

INTRODUCTORY GRADUATE ECONOMETRICS

Craig Burnside

Department of Economics

University of Pittsburgh

Pittsburgh, PA 15260

January 2, 1994

*These are incomplete notes intended for use in an introductory graduate econometrics course. As notes, the style of presentation is deliberately informal and lacking in proper citations. Please point out any errors you find.

1. Ordinary Least Squares

We have a sequence of observations on a random variable

$$y_t, \quad t = 1, 2, \dots, T.$$

The T indicates I'm a time series guy. Furthermore we have an economic model which tells us that y is a linear function of explanatory variables plus a random component. I.e.

$$y_t = x_t' \beta + \epsilon_t$$

where x_t is a $k \times 1$ vector of explanatory variables, β is a $k \times 1$ vector and y_t and ϵ_t are scalars. Written out this is

$$y_t = (x_{1t} \quad x_{2t} \quad \dots \quad x_{kt}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \epsilon_t.$$

Typically $x_{1t} = 1$, for all t . If the observations are stacked we get

$$y_1 = \beta_1 x_{11} + \beta_2 x_{21} + \dots + \beta_k x_{k1} + \epsilon_1$$

$$y_2 = \beta_1 x_{12} + \beta_2 x_{22} + \dots + \beta_k x_{k2} + \epsilon_2$$

$$\vdots$$

$$y_T = \beta_1 x_{1T} + \beta_2 x_{2T} + \dots + \beta_k x_{kT} + \epsilon_T$$

or

$$y = X\beta + \epsilon$$

with y and ϵ $T \times 1$ vectors, X a $T \times k$ matrix.

1.1. Assumptions

- (1) X is a matrix of fixed variables (unrealistic) and has full column rank (i.e. the columns of X are linearly independent)
- (2) $E(\epsilon) = 0$
- (3) $E(\epsilon\epsilon') = \sigma^2 I_T$ or more strongly, the ϵ_t are i.i.d.
- (4) β is an unknown constant parameter vector, and σ^2 is an unknown scalar.

1.2. Estimation

We want to estimate β . The first method is to use the least squares criterion. If I choose some estimate $\hat{\beta}$, the difference between y and $X\hat{\beta}$ is called a *residual*.

Residuals $e = y - X\hat{\beta}$. Minimize the sum of the squared residuals by choice of $\hat{\beta}$.

$$SSE = \sum_{i=1}^T e_t^2 = e'e = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

Minimize SSE by choosing $\hat{\beta}$ so that $\partial SSE / \partial \hat{\beta} = 0$.

$$\frac{\partial SSE}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

Solution is $\hat{\beta} = (X'X)^{-1}X'y$. To verify that this is a minimum compute the matrix of second derivatives

$$\frac{\partial^2 SSE}{\partial \hat{\beta} \partial \hat{\beta}'} = 2X'X$$

which is positive definite. See Rule 5 p.961 Red Judge to see that this follows from our assumptions. Our solution is therefore the unique minimum.

We will also want to have an estimate of the variance σ^2 . The estimator we will use is

$$\hat{\sigma}^2 = \frac{1}{T-k} \sum_{t=1}^T e_t^2 = \frac{e'e}{T-k}$$

1.3. Sampling Properties

Since the vector y is random, this means $\hat{\beta}$ is a vector-valued random variable. Similarly, $\hat{\sigma}^2$ is a random variable. Therefore, we can ask questions about the distributions of these random variables. The first property we will consider is the mean of these random variables. We will show that both estimators are *unbiased*. An estimator $\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta$.

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

$$\begin{aligned}
E(\hat{\sigma}^2) &= E\left[\frac{1}{T-k}e'e\right] \\
&= \frac{1}{T-k}E(e'e)
\end{aligned}$$

Consider the definition of the residuals.

$$\begin{aligned}
e &= y - X\hat{\beta} = y - X(X'X)^{-1}X'y \\
&= \left[I - X(X'X)^{-1}X'\right]y = My \\
&= \left[I - X(X'X)^{-1}X'\right](X\beta + \epsilon) = M\epsilon
\end{aligned}$$

Some facts about M . It is $T \times T$ but has rank $T - k$. It is *symmetric* as $M = M'$. Furthermore it is *idempotent* as $MM = M$. As a result of all this (trace is sum of elements on diagonal of a square matrix)

$$\begin{aligned}
e'e &= \text{tr}(e'e) = \text{tr}(\epsilon'M'M\epsilon) = \text{tr}(\epsilon'M\epsilon) \\
E(e'e) &= E\left[\text{tr}(\epsilon'M\epsilon)\right] \\
&= E\left[\text{tr}(M\epsilon\epsilon')\right] \quad \text{Rule : } \text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA) \\
&= \text{tr}\left[ME(\epsilon\epsilon')\right] \\
&= \text{tr}(M\sigma^2I) = \sigma^2\text{tr}(M) \\
&= \sigma^2(T - k)
\end{aligned}$$

The last step follows from the fact that an idempotent matrix of rank s has trace equal to s . Therefore $E(\hat{\sigma}^2) = \sigma^2$.

We will also derive the variance-covariance matrix of $\hat{\beta}$. I.e. we will compute (give explanation of what's in there) $V(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$. To do this, note that $\hat{\beta} - \beta =$

$(X'X)^{-1}X'y - \beta = (X'X)^{-1}X'(X\beta + \epsilon) - \beta = (X'X)^{-1}X'\epsilon$. Therefore,

$$\begin{aligned} V(\hat{\beta}) &= E\left[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'E(\epsilon\epsilon')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

1.4. The Gauss-Markov Theorem

Recall that the elements of $\hat{\beta}$ are linear combinations of the elements of y . I.e. it is a linear estimator. Suppose we consider any other linear estimator $\bar{\beta} = Ay$ which is also unbiased. I.e. $E(\bar{\beta}) = \beta$. Then $V(\bar{\beta}) \geq V(\hat{\beta})$. Proof: First note that A is a $k \times T$ matrix. Since $\bar{\beta} = Ay = AX\beta + A\epsilon$ we have $E(\bar{\beta}) = AX\beta = \beta$ by the assumption of unbiasedness. This means $AX = I_k$. However this does not imply $A = X^{-1}$ since neither A nor X is invertible. Notice that the last result means that $\bar{\beta} = \beta + A\epsilon$ or $\bar{\beta} - \beta = A\epsilon$. Therefore

$$\begin{aligned} V(\bar{\beta}) &= E\left[(\bar{\beta} - \beta)(\bar{\beta} - \beta)'\right] \\ &= E(A\epsilon\epsilon'A') \\ &= \sigma^2AA' \end{aligned}$$

Now define $C = A - (X'X)^{-1}X'$. This means that $A = C + (X'X)^{-1}X'$. Some neat facts about C . CC' is a $k \times k$ positive semi-definite matrix. $CX = AX - (X'X)^{-1}X'X = 0$! Therefore,

$$\begin{aligned} V(\bar{\beta}) &= \sigma^2[C + (X'X)^{-1}X'] [C + (X'X)^{-1}X']' \\ &= \sigma^2[C + (X'X)^{-1}X'] [C' + X(X'X)^{-1}] \\ &= \sigma^2[CC' + (X'X)^{-1}X'C' + CX(X'X)^{-1} + (X'X)^{-1}X'X(X'X)^{-1}] \\ &= \sigma^2[CC' + (X'X)^{-1}] \end{aligned}$$

Therefore, $V(\bar{\beta}) - V(\hat{\beta})$ is a positive semidefinite matrix. This proves that OLS is BLUE given our assumptions.

1.5. Further Statistical Properties

Suppose we make a further assumption that

- (5) The errors are normally distributed.

This last assumption implies that $\hat{\beta}$ is a normally distributed random vector since it is a linear combination of the ϵ_t .

Also define $C^2 = (T - k)\hat{\sigma}^2/\sigma^2 = e'e/\sigma^2$. Recall that $e = M\epsilon$. Therefore,

$$\begin{aligned} C^2 &= \frac{\epsilon' M' M \epsilon}{\sigma^2} \\ &= \frac{\epsilon'}{\sigma} M \frac{\epsilon}{\sigma}. \end{aligned}$$

Notice that $\epsilon/\sigma \sim N(0, I)$ and that M is idempotent of rank k . This implies that $C^2 \sim \chi^2(T - k)$. Check Red Judge for this trivia tidbit!

We could use both of these last results to conduct hypothesis tests if we so desired. Unfortunately we can't use the first result directly since $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ involves the unknown parameter vector σ^2 . Recall that a t -distributed random variable is defined as $t(n) = z/\sqrt{x/n}$ where z is standard normal, x is $\chi^2(n)$ and z and x are independent. Thus, you see why we use the t -statistics

$$\begin{aligned} t_i &= \frac{(\hat{\beta}_i - \beta_i)}{\sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{ii}}} \\ &= \frac{(\hat{\beta}_i - \beta_i)}{\sqrt{\sigma^2[(X'X)^{-1}]_{ii}}} / \sqrt{\hat{\sigma}^2/\sigma^2} \\ &= \frac{(\hat{\beta}_i - \beta_i)}{\sqrt{\sigma^2[(X'X)^{-1}]_{ii}}} / \sqrt{C^2/(T - k)} \\ &= z/\sqrt{x/n}. \end{aligned}$$

Thus if we wanted to test $H_0 : \beta_i = \beta_{i0}$, we could exploit the fact that t_i is distributed $t(T - k)$. Notice that I didn't verify independence in my proof. This would make a good assignment question.

We might also want to conduct tests such as $H_0 : R\beta = r$, where R is a $j \times k$ matrix.

I.e. we're looking at j joint linear restrictions on β . Notice that

$$\begin{aligned} R\hat{\beta} - R\beta &= R(\hat{\beta} - \beta) \\ &= R(X'X)^{-1}X'\epsilon \\ &\sim N\left(0, \sigma^2 R(X'X)^{-1}R'\right) \end{aligned}$$

The next step involves finding the Cholesky decomposition of $\sigma^2 R(X'X)^{-1}R'$. I.e. find C , lower triangular, and invertible such that $CC' = \sigma^2 R(X'X)^{-1}R'$. Now define $Z = C^{-1}(R\hat{\beta} - R\beta)$. From the definition of C we have $Z \sim N(0, I_j)$. This means $Z'Z$ is $\chi^2(j)$.

Or

$$(R\hat{\beta} - R\beta)'C^{-1'}C^{-1}(R\hat{\beta} - R\beta) = \frac{(R\hat{\beta} - R\beta)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - R\beta)}{\sigma^2} \sim \chi^2(j).$$

Again we have a problem since we don't know σ^2 . When we use $\hat{\sigma}^2$ instead we get the familiar F -statistic. To see this recall that $F(n_1, n_2) = (\chi_1^2/n_1)/(\chi_2^2/n_2)$ where χ_1^2 and χ_2^2 are independent. Therefore,

$$\begin{aligned} F(j, T - k) &= \frac{(R\hat{\beta} - R\beta)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - R\beta)/\sigma^2 j}{[(T - k)\hat{\sigma}^2/\sigma^2]/(T - k)} \\ &= \frac{(R\hat{\beta} - R\beta)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - R\beta)/j}{\hat{\sigma}^2} \end{aligned}$$

Are they independent χ^2 ? Good assignment question!

1.6. Measures of Fit

Measures of fit are used to summarize the extent to which the estimated model ‘fits’ the data. Notice that by the definition of the residuals $y = X\hat{\beta} + e$. This implies that $y - \bar{y}1 = X\hat{\beta} - \bar{y}1 + e$. Thus the total sum of squared (SST) deviations of y_t around its mean is given by

$$\begin{aligned}
 SST &= (y - \bar{y}1)'(y - \bar{y}1) \\
 &= (X\hat{\beta} - \bar{y}1 + e)'(X\hat{\beta} - \bar{y}1 + e) \\
 &= (\hat{y} - \bar{y}1 + e)'(\hat{y} - \bar{y}1 + e) \\
 &= (\hat{y} - \bar{y}1)'(\hat{y} - \bar{y}1) + e'e - 2\bar{y}1'e \\
 &= (\hat{y} - \bar{y}1)'(\hat{y} - \bar{y}1) + e'e \quad \text{if } 1 \in X \\
 &= SSR + SSE
 \end{aligned}$$

The R^2 is defined to be $R^2 = SSR/SST = 1 - SSE/SST$. Clearly this can simply be increased by adding regressors indiscriminately because of the extra degree of freedom to reduce SSE that is introduced by adding a parameter. Thus the adjusted R^2 is introduced to adjust for the number of regressors. It is defined as

$$\begin{aligned}
 \bar{R}^2 &= \left(\frac{T-1}{T-k} \right) R^2 - \left(\frac{k-1}{T-k} \right) \\
 &= 1 - \left(\frac{T-1}{T-k} \right) (1 - R^2).
 \end{aligned}$$

1.7. Geometry

We saw earlier that $e = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = [I - X(X'X)^{-1}X']y = My$. M was shown to be symmetric and idempotent. Now, let's also consider the fitted values $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Py$. P is also symmetric and idempotent, but this time with rank k .

The matrix P projects a $T \times 1$ vector y onto the space spanned by the k $T \times 1$ vectors which are the columns of X . I.e. P produces a linear combination of the columns of X , \hat{y} ,

whose deviation from y , e , is orthogonal to the space spanned by those columns. M would project any vector onto the orthogonal complement of that space (because it produces the residuals which are orthogonal to X). Notice that $\hat{y} = Py$ so that $y - \hat{y} = y - Py = (I - P)y = My = e$.

Things to note are the following.

1. $X'e = X'My = X'[I - X(X'X)^{-1}X']y = [X' - X'X(X'X)^{-1}X']y = 0$. Notice that only if X contains a 1 vector does $\sum_{t=1}^T e_t = 1'e = 0$. Furthermore we have $X'\hat{y} = X'Py = X'X(X'X)^{-1}X'y = X'y$.

2.

$$\begin{aligned} y'y &= (X\hat{\beta} + e)'(X\hat{\beta} + e) \\ &= \hat{\beta}'X'X\hat{\beta} + e'X\hat{\beta} + \hat{\beta}'X'e + e'e \\ &= \hat{\beta}'X'X\hat{\beta} + e'e. \end{aligned}$$

I.e. (raw) $SST = SSR + SSE$.

3. Suppose I transform the variables before regression by applying a nonsingular $k \times k$ matrix A to X . I.e. I generate $X^* = XA$. Then notice that if you regress y on X^* you get

$$\begin{aligned} P^* &= X^*(X^{*'}X^*)^{-1}X^{*'} \\ &= XA(A'X'XA)^{-1}A'X' \\ &= XAA^{-1}(X'X)^{-1}A'^{-1}A'X' \\ &= X(X'X)^{-1}X' = P. \end{aligned}$$

Thus, the predicted values from the two regressions will be the same. What about the coefficients? $\hat{\beta}^* = (X^{*'}X^*)^{-1}X^{*'}y = A^{-1}(X'X)^{-1}A'^{-1}A'Xy = A^{-1}\hat{\beta}$. The coefficients are changed by the inverse of the linear transformation.

4. Suppose I regress y on X and Z where the model is $y = X\beta + Z\gamma + \epsilon$. Define $W = (X \ Z)$. Also define $\theta' = (\beta' \ \gamma')'$. Then $y = W\theta + \epsilon$. Thus we'll get $\hat{\theta} = (W'W)^{-1}W'y$.

Expanding this we get

$$\begin{aligned} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} &= \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X' \\ Z' \end{pmatrix} y \\ &= H^{-1} \begin{pmatrix} X' \\ Z' \end{pmatrix} y \end{aligned}$$

This means that the estimate of γ is given by

$$\begin{aligned}
 \hat{\gamma} &= H_{21}^{-1}X'y + H_{22}^{-1}Z'y \\
 &= -(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}H_{21}H_{11}^{-1}X'y + (H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}Z'y \\
 &= -[Z'Z - Z'X(X'X)^{-1}X'Z]^{-1}(Z'X)(X'X)^{-1}X'y + \\
 &\quad [Z'Z - Z'X(X'X)^{-1}X'Z]^{-1}Z'y \\
 &= -\left(Z'[I - X(X'X)^{-1}X']Z\right)^{-1}Z'P_x y + \left(Z'[I - X(X'X)^{-1}X']Z\right)^{-1}Z'y \\
 &= -(Z'M_x Z)^{-1}Z'P_x y + (Z'M_x Z)^{-1}Z'y \\
 &= (Z'M_x Z)^{-1}Z'(I - P_x)y \\
 &= (Z'M_x Z)^{-1}Z'M_x y
 \end{aligned}$$

Suppose that instead I did the following. First regress y onto X and compute the residuals $y^* = M_x y$ from that regression. Also regress each of the columns of Z onto X and make a matrix out of all the different residual series $Z^* = M_x Z$. Now regress y^* onto Z^* . You'll get

$$\begin{aligned}
 \hat{\gamma} &= (Z^{*'}Z^*)^{-1}Z^{*'}y^* \\
 &= (Z'M'_x M_x Z)^{-1}Z'M'_x M_x y \\
 &= (Z'M_x Z)^{-1}Z'M_x y
 \end{aligned}$$

Notice the equivalence. This result is called the Frisch-Waugh-Lovell Theorem.

2. Hypothesis Testing

There are different approaches to hypothesis testing. The ones you will be introduced to in this course are all *classical* hypothesis tests. The term *classical* refers to the basic approach to testing. Essentially, the *classical* approach works as follows.

1. Determine the null hypothesis to be tested.
2. Find some testable implication of the null hypothesis. Usually this will pertain to the distribution of some statistic under the null.
3. Set up the test. What this means is that you have to have a rejection region for the test. I.e. you must define a region such that if the statistic ever lies in that region you will reject the null hypothesis.
4. Actually compute the test statistic for your sample and determine the outcome of the test.

How does one actually construct the test? In the classical approach, the parameters of the model are treated as unknown fixed constants. This is as opposed to the Bayesian approach where the econometrician actually assigns a prior probability distribution to the parameters, and then updates this to form a posterior given the observed sample of data. Because of the inherent assumption of the classical approach there is always assumed to exist some unknown true value of the parameter in question, unless the model is misspecified.

Classical hypothesis tests are constructed with the following two considerations in mind.

1. SIZE - this is the probability of rejecting the null hypothesis even though the null hypothesis is true, i.e. it is the probability of type I error.
2. POWER - this is the probability of rejecting the null hypothesis when it is false, i.e. it is 1 minus the probability of type II error.

To evaluate the size or power of tests one clearly needs to compute the probabilities

in question. One way to do this would be to perform the conceptual experiment of an infinite number of repeated samples. For example, suppose I wanted to know the weight assigned to heads on a coin. I either know the true weight (i.e. I know the probability distribution governing the outcome) or I can flip the coin an infinite number of times. We know that the proportion of outcomes which are heads will converge to the true proportion with a large enough number of flips. That's what's really going on in hypothesis testing.

Size is determined by assuming the null hypothesis is true. If it is true, then the question is what is the probability in any given sample that I will end up rejecting the null. Sometimes, this can be calculated by carefully working out the probability distribution of the test statistic. Other times, it can be determined by doing a lot of repeated sampling in a controlled experiment. This is what is referred to as *Monte-Carlo simulation*.

For a given null hypothesis, size is a fixed number α . However, power clearly depends on what the true parameter actually is. For example, suppose I set up the following test for whether the mean of a normally distributed random variable was zero or not. Assume that we know the variable is distributed normally, and that the variance is 1. The null hypothesis is $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Suppose I decide to reject the null hypothesis whenever \bar{x} from T observations on X below -0.1 or above 0.1 . What do we know about the sample mean when the draws are i.i.d? A testable implication is that if H_0 is true, the sample mean \bar{x} is distributed as $N(0, 1/T)$. This is a standard result. We know that the distribution of the sample mean of a $N(\mu, \sigma^2)$ is $N(\mu, \sigma^2/T)$. Therefore the size in my example is just

$$1 - \Pr(-0.1 < \bar{x} < 0.1) = 1 - \Pr\left(\frac{-0.1}{\sqrt{T}} < Z < \frac{0.1}{\sqrt{T}}\right)$$

which is just some number for any fixed T . However what is the power of the test? It is the probability of rejecting the null hypothesis when it is false. When it is false, we want to standardize \bar{x} with the correct (true) mean, rather than 0. Therefore power is

$$1 - \Pr(-0.1 < \bar{x} < 0.1) = 1 - \Pr\left(\frac{-0.1 - \mu}{\sqrt{T}} < Z < \frac{0.1 - \mu}{\sqrt{T}}\right)$$

which is clearly a function of μ . Suppose μ is very close to zero. Then power will approximately equal the size of the test. For the test I've set up, the power gets bigger the further μ gets from zero.

The way hypothesis tests are typically designed is to choose a test with a given amount of size. Then among tests of a given size the natural thing to do would be to look for the most powerful test. This isn't always the approach, especially when asymptotic tests are being used. However, there is a theory of *uniformly most powerful tests* which tries to find exactly that. It turns out that usually such a test is not available. However, for certain classes of distributions it's often the case that there is a UMP test among the class of *unbiased* tests, i.e. tests for which the power is always greater than or equal to the size.

This is what provides the logic behind the typical tail rejection regions that we use. For example, to test whether $\mu = 0$, in the example above you usually want to set up the statistic $Z_S = \frac{\bar{x}}{\sigma/\sqrt{T}}$. Usually we set up a 5% test where the rejection region is $\{z \mid |z| > 1.96\}$. Clearly when the null hypothesis is true the probability of rejection is 5%. However, there are an infinite number of critical regions which give tests whose sizes are 5%. To be specific our rejection region, C , would only have to satisfy the condition that $\int_C \phi(z) dz = .05$. There are many such regions. Lets try out a few, and examine their power properties.

- i. $C_1 = \{z \mid |z| > 1.96\}$,
- ii. $C_2 = \{z \mid |z| < 0.0625\}$,
- iii. $C_3 = \{z \mid 0.5960 < |z| < 0.7535\}$.

All three critical regions have size of 5%, but their power properties are dramatically different. Look at the power functions on the following page. These are constructed for the case where σ is known to be 1 and $T = 100$, and x_t is normally distributed and i.i.d. The region C_1 is the standard rejection region, and is seen to have maximal power, approaching 1, against the most distant alternatives to $H_0 : \mu = 0$, and weakest power, approaching

0.05 against those alternatives which are closest to H_0 . The region C_2 leads to size of 5% but has obviously undesirable power properties. Because it is centered around zero, power is maximal at $\mu = 0$, and is never greater than 0.05. Furthermore, power is zero against the most distant alternatives. The region C_3 , which is off-centered but involves no tail area, again leads to terrible power against the most distant alternatives, although power is now maximized to the right of H_0 . However, it has slightly higher power against some nearby alternatives than does C_1 , although this gain in power is very hard to discern in the graphs.

Figure 2.1: Power

3. Maximum Likelihood

We have a sequence of observations on a random variable

$$y_t, \quad t = 1, 2, \dots, T.$$

Suppose the joint density function of this sequence of random variables is given by

$$f(y_1, y_2, \dots, y_T | \theta)$$

where θ is some parameter (vector) of that joint distribution. f is called the *likelihood function* when written as a function of θ , i.e. $L(\theta | y_1, y_2, \dots, y_T) = f(y_1, y_2, \dots, y_T | \theta)$. The *log-likelihood function* is given by $\mathcal{L}(\theta | y) = \ln[L(\theta | y)]$, where y is the vector of observations as in section 1.

3.1. The Linear Model

In the linear model we had $y_t = x_t' \beta + \epsilon_t$. To determine the likelihood function for y let us first consider the joint distribution or likelihood of the vector ϵ . Given assumptions (1)–(5), in section 1, we have

$$\begin{aligned} f(\epsilon) &= \prod_{t=1}^T \phi(\epsilon_t) \quad \text{by independence} \\ &= \prod_{t=1}^T \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} \epsilon_t^2\right) \\ &= (2\pi)^{-\frac{T}{2}} \sigma^{-T} \exp\left(-\frac{1}{2\sigma^2} \epsilon' \epsilon\right) \end{aligned}$$

Now to get the distribution of y we have to use the following fact. If any variable y is a function of ϵ , $g(\epsilon)$, then the joint distribution of the y 's is

$$f_Y(y) = f[g^{-1}(y)] |J| \quad \text{where} \quad J = \frac{\partial \epsilon}{\partial y} = \frac{\partial g^{-1}(y)}{\partial y}.$$

In the linear model $y = X\beta + \epsilon$ so that $\epsilon = y - X\beta$. Therefore $J = I$ and $|J| = 1$. Thus

$$L(\theta | y) = f(y) = (2\pi)^{-\frac{T}{2}} \sigma^{-T} \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta)\right)$$

and the log-likelihood is

$$\begin{aligned}\mathcal{L}(\theta|y) &= -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \\ &= -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - x_t'\beta)^2.\end{aligned}$$

In this case, $\theta = (\beta' \ \sigma^2)'$. The *maximum likelihood estimator* (MLE) for θ is obtained by maximizing the likelihood function. Since \ln is a monotonic transformation we can maximize the log-likelihood instead. We will do this by solving the first order conditions for choices of β and σ^2 .

First expand the definition of \mathcal{L} :

$$\mathcal{L} = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y'y - 2\beta'X'y + \beta'X'X\beta)$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta} &= -\frac{1}{2\sigma^2} (-2X'y + 2X'X\beta) \\ \frac{\partial \mathcal{L}}{\partial \sigma^2} &= -\frac{T}{2} \sigma^{-2} + \frac{1}{2} \sigma^{-4} (y - X\beta)'(y - X\beta)\end{aligned}$$

If we solve these FONCs for β and σ^2 we get MLE estimators $\tilde{\beta} = (X'X)^{-1}X'y$ and $\tilde{\sigma}^2 = \frac{1}{T}e'e$ where $e = y - X\tilde{\beta}$. Notice that $\tilde{\beta}$ can be solved for independently of σ^2 and is the same as the OLS estimator. We already know that $E(\tilde{\beta}) = \beta$. Since $\tilde{\sigma}^2 = (T - k)\hat{\sigma}^2/T$ we also know that $E(\tilde{\sigma}^2) = (T - k)\sigma^2/T$. Furthermore, $V(\tilde{\beta}) = \sigma^2(X'X)^{-1}$. The variance of the variance estimator is

$$\begin{aligned}E\left((\tilde{\sigma}^2 - E\tilde{\sigma}^2)^2\right) &= E\left(\left(\frac{1}{T}e'e - \frac{T - k}{T}\sigma^2\right)^2\right) \\ &= \frac{\sigma^4}{T^2} E\left[\left(\frac{e'e}{\sigma^2} - (T - k)\right)^2\right] \\ &= \frac{\sigma^4}{T^2} E\left[\left((T - k)\frac{\hat{\sigma}^2}{\sigma^2} - (T - k)\right)^2\right] \\ &= \frac{\sigma^4}{T^2} E\left[\left(\chi^2 - (T - k)\right)^2\right] \\ &= 2(T - k)\sigma^4/T^2\end{aligned}$$

because the mean of a $\chi^2(T - k)$ is $T - k$ and its variance is $2(T - k)$. As for covariance

$$\begin{aligned} \text{Cov}(\tilde{\beta}, \tilde{\sigma}^2) &= E\left((\tilde{\beta} - \beta)(\tilde{\sigma}^2 - E\tilde{\sigma}^2)\right) \\ &= E\left[\left((X'X)^{-1}X'\epsilon\right)\left(\frac{1}{T}e'e - \frac{T - k}{T}\sigma^2\right)\right] \\ &= 0 \end{aligned}$$

This is because the first term is a linear combination of a normal, while the second term is a quadratic form (based on M) in the same normal. These are independent if the two matrices multiplied give you 0. Of course $(X'X)^{-1}X'M = 0$. Therefore the variance covariance matrix

$$V(\theta) = \begin{pmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & 2\frac{T-k}{T}\sigma^4 \end{pmatrix}$$

3.2. Cramer-Rao Lower Bound

The Cramer-Rao Lower Bound simply states that any unbiased estimator of a parameter vector θ will have greater variance than the inverse of a matrix called $I(\theta)$. I.e. if $\bar{\theta}$ is an unbiased estimator of θ , $V(\bar{\theta}) - I(\theta)^{-1}$ is a p.s.d. matrix, where

$$I(\theta) = -E\left(\frac{\partial^2 \mathcal{L}}{\partial\theta\partial\theta'}\right).$$

This is a general principle applying to any model with a likelihood. To find the matrix I called the *information matrix*, we need to compute the second derivatives.

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial\beta\partial\beta'} &= -\frac{1}{\sigma^2}X'X \\ \frac{\partial^2 \mathcal{L}}{\partial\beta\partial\sigma^2} &= \frac{1}{\sigma^4}(-X'y + X'X\beta) \\ \frac{\partial^2 \mathcal{L}}{\partial\sigma^2\partial\beta'} &= \frac{1}{\sigma^4}(-y'X + \beta'X'X) \\ \frac{\partial^2 \mathcal{L}}{\partial(\sigma^2)^2} &= \frac{T}{2}\sigma^{-4} + -\sigma^{-6}(y - X\beta)'(y - X\beta) \end{aligned}$$

The negation of the expectation of these terms is

$$\begin{aligned} I(\theta) &= \begin{pmatrix} \sigma^{-2}X'X & 0 \\ 0 & -\frac{T}{2}\sigma^{-4} + \sigma^{-6}T\sigma^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma^{-2}X'X & 0 \\ 0 & \frac{T}{2}\sigma^{-4} \end{pmatrix}. \end{aligned}$$

Thus the CRLB for θ is

$$I(\theta)^{-1} = \begin{pmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2}{T}\sigma^4 \end{pmatrix}.$$

Another useful concept in maximum likelihood theory is the sufficient statistic. A *sufficient statistic* is a statistic which the likelihood can be completely characterized in terms of. I.e. the random data no longer appear in the formula for the likelihood. In this case

$$L(y|\theta) = (2\pi)^{-T/2}\sigma^{-T} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right).$$

Now $y - X\beta = (y - X\hat{\beta}) + (X\hat{\beta} - X\beta) = e + X(\hat{\beta} - \beta)$, so that

$$\begin{aligned} (y - X\beta)'(y - X\beta) &= e'e + 2e'X(\hat{\beta} - \beta) + (\hat{\beta}' - \beta')X'X(\hat{\beta} - \beta) \\ &= (T - k)\hat{\sigma}^2 + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \end{aligned}$$

Substituting this into L gets rid of the data y . Thus $\hat{\beta}$ and $\hat{\sigma}^2$ are jointly sufficient statistics for L . A theorem in statistics states that if a sufficient statistic is unbiased it is minimum variance unbiased. Therefore, $\hat{\beta}$ and $\hat{\sigma}^2$ are the MVUE of β and σ^2 .

Since the OLS estimators are the MVUEs, this might raise the question: why use ML? It turns out that in cases where we have a growing sample size and we need to use asymptotic theory (we'll see cases like this later), ML estimators have the following properties:

1. Consistency (sort of asymptotic unbiasedness)
2. Asymptotic normality of the estimator regardless of distribution of ϵ
3. Asymptotic efficiency - achieving the CRLB.

We will discuss asymptotic theories more completely when we get to cases with stochastic regressors.

3.3. A Brief Discussion of Asymptotics

What is all this asymptotic stuff? Well, we have to diverge a little bit to talk about it. We need definitions. This is different from what we've been talking about because the sample size T is treated as variable. What does this mean for things like X fixed? It just means that in repeated samples of any size we'd get the same sequence of x_t 's. We will also need to redefine some things because of scale factors.

CONSISTENCY: An estimator $\hat{\theta}_T$ converges in probability to c if

$$\lim_{T \rightarrow \infty} \Pr(|\hat{\theta}_T - c| > \epsilon) = 0.$$

This is sometimes denoted $c = \text{plim } \hat{\theta}_T$. $\hat{\theta}_T$ is consistent if $\theta = \text{plim } \hat{\theta}_T$.

ASYMPTOTIC NORMALITY: A random variable X_T converges in distribution to a random variable X if

$$\lim_{T \rightarrow \infty} |F_T(x) - F(x)| = 0$$

at all continuity points of X . Typically, we'll see cases where $\hat{\theta}_T$ is asymptotically normal if $\sqrt{T}(\hat{\theta}_T - \theta)$ converges in distribution to $N(0, V)$.

To illustrate this let's take a scalar i.i.d. random variable $x_t \sim N(\mu, \sigma^2)$. The estimator I'll consider is $\bar{x}_T = \sum_{t=1}^T x_t/T$. Because I've assumed normality we're obviously going to get an exact normal distribution for \bar{x}_T for any T , because it's a linear combination of normal random variables. The mean is clearly μ and because the x_t are i.i.d. we easily get $V(\bar{x}_T) = \sigma^2/T$. Can we prove consistency? Yes. $\bar{x}_T - \mu = w_T \sim N(0, \sigma^2/T)$. Therefore, $\Pr(|\bar{x}_T - \mu| > \epsilon) = \Pr(\bar{x}_T - \mu > \epsilon) + \Pr(\bar{x}_T - \mu < -\epsilon) = \Pr(Tw_T/\sigma > T\epsilon/\sigma) + \Pr(Tw_T/\sigma < -T\epsilon/\sigma) = \Pr(Z > T\epsilon/\sigma) + \Pr(Z < -T\epsilon/\sigma)$. The limit as $T \rightarrow \infty$ of this quantity is clearly 0.

The important thing about asymptotic normality is that you must blow up by \sqrt{T} . Note in the example the variance goes to zero if you don't blow it up, while it is σ^2 if you do. This would be unnecessary if you always had an exact small sample distribution as

we have in the example and all the work so far, but it is not helpful in cases where we can't characterize the small sample distribution. When this happens we use the asymptotic distribution to approximate the small sample distribution, which is only useful if the asymptotic distribution is nonredundant.

3.4. The 3 Tests

We will now discuss three methods for testing restrictions in the context of ML estimation. Specifically we're interested in tests of the form $H_0 : g(\theta) = 0$ vs $H_A : g(\theta) \neq 0$. The function $g : \mathbf{R}^k \rightarrow \mathbf{R}^j$.

3.4.1. The Likelihood Ratio Test

Denote $\tilde{\theta} = \arg \max_{\theta} \mathcal{L}(\theta|y)$ while $\ddot{\theta} = \arg \max_{\theta} \mathcal{L}(\theta|y)$ s.t. $g(\theta) = 0$. Clearly $\mathcal{L}(\tilde{\theta}|y) \geq \mathcal{L}(\ddot{\theta}|y)$. The test is based on how big the difference is. If it is large the hypothesis is brought into question.

$$LR = -2[\mathcal{L}(\ddot{\theta}|y) - \mathcal{L}(\tilde{\theta}|y)] = -2 \log(r) \quad \text{where} \quad r = \frac{L(\ddot{\theta}|y)}{L(\tilde{\theta}|y)}.$$

It turns out that LR converges in distribution to $\chi^2(j)$ as $T \rightarrow \infty$. The LR test requires that you estimate the model under both the null and the alternative. Sometimes this isn't practical depending on the form of the function g .

In our linear model suppose we used the LR test. Having estimated a restricted and an unrestricted model we would have two parameter vectors $\tilde{\beta}$ and $\ddot{\beta}$. Then LR is given by

$$\begin{aligned} LR &= -2[\mathcal{L}(\ddot{\beta}) - \mathcal{L}(\tilde{\beta})] \\ &= -2\left(-\frac{T}{2} \log(\ddot{\sigma}^2) - \frac{1}{2\ddot{\sigma}^2} \ddot{e}'\ddot{e} + \frac{T}{2} \log(\tilde{\sigma}^2) + \frac{1}{2\tilde{\sigma}^2} \tilde{e}'\tilde{e}\right) \\ &= -2\left(-\frac{T}{2} \log(\ddot{e}'\ddot{e}) + \frac{T}{2} \log(\tilde{e}'\tilde{e})\right) \\ &= T \log(\tilde{e}'\tilde{e}/\ddot{e}'\ddot{e}) \end{aligned}$$

3.4.2. The Wald Test

The Wald test, rather than using the fact that the likelihoods should be close if the hypothesis is true, uses the fact that when we estimate the unrestricted model, if the restriction is true, then $g(\tilde{\theta})$ should be close to zero anyway. Redefine the information matrix as

$$I_A(\theta) = -E\left(\frac{1}{T} \frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}\right)$$

Notice the $1/T$ part. Then

$$W = Tg(\tilde{\theta})' \left[\frac{\partial g(\theta)}{\partial \theta'} \Big|_{\tilde{\theta}} I_A(\tilde{\theta})^{-1} \frac{\partial g(\theta)}{\partial \theta} \Big|_{\tilde{\theta}} \right]^{-1} g(\tilde{\theta}).$$

What does this look like when the restriction is $R\beta = r$ (i.e. when the restrictions are linear)? $g(\cdot) = (R_{j \times k} \quad 0_{j \times 1})\theta - r_{j \times 1}$. As a result $\partial g/\partial \theta = (R \quad 0)'$,

$$I_A(\tilde{\theta}) = \begin{pmatrix} \tilde{\sigma}^{-2} T^{-1} X'X & 0 \\ 0 & \frac{1}{2} \tilde{\sigma}^{-4} \end{pmatrix}$$

As a result

$$W = (R\tilde{\beta} - r)' [R(X'X)^{-1}R']^{-1} (R\tilde{\beta} - r) / \tilde{\sigma}^2.$$

Under some assumptions, W is $\chi^2(j)$ asymptotically, but it only relies on estimation of the model under the alternative (or unrestricted) model. Thus, it is an especially convenient test when estimation under the null is problematic.

3.4.3. The Lagrange Multiplier Test

Finally, we have a test which only requires estimates of the restricted model. It is based on the fact that if the restriction is valid the slope of the likelihood function shouldn't be much different at the restricted estimator than at the unrestricted estimator (where it is clearly zero). Let's set up the restricted estimation in Lagrangean form

$$\Lambda(\theta) = \mathcal{L}(\theta|y) - \lambda'_{j \times 1} g(\theta).$$

The first order condition for this problem is

$$\frac{\partial \Lambda(\theta)}{\partial \theta} \Big|_{\check{\theta}} = \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \Big|_{\check{\theta}} - \check{\lambda}' \frac{\partial g(\theta)}{\partial \theta} \Big|_{\check{\theta}} = 0.$$

The test is measured as

$$LM = \frac{1}{T} \check{\lambda}' \frac{\partial g(\theta)}{\partial \theta} \Big|_{\check{\theta}} I_A(\check{\theta})^{-1} \frac{\partial g(\theta)}{\partial \theta'} \Big|_{\check{\theta}} \check{\lambda}.$$

As you might expect, given our results for the other tests, LM is asymptotically $\chi^2(j)$. What is the form of the test when the restriction is $R\beta = r$? Since in this case σ^2 is unrestricted the partial of g (and thus of \mathcal{L}) with respect to σ^2 is 0. However, the partial of \mathcal{L} with respect to β is $(X'y - X'X\check{\beta})/\check{\sigma}^2 = X'\check{e}/\check{\sigma}^2$. These two form the leading term in the formula. Also

$$I_A(\check{\theta}) = \begin{pmatrix} \check{\sigma}^{-2} T^{-1} X'X & 0 \\ 0 & \frac{1}{2} \check{\sigma}^{-4} \end{pmatrix}$$

Therefore

$$\begin{aligned} LM &= \frac{1}{T} \begin{pmatrix} X'y/\check{\sigma}^2 & 0 \end{pmatrix} \begin{pmatrix} T\check{\sigma}^2(X'X)^{-1} & 0 \\ 0 & 2\check{\sigma}^4 \end{pmatrix} \begin{pmatrix} X'\check{e}/\check{\sigma}^2 \\ 0 \end{pmatrix} \\ &= \check{e}' X (X'X)^{-1} X' \check{e} / \check{\sigma}^2 \end{aligned}$$

Under certain conditions, which happen to be satisfied by any linear model, although the tests have the same distributional form in large samples, in small samples we have the following ordering

$$W \geq LR \geq LM.$$

Therefore, the Wald test will tend to reject more often than the likelihood ratio test, which will tend to reject more often than the Lagrange multiplier test.

4. Generalized Least Squares

We now change one of the assumptions of the linear model. In particular we change the assumption made about the covariance matrix of the error term. We now assume that $E(\epsilon\epsilon') = \sigma^2\Omega$ where Ω is a $T \times T$, positive definite, symmetric matrix.

4.1. Ω Known

First consider the case where Ω is known to the econometrician. Let's reconsider the OLS estimator, $\hat{\beta}_O = (X'X)^{-1}X'y$. It's still unbiased using the same proof. However, what about its variance?

$$\begin{aligned} V(\hat{\beta}_O) &= E[(\hat{\beta}_O - \beta)(\hat{\beta}_O - \beta)'] \\ &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] \\ &= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}. \end{aligned}$$

This doesn't look so good.

Since Ω is positive definite we can decompose it as $P\Omega P' = I$ where P is a $T \times T$ nonsingular matrix, or $\Omega = P^{-1}P'^{-1}$, or $\Omega^{-1} = P'P$. Since P is known, consider premultiplying the data by P so that we have

$$\begin{aligned} Py &= PX\beta + P\epsilon \\ y^* &= X^*\beta + \epsilon^* \end{aligned}$$

where $E(\epsilon^*\epsilon^{*'}) = E(P\epsilon\epsilon'P') = \sigma^2P\Omega P' = \sigma^2I$. Now if we use OLS on this transformed model we'll get $\hat{\beta}_G = (X^{*'}X^*)^{-1}X^{*'}y^*$. Clearly $\hat{\beta}_G$ is BLUE because the transformed model satisfies the conditions of the Gauss-Markov Theorem. It's clearly unbiased and its covariance matrix is clearly $\sigma^2(X^{*'}X^*)^{-1}$.

Now going back to untransformed variables we see that

$$\begin{aligned} \hat{\beta}_G &= (X^{*'}X^*)^{-1}X^{*'}y^* \\ &= (X'P'PX)^{-1}X'P'y \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \end{aligned}$$

and $V(\hat{\beta}_G) = \sigma^2(X^*X^*)^{-1} = \sigma^2(X'P'PX)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1}$. How about proving that this means GLS is more efficient than OLS.

$$\begin{aligned} V(\hat{\beta}_O) - V(\hat{\beta}_G) &= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} - \sigma^2(X'\Omega^{-1}X)^{-1} \\ &= \sigma^2[(X'X)^{-1}X' - (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}]\Omega \\ &\quad [(X'X)^{-1}X' - (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}]' \\ &= \sigma^2 A\Omega A' \end{aligned}$$

which is p.s.d. because Ω is p.d.

Good assignment question: Prove that $\hat{\sigma}_O^2$ is biased. What is the GLS estimator for the variance. Well you go back to the transformed model

$$\begin{aligned} \hat{\sigma}_G^2 &= e^{*'}e^*/(T - k) \\ &= (y^* - X^*\hat{\beta}_G)'(y^* - X^*\hat{\beta}_G)/(T - k) \\ &= (y'P' - \hat{\beta}_G'X'P')(Py - PX\hat{\beta}_G)/(T - k) \\ &= (y - X\hat{\beta}_G)'P'P(y - X\hat{\beta}_G)/(T - k) \\ &= e_G'\Omega^{-1}e_G/(T - k) \end{aligned}$$

Clearly, this estimator can be shown to be unbiased using the standard argument (the OLS proof) on the transformed model.

4.2. Ω Unknown

In general Ω could have how many unknown elements? It is a $T \times T$ matrix but it is symmetric so that there are actually only $T + (T - 1) + (T - 2) + \dots + 1 = T(T + 1)/2$ different elements of Ω . Actually, since we've normalized by pulling out σ^2 there is one less than that. You cannot model Ω freely when it is unknown because you only have T observations to work with and you've already used up k of those degrees of freedom in β . Therefore, you need assumptions as to the form of Ω . Then using those assumptions we'll construct estimators for Ω , say $\hat{\Omega}$.

The GLS estimator in this case will become

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

It is no longer possible to pass the expectations operator all the way through to y because now there is an additional random component due to $\hat{\Omega}$. At this point we must rely on asymptotic arguments.

Take for example the GLS estimator

$$\begin{aligned}\hat{\beta}_G &= (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y \\ &= (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}(X\beta + \epsilon) \\ &= \beta + (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}\epsilon \\ \hat{\beta}_G - \beta &= (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}\epsilon \\ \sqrt{T}(\hat{\beta}_G - \beta) &= \left(\frac{X'\hat{\Omega}^{-1}X}{T}\right)^{-1} \frac{1}{\sqrt{T}}X'\hat{\Omega}^{-1}\epsilon\end{aligned}$$

To show that this thing will be asymptotically normal you need

1. a law of large numbers so that $(X'\hat{\Omega}^{-1}X)/T$ converges in probability to some fixed positive definite matrix D . This is usually the same matrix which $(X'\Omega X)/T$ would converge to deterministically if Ω were known and
2. a central limit theorem so that $(X'\hat{\Omega}^{-1}\epsilon)/\sqrt{T}$ converges in distribution to the same random variable as $(X'\Omega^{-1}\epsilon)/\sqrt{T}$ would converge to, say a $N(0, \sigma^2 V)$, for some positive definite symmetric matrix V .

Generally we can get these two results if we obtain $\hat{\Omega}$ in such a way that it converges in probability to Ω .

4.3. Heteroskedasticity

Suppose that we have the same linear model, but $E(\epsilon_t^2) = \sigma_t^2$ where $\ln \sigma_t^2 = \ln \sigma^2 + z_t'\alpha$, but there is no covariance across ϵ 's. Typical procedure works like this

1. Estimate the model by OLS, and obtain the OLS residuals $e = y - X\hat{\beta}_O$.
2. Notice that $\ln e_t^2 + \ln \sigma_t^2 = \ln \sigma^2 + z_t'\alpha + \ln e_t^2$ or $\ln e_t^2 = \ln \sigma^2 + z_t'\alpha + \ln(e_t^2/\sigma_t^2) = \ln \sigma^2 + z_t'\alpha + v_t$. Now regress $\ln e^2$ on $[1 \ Z]$ and obtain OLS estimates $\hat{\alpha}$ and $\widehat{\ln \sigma^2}$.
3. Construct $\hat{\sigma}_t^2 = \exp(\widehat{\ln \sigma^2} + z_t'\hat{\alpha})$, and a matrix $\hat{\Omega}$ with the $\exp(z_t'\hat{\alpha})$'s on the diagonal.

4. Compute $\hat{\beta}_G$ and $\hat{\sigma}_G^2$ using $\hat{\Omega}$ in place of Ω in the formulas.

Generally, these estimators will have the desired properties.

4.4. Autocorrelation

The classic case you've all seen is where the error term is a first-order autoregression.

I.e. we have a model of the form

$$\begin{aligned} y_t &= x_t' \beta + u_t \\ u_t &= \rho u_{t-1} + \epsilon_t \quad -1 < \rho < 1 \end{aligned}$$

where x_t satisfies all the assumptions we've made up until now and ϵ_t is i.i.d. $N(0, \sigma_\epsilon^2)$. Since the model is now $y = X\beta + u$ we must of course be concerned with the properties of u . It is clearly mean zero. However what is $E(uu')$?

$$\begin{aligned} u_t &= \rho u_{t-1} + \epsilon_t \\ &= \rho^2 u_{t-2} + \epsilon_t + \rho \epsilon_{t-1} \quad \dots \\ &= \sum_{i=0}^{\infty} \rho^i \epsilon_{t-i} \end{aligned}$$

As a result of this we can compute the variances and covariances of the u_t 's.

$$\sigma^2 = V(u_t) = E(u_t^2) = \sum_{i=0}^{\infty} \rho^{2i} \sigma_\epsilon^2 = \sigma_\epsilon^2 / (1 - \rho^2)$$

Furthermore, we have $E(u_t u_{t-1}) = \rho \sigma^2$, etc. with $E(u_t u_{t-i}) = \rho^i \sigma^2$. Therefore,

$$\Omega = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{pmatrix}$$

To get a consistent estimator for Ω we just need a consistent estimate of one parameter: ρ . We could use the following steps.

1. Estimate the model by OLS and obtain the residuals.

2. Compute a consistent estimator for ρ . A bunch of possibilities are available

$$r = \frac{\sum_{t=2}^T e_t e_{t-1}}{\sum_{t=1}^T e_t^2} \quad \text{Theil's } r^* = \frac{T-k}{T-1} r \quad r^{**} = 1 - \frac{d}{2}$$

where d is the Durbin-Watson statistic, $d = \sum_{t=2}^T (e_t - e_{t-1})^2 / \sum_{t=1}^T e_t^2$.

3. Then in the last step there are various methods used

- a. Full GLS,
- b. Full GLS dropping the first observation, or
- c. Full ML, Beach and Mackinnon (1978).

The rationale for (3b) versus (3a) is that (3b) involves a simpler transformation of the data. It turns out that in this case

$$P = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 \\ -\rho & 1 & 0 & \dots & 0 \\ 0 & -\rho & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

As a result applying the transformation P to the data just means taking any variable z_t and replacing it with $z_t - \rho z_{t-1}$. However the first observation is treated differently.

To do full ML note that the likelihood of the u vector is a multivariate normal

$$f_U(u) = (2\pi\sigma^2)^{-T/2} |\Omega|^{-1/2} \exp\left(-\frac{1}{2\sigma^2} u' \Omega^{-1} u\right)$$

but since the Jacobian of the transformation between y and u is just I , we have

$$f_Y(y) = (2\pi\sigma^2)^{-T/2} |\Omega|^{-1/2} \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)' \Omega^{-1} (y - X\beta)\right)$$

The likelihood is maximized by choosing β , σ^2 and ρ to maximize f or \mathcal{L} . Similarly, one could do ML for any heteroscedastic model. However, ML tends to be computationally burdensome.

4.5. Testing for Heteroskedasticity

4.5.1. White's Test

Take the residuals from OLS (the restricted model), and square them to get e_t^2 . Regress these on all unique combinations in $x_t \otimes x_t$. If there are p regressors in all, then TR^2 from that regression will have a $\chi^2(p-1)$ distribution under the null of homoskedasticity. Effectively the hypotheses being compared here are $H_0 : \sigma_t^2 = \sigma^2, \forall t$, vs. $H_1 : \sigma_t^2 \neq \sigma_s^2$ for at least one $s \neq t$. Unfortunately (?), this test tends to pick up all kinds of specification error, and is not suggestive about the source of heteroscedasticity.

4.5.2. Goldfeld-Quandt (A Special Case)

Here the null is $H_0 : \sigma_t^2 = \sigma^2, \forall t$ while the alternative is $H_1 : \sigma_t^2 = \sigma_1^2, \forall t \in \tau, \sigma_t^2 = \sigma_2^2, \forall t \notin \tau$, where τ is some subsample. Here the idea is to estimate 2 separate OLS regressions on the two subsamples and compute separate estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. Under H_0 the ratio $[(T_1 - k)\hat{\sigma}_1^2/\sigma^2(T_1 - k)]/[(T_2 - k)\hat{\sigma}_2^2/\sigma^2(T_2 - k)] = \hat{\sigma}_1^2/\hat{\sigma}_2^2$ will be distributed $F(T_1 - k, T_2 - k)$. This would be tested using a 2-sided rejection region since F could be small or large under the alternative.

4.5.3. Breusch-Pagan

Here the model for the variance is $\sigma_t^2 = \exp(\alpha_1 + z_t' \alpha_2)$. The relevant hypotheses are $H_0 : \alpha_2 = 0$ versus $H_1 : \alpha_2 \neq 0$. Assuming that α_2 is a $s \times 1$ vector it turns out that if you regress $\ln(e_t^2/\tilde{\sigma}^2)$ on z_t then $(SST - SSE)/2$ from this regression is asymptotically $\chi^2(s)$. As it turns out so is TR^2 from regression of squared residuals on constant and z_t .

4.6. Testing for Autocorrelation

4.6.1. Durbin-Watson

The test statistic is given by

$$\begin{aligned} d &= \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2} \\ &= 2(1 - r) + (e_1^2 + e_T^2) / \sum_{t=1}^T e_t^2 \end{aligned}$$

Note that $0 < d < 4$. Near 0 suggests positive serial correlation. Near 4 suggests negative serial correlation. The distribution of d in any finite sample depends on the data matrix. However, upper and lower bounds for relevant critical values have been obtained for different sample sizes and numbers of regressors. SHAZAM can actually compute the critical value appropriate for a particular case numerically when you run your job. This test is not valid when the X data matrix is not fixed, especially with lagged dependent variables.

4.6.2. Box-Pierce

This is a test for general autocorrelation of any variety in the data. I.e. it tests a null of no autocorrelation versus an alternative of any autocorrelation. The test statistic is $Q = T \sum_{j=1}^L r_j^2$, where $r_j^2 = \sum_{t=j+1}^T e_t e_{t-j} / \sum_{t=1}^T e_t^2$. It turns out that in large samples $Q \sim \chi^2(L)$.

4.6.3. Ljung-Box

Just like the Box-Pierce test except $Q = T(T+2) \sum_{j=1}^L r_j^2 / (T-j)$.

You might not see any reason to choose between (2) and (3). These tests have different small sample properties. There is also the problem of how to choose L . It is an arbitrary choice. Obviously it cannot be chosen too large because with large L we are using fewer and fewer observations to estimate r_j^2 . On the other hand too small can miss

autocorrelation at more distant lags. The Ljung-Box appears to be closer to a $\chi^2(L)$ in small samples.

In general you do not need to use any of these tests. You can use any convenient LR, Wald, or LM test if you estimate by ML. One of the easiest ways to do this is to estimate under the null (restricted) model. This lets you use OLS to estimate β . Typically it is straightforward to derive an appropriate LM test which lets you do some easy regression involving the residuals, and then to take TR^2 as the test statistic.

5. Systems of Equations

5.1. Seemingly Unrelated Regression Equations

Suppose we have a set of equations rather than a single equation. For example, forgetting about simultaneous equations bias suppose you had a set of demand equations for several goods.

$$y_{it} = x'_{it}\beta_i + \epsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T$$

or

$$y_1 = X_1\beta_1 + \epsilon_1$$

$$y_2 = X_2\beta_2 + \epsilon_2 \quad \dots$$

$$y_n = X_n\beta_n + \epsilon_n,$$

where the y_i are $T \times 1$ vectors, each X_i is a $T \times k_i$ vector, each β_i is a $k_i \times 1$ vector and the ϵ_i are $T \times 1$. This can be rewritten as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & X_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

or

$$y = X\beta + \epsilon,$$

where y and ϵ are $nT \times 1$ vector, X is a $nT \times \sum_i k_i$ matrix, and β is a $\sum_i k_i \times 1$ vector.

Assumptions about the error terms: $E(\epsilon_{it}) = 0, \forall i, t$.

$$E(\epsilon_{it}\epsilon_{js}) = \begin{cases} \sigma_{ij} & \text{if } i \neq j, t = s \\ 0 & \text{otherwise} \end{cases}.$$

Let

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}$$

Then

$$\begin{aligned}
E(\epsilon\epsilon') &= \begin{pmatrix} \epsilon_1\epsilon'_1 & \epsilon_1\epsilon'_2 & \dots & \epsilon_1\epsilon'_n \\ \epsilon_2\epsilon'_1 & \epsilon_2\epsilon'_2 & \dots & \epsilon_2\epsilon'_n \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_n\epsilon'_1 & \epsilon_n\epsilon'_2 & \dots & \epsilon_n\epsilon'_n \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11}I_T & \sigma_{12}I_T & \dots & \sigma_{1n}I_T \\ \sigma_{21}I_T & \sigma_{22}I_T & \dots & \sigma_{2n}I_T \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}I_T & \sigma_{n2}I_T & \dots & \sigma_{nn}I_T \end{pmatrix} \\
&= \Sigma \otimes I_T = \Phi
\end{aligned}$$

If Σ is known, then Φ is known and we just have a huge GLS model. $\hat{\beta}_G = (X'\Phi^{-1}X)^{-1}X'\Phi^{-1}y$. If Σ is not known we do something like what we did in simple GLS. Estimate the model by OLS equation by equation. Get the n residual series e_j . Estimate $\hat{\sigma}_{ij} = e'_i e_j / T$. Then construct $\hat{\Sigma} = [\hat{\sigma}_{ij}]$ and $\hat{\Phi} = \hat{\Sigma} \otimes I_T$. Then construct $\hat{\beta}_G = (X'\hat{\Phi}^{-1}X)^{-1}X'\hat{\Phi}^{-1}y$. Why do we do it all simultaneously rather than equation by equation? Because, you can exploit the information in the error structure to help predict the errors in the other equations.

Some rules about Kronecker products, when $A_{n \times n}$, and $B_{T \times T}$.

1. $(A \otimes B)(C \otimes D) = AC \otimes BD$
2. $A \otimes (B \otimes C) = (A \otimes B) \otimes C$
3. $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$
4. $(B + C) \otimes A = (B \otimes A) + (C \otimes A)$
5. $(A \otimes B)' = A' \otimes B'$
6. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

When is OLS equation by equation the same as GLS.

1. When Σ is a diagonal matrix. Notice that in this case the matrix Φ is a block diagonal matrix. As a result

$$\Phi^{-1} = \begin{pmatrix} \sigma_{11}^{-1}I_T & 0 & \dots & 0 \\ 0 & \sigma_{22}^{-1}I_T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{nn}^{-1}I_T \end{pmatrix}$$

Therefore

$$X'\Phi^{-1}X = \begin{pmatrix} \sigma_{11}^{-1}X_1'X_1 & 0 & \dots & 0 \\ 0 & \sigma_{22}^{-1}X_2'X_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{nn}^{-1}X_n'X_n \end{pmatrix}$$

and

$$(X'\Phi^{-1}X)^{-1} = \begin{pmatrix} \sigma_{11}(X_1'X_1)^{-1} & 0 & \dots & 0 \\ 0 & \sigma_{22}(X_2'X_2)^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{nn}(X_n'X_n)^{-1} \end{pmatrix}$$

and

$$X'\Phi^{-1}y = \begin{pmatrix} \sigma_{11}^{-1}X_1'y_1 & 0 & \dots & 0 \\ 0 & \sigma_{22}^{-1}X_2'y_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{nn}^{-1}X_n'y_n \end{pmatrix}$$

so that

$$\hat{\beta}_G = \begin{pmatrix} (X_1'X_1)^{-1}X_1'y_1 \\ (X_2'X_2)^{-1}X_2'y_2 \\ \vdots \\ (X_n'X_n)^{-1}X_n'y_n \end{pmatrix}$$

2. $X_i = Z, \forall i$. I.e. the regressors are the same in every equation. $X = I_n \otimes Z$.

$$\begin{aligned} \hat{\beta}_G &= [(I_n \otimes Z)'(\Sigma \otimes I_T)^{-1}(I_n \otimes Z)]^{-1}(I_n \otimes Z)'(\Sigma \otimes I_T)^{-1}y \\ &= [(I_n \otimes Z)'(\Sigma^{-1} \otimes I_T)(I_n \otimes Z)]^{-1}(I_n \otimes Z)'(\Sigma^{-1} \otimes I_T)y \\ &= [(\Sigma^{-1} \otimes Z')(I_n \otimes Z)]^{-1}(\Sigma^{-1} \otimes Z')y \\ &= (\Sigma^{-1} \otimes Z'Z)^{-1}(\Sigma^{-1} \otimes Z')y \\ &= [\Sigma \otimes (Z'Z)^{-1}](\Sigma^{-1} \otimes Z')y \\ &= [I_n \otimes (Z'Z)^{-1}Z']y \end{aligned}$$

3. Kruskal's Theorem. Applies to standard GLS problems but this one also. If \exists non-singular $G_{k \times k}$ such that $\Omega X = XG$ then $\hat{\beta}_G = \hat{\beta}_O$. Notice that $\Omega XG^{-1} = X$.

$$\begin{aligned} \hat{\beta}_G &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \\ &= (G^{-1}'X'X)^{-1}G^{-1}'X'y \\ &= (X'X)^{-1}G'G^{-1}'X'y \\ &= (X'X)^{-1}X'y \end{aligned}$$

So in the context of SURE we're looking for $G_{nk \times nk}$ such that $(\Sigma \otimes I_T)X = XG$.

5.2. Panel Data

In this case we have observations which are sorted both by time t and by individual i . However, the coefficients do not vary across the individuals, except in very specific ways.

$$y_{it} = x'_{it}\beta + z'_i\gamma + \alpha_i + \epsilon_{it}.$$

The α_i are unobservable effects which vary by individual. Therefore, we can think of them as unobservable person specific regression coefficients. However, β and γ do not vary by individual.

We will assume that $E(\epsilon_{it}) = 0$, $E(\epsilon_{it}\epsilon_{js}) = 0$, for $i \neq j$ or $t \neq s$, $E(\epsilon_{it}^2) = \sigma^2$, $\forall i, t$.

There are two ways to approach the estimation

- a. Fixed Effects - treat the α_i as unknown constants, and put in dummy variables, d_i , which are 1 in observations involving person i and 0 otherwise.
- b. Random effects - assume that the differences in α across individuals are random, and make assumptions about the distribution they come from.

5.2.1. Fixed Effects

The model can be written in stacked form as

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{nT} \end{pmatrix} = \begin{pmatrix} x'_{11} \\ x'_{12} \\ \vdots \\ x'_{1T} \\ x'_{21} \\ \vdots \\ x'_{2T} \\ \vdots \\ x'_{nT} \end{pmatrix} \beta + \begin{pmatrix} z'_1 \\ z'_1 \\ \vdots \\ z'_1 \\ z'_2 \\ \vdots \\ z'_2 \\ \vdots \\ z'_n \end{pmatrix} \gamma + \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1T} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2T} \\ \vdots \\ \epsilon_{nT} \end{pmatrix}.$$

If we could estimate the whole thing in one big block it would be BLUE. Least squares involves minimizing $S = \sum_i \sum_t e_{it}^2 = \sum_i \sum_t (y_{it} - x'_{it}\beta - z'_i\gamma - \alpha_i)^2$. Differentiate

$$\frac{\partial S}{\partial \alpha_j} = -2 \sum_t (y_{jt} - x'_{jt}\beta - z'_j\gamma - \alpha_j) = 0.$$

This is easily solved for α_j .

$$\begin{aligned}\alpha_j &= \frac{1}{T} \left(\sum_t y_{jt} - \sum_t x'_{jt} \beta - \sum_t z'_j \gamma \right) \\ &= \bar{y}_j - \bar{x}'_j \beta - z'_j \gamma\end{aligned}$$

When I substitute this back into S I get $S = \sum_i \sum_t [y_{it} - \bar{y}_i - (x_{it} - \bar{x}_i)' \beta]^2$. As a result, γ is not identified because z_j is time invariant. Therefore, in fixed effects stuff you have to drop any variables which are individual specific and don't vary in the time dimension. Just do least squares after subtracting the means across individuals from y and x as suggested in the solution for α_j above.

A simple test for whether fixed effects exist would be to test whether all the α 's are the same. This is $n - 1$ restrictions. Just run the restricted regression and the unrestricted regression and do an F -test of the restrictions.

Summarizing fixed effects: Individual specific effects are constants-implies you can't estimate γ . Furthermore, the assumed lack of covariance across individuals means you can't exploit it to improve estimates of β .

5.2.2. Random Effects

In this case the α 's are random variables. The model is $y_{it} = x'_{it} \beta + z'_i \gamma + \alpha_i + \epsilon_{it} = \alpha + x'_{it} \beta + z'_i \gamma + u_{it}$, where $u_{it} = (\alpha_i - \alpha) + \epsilon_{it}$. The same assumptions are made about ϵ_{it} . In addition it is independent of α_i . The distribution of α_i is such that $E(\alpha_i) = \alpha$, $\forall i$, and $E(\alpha_i - \alpha)^2 = \sigma_\alpha^2$ and they are i.i.d. across individuals. Therefore, $E(u_{it}) = 0$, $E(u_{it}^2) = \sigma_\alpha^2 + \sigma^2$, $E(u_{it} u_{is}) = \sigma_\alpha^2$, 0 otherwise.

If we write the model as a whole we have $y = \alpha + X\beta + Z\gamma + u = W\theta + u$. Do the stacking by individual. $E(uu') = I_n \otimes \Sigma = \Omega$, where

$$\Sigma = \begin{pmatrix} \sigma_\alpha^2 + \sigma^2 & \sigma_\alpha^2 & \dots \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma^2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} = \sigma^2 I_T + \sigma_\alpha^2 11'$$

To estimate the model you just do a big GLS.

6. Nonlinear Models

What if the model we have is nonlinear of the form $y_t = f(x_t, \beta) + \epsilon_t$. We will have two ways of estimating models like this. The first method is called *nonlinear least squares*, while the other method is maximum likelihood.

6.1. Nonlinear Least Squares

The criterion function as before is the sum of squared residuals, $S(\beta) = \sum_t e_t^2 = \sum_t (y_t - f(x_t, \beta))^2 = [y - f(X, \beta)]' [y - f(X, \beta)]$. We will make the usual assumptions about X and ϵ . If we minimize the sum of squared residuals by choosing β the first order condition is

$$\frac{\partial S(\beta)}{\partial \beta} = -2 \frac{\partial f(X, \beta)'}{\partial \beta} [y - f(X, \beta)] = 0.$$

Do a first-order Taylor series expansion of $f(X, \beta)$ around $\beta = \beta_1$.

$$f(X, \beta) \approx f(X, \beta_1) + \frac{\partial f(X, \beta_1)}{\partial \beta} (\beta - \beta_1)$$

which implies that

$$\begin{aligned} y &\approx f(X, \beta_1) + \frac{\partial f(X, \beta_1)}{\partial \beta} (\beta - \beta_1) + \epsilon \\ y - f(X, \beta_1) + \frac{\partial f(X, \beta_1)}{\partial \beta} \beta_1 &\approx \frac{\partial f(X, \beta_1)}{\partial \beta} \beta + \epsilon \\ y^* &= \frac{\partial f(X, \beta_1)}{\partial \beta} \beta + \epsilon \end{aligned}$$

Notice that for any β_1 , $\frac{\partial f(X, \beta_1)}{\partial \beta}$ and y^* are data. Therefore starting from some arbitrary β_1 , create $\frac{\partial f(X, \beta_1)}{\partial \beta}$ and y^* . Then run a regression and get a new estimate

$$\begin{aligned} \beta_2 &= \left(\frac{\partial f(X, \beta_1)'}{\partial \beta} \frac{\partial f(X, \beta_1)}{\partial \beta} \right)^{-1} \frac{\partial f(X, \beta_1)'}{\partial \beta} y^* \\ &= \left(\frac{\partial f(X, \beta_1)'}{\partial \beta} \frac{\partial f(X, \beta_1)}{\partial \beta} \right)^{-1} \frac{\partial f(X, \beta_1)'}{\partial \beta} (y - f(X, \beta_1)) + \beta_1 \end{aligned}$$

In general you keep going, following the rule

$$\beta_{n+1} = \left(\frac{\partial f(X, \beta_n)'}{\partial \beta} \frac{\partial f(X, \beta_n)}{\partial \beta} \right)^{-1} \frac{\partial f(X, \beta_n)'}{\partial \beta} (y - f(X, \beta_n)) + \beta_n$$

Suppose that at some point $\beta_{n+1} = \beta_n$, i.e. the algorithm converges. What does that imply? It must be the case that

$$\frac{\partial f(X, \beta_n)'}{\partial \beta} (y - f(X, \beta_n)) = 0$$

which simply means the first-order condition for maximization is satisfied. Can we be sure it is a minimum? It can't be a maximum because we'll assume that $\frac{\partial f(X, \beta)}{\partial \beta}$ has full column rank which means $[\partial f' \partial f]^{-1}$ is positive definite. This means you are always going downhill. On the other hand, you can't show that it's not just a local minimum. Under regularity conditions the asymptotic variance covariance matrix of $\hat{\beta}_N$ is consistently estimated by

$$\hat{\sigma}^2 \left(\frac{\partial f(X, \hat{\beta})'}{\partial \beta} \frac{\partial f(X, \hat{\beta})}{\partial \beta} \right)^{-1}$$

where $\hat{\sigma}^2 = S(\hat{\beta}_N)/(T - k)$. What we have just described is the Gauss-Newton Algorithm.

Next we will examine the NEWTON-RAPHSON algorithm. Here we do a second-order Taylor series expansion of $S(\beta)$ around $\beta = \beta_1$.

$$S(\beta) \approx S(\beta_1) + \frac{\partial S(\beta_1)'}{\partial \beta} (\beta - \beta_1) + \frac{1}{2} (\beta - \beta_1)' \frac{\partial^2 S(\beta_1)}{\partial \beta \partial \beta'} (\beta - \beta_1).$$

Suppose we tried to minimize with respect to β

$$\frac{\partial S(\beta)}{\partial \beta} \approx \frac{\partial S(\beta_1)'}{\partial \beta} + \frac{\partial^2 S(\beta_1)}{\partial \beta \partial \beta'} (\beta - \beta_1) = 0$$

$$\beta - \beta_1 \approx - \left(\frac{\partial^2 S(\beta_1)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial S(\beta_1)'}{\partial \beta}$$

$$\beta_2 = \beta_1 - H(\beta_1)^{-1} \frac{\partial S(\beta_1)'}{\partial \beta}$$

or in general

$$\beta_{n+1} = \beta_n - H(\beta_n)^{-1} \frac{\partial S(\beta_n)'}{\partial \beta}.$$

If the algorithm ever converges what will be true? Clearly, $\frac{\partial S(\beta_n)}{\partial \beta} = 0$ which solves the first order condition. Is it a minimum? Depends on whether H is positive definite. If it is positive definite everywhere then you could guarantee it, but usually it's only locally a minimum because H is non-pd in places. There are some methods for always ensuring that at each step you go downhill. These are usually of the form

$$\beta_{n+1} = \beta_n - t_n H(\beta_n)^{-1} \frac{\partial S(\beta_n)}{\partial \beta}$$

where at each n , t_n is varied until $S(\beta_{n+1}) < S(\beta_n)$. Mention the often used 1, 0.5, 0.25, 0.125, ... method. The variance of $\hat{\beta}_N$ is consistently estimated by $2\hat{\sigma}^2 H(\hat{\beta}_N)^{-1}$ where $\hat{\sigma}^2 = S(\hat{\beta}_N)/(T - k)$.

What is the relationship between Gauss-Newton and Newton-Raphson? Notice that

$$\begin{aligned} H(\beta) &= \frac{\partial^2 S}{\partial \beta \partial \beta'} = \frac{\partial}{\partial \beta'} \left(-2 \sum_{t=1}^T [y_t - f(x_t, \beta)] \frac{\partial f(x_t, \beta)}{\partial \beta} \right) \\ &= 2 \sum_{t=1}^T \frac{\partial f(x_t, \beta)}{\partial \beta} \frac{\partial f(x_t, \beta)'}{\partial \beta} - 2 \sum_{t=1}^T [y_t - f(x_t, \beta)] \frac{\partial^2 f(x_t, \beta)}{\partial \beta \partial \beta'} \end{aligned}$$

Therefore,

$$E[H(\beta)] = 2 \frac{\partial f(X, \beta)'}{\partial \beta} \frac{\partial f(X, \beta)}{\partial \beta},$$

which shows you that the methods are really quite similar. In general there are a multitude of methods of the form

$$\beta_{n+1} = \beta_n - t_n P_n \frac{\partial S(\beta_n)}{\partial \beta}.$$

Note that even the Gauss-Newton method fits into this form. See Judge Appendix B for a description of all the methods.

6.2. Maximum Likelihood

Since the model is of the form $y = f(X, \beta) + \epsilon$ we can simply take the likelihood for the error terms and replace them by $y - f()$. Therefore,

$$\mathcal{L} = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} S(\beta).$$

As usual we'll get $\tilde{\sigma}^2 = S(\tilde{\beta})/T$, but the formula for $\tilde{\beta}$ is not a closed form. Plug the solution for $\tilde{\sigma}^2$ into \mathcal{L} and reoptimize.

$$\mathcal{L}^* = -\frac{T}{2} \ln 2\pi - \frac{T}{2}(1 + \ln T^{-1}) - \frac{T}{2} \ln S(\beta).$$

What's clear is that maximizing the likelihood is equivalent to minimizing the sum of squares. There are three methods for finding the MLE for β . Going back to the original likelihood

6.2.1. Newton-Raphson

In this case you really just do the same thing as you do in NLLS. Notice that

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta} &= -\frac{1}{2\sigma^2} \frac{\partial S(\beta)}{\partial \beta} \\ \frac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta'} &= -\frac{1}{2\sigma^2} \frac{\partial^2 S(\beta)}{\partial \beta \partial \beta'} \end{aligned}$$

Therefore,

$$\begin{aligned} \beta_{n+1} &= \beta_n - \left(\frac{\partial^2 \mathcal{L}(\beta_n)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial \mathcal{L}(\beta_n)}{\partial \beta} \\ &= \beta_n - \left(\frac{\partial^2 S(\beta_n)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial S(\beta_n)}{\partial \beta} \end{aligned}$$

Here the estimator for the covariance matrix is given by

$$2\tilde{\sigma}^2 \left(\frac{\partial^2 S(\tilde{\beta})}{\partial \beta \partial \beta'} \right)^{-1}$$

6.2.2. Method of Scoring

This method uses the inverse of the expectation of the Hessian instead of the inverse of the Hessian to weight the first derivatives.

$$\begin{aligned} \beta_{n+1} &= \beta_n - \left(E \frac{\partial^2 \mathcal{L}(\beta_n)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial \mathcal{L}(\beta_n)}{\partial \beta} \\ &= \beta_n - \left(-E \frac{1}{2\sigma^2} \frac{\partial^2 S(\beta_n)}{\partial \beta \partial \beta'} \right)^{-1} \left(-\frac{1}{2\sigma^2} \right) \frac{\partial S(\beta_n)}{\partial \beta} \\ &= \beta_n - \frac{1}{2} \left(\frac{\partial f(X, \beta_n)'}{\partial \beta} \frac{\partial f(X, \beta_n)}{\partial \beta} \right) \frac{\partial S(\beta_n)}{\partial \beta} \end{aligned}$$

Here the estimator for the covariance matrix is given by

$$\tilde{\sigma}^2 \left(\frac{\partial f(x, \tilde{\beta})'}{\partial \beta} \frac{\partial f(x, \tilde{\beta})}{\partial \beta} \right)^{-1}$$

6.2.3. Berndt, Hall, Hall and Hausman (BHHH)

This is a method with a seemingly arbitrary choice of weighting matrix but it does relate to the method of scoring. Define

$$l_t = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \left(\frac{[y_t - f(x_t, \beta)]^2}{2\sigma^2} \right)$$

so that $\sum_t l_t = \mathcal{L}$. Here

$$\begin{aligned} \beta_{n+1} &= \beta_n - \left(-\sum_t \frac{\partial l_t}{\partial \beta} \frac{\partial l_t}{\partial \beta'} \right)_{\beta_n, \sigma_n}^{-1} \frac{\partial \mathcal{L}(\beta_n)}{\partial \beta} \\ &= \beta_n - \left(\sum_t \frac{[y_t - f(x_t, \beta)]^2}{\sigma^4} \frac{\partial f(x, \beta)}{\partial \beta} \frac{\partial f(x, \beta)'}{\partial \beta} \right)_{\beta_n, \sigma_n}^{-1} \left(-\frac{1}{2\sigma^2} \frac{\partial S(\beta_n)}{\partial \beta} \right) \\ &= \beta_n - \frac{\sigma_n^2}{2} \left(\sum_t \frac{[y_t - f(x_t, \beta)]^2}{\sigma^4} \frac{\partial f(x, \beta)}{\partial \beta} \frac{\partial f(x, \beta)'}{\partial \beta} \right)_{\beta_n, \sigma_n}^{-1} \frac{\partial S(\beta_n)}{\partial \beta} \end{aligned}$$

Notice that

$$E \left(\sum_t \frac{\partial l_t}{\partial \beta} \frac{\partial l_t}{\partial \beta'} \right) = \frac{1}{\sigma^2} \sum_t \left(\frac{\partial f(x, \beta)}{\partial \beta} \frac{\partial f(x, \beta)'}{\partial \beta} \right)$$

which looks just like method of scoring. Therefore, the estimate of the variance-covariance matrix is just

$$\tilde{\sigma}^4 \left(\sum_t \frac{[y_t - f(x_t, \beta)]^2}{\sigma^4} \frac{\partial f(x, \beta)}{\partial \beta} \frac{\partial f(x, \beta)'}{\partial \beta} \right)_{\tilde{\beta}}^{-1}$$

7. Stochastic Regressors

7.1. Independent Regressors

We now change one of the previously vital assumptions. Up until now we have been assuming that the matrix X is fixed. Now we will instead assume that it is stochastic, but also that X is independent of ϵ . This implies that any function of X is independent of any function of ϵ . In the plain linear model we have $y = X\beta + \epsilon$. Therefore, the OLS estimator is $\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\epsilon$. Therefore, $E(\hat{\beta}) = \beta + E[(X'X)^{-1}X']E(\epsilon) = \beta$. Furthermore, we have

$$\begin{aligned}
 E(e'e) &= E(\epsilon'M\epsilon) \\
 &= E[\text{tr}(\epsilon'M\epsilon)] \\
 &= E[\text{tr}(M\epsilon\epsilon')] \\
 &= \text{tr}[E(M)E(\epsilon\epsilon')] \\
 &= \sigma^2\text{tr}(EM) = \sigma^2 E\text{tr}(M) \\
 &= \sigma^2(T - k).
 \end{aligned}$$

Therefore, $E(\hat{\sigma}^2) = \sigma^2$. And

$$\begin{aligned}
 V(\hat{\beta}) &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] \\
 &= \sigma^2 E(X'X)^{-1}.
 \end{aligned}$$

One can approximate this by $\hat{\sigma}^2(X'X)^{-1}$ which is also unbiased.

7.2. Nonindependent Regressors (of a particular form)

In this case we imagine a world where $y = X\beta + \epsilon$ but our previous assumptions don't hold. Now we only assume that ϵ_t is i.i.d. mean zero, variance σ^2 with $\text{plim } \epsilon'\epsilon/T = \sigma^2$, $\text{plim } X'X/T = \Sigma_{xx}$ and $\text{plim } X'\epsilon/T = 0$. Under these circumstances

$$\begin{aligned}\text{plim } \hat{\beta} &= \text{plim}(X'X)^{-1}X'y \\ &= \beta + \text{plim}\left(\frac{X'X}{T}\right)^{-1} \frac{X'\epsilon}{T} \\ &= \beta + \text{plim}\left(\frac{X'X}{T}\right)^{-1} \text{plim} \frac{X'\epsilon}{T} \\ &= \beta + \Sigma_{xx}^{-1}0 = \beta.\end{aligned}$$

You can check the textbook for proofs that the estimates of the variances converge in probability to the right objects. One last thing is the asymptotic distribution. We have

$$\sqrt{T}(\hat{\beta} - \beta) = \left(\frac{X'X}{T}\right)^{-1} \frac{X'\epsilon}{\sqrt{T}}.$$

Therefore, by the quick and dirty asymptotic method we have an estimator which will be asymptotically normal and the vcv matrix will be $\sigma^2\Sigma_{xx}^{-1}$.

7.3. Instrumental Variables

7.3.1. Errors in Variables

Let us discuss all this in a univariate setting. Suppose the model is $y_t = x_t\beta + \epsilon_t$. However, you do not observe x_t but a measurement error ridden version of it called $z_t = x_t + u_t$. Therefore, we can write $y_t = z_t\beta + \epsilon_t - u_t\beta = z_t\beta + v_t$. We'll assume the ϵ 's and the u 's are i.i.d. and independent of each other and x_t . If the econometrician regresses y_t on z_t he gets

$$\begin{aligned}\hat{\beta} &= \frac{\sum z_t y_t}{\sum z_t^2} \\ &= \frac{\sum z_t (z_t \beta + v_t)}{\sum z_t^2} \\ &= \beta + \frac{\sum z_t v_t}{\sum z_t^2} \\ \hat{\beta} - \beta &= \frac{\frac{1}{T} \sum z_t v_t}{\frac{1}{T} \sum z_t^2}\end{aligned}$$

Now $\frac{1}{T} \sum z_t v_t = \frac{1}{T} \sum (x_t + u_t)(\epsilon_t - u_t \beta) \xrightarrow{p} -\beta \sigma_u^2$. Also, $\frac{1}{T} \sum z_t^2 = \frac{1}{T} \sum (x_t + u_t)^2 \xrightarrow{p} \sigma_x^2 + \sigma_u^2$. Therefore, $\hat{\beta} - \beta \xrightarrow{p} -\sigma_u^2 \beta / (\sigma_x^2 + \sigma_u^2)$. This means that $\hat{\beta}$ is not consistent except when $\beta = 0$. As the signal to noise ratio σ_x^2 / σ_u^2 ratio gets small $\text{plim } \hat{\beta} \rightarrow 0$, while as it gets large $\text{plim } \hat{\beta} \rightarrow \beta$. The basic problem arises because the regressor z_t is not orthogonal to the error term v_t .

7.3.2. Instrumental Variables

Now we will consider general models of the form $y_t = x_t \beta + \epsilon_t$ where $\text{plim } \frac{1}{T} \sum x_t \epsilon_t = \sigma_{x\epsilon} \neq 0$. This will imply that OLS is inconsistent.

What is *instrumental variables* estimation? The trick is to find some variable z_t s.t.

- (1) $\frac{1}{T} \sum z_t \epsilon_t \xrightarrow{p} 0$.
- (2) $\frac{1}{T} \sum z_t x_t \xrightarrow{p} \sigma_{zx} \neq 0$.

The IV estimator is generated by the following formula

$$\hat{\beta}_{IV} = \frac{\sum z_t y_t}{\sum z_t x_t}.$$

As a result, if we substitute in the definition of y_t we get

$$\begin{aligned} \hat{\beta}_{IV} &= \frac{\sum z_t (x_t \beta + \epsilon_t)}{\sum z_t x_t} \\ &= \beta + \frac{\sum z_t \epsilon_t}{\sum z_t x_t} \\ \hat{\beta}_{IV} - \beta &= \frac{\frac{1}{T} \sum z_t \epsilon_t}{\frac{1}{T} \sum z_t x_t} \\ &\xrightarrow{p} 0. \end{aligned}$$

Also as far as distribution theory is concerned we'll get

$$\sqrt{T}(\hat{\beta}_{IV} - \beta) = \frac{\frac{1}{\sqrt{T}} \sum z_t \epsilon_t}{\frac{1}{T} \sum z_t x_t}.$$

We can for the purposes of this course assert that $\frac{1}{\sqrt{T}} \sum z_t \epsilon_t \xrightarrow{d} N(0, \sigma_\epsilon^2 \sigma_z^2)$ while $\frac{1}{T} \sum z_t x_t \xrightarrow{p} \sigma_{xz}$. Therefore,

$$\sqrt{T}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N(0, \sigma_\epsilon^2 \sigma_{xz}^{-1} \sigma_z^2 \sigma_{xz}^{-1}).$$

Notice that the variance is equal to

$$\begin{aligned}\frac{\sigma_\epsilon^2 \sigma_z^2}{\sigma_{xz}^2} &= \frac{\sigma_\epsilon^2}{\sigma_x^2} \frac{\sigma_x^2 \sigma_z^2}{\sigma_{xz}^2} \\ &= V_{OLS} / \rho_{xz}^2\end{aligned}$$

Obviously, OLS is just IV with X as the instrument. We want a Z which is as highly correlated with X as possible without being correlated with ϵ .

In a multivariate setting the story is really no different. We have the model $y = X\beta + \epsilon$. We define $\Sigma_{xx} = \text{plim} \frac{1}{T} X'X$ and $\Sigma_{x\epsilon} = \text{plim} \frac{1}{T} X'\epsilon \neq 0$. Suppose we have the replacements for the X matrix called $Z_{T \times k}$. We'll define $\Sigma_{zx} = \text{plim} \frac{1}{T} Z'X$ and $\Sigma_{zz} = \text{plim} \frac{1}{T} Z'Z$ which are assumed to have full rank. Furthermore, $\Sigma_{z\epsilon} = \text{plim} \frac{1}{T} Z'\epsilon = 0$. Also assume enough regularity conditions such that $\frac{1}{\sqrt{T}} Z'\epsilon \xrightarrow{d} N(0, \sigma^2 \Sigma_{zz})$. Then

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

Clearly $\text{plim} \hat{\beta}_{IV} - \beta = 0$. Furthermore,

$$\sqrt{T}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N(0, \sigma^2 \Sigma_{zx}^{-1} \Sigma_{zz} \Sigma_{xz}^{-1})$$

One natural question that arises is how to choose the instrument which replaces each column in the matrix X . Suppose I have l variables (called W 's) which I can reasonably assume to be uncorrelated with the ϵ 's and which are correlated with the X 's. If $l > k$, then clearly we have more candidates than we have room for. How can we reduce the number from l to k in the optimal way. One way to think about this is to think about the k linear combinations of the l variables which minimizes the size of the variance covariance matrix of $\hat{\beta}_{IV}$. I.e. the goal is to construct a matrix $Z_{T \times k} = W_{T \times l} A_{l \times k}$ in such way as to minimize the variance covariance matrix. Clearly this requires choosing

A optimally. Now it's clear that $\Sigma_{zz} = \text{plim } \frac{1}{T} Z'Z = \text{plim } \frac{1}{T} A'W'WA = A'\Sigma_{ww}A$. Also, $\Sigma_{zx} = \text{plim } \frac{1}{T} Z'X = \text{plim } \frac{1}{T} A'W'X = A'\Sigma_{wx}$. Therefore, the variance of the IV estimator is

$$V = \sigma^2(A'\Sigma_{wx})^{-1}A'\Sigma_{ww}A(\Sigma_{xw}A)^{-1}.$$

which can be minimized by choosing $A^* = \Sigma_{ww}^{-1}\Sigma_{wx}$ so that the variance is

$$V^* = \sigma^2(\Sigma_{xw}\Sigma_{ww}^{-1}\Sigma_{wx})^{-1}.$$

To prove that this is smallest in a cheesy way, consider the difference between its inverse and the inverse of any other matrix for any other choice of A .

$$\begin{aligned} V^{*-1} - V^{-1} &\propto \Sigma_{xw}\Sigma_{ww}^{-1}\Sigma_{wx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx} \\ &= X'[W(W'W)^{-1}W' - Z(Z'Z)^{-1}Z']X \end{aligned}$$

which is positive semidefinite as the matrix in the middle is idempotent. To check this notice that

$$\begin{aligned} [W(W'W)^{-1}W' - Z(Z'Z)^{-1}Z']^2 &= W(W'W)^{-1}W' - W(W'W)^{-1}W'Z(Z'Z)^{-1}Z' \\ &\quad - Z(Z'Z)^{-1}Z'W(W'W)^{-1}W' + Z(Z'Z)^{-1}Z' \\ &= W(W'W)^{-1}W' - WA(Z'Z)^{-1}Z' - Z(Z'Z)^{-1}A'W' + Z(Z'Z)^{-1}Z' \\ &= W(W'W)^{-1}W' - Z(Z'Z)^{-1}Z'. \end{aligned}$$

8. Time Series

We'll be looking at univariate time series. We're looking at a sequence $\{y_t\}_{t=1}^T$. We're going to look at autoregressive moving average (ARMA) processes. These are processes which allow us to characterize the distribution of y_t in terms of its own past. First we need to define a certain type of *stationarity*.

A stochastic process is *stationary* if

$$f(y_t, y_{t-1}, \dots, y_{t-k}) = f(y_{t+s}, y_{t+s-1}, \dots, y_{t+s-k})$$

for all s, k .

A stochastic process $\{y_t\}_{t=1}^T$ is *covariance stationary* if $E(y_t) = \mu$ and $\text{Var}(y_t) = \gamma_0 < \infty$ and $E[(y_t - \mu)(y_{t-s} - \mu)] = \gamma_s, \forall t$.

Let's look at different types of time series processes.

8.1. Time Series Processes

8.1.1. White Noise

We will generically represent a white noise process with the symbol (ϵ_t) . It has the following properties

$$E(\epsilon_t) = 0$$

$$\gamma_s = \begin{cases} \sigma^2 & \text{if } s = 0 \\ 0 & \text{if } s \neq 0 \end{cases}$$

8.1.2. Autoregressive 1st Order - AR(1)

$$\begin{aligned} y_t &= \phi y_{t-1} + \epsilon_t \\ &= \phi(\phi y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \phi^2 y_{t-2} + \epsilon_t + \phi \epsilon_{t-1} \\ &= \phi^T y_{t-T} + \sum_{j=0}^{T-1} \phi^j \epsilon_{t-j} \end{aligned}$$

If y_{t-T} were fixed then $E(y_t) = \phi^T y_{t-T}$ which implies that if $\phi \neq 1$ then $E(y_t) \neq E(y_s)$ for $t \neq s$. Suppose then that y_{t-T} is not fixed for any T . If $|\phi| < 1$ we can write

$$y_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$$

and thus $Ey_t = Ey_s = 0, \forall s, t$.

$$\begin{aligned} \text{Var}(y_t) &= E\left[\sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}\right]^2 \\ &= E\left[\sum_{j=0}^{\infty} \phi^{2j} \epsilon_{t-j}^2\right] \\ &= (1 + \phi^2 + \phi^4 + \dots)\sigma^2 \\ &= \sigma^2/(1 - \phi^2) = \gamma_0 \end{aligned}$$

$$\begin{aligned} \gamma_s &= E(y_t y_{t-s}) \\ &= E[(\phi^s y_{t-s} + \epsilon_t + \phi \epsilon_{t-1} + \phi^{s-1} \epsilon_{t-s+1}) y_{t-s}] \\ &= \phi^s \gamma_0 \end{aligned}$$

We're going to use *lag operator* notation. $Ly_t = y_{t-1}$. $L^i y_t = y_{t-i}$.

$$\begin{aligned} y_t &= \phi Ly_t + \epsilon_t \\ (1 - \phi L)y_t &= \epsilon_t \\ y_t &= (1 - \phi L)^{-1} \epsilon_t \\ y_t &= (1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \dots) \epsilon_t \end{aligned}$$

The polynomial $(1 - \phi L)^{-1}$ exists iff $\forall |Z| \leq 1, (1 - \phi Z) \neq 0$. All the roots of the polynomial lie outside the unit circle. Furthermore, an AR(1) process is covariance stationary iff $\forall |Z| \leq 1, 1 - \phi Z \neq 0$.

8.1.3. Autoregressive p th Order - AR(p)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t$$

$$\phi(L)y_t = \epsilon_t$$

The polynomial $\phi(L)^{-1}$ exists iff $\forall |Z| \leq 1, \phi(Z) \neq 0$. All the roots of the polynomial lie outside the unit circle. Furthermore, an AR(p) process is covariance stationary iff $\forall |Z| \leq 1, \phi(Z) \neq 0$.

8.1.4. Moving Average 1st Order - MA(1)

$$y_t = \epsilon_t + \theta \epsilon_{t-1}$$

$$E y_t = 0$$

$$\begin{aligned} \text{Var}(y_t) &= E y_t^2 = E(\epsilon_t^2 + 2\theta \epsilon_t \epsilon_{t-1} + \theta^2 \epsilon_{t-1}^2) \\ &= \sigma^2(1 + \theta^2) = \gamma_0 \end{aligned}$$

$$\begin{aligned} \gamma_1 &= \epsilon_t \epsilon_{t-1} + \theta \epsilon_{t-1}^2 + \theta^2 \epsilon_{t-1} \epsilon_{t-2} \\ &= \theta \sigma^2 \end{aligned}$$

$$\gamma_s = 0, \quad s > 1$$

An MA(1) process is always covariance stationary. On the other hand the value of θ will matter for *invertibility*. We want to write $(1 + \theta L)^{-1} y_t = \theta(L) y_t = \epsilon_t$. This MA process is invertible iff $\forall |Z| \leq 1, 1 + \theta Z \neq 0$.

8.1.5. Moving Average q th Order - MA(q)

$$\begin{aligned} y_t &= \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q} \\ &= \theta(L) \epsilon_t \end{aligned}$$

$$Ey_t = 0$$

$$\text{Var}(y_t) = Ey_t^2 = \sigma^2(1 + \theta_1^2 + \dots + \theta_q^2) = \gamma_0$$

$$\gamma_i = 0, \quad i > q$$

An MA(q) process is always covariance stationary. We want to write $\theta(L)y_t = \epsilon_t$.

This MA process is invertible iff $\forall |Z| \leq 1, \theta(Z) \neq 0$.

8.1.6. ARMA(p, q)

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$$

$$\phi(L)y_t = \theta(L)\epsilon_t$$

$$\text{STATIONARY :} \quad \forall |Z| \leq 1 \quad \phi(Z) \neq 0 \quad y_t = \phi(L)^{-1} \theta(L) \epsilon_t$$

$$\text{INVERTIBLE :} \quad \forall |Z| \leq 1 \quad \theta(Z) \neq 0 \quad \epsilon_t = \theta(L)^{-1} \phi(L) y_t$$

8.2. Wold Decomposition Theorem

Any covariance stationary stochastic process $\{x_t\}$ can be represented as $x_t = y_t + z_t$ where z_t is linearly deterministic [i.e. $\text{Var}(z_t | z_{t-1}, \dots) = 0$] and y_t is purely non-deterministic with an MA(∞) representation $y_t = \psi(L)\epsilon_t$ where $\sum_{j=1}^{\infty} \psi_j^2 < \infty$.

ARMA models represent approximations to the MA(∞). I.e. $\psi(L) \approx \phi(L)^{-1} \theta(L)$.

8.3. Box-Jenkins Procedure

The procedure consists of three basic parts

- (1) Identification (Model Selection) – make a judgment based on autocorrelations and partial autocorrelations that leads to a model that could have generated them,

- (2) Estimation – estimate model chosen in (1),
- (3) Diagnostic Checking – do various diagnostic checks to determine whether the residuals resemble white noise.

8.3.1. Model Selection

We'll select models based on the autocorrelations and partial autocorrelations of a time series. We first need to define what these objects are.

The *autocorrelation function* (of i) is defined as $\rho_i = \gamma_i/\gamma_0$. Notice that $|\rho_i| \leq 1$, $\rho_{-i} = \rho_i$ and $\rho_0 = 1$.

$$\begin{aligned} \text{WN :} \quad \rho_0 &= 1 \\ \rho_i &= 0 \quad i \neq 0 \end{aligned}$$

$$\begin{aligned} \text{AR}(1) : \quad \rho_0 &= 1 \\ \rho_i &= \phi^i \end{aligned}$$

For $\text{AR}(p)$ there is some more complex form of decay.

$$\begin{aligned} \text{MA}(1) : \quad \rho_0 &= 1 \\ \rho_1 &= \theta/(1 + \theta^2) \\ \rho_i &= 0 \quad i > 1 \end{aligned}$$

$$\text{MA}(q) : \quad \rho_i = 0 \quad i > q$$

So if you thought your data followed an $\text{MA}(q)$ process you would look for a drop off in the autocorrelation function at some lag in order to identify q . How can you tell whether an autocorrelation is 0 or not. Do a statistical test! If y_t is mean zero then the i th autocovariance is $\hat{\gamma}_i = \frac{1}{T-i} \sum_{t=i+1}^T y_t y_{t-i}$. You would estimate $\hat{\rho}_i = \hat{\gamma}_i/\hat{\gamma}_0$. To test the null hypothesis $H_0 : \rho_i = 0$ for some $i > 0$ it turns out that under the null hypothesis $\sqrt{T}\hat{\rho}_i \xrightarrow{d} N(0, 1)$. This is because the numerator converges in distribution to $N(0, \sigma_y^4)$ and the denominator converges in probability to σ_y^2 . Thus an easy way to check whether the autocorrelations are significant is to construct an acceptance region based on the 5% significance level around 0, equal to $\pm 1.96/\sqrt{T}$.

Suppose you want to do a joint test $H_0 : \rho_i = 0, i = 1, \dots, N$. Then construct the Box-Pierce statistic $Q_{BP} = T \sum_{i=1}^N \hat{\rho}_i^2$ or the Ljung-Box $Q_{LB} = T(T-2) \sum_{i=1}^N \hat{\rho}_i^2 / (T-i)$. Both of these are χ_N^2 asymptotically although the Ljung-Box is closer to a χ^2 is small samples. It's because both are sums of squared standard normals which are independent asymptotically.

Obviously these methods are more helpful for MA processes than AR processes.

The *partial autocorrelation* function ϕ_{kk} is defined by the *Yule-Walker* equations

$$\rho_j = \sum_{i=1}^k \phi_{ki} \rho_{i-j}, \quad j = 1, \dots, k.$$

Notice that we have k equations and there are k unknowns $\phi_{ki}, i = 1, \dots, k$. This can be written in matrix notation

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \cdots & \cdots & \cdots & 1 \end{pmatrix} \begin{pmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{pmatrix}$$

or $\bar{\rho}_k = \bar{P}_k \bar{\phi}_k$ or $\bar{\phi}_k = \bar{P}_k^{-1} \bar{\rho}_k$, but remember we're really only interested in ϕ_{kk} the last element. Notice that for $k = 1$, this just reduces to $\rho_1 = \phi_{11}$, so that the first partial autocorrelation is the same as the first autocorrelation. Might want to illustrate $k = 2$ also.

Of course we'd be doing this in sample, so all the quantities would be sample estimates. It looks pretty formidable but it isn't really. In fact, these things can be computed from simple OLS regressions. Start with ϕ_{11} . Since it equals ρ_1 it is clear that $\hat{\phi}_{11} = \hat{\rho}_1 = \hat{\phi}_1$ the OLS estimator in a regression of y_t on y_{t-1} . I.e. $\hat{\phi}_1 = (\sum y_{t-1} y_t / T) / (\sum y_{t-1}^2 / T) \xrightarrow{p} \gamma_1 / \gamma_0 = \rho_1 = \phi_{11}$. Suppose we estimated an AR(2) model $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$. Notice we'd have

$$\begin{aligned} \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix} &= \begin{pmatrix} \sum y_{t-1}^2 / T & \sum y_{t-1} y_{t-2} / T \\ \sum y_{t-1} y_{t-2} / T & \sum y_{t-2}^2 / T \end{pmatrix}^{-1} \begin{pmatrix} \sum y_{t-1} y_t / T \\ \sum y_{t-2} y_t \end{pmatrix} \\ &\xrightarrow{p} \begin{pmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \\ &= \begin{pmatrix} \rho_0 & \rho_1 \\ \rho_1 & \rho_0 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} \end{aligned}$$

Therefore $\hat{\phi}_2$ acts as an estimate of ϕ_{22} . In general, $\hat{\phi}_k$ in an OLS of an AR(k) acts as an estimate of ϕ_{kk} . In other words, $\phi_{kk} = \text{Corr}(y_t, y_{t-k} | y_{t-1}, \dots, y_{t-k+1})$. An implication of all of this is that $\phi_{kk} = 0$ for $k > p$ if y_t is an AR(p) process. On the other hand the ϕ_{kk} will decay in some manner if y_t is an MA process.

Another small point is that when we're selecting a model we may need to difference the time series to render it stationary. Most macroeconomic time series look like this because they grow steadily through time. There is an issue as to whether this growth can be modelled as a deterministic trend or must involve explicitly modelling the first difference of the log of a series. We will not discuss this issue. Let me just mention a definition.

The stochastic process $\{y_t\}$ is *integrated* of order d iff $\Delta^d y_t = (1 - L)^d y_t$ is stationary. This definition leads us to refer to ARIMA(p, d, q) processes which have the form $\phi(L)\Delta^d y_t = \theta(L)\epsilon_t$. Two special ARIMA processes are the (misnamed by economists) *random walk* $\Delta y_t = \epsilon_t$ and the *random walk with drift* $\Delta y_t = \mu + \epsilon_t$. Use of these models is especially prevalent in finance.

8.3.2. Estimation

We will consider maximum likelihood estimation. This requires the assumption that $\epsilon_t \sim N(0, \sigma^2)$. The estimation of autoregressive models is simple. To estimate them we need to write down the joint density function of the y 's, $L(y) = f(y_1, y_2, \dots, y_T)$. Let me introduce the notation $Y_t = (y_1, y_2, \dots, y_t)$. We can use the definition of conditional density functions to derive f .

$$\begin{aligned} f(Y_T) &= f(y_T | Y_{T-1}) f(Y_{T-1}) \\ &= f(y_T | Y_{T-1}) f(y_{T-1} | Y_{T-2}) f(Y_{T-2}) \\ &= \left(\prod_{t=2}^T f(y_t | Y_{t-1}) \right) f(y_1) \end{aligned}$$

Since each of these conditional distributions is easy to characterize in an AR model, this is a convenient way to proceed.

Take for example the AR(1) model. What is $f(y_t|Y_{t-1})$? Well, in the AR(1) model $y_t = \phi y_{t-1} + \epsilon_t$. Given the information on all y 's up to time $t - 1$ it is clear that $E_{t-1}y_t = \phi y_{t-1}$ and that $\text{Var}_{t-1}y_t = \sigma^2$. Therefore, $(y_t|Y_{t-1}) \sim N(\phi y_{t-1}, \sigma^2)$. We also need to consider y_1 . We know that $y_1 = \sum_{i=0}^{\infty} \phi^i \epsilon_{1-i}$. Therefore, the marginal distribution of y_1 is $N[0, \sigma^2/(1 - \phi^2)]$. Thus the joint density is

$$f(Y_T) = \left[\prod_{t=2}^T (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_t - \phi y_{t-1})^2\right) \right] \times \\ (2\pi\sigma^2)^{-1/2} (1 - \phi^2)^{1/2} \exp\left(-\frac{1 - \phi^2}{2\sigma^2} y_1^2\right)$$

so that the log of the likelihood function is simply

$$\mathcal{L} = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) + \frac{1}{2} \ln(1 - \phi^2) - \frac{1}{2\sigma^2} \sum_{t=2}^T (y_t - \phi y_{t-1})^2 - \frac{(1 - \phi^2)}{2\sigma^2} y_1^2.$$

Notice that if we were to drop the two extra terms from the likelihood we'd have the OLS criterion function. Since that term dominates as the sample size grows there is going to be numerical similarity between the ML estimate and the OLS estimate in large samples. They are asymptotically equivalent. You should verify by computing the first derivatives, second derivatives of the likelihood function and then the information matrix that the asymptotic variance of the ML and OLS estimators is $\sigma^2 \gamma_0^{-1}$, which is consistently estimated by $\sigma^2 (\sum y_{t-1}^2 / T)^{-1}$ which in small samples means use $\sigma^2 (X'X)^{-1}$ from OLS! For higher order AR processes the story is the same although there are more complex issues to deal with concerning the first p observations.

For a moving average model it is difficult to write y in terms of lagged y 's and current shocks. For example, if we have an MA(1) model, $y_t = \epsilon_t + \theta \epsilon_{t-1}$ the appropriate way to do it is to invert the MA process to get $(1 + \theta L)^{-1} y_t = (1 - \theta L + \theta^2 L^2 - \dots) y_t = \epsilon_t$. Even in an MA(1) y depends on an infinite sequence of lagged y 's when the process is inverted. Therefore, the easier way to go is just to directly write down the likelihood for the process. Clearly, since the ϵ 's are normal, the y 's are normal. Each y has mean zero,

variance $\sigma^2(1 + \theta^2)$ and first order covariance $\theta\sigma^2$. Therefore, $Y_T \sim N(0, \sigma^2\Omega)$ where

$$\Omega = \begin{pmatrix} 1 + \theta^2 & \theta & 0 & \dots & 0 \\ \theta & 1 + \theta^2 & \theta & \dots & 0 \\ 0 & \theta & 1 + \theta^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 + \theta^2 \end{pmatrix}$$

So that

$$\mathcal{L} = -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \ln |\Omega| - \frac{1}{2\sigma^2} Y_T' \Omega^{-1} Y_T.$$

There are some kinky ways to estimate this guy but you wont learn them here. Shazam seems a little useless to me, try RATS.

8.3.3. Diagnostic Testing

One thing you could do which is straightforward is to redo the Ljung-Box or Box-Pierce tests on the residuals to check for leftover autocorrelations in the data. Unfortunately, these tests have very low power, so they tend to often accept the null when it is false. You really can only get much out of these tests if you have a large sample and you can sum over many autocorrelations.

A better approach is to use LM tests. You will have chosen a particular model which you might think of as a restricted version of a more general model. If you want to test whether your version is adequate you can use an LM test, which only requires estimation of the restricted model along with the derivatives of the likelihood function at the parameter estimates. We will only examine a test based on a selected model of AR(1). I will test this specification versus an AR(2). Dropping the starting values from consideration the likelihood is just

$$\mathcal{L} = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2})^2.$$

In this case the usual block separation of the information matrix ϕ components and σ^2

component will hold so we only need

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \phi_1} &= \frac{1}{\sigma^2} \sum_{t=1}^T (y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2}) y_{t-1} \\ \frac{\partial \mathcal{L}}{\partial \phi_2} &= \frac{1}{\sigma^2} \sum_{t=1}^T (y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2}) y_{t-2} \\ \frac{\partial^2 \mathcal{L}}{\partial \phi_1^2} &= -\frac{1}{\sigma^2} \sum_{t=1}^T y_{t-1}^2 \\ \frac{\partial^2 \mathcal{L}}{\partial \phi_1 \phi_2} &= -\frac{1}{\sigma^2} \sum_{t=1}^T y_{t-1} y_{t-2} \\ \frac{\partial^2 \mathcal{L}}{\partial \phi_2^2} &= -\frac{1}{\sigma^2} \sum_{t=1}^T y_{t-2}^2\end{aligned}$$

At the AR(1) estimates $\phi_2 = 0$ and ϕ_1 renders the first order condition for ϕ_1 to zero.

$$\frac{\partial \mathcal{L}}{\partial \phi_1} = \frac{1}{\sigma^2} \sum_{t=1}^T (y_t - \phi_1 y_{t-1}) y_{t-1} = \frac{1}{\sigma^2} \sum_{t=1}^T e_t y_{t-1} = 0$$

The first order condition for ϕ_2 reduces to

$$\frac{\partial \mathcal{L}}{\partial \phi_2} = \frac{1}{\sigma^2} \sum_{t=1}^T (y_t - \phi_1 y_{t-1}) y_{t-2} = \frac{1}{\sigma^2} \sum_{t=1}^T e_t y_{t-2}$$

This makes the LM statistic

$$\begin{aligned}LM &= \begin{pmatrix} \sum e_t y_{t-1} & \sum e_t y_{t-2} \end{pmatrix} \begin{pmatrix} \sum y_{t-1}^2 & y_{t-1} y_{t-2} \\ y_{t-1} y_{t-2} & \sum y_{t-2}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum e_t y_{t-1} \\ \sum e_t y_{t-2} \end{pmatrix} / \hat{\sigma}^2 \\ &= T \begin{pmatrix} \sum e_t y_{t-1} & \sum e_t y_{t-2} \end{pmatrix} \begin{pmatrix} \sum y_{t-1}^2 & y_{t-1} y_{t-2} \\ y_{t-1} y_{t-2} & \sum y_{t-2}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum e_t y_{t-1} \\ \sum e_t y_{t-2} \end{pmatrix} / \sum e_t^2 \\ &= T e' X (X' X)^{-1} X' e / (e' e) \\ &= TR^2\end{aligned}$$

from the regression of e on y_{t-1} and y_{t-2} .

Suppose I tried an ARMA(1,1) model. In this case (and I won't go into why) it turns out that the relevant regression is the residuals e_t on y_{t-1} and e_t . See if you can decide on analogs for more complex models.

9. Vector Autoregressions

At this point we move to a multivariate setting since univariate models aren't particularly interesting from an economic perspective. If we just wrote down a multivariate ARMA model it would look like

$$\Phi_0 x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \dots + \Phi_p x_{t-p} + \epsilon_t + \Theta_1 \epsilon_{t-1} + \dots + \Theta_q \epsilon_{t-q}$$

where Φ_0 isn't necessarily I since we might want to allow contemporaneous responses of one variable to another and $E(\epsilon_t \epsilon_t') = \Sigma$. Notice how many free parameters we have $n^2(1 + p + q) + n(n + 1)/2$. This could be hideous. I won't discuss the restrictions on the matrix polynomials $\Phi(Z)$ and $\Theta(Z)$ required to render this process stationary and invertible, suffice it to say that they are analagous to the univariate setting.

9.1. Identification

Identification is going to be a problem, just as it is in simultaneous equations models. Suppose we restrict our attention to models with no MA components (they'd be hideous to deal with) and write a *vector autoregression* (VAR)

$$\Phi_0 x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \dots + \Phi_p x_{t-p} + \epsilon_t.$$

The likelihood function for the ϵ 's is

$$L = \prod_{t=1}^T (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\epsilon_t' \Sigma^{-1} \epsilon_t\right)$$

The Jacobian of the transformation from x_t to ϵ_t is Φ_0 so that the likelihood of X is

(ignoring starting values)

$$\begin{aligned}
 L &= \prod_{t=1}^T (2\pi)^{-n/2} |\Sigma|^{-1/2} |\Phi_0| \exp\left(-[\Phi(L)x_t]' \Sigma^{-1} [\Phi(L)x_t]\right) \\
 &= \prod_{t=1}^T (2\pi)^{-n/2} |\Phi_0^{-1} \Sigma \Phi_0^{-1'}|^{-1/2} \exp\left([x_t' \Phi_0' - x_{t-1}' \Phi_1' - \dots - x_{t-p}' \Phi_p'] \Sigma^{-1} \right. \\
 &\quad \left. [\Phi_0 x_t - \Phi_1 x_{t-1} - \dots - \Phi_p x_{t-p}]\right) \\
 &= \prod_{t=1}^T (2\pi)^{-n/2} |\Phi_0^{-1} \Sigma \Phi_0^{-1'}|^{-1/2} \exp\left([x_t' - x_{t-1}' \Phi_1' \Phi_0^{-1'} - \dots - x_{t-p}' \Phi_p' \Phi_0^{-1'}] \Phi_0' \Sigma^{-1} \right. \\
 &\quad \left. \Phi_0 [x_t - \Phi_0^{-1} \Phi_1 x_{t-1} - \dots - \Phi_0^{-1} \Phi_p x_{t-p}]\right) \\
 &= \prod_{t=1}^T (2\pi)^{-n/2} |V|^{-1/2} \exp\left([x_t' - x_{t-1}' C_1' - \dots - x_{t-p}' C_p'] V^{-1} \right. \\
 &\quad \left. [x_t - C_1 x_{t-1} - \dots - C_p x_{t-p}]\right)
 \end{aligned}$$

Suppose I generated new parameters using any non-singular $n \times n$ matrix G ,

$$\bar{\Phi}_i = G\Phi_i \quad \bar{\Sigma} = G\Sigma G'$$

then

$$\begin{aligned}
 \bar{C}_i &= \bar{\Phi}_0^{-1} \bar{\Phi}_i \\
 &= (G\Phi_0)^{-1} G\Phi_i \\
 &= \Phi_0^{-1} G^{-1} G\Phi_i \\
 &= \Phi_0^{-1} \Phi_i = C_i \\
 \bar{V} &= \bar{\Phi}_0^{-1} \bar{\Sigma} \bar{\Phi}_0'^{-1} \\
 &= \Phi_0^{-1} G^{-1} G\Sigma G' G'^{-1} \Phi_0'^{-1} \\
 &= \Phi_0^{-1} \Sigma \Phi_0'^{-1} = V
 \end{aligned}$$

This means that just as in the simultaneous equations models we must put some restrictions upon the parameters to achieve identification. The idea is to put enough restrictions on the Φ_i and Σ to imply that the only admissible G that delivers C_i and V is $G = I$. Unlike simultaneous equations models, people who use VAR models don't typically have in mind restrictions on the Φ_i matrices for $i \geq 1$. This leaves us with two possibilities.

- i) Restrictions on Φ_0 and/or
- ii) Restrictions on Σ .

The macroeconomist typically wants an interpretation of his model which allows him to identify error terms as particular kinds of shocks. For example, money shocks, supply shocks, technology shocks etc. In choosing what restrictions to impose, the economist needs to be aware of what he wants to do with his model once it's identified.

9.1.1. Restrictions are $\Phi_0 = I$.

Clearly this achieves identification. Now, this means that the model can be written as

$$x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \dots + \Phi_p x_{t-p} + \epsilon_t.$$

This model has an infinite moving average representation which can be written by inverting the autoregressive matrix polynomial in the lag operator. If we write $\Phi(L)x_t = \epsilon_t$ then we have $x_t = \Phi(L)^{-1}\epsilon_t = \Lambda(L)\epsilon_t$ or

$$x_t = \Lambda_0 \epsilon_t + \Lambda_1 \epsilon_{t-1} + \Lambda_2 \epsilon_{t-2} + \dots$$

One of the ways in which economists like to interpret VARs is to look at this moving average representation. It's quickly obvious that

$$\frac{\partial x_{i,t}}{\partial \epsilon_{j,t-k}} = \Lambda_k[ij] = \lambda_{k,ij}.$$

Typically $\lambda_{k,ij}$ is plotted as a function of k and is called the *impulse response function* (IRF) of variable i with respect to shocks in equation j . The only problem with this concept is that with no restrictions on Σ this is a rather wierd thing to do. If there's a shock to ϵ_{jt} and there's covariance between ϵ_j and ϵ_l how can we look at the response to one shock in isolation from the responses to other shocks which are correlated with it? To be able to isolate individual shocks we need to identify shocks which are uncorrelated with each other. Suppose we do a decomposition of $\Sigma = C'DC$, C is invertible and upper

triangular with 1's on the diagonal and D is a diagonal matrix. If we create new shocks $v_t = C^{-1}'\epsilon_t$ notice that $E(v_tv_t') = E(C^{-1}'\epsilon_t\epsilon_t'C^{-1}) = D$. We can then write

$$\begin{aligned} x_t &= \Lambda_0 C' v_t + \Lambda_1 C' v_{t-1} + \Lambda_2 C' v_{t-2} + \dots \\ &= \bar{\Lambda}_0 v_t + \bar{\Lambda}_1 v_{t-1} + \bar{\Lambda}_2 v_{t-2} + \dots \end{aligned}$$

which creates a new impulse response function with respect to the elements of v_t given by $\bar{\lambda}_{k,ij}$. This does allow the computation of a response function to shocks which are mutually orthogonal but it still creates another problem of interpretation. Unfortunately, the impulse responses will be sensitive to the ordering of the equations. The decomposition we did will lead to different results if the variables are ordered differently in the VAR. This leads us to wonder what the heck getting impulse responses to v_t 's really means. Sometimes theory can shed light on what ordering is appropriate but clear choices are rare.

A second thing people like to look at is called the *variance decomposition*. Let's go back to our representation

$$x_t = \Lambda_0 \epsilon_t + \Lambda_1 \epsilon_{t-1} + \Lambda_2 \epsilon_{t-2} + \dots$$

Suppose we wanted to forecast x_t with information up to time $t - k$. The forecast error would be

$$\begin{aligned} e_{k,t} &= x_t - E(x_t | \mathcal{I}_{t-k}) \\ &= \Lambda_0 \epsilon_t + \dots + \Lambda_{k-1} \epsilon_{t-k+1} \end{aligned}$$

The variance of this k -step ahead forecast error¹ is given by

$$V_k = \Lambda_0 \Sigma \Lambda_0' + \dots + \Lambda_{k-1} \Sigma \Lambda_{k-1}'.$$

Clearly V_k is a linear function of σ_{ij} , $\forall i, j$. The variance decomposition looks at the percentage of the variance of the k -step ahead forecast error in forecasting variable i which is due to variation in shock j . This is not well defined if there are covariance terms floating around in the expressions on the diagonal of V_k . We'd want to look at $V_k(ii)$ and see

¹ Notice that at $k = \infty$ you get the unconditional variance in x_t .

something like $a_1\sigma_{11} + a_2\sigma_{22} + \dots + a_n\sigma_{nn}$ and take the ratio of $a_j\sigma_{jj}$ to $V_k(ii)$. The only way to ensure that the diagonal of V_k has this appearance is to have a diagonal Σ such as we had in our decomposed model.

9.1.2. Restrict Φ_0 and Σ

In this case, the most common restriction is to assume that Φ_0 is lower triangular with 1's on the diagonal and that Σ is diagonal. This setup has the advantage of allowing contemporaneous relationships between the variables. For this reason it is often referred to as a *structural VAR*. Because Σ is assumed to be diagonal we can interpret the shocks in the model as independent processes. This helps in thinking about the impulse response functions.

$$\begin{aligned}\Phi_0 x_t &= \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \dots + \Phi_p x_{t-p} + \epsilon_t \\ x_t &= \Phi_0^{-1} \Phi_1 x_{t-1} + \Phi_0^{-1} \Phi_2 x_{t-2} + \dots + \Phi_0^{-1} \Phi_p x_{t-p} + \Phi_0^{-1} \epsilon_t \\ &= A_1 x_{t-1} + A_2 x_{t-2} + \dots + A_p x_{t-p} + u_t\end{aligned}$$

So that our structural model can be written as a plain VAR model with variance covariance matrix for the u 's equal to $\Phi_0^{-1} \Sigma \Phi_0^{-1'}$. So suppose we then wrote down the moving average representation of x_t in terms of u_t . We'd have

$$x_t = \sum_{k=0}^{\infty} B_k u_{t-k}$$

To get the moving average of the representation of x_t in terms of the ϵ_t 's is quite simple.

It is simply

$$x_t = \sum_{k=0}^{\infty} B_k \Phi_0^{-1} \epsilon_{t-k} = \sum_{k=0}^{\infty} \bar{B}_k \epsilon_{t-k}.$$

Notice that when we computed the IRF in the $\Phi_0 = I$ model we took it with respect to the v_t which were orthogonal, by getting a decomposition of the nonspherical ϵ 's using $C^{-1'}$. Here we are decomposing the nonspherical u_t 's using the matrix Φ_0^{-1} which by assumption looks a lot like C did. In both cases the model when written in terms of nonspherical

disturbances was a plain VAR model. Thus if we estimated a plain VAR model with nonspherical disturbances we would simply have to decide how to decompose the variance covariance matrix of the errors. Either by obtaining an estimate of Φ_0 or of C . Notice that in the structural model, if the variance decomposition is computed as

$$V_k = \bar{B}_0 \Sigma \bar{B}'_0 + \dots + \bar{B}_{k-1} \Sigma \bar{B}'_{k-1}$$

then the expression will be linear in the σ_{ii} and no covariance terms will appear since Σ is assumed to be diagonal.

9.2. Estimation

We're going to work with either set of restrictions and we'll write both models in their plain VAR form, generically as

$$x_t = D_1 x_{t-1} + D_2 x_{t-2} + \dots + D_p x_{t-p} + w_t$$

Notice that in either setup, this implies that w_t is an error term which is nonspherical. As a result GLS seems appropriate. However, since in each equation (remember x_t is a vector here) the regressors are identical, GLS (which is SURE) is going to be numerically identical to OLS equation by equation. Therefore in either case simply compute \hat{D}_i by OLS and get an estimate, $\hat{\Sigma}_w = \sum \hat{w}_t \hat{w}'_t / T$, where the \hat{w}_t 's are the residuals.

9.2.1. Restriction 9.1.1.

In this case the \hat{D}_i represent estimates of the Φ_i and $\hat{\Sigma}_w$ is an estimate of Σ . To do the impulse response functions simply requires the computation of a \hat{C} and \hat{D} such that $\hat{C}' \hat{D} \hat{C} = \hat{\Sigma}_w$.

9.2.2. Restriction 9.1.2.

In this case the \hat{D}_i represent estimates of the $\Phi_0^{-1}\Phi_i$ and $\hat{\Sigma}_w$ is an estimate of $\Phi_0^{-1}\Sigma\Phi_0'^{-1}$. Clearly to get out the parameters we're interested in requires an estimate of Φ_0 . But remember simultaneous equations. If Φ_0 is lower triangular we can do OLS equation by equation and get the right estimates.

9.3. Structural VARs

Suppose we didn't have such a wonderful situation. Suppose that Φ_0 had enough restrictions in it to achieve identification but it wasn't lower triangular. This would be tricky. It's hard because the triangularity bought a lot. It meant that if you looked in any equation like the one for x_{it} you'd see

$$x_{it} = \beta_{1i}x_{1t} + \dots + \beta_{i-1i}x_{i-1t} + f(\text{lags}) + \epsilon_{it}.$$

The whole secret to a system like this is that everything on the right hand side is either exogenous or is determined by an equation higher in the order **which does not depend on** x_{it} . Without triangularity OLS wont work. So what will we do. Blanchard and Watson (1986) suggest a way of doing this which only works for some kinds of nontriangular systems. Let me illustrate their procedure with a 4-variable system.

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -\alpha \\ 0 & 0 & 1 & -\beta \\ -\phi_1 & -\phi_2 & -\phi_3 & 1 \end{pmatrix} y_t = Z_t\gamma + \epsilon_t$$

where Z_t represents all the lags. This can be written as 4 equations,

$$y_1 = Z\gamma_1 + \epsilon_1$$

$$y_2 = \alpha y_4 + Z\gamma_2 + \epsilon_1$$

$$y_3 = \beta y_4 + Z\gamma_3 + \epsilon_1$$

$$y_4 = \phi_1 y_1 + \phi_2 y_2 + \phi_3 y_3 + Z\gamma_4 + \epsilon_1$$

If you work out the reduced form for y_4 it is

$$y_4 = [Z(\phi_1\gamma_1 + \phi_2\gamma_2 + \phi_3\gamma_3 + \gamma_4) + (\phi_1\epsilon_1 + \phi_2\epsilon_2 + \phi_3\epsilon_3 + \epsilon_4)] / (1 - \alpha\phi_2 - \beta\phi_3)$$

Now suppose I estimated each equation by OLS but left out the endogenous RHS variables creating what are really reduced form coefficient estimates which I'll call $\hat{\gamma}_i$. These guys wouldn't converge to the Φ_i 's rather they'd converge to the $\Phi_0^{-1}\Phi_i$'s. Create four residual series

$$e_1 = y_1 - Z(Z'Z)^{-1}Z'y_1 = My_1$$

$$= M(Z\gamma_1 + \epsilon_1) = M\epsilon_1$$

$$e_2 = My_2 = M(Z\gamma_2 + \alpha y_4 + \epsilon_2)$$

$$= \alpha My_4 + M\epsilon_2 = \alpha e_4 + M\epsilon_2$$

$$e_3 = My_3 = M(Z\gamma_3 + \beta y_4 + \epsilon_3)$$

$$= \beta My_4 + M\epsilon_3 = \beta e_4 + M\epsilon_3$$

$$e_4 = My_4 = M(Z\gamma_4 + \phi_1 y_1 + \phi_2 y_2 + \phi_3 y_3 + \epsilon_4) = \phi_1 My_1 + \phi_2 My_2 + \phi_3 My_3 + M\epsilon_4$$

$$= \phi_1 e_1 + \phi_2 e_2 + \phi_3 e_3 + M\epsilon_4$$

How could we get $\hat{\alpha}$? Regress e_2 on e_4 ?

$$\begin{aligned}\hat{\alpha} &= (e_4' e_4)^{-1} e_4' e_2 \\ &= (e_4' e_4)^{-1} e_4' (\alpha e_4 + M\epsilon_2) \\ &= \alpha + (y_4' M y_4)^{-1} y_4' M \epsilon_2\end{aligned}$$

Clearly this means there will be bias in $\hat{\alpha}$ because look at covariance in reduced form for y_4 with ϵ_2 . Could we instrument it out? Yes. Use e_1 as an instrument for e_4 !

$$\begin{aligned}\hat{\alpha}_{IV} &= (e_1' e_4)^{-1} e_1' e_2 \\ &= (e_1' e_4)^{-1} e_1' (\alpha e_4 + M\epsilon_2) \\ &= \alpha + (\epsilon_1' M y_4)^{-1} \epsilon_1' M \epsilon_2\end{aligned}$$

This is going to work because the second part is zero in expectation and the first part is nonsingular because ϵ_1 appears in y_4 's reduced form. This last part is critical! We could

continue like this by next taking e_1 and $\bar{e}_2 = e_2 - \hat{\alpha}_{IV}e_4$ as instruments (using optimal IV methods) for e_4 to get an estimate of β . This would lead to $\bar{e}_3 = e_3 - \hat{\beta}_{IV}e_4$. Finally, e_1 , \bar{e}_2 and \bar{e}_3 could be (using optimal IV methods) to get instruments for e_1 , e_2 and e_3 and thereby get estimates of the ϕ_i 's.

10. Generalized Method of Moments

There is a large class of macroeconomic models which lead to Euler equations, or in general, nonlinear restrictions of the form

$$E_{t-i}f(Y_{t-1}, y_t|\theta) = 0.$$

for some $i \geq 0$. Because of the law of iterated expectations it follows that $E_{t-i}[f(Y_{t-1}, y_t|\theta)x_{t-i}] = 0$ for any x_{t-i} which is in the time $t-i$ information set. Furthermore we can write $E[f(Y_{t-1}, y_t|\theta)x_{t-i}] = 0$. With restrictions like these we could write $f(Y_{t-1}, y_t|\theta)x_{t-i} = u_t$, where restriction is rewritten as $Eu_t = 0$. This just looks like a nonlinear model. The problem that GMM estimation is meant to solve is that the model in its most primitive form often makes no statement as to the distribution function of u_t . I.e. we don't know that it's normal or virtually anything about it (e.g. we don't know whether it is heteroskedastic or autocorrelated). We do know that $E(u_t) = 0$. Furthermore, $E(u_t u_{t-j}) = 0$ for $|j| \geq i$, otherwise nonzero, because u_t is orthogonal to anything known up to time $t-i$.

10.1. Estimation

The intuition behind GMM estimation is to have an estimator of θ which sets the sample equivalent of the moment condition equal to zero. I.e. choose $\hat{\theta}$ so that

$$\frac{1}{T} \sum_{t=1}^T f(Y_{t-1}, y_t|\theta) \otimes x_{t-i} = \frac{1}{T} \sum_{t=1}^T u_t(\theta)$$

is set equal to zero. I'm now making explicit the dependence of the error on the choice of θ . Suppose f is $m \times 1$ and x_{t-i} is $n \times 1$. If θ is $k \times 1$ then, in general, we can only set the average of the u_t equal to zero if $k = mn$. Usually what we have is a situation where $k < mn$. As a result, we can only set $(1/T) \sum u_t$ close to zero. But how? One way to proceed is to minimize

$$J = \left[\frac{1}{T} \sum_{t=1}^T u_t(\theta) \right]' W_T \left[\frac{1}{T} \sum_{t=1}^T u_t(\theta) \right]$$

where W_T is some positive definite symmetric matrix which may be sample dependent.

What is the first order condition?

$$\frac{\partial J}{\partial \theta} = \frac{1}{T} \sum_{t=1}^T \frac{\partial u_t(\theta)'}{\partial \theta}{}_{mn \times k} W_{mn \times mn} \frac{1}{T} \sum_{t=1}^T u_t(\theta)_{mn \times 1} = 0$$

As you can see, the first-order condition sets k linear combinations of the mn restrictions to zero.

10.2. Asymptotics

It turns out that as long as $u_t(\theta_0)$ is stationary the $\hat{\theta}$ which solves the first order condition will be consistent. As a result

$$\frac{1}{T} \sum_{t=1}^T \frac{\partial u_t(\hat{\theta})}{\partial \theta} \xrightarrow{p} = D_0$$

where

$$D_0 = E \frac{\partial u_t(\theta_0)}{\partial \theta}.$$

Furthermore, with the assumption that $u_t(\theta_0)$ is stationary we can assume that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t(\theta_0) \xrightarrow{d} = N(0, S_0).$$

where

$$S_0 = E \left(\sum_{j=-\infty}^{\infty} u_t(\theta_0) u_{t-j}(\theta_0)' \right).$$

It turns out that if W_T is chosen (optimally) to be a consistent estimator for S_0^{-1} , then

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N \left[0, (D_0' S_0^{-1} D_0)^{-1} \right].$$

Of course you'd have to estimate the variance covariance matrix as usual. This is usually done by computing the sample counterparts of D_0 and S_0 .

10.3. Examples

Some examples serve to illustrate GMM.

10.3.1. Estimating the Mean of a Random Variable

Suppose we are interested in estimating the mean of a scalar random variable x_t . If we parameterize its mean as μ , we have the moment restriction

$$E(x_t - \mu) = 0.$$

As a result, $mn = 1$ and $k = 1$, so that μ is exactly identified. The GMM estimator for μ is simply the sample mean

$$\hat{\mu} = \bar{x} = \frac{1}{T} \sum_{t=1}^T x_t.$$

Clearly, $D_0 = -1$. Also, assuming x_t is stationary, we have

$$S_0 = E\left(\sum_{j=-\infty}^{\infty} (x_t - \mu_0)(x_{t-j} - \mu_0)\right).$$

One special case would be that x_t is i.i.d. In this case,

$$S_0 = E\left[(x_t - \mu_0)^2\right] = \sigma^2 \equiv \text{Var}(x_t).$$

Thus, in the i.i.d. case we have

$$\sqrt{T}(\hat{\mu} - \mu_0) \xrightarrow{d} N(0, \sigma^2),$$

which is the familiar result from undergraduate statistics where $\hat{\mu}$ is casually written as having a normal distribution with variance σ^2/T in large samples.

10.3.2. The Linear Model With Stochastic Regressors

Suppose the model is $y_t = x_t' \beta + \epsilon_t$ where ϵ_t is white noise. Clearly $E_t(y_t - x_t' \beta) = 0$. Of course this implies that $E_t x_t (y_t - x_t' \beta) = E x_t (y_t - x_t' \beta) = 0$. This is a case where $m = 1$ and $n = k$ so we have exact identification. Therefore we can set

$$\frac{1}{T} \sum_{t=1}^T x_t (y_t - x_t' \beta) = 0$$

by choosing $\hat{\beta} = (\sum_{t=1}^T x_t x_t')^{-1} \sum_{t=1}^T (x_t y_t)$. What's the variance-covariance matrix? It had better be the usual one.

$$\frac{\partial u_t}{\partial \beta} = -x_t x_t'$$

so that $D_0 = -\Sigma_{xx}$. What about S_0 ? We have $E(u_t u_t') = E(x_t \epsilon_t^2 x_t') = \sigma^2 \Sigma_{xx}$, while $E(u_t u_{t-i}') = 0$ for $i \neq 0$. Therefore, $S_0 = \sigma^2 \Sigma_{xx}$. Therefore the variance-covariance matrix is $\sigma^2 \Sigma_{xx}^{-1}$.

10.3.3. The Consumption-Based Asset Pricing Model

When consumers have time-separable constant relative risk aversion preferences, so that $u(c) = c^{1-\gamma}/(1-\gamma)$. Then the intertemporal Euler equation which governs the behavior of returns is

$$1 = E_t \beta z_{t+1}^{-\gamma} R_{t+1},$$

where $z_{t+1} = c_{t+1}/c_t$. In this case, the parameters of interest are β and γ . As a result, to estimate these parameters, the econometrician can use Euler equations for different assets, thus increasing the number of moment restrictions, or the econometrician can identify different variables x_t that are in the time t information set. As a result, the econometrician uses a set of equations given by

$$E \left[(\beta z_{t+1}^{-\gamma} R_{t+1} - 1) \otimes x_t \right],$$

where x_t is $n \times 1$. Typical choices for the elements of x_t would be lagged values of consumption growth and the asset returns. In general, the variable R can be an $m \times 1$ vector, so that we have mn moment restrictions. As a result, there will be $mn - 2$ overidentifying restrictions.

10.4. Testing With Overidentifying Restrictions

Last, but not least, when $mn > k$ and the weighting matrix is chosen optimally, it turns out that $TJ(\hat{\theta}) \xrightarrow{d} \chi^2(mn-k)$. This provides a statistic which can be used to test the overidentifying moment conditions. The idea is that when the overidentifying restrictions hold, the minimized value of J should be small even though it won't be exactly zero. The factor T is required, since J converges to 0 with probability 1 under the null hypothesis that the moment restrictions all hold. On the other hand TJ converges to a random variable which can be used to test this hypothesis. When TJ is large, the hypothesis is rejected.