 \\
&= \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right]^2 h(z)dz - \gamma_n^k E\left[\widehat{h}(x)\right]^2
\end{aligned}
$$

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

$$
\lim_{n\to\infty} n\gamma_n^k V\left[\hat{h}(x)\right] = h(x) \int \left[K(z^*)\right]^2 dz^*.
$$

365

Since both $\int [K(z^*)]^2 dz^*$ 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).

22.4.2 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:

$$\hat{f}(x,y) = \frac{1}{n} \sum_{t=1}^{n} \frac{K_* \left[(y-y_t)/\gamma_n, (x-x_t)/\gamma_n\right]}{\gamma_n^{k+1}}$$

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.

22.4.3 Discussion

- The kernel regression estimator for $g(x_t)$ is a weighted average of the $y_j, j = 1, 1, ..., n$, where higher weights are associated with points that are closer to x_t. The weights sum to 1.

- 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.

22.4.4 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.

22.5 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

$$
\begin{aligned}
\widehat{f}_{y|x} &= \frac{\hat{f}(x,y)}{\hat{h}(x)} \\
&= \frac{\frac{1}{n}\sum_{t=1}^{n}\frac{K_*[(y-y_t)/\gamma_n,(x-x_t)/\gamma_n]}{\gamma_n^{k+1}}}{\frac{1}{n}\sum_{t=1}^{n}\frac{K[(x-x_t)/\gamma_n]}{\gamma_n^{k}}} \\
&= \frac{1}{\gamma_n}\frac{\sum_{t=1}^{n}K_*\left[(y-y_t)/\gamma_n,(x-x_t)/\gamma_n\right]}{\sum_{t=1}^{n}K\left[(x-x_t)/\gamma_n\right]}
\end{aligned}
$$

where we obtain the expressions for the joint and marginal densities from the section on kernel regression.

22.6 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 http://www.econ.duke.edu/~get/snp.html. 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

$$
g_P(y|x,\phi,\theta) = \frac{h_p^2(y|\theta)f(y|x,\phi)}{\eta_p(x,\phi,\theta)}
$$

where

$$
h_p(y|\theta) = \sum_{k=0}^{p}\theta_k y^k
$$

369

and $\eta_p(x, \phi, \theta)$ is a normalizing factor to make the density integrate (sum) to one. Because $h_p^2(y|\theta)/\eta_p(x, \phi, \theta)$ is a homogenous function of θ it is necessary to impose a normalization: θ_0 is set to 1.

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 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, $E(Y) = \lambda$.

To obtain a more flexible density, we may reshape the negative binomial density using a squared polynomial

$$h_p(y|\gamma) = \sum_{k=0}^{p} \gamma_k y^k, \tag{43}$$

The new density, with normalization to sum to one, is

$$f_Y(y|\phi, \gamma) = \frac{[h_p(y|\gamma)]^2}{\eta_p(\phi, \gamma)} \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^{\psi} \left(\frac{\lambda}{\psi+\lambda}\right)^{y}, \tag{44}$$

The normalization factor $\eta_p(\phi, \gamma)$ is calculated (following Cameron and Johansson)

using

$$
\begin{aligned}
E(Y^r) &= \sum_{y=0}^{\infty} y^r f_Y(y|\phi,\gamma) \\
&= \sum_{y=0}^{\infty} y^r \frac{[h_p(y|\gamma)]^2}{\eta_p(\phi,\gamma)} f_Y(y|\phi) \\
&= \sum_{y=0}^{\infty} \sum_{k=0}^{p} \sum_{l=0}^{p} y^r f_Y(y|\phi)\gamma_k\gamma_l y^k y^l / \eta_p(\phi,\gamma) \\
&= \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k\gamma_l \left\{ \sum_{y=0}^{\infty} y^{r+k+l} f_Y(y|\phi) \right\} / \eta_p(\phi,\gamma) \\
&= \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k\gamma_l m_{k+l+r} / \eta_p(\phi,\gamma).
\end{aligned}
$$

By setting $r = 0$ we get that the normalizing factor is

$$
\eta_p(\phi,\gamma) = \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k\gamma_l m_{k+l} \tag{45}
$$

Recall that γ_0 is set to 1 to achieve identification. The $m_r(\lambda,\psi)$ in equation 45 are the negative binomial raw moments, which may be obtained from the moment generating function

$$
M_Y(t) = \psi^{\psi} \left(\lambda - e^t \lambda + \psi \right)^{-\psi}. \tag{46}
$$

To illustrate, here are the first through fourth raw moments of the NB density, calculated using *Mathematica* and then programmed in *Ox*. These are the moments you would need to use a second order polynomial $(p = 2)$.

```
if(k_gam >= 1)
    {
      m[][0] = lambda;
      m[][1] = (lambda .* (lambda + psi + lambda .* psi)) ./ psi;
    }
```

```
if(k_gam >= 2)

   {

       m[][2] = (lambda .* (psi .^ 2 + 3 .* lambda .* psi .* (1 +
psi) + lambda .^          2 .* (2 + 3 .* psi + psi .^ 2))) ./ psi
.^ 2;

       m[][3] = (lambda .* (psi .^ 3 + 7 .* lambda .* psi .^ 2 .*
(1 + psi) +

           6 .* lambda .^ 2 .* psi .* (2 + 3 .* psi + psi .^ 2) +

           lambda .^ 3 .* (6 + 11 .* psi + 6 .* psi .^ 2 + psi .^ 3)))
./ psi .^ 3;

   }
```

After calculating the raw moments, the normalization factor is calculated using equation 45, again with the help of *Mathematica*.

```
if(k_gam == 1)

   {

      norm_factor = 1 + gam[0][] .* (2 .* m[][0] + gam[0][] .* m[][1]);

   }

   else

   if(k_gam == 2)

     {

        norm_factor = 1 + gam[0][] .^ 2 .* m[][1] + 2 .* gam[0][]
.* (m[][0] +          gam[1][] .* m[][2]) +

           gam[1][] .* (2 .* m[][1] + gam[1][] .* m[][3]);

     }
```

For $p = 6$, the analogous formulae are impressively long. This is an example of a model that would be difficult ot formulate without the help of a program like *Mathe-*

372

matica.

It is possible that there is conditional heterogeneity such that the appropriate re-shaping should be more local. This can be accomodated by allowing the θ_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 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. The baseline model is a negative binomial density.

Case 1

Case 2

Case 3

Case 4

374
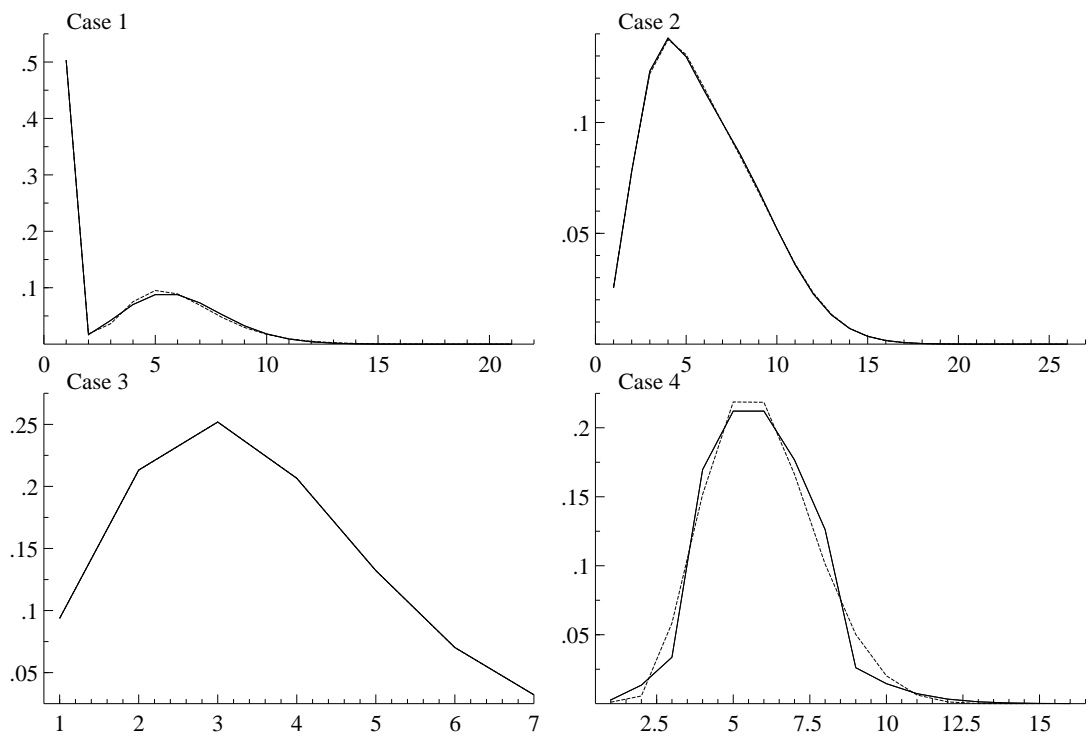
23 Simulation-based estimation

Readings: In addition to the book mentioned previously, articles include Gallant and Tauchen (1996), "Which Moments to Match?", ECONOMETRIC THEORY, Vol. 12, 1996, pages 657-681;ă Gourieroux, Monfort and Renault (1993), "Indirect Inference," *J. Apl. Econometrics;* Pakes and Pollard (1989) *Econometrica*; McFadden (1989) *Econometrica.*

23.1 Motivation

Simulation methods are of interest when the DGP is fully characterized by a parameter vector, but the likelihood function is not calculable. If it were available, we would simply estimate by MLE, which is asymptotically fully efficient.

23.1.1 Example: Multinomial and/or dynamic discrete response models

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

$$\varepsilon_i \sim N(0,\Omega) \tag{47}$$

Henceforth drop the i subscript when it is not needed for clarity.

- y^* is not observed. Rather, we observe a many-to-one mapping

$$y = \tau(y^*)$$

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 ith 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 infeasi-

ble 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.

23.1.2 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:

$$\mathcal{E}(y) = V(y) = \lambda.$$

378

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

$$\Pr(y = y_i) = \int_S \frac{\exp[-\exp(X_i\beta + \eta_i)]\left[\exp(X_i\beta + \eta_i)\right]^{y_i}}{y_i!} d\mu(\eta_i)$$

will have a closed-form solution (one can derive the negative binomial distribution in the way if η has an exponential distribution), but often this will not be possible. In this case, simulation is a means of calculating $\Pr(y = i)$, 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.

23.1.3 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. 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.

380

One can think of Brownian motion the accumulation of independent normally distributed shocks with 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

$$
\begin{aligned}
y_t - y_{t-1} &= g(\phi, y_{t-1}) + h(\phi, y_{t-1})\varepsilon_t \\
\varepsilon_t &\sim N(0, 1)
\end{aligned}
$$

The discretization induces a new parameter, ϕ (that is, the ϕ^0 which defines the best approximation of the discretization to the actual (unknown) discrete time version of the model is not equal to θ^0 which is the true parameter value). This is an approximation, and as such "ML" estimation of ϕ (which is actually quasi-maximum likelihood, QML) based upon this equation is in general biased and inconsistent for the original parameter, θ. Nevertheless, the approximation shouldn't be too bad, which will be useful, as we will see.

- 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.

23.2 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 tth 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

$$\tilde{p}(y_t, X_t, \theta) = \frac{1}{H} \sum_{s=1}^{H} f(\nu_{ts}, y_t, X_t, \theta)$$

is unbiased for $p(y_t|X_t, \theta)$.

- The SML simply substitutes $\tilde{p}(y_t, X_t, \theta)$ in place of $p(y_t|X_t, \theta)$ in the log-likelihood function, that is

$$\hat{\theta}_{SML} = \arg\max s_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln \tilde{p}(y_t, X_t, \theta)$$

23.2.1   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 $\Pr(y_j = 1)$ 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

$$\widetilde{\pi}_{ij} = \frac{\sum_{h=1}^{H} \tilde{y}_{ijh}}{H}$$

- Define $\widetilde{\pi}_i$ as the m-vector formed of the $\widetilde{\pi}_{ij}$. Each element of $\widetilde{\pi}_i$ is between 0 and 1, and the elements sum to one.

- Now $\tilde{p}(y_i, X_i, \theta) = y_i' \frac{1}{H} \sum_{s=1}^{H} \ln \widetilde{\pi}_i(\beta, \Omega)$

- The SML multinomial probit log-likelihood function is

$$\ln \mathcal{L}(\beta, \Omega) = \frac{1}{n} \sum_{i=1}^{n} y_i' \ln \tilde{p}(y_i, X_i, \theta)$$

383

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 $\widetilde{\pi}_i$ are zero or one. In this case, taking the logarithm is going to cause problems.

- 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 $u_{ij} = 1$

if it 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, use the slog function distributed on the web page. Also, increase H if this is a serious problem.

23.2.2 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 68 *[Lee] 1) if* $\lim_{n \to \infty} n^{1/2}/H = 0$, *then*

$$\sqrt{n}\left(\hat{\theta}_{SML} - \theta^0\right) \xrightarrow{d} N(0, I^{-1}(\theta^0))$$

2) if $\lim_{n \to \infty} n^{1/2}/H = \lambda$, λ *a finite constant, then*

$$\sqrt{n}\left(\hat{\theta}_{SML} - \theta^0\right) \xrightarrow{d} N(B, I^{-1}(\theta^0))$$

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.

23.3   Method of simulated moments (MSM)

Suppose we have a $\text{DGP}(y|x,\theta)$ 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

$$\widetilde{k}(x_t, \theta) = \frac{1}{H}\sum_{h=1}^{H} K(\widetilde{y}_t^h, x_t)$$

- By the law of large numbers, $\widetilde{k}(x_t, \theta) \overset{a.s.}{\to} k(x_t, \theta)$, 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

$$\widetilde{m_t}(\theta) = \left[K(y_t, x_t) - \widetilde{k}(x_t, \theta) \right] z_t \tag{48}$$

where z_t is drawn from the information set. As before, form

$$
\begin{aligned}
\widetilde{m}(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \widetilde{m_t}(\theta) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[K(y_t, x_t) - \frac{1}{H} \sum_{h=1}^{H} k(\widetilde{y}_t^h, x_t) \right] z_t \tag{49}
\end{aligned}
$$

with which we form the GMM criterion and estimate as usual. Note that the unbiased simulator $k(\widetilde{y}_t^h, x_t)$ appears linearly within the sums.

23.3.1 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,

$$\sqrt{n} \left(\hat{\theta}_{MSM} - \theta^0 \right) \xrightarrow{d} N \left[0, \left(1 + \frac{1}{H} \right) \left(D_\infty \Omega^{-1} D_\infty' \right)^{-1} \right] \tag{50}$$

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. Simulated GMM can be applied to moment conditions of any form.

23.3.2 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

388

sum over n (see equation [**??**]). 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

$$\tilde{m}(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[K(y_t,x_t)-\frac{1}{H}\sum_{h=1}^{H}k(\tilde{y}_t^h,x_t)\right]z_t \tag{51}$$

$$=\; \frac{1}{n}\sum_{i=1}^{n}\left[k(x_t,\theta^0)+\varepsilon_t-\frac{1}{H}\sum_{h=1}^{H}\left[k(x_t,\theta)+\tilde{\varepsilon}_{ht}\right]\right]z_t \tag{52}$$

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 52 a bit, you will see why the variance inflation factor is $(1+\frac{1}{H})$.

23.4 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).

389

- 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 $D_\lambda \ln f(y_t|x_t, \hat{\lambda})$.

390

- 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

$$m_n(\theta, \hat{\lambda}) = \frac{1}{n} \sum_{t=1}^{n} \int D_\lambda \ln f_t(\hat{\lambda}) p_t(\theta) dy \tag{53}$$

- 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.

23.4.1 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 of 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 58 to conclude that

$$\sqrt{n}\left(\hat{\lambda} - \lambda^0\right) \xrightarrow{d} N\left[0, \mathcal{J}(\lambda^0)^{-1} I(\lambda^0) \mathcal{J}(\lambda^0)^{-1}\right] \tag{54}$$

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 $\mathcal{J}(\lambda^0)^{-1} I(\lambda^0)$ 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 53, we see that

$$\mathcal{J}(\lambda^0) = D_{\lambda'} m(\theta^0, \lambda^0).$$

As in Theorem 58,

$$I(\lambda^0) = \lim_{n \to \infty} \mathcal{E} \left[n \left. \frac{\partial s_n(\lambda)}{\partial \lambda} \right|_{\lambda^0} \left. \frac{\partial s_n(\lambda)}{\partial \lambda'} \right|_{\lambda^0} \right].$$

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 $D_{\lambda'} \tilde{m}_n(\theta^0, \lambda^0) \overset{a.s.}{\to} \mathcal{J}(\lambda^0)$, 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 54

$$\sqrt{n} \mathcal{J}(\lambda^0) \left(\hat{\lambda} - \lambda^0 \right) \overset{a}{\sim} N \left[0, I(\lambda^0) \right]$$

Now, combining the results for the first and second terms,

$$\sqrt{n} \tilde{m}_n(\theta^0, \hat{\lambda}) \overset{a}{\sim} N \left[0, \left(1 + \frac{1}{H} \right) I(\lambda^0) \right]$$

Suppose that $\widehat{I(\lambda^0)}$ 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

393

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 $I(\lambda^0)$ 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 61, we see that defining $\hat{\theta}$ as

$$\hat{\theta} = \arg\min_{\Theta} m_n(\theta, \hat{\lambda})' \left[\left(1 + \frac{1}{H} \right) \widehat{I(\lambda^0)} \right]^{-1} m_n(\theta, \hat{\lambda})$$

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).

23.4.2 Asymptotic distribution

Since we use the optimal weighting matrix, the asymptotic distribution is as in Equation 24, so we have (using the result in Equation 54):

$$\sqrt{n} \left(\hat{\theta} - \theta^0 \right) \xrightarrow{d} N \left[0, \left(D_\infty \left[\left(1 + \frac{1}{H} \right) I(\lambda^0) \right]^{-1} D_\infty' \right)^{-1} \right],$$

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

$$\hat{D} = D_\theta m'_n(\hat{\theta}, \hat{\lambda})$$

23.4.3 Diagnotic testing

The fact that

$$\sqrt{n} m_n(\theta^0, \hat{\lambda}) \overset{a}{\sim} N\left[0, \left(1 + \frac{1}{H}\right) I(\lambda^0)\right]$$

implies that

$$n m_n(\hat{\theta}, \hat{\lambda})' \left[\left(1 + \frac{1}{H}\right) I(\hat{\lambda})\right]^{-1} m_n(\hat{\theta}, \hat{\lambda}) \overset{a}{\sim} \chi^2(q)$$

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:

$$\left(\text{diag}\left[\left(1 + \frac{1}{H}\right) I(\hat{\lambda})\right]^{1/2}\right)^{-1} \sqrt{n} m_n(\hat{\theta}, \hat{\lambda})$$

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.

23.5 Application I: estimation of auction models

References: Laffont, Ossard and Vuong, "The Econometrics of First Price Auctions," *Econometrica*, 1995.

The above estimators open up interesting research possibilities in areas that are relatively undeveloped empirically. An example is models of auctions, which are well developed theoretically but much less so empirically. To see whether a theoretical model is compatible with observed behavior, one needs an econometric model sufficiently rich so that it can embed the complicated interactions between values, strategic behavior and attitudes toward risk. The following illustrates how a Sealed Bid First Price auction could be modeled econometrically.

Assumptions:

- B bidders (known before bidding).

- r_o: reservation price. If the highest bid is below r_0 the good is not sold.

- q : vector of characteristics of the auctioned good

- Bidders seal their bids, and envelopes are opened after all bids collected.

- Each bidder has private valuation $v_i(q, \alpha^0)$, $i = 1, 2, ..., B$.

- Bidders know their own valuation and the distribution of valuations in the population, $f(v|q, \beta^0)$.

- The bidders at time t are assumed to be drawn randomly from the population of bidders.

- Bidders do not know other bidders' valuations.

- Let $\theta^0 = (\alpha^{0\prime}, \beta^{0\prime})' \in \Theta$.

- Bidders are risk neutral, and form their bids under the assumption that all bidders play a symmetric Bayesian Nash strategy.

The problem for the econometrician is to estimate θ^0, which allows prediction of the distribution of bids and of the selling price, as a function of q and B.

- Under the above assumptions, the winning bid (for 2 or more bidders, and assuming the item is sold) is

$$y = E\left\{\max\left[v_{(B-1:B)}, r_0\right] | v_{(B:B)}\right\}$$

where $v_{(1:B)} \leq v_{(2:B)} \leq \cdots \leq v_{(B-1:B)} \leq v_{(B:B)}$ are the order statistics of $v_1, ..., v_B$, which are B random draws from $f(v|q, \theta^0)$.

  - Intuitively, a bidder will bid the value of the order statistic that is less than his/her private value, since this bid is the lowest bid that can be expected to win, conditional on the winning bid being below the private valuation.

- Let $p(y|r_0, B, q, \theta)$ be the density of the winning bid. This density is ordinary except at $y = 0$ and $y = r_0$, where there are concentrations of probability (atoms).

- In general,
$$p(y|r_0, B, q, \theta) \equiv p(y|x, \theta)$$

is not calculable.

- However, $p(y|x, \theta)$ is easily simulable, given θ.

- Indirect inference would supply some tractable pseudo-density $f(y|x, \lambda)$ as the score generator in place of $p(y|x, \theta)$, and would form moment conditions as above.

- The data necessary to estimate this model are simply the characteristics of the good, the reservation price, and the winning bid. A more efficient (and complicated) model would use all of the bid information, were it available.

- A potential application of this sort of model would be the supply of generation of electrical power: generating companies in Norway and the UK bid daily for the price at which power is supplied to the electrical network.

23.6   Application II: estimation of stochastic differential equations

It is often convenient to formulate theoretical models in terms of differential equations, and when the observation frequency is high (e.g., weekly, daily, hourly or real-time) it may be more natural to adopt this framework for econometric models of time series.

The most common approach to estimation of stochastic differential equations is to "discretize" the model, as above, and estimate using the discretized version. However, since the discretization is only an approximation to the true discrete-time version of the model (which is not calculable), the resulting estimator is in general biased and inconsistent.

An alternative is to use indirect inference: The discretized model is used as the score generator. That is, one estimates by QML to obtain the scores of the discretized approximation:

$$
\begin{aligned}
y_t - y_{t-1} &= g(\phi, y_{t-1}) + h(\phi, y_{t-1})\varepsilon_t \\
\varepsilon_t &\sim N(0,1)
\end{aligned}
$$

Indicate these scores by $m_n(\theta, \hat{\phi})$. Then the system of stochastic differential equations

$$dy_t = g(\theta, y_t)dt + h(\theta, y_t)dW_t$$

is simulated over θ, and the scores are calculated and averaged over the simulations

$$\tilde{m}_n(\theta, \hat{\phi}) = \frac{1}{N}\sum_{i=1}^{N} m_{in}(\theta, \hat{\phi})$$

$\hat{\theta}$ is chosen to set the simulated scores to zero

$$\tilde{m}_n(\hat{\theta}, \hat{\phi}) \equiv 0$$

(since θ and ϕ are of the same dimension).

This method requires simulating the stochastic differential equation. There are many ways of doing this. Basically, they involve doing very fine discretizations:

$$
\begin{aligned}
y_{t+\tau} &= y_t + g(\theta, y_t) + h(\theta, y_t)\eta_t \\
\eta_t &\sim N(0, \tau)
\end{aligned}
$$

By setting τ very small, the sequence of η_t approximates a Brownian motion fairly well.

This is only one method of using indirect inference for estimation of differential equations. There are others (see Gallant and Long, 1995 and Gourieroux *et. al.*). Use of a series approximation to the transitional density as in Gallant and Long is an interesting possibility since the score generator may have a higher dimensional parameter than the model, which allows for diagnostic testing. In the method described

above the score generator's parameter ϕ is of the same dimension as is θ, so diagnostic testing is not possible.

23.7 Application III: estimation of a multinomial probit panel data model

For selection of one alternative out of G, let the vector y be G-dimensional (high enough so that direct probit is not feasible). Only one element is equal to 1, indicating the alternative chosen, while the rest are zero. The choice depends upon the characteristics of the alternatives, $\mathbf{x}_i, i = 1, 2, ..., G$. While one can estimate a multinomial probit (MNP) model using SML or MSM, one looses the diagnostic testing possibilities of indirect inference.

For example, the score generator could be a multinomial logit model (MNL) model, characterized by choice probabilities of the form

$$\Pr(y_i = 1) = \frac{\exp(\mathbf{x}_i'\beta)}{\sum_{j=1}^{G} \exp(\mathbf{x}_j'\beta)}.$$

These are tractable for any dimension G. The reason the multinomial probit is to be preferred over the multinomial logit is that the MNL suffers from a problem of lack of "independence of irrelevant alternatives". For example, if we have a problem of choice between travel to work by car and red bus, the probabilities of selection of these modes of transit are P_C and P_{RB}. According to the MNL model, if we add the possibility of travel by blue bus, P_C will drop, since the numerator doesn't change but the denominator does. The MNP model is more satisfactory since the covariance matrix Ω of the errors (see equation [47]) allows for complementarity and substitutability of alternatives).

24 Thanks

The following is a list of people who have contributed to these notes in some form.

A number of *IDEA* students - error corrections

Montserrat Farell - error corrections

25 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 COPY-ING, 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. (Here-

402

inafter, 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

403

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) 19yy <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) 19yy 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.

Index