

Economics 624: Econometrics *

Matthew Chesnes

Updated: May 12, 2005

*These are Matthew Chesnes' notes from a course taught by Ingmar Prucha.

1 Lecture 1: January 27, 2005

1.1 The Nature of Econometric Modelling

- 3 input streams:

Economic Theory \Rightarrow Economic Model \Rightarrow Econometric Model.

Facts \Rightarrow Data \Rightarrow Refined Data.

Statistical Theory \Rightarrow Econometric Techniques.

- Then we estimate the econometric model with refined data using econometric techniques. Out of this we can do structural analysis, forecasting, and policy analysis.
- Example. Factor demand for labor and capital of a competitive firm. The problem might look like:

$$\text{Max } F(K_t, L_t) - w_t L_t - c_t K_t.$$

FOCs:

$$\frac{\partial F}{\partial L_t} - w_t = 0.$$
$$\frac{\partial F}{\partial K_t} - c_t = 0.$$

Assume a quadratic form with no cross product term:

$$F(K_t, L_t) = \alpha_0 + \alpha_K K_t + \alpha_L L_t + \frac{1}{2} \alpha_{KK} K_t^2 + \frac{1}{2} \alpha_{LL} L_t^2.$$

FOCs become:

$$\alpha_L + \alpha_{LL} L_t - w_t = 0.$$
$$\alpha_K + \alpha_{KK} K_t - c_t = 0.$$

Solve for L and K :

$$L_t = \frac{1}{\alpha_{LL}} (w_t - \alpha_L). \quad (1)$$

$$K_t = \frac{1}{\alpha_{KK}} (c_t - \alpha_K). \quad (2)$$

So the economic theory gives us these two equations along with our quadratic production function. But to do econometric testing, the data almost surely will not fit these equations exactly so we need to introduce some error terms. Suppose we interpret α_L as $\alpha_L - \eta_t^L$, so labor's share of production is reduced (or increased) by productivity shocks. Suppose η_t^L is known to the firm but unknown to us, the researchers. Suppose we have a similar adjustment to α_K . Equations 1 and 2 become:

$$L_t = \frac{1}{\alpha_{LL}} (w_t - \alpha_L + \eta_t^L). \quad (1')$$

$$K_t = \frac{1}{\alpha_{KK}}(c_t - \alpha_K + \eta^K). \quad (2')$$

So these η terms become our error terms in our regressions.

1.2 Classical Linear Regression Model

- Suppose we have the regression relation:

$$y_t = x_{t1}\beta_1 + \cdots + x_{tK}\beta_K + u_t, \quad t = 1, \dots, T.$$

- If we take the X 's as given constants and make the following three assumptions:

$$E[u_t] = 0, \text{ Mean Zero,}$$

$$E[u_t^2] = \sigma^2, \text{ Constant Variance,}$$

$$E[u_t x_s] = 0, \text{ No Correlation.}$$

Then we are working with the ‘‘Classical’’ linear regression model.

- In matrix form:

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

where y and u are $T \times 1$, X is $T \times K$ and β is $K \times 1$.

- The OLS estimator becomes:

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

- We will also study the nonlinear regression model in the later part of the course which will involve a regression equation:

$$y_t = f(X_t, \beta) + u_t.$$

Here we won't be able to solve directly for the parameters but will utilize asymptotic theory to determine our best estimate of the true parameters.

2 Lecture 2: February 1, 2005

2.1 Classical Linear Regression Model

Specification

- We start with a simple 2 variable regression:

$$C_t = f(Y_t) = a + bY_t.$$

- The specification might be:

$$C_t = a + bY_t + u_t,$$

$$E[C_t|Y_t] = a + bY_t,$$

$$E[u_t|Y_t] = 0.$$

The expected value of consumption given income is the value of the average relationship (or fitted value).

- We also might assume:

$$E[u_t^2|Y_t] = h(Y_t) = \sigma^2, \text{ constant.}$$

This is an assumption of the classical model. We say the error terms are homoskedastic (versus heteroskedastic).

- In general, the relationship between two variables will NOT be deterministic, ie observations will not lie exactly on the regression line. We still do a linear regression sometimes if we think there might be error coming into the model from some unknown source. Reasons this might happen could be missing explanatory variables on the RHS of the regression, measurement error in the independent variable, or stochastic coefficients:

$$C_t = a_t + b_t Y_t, \quad a_t = a + v_t, \quad b_t = b + w_t, \quad E[v_t] = E[w_t] = 0.$$

Substituting:

$$c_t = a + bY_t + \underbrace{v_t + w_t Y_t}_{u_t}.$$

Note $E[u_t] = 0$.

- Typical regression notation is:

$$y_t = a + bx_t + u_t.$$

Where x_t is the regressor, u_t is the disturbance or error term, and y_t is the independent variable or regressand.

- Assumptions of the classical model:

- (a.1) $E[u_t|x_t] = 0 \forall t$.
 - (a.2a) $E[u_t^2|x_t] = \sigma^2 \forall t$.
 - (a.2b) $Cov[u_t u_s | x_t, x_s] = 0 \forall t \neq s$.
 - (a.3) $x_t \neq x_s$ for some $t \neq s$.
- Note that a.1 is not very strong but a.2 gives us a lot. If the variance was different for different observations, then it would not make sense to weight all the observations the same in the regression. With a.2, we treat all observations equally. Assumption a.3 is equivalent to saying that the X matrix of regressors is of full column rank, (rank K).
 - We usually impose STRICT exogeneity of the X 's as follows:
 - (a.1') $E[u_t|x_1, \dots, x_T] = 0 \forall t$.
 - (a.2a') $E[u_t^2|x_1, \dots, x_T] = \sigma^2 \forall t$.
 - (a.2b') $Cov[u_t u_s | x_1, \dots, x_T] = E[u_t u_s | x_1, \dots, x_T] = 0 \forall t \neq s$.
 - Example. Consider the simple regression equation with no intercept:

$$y_t = bx_t + u_t.$$

One estimator of b might be:

$$\hat{b} = \frac{\sum x_t y_t}{\sum x_t^2}.$$

Substituting:

$$\hat{b} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum x_t (bx_t + u_t)}{\sum x_t^2} = b + \frac{\sum x_t u_t}{\sum x_t^2}.$$

Now lets try to show that the expected value of our estimator is the true parameter:

$$E[\hat{b}] = b + E\left[\frac{\sum x_t u_t}{\sum x_t^2}\right].$$

Suppose we could separate the expectation operator (we can't):

$$\begin{aligned} E[\hat{b}] &= b + \frac{E[\sum x_t u_t]}{E[\sum x_t^2]} \\ &= b + \frac{E_{x_t} E[\sum x_t u_t | x_t]}{E[\sum x_t^2]} \\ &= b + \frac{E_{x_t} x_t \overbrace{E[\sum u_t | x_t]}^0}{E[\sum x_t^2]} \\ &= b \end{aligned}$$

And thus \hat{b} is unbiased. But since we cannot separate the expectations operator, we have to try something different:

$$\begin{aligned}
 E[\hat{b}] &= b + E\left[\frac{\sum x_t u_t}{\sum x_t^2}\right]. \\
 &= b + E_{x_1 \dots x_T} E\left[\frac{\sum x_t u_t}{\sum x_t^2} \middle| x_1 \dots x_T\right]. \\
 &= b + E_{x_1 \dots x_T} \frac{1}{\sum x_t^2} E\left[\sum x_t u_t \middle| x_1 \dots x_T\right]. \\
 &= b + E_{x_1 \dots x_T} \frac{1}{\sum x_t^2} \sum x_t \underbrace{E[u_t | x_1 \dots x_T]}_0. \\
 &= b
 \end{aligned}$$

So we needed strict exogeneity to get unbiasedness. This is the same as saying that the X 's are non-stochastic constants.

3 Lecture 3: February 3, 2005

3.1 Classical Linear Model - Multiple Variables

Notation

- Denote a matrix A with elements a_{ij} , a_i the i^{th} column of A .
- Denote the inverse matrix of A , A^{-1} , with elements a^{ij} , a^i the i^{th} column of A^{-1} .
- Denote $E[A] = (E[a_{ij}])$, the expectation of a matrix is the matrix of expected values of each element.

Preliminaries

- Consider the following regression model:

$$y_t = \beta_1 x_{t1} + \dots + \beta_K x_{tK} + u_t, \quad t = 1, \dots, T.$$

Or in matrix form:

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

Where y and u are $T \times 1$, β is $K \times 1$, and X is $T \times K$.

- Assumptions:
 - (A.1) $E[u|X] = 0$.
 - (A.2) $E[uu'|X] = \sigma^2 I_T$.
 - (A.3) $\text{Rank}(X) = K$, full rank.

To simplify the notation, WLOG, we can assume the X 's are nonstochastic so our assumptions become:

- (A.1) $E[u] = 0$.
 - (A.2) $E[uu'] = \sigma^2 I_T$.
 - (A.3) $\text{Rank}(X) = K$, X 's nonstochastic.
- Note that A.2 says that the variance/covariance matrix has a constant variance down the diagonal and 0 covariances off the diagonal. This comes from:

$$E[uv] = \underbrace{E[u]}_0 \underbrace{E[v]}_0 + \text{Cov}(u, v) = \text{Cov}(u, v).$$

- The full rank assumption means that we cannot have an explanatory variable that is a linear combination of one or more of the other explanatory variables. We might even have a case where $x_1 = ax_2 + bx_3$, and we call this perfect multicollinearity. In this situation, we can really only estimate the effects of say x_1 and x_2 or of x_1 and x_3 , assuming that there wasn't collinearity in these two variable regressions.

- Another way of saying that the X matrix must have full rank is to say that the parameters in the model must be fully identified.

3.2 The algebra of OLS with no parameter restrictions

- Model:

$$y = X\beta + u.$$

We will forgo the assumption about the full rank of X for now.

- Note the following about sample and population means of y . Take the simple regression model:

$$y_t = a + bx_t + u_t.$$

This implies:

$$E[y_t] = a + bx_t.$$

But as usual:

$$\bar{y} = \frac{1}{T} \sum_{t=1}^T y_t,$$

the sample mean. The expectation of y is actually a mean vector which may change over time depending on the value of x . In this situation, clearly the sample mean is not an estimator of the population mean.

- Back to our model. Let $\tilde{\beta}$ be *some* estimator of β . The following implications follow:

$$u = y - X\beta.$$

$$E[y] = X\beta.$$

$$\tilde{y} = X\tilde{\beta}.$$

$$\tilde{u} = y - X\tilde{\beta} = y - \tilde{y}.$$

- See G-3.1. Consider a true relationship, $y_t = a + bx_t + u_t$, and our estimated relationship:

$$\tilde{y}_t = \tilde{a} + \tilde{b}x_t.$$

From the graph it is clear that the vertical distance between the point and the line represents some sort of residual in our fit.

- Define the OLS estimator, $\hat{\beta}$, as the vector that minimizes:

$$\begin{aligned}
 S(\tilde{\beta}) &= \sum_{t=1}^T \tilde{u}^2 \\
 &= \tilde{u}'\tilde{u} = \|\tilde{u}\|^2 \\
 &= (y - \tilde{y})'(y - \tilde{y}) \\
 &= (y - X\tilde{\beta})'(y - X\tilde{\beta}) \\
 &= y'y - 2y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta}.
 \end{aligned}$$

Or in the simple case:

$$\text{Min}_{\tilde{a}, \tilde{b}} \sum_{t=1}^T (y_t - \tilde{a} - \tilde{b}x_t)^2.$$

- Note we could also use the sum of absolute deviations (LAD estimator) but this will have issues which we will talk about later. In general, the OLS estimator is much more sensitive to outliers than the LAD estimator.
- Two final notes: Given random n -vectors, z and y ,

$$\|z\| = \left(\sum_i z_i^2 \right)^{1/2}.$$

$$\|z - y\| = \left(\sum_i (z_i - y_i)^2 \right)^{1/2}.$$

This way of representing the OLS estimation procedure, minimizing $\|\tilde{u}\|^2$, follows from the interpretation that the OLS technique can also be interpreted as an orthogonal projection of the y vector onto the space spanned by the X vectors. Because we are minimizing the distance, we end up with an orthogonal relationship between the X 's and the residuals.

4 Lecture 4: February 8, 2005

4.1 Classical Linear Model - Estimator

OLS Estimation, No Restrictions

- Recall we seek to minimize:

$$S(\tilde{\beta}) = \text{Min} \{(y - X\tilde{\beta})'(y - X\tilde{\beta})\}.$$

- Note that when you have a multidimensional variable and you want to take a derivative, the results are almost the same as with scalars. Specifically if y is $px1$, A is pxs , and x is $sx1$ with $y = Ax$, then:

$$\frac{\partial y}{\partial x} = A.$$

Suppose now that $y = x'Ax$, a quadratic form. Then if A is symmetric:

$$\frac{\partial y}{\partial x} = 2x'A.$$

If A is not symmetric:

$$\frac{\partial y}{\partial x} = x'(A + A').$$

The second result follows from the fact that since $x'Ax$ is a scalar, $x'Ax = (x'Ax)' = x'A'x$. So:

$$y = \frac{1}{2}x'(\underbrace{A + A'}_{\text{Symmetric}})x.$$

So,

$$\frac{\partial y}{\partial x} = \frac{1}{2}2x'(A + A') = x'(A + A').$$

- So minimizing over $\tilde{\beta}$ gives FOC:

$$-2y'X + 2\tilde{\beta}'X'X = 0.$$

Or,

$$X'X\tilde{\beta} = X'y.$$

- This is the normal equation for OLS. In fact it is a set of K equations. If $X'X$ was non-singular, we could invert and solve for $\tilde{\beta}$. If $X'X$ was singular, we would have a whole manifold of solutions. When is $X'X$ non-singular? When X has full rank (K in this case). Thus, if X has rank K ,

$$\tilde{\beta} = (X'X)^{-1}X'y,$$

the OLS estimator. Note that $X'X$ must be positive semi-definite, ie if $A = X'X$, then $x'Ax \geq 0 \forall x \neq 0$.

- Note that if $Ab = a$, then $b = A^{-1}a$ if A can be inverted. If not, we denote the generalize inverse, or Moor-Penrose Inverse as A^+ such that:

$$AA^+A = A,$$

$$A^+AA^+ = A^+,$$

$$AA^+ \text{ symmetric,}$$

$$A^+A \text{ symmetric.}$$

Moreover the Moor-Penrose inverse is unique, and if A is square and non-singular, then $A^+ = A^{-1}$.

- In scalar notation, the LHS and RHS of the normal equations become:

$$\frac{1}{T}(X'X)_{ij} = \frac{1}{T} \sum_t x_{ti}x_{tj},$$

$$\frac{1}{T}(X'y)_{ij} = \frac{1}{T} \sum_t x_{ti}y_t.$$

But these are just the non-central sample moments so:

$$M_{xx}\tilde{\beta} = M_{xy}.$$

- Now suppose that:

$$X = [e_T, \underline{X}],$$

$$\beta = (\tilde{\beta}_1, \tilde{\beta})'.$$

Where e_T is a vector of ones. Thus we are including an intercept in our regression, $\tilde{\beta}_1$, and the slope parameters are all in $\tilde{\beta}$. Then:

$$X'X = \begin{bmatrix} e_T' \\ \underline{X}' \end{bmatrix} * \begin{bmatrix} e_T & \underline{X} \end{bmatrix} = \begin{bmatrix} e_T'e_T & e_T'\underline{X} \\ \underline{X}'e_T & \underline{X}'\underline{X} \end{bmatrix}.$$

And,

$$X'y = \begin{bmatrix} e_T' \\ \underline{X}' \end{bmatrix} * y = \begin{bmatrix} e_T'y \\ \underline{X}'y \end{bmatrix}.$$

Plugging these values into our normal equation and simplifying yields:

$$\underbrace{\underline{X}'(I - \frac{1}{T}e_Te_T')\underline{X}}_{\text{Subtracts the Mean}}\tilde{\beta} = \underline{X}'(I - \frac{1}{T}e_Te_T')y.$$

Note also that $I - \frac{1}{T}e_T e_T'$ is idempotent, so,

$$\underbrace{\underline{X}'(I - \frac{1}{T}e_T e_T')}_{\underline{X}'_*} \underbrace{(I - \frac{1}{T}e_T e_T')X}_{\underline{X}_*} \tilde{\underline{\beta}} = \underbrace{\underline{X}'(I - \frac{1}{T}e_T e_T')}_{\underline{X}'_*} \underbrace{(I - \frac{1}{T}e_T e_T')y}_{y_*}.$$

Or,

$$\underline{X}'_* \underline{X}_* \tilde{\underline{\beta}} = \underline{X}'_* y_*.$$

Then,

$$\frac{1}{T} \underline{X}'_* \underline{X}_* \tilde{\underline{\beta}} = \frac{1}{T} \underline{X}'_* y_*.$$

$$S_{ij} \tilde{\underline{\beta}} = \frac{1}{T} \underline{X}'_* y_*,$$

Where S_{ij} are the central sample moments.

- Example. Suppose we have the following regression model:

$$y_t = a + bx_t + u_t.$$

So X contains a column of ones and then the single x_t variable. Thus our normal equations are:

$$X'X * \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} = X'y.$$

Multiplying out, we have:

$$T\tilde{a} + \sum x_t \tilde{b} = \sum y_t,$$

and,

$$\sum x_t \tilde{a} + \sum x_t^2 \tilde{b} = \sum x_t y_t.$$

Divide the first through by T :

$$\tilde{a} + \bar{x}\tilde{b} = \bar{y},$$

So this shows that if you have an intercept, the regression line will always go through the point (\bar{x}, \bar{y}) .

- Multiplying the first through by \bar{x} and subtracting the second yields:

$$\tilde{b} = \frac{\sum_t (x_t - \bar{x})(y_t - \bar{y})}{\sum_t (x_t - \bar{x})^2}.$$

And from above:

$$\tilde{a} = \bar{y} - \bar{x}\tilde{b}.$$

So these are our OLS estimators.

- **Proposition 1** If X has full column rank, then the OLS estimator is unique and given by:

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

If you include an intercept, we can pull out the column of ones from the X matrix and write:

$$\underline{\hat{\beta}} = (\underline{X}'\underline{X}_*)^{-1}\underline{X}'\underline{y}_*,$$

$$\hat{\beta}_1 = \bar{y} - \bar{x}_2\hat{\beta}_2 - \dots - \bar{x}_K\hat{\beta}_K.$$

- **Proposition 2** Any solution of the normal equations minimizes the sum of the squared residuals. For any other vector of coefficients, b , $(y - Xb)'(y - Xb) \geq (y - X\hat{\beta})'(y - X\hat{\beta})$. See notes for proof.

5 Lecture 5: February 10, 2005

5.1 More on OLS with no restrictions

- **Proposition 3** If $\hat{\beta}$ and $\tilde{\beta}$ are any two solutions to the normal equations, then,

$$E[y] = X\hat{\beta} = X\tilde{\beta}.$$

Proof: By assumption:

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

$$X'X\tilde{\beta} = X'y.$$

So,

$$X'X(\hat{\beta} - \tilde{\beta}) = 0.$$

$$\underbrace{(\hat{\beta} - \tilde{\beta})' X'}_{Z'} \underbrace{X(\hat{\beta} - \tilde{\beta})}_Z = 0.$$

$$Z'Z = 0 \iff \sum z_i^2 = 0 \implies Z = 0.$$

So,

$$X(\hat{\beta} - \tilde{\beta}) = 0 \iff X\hat{\beta} = X\tilde{\beta}.$$

QED. This says that though the estimator need not be unique (as is the case when $X'X$ is singular), the systematic component of the relationship between y and X will be unique!

- Suppose X does not have full column rank. Thus,

$$X\gamma = 0, \gamma \neq 0.$$

Suppose $\hat{\beta}$ and $\tilde{\beta} = \hat{\beta} + c\gamma$ are both solutions to the normal equations. Then

$$X'X\tilde{\beta} = X'X\hat{\beta} + cX' \underbrace{X\gamma}_0 = X'X\hat{\beta} = X'y.$$

Thus β is not necessarily unique.

- **Lemma 1** If X does not have full rank then:

$$\hat{\beta} = (X'X)^+ X'y,$$

has NO economic meaning; the coefficient is NOT unique.

- **Proposition 4**

$$\hat{y} = X\hat{\beta} = X(X'X)^+ X'y$$

is a UNIQUE estimator for $X\beta$ no matter if X has full rank or not. Then:

$$\hat{u} = y - X\hat{\beta}$$

is also UNIQUE regardless whether or not β is identified. Further more, if X has full column rank, then the OLS estimators for β , $X\beta$, and u are:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y, \\ \hat{y} &= X\hat{\beta} = X(X'X)^{-1}X'y, \\ \hat{u} &= y - X\hat{\beta} = \underbrace{[I - X(X'X)^{-1}X']}_M y.\end{aligned}$$

- Note that $X(X'X)^{-1}X'$ is a projection matrix and M , above, will be defined next.
- **Lemma 2** $M = I - X(X'X)^{-1}X'$. M is idempotent, symmetric, and $MX = X'M' = X'M = 0$.

Proof: M is clearly symmetric since $(M')^+ = (M^+)'$. $MM = M$ just by multiplying out. Finally,

$$MX = X - X(X'X)^{-1}X'X = X - X = 0.$$

- **Proposition 5** $X'\hat{u} = \hat{y}'\hat{u} = 0$. So the estimated residuals are orthogonal to X and the fitted values.

Proof:

$$\begin{aligned}X'\hat{u} &= X'(y - X\hat{\beta}) = X'(y - X(X'X)^{-1}X'y) = X'My = 0. \\ \hat{y}'\hat{u} &= (X\hat{\beta})'\hat{u} = \hat{\beta}'\underbrace{X'\hat{u}}_0 = 0.\end{aligned}$$

- **Remark** Regardless of whether or not X has full rank:

$$y = \hat{y} + \hat{u},$$

where \hat{y} and \hat{u} are orthogonal to each other.

5.2 Coefficient of Determination (Goodness of Fit)

- First some notation. Define the following two $T \times 1$ vectors:

$$\begin{aligned}v &= (v_1, \dots, v_T)', \\ w &= (w_1, \dots, w_T)'. \end{aligned}$$

Let the sample mean and variance be denoted:

$$\begin{aligned}\bar{v} &= \frac{1}{T} \sum_t v_t = T^{-1}e_T'v. \\ s_{vv} &= \frac{1}{T} \sum_t (v_t - \bar{v})^2 = T^{-1}v'(I - \frac{e_T e_T'}{T})v.\end{aligned}$$

Denote the non-central second order sample moments between v and w :

$$m_{vw} = \frac{1}{T} \sum_t v_t w_t = T^{-1} v' w.$$

Denote the central second order sample moments between v and w :

$$s_{vw} = \frac{1}{T} \sum_t (v_t - \bar{v})(w_t - \bar{w}) = T^{-1} v' (I_T - \frac{e_T e_T'}{T}) w.$$

• **Proposition 6**

$$y'y = \hat{y}'\hat{y} + \hat{u}'\hat{u} \iff m_{yy} = m_{\hat{y}\hat{y}} + m_{\hat{u}\hat{u}}.$$

Proof:

$$y'y = (\hat{y} + \hat{u})'(\hat{y} + \hat{u}) = \hat{y}'\hat{y} + \hat{u}'\hat{u},$$

because $\hat{y}'\hat{u} = 0$.

• **Proposition 7** If $X = [e_T, X^*]$, ie there is an intercept, then:

$$X'\hat{u} = 0 \implies (e_T', X^{*'})'\hat{u} = (e_T'\hat{u}, X^{*'}\hat{u}) = 0.$$

Thus,

$$\bar{\hat{u}} = T^{-1} e_T' \hat{u} = 0,$$

from above. And, since $y = \hat{y} + \hat{u}$,

$$T^{-1} e_T' y = T^{-1} e_T' \hat{y} + \underbrace{T^{-1} e_T' \hat{u}}_0.$$

$$\bar{y} = \bar{\hat{y}}.$$

So if we have an intercept in our regression, the regression line goes through the point of sample averages. Both the mean residual is zero and the mean of the y 's equals the mean of the fitted values. Finally,

$$\begin{aligned} y'y &= \hat{y}'\hat{y} + \hat{u}'\hat{u} \\ T^{-1}y'y &= T^{-1}\hat{y}'\hat{y} + T^{-1}\hat{u}'\hat{u} \\ \underbrace{T^{-1}y'y - \bar{y}^2}_{s_{yy}} &= T^{-1}\hat{y}'\hat{y} - \underbrace{\bar{y}^2}_{\bar{\hat{y}}^2} + T^{-1}\hat{u}'\hat{u} - \underbrace{\bar{\hat{u}}}_0 \\ s_{yy} &= \underbrace{T^{-1}\hat{y}'\hat{y} - \bar{\hat{y}}^2}_{s_{\hat{y}\hat{y}}} + T^{-1}\hat{u}'\hat{u} - \bar{\hat{u}} \\ s_{yy} &= s_{\hat{y}\hat{y}} + \underbrace{T^{-1}\hat{u}'\hat{u} - \bar{\hat{u}}}_{s_{\hat{u}\hat{u}}} \\ s_{yy} &= s_{\hat{y}\hat{y}} + s_{\hat{u}\hat{u}} \end{aligned}$$

- **Definition** Coefficient of Determination. Denote:

$$R^2 = r_{y\hat{y}}^2,$$

the square of the simple correlation coefficient between the y 's and the fitted values. This can be written:

$$R^2 = \frac{s_{y\hat{y}}^2}{s_{yy}s_{\hat{y}\hat{y}}}.$$

So $0 \leq R^2 \leq 1$.

- **Proposition 8** Suppose $X = [e_T, X^*]$. Note:

$$\hat{y}'\hat{y} = y'\hat{y}.$$

Thus,

$$s_{y\hat{y}} = \frac{1}{T}y'\hat{y} - \bar{y}\bar{\hat{y}} = \frac{1}{T}\hat{y}'\hat{y} - \bar{\hat{y}}^2 = s_{\hat{y}\hat{y}}.$$

So,

$$R^2 = \frac{s_{y\hat{y}}^2}{s_{yy}s_{\hat{y}\hat{y}}} = \frac{s_{\hat{y}\hat{y}}^2}{s_{yy}s_{\hat{y}\hat{y}}} = \frac{s_{\hat{y}\hat{y}}}{s_{yy}} = \frac{s_{yy} - s_{\hat{u}\hat{u}}}{s_{yy}} = 1 - \frac{s_{\hat{u}\hat{u}}}{s_{yy}}.$$

Or,

$$R^2 = 1 - \frac{\sum \hat{u}_t^2}{\sum (y_t - \bar{y})^2}.$$

- Note that using this second definition and then not including an intercept would be problematic since the coefficient would no longer be bounded between zero and one. The first definition works always.
- Important remark:

$$ESS = \text{Error (Residual) Sum of Squares: } \sum \hat{u}_t^2.$$

$$RSS = \text{Regression (Explained) Sum of Squares: } \sum (\hat{y}_t - \bar{\hat{y}})^2.$$

$$TSS = \text{Total Sum of Squares: } \sum (y_t - \bar{y})^2.$$

E's and R's can sometimes be switched.

- In this notation,

$$s_{yy} = s_{\hat{y}\hat{y}} + s_{\hat{u}\hat{u}} \iff TSS = RSS + ESS.$$

So if we have an intercept,

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}.$$

- Note that R^2 will always be increasing in the number of parameters you include in the model. Thus, define the adjusted coefficient of determination:

$$\bar{R}^2 = R^2 - \left(\frac{K-1}{T-K}\right)(1-R^2).$$

This form of the coefficient penalizes the addition of K 's.

5.3 OLS Estimation with Parameter Restrictions

- Consider the set of G “linear” restrictions on β . Suppose $K = 5$ and we have restrictions:

$$\begin{aligned}\beta_2 + \beta_4 &= 5, \\ \beta_1 + 3\beta_2 + \beta_5 &= 7.\end{aligned}$$

Then write $R\beta = r$ or:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 3 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}.$$

So R is $G \times K$, β is $K \times 1$ and r is $G \times 1$. Assume $\text{rank}(R) = K$. More next time.

6 Lecture 6: February 15, 2005

6.1 More on OLS with Parameter Restrictions

- Recall we wrote our linear restrictions in the form:

$$R\beta = r,$$

where R is $G \times K$ and r is $G \times 1$. R must have full row rank and r must be in the range space of R , but otherwise, we do not impose anything else.

- To determine our optimal estimator for β , call it β^* , we minimize the following:

$$\text{Min}_{\beta^*} \sum u_t^{*2} = \text{Min} (y - X\beta)'(y - X\beta) = \text{Min} y'y - 2\beta^{*'}X'y + \beta^{*'}X'X\beta^*,$$

such that:

$$R\beta^* = r.$$

So we write out our lagrangian:

$$\mathcal{L} = y'y - 2\beta^{*'}X'y + \beta^{*'}X'X\beta^* - 2\lambda'(R\beta^* - r).$$

- FOCs imply:

$$\beta^* = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}\{r - R\hat{\beta}\}.$$

So β^* is the constrained estimator which you can see is equal to the unconstrained estimator, $\hat{\beta}$, plus this other term which is multiplied by $r - R\hat{\beta}$. So if $r - R\hat{\beta} = 0$, $\beta^* = \hat{\beta}$ and the restriction is already satisfied. The further away we are, ie, the more we restrict, the larger is the second term in the expression.

- In practice, we probably wouldn't use the above equation but rather just impose the restriction right in the regression equation and then run an unrestricted regression on the new variables.
- Proposition 10** Denote the unrestricted residuals as $\hat{u} = y - X\hat{\beta}$ and the restricted residuals as:

$$u^* = y - X\beta^* = y - X\hat{\beta} + X\hat{\beta} - X\beta^* = \hat{u} + X(\hat{\beta} - \beta^*).$$

Thus,

$$\begin{aligned} u^{*'}u^* &= [\hat{u} + X(\hat{\beta} - \beta^*)]'[\hat{u} + X(\hat{\beta} - \beta^*)]. \\ \underbrace{u^{*'}u^*}_{ESS_R} &= \underbrace{\hat{u}'\hat{u}}_{ESS_U} + \underbrace{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}_{\geq 0}. \end{aligned}$$

Thus (following intuition), the Restricted Error Sum of Squares is at least as big as the Unrestricted Error Sum of Squares.

6.2 Statistical Background

- A few facts from EC623. If X is an $n \times 1$ vector $\sim N(\mu, \Sigma)$, then $Z = AX + C \sim N(A\mu + C, A\Sigma A')$.
- The marginals of a multivariate normal are also normal.
- For normally distributed random variables, independence \Leftrightarrow zero-covariance.
- Quadratic forms of standard normals are chi-squared. Thus, if Z is standard normal,

$$\frac{Z'AZ}{\sigma^2} = \eta' A \eta \sim \chi^2(r),$$

with $\eta = Z/\sigma \sim N(0, I)$.

- Linear forms and quadratic forms in X are stochastically independent. Two quadratic forms in X are also stochastically independent if $BA = 0$.

6.3 Unbiasedness and Efficiency

- Consider a random sample, z_1, \dots, z_n , with distribution $F(z_1, \dots, z_n; \theta)$. Consider an estimator:

$$\hat{\theta} = \hat{\theta}(z_1, \dots, z_n).$$

Where $\theta \in \Theta$.

- Then we say that $\hat{\theta}$ is unbiased if:

$$E_{\theta}[\hat{\theta}] = \int \hat{\theta}(z_1, \dots, z_n) \underbrace{dF(z_1, \dots, z_n; \theta)}_{f(z_1, \dots, z_n, \theta) dz_1 dz_2 \dots dz_n} = \theta \quad \forall \theta \in \Theta.$$

- What about efficiency? One thing we could do is compare the efficiency between two estimators. Consider the variance/covariance matrices of two unbiased estimators:

$$\Sigma_{\hat{\theta}} = E[(\hat{\theta} - E[\hat{\theta}])(\hat{\theta} - E[\hat{\theta}])'] = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'].$$

$$\Sigma_{\tilde{\theta}} = E[(\tilde{\theta} - E[\tilde{\theta}])(\tilde{\theta} - E[\tilde{\theta}])'] = E[(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)'].$$

We say that $\hat{\theta}$ is efficient relative to $\tilde{\theta}$ if and only if $\Sigma_{\tilde{\theta}} - \Sigma_{\hat{\theta}}$ is a POSITIVE semidefinite matrix for all $\theta \in \Theta$.

- If θ is simply one parameter, we are just saying that the estimator with the lowest variance is the most efficient. For many parameters, we also include this criteria but make considerations about the covariance terms as well.
- Suppose we want to estimate a linear combination of the θ 's. So,

$$\phi = c'\theta = c_1\theta_1 + \dots + c_m\theta_m.$$

Two estimators might be:

$$\begin{aligned}\hat{\phi} &= c'\hat{\theta}, \\ \tilde{\phi} &= c'\tilde{\theta}.\end{aligned}$$

So,

$$\begin{aligned}\text{Var}(\hat{\phi}) &= E[(\hat{\phi} - \phi)^2] \\ &= E[(c'\hat{\theta} - c'\theta)^2] \\ &= E[\underbrace{(c'(\hat{\theta} - \theta))^2}_{\text{scalar}}] \\ &= E[(c'(\hat{\theta} - \theta))(c'(\hat{\theta} - \theta))'] \\ &= E[(c'(\hat{\theta} - \theta))((\hat{\theta} - \theta)'c)] \\ &= c'E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)']c \\ &= c'\Sigma_{\hat{\theta}}c\end{aligned}$$

Similarly,

$$\text{Var}(\tilde{\phi}) = c'\Sigma_{\tilde{\theta}}c.$$

So, $\hat{\phi}$ is efficient relative to $\tilde{\phi}$ if:

$$\text{Var}(\tilde{\phi}) - \text{Var}(\hat{\phi}) = c'\Sigma_{\tilde{\theta}}c - c'\Sigma_{\hat{\theta}}c = c'(\Sigma_{\tilde{\theta}} - \Sigma_{\hat{\theta}})c \geq 0.$$

- **Definition:** $\hat{\beta} = (X'X)^{-1}X'y = Ay$ is BLUE, Best Linear Unbiased Estimator, if it is efficient for all A .

6.4 Cramer-Rao Lower Bound

- Is there a lower bound to the variance of an unbiased estimator ? Yes.
- Suppose $z = (z_1, \dots, z_n)'$. Consider some parameter that is a function of the z 's with:

$$E[\hat{\theta}(z)] = \int \hat{\theta}(z) \underbrace{L(z, \theta)}_{\text{density}} dz = \theta \quad \forall \theta \in \Theta,$$

so $\hat{\theta}$ is unbiased.

- Assume $\int L(z, \theta) dz = 1$.
- More next time.

7 Lecture 7: February 17, 2005

7.1 Cramer-Rao Inequality

- Cramer-Rao provides a lower bound on the variance of an unbiased estimator.
- Consider an estimator, $\hat{\theta} = \hat{\theta}(Z)$, with likelihood, $L(Z, \theta)$. If integration and differentiation are interchangeable, we have three results:
 - (1) The gradient of the likelihood function is zero:

$$E \left[\frac{\partial \log L(Z, \theta)}{\partial \theta} \right] = 0.$$

- (2) Fisher information matrix, J :

$$\underbrace{-E \left[\frac{\partial^2 \log L(Z, \theta)}{\partial \theta \partial \theta'} \right]}_{J(\theta)} = E \left[\frac{\partial \log L(Z, \theta)}{\partial \theta'} \frac{\partial \log L(Z, \theta)}{\partial \theta} \right].$$

So on the right we have the variance/covariance matrix of the gradient of the likelihood function which must be positive semidefinite, and on the left we have the negative of the hessian. Since when we do ML estimation, we maximize, it makes sense the matrix of SOC's should be negative semidefinite, so clearly the negative of the Hessian should be positive semidefinite.

- (3) The covariance between $\hat{\theta}$ and the gradient is equal to the identity matrix:

$$\int \hat{\theta}(Z) \frac{\partial L(Z, \theta)}{\partial \theta} dZ = \int \hat{\theta}(Z) \frac{\partial \log L(Z, \theta)}{\partial \theta} L(Z, \theta) dZ = E \left[\hat{\theta}(Z) \frac{\partial \log L(Z, \theta)}{\partial \theta} \right] = I.$$

This follows because:

$$\int \hat{\theta}(Z) L(Z, \theta) dZ = \theta,$$

because $\hat{\theta}$ is unbiased. And then differentiate with respect to θ to get the result in (3).

- **Proposition** So Cramer-Rao says that suppose $\hat{\theta}$ is unbiased and suppose the three conditions above hold (since differentiability and integration are interchangeable). Also assume that $L(Z, \theta)$ is C^2 . Let $\Sigma_{\hat{\theta}}(\theta)$ be the variance/covariance matrix of $\hat{\theta}$. Assuming $J(\theta)$ is non-singular:

$$\Sigma_{\hat{\theta}}(\theta) - J^{-1}(\theta) \geq 0, \text{ Positive Semidefinite.}$$

So the inverse of J , the Fisher information matrix, is a lower bound on the variance/covariance matrix of $\hat{\theta}$, an unbiased estimator.

- **Proposition** Assume the model is $y = X\beta + u$ with $u \sim N(0, \sigma^2 I)$. Assume X has full rank. Then the ML estimators are given by:

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

$$\tilde{\sigma}^2 = \frac{1}{T} \hat{u}'\hat{u}.$$

Note we are now assuming NORMALITY! The OLS estimator for the variance is going to be $\hat{\sigma}^2 = \frac{1}{T-K} \hat{u}'\hat{u}$. We will show that $\hat{\sigma}^2$ is unbiased while the ML estimate is biased. We also know that both the ML estimator and the OLS estimator for the variance does NOT attain the Cramer-Rao lower bound, while the OLS and ML estimator for β , both $\hat{\beta}$, does attain the lower bound. What is the lower bound? It's this:

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

- Under normality, we have the model:

$$y = X\beta + u, \quad u \sim N(0, \sigma^2 I).$$

Then:

$$y \sim N(X\beta, \sigma^2 I).$$

To find the J inverse matrix above, right out the likelihood function (multivariate normal) and differentiate twice!

- Proof that the J matrix is a lower bound. Consider the following variance/covariance matrix:

$$\Theta = VC \left[\begin{array}{c} \hat{\theta} \\ \frac{\partial \log L(Z, \theta)}{\partial \theta'} \end{array} \right] = \begin{bmatrix} VC(\hat{\theta}) & Cov(\hat{\theta}, \frac{\partial \log L}{\partial \theta}) \\ Cov(\hat{\theta}, \frac{\partial \log L}{\partial \theta'}) & VC(\frac{\partial \log L}{\partial \theta'}) \end{bmatrix} = \begin{bmatrix} \Sigma_{\hat{\theta}}(\theta) & I \\ I & J \end{bmatrix} \geq 0.$$

Where the last matrix comes from our three results above and since the variance/covariance matrix positive semidefinite, Θ is positive semidefinite. Now consider:

$$[I \quad -J^{-1}] \cdot \Theta \cdot [I \quad -J^{-1}]' = \Sigma_{\hat{\theta}}(\theta) - J^{-1}(\theta) \geq 0.$$

Again since Θ is positive semidefinite, this signs our result. Since the result is positive or 0, we have shown that the J inverse matrix clearly is a lower bound for the variance/covariance matrix of our unbiased estimator. Well, maybe not clearly!

7.2 Small Sample Properties of the OLS Estimators

- Here, we do not necessarily assume normality. We do assume the following:

- (A1) $E(u) = 0$.
- (A2) $E[uu'] = \sigma^2 I$.
- (A3) X is non-stochastic and has full column rank.

- Our model is $y = X\beta + u$ where the distribution of u is not necessarily normal.

Properties of $\hat{\beta}$

- **Proposition 1** $\hat{\beta}$ is an unbiased estimate for β .

Proof:

$$E[\hat{\beta}] = E[(X'X)^{-1}X'y] = (X'X)^{-1}X'E[X\beta + u] = \beta.$$

Note we need that the X 's are non-stochastic and that the $E[u] = 0$ to complete this proof.

- **Proposition 2** Let $\tilde{\beta}$ be any solution to the normal equations. Then $X\tilde{\beta}$ is an unbiased estimator for $X\beta$ even if the $Rank(X)$ is less than K . Thus the systematic component is unbiased.

- **Proposition 3** The variance/covariance matrix of $\hat{\beta}$. Note that $\hat{\beta} - \beta = (X'X)^{-1}X'u$. So,

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

So this depended crucially on A2. If we didn't have A2, we would get:

$$VC(\hat{\beta}) = (X'X)^{-1}X'\Omega X(X'X)^{-1},$$

where $\Omega = E[uu']$.

- **Corollary** So if $u \sim N(0, \sigma^2 I)$, then:

$$y \sim N(X\beta, \sigma^2 I),$$

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}).$$

- **Proposition 4** The Gauss Markov theorem. Given A1-A3 are satisfied, the OLS estimator, $\hat{\beta}$ is the best linear unbiased estimator (BLUE) for β .

Proof: Suppose $\tilde{\beta} = Cy$ is some linear unbiased estimator. Note that for OLS, $C = (X'X)^{-1}X'$. Let D be the difference between C and the OLS "C" term:

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

Or,

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

Thus,

$$E[\tilde{\beta}] = E[Cy] = CE[y] = CX\beta = [(X'X)^{-1}X' - D]X\beta = \beta - DX\beta.$$

Thus since $\tilde{\beta}$ is unbiased, $D\tilde{\beta} = 0$.

Now consider the variance/covariance matrix of $\tilde{\beta}$. Note that:

$$\tilde{\beta} = Cy = C[X\beta + u] = CX\beta + Cu = [(X'X)^{-1}X' - D]X\beta + Cu = \beta + Cu.$$

Thus,

$$\begin{aligned} VC(\tilde{\beta}) &= E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'] \\ &= E[(Cu)(Cu)'] \\ &= CE[uu']C' \\ &= \sigma^2((X'X)^{-1}X' - D)((X'X)^{-1}X' - D)' \\ &= \sigma^2((X'X)^{-1}X' - D)(X(X'X)^{-1} - D)' \\ &= \sigma^2[(X'X)^{-1}X'X(X'X)^{-1} + DD' + (X'X)^{-1}\underbrace{X'D'}_0 - \underbrace{DX(X'X)^{-1}}_0] \\ &= \sigma^2[(X'X)^{-1} + DD'] \\ &= VC(\hat{\beta}) + \sigma^2DD' \end{aligned}$$

Thus,

$$VC(\tilde{\beta}) - VC(\hat{\beta}) = \sigma^2DD' \geq 0.$$

Thus $\hat{\beta}$ satisfies the 'best' part of BLUE. It has a lower variance than any other linear unbiased estimator.

8 Lecture 8: February 22, 2005

8.1 Small Sample Properties of the OLS Estimators

Properties of $\hat{\sigma}^2$

- Denote:

$$E[u_t^2] = \sigma^2.$$

- How should we estimate σ^2 ? If we know the u 's, then:

$$\hat{\sigma}^2 = \frac{1}{T} \sum u_t^2$$

is the natural estimator. Then:

$$E[\hat{\sigma}^2] = \frac{1}{T} \sum E[u_t^2] = \frac{1}{T} T \sigma^2 = \sigma^2.$$

And,

$$\text{Var}(\hat{\sigma}^2) = \frac{1}{T^2} \sum \text{Var}(u_t^2) = \frac{T(\mu_4 - \sigma^4)}{T^2} = \frac{\mu_4 - \sigma^4}{T}.$$

Where does this last equality come from? Denote $E[u_t^4] = \mu_4$. Then:

$$\text{Var}(u_t^2) = E[(u_t^2 - E(u_t^2))^2] = E[u_t^4] - [E(u_t^2)]^2 = \mu_4 - [\sigma^2]^2 = \mu_4 - \sigma^4.$$

- If we do NOT know the u 's, then:

$$s^2 = \frac{1}{T - K} \sum \hat{u}_t^2$$

is our natural estimator. The reason we divide by $T - K$ will be apparent soon. We can rewrite \hat{u} as follows:

$$\hat{u} = y - X\hat{\beta} = (I - X(X'X)^{-1}X')y = My = M(X\beta + u) = \underbrace{MX}_0\beta + Mu = Mu.$$

So,

$$s^2 = \frac{\hat{u}'\hat{u}}{T - K} = \frac{u'Mu}{T - K}$$

- **Proposition 5** $E[s^2] = \sigma^2$.

Proof:

$$\begin{aligned} E[u'Mu] &= E[\text{tr}(u'Mu)] = E[\text{tr}(Mu u')] = \text{tr}[E(Mu u')] = \text{tr}[ME(uu')] = \\ &= \text{tr}[M\sigma^2 I] = \sigma^2 \text{tr}[M] = \sigma^2(T - K). \end{aligned}$$

So,

$$E[s^2] = E\left[\frac{u'Mu}{T-K}\right] = \sigma^2,$$

unbiased.

- **Proposition 6** Given $u \sim N(0, \sigma^2 I)$. Then,

$$\frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2(T-K).$$

Proof:

$$\frac{\hat{u}'\hat{u}}{\sigma^2} = \frac{u'Mu}{\sigma^2} = \underbrace{\left(\frac{u}{\sigma}\right)'}_{Std\ Norm} M \underbrace{\left(\frac{u}{\sigma}\right)}_{Std\ Norm} \sim \chi^2(T-K).$$

This follows since a quadratic form in standard normals with M , a symmetric, idempotent matrix, results in a chi-squared with degrees of freedom equal to: $rank(M) = tr(M) = T - K$.

8.2 Prediction

- Suppose we have the following model:

$$y_t = a + bx_t + u_t, \quad t = 1, \dots, T.$$

And suppose we have S future (out of sample) values of x_t that we would like to look at and determine both the expected value of y and the estimate of y itself. Thus our future relationship might be:

$$y_0 = X_0\beta + u_0.$$

Then,

$$E[y_0] = X_0\beta.$$

Then a natural estimator for $E[y_0]$ of course is:

$$y_0^p = X_0\hat{\beta}.$$

So clearly,

$$E[y_0^p] = X_0\beta.$$

Our guess for $y_0 = X_0\beta + u_0$ would also be $X_0\hat{\beta}$ but only if we can make the following assumptions:

$$E[u] = 0, \quad E[u_0] = 0, \quad E[uu'] = \sigma^2 I_T, \quad E[u_0u_0'] = \sigma^2 I_S, \quad E[uu_0'] = 0.$$

Thus,

$$E[y_0] = y_0^p = X_0\hat{\beta}.$$

- So now consider properties of y_0^p . We have shown $E[y_0^p] = X_0\beta$. Also,

$$\begin{aligned}
VC(y_0^p) &= E[(y_0^p - E(y_0^p))(y_0^p - E(y_0^p))'] \\
&= E[(X_0\hat{\beta} - X_0\beta)(X_0\hat{\beta} - X_0\beta)'] \\
&= E[X_0(\hat{\beta} - \beta)(X_0(\hat{\beta} - \beta))'] \\
&= E[X_0(\hat{\beta} - \beta)(\hat{\beta} - \beta)'X_0'] \\
&= X_0\sigma^2(X'X)^{-1}X_0' \\
&= \sigma^2X_0(X'X)^{-1}X_0'
\end{aligned}$$

- How about the error in our prediction? Consider:

$$v_0 = y_0 - y_0^p = y_0 - X_0\hat{\beta} = X_0\beta + u_0 - X_0\hat{\beta} = u_0 - X_0(\hat{\beta} - \beta).$$

So,

$$E[v_0] = E[u_0 - X_0(\hat{\beta} - \beta)] = 0.$$

Also, since u_0 are the errors from the out of sample observations and $\hat{\beta}$ is estimated using the original sample, u_0 and $X_0(\hat{\beta} - \beta)$ are independent. Thus,

$$\begin{aligned}
VC(v_0) &= VC(u_0) + VC(X_0(\hat{\beta} - \beta)) \\
&= \sigma^2I_s + X_0VC(\hat{\beta} - \beta)X_0' \\
&= \sigma^2I_s + X_0\sigma^2(X'X)^{-1}X_0' \\
&= \sigma^2(I_s + X_0(X'X)^{-1}X_0')
\end{aligned}$$

So the variance of the prediction error comes from the variance of the out of sample observation as well as the error in estimating the original coefficient.

- Consider the simple regression:

$$y_t = a + bx_t + u_t,$$

and consider estimating a single out of sample observation x_0 . Then, it can be shown – I tried and failed – that:

$$\sigma_{v_0}^2 = \sigma^2\left[1 + \frac{1}{T} + \frac{x_0 - \bar{x}}{\sum(x_t - \bar{x})^2}\right].$$

See G-8.1. It is clear from the formula that more observations in sample will reduce the variability of our prediction error. Also, the further x_0 is from \bar{x} , the worse we do, ie the more out of sample we try to predict, the more possibility of error.

- **Proposition 9** The estimator of $y_0^p = X_0\hat{\beta}$ is BLUE. Ie,

$$y_0^p = X_0\hat{\beta} = X_0(X'X)^{-1}X'y = Py,$$

is linear in y and $P = X_0(X'X)^{-1}X'$ is the best (smallest variance) estimator.

8.3 Hypothesis Testing

- Assume A1-A3 holds plus:

$$(A4) : u \sim N(0, \sigma^2 I).$$

- **Lemma** Given A1-A4, $y \sim N(X\beta, \sigma^2 I)$ and $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$.
- Hypothesis:

$$H_0 : R\beta = r, \text{ versus } H_1 : R\beta \neq r.$$

- Recall:

$$ESS_R = ESS_U + (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r).$$

- **Proposition** F-Test. Suppose R is $G \times K$ and $\text{rank}(R) = G$. r is $G \times 1$ and $\hat{\beta}$ is our OLS estimator. Then under the null that $R\beta = r$,

$$F = \frac{(ESS_R - ESS_U)/G}{ESS_U/(T - K)} = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/G}{(y - X\hat{\beta})'(y - X\hat{\beta})/(T - K)} \sim F(G, T - K).$$

Proof: Under the null,

$$\begin{aligned} R\hat{\beta} - r &= R(X'X)^{-1}X'y - r \\ &= R(X'X)^{-1}X'(X\beta + u) - r \\ &= R\beta + R(X'X)^{-1}X'u - r \\ &= R(X'X)^{-1}X'u + \underbrace{R\beta - r}_{=0 \text{ under } H_0} \\ &= R(X'X)^{-1}X'u \end{aligned}$$

So consider the numerator of our test statistic:

$$\begin{aligned} ESS_R - ESS_U &= (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) \\ &= (R(X'X)^{-1}X'u)'[R(X'X)^{-1}R']^{-1}(R(X'X)^{-1}X'u) \\ &= \underbrace{u'X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'u}_A \\ &= u'Au \end{aligned}$$

It is clear that A is idempotent and symmetric with $tr(A) = rank(A) = G$. Now consider the denominator:

$$\begin{aligned} ESS_U &= \hat{u}'\hat{u} \\ &= u'Mu \end{aligned}$$

So our test statistic is:

$$F = \frac{u'Au/G}{u'Mu/(T-K)}.$$

Further note that:

$$MA = MX(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' = 0,$$

because $MX = 0$. Thus, rewriting our statistic:

$$F = \frac{\frac{u'Au}{\sigma^2}/G}{\frac{u'Mu}{\sigma^2}/(T-K)} = \frac{\chi^2(G)}{\chi^2(T-K)} \sim F(G, T-K).$$

So we have a quadratic form in standard normals on top (divided by the rank of A) which gives us a $\chi^2(G)$ and on the bottom is a quadratic form in standard normals (divided by the rank of M) which gives us a $\chi^2(T-K)$. Since $MA = 0$, the ratio of these chi-squares is $F(G, T-K)$. Nice.

9 Lecture 9: March 1, 2005

9.1 More on Hypothesis Testing

F-tests and *t*-tests

- Recall, under the null of the restricted model:

$$F = \frac{(ESS_R - ESS_U)/G}{ESS_U/(T - K)} \sim F(G, T - K).$$

See G-9.1. If F is large, maybe the restriction is NOT valid so we should reject the null.

- Note that the F -statistic corresponds to other tests of model restrictions including the likelihood ratio test, the Lagrange multiplier test, and the Wald test.
- **Proposition 3** Test of significance of β_i . Under $H_0 : \beta_i = \beta_i^0$, the so called “t-ratio” is:

$$t = \frac{\hat{\beta}_i - \beta_i^0}{s_{\hat{\beta}_i}} \sim t(T - K).$$

Or,

$$t = \frac{\hat{\beta}_i - \beta_i^0}{\sqrt{s^2(X'X)^{ii}}} \sim t(T - K),$$

where $s^2 = \hat{u}'\hat{u}/(T - K)$.

- Note that $F = t^2$ where F is the F -stat for $\beta_i = \beta_i^0$. Why?

$$\begin{aligned} t &= \frac{\hat{\beta}_i - \beta_i^0}{\sqrt{s^2(X'X)^{ii}}} \\ &= \frac{\hat{\beta}_i - \beta_i^0/\sigma\sqrt{(X'X)^{ii}}}{s/\sigma} \\ &= \frac{N(0, 1)}{s/\sigma} \\ &= \frac{N(0, 1)}{\sqrt{\hat{u}'\hat{u}/(\sigma^2(T - K))}} \\ &= \frac{N(0, 1)}{\sqrt{u'Mu/(\sigma^2(T - K))}} \\ &= \frac{N(0, 1)}{\sqrt{\chi^2(T - K)}} \sim t(T - K) \end{aligned}$$

The last line follows from the fact that the numerator and denominator are independent. Thus $t^2 = F(1, T - K)$.

- See G-9.2. Given the hypothesis $H_0 : \beta_i = \beta_i^0$ versus $H_1 : \beta_i \neq \beta_i^0$, we reject at the 5% level if:

$$|t| \geq t_{0.975},$$

or,

$$|\hat{\beta}_i - \beta_i| \geq t_{0.975} * s_{\hat{\beta}_i},$$

or,

$$\beta_i^0 \notin (\hat{\beta}_i \pm t_{0.975} s_{\hat{\beta}_i}).$$

- **Definition:** The p -value is defined as:

$$p = Prob\{|t| \geq t_{obs}\}.$$

- Rule of thumb. Accept if $t > 2$. Note for $N > 30$, $t \approx Z$ so $t_{0.975} = Z_{0.975} = 1.96$.
- Only do one-sided tests if you know for sure the coefficient should be positive, negative, or on some side of a constant. One-sided tests are more efficient.
- **Proposition 4** Test of the existence of a relationship. Consider the model:

$$y_t = \beta_1 + \beta_2 x_{2t} + \dots + \beta_K x_{tK} + u_t.$$

Hypotheses:

$$H_0 : \beta_2 = \dots = \beta_K = 0, \text{ versus } H_1 : \text{at least one coefficient not zero.}$$

Thus the restricted model would be:

$$y_t = \beta_1 + u_t \implies \hat{\beta}_{1R} = \bar{y}, u_R = y - e_T \bar{y}.$$

Thus the restricted sum of squared error:

$$ESS_R = u_R' u_R = \sum (y_t - \bar{y})^2 = TSS.$$

And the unrestricted sum of squared error:

$$ESS_U = \hat{u}' \hat{u} = ESS.$$

Note $TSS = RSS + ESS$. So consider the following statistic:

$$\begin{aligned}
F &= \frac{(ESS_R - ESS_U)/G}{ESS_U/(T - K)} \\
&= \frac{(ESS_R - ESS_U)/(TSS * G)}{ESS_U/(TSS * (T - K))} \\
&= \frac{(TSS - ESS)/(TSS * G)}{ESS/(TSS * (T - K))} \\
&= \frac{(1 - \frac{ESS}{TSS})/G}{\frac{ESS}{TSS}/(T - K)} \\
&= \frac{R^2/G}{(1 - R^2)/(T - K)}
\end{aligned}$$

And this is the F -test for a regression relationship.

- **Proposition 5** Chow Test for a Structural Break. Suppose we think there might be a break in the data such that the coefficients might change in the two subsets. (Say pre-war and post-war time series data). Consider two regressions:

$$y^1 = X^1\beta + u^1, \quad (T_1 \text{ observations}).$$

$$y^2 = X^2\beta^* + u^2, \quad (T_2 \text{ observations}).$$

So $T_1 + T_2 = T$ and $T_1, T_2 \geq K$. So we are testing the following hypothesis:

$$H_0 : \beta = \beta^*, \text{ versus } H_1 : \beta \neq \beta^*.$$

We can write the unrestricted regressions as:

$$\begin{bmatrix} y^1 \\ y^2 \end{bmatrix} = \begin{bmatrix} X^1 & 0 \\ 0 & X^2 \end{bmatrix} \begin{bmatrix} \beta \\ \beta^* \end{bmatrix} + \begin{bmatrix} u^1 \\ u^2 \end{bmatrix}.$$

These two regressions yield ESS_1 and ESS_2 such that $ESS_U = ESS_1 + ESS_2$. The restricted regression could be written:

$$\begin{bmatrix} y^1 \\ y^2 \end{bmatrix} = \begin{bmatrix} X^1 \\ X^2 \end{bmatrix} \beta + \begin{bmatrix} u^1 \\ u^2 \end{bmatrix}.$$

This regression yields ESS_R . Given these two regressions, consider the following test:

$$F = \frac{[ESS_R - (ESS_1 + ESS_2)]/K}{[ESS_1 + ESS_2]/(T - 2K)} \sim F(K, T - 2K).$$

So if F is large, reject H_0 and conclude there must be a structural break.

- A note on the Chow test when $T_2 \leq K$. In this case, our second restricted regression is NOT identified. Thus $ESS_2 = 0$. Essentially, we are only estimating T_2 parameters in the second regression so there should be only T_2 constraints. Also, for the unrestricted ESS , we have T observations less K parameters in the first regression and T_2 parameters in the second, thus, we should be dividing by $T - K - T_2 = T_1 - K$. So the statistic becomes:

$$F^* = \frac{[ESS_R - ESS_1]/T_2}{ESS_1/(T_1 - K)} \sim F(T_2, T_1 - K).$$

10 Lecture 10: March 3, 2005

10.1 Asymptotic Theory

- So far, we have been assuming normality of the disturbances to do hypothesis testing. Is it necessary? Consider an estimator:

$$\hat{\theta} = h(Y_1, \dots, Y_n),$$

where $h(\cdot)$ is a non-linear function. To determine properties of $\hat{\theta}$, we need to know something about the distribution of the Y 's. We would like to have a distribution free way of considering properties of our estimators.

- Two ideas will be considered: Consistency and Asymptotic normality. It is *unanimously* accepted that consistency is a fairly minimal restriction to place on our estimator.

- Suppose:

$$\hat{\theta}_n = \frac{1}{n} \sum_i Y_i, \quad Y_i \sim iid(\theta, \sigma^2).$$

Thus,

$$E[\hat{\theta}_n] = \theta \dots \text{consistent.}$$

$$Var(\hat{\theta}_n) = \frac{\sigma^2}{n}.$$

- We also might consider the distribution function of our estimator:

$$G_n(y) = Pr\{\hat{\theta}_n \leq y\}.$$

Since the variance of our estimator goes to zero as $n \rightarrow \infty$, the distribution $G_n(y)$ will become degenerate at θ . See G-10.1.

- Denote:

$$Z_n = \sqrt{n}(\hat{\theta}_n - \theta).$$

Thus:

$$E[Z_n] = 0, \quad Var[Z_n] = \sigma^2.$$

By scaling our estimator by just the right amount, we can get a new random variable (estimator) that will have a non-zero and non-explosive variance in the limit.

- Now consider the distribution of Z_n :

$$F_n(Z) = Pr(Z_n \leq Z).$$

$F_n(Z)$ will NOT collapse as n gets large and it will be shown (via the Central Limit Theorem):

$$F_n(Z) \rightarrow F(Z) \text{ where, } F(Z) \rightsquigarrow N(0, \sigma^2).$$

- Thus, for finite n , (since in the limit, it is exactly normal)

$$Z_n \approx N(0, \sigma^2).$$

Solving Z_n for $\hat{\theta}_n$:

$$\hat{\theta}_n = \theta + \frac{1}{\sqrt{n}} Z_n.$$

Thus,

$$\hat{\theta}_n \approx N\left(\theta, \frac{\sigma^2}{n}\right).$$

Finally, if $Y_i \sim iid N(\theta, \sigma^2)$, then $\hat{\theta}_n \sim N\left(\theta, \frac{\sigma^2}{n}\right)$ exactly.

Modes of Convergence

- Note that estimators are always random variables so we will usually talk about random variables (or vectors) but the analysis also applies to estimators. Let Z_n be a sequence of real numbers such that:

$$\lim_{n \rightarrow \infty} Z_n = Z.$$

What does this mean? Precisely,

$$\forall \epsilon > 0, \exists N_\epsilon \ni |Z_n - Z| \leq \epsilon \forall n \geq N_\epsilon.$$

- For random variables, we say that the random variable converges if some sequence of reals converge.
- **Definition:** Convergence in Probability. A sequence of random variables, Z_n , (weakly) converges in probability to a random variable, Z , iff:

$$\lim_{n \rightarrow \infty} Pr\{|Z_n - Z| \leq \epsilon\} = 1 \forall \epsilon > 0.$$

Also denoted:

$$plim_{n \rightarrow \infty} Z_n = Z.$$

$$Z_n \xrightarrow{p} Z.$$

- **Definition:** Almost Sure Convergence. A sequence of random variables, Z_n , almost surely converges to a random variable, Z , iff:

$$Pr(\{\omega \in \Omega : \lim_{n \rightarrow \infty} Z_n(\omega) = Z(\omega)\}) = 1.$$

Also denoted:

$$Z_n \xrightarrow{as} Z.$$

- **Theorem:** A sequence of random variables, Z_n , almost surely converges to a random variable, Z , iff:

$$\lim_{n \rightarrow \infty} Pr\{|Z_i - Z| \leq \epsilon \forall i \geq n\} = 1 \forall \epsilon > 0.$$

So this is an alternate definition for almost sure convergence and from this you can see that it is a stronger condition than convergence in probability.

- **Definition:** Convergence in the r^{th} Mean. A sequence of random variables, Z_n , converges in the r^{th} mean to a random variable, Z , iff:

$$\lim_{n \rightarrow \infty} E[|Z_n - Z|^r] = 0.$$

Also denoted:

$$Z_n \xrightarrow{r\text{-th}} Z.$$

If $r = 2$, we say that the sequence converges in the quadratic mean.

- **Theorem:** If $Z_n \xrightarrow{r\text{-th}} Z$ for some $r > 0$, then $Z_n \xrightarrow{p} Z$.

Proof: Consider the following:

$$0 \leq \Pr(|Z_n - Z| \geq \epsilon) = \Pr(|Z_n - Z|^r \geq \epsilon^r) \leq \frac{E[(Z_n - Z)^r]}{\epsilon^r}.$$

Take limits:

$$0 \leq \lim \Pr(|Z_n - Z| \geq \epsilon) = \lim \Pr(|Z_n - Z|^r \geq \epsilon^r) \leq \underbrace{\frac{\lim E[(Z_n - Z)^r]}{\lim \epsilon^r}}_{0 \text{ by ass.}}$$

$$0 \leq \lim \Pr(|Z_n - Z| \geq \epsilon) \leq 0.$$

Thus,

$$\Pr(|Z_n - Z| \geq \epsilon) = 0 \implies Z_n \xrightarrow{p} Z.$$

- **Corollary** If $E[Z_n] \rightarrow c$ and $\text{Var}(Z_n) \rightarrow 0$, then $Z_n \xrightarrow{p} c$.

Proof: Note that we can write $E[(Z_n - c)^2] = E[(Z_n - E[Z_n])^2] + (E[Z_n] - c)^2$, or the MSE equals variance plus square of the bias. This must go to zero by assumption. Thus,

$$0 \leq \Pr(|Z_n - c| \geq \epsilon) = \Pr(|Z_n - c|^2 \geq \epsilon^2) \leq \frac{E[(Z_n - c)^2]}{\epsilon^2}.$$

Taking limits:

$$\Pr(|Z_n - c| \geq \epsilon) = 0 \implies Z_n \xrightarrow{p} c.$$

- So, we have the following implications:

$$Z_n \xrightarrow{as} Z \implies Z_n \xrightarrow{p} Z.$$

$$Z_n \xrightarrow{r\text{-th}} Z \implies Z_n \xrightarrow{p} Z.$$

$$Z_n \xrightarrow{p} Z \not\Rightarrow Z_n \xrightarrow{r\text{-th}} Z.$$

$$Z_n \xrightarrow{as} Z \not\Rightarrow Z_n \xrightarrow{r\text{-th}} Z.$$

- So far, we have been dealing with a single random variable. Suppose now that Z_n is a k dimensional random vector and Z is k dimensional. Then convergence in probability

could be written one of two ways:

$$Pr(\underbrace{|Z_n - Z|}_{\text{euclidean dist}} \leq \epsilon) = 1 \quad \forall \epsilon > 0.$$

$$Pr(|Z_n^{(i)} - Z^{(i)}| \leq \eta) = 1 \quad \forall \eta > 0, \quad i = 1, \dots, k.$$

Where the euclidean distance is as usual:

$$d(x, y) = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}.$$

Convergence of OLS Estimators

- To do this example, we need some results which will be studied later in the course:
 - (1) $plim (W_n + V_n) = plim W_n + plim V_n$.
 - (2) $plim (W_n * V_n) = plim W_n * plim V_n$.
 - (3) $plim (W_n/V_n) = plim W_n/plim V_n$ if $plim V_n \neq 0$.
 - (4) $plim g(W_n, V_n) = g(plim W_n, plim V_n)$ if $g(\cdot)$ is continuous at the probability limits. Note how this is different from the expectations operator which you can't bring inside. A nice advantage of plims.
- Consider the regression model:

$$y_t = ax_t + u_t, \quad u_t \sim iid(0, \sigma^2).$$

Assume the x 's are non-stochastic and:

$$\frac{1}{T} \sum x_t^2 \rightarrow q.$$

Thus, the OLS estimator for a is:

$$\hat{a} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum x_t (ax_t + u_t)}{\sum x_t^2} = a + \frac{\sum x_t u_t}{\sum x_t^2} = a + \frac{1/T \sum x_t u_t}{1/T \sum x_t^2}.$$

Take the probability limit:

$$\begin{aligned} plim \hat{a} &= plim a + plim \frac{1/T \sum x_t u_t}{1/T \sum x_t^2} \\ &= a + \frac{plim 1/T \sum x_t u_t}{plim 1/T \sum x_t^2} \\ &= a + \frac{plim 1/T \sum x_t u_t}{q} \end{aligned}$$

Now let $\psi_n = 1/T \sum x_t u_t$. Thus, $E[\psi_n] = 0$ and:

$$\text{Var}(\psi_n) = \frac{1}{T^2} \sum \text{Var}(x_t u_t) = \frac{1}{T^2} \sum x_t^2 \sigma^2 = \frac{\sigma^2}{T} \frac{1}{T} \sum x_t^2.$$

So as $T \rightarrow \infty$,

$$\lim_{T \rightarrow \infty} \text{Var}(\psi_n) = \lim_{T \rightarrow \infty} \underbrace{\frac{\sigma^2}{T}}_0 \underbrace{\frac{1}{T} \sum x_t^2}_q = 0.$$

Thus, by our previous corollary, $\text{plim } \psi_n = 0$. So,

$$\text{plim } \hat{a} = a + \frac{0}{q} = a.$$

So we say that \hat{a} is a consistent estimator for a .

11 Lecture 11: March 8, 2005

11.1 Convergence in Distribution

- Motivation. Suppose $\hat{\theta}_n \rightarrow^p \theta$. Let $G_n(z) = Pr(\hat{\theta}_n \leq z)$, the cumulative distribution function of $\hat{\theta}_n$. Then if $\hat{\theta}_n$ converges in probability to θ , we can think of the CDF and PDF acting as shown in G-11.1. The density will peak at θ and the cumulative density will become a degenerate step function at θ . Thus the limiting distribution is:

$$G(z) = \begin{cases} 0 & z < \theta \\ 1 & z > \theta \end{cases}$$

Thus, $\lim_{n \rightarrow \infty} G_n(z) = G(z)$ for $z > \theta$ and $z < \theta$. However,

$$\lim G_n(z) \neq G(z) \text{ for } z = \theta.$$

Why? Because, at this point, $G(z)$ jumps from 0 to 1 and we can't say anything about what happens to $G(\theta)$.

- So whenever $\hat{\theta}_n \rightarrow^p \theta$, looking at $G_n(z)$ doesn't give us much because the distribution is degenerate.
- Now consider a transformed random variable:

$$Z_n = \sqrt{n}(\hat{\theta}_n - \theta).$$

Suppose Z_n has distribution function $F_n(z)$ and suppose:

$$F_n(z) \rightarrow F(z) = \Phi(z), \text{ normal.}$$

Then we have a scaled version of our estimator that does NOT degenerate! Trick is choosing the right scaling factor.

- **Definition:** Let F_1, F_2, \dots and F denote CDFs on \mathfrak{R} . Then F_n weakly converges to F if:

$$\lim_{n \rightarrow \infty} F_n(z) = F(z), \forall z \in \mathfrak{R} \text{ that are continuity points of } F.$$

Note that this qualifier about continuity points deals with the situation of degenerate distributions. So if Z_1, Z_2, \dots and Z are random variables with distribution functions F_1, F_2, \dots , and F then we say that Z_n converges in distribution to Z if $F_n(z) \rightarrow F(z)$. Also written:

$$Z_n \rightarrow^d Z.$$

- Results:

Convergence in Probability \implies Convergence in Distribution.

Convergence in Probability $\not\Leftarrow$ Convergence in Distribution.

- Consider an example. (u_t, x_t) are iid sequences, mutually independent. $E[u_t] = 0$. $E[u_t^2] = \sigma^2$. $E[x_t^4] < \infty$. Denote:

$$\hat{\theta}_t = \theta + \frac{\sum x_t u_t}{\sum x_t^2}.$$

Let $V_t = \frac{1}{T} \sum x_t u_t$ and $W_t = \frac{1}{T} \sum x_t^2$. Thus,

$$E[V_t] = 0, \text{ and } Var[V_t] = \frac{E[x_t^2]\sigma^2}{T} \rightarrow 0 \text{ as } T \rightarrow \infty.$$

$$E[W_t] = E[x_t^2], \text{ and } Var[W_t] = \frac{Var[x_t^2]}{T} \rightarrow 0 \text{ as } T \rightarrow \infty.$$

So, from previous results,

$$V_T \xrightarrow{p} 0.$$

$$W_T \xrightarrow{p} E[x_t^2].$$

Now consider a transformed estimator:

$$Z_t = \sqrt{T}(\hat{\theta}_t - \theta).$$

This can be written:

$$Z_t = \frac{1/\sqrt{T} \sum x_t u_t}{1/T \sum x_t^2}.$$

It can be shown that:

$$Z_t \rightarrow^d N\left(0, \frac{\sigma^2}{m_x^2}\right),$$

where $m_x^2 = E[x_t^2]$. Thus,

$$Z_t \approx N\left(0, \frac{\sigma^2}{m_x^2}\right).$$

Now rewrite Z_t for $\hat{\theta}_t$:

$$\hat{\theta}_t = \theta + \frac{1}{\sqrt{T}} Z_t \approx N\left(\theta, \frac{1}{T} \frac{\sigma^2}{m_x^2}\right).$$

What is a good estimator for m_x^2 ? Answer: $1/T \sum x_t^2 = 1/T x'x$. Thus,

$$\hat{\theta}_t \approx N\left(\theta, \sigma^2 (x'x)^{-1}\right).$$

So we say that $\hat{\theta}_t$ is a consistent estimator for θ .

- **Theorem:** Let $c \in \mathfrak{R}$, a constant. Then,

$$Z_n \rightarrow^d c \iff Z_n \rightarrow^p c.$$

So convergence in distribution to a constant is equivalent to convergence in probability to a constant.

Proof (partial). Assume convergence in distribution and show convergence in probability. Write:

$$\begin{aligned}
P(|Z_n - c| > \epsilon) &= P(Z_n - c < -\epsilon) + P(Z_n - c > \epsilon) \\
&= P(Z_n - c < -\epsilon) + (1 - P(Z_n - c < \epsilon)) \\
&= P(Z_n < c - \epsilon) + 1 - P(Z_n < c + \epsilon) \\
&\leq P(Z_n \leq c - \epsilon) + 1 - P(Z_n \leq c + \epsilon) \\
&= F_n(c - \epsilon) - F_n(c + \epsilon) + 1
\end{aligned}$$

Where $F_n(z)$ is the CDF of Z_n , which equals:

$$F_n(z) = \begin{cases} 0 & z < c \\ 1 & z \geq c \end{cases}$$

Thus $c \pm \epsilon$ are continuity points of F . Since $Z_n \rightarrow^d c$,

$$F_n(c - \epsilon) \rightarrow F(c - \epsilon) = 0.$$

$$F_n(c + \epsilon) \rightarrow F(c + \epsilon) = 1.$$

Thus,

$$0 \leq P(|Z_n - c| > \epsilon) \leq F_n(c - \epsilon) - F_n(c + \epsilon) + 1 \rightarrow 0 - 1 + 1 = 0.$$

So $P(|Z_n - c| > \epsilon) = 0$ and $Z_n \rightarrow^p c$. QED.

- **Definition:** Let F_1, F_2, \dots and F denote CDFs on \mathfrak{R}^k . Then F_n converges weakly to F if:

$$\lim_{n \rightarrow \infty} F_n(z) = F(z) \quad \forall z \in \mathfrak{R}^k \text{ that are continuity points of } F.$$

- **Theorem:** Weak convergence of F_n to F implies weak convergence of $F_n^{(i)}$ to $F^{(i)}$ and $Z_n \rightarrow^d Z$ implies $Z_n^{(i)} \rightarrow^d Z^{(i)}$. So convergence of the multivariate distributions implies convergence of the marginals. The opposite is NOT true!
- **Theorem:** Cramer-Wold Device. Let Z_1, Z_2, \dots and Z denote random vectors taking their values in \mathfrak{R}^k . The following are equivalent statements:

- (1) $Z_n \rightarrow^d Z$.
- (2) $\alpha' Z_n \rightarrow^d \alpha' Z, \forall \alpha \in \mathfrak{R}^k$.
- (3) $\alpha' Z_n \rightarrow^d \alpha' Z, \forall \alpha \in \mathfrak{R}^k$ with $\|\alpha\| = 1$.

12 Lecture 12: March 10, 2005

12.1 Cramer-Wold Example

- Consider a_i non-stochastic with $\lim \frac{1}{n} \sum a_i^2 = q > 0$. Let $u_t \sim iid(0, \sigma^2)$. Then by the central limit theorem:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n a_i u_i \rightarrow^d N(0, \sigma^2 q).$$

- Given this, consider X_n , a sequence of non-stochastic $n \times k$ matrices with,

$$\lim \frac{1}{n} X'X = Q, \text{ positive definite.}$$

Let $u = (u_1, \dots, u_n)'$ with $u_i \sim iid(0, \sigma^2)$. Then,

$$\frac{1}{\sqrt{n}} X'u \rightarrow N(0, \sigma^2 Q).$$

- Proof: Consider:

$$Z_n = \frac{1}{\sqrt{n}} X'u,$$

a $k \times 1$ vector, and:

$$Z \sim N(0, \sigma^2 Q).$$

Want to show that $Z_n \rightarrow Z$. Pick an arbitrary α . Thus,

$$\alpha' Z_n = \frac{1}{\sqrt{n}} \alpha' X'u.$$

Note that $X\alpha = a$ is $n \times 1$. Thus,

$$\alpha' Z_n = \frac{1}{\sqrt{n}} a'u = \frac{1}{\sqrt{n}} \sum a_i u_i.$$

Note that:

$$\frac{1}{n} \sum a_i^2 = \frac{1}{n} a'a = \frac{1}{n} \alpha' X'X \alpha = \alpha' \frac{1}{n} X'X \alpha.$$

So,

$$\lim \frac{1}{n} \sum a_i^2 = \alpha' Q \alpha.$$

We need to use Cramer-Wold to show that:

$$\alpha' Z_n \rightarrow^d \alpha' Z \forall \alpha \iff Z_n \rightarrow^d Z.$$

If $\alpha = 0$, this holds trivially. If $\alpha \neq 0$,

$$\lim \frac{1}{n} \sum a_i^2 = \alpha' Q \alpha > 0,$$

So the condition on the a_i 's is satisfied. Thus, since $u_i \sim iid(0, \sigma^2)$,

$$\alpha'Z_n = \frac{1}{\sqrt{n}}a'u \sim N\left(0, \frac{1}{n}a'\sigma^2a\right) \equiv N(0, \sigma^2\alpha'Q\alpha).$$

Also,

$$\alpha'Z \sim N(0, \sigma^2\alpha'Q\alpha).$$

Thus,

$$\alpha'Z_n \rightarrow^d \alpha'Z \forall \alpha \implies Z_n \rightarrow^d Z.$$

This establishes the result. Note that most CLT's are written in scalar form, but we can scale up to any vector or matrix using the Cramer-Wold Device.

12.2 Properties of Limits

- Let W_n , V_n and U_n be random vectors that converge in probability to W , V , and U . We have the following results:

- If $W_n \rightarrow^p W$ and $V_n \rightarrow^p V$ then $Z_n = [W_n, V_n]' \rightarrow^p [W, V]'$.
- $W_n \pm V_n \rightarrow^p W \pm V$.
- $W_n'V_n \rightarrow^p W'V$.
- $W_n/V_n \rightarrow^p W/V$ if $V \neq 0$ with probability 1.

- **Theorem:** Slutsky's theorem. Suppose $Z_n \rightarrow^p Z$. Then,

$$g(Z_n) \rightarrow^p g(Z) \text{ provided } g(\cdot) \text{ is continuous at } Z.$$

Or written another way,

$$plim g(Z_n) = g(plim(Z_n)) \text{ provided } g(\cdot) \text{ is continuous at } Z.$$

This a pretty important result since you usually can't interchange the operators like this. It comes from the calculus result:

$$\lim_{n \rightarrow \infty} g(a_n) = g(\underbrace{plim a_n}_a) \text{ provided } g(\cdot) \text{ is continuous at } a.$$

So if $Z_n = (W_n, V_n, U_n)'$ and $Z = (W, V, U)'$ with $Z_n \rightarrow^p Z$, then:

$$g(Z_n) = 3W_n^2 + 2U_n/V_n \rightarrow^p 3W^2 + 2U/V, \text{ provided } V \neq 0.$$

- Slutsky's theorem also holds for Almost Sure Convergence and Convergence in Distribution. A note on convergence in distribution though. Suppose $Z_n = [W_n, V_n, U_n]' \rightarrow^d Z$. Then,

$$g(Z_n) \rightarrow^d g(Z).$$

But!! :

$$W_n \rightarrow^d W, V_n \rightarrow^d V, U_n \rightarrow^d U \not\Rightarrow Z_n \rightarrow Z.$$

So even though the marginals converge, the multivariate distribution of Z_n need not converge! For *i.p.* and *a.s.* convergence, we're ok, the causation goes both ways, but for convergence in distribution, we can't go from marginals to multivariate.

- If W_n , V_n , and U_n all converge in distribution to constants, then Z_n does necessarily converge in distribution to Z .
- All of these properties apply to random matrices as well as random vectors.
- **Corollary** Important result:

$$plim W_n^{-1} = (plim W_n)^{-1},$$

only if W_n is non-singular.

- **Corollary** Let W_n , V_n be sequences of $k \times 1$ and $l \times 1$ random vectors and let:

$$W_n \rightarrow^d W,$$

$$V_n \rightarrow^p c.$$

Then,

$$W_n + V_n \rightarrow^d W + c, \text{ and } W_n' V_n \rightarrow^d W' c,$$

and specifically for $c = 0$,

$$W_n' V_n \rightarrow^p 0.$$

- Example. Consider:

$$\hat{\beta}_n = \underbrace{\beta}_{V_n} + \underbrace{\left(\frac{1}{n} X' X\right)^{-1}}_{A_n} \underbrace{\frac{1}{n} X' u}_{W_n}.$$

Suppose $\frac{1}{n} X' X \rightarrow^p Q$, positive definite. Then $A_n \rightarrow^p Q^{-1}$. Under typical assumptions, $W_n \rightarrow^p 0$. Thus,

$$plim \hat{\beta}_n = plim(V_n + A_n W_n) \rightarrow V + AW = \beta + Q^{-1} * 0 = \beta.$$

- **Corollary** Quadratic form convergence. Suppose:

$$W_n \rightarrow^d W, B_n \rightarrow^p B.$$

Then,

$$W_n' B_n W_n \rightarrow^d W' B W.$$

Thus, as an example,

$$u'X(X'X)^{-1}X'u = \underbrace{\frac{1}{\sqrt{n}}u'X}_{\rightarrow^d Z'} \underbrace{\left(\frac{1}{n}X'X\right)^{-1}}_{\rightarrow^p Q^{-1}} \underbrace{\frac{1}{\sqrt{n}}X'u}_{\rightarrow^d Z} \rightarrow Z'Q^{-1}Z.$$

12.3 Laws of Large Numbers

- Consider a random vector Z_t with $E[Z_t] = \mu_t$. Then under what conditions does:

$$\frac{1}{n} \sum_{t=1}^n (Z_t - \mu_t) \rightarrow^p 0.$$

These conditions are called the Law of Large Numbers (LLNs). We could also write the question as follows. Suppose:

$$\frac{1}{n} \sum \mu_t \rightarrow \bar{\mu}.$$

When does:

$$\frac{1}{n} \sum Z_t \rightarrow \bar{\mu}.$$

Many consistency results involve sample averages so these laws give us a lot of information.

- **Theorem:** Kolmogorov Strong Law of Large Numbers. Given:

$$Z_t \sim iid \text{ with } E[Z_t] = \mu < \infty.$$

Then,

$$\frac{1}{n} \sum_{t=1}^n (Z_t - \mu) \rightarrow^{a.s.} 0,$$

Or,

$$\frac{1}{n} \sum_{t=1}^n Z_t \rightarrow^{a.s.} \mu.$$

So for *iid* random variables, we just need a finite mean to get this LLN result.

- **Corollary** Suppose:

$$E[|Y_t|] = E[|g(Z_t)|] < \infty.$$

Ie, $Y_t = g(Z_t) = Z_t^2$, where Y_t is *iid* since Z_t is *iid*. Then:

$$\frac{1}{n} \sum_{t=1}^n Y_t \rightarrow^{a.s.} E[Y_t],$$

Or,

$$\frac{1}{n} \sum_{t=1}^n Z_t^2 \xrightarrow{a.s.} E[Z_t^2],$$

Thus if the second non-central moment is finite, then we can show the average second sample moment converges to the population moment. We could extend this to higher moments as well.

13 Lecture 13: March 15, 2005

13.1 More Laws of Large Numbers

- **Theorem:** Chebychev's LLN. Consider random variables, Z_t , with $E[Z_t] = \mu_t$, and $Var(Z_t) = \sigma_t^2$. Assume:

$$\frac{1}{n} \sum \sigma_t^2 \rightarrow \sigma^2 \implies \frac{1}{n^2} \sum \sigma_t^2 \rightarrow 0,$$

or the average variance converges. Then:

$$\frac{1}{n} \sum (Z_t - \mu_t) \xrightarrow{p} 0.$$

13.2 Central Limit Theorems

- Consider Z_t independent with $Var(Z_t) = \sigma^2$. Define:

$$Y_n = \frac{1}{\sqrt{n}} \sum (Z_t - \mu_t).$$

Then $E[Y_n] = 0$ and $Var[Y_n] = \sigma^2$. So clearly $Y_n \sim (0, \sigma^2)$. What is the probability law of Y_n ? We appeal to the following CLT.

- **Theorem:** Lindeberg-Levy CLT. Consider Z_t iid with $E[Z_t] = 0$ and $Var(Z_t) = \sigma^2 < \infty$. (Note if the mean was not zero, we could just demean). It follows that:

$$\frac{1}{\sqrt{n}} \sum_t Z_t \rightarrow^d N(0, \sigma^2),$$

regardless of the distribution of Z_t !! Z_t just has to be iid.

- **Corollary** Suppose Z_t is $k \times 1 \sim iid(0, \Sigma)$ with Σ finite. Then,

$$Y_n = \frac{1}{\sqrt{n}} \sum_t Z_t \rightarrow^d Y \sim N(0, \Sigma).$$

Proof (Via Cramer-Wold). Consider $\alpha'Y_n$ and $\alpha'Y$.

$$\alpha'Y_n = \frac{1}{\sqrt{n}} \sum_t \underbrace{\alpha'Z_t}_{V_t}.$$

Then,

$$V_t \sim iid(0, \alpha'\Sigma\alpha).$$

And by the CLT above,

$$\alpha'Y_n = \frac{1}{\sqrt{n}} \sum_t V_t \rightarrow^d N(0, \alpha'\Sigma\alpha).$$

Also,

$$\alpha'Y \sim N(0, \alpha'\Sigma\alpha).$$

Thus, by Cramer-Wold,

$$\alpha'Y_n \rightarrow^d \alpha'Y \implies Y_n \rightarrow^d Y.$$

- **Theorem:** Lindeberg-Feller CLT. Consider Z_t , an independent random variable with $E[Z_t] = 0$ and $Var(Z_t) = \sigma_t^2 < \infty$ (so not necessarily identical). Denote:

$$\sigma_{(n)}^2 = \sum_{t=1}^n \sigma_t^2,$$

and,

$$\sigma_{(n)} = \sqrt{\sigma_{(n)}^2} = \sqrt{\sum_{t=1}^n \sigma_t^2}.$$

Then if the Lindeberg condition holds (below),

$$\frac{1}{\sigma_{(n)}} \sum_t Z_t \rightarrow^d N(0, 1).$$

The Lindeberg condition states:

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma_{(n)}^2} E \left[|Z_t|^2 \mathbf{1}(|Z_t| \geq \epsilon \sigma_{(n)}) \right] = 0 \quad \forall \epsilon > 0.$$

So this is a pretty strange condition. Consider the case when Z_t is *iid*. The condition becomes:

$$\lim_{n \rightarrow \infty} \frac{1}{n\sigma^2} n * \underbrace{E \left[|Z|^2 \mathbf{1}(|Z| \geq \epsilon \sqrt{n}\sigma) \right]}_Q = 0 \quad \forall \epsilon > 0.$$

And when is the indicator function equal to 1? Only when $|Z|^2 > \epsilon \sqrt{n}\sigma$. Or,

$$E[Q] = 2 \int_{\epsilon \sqrt{n}\sigma}^{\infty} |Z|^2 f(Z) dZ.$$

So this is saying that this expectation cannot be too large! It means that the tails of the distribution cannot contain too much mass for this condition to hold. A better condition follows.

- **Definition:** Lyapounov Condition. This is stronger than the Lindeberg condition but

easier to verify:

$$\frac{1}{\sigma_{(n)}^{2+\delta}} \sum |Z_t|^{2+\delta} \rightarrow 0, \text{ for some } \delta > 0.$$

- Finally, we can write the condition as the following two sufficient conditions:

– (1)

$$\frac{1}{n} \sigma_{(n)}^2 = \frac{1}{n} \sum_t \sigma_t^2 \rightarrow \psi, \text{ with } 0 < \psi < \infty.$$

– (2)

$$\text{Sup}_n \frac{1}{n} \sum_{t=1}^n E[|Z_t|^{2+\delta}] \leq c < \infty.$$

If these two conditions are satisfied, then the Lindeberg-Feller CLT holds. Basically we need the average variance to be finite and a “bit more” (delta) than the second moments to exist.

- **Theorem:** Consider \underline{Z}_t as independent random variables with $E[\underline{Z}_t] = 0$ and $\text{Var}(\underline{Z}_t) = 1$. Denote some constants: $\sigma_1^2, \sigma_2^2, \dots$. Denote:

$$Z_t = \sigma_t \underline{Z}_t \implies E[Z_t] = 0, \text{ Var}[Z_t] = \sigma_t^2.$$

Make the following assumption:

$$\frac{\text{Max}_{1 \leq t \leq n} \{\sigma_t^2\}}{\sum_{t=1}^n \sigma_t^2} \rightarrow 0, \text{ as } n \rightarrow \infty.$$

This means that one variance cannot dominate the others. The distribution must not be too skewed. Then it follows:

$$\frac{1}{(\sum \sigma_t^2)^{1/2}} \sum_{t=1}^n Z_t \rightarrow^d N(0, 1).$$

- Example. Consider the regression model:

$$y_t = x_t \beta + u_t, \quad u_t \sim iid(0, \sigma^2), \quad x_t \text{ nonstochastic.}$$

Then,

$$\hat{\beta}_n = \beta + \frac{\sum x_t u_t}{\sum x_t^2}.$$

Make the following assumption:

$$\frac{\text{Max}_{1 \leq t \leq n} \{x_t^2\}}{\sum_{t=1}^n x_t^2} \rightarrow 0, \text{ as } n \rightarrow \infty.$$

Note, that even with this restriction (that none of the x 's dominate), we still can have growing regressors: let $x_t = \sqrt{t}$. Then,

$$\frac{\text{Max}_{1 \leq t \leq n} \{x_t^2\}}{\sum_{t=1}^n x_t^2} = \frac{n}{\sum_{t=1}^n t} = \frac{n}{n(n+1)/2} \rightarrow 0.$$

So what happens to $\sum x_t u_t$? Denote:

$$\underline{Z}_t = \frac{u_t}{\sigma} \sim iid(0, 1).$$

And,

$$\sigma_t = \sigma x_t.$$

Thus,

$$Z_t = \sigma_t \underline{Z}_t = \sigma x_t \frac{u_t}{\sigma} = x_t u_t.$$

Notice that $E[\underline{Z}_t] = 0$, $Var(\underline{Z}_t) = 1$ and $E[Z_t] = 0$, $Var[Z_t] = \sigma_t^2$. We also have the assumption that none of the x 's dominate. Thus, using the above result with $Z_t = x_t u_t$ and $\sigma_t^2 = x_t^2 \sigma^2$, we have the result:

$$\frac{1}{(\sigma^2 \sum x_t^2)^{1/2}} \sum_{t=1}^n x_t u_t \rightarrow^d N(0, 1).$$

- Now consider,

$$\hat{\beta}_n - \beta = \frac{\sum x_t u_t}{\sum x_t^2}.$$

Rewrite:

$$\hat{\beta}_n - \beta = \frac{\frac{1}{(\sigma^2 \sum x_t^2)^{1/2}} \sum x_t u_t}{\frac{1}{(\sigma^2 \sum x_t^2)^{1/2}} \sum x_t^2}.$$

Or,

$$\hat{\beta}_n - \beta = \frac{\frac{1}{(\sigma^2 \sum x_t^2)^{1/2}} \sum x_t u_t}{\frac{1}{\sigma} (\sum x_t^2)^{1/2}}.$$

Moving the term on the RHS over:

$$\left(\sum x_t^2 \right)^{1/2} (\hat{\beta}_n - \beta) = \sigma \underbrace{\frac{1}{(\sigma^2 \sum x_t^2)^{1/2}} \sum x_t u_t}_{\rightarrow^d N(0,1)} \rightarrow^d N(0, \sigma^2).$$

- Thus, without any assumptions on $\frac{1}{n} \sum x_t^2$, we have:

$$C_n = \left(\sum x_t^2 \right)^{1/2} (\hat{\beta}_n - \beta) \rightarrow^d N(0, \sigma^2),$$

or,

$$C_n \approx N(0, \sigma^2).$$

Solving for $\hat{\beta}_n$,

$$\hat{\beta}_n = \beta + \frac{1}{\left(\sum x_t^2 \right)^{1/2}} C_n \approx N\left(\beta, \sigma^2 \frac{1}{\sum x_t^2}\right).$$

- Consider the special case where $\frac{1}{n} \sum x_t^2 \rightarrow q$. Then consider:

$$A_n = \underbrace{\left(\frac{1}{n} \sum x_t^2 \right)^{1/2}}_{\rightarrow q^{1/2}} n^{1/2} (\hat{\beta}_n - \beta) \rightarrow^d N(0, \sigma^2).$$

Rearrange:

$$A_n \underbrace{\left(\frac{1}{n} \sum x_t^2 \right)^{-1/2}}_{\rightarrow q^{-1/2}} = n^{1/2} (\hat{\beta}_n - \beta).$$

So, call this new expression D_n :

$$D_n = n^{1/2} (\hat{\beta}_n - \beta) = \underbrace{A_n}_{\rightarrow N(0, \sigma^2)} \underbrace{\left(\frac{1}{n} \sum x_t^2 \right)^{-1/2}}_{\rightarrow q^{-1/2}} \rightarrow^d N(0, q^{-1} \sigma^2).$$

A good, consistent estimate of q is $\frac{1}{n} \sum x_t^2$. So,

$$D_n \approx N\left(0, \sigma^2 \frac{1}{\frac{1}{n} \sum x_t^2}\right).$$

So finally,

$$n^{-1/2} D_n = n^{-1/2} n^{1/2} (\hat{\beta}_n - \beta) = \hat{\beta}_n - \beta \approx N\left(0, \sigma^2 \frac{1}{\sum x_t^2}\right).$$

Hence,

$$\hat{\beta}_n \approx N\left(\beta, \sigma^2 \frac{1}{\sum x_t^2}\right).$$

So our assumption on the convergence of $n^{-1} x'x$ didn't matter because of the CLT. We get the same result for the distribution of $\hat{\beta}_n$.

14 Lecture 14: March 17, 2005

14.1 More on Central Limit Theorems

- **Theorem:** Suppose X is $n \times k$, non-stochastic, and $u_t \sim iid(0, \sigma^2)$. Note we do not assume normality. The OLS estimator as usual:

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

or,

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

Scaling:

$$\sqrt{n}(\hat{\beta} - \beta) = (1/nX'X)^{-1} \frac{1}{\sqrt{n}}X'u.$$

Assume:

$$\frac{1}{n}X'X \rightarrow Q, \text{ positive definite.}$$

Then, the theorem says:

$$\sqrt{n}(\hat{\beta} - \beta) = \underbrace{(1/nX'X)^{-1}}_{\rightarrow Q^{-1}} \underbrace{\frac{1}{\sqrt{n}}X'u}_{\rightarrow N(0, \sigma^2 Q)} \rightarrow N(0, Q^{-1}\sigma^2 Q Q^{-1}) \equiv N(0, \sigma^2 Q^{-1}).$$

- From this result, we can go further. Denote:

$$\eta_n = \sqrt{n}(\hat{\beta} - \beta).$$

We have just shown that η_n converges in distribution to a random variable. Thus,

$$\hat{\beta}_n = \beta + \underbrace{\frac{1}{\sqrt{n}}}_{\rightarrow 0} \underbrace{\eta_n}_{\rightarrow R.V.} \xrightarrow{p} \beta.$$

Thus $\hat{\beta}_n$ is consistent for β . Since we scaled by \sqrt{n} , we say that $\hat{\beta}_n$ is \sqrt{n} -consistent.

14.2 Large Sample Properties of Estimators

- Suppose we have random variables, Y_1, \dots, Y_n , with joint CDF and PDF: $F(y_1, \dots, y_n; \theta)$ and $f(y_1, \dots, y_n; \theta)$. Suppose we have an estimator of θ :

$$\hat{\theta}_n = \hat{\theta}_n(Y_1, \dots, Y_n).$$

We say that $\hat{\theta}_n$ is unbiased if $E[\hat{\theta}_n] = \theta$, or:

$$E_\theta[\hat{\theta}_n] = \int \hat{\theta}_n(y_1, \dots, y_n) f(y_1, \dots, y_n; \theta) dy_1 \dots dy_n = \theta, \forall \theta \in \Theta, \forall n.$$

- **Definition:** Let $plim_{n \rightarrow \infty, \theta}$ denote the limit based on convergence in probability and $p1lim_{n \rightarrow \infty, \theta}$ denote the limit based on almost sure convergence. Then we say $\hat{\theta}_n$ for θ is:

- (1) asymptotically consistent iff $lim_{n \rightarrow \infty} E_{\theta}[\hat{\theta}_n] = \theta \forall \theta \in \Theta$.
- (2) weakly consistent iff $plim_{\theta} \hat{\theta}_n = \theta \forall \theta \in \Theta$.
- (3) strongly consistent iff $p1lim_{\theta} \hat{\theta}_n = \theta \forall \theta \in \Theta$.

Weak consistency of an estimator is often just referred to as consistency.

- Note that none of the above definitions imply each other. All we do know is that if an estimator is unbiased in the limit and its variance goes to zero in the limit, then the estimator is (weakly) consistent.

14.3 Small Sample Efficiency

- **Theorem:** Cramer-Rao (again). Consider random variables, Y_1, \dots, Y_n , with joint PDF, $f_n(y_1, \dots, y_n; \theta)$, $\theta \in \Theta$. Assume f_n is C^2 with respect to θ . Consider an estimator $\hat{\theta}_n$ with $E[\hat{\theta}_n] = \theta$ and variance/covariance matrix, $\Sigma_{\hat{\theta}_n}(\theta)$. Denote the Fisher information matrix as:

$$I_n(\theta) = -E \left[\frac{\partial^2 L_n L(\theta; y_1, \dots, y_n)}{\partial \theta \partial \theta'} \right].$$

Then,

$$\Sigma_{\hat{\theta}_n}(\theta) - (I_n(\theta))^{-1}$$

is positive semi-definite. That is, the inverse of the Fisher information matrix is a lower bound for the variance/covariance matrix of ANY estimator of θ .

- Recall that the C-R lower bound need NOT be attained! You might get an efficient estimator which does not attain the lower bound. For example, the OLS (and ML) estimator for β does attain the C-R lower bound. However, the estimate of σ^2 does not attain the lower bound (either OLS or ML), though it can be shown that they are both efficient (or at least OLS is?). So we can't do any better but we never reach the lower bound.
- **Definition:** Asymptotic Normality and Asymptotic Efficiency. Suppose:

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow^d N(0, \Sigma).$$

Thus,

$$\hat{\theta}_n \approx N\left(\theta, \frac{1}{n}\Sigma\right).$$

Then we say,

$$\hat{\theta}_n \sim A N\left(\theta, \frac{1}{n}\Sigma\right),$$

or that our estimator is distributed asymptotically normal with mean θ and an “asymptotic variance/covariance matrix” of Σ/n .

- **Definition:** An estimator, $\hat{\theta}_n$ for $\theta \in \Theta \subseteq \mathfrak{R}^K$ is said to be Consistently Asymptotically Normal (CAN) with mean θ and variance/covariance, Σ if:

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow^d N(0, \Sigma).$$

Also, the estimator is said to be Consistently Uniformly Asymptotically Normal (CUAN) if this convergence is uniform over any compact subset of the parameter space. Usually all our estimators we deal with are CUAN.

15 Lecture 15: March 29, 2005

15.1 More on the Asymptotic Distribution

- Suppose we have:

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow^d N(0, \Sigma_{\hat{\theta}}).$$

Then, the asymptotic distribution of $\hat{\theta}_n$ is :

$$\hat{\theta}_n \approx N\left(\theta, \frac{1}{n}\Sigma_{\hat{\theta}}\right).$$

- Now consider a different estimator:

$$\sqrt{n}(\tilde{\theta}_n - \theta) \rightarrow^d N(0, \Sigma_{\tilde{\theta}}).$$

Then we say that $\hat{\theta}$ is more asymptotically efficient if:

$$\Sigma_{\tilde{\theta}} - \Sigma_{\hat{\theta}} \geq 0.$$

- **Theorem** Rao-Cramer. In the limit, the ML estimator attains the R-C lower bound. Let $\hat{\theta}_n$ be the MLE for θ . Then $\hat{\theta}_n$ is a consistent estimator for θ and:

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow^d N(0, \Phi),$$

with,

$$\Phi = \lim_{n \rightarrow \infty} \left[\frac{1}{n} I_n(\theta) \right]^{-1}.$$

Where I_n is the fisher information matrix. This implies:

$$\hat{\theta}_n \approx N(\theta, \Phi/n).$$

Which we could estimate with:

$$\hat{\theta}_n \approx N(\theta, \hat{\Phi}/n).$$

But note also that:

$$\hat{\Phi}/n \approx \left[\frac{1}{n} I_n(\theta) \right]^{-1} / n = [I_n(\theta)]^{-1}.$$

Thus,

$$\hat{\theta}_n \approx N(\theta, [I_n(\theta)]^{-1}).$$

- **Theorem** Suppose $\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow^d N(0, \Sigma(\theta))$. Then,

$$\Sigma(\theta) - \lim_{n \rightarrow \infty} \left[\frac{1}{n} I_n(\theta) \right]^{-1} \geq 0 \quad \forall \theta \in \Theta.$$

Thus, if the errors are normal, the ML estimator is the best we can do. It may not be the best under other distributions however.

15.2 Limiting Distributions of the OLS estimators

- Consider the following model:

$$y = X\beta + u.$$

Also consider the following assumptions:

- (A1) $E[u] = 0$.
- (A2) $VC(u) = E[uu'] = \sigma^2 I, \sigma^2 > 0$.
- (A3) X is non-stochastic and full column rank and:

$$\lim_{T \rightarrow \infty} \frac{1}{T} X'X = \underline{Q},$$

nonsingular, finite.

- (A4) $u_t \sim \text{iid}$.

Assume A1-A4 hold in all that follows.

- Consider our OLS estimator:

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

And estimated variance:

$$s_T^2 = \frac{\hat{u}'\hat{u}}{T - K}.$$

Thus $\hat{u} = y - X\hat{\beta} = Mu$ with $M = I - X(X'X)^{-1}X'$.

- Preliminary results:

$$E[1/T X'u] = 0,$$

$$VC[1/T X'u] = E\left[\frac{1}{T} X'uu'X \frac{1}{T}\right] = \frac{1}{T} \frac{1}{T} X'X E[uu'] = \frac{1}{T} \underbrace{\frac{\sigma^2}{T} X'X}_{\rightarrow \underline{Q}} \rightarrow 0.$$

Thus, we have:

$$\frac{1}{T} X'u \rightarrow^p 0. \quad (*)$$

Since the expectation and the variance both go to zero (this has been shown via Chebychev in previous lectures).

- Another result:

$$\frac{1}{\sqrt{T}} X'u \rightarrow^d N(0, \sigma^2 \underline{Q}). \quad (**)$$

This is by the CLT 4.7. Note we could have used this second result to attain the first:

$$\frac{1}{T} X'u = \frac{1}{\sqrt{T}} \underbrace{\frac{1}{\sqrt{T}} X'u}_{\rightarrow^d r.v.} \rightarrow^p 0.$$

- Thus (*) and (**) are the two key conditions that we need. In fact any set of assumptions (including A1-A4 above) that generates equations (*) and (**) are sufficient for examining the limiting distributions of the OLS estimators. Consider an alternative set of assumptions.
- Replace (A3) with the following:
 - (a): regressors $x_t = [x_{t1} \cdots x_{tK}] \sim iid$ with $E[x_t' x_t] = \underline{Q}$, finite and nonsingular.
 - (b) (x_t) and (u_t) independent.

Then, by Khinchines LLN, we have (*) and by Lindeberg-Levy CLT for iid RVs, we have (**).

16 Lecture 16: March 31, 2005

16.1 Asymptotic Properties of OLS

- Recall that our four assumptions from last time imply two key results:

$$\frac{1}{T}X'X \rightarrow \underline{Q}, \text{ non-singular.}$$

$$\frac{1}{\sqrt{T}}X'u \rightarrow^d N(0, \sigma^2 \underline{Q}).$$

Regression Coefficient: β

- **Theorem** $\hat{\beta}$ is consistent for β . Consider:

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

Thus,

$$\begin{aligned} \text{plim } \hat{\beta}_T &= \text{plim } [\beta + (X'X)^{-1}X'u] \\ &= \beta + \text{plim } (X'X)^{-1} \text{plim } X'u \\ &= \beta + \text{plim } \left(\frac{1}{T}X'X\right)^{-1} \text{plim } \frac{1}{T}X'u \\ &= \beta + Q^{-1} * 0 = \beta \end{aligned}$$

Note this relied on Slutsky's theorem a bunch of times including the fact that the plim of the inverse of a matrix is the inverse of the plim, provided that plim of the matrix is non-singular. So we say:

$$\hat{\beta}_T \rightarrow^p \beta \forall \beta \in \mathfrak{R}^K.$$

- **Theorem** Limiting distribution of $\hat{\beta}_T$. Consider:

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

Thus,

$$\sqrt{T}(\hat{\beta}_T - \beta) = \underbrace{\left(\frac{1}{T}X'X\right)^{-1}}_{\rightarrow^p \underline{Q}^{-1}} \underbrace{\frac{1}{\sqrt{T}}X'u}_{\rightarrow^d N(0, \sigma^2 \underline{Q})}.$$

So,

$$\sqrt{T}(\hat{\beta}_T - \beta) \rightarrow^d N(0, \underline{Q}^{-1} \sigma^2 \underline{Q} \underline{Q}^{-1}) \equiv N(0, \sigma^2 \underline{Q}^{-1}),$$

because $Q^{-1'} = Q^{-1}$, symmetric. Thus, we can solve for $\hat{\beta}_t$ and say that:

$$\hat{\beta}_T \approx N\left(\beta, \frac{1}{T}\sigma^2 \underline{Q}^{-1}\right).$$

However, $\sigma^2 \underline{Q}^{-1}$ is unknown. But we can estimate σ^2 with s^2 , and \underline{Q} with $\frac{1}{T}X'X$, so,

$$\hat{\beta}_T \approx N\left(\beta, \frac{1}{T}s^2\left(\frac{1}{T}X'X\right)^{-1}\right) \equiv N\left(\beta, s^2(X'X)^{-1}\right).$$

We also call this the asymptotic distribution of $\hat{\beta}_T$.

Regression Variance: σ^2

– **Theorem** s^2 is consistent for σ^2 . Consider:

$$\begin{aligned} s_T^2 &= \frac{\hat{u}'\hat{u}}{T-K} \\ &= \frac{T}{T-K} \frac{1}{T} u' M u \\ &= \frac{T}{T-K} \frac{1}{T} u' (I - X(X'X)^{-1}X') u \\ &= \frac{T}{T-K} \frac{1}{T} [u'u - u'X(X'X)^{-1}X'u] \\ &= \frac{T}{T-K} \left[\underbrace{\frac{u'u}{T}}_{\rightarrow 1} - \underbrace{\frac{1}{T}u'X}_{\rightarrow 0} \underbrace{\left(\frac{1}{T}X'X\right)^{-1}}_{\rightarrow \underline{Q}^{-1}} \underbrace{\frac{1}{T}X'u}_{\rightarrow 0} \right] \\ &\rightarrow \sigma^2 \end{aligned}$$

Note that $u'u/T = \frac{1}{T} \sum u_t^2 \xrightarrow{as} \sigma^2$ by Khinchine's LLN because $E[u_t^2] = \sigma^2$. So we say:

$$s_T^2 \xrightarrow{p} \sigma^2 \quad \forall \sigma^2 \in \mathfrak{R}.$$

– **Theorem** Limiting distribution of s^2 . Consider:

$$\sqrt{T}(s_T^2 - \sigma^2) = \underbrace{\sqrt{T}\left(s_T^2 - \frac{u'u}{T}\right)}_{W_t - V_t} + \underbrace{\sqrt{T}\left(\frac{u'u}{T} - \sigma^2\right)}_{V_t}.$$

Where,

$$\begin{aligned} W_t &= \sqrt{T}(s_T^2 - \sigma^2), \\ V_t &= \sqrt{T}\left(\frac{u'u}{T} - \sigma^2\right). \end{aligned}$$

We want to show that V_t converges in distribution to some random variable and $W_t - V_t$ converges in probability to zero. Thus,

$$V_t = \sqrt{T} \left(\frac{u'u}{T} - \sigma^2 \right) = \sqrt{T} \left(\frac{1}{T} \sum u_t^2 - \sigma^2 \right) = \sqrt{T} \frac{1}{T} \sum [u_t^2 - \sigma^2].$$

Or,

$$V_t = \frac{1}{\sqrt{T}} \sum [u_t^2 - \sigma^2].$$

By a CLT, $z_t = \frac{1}{\sqrt{T}} \sum (u_t^2 - \sigma^2) \rightarrow N(0, \text{Var}(z_t))$. Thus,

$$V_t \rightarrow^d N(0, \text{Var}(z_t)) \equiv N(0, \mu_4 - \sigma^4).$$

It can also be shown that $W_t - V_t \rightarrow 0$. (See handout). Thus,

$$\sqrt{T}(s_T^2 - \sigma^2) \rightarrow^d N(0, \mu_4 - \sigma^4).$$

This is the asymptotic distribution of s_T^2 .

Test Statistic: t

– **Theorem** Asymptotic distribution of the t -statistic. Consider, under $H_0 : \beta_i = \beta_i^*$,

$$t = \frac{\hat{\beta}_{i,T} - \beta_i^*}{s_{\hat{\beta}_{i,T}}} \rightarrow^d N(0, 1).$$

Proof: Note we have:

$$\sqrt{T}(\hat{\beta}_T - \beta^*) \rightarrow^d N(0, \sigma^2 \underline{Q}^{-1}).$$

Therefore, for an individual coefficient:

$$\sqrt{T}(\hat{\beta}_{i,T} - \beta_i^*) \rightarrow^d N(0, \sigma^2 \underline{q}^{ii}).$$

And we can write the estimated variance of the estimator as:

$$s_{\hat{\beta}_{i,T}}^2 = s^2 (X'X)^{ii}.$$

[Note the s^2 terms on either side of the equation are different.] So we can write our statistic:

$$\begin{aligned}
t &= \frac{\hat{\beta}_{i,T} - \beta_i^*}{\sqrt{s^2(X'X)^{ii}}} \\
&= \frac{\hat{\beta}_{i,T} - \beta_i^*}{\sqrt{\frac{1}{T}s^2(\frac{1}{T}X'X)^{ii}}} \\
&= \frac{\sqrt{T}(\hat{\beta}_{i,T} - \beta_i^*)}{\sqrt{s^2(\frac{1}{T}X'X)^{ii}}} \\
&= \underbrace{\frac{1}{\sqrt{s^2(\frac{1}{T}X'X)^{ii}}}}_{\rightarrow 1/\sqrt{\sigma^2 \underline{q}^{ii}}} * \underbrace{\sqrt{T}(\hat{\beta}_{i,T} - \beta_i^*)}_{\rightarrow^d N(0, \sigma^2 \underline{q}^{ii})} \\
&\rightarrow N\left(0, \frac{1}{\sigma \sqrt{\underline{q}^{ii}}} \sigma^2 \underline{q}^{ii} \frac{1}{\sigma \sqrt{\underline{q}^{ii}}}\right) \equiv N(0, 1)
\end{aligned}$$

Note that $t(T - K) \rightarrow^d N(0, 1)$ as $T \rightarrow \infty$.

Test Statistic: F

– **Theorem** Asymptotic distribution of the F -statistic. Consider, under $H_0 : R\beta = r$,

$$F = \frac{(r - R\hat{\beta})'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})/G}{\hat{u}'\hat{u}/(T - K)} \rightarrow^d \chi^2(G)/G \text{ as } T \rightarrow \infty.$$

Proof: Under H_0 ,

$$R\hat{\beta} - r = R\hat{\beta} - R\beta = R(\hat{\beta} - \beta) = R(X'X)^{-1}X'u.$$

Move over the G to the other side and consider $G * F$:

$$\begin{aligned}
GF &= \frac{u'X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'u}{s_T^2} \\
&= \frac{u'X(X'X)^{-1}R'}{s_T} [R(X'X)^{-1}R']^{-1} \frac{R(X'X)^{-1}X'u}{s_T} \\
&= \frac{\frac{1}{T}u'X(\frac{1}{T}X'X)^{-1}R'}{s_T} [R(X'X)^{-1}R']^{-1} \frac{\frac{1}{T}R(X'X)^{-1}\frac{1}{T}X'u}{s_T} \\
&= \frac{\frac{1}{T}u'X(\frac{1}{T}X'X)^{-1/2}(\frac{1}{T}X'X)^{-1/2}R'}{s_T} [R(X'X)^{-1}R']^{-1} \frac{\frac{1}{T}R(X'X)^{-1/2}(\frac{1}{T}X'X)^{-1/2}\frac{1}{T}X'u}{s_T} \\
&= \frac{\frac{1}{\sqrt{T}}u'X(\frac{1}{T}X'X)^{-1/2}(\frac{1}{T}X'X)^{-1/2}R'}{s_T} [R\frac{1}{T}(X'X)^{-1}R']^{-1} \frac{\frac{1}{T}R(X'X)^{-1/2}(\frac{1}{T}X'X)^{-1/2}\frac{1}{\sqrt{T}}X'u}{s_T} \\
&= W'_t B_t W_t
\end{aligned}$$

With

$$W_t = \frac{1}{s_T} \left(\frac{1}{T} X'X \right)^{-1/2} \frac{1}{\sqrt{T}} X'u,$$

and

$$B_t = \left(\frac{1}{T} X'X \right)^{-1/2} R' [R \frac{1}{T} (X'X) R']^{-1} R \left(\frac{1}{T} X'X \right)^{-1/2}.$$

So,

$$W_t \rightarrow^d N\left(0, \frac{1}{\sigma} \underline{Q}^{-1/2} \sigma^2 \underline{Q} \frac{1}{\sigma} \underline{Q}^{-1/2}\right) \equiv N(0, I).$$

And,

$$B_t \rightarrow^p B, \text{ symmetric, idempotent with rank } G.$$

Thus,

$$G * F = W'_t B_t W_t \rightarrow^d \chi^2(G) \implies F \rightarrow^d \chi^2(G)/G.$$

This is because we have a quadratic form in standard normals which gives us a chi-squared where the degrees of freedom equal the rank of the matrix in the middle. Note that $F(G, T - K) \rightarrow^d \chi^2(G)/G$ as $T \rightarrow \infty$.

17 Lecture 17: April 5, 2005

17.1 Nonlinear Econometric Modeling

- In all that follows, we assume that the data is iid. Denote the exogenous data as z_i , a $1 \times p_z$ vector. Then our objective function is:

$$Q_n(z_1, \dots, z_n, \hat{\tau}_n, \beta).$$

Where $\hat{\tau}_n$ is our estimated nuisance parameter. Thus,

$$\hat{\beta}_n = \arg \min_{\beta \in B} Q(z_1, \dots, z_n, \hat{\tau}_n, \beta).$$

This is called an M-estimator for β .

- What is Q? There are two major classes of the objective function, Q:
 - (1) Least mean Distance Estimators.
 - (2) Generalized Method of Moments Estimators.

Example 1: Nonlinear Least Squares (NLS)

- Let $g : X * A \mapsto \mathfrak{R}$ be a borel measurable function with $X \subseteq \mathfrak{R}^{p_x}$, $A \subseteq \mathfrak{R}^{p_a}$ as borel sets. Assume y_i 's are generated according to:

$$y_t = g(x_i, \alpha_0) + \epsilon_i, \quad i \in N.$$

Here the x_i 's are exogenous and $\alpha_0 \in A$ is the true parameter vector, so in the special case of a linear function, $g(x_i, \alpha_0) = \alpha_0 x_i$.

- Thus the objective function of the NLS estimator is:

$$Q_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_{i=1}^n q(z_i, \beta) = \frac{1}{n} \sum_{i=1}^n [y_i - g(x_i, \alpha)]^2,$$

with $z_i = [y_i, x_i]$ and $\beta = \alpha$.

Example 2: Simultaneous Equations Estimation

- Suppose (y_i) are generated as:

$$y_{im} = y_i b_{.m} + x_i c_{.m} + \epsilon_{im},$$

a linear function. Where:

$$y_i = [y_{i1}, \dots, y_{ip_y}],$$
$$x_i = [x_{i1}, \dots, x_{ip_k}],$$

$$\epsilon_i = [\epsilon_{i1}, \dots, \epsilon_{ip_y}].$$

Note we have the y 's on both sides of the equation. The element of b corresponding to y_{im} would be zero.

- Consider the simple example:

$$[y_{i1}, y_{i2}] = [y_{i1}, y_{i2}] \begin{bmatrix} 0 & b_{12} \\ b_{21} & 0 \end{bmatrix} + [x_{i1} \ x_{i2} \ x_{i3}] \begin{bmatrix} c_{11} & c_{21} \\ c_{12} & c_{22} \\ c_{12} & c_{23} \end{bmatrix} + [\epsilon_{i1} \ \epsilon_{i2}].$$

Multiply out and we have two simultaneous equations.

- Thus we could also write:

$$y_{im} = \sum_{j=1}^{p_y} y_{ij} b_{jm} + \sum_{k=1}^{p_x} x_{ik} c_{km} + \epsilon_{im}, \quad m = 1 \dots p_y.$$

So we have p_y simultaneous equations. We could also write this as:

$$[y_{i1}, \dots, y_{ip_y}] = y_i [b_{\cdot 1}, \dots, b_{\cdot p_y}] + x_i [c_{\cdot 1}, \dots, c_{\cdot p_y}] + [\epsilon_{i1}, \dots, \epsilon_{ip_y}].$$

Or,

$$y_i = y_i B + x_i C + \epsilon_i.$$

All along, we have been assuming the simultaneous equations are linear but they could be nonlinear. In this case we could write in general:

$$f(y_i, x_i, \alpha) = \epsilon_i.$$

- Thus we could solve this equation to get:

$$y_i = g(x_i, \alpha, \epsilon_i),$$

our data generating function.

- Now assume $\epsilon_t \sim iid N(0, \Sigma)$. Then the density of ϵ_i is $f_\epsilon(\epsilon_i)$ and by the transformation technique, the pdf of y is:

$$f_y(y_i) = f_\epsilon(f(y_i, x_i, \alpha)) \left| \frac{\partial f(y_i, x_i, \alpha)}{\partial y_i} \right|,$$

where the $f(\cdot)$ function inside the jacobian is the regression function, not a density function.

- So the objective function of the ML estimator (or the Normal Full Information Maximum Likelihood (NFIML) estimator) is the normal log-likelihood function of $f_y(y)$

(above):

$$Q_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_{i=1}^n q(z_i, \beta) = \frac{1}{n} \sum [\ln(|\partial f_i / \partial y|) + 0.5 \ln(\det(\Sigma)) + 0.5 f_i' \Sigma^{-1} f_i],$$

with $\beta = (\alpha, \Sigma)$. We have the jacobian term and then the rest is the log-likelihood of the normal. So we are just maximizing the log-likelihood function of y .

- Note this estimator and the NLS estimator are of the “least mean distance” type of estimator.

18 Lecture 18: April 18, 2005

18.1 More on Nonlinear Estimation

- Recall our objective function will be one of two forms:

$$Q(z_1, \dots, z_n, \hat{\tau}_n, \beta) = \begin{cases} \frac{1}{n} \sum_{i=1}^n q(z_t, \beta) \\ \left[\frac{1}{n} \sum_{i=1}^n q(z_t, \beta) \right]' \hat{\Xi}_n \left[\frac{1}{n} \sum_{i=1}^n q(z_t, \beta) \right] \end{cases}$$

Where $q : ZxB \mapsto \mathbb{R}^p$.

- Note the first type of objective function is called the class of least mean distance estimators, and q is 1×1 . In the second class we have generalized method of moments estimators where q is $p \times 1$.
- Consider our usual linear model:

$$y_i = x_i \alpha_0 + \epsilon_i,$$

where y_i is 1×1 and x_i is $1 \times K$. Then to run OLS we would:

$$\text{Min}_{\alpha} \sum_i (y_i - x_i \alpha)^2.$$

This is equivalent to:

$$\text{Min}_{\beta} \frac{1}{n} \sum_i q(z, \beta),$$

where $q(z, \beta) = (y - x\alpha)^2$, $z = [y, x]$, and $\beta = \alpha$.

- Now consider a nonlinear model:

$$y_i = g(x_i, \alpha_0) + \epsilon_i,$$

where y_i is 1×1 and x_i is $1 \times K$. Then to run OLS we would:

$$\text{Min}_{\beta} \frac{1}{n} \sum_i q(z, \beta),$$

where $q(z, \beta) = (y - g(x, \alpha))^2$, $z = [y, x]$, and $\beta = \alpha$.

- Suppose $f(y_i, x_i, \alpha_0) = \epsilon_i$ where y_i is $1 \times p_y$, x_i is $1 \times p_x$, α_0 is $1 \times p_\alpha$, and ϵ_i is $1 \times p_y$. Then $f(\cdot)$ is a vector valued function so:

$$f = (f_1, \dots, f_{p_y})'$$

Thus we have a system of p_y equations.

- Example. Suppose we have the supply and demand equations for wheat.

$$\text{Demand: } q_i = a + bp_t + es_i + \epsilon_i^d,$$

$$\text{Supply: } q_i = c + dp_t + fr_i + \epsilon_i^s.$$

So s_i is income and r_i is rainfall. In this setup, we have two endogenous variables,

$$y_i = [q_i, p_i],$$

and two exogenous,

$$x_i = [s_i, r_i],$$

with,

$$\epsilon_i = [\epsilon_i^d, \epsilon_i^s],$$

and,

$$\alpha_0 = [a, b, c, d, e, f].$$

- In general we can write:

$$f(y_i, x_i, \alpha_0) = \epsilon_i.$$

Assume:

$$\epsilon_i | x_i \sim iid N(0, \Sigma).$$

- To do maximum likelihood, we need the distribution of the y_i 's. Thus write:

$$y_i = g(\epsilon_i, x_i, \alpha_0),$$

and let $h_\epsilon(\epsilon_i)$ denote the density of the ϵ_i 's. Then by the change of variables technique:

$$h_y(y_i) = h_\epsilon(f(y_i, x_i, \alpha_0)) \cdot \left\| \frac{\partial f(y_i, x_i, \alpha_0)}{\partial y_i} \right\|.$$

If we compute this using the normal pdf and take logs, we get:

$$q(z_i, \beta) = -\ln|\det(\partial f_i / \partial y)| + \frac{1}{2} \ln|\Sigma| + \frac{1}{2} f_i' \Sigma^{-1} f_i.$$

And thus our objective function becomes:

$$Q_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_{i=1}^n q(z_i, \beta),$$

where $z_i = [y_i, x_i]$ and $\beta = \{\text{elements of } \alpha \text{ and } \Sigma\}$.

- Consider another example. Suppose $y_i = \{0, 1\}$ and,

$$Pr(y_i = 1 | x_i) = G(x_i, \alpha_0),$$

$$Pr(y_i = 0 | x_i) = 1 - G(x_i, \alpha_0).$$

So $G(\cdot)$ is a CDF that could be either logit or probit if the distribution is logistic or normal respectively. Then,

$$f(y_i) = G(x_i, \alpha_0)^{y_i} (1 - G(x_i, \alpha_0))^{1-y_i}.$$

Then the log likelihood function would be:

$$\ln \mathcal{L}_i = y_i \ln G(x_i, \alpha_0) + (1 - y_i) \ln G(x_i, \alpha_0).$$

And our objective function is:

$$Q_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_{i=1}^n \ln \mathcal{L}_i = \frac{1}{n} \sum_{i=1}^n [y_i \ln G(x_i, \alpha_0) + (1 - y_i) \ln G(x_i, \alpha_0)].$$

- Note that when we do OLS, the model is $y_i = x_i \alpha_0 + \epsilon_i$ with $\epsilon_i | x_i \sim iid(0, \sigma^2)$. We then minimize: $\sum (y_i - x_i \alpha)^2$. We know in OLS that the population moment condition is:

$$E[x'_i \epsilon_i] = 0,$$

which is actually K moment conditions, one for each of the K regressors: $E[x_{i1} \epsilon_i] = 0, \dots, E[x_{ik} \epsilon_i] = 0$. So inserting the model we have:

$$E[x'_i (y_i - x_i \alpha)] = E[q(z_i, \beta)] = 0 \text{ at } \alpha = \alpha_0.$$

When we run OLS, we could achieve the same estimator as usual using the sample moments:

$$\frac{1}{n} \sum_i x'_i (y_i - x_i \alpha).$$

If we pick α to get this as close to zero as possible, we have found our usual OLS estimator:

$$\hat{\alpha}_{ols} = \frac{\sum x'_i y_i}{\sum x'_i x_i}.$$

As long as we have as many moment conditions as we have unknown parameters, we can set the sample moment equal to zero and find our estimator. This is the least mean distance type of estimator. When we have more moments than parameters, we use GMM.

- In an even more general setting, we consider a set of instruments, a_i , for the x 's. So if the model is $y_i - x_i \alpha_0 = \epsilon_i$ or:

$$f(y_i, x_i, \alpha_0) = \epsilon_i,$$

there might be endogenous variables (inside x_i) that are correlated with ϵ_i that we need to find instruments for. We can use the x 's from the ENTIRE system as instruments (and possibly other exogenous variables outside the system). So a_i is a $p_a \times 1$ vector of

x_i 's that are uncorrelated with ϵ_i . We need the number of uncorrelated instruments to be at least as big as the number of parameters we choose to estimate.

- So our population moment condition might be:

$$E[a_i'\epsilon] = E[a_i'f(y_i, x_i, \alpha_0)] = 0,$$

noting that we have substituted in the true parameter vector, α_0 .

- In this case our objective function would become (and this is somehow equivalent to 2-stage least squares):

$$Q_n(z_1, \dots, z_n, \beta) = \left[\frac{1}{n} \sum_{i=1}^n a_i' f(y_i, x_i, \alpha) \right]' \left(\frac{1}{n} \sum_{i=1}^n a_i' a_i \right)^{-1} \left[\frac{1}{n} \sum_{i=1}^n a_i' f(y_i, x_i, \alpha) \right].$$

19 Lecture 19: April 19, 2005

19.1 More on Nonlinear Estimation

- Suppose we have data, $x = (x_1, \dots, x_n)'$ and:

$$\frac{1}{n}x'\epsilon = \frac{1}{n}[x'_1, \dots, x'_n][\epsilon_1, \dots, \epsilon_n]' = \frac{1}{n} \sum_{i=1}^n x'_i \epsilon_i.$$

- Assume we have the least squares moment conditions:

$$E\left[\frac{1}{n}x'\epsilon\right] = \frac{1}{n} \sum E[x'_i \epsilon_i] = 0.$$

This is actually a set of M moment conditions, one for each regressor:

$$\frac{1}{n} \sum_{i=1}^n E[X_{i1} \epsilon_i] = 0$$

⋮

$$\frac{1}{n} \sum_{i=1}^n E[X_{iM} \epsilon_i] = 0$$

- In the linear case, $\epsilon_i = y_i - x_i\beta$, we would have empirical moments:

$$\frac{1}{n} \sum_i X'_i (y_i - X_i \beta),$$

and we would choose β to make this as close to zero as possible.

- In general, we have:

$$\epsilon_i = y_i - g(x_i, \beta_0),$$

with empirical moments:

$$\frac{1}{n} \sum x'_i (y_i - g(x_i, \beta)) = 0 \text{ only if } \dim(\beta) = \dim(x_i).$$

If the dimensions do not match, ie we have more moments than parameters, we will use some quadratic form with a positive semi-definite matrix in the middle.

- Now consider the case were we have a bunch of instruments, a_i , for the x_i 's. If there is one equation, then the population moment condition is:

$$E[a'_i \epsilon_i] = 0.$$

Now assume there are M equations so $\epsilon_i = [\epsilon_{i1}, \dots, \epsilon_{iM}]$. Thus our population moments are:

$$\begin{aligned} E[a'_i \epsilon_{i1}] &= 0 \\ &\vdots \\ E[a'_i \epsilon_{iM}] &= 0 \end{aligned}$$

- Consider an example of an s equation system with $G + 1$ endogenous variables and K exogenous variables. Thus the system is:

$$\begin{aligned} \underbrace{y_1}_{nx1} &= \underbrace{Y_1}_{nxG1} \underbrace{\beta_1^0}_{G1x1} + \underbrace{X_1}_{nxK1} \underbrace{\gamma_1^0}_{K1x1} + \underbrace{\epsilon_1}_{nx1} \\ &\vdots \\ \underbrace{y_s}_{nx1} &= \underbrace{Y_s}_{nxGs} \underbrace{\beta_s^0}_{Gsx1} + \underbrace{X_s}_{nxKs} \underbrace{\gamma_s^0}_{Ksx1} + \underbrace{\epsilon_s}_{nx1} \end{aligned}$$

So the Y 's are the endogenous variables (plus y) makes $G + 1$ of them. And the X 's are exogenous. Now suppose we want to estimate the first equation and we use the moment condition:

$$E[X'_1 \epsilon_1] = 0.$$

This is only $K1$ conditions and we have $K1 + G1$ parameters to estimate so we need more moments. Thus use:

$$\begin{aligned} E[X'_1 \epsilon_1] &= 0 \\ &\vdots \\ E[X'_s \epsilon_s] &= 0 \end{aligned}$$

for each of the s equations. Note that X matrix is ALL the exogenous variables (nxK) in the system, not just those in each equation separately. Thus, to get identification, we need $K > Gi + Ki$ for $i = 1 \dots s$. So what are empirical moments? They are:

$$q(z_i, \beta) = \begin{bmatrix} \frac{1}{n} \sum X'_i [y_1 - Y_1 \beta_1 - X_1 \gamma_1] \\ \vdots \\ \frac{1}{n} \sum X'_i [y_s - Y_s \beta_s - X_s \gamma_s] \end{bmatrix}.$$

Where $\beta = (\beta_1, \gamma_1, \dots, \beta_s, \gamma_s)$ and $z = (y_1, \dots, y_s, X)$.

19.2 Consistency of an Estimator

- Recall the two forms of our objective function (Least Mean Distance and Generalized Method of Moments):

$$Q_n(z_1, \dots, z_n, \hat{\tau}_n, \beta) = \begin{cases} \frac{1}{n} \sum_{i=1}^n \underbrace{q}_{1 \times 1}(z_i, \beta) \\ \left[\frac{1}{n} \sum_{i=1}^n \underbrace{q}_{p \times 1}(z_i, \beta) \right]' \hat{\Xi}_n \left[\frac{1}{n} \sum_{i=1}^n \underbrace{q}_{p \times 1}(z_i, \beta) \right] \end{cases}$$

- Denote:

$$Q_n(z_1, \dots, z_n, \beta) = R(\omega, \beta),$$

where $z_i = z_i(\omega)$.

- Thus, our M-estimator, $\hat{\beta}_n$ solves:

$$Q_n(z_1, \dots, z_n, \hat{\beta}_n) = \min_{\beta \in B} Q_n(z_1, \dots, z_n, \beta).$$

- Denote the limiting distribution of the objective function as $\bar{Q}(\beta)$. Let β_0 be the minimizer of \bar{Q} .
- So to show consistency, we will try to show that $Q_n \rightarrow \bar{Q}$ and if the convergence is uniform, then the minimizer of each function should also converge.
- But what form does \bar{Q} take? Consider least squares:

$$y_i = x_i \beta_0 + \epsilon_i, \quad E[\epsilon_i | x_i] = 0, \quad E[\epsilon_i^2 | x_i] = \sigma^2.$$

Note that as with everything we do in nonlinear analysis, the y 's and x 's are both iid. Then the objective function is:

$$Q_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_i (y_i - x_i \beta)^2.$$

Define the limiting distribution as:

$$\begin{aligned} \bar{Q}(\beta) &= E[Q_n(z_1, \dots, z_n, \beta)] \\ &= E\left[\frac{1}{n} \sum_i (y_i - x_i \beta)^2\right] \\ &= E[(y_i - x_i \beta)^2] \\ &= E[(x_i \beta_0 + \epsilon_i - x_i \beta)^2] \\ &= E[(\epsilon_i - x_i(\beta - \beta_0))^2] \\ &= E[\epsilon_i^2] - E[2\epsilon_i x_i(\beta - \beta_0)] + E[(\beta - \beta_0)' x_i' x_i (\beta - \beta_0)] \\ &= \sigma^2 + (\beta - \beta_0)' E[x_i' x_i] (\beta - \beta_0) \end{aligned}$$

Thus $\bar{Q}(\beta)$ is minimized at the true parameter value, ie $\beta = \beta_0$.

- See G-19.1. We can see here that the Q_n function might be minimized at a different value than \bar{Q} is minimized. If $Q_n \rightarrow \bar{Q}$ and this convergence happens uniformly, then we will get consistency of our estimator.
- See G-19.2 for a picture of what non-uniform convergence might look like. Note that Q_{100} is closer to \bar{Q} than Q_{10} at most points except for the dip where the convergence is NOT uniform. You might get divergence of the minimizer away from β_0 as shown in the graph.

20 Lecture 20: April 21, 2005

20.1 More on Nonlinear Estimation

Consistency of the Nonlinear Estimator

- Recall our finite and limiting objective functions:

$$Q_n(z_1, \dots, z_n) = R(\omega, \beta),$$

$$\bar{Q}(\beta) = \bar{R}(\beta).$$

Which we can write as:

$$Q_n = \frac{1}{n} \sum_{i=1}^n q(z_i, \beta),$$

$$\bar{Q} = \frac{1}{n} \sum_{i=1}^n E[q(z_i, \beta)] = E[q(z_i, \beta)],$$

where the final equality follows because z_i is iid.

- We also have the vector valued objective function (as with GMM estimators),

$$Q_n = \left[\frac{1}{n} \sum_{i=1}^n q(z_i, \beta) \right]' \hat{\Xi}_n \left[\frac{1}{n} \sum_{i=1}^n q(z_i, \beta) \right],$$

and,

$$\bar{Q} = [E[q(z_i, \beta)]]' \hat{\Xi}_n [E[q(z_i, \beta)]].$$

- See G-20.1. We can see that \bar{Q} is uniquely identified at the true β_0 . We say that β_n is identifiably unique if:

$$\liminf_{n \rightarrow \infty} \left[\inf_{\beta \ni |\beta - \beta_n| \geq \epsilon} \bar{Q}_n(\beta) - \bar{Q}_n(\beta_n) \right] > 0 \quad \forall \epsilon > 0.$$

Note we put the subscripts on the β 's because it's possible that we could have different estimators (minimizers) for different n 's. Thus the distance between the value of our limiting objective function at the estimator must be smaller than the value of the limiting objective function at any other parameter value in the parameter range outside any closed ball around the estimator. This is a long and complicated way of saying the objective function has a true min.

- **Remark** A sufficient set of conditions for an estimator to be identifiably unique is:
 - (1) \bar{Q} continuous.
 - (2) B compact (the parameter space).
 - (3) β_0 must be a unique minimizer.

So we need to have an objective function which has a unique minimum and this expression above guarantees that for any n , (particularly for large n), the minimizer converges (ie, we can identify it).

- **Theorem** This seems like the key theorem so far. Recall our notation:

$$R_n(\omega, \beta) = Q_n(z_1, \dots, z_n, \beta),$$

$$\bar{R}(\beta) = \bar{Q}(\beta).$$

Note we are assuming there is no nuisance parameter. Note also that both of these functions are REAL valued. No matter if we have to do some sort of quadratic form to get there, what comes out is just one number. Now assume:

$$\text{Sup}_{\beta \in B} |R_n(\omega, \beta) - \bar{R}(\beta)| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty.$$

Thus the biggest difference between these two functions goes to zero (uniform convergence). Suppose also that B is compact, \bar{R} is continuous, and β_0 is a unique minimizer. Then for our minimizer, $\hat{\beta}_n$, such that:

$$R(\omega, \hat{\beta}_n) = \text{Inf}_{\beta \in B} R(\omega, \beta),$$

we have:

$$\hat{\beta}_n \xrightarrow{p} \beta_0 \text{ as } n \rightarrow \infty.$$

So $\hat{\beta}_n$ is consistent for β_0 . See notes for proof.

- Remark. Consider the case of Least Mean Distance Estimators. We have:

$$R_n(\omega, \beta) = \frac{1}{n} \sum_i q(z_i, \beta),$$

$$\bar{R}(\beta) = E[q(z_i, \beta)].$$

Then,

$$|R_n - \bar{R}| = \left| \frac{1}{n} \sum q(z_i, \beta) - E[q(z_i, \beta)] \right| \xrightarrow{p} 0,$$

for EACH $\beta \in B$. This is exactly a Law of Large Numbers result. But since we consider all parameters at once, the result above simplifies really to a Uniform Law of Large Numbers or ULLN.

- Note the same thing can be shown for GMM estimators. We really need a ULLN result to show consistency.

- **Theorem** Uniform Law of Large Numbers. Assume the following holds:

- (1) z_i are iid.
- (2) q can be either scalar or vector valued.
- (3) $q(\cdot, \beta)$ is measurable for each $\beta \in B$.

- (4) $q(z, \cdot)$ is continuous for each $z \in Z$.
- (5) B is compact.
- (6) Domination Condition: $|q(z, \beta)| \leq h(z)$ with $E[h(z)] < \infty$.

Then we have:

$$\text{Uniform Convergence: } \text{Sup}_{\beta \in B} \left| \frac{1}{n} \sum q(z_i, \beta) - E[q(z_i, \beta)] \right| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty,$$

and,

$$E[q(z_i, \beta)] \text{ if finite and continuous in } \beta.$$

- Finally, we can switch things around and have another useful result. Suppose $\hat{\beta}_n \rightarrow \beta_0$. When does:

$$R(\omega, \hat{\beta}_n) \rightarrow \bar{R}(\beta_0) ??$$

The answer is when the following two conditions are satisfied:

- (1) A ULLN is satisfied: $\text{Sup}_{\beta \in B} |R(\omega, \beta) - \bar{R}(\beta)| \xrightarrow{p} 0$.
- (2) \bar{R} is continuous.

You might use a result like this to check to see if the elements of the hessian converge to their expected value in order to attain an estimator for the variance/covariance matrix of your estimator, $\hat{\beta}_n$.

21 Lecture 21: April 21, 2005

21.1 Consistency - Catalogues of Assumptions

Least Mean Distance Estimators - General

- Consider the objective function for LMD estimators:

$$Q_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_i q(z_i, \beta),$$

where q is 1×1 .

- **Assumptions 4.1**

- (a) $q : Z \times B \mapsto \mathfrak{R}$. q is real valued.
- (b) B is compact.
- (c) $q(\cdot, \beta)$ is measurable for each $\beta \in B$. $q(z, \cdot)$ is continuous for each $z \in Z$.
- (d) z_i is iid.
- (e) Domination: $\text{Sup}_{\beta \in B} |q(z_i, \beta)| < \infty$.

- If assumptions 4.1 are satisfied then we have the following ULLN:

$$\text{Sup}_{\beta \in B} \left| \frac{1}{n} \sum q(z_i, \beta) - E[q(z_i, \beta)] \right| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty,$$

and,

$$E[q(z_i, \beta)] \text{ is continuous.}$$

- If we add to assumptions 4.1 the assumption that β_0 is a UNIQUE minimizer of \bar{R} , then we get a consistent estimator:

$$\hat{\beta}_n \xrightarrow{as} \beta_0 \text{ as } n \rightarrow \infty.$$

Least Mean Distance Estimators - Nonlinear Least Squares

- Consider iid data processes, (y_i, x_i) . Suppose we think that the following holds:

$$E[y_i | x_i] = g(x_i, \beta_0).$$

- Define a residual:

$$\epsilon_i = y_i - g(x_i, \beta_0) \implies E[\epsilon_i] = 0.$$

- Thus the nonlinear model becomes:

$$y_i = g(x_i, \beta_0) + \epsilon_i.$$

Why did we do this in such a strange way? Because it allows for $E[\epsilon_i^2|x_i] = h(x_i)$, ie, conditional heteroskedasticity. If we start with the model, we wouldn't be allowing for this.

- Thus, our objective function is then:

$$R_n(\omega, \beta) = Q_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_i q(z_i, \beta) = \frac{1}{n} \sum_i (y_i - g(x_i, \beta))^2.$$

And the limiting distribution is:

$$\bar{R}(\beta) = E[(y_i - g(x_i, \beta))^2].$$

Note that usually the limiting distribution involves taking a limit as n goes to infinity, but since we assume iid data, this is equivalent to taking expectations.

- Consider rewriting the limiting distribution:

$$\begin{aligned} \bar{R}(\beta) &= E[(y_i - g(x_i, \beta))^2] \\ &= E[(g(x_i, \beta_0) + \epsilon_i - g(x_i, \beta))^2] \\ &= E[\epsilon_i^2] + E[(g(x_i, \beta_0) - g(x_i, \beta))^2] \end{aligned}$$

So the second term is minimized when $\beta = \beta_0$. If we assume $E[(g(x_i, \beta_0) - g(x_i, \beta))^2] > 0 \forall \beta \neq \beta_0$, we should get a unique minimizer. See theorem below.

- **Assumptions 4.2**

- (a) g is real valued.
- (b) B is compact.
- (c) $g(\cdot, \beta)$ is measurable for each $\beta \in B$. $g(x, \cdot)$ is continuous for each $x \in p_x$.
- (d) $z_i = [y_i, x_i]$ is iid.
- (e) $E[y_i|x_i] = g(x_i, \beta_0)$.
- (f) Domination: $E[(y_i - g(x_i, \beta))^2] < \infty$ and $Sup_{\beta \in B} g(x_i, \beta)^2 < \infty$.

- **Theorem** If assumptions 4.2 are satisfied and $E[(g(x_i, \beta_0) - g(x_i, \beta))^2] > 0 \forall \beta \neq \beta_0$, then:

$$\hat{\beta}_n \rightarrow^{as} \beta_0 \text{ as } n \rightarrow \infty.$$

- Note the equivalent condition to assumption 4.2.f in the linear case is for $X'X$ to be non-singular. Assumptions 4.2 focus on the $g(\cdot, \cdot)$ function rather than the $q(\cdot, \cdot)$ function mostly for convenience. You could work with either.

Least Mean Distance Estimators - Maximum Likelihood

- Suppose we have a conditional density function:

$$f(y|x; \beta),$$

where y and x could be scalar or vector valued.

- Consider our objective function:

$$R_n(\omega, \beta) = Q_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_i q(z_i, \beta) = \frac{1}{n} \sum_i -\log[f(y_i|x_i, \beta)].$$

- **Assumptions 4.3**

- (a) q is real valued.
- (b) B is compact.
- (c) $q(\cdot, \beta)$ is measurable for each $\beta \in B$. $g(z, \cdot)$ is continuous for each $z \in Z$.
- (d) $z_i = [y_i, x_i]$ is iid.
- (e) Domination: $E[\text{Sup}_{\beta \in B} |q(z_i, \beta)|] < \infty$.

- **Theorem** If assumptions 4.3 are satisfied and β_0 is a unique minimizer for $E[q(z_i, \beta)]$, then:

$$\hat{\beta}_n \rightarrow^{as} \beta_0 \text{ as } n \rightarrow \infty.$$

- **Remark** Final point. When is β_0 unique when we're doing maximum likelihood? Consider the Kullback-Leibler Divergence:

$$E \left[\ln \frac{f(y, \beta_0)}{f(y, \beta)} \right] = E[q(z_i, \beta)] - E[q(z_i, \beta_0)] \geq 0.$$

Thus β_0 is always going to be a minimizer when we do MLE. However, we still have to assume that it's UNIQUE!

22 Lecture 22: April 26, 2005

22.1 Consistency - Catalogues of Assumptions

General Method of Moments Estimators - General

- Example 1. We are seeking moment conditions of the form: $E[q(z_i, \theta)] = 0$. Suppose we have a model of production:

$$Q = F(K, L), \quad L = G(K, Q, \theta_0).$$

So we might want to minimize:

$$\text{Min } E[wL + cK],$$

expected cost. A FOC would then be:

$$E\left[w_i \frac{\partial L_i}{\partial K_i} + c_i\right] = E[f(z_i, \theta)] = 0, \quad z_i = [w_i, c_i, K_i].$$

So these would be our moment conditions.

- Example 2. Suppose our model was the general nonlinear model: $y_i = g(x_i, \beta_0) + \epsilon_i$. We might use moments:

$$E[x'_i \epsilon_i] = 0 \iff E[x'_i (y_i - g(x_i, \beta_0))] = 0 \iff E[q(z_i, \beta_0)] = 0.$$

Now, what should we do with these moments? We have k of them, so calculate:

$$Q = \text{Min} \left[\frac{1}{n} \sum_i q(z_i, \beta_0) \right]' \hat{\Xi}_n \left[\frac{1}{n} \sum_i q(z_i, \beta_0) \right],$$

where $\hat{\Xi}_n$ is a positive semi-definite matrix. It could be the identity matrix, but we will show that the most efficient matrix to put in there is the inverse of the variance/covariance matrix of the moments.

- Example 3. Consider the ordinary linear model: $y_i = x_i \beta_0 + \epsilon_i$. Assume $E[x'_i \epsilon_i] = 0$, our k moment conditions. Then construct a GMM type objective function:

$$Q = \left[\frac{1}{n} \sum_i x'_i (y_i - x_i \beta_0) \right]' I_k \left[\frac{1}{n} \sum_i x'_i (y_i - x_i \beta_0) \right].$$

Clearly, in this case, minimizing the quadratic form yields the same estimator as minimizing the first term which means setting it equal to zero:

$$\frac{1}{n} \sum_i x'_i (y_i - x_i \hat{\beta}) = 0 \implies \frac{1}{n} X' y - \frac{1}{n} X' X \hat{\beta} = 0.$$

So the OLS estimator, $\hat{\beta}$, can be viewed as a GMM estimator when we utilize the moment condition that the x 's and the errors are orthogonal.

- So consider the objective function of the GMM estimator:

$$R_n(\omega, \beta) = Q_n(z_1, \dots, z_n, \beta) = \left[\frac{1}{n} \sum_i q(z_i, \beta) \right]' \hat{\Xi}_n \left[\frac{1}{n} \sum_i q(z_i, \beta) \right].$$

With $\hat{\Xi}_n \rightarrow^p \Xi_0$. Thus, the limiting distribution (non-stochastic analogue) is:

$$\bar{R}(\beta) = [E[q(z_i, \beta)]]' \Xi_0 E[q(z_i, \beta)].$$

- **Assumptions 4.4**

- (a) $q : Z \times B \mapsto \mathfrak{R}^{p_q}$ is a real vector valued function.
- (b) B is compact.
- (c) $q(\cdot, \beta)$ is measurable for each $\beta \in B$. $q(z, \cdot)$ is continuous for each $z \in Z$.
- (d) z_i is iid.
- (e) Domination: $E[\text{Sup}_{\beta \in B} \|q(z_i, \beta)\|] < \infty$. where $\|\cdot\|$ is the euclidean norm.

- **Theorem** If assumptions 4.4 are satisfied and,

$$\bar{R}(\beta) = [E[q(z_i, \beta)]]' \Xi_0 E[q(z_i, \beta)] \begin{cases} \neq 0 & \text{if } \beta \neq \beta_0 \\ = 0 & \text{if } \beta = \beta_0 \end{cases}$$

Then:

$$\hat{\beta}_n \rightarrow^{as} \beta_0 \text{ as } n \rightarrow \infty.$$

- Proof. Given assumptions 4.4, we get a ULLN:

$$\text{Sup}_{\beta} \left\| \frac{1}{n} \sum q(z_i, \beta) - E[q(z_i, \beta)] \right\| \rightarrow^p 0 \implies \text{Uniform Convergence,}$$

and,

$$E[q(z_i, \beta)] \text{ continuous.}$$

It follows that:

$$\text{Sup}_{\beta} |R_n(\omega, \beta) - \bar{R}(\beta)| \rightarrow^{as} 0 \text{ as } n \rightarrow \infty.$$

Since $\bar{R}(\beta) = 0$ only at the true parameter value and Ξ is positive semidefinite, then the quadratic form above defines a unique minimizer.

23 Lecture 23: April 28, 2005

23.1 Asymptotic Normality of an Estimator

Least Mean Distance Estimators

- First some notation:

$$\frac{\partial f}{\partial x} = \left[\frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_p} \right] = \underbrace{\nabla_x f}_{1 \times p}$$

$$\frac{\partial^2 f}{\partial x \partial x} = \underbrace{\nabla_{xx} f}_{p \times p}$$

- We now assume that we have a consistent estimator and we seek conditions such that it is asymptotically normal. Consider the objective function of the Least Mean Distance estimator:

$$Q(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_{i=1}^n q(z_i, \beta).$$

Normalize:

$$n^{1/2} Q(z_1, \dots, z_n, \beta) = n^{-1/2} \sum_{i=1}^n q(z_i, \beta).$$

Differentiate:

$$n^{1/2} \frac{\partial Q(\hat{\beta}_n)}{\partial \beta'} = n^{-1/2} \sum_{i=1}^n \frac{\partial q(z_i, \hat{\beta}_n)}{\partial \beta'} = 0.$$

(Setting this equal to zero gives us our estimator, $\hat{\beta}_n$).

- Now, consider a function $g(\hat{\beta}_n)$, and consider a first order Taylor series expansion around the true parameter value β_0 :

$$g(\hat{\beta}_n) \approx g(\beta_0) + \left. \frac{\partial g}{\partial \beta} \right|_{\beta=\beta_0} (\hat{\beta}_n - \beta_0).$$

For this to hold with equality, we would need an infinite number of terms on the RHS. We can also exploit the mean value theorem which says:

$$g(\hat{\beta}_n) = g(\beta_0) + \left. \frac{\partial g}{\partial \beta} \right|_{\beta=\tilde{\beta}} (\hat{\beta}_n - \beta_0).$$

where $\tilde{\beta} \in [\hat{\beta}_n, \beta_0]$. So there is some value of the parameter that if we evaluate just this first term at that parameter, we get the exact value of the function. So applying this idea to the LHS of our equation:

$$0 = n^{1/2} \frac{\partial Q(\hat{\beta}_n)}{\partial \beta'} = n^{1/2} \frac{\partial Q(\beta_0)}{\partial \beta'} + n^{1/2} \frac{\partial^2 Q(\tilde{\beta})}{\partial \beta \partial \beta'} (\hat{\beta}_n - \beta_0).$$

- Assuming invertibility, we can solve for $\hat{\beta}_n - \beta_0$:

$$\hat{\beta}_n - \beta_0 = - \left(n^{1/2} \frac{\partial^2 Q(\tilde{\beta})}{\partial \beta \partial \beta'} \right)^{-1} n^{1/2} \frac{\partial Q(\beta_0)}{\partial \beta'}.$$

Or in summation notation (equivalent to doing the expansion on the RHS of our equation above):

$$\hat{\beta}_n - \beta_0 = - \left(n^{-1/2} \sum_{i=1}^n \frac{\partial^2 q(z_i, \tilde{\beta})}{\partial \beta \partial \beta'} \right)^{-1} n^{-1/2} \sum_{i=1}^n \frac{\partial q(z_i, \beta_0)}{\partial \beta'}.$$

Normalize:

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = - \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 q(z_i, \tilde{\beta})}{\partial \beta \partial \beta'} \right)^{-1} n^{-1/2} \sum_{i=1}^n \frac{\partial q(z_i, \beta_0)}{\partial \beta'}.$$

- So consider the gradient term. At the true parameter:

$$\frac{\partial q(z_i, \beta_0)}{\partial \beta'} = \left[\frac{\partial q(z_i, \beta_0)}{\partial \beta_1} \dots \frac{\partial q(z_i, \beta_0)}{\partial \beta_p} \right]' = [\psi_{1i} \dots \psi_{pi}]' = \underbrace{\psi_i}_{p \times 1}.$$

Since z_i are iid, ψ_i are also iid (being functions of iid random variables). So, if $\psi_i \sim iid(0, \Sigma)$, then by a CLT,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i \rightarrow^d N(0, \Sigma).$$

In our case, we typically have:

$$E \left[\frac{\partial q(z_i, \beta_0)}{\partial \beta} \right] = 0,$$

at the true parameter value. Hence,

$$n^{-1/2} \sum_{i=1}^n \frac{\partial q(z_i, \beta_0)}{\partial \beta'} \rightarrow^d \xi \sim N(0, \Lambda_0).$$

- So consider the hessian term. At the true parameter:

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 q(z_i, \beta_0)}{\partial \beta \partial \beta'} = \begin{bmatrix} \ddots & & & & \\ \ddots & \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 q(z_i, \beta_0)}{\partial \beta_k \partial \beta_l} & & & \\ \ddots & & \ddots & & \\ \ddots & & & \ddots & \\ \ddots & & & & \ddots \end{bmatrix}.$$

So element by element we have iid random variables which by Khinchines LLN:

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 q(z_i, \beta_0)}{\partial \beta \partial \beta'} \rightarrow^p E \left[\frac{\partial^2 q(z_i, \beta_0)}{\partial \beta \partial \beta'} \right].$$

BUT we are not evaluating at the true parameter but rather our “mean value”, $\tilde{\beta}$. Well, since we have assumed consistency:

$$\hat{\beta}_n \rightarrow \beta_0 \text{ and } \tilde{\beta} \in [\hat{\beta}_n, \beta_0], \text{ then } \tilde{\beta}_n \rightarrow \beta_0.$$

But this isn't quite enough. See the end of lecture 20, which says that if we have a consistent estimator, then we can get the objective function to converge if we have a ULLN. So if things converge nicely here, we get:

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 q(z_i, \tilde{\beta})}{\partial \beta \partial \beta'} \rightarrow^p E \left[\frac{\partial^2 q(z_i, \beta_0)}{\partial \beta \partial \beta'} \right] = A_0.$$

- So, we have our result:

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \underbrace{- \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 q(z_i, \tilde{\beta})}{\partial \beta \partial \beta'} \right)^{-1}}_{\rightarrow^p A_0^{-1}} \underbrace{n^{-1/2} \sum_{i=1}^n \frac{\partial q(z_i, \beta_0)}{\partial \beta'}}_{\rightarrow^d \xi \sim N(0, \Lambda_0)} \rightarrow^d N(0, A_0^{-1} \Lambda_0 A_0^{-1}).$$

Asymptotic Normality Result for LMD Estimators

• Assumptions 5.1

- (a) Parameter Spaces are compact
- (b) Q_n is C^2 .
- (c) $\hat{\beta}_n \rightarrow^p \beta_0$ as $n \rightarrow \infty$ and $\beta_0 \in B$. We need the true parameter to be in the strict interior so we can do a Taylor expansion around it. Also, $n^{1/2}(\hat{\tau}_n - \tau_0) = O_p(1)$, a random variable.
- (d) $n^{1/2} \nabla_{\beta'} Q(\hat{\beta}_n) = o_p(1)$. This means that our estimator, $\hat{\beta}_n$ satisfies our FOC up to an error of magnitude $o_p(1)$.
- (e) $\nabla_{\beta\beta} Q_n(\tilde{\beta}_n) \rightarrow^p A_0$. This we have shown above.
- (f) $\nabla_{\beta\tau} Q_n(\tilde{\tau}_n, \tilde{\beta}_n) \rightarrow^p 0$, so the cross derivative terms are 0. τ is truly a nuisance. This was clear before we started ...
- (g) There exists a real matrix D_0 , such that:

$$-n^{1/2} \nabla_{\beta'} Q_n(z_1, \dots, z_n, \tau_0, \beta_0) = D_0 \xi_n + o_p(1).$$

- **Theorem** If assumptions 5.1 hold, then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow^d A_0^{-1} D_0 \xi.$$

And if $\xi \sim N(0, \Lambda_0)$, then as above,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}D_0\Lambda_0D_0'A_0^{-1}) \equiv N(0, A_0^{-1}B_0A_0^{-1}),$$

with $B_0 = D_0\Lambda_0D_0'$.

- Note for least mean distance estimators, $D_0 = I$, so,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}\Lambda_0A_0^{-1}).$$

- So if assumptions 5.1 are satisfied then the LMD estimator is consistent and asymptotically normal.

24 Lecture 24: May 3, 2005

24.1 Asymptotic Normality of an Estimator

GMM Estimators

- Before we start into deriving the asymptotic distribution of the GMM estimator, we start with an example. Suppose we have the following three equation system:

$$\text{Technology: } y_{i1} = Ay_{i2}^{\alpha_1}y_{i3}^{\alpha_2} + u_{i1}.$$

$$\text{Labor Demand: } y_{i2} = b_0 + b_1x_{i1} + u_{i2}.$$

$$\text{Capital Demand: } y_{i3} = c_0 + c_1x_{i2} + u_{i3}.$$

Where x_{i1} might be the wage and x_{i2} might be the rental rate of capital. Suppose we have to estimate the first equation and we are given that the x 's and the u 's are independent. This gives us the following population moment conditions:

$$E[u_{i1}] = 0.$$

$$E[x_{i1}u_{i1}] = 0.$$

$$E[x_{i2}u_{i1}] = 0.$$

Since the first equation has three parameters, we need at least three moment conditions. We also need the moment conditions to be RELEVANT! A moment condition like $E[x_{i1}u_{i2}] = 0$ would be ok but it wouldn't help explain the first equation. Thus our moments would be of the form $q(z_i, \beta)$ with $z_i = (y_{i1}, y_{i2}, y_{i3}, x_{i1}, x_{i2})$ and $\beta = (A, \alpha_1, \alpha_2)'$. Specifically,

$$q(z_i, \beta) = \begin{bmatrix} y_{i1} - Ay_{i2}^{\alpha_1}y_{i3}^{\alpha_2} \\ x_{i1}[y_{i1} - Ay_{i2}^{\alpha_1}y_{i3}^{\alpha_2}] \\ x_{i2}[y_{i1} - Ay_{i2}^{\alpha_1}y_{i3}^{\alpha_2}] \end{bmatrix}.$$

With $E[q(z_i, \beta)] = 0$. We would then construct our GMM estimator as a quadratic form in the moments but what goes in the middle? That comes next.

- Consider the objective function of the GMM estimator:

$$Q_n(z_1, \dots, z_n, \beta) = S_n(z_1, \dots, z_n, \beta)' \hat{\Xi}_n S_n(z_1, \dots, z_n, \beta).$$

Where usually we have:

$$S_n(z_1, \dots, z_n, \beta) = \frac{1}{n} \sum_{i=1}^n q(z_i, \beta),$$

with,

$$E[S_n(\beta_0)] = \frac{1}{n} \sum_{i=1}^n E[q(z_i, \beta_0)] = 0.$$

- Note that when we have a quadratic form like $X(\beta)'AX(\beta)$ and we differentiate:

$$\frac{\partial}{\partial \beta}[X(\beta)'AX(\beta)] = [2\frac{\partial X}{\partial \beta}AX(\beta)]'.$$

- So consider the FOC of the objective function for the GMM estimator (we omit the scaling factor $n^{1/2}$):

$$\frac{\partial S_n(z_1, \dots, z_n, \hat{\beta}_n)}{\partial \beta'} \hat{\Xi}_n S_n(z_1, \dots, z_n, \hat{\beta}_n) = 0.$$

Or written more compactly,

$$\nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n S_n(\hat{\beta}_n) = 0.$$

- Now consider a first order Taylor series expansion of the last term (utilizing the mean value theorem so we get equality):

$$S_n(\hat{\beta}_n) = S_n(\beta_0) + \nabla_{\beta} S_n(\tilde{\beta})(\hat{\beta}_n - \beta_0).$$

- Substituting in:

$$\begin{aligned} 0 &= \nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n \left[S_n(\beta_0) + \nabla_{\beta} S_n(\tilde{\beta})(\hat{\beta}_n - \beta_0) \right] \\ 0 &= \nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n S_n(\beta_0) + \nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n \nabla_{\beta} S_n(\tilde{\beta})(\hat{\beta}_n - \beta_0) \\ \nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n \nabla_{\beta} S_n(\tilde{\beta})(\hat{\beta}_n - \beta_0) &= -\nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n S_n(\beta_0) \\ (\hat{\beta}_n - \beta_0) &= -\left(\nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n \nabla_{\beta} S_n(\tilde{\beta}) \right)^{-1} \nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n S_n(\beta_0) \end{aligned}$$

- So we have:

$$\underbrace{\hat{\beta}_n - \beta_0}_{p_{\beta} \times 1} = -\left(\underbrace{\nabla_{\beta'} S_n(\hat{\beta}_n)}_{p_{\beta} \times p_q} \underbrace{\hat{\Xi}_n}_{p_q \times p_q} \underbrace{\nabla_{\beta} S_n(\tilde{\beta})}_{p_q \times p_q} \right)^{-1} \nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n \underbrace{S_n(\beta_0)}_{p_q \times 1}.$$

Asymptotic Normality Result for GMM Estimators

- **Assumptions 5.2**

- (a) Parameter Spaces are compact

- (b) S_n is C^1 .
- (c) $\hat{\beta}_n \xrightarrow{p} \beta_0$ as $n \rightarrow \infty$ and $\beta_0 \in B$. We need the true parameter to be in the strict interior so we can do a Taylor expansion around it. Also, $\hat{\Xi}_n \xrightarrow{p} \Xi_0$.
- (d) $\nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n S_n(\hat{\beta}_n) = o_p(1)$. This means that our estimator, $\hat{\beta}_n$ satisfies our FOC up to an error of magnitude $o_p(1)$.
- (e) $\nabla_{\beta} S_n(\tilde{\beta}) \xrightarrow{p} G_0$, for any consistent estimator, $\tilde{\beta}$. Note $\hat{\beta}_n$ is also consistent.
- (f) $n^{1/2} S_n(\beta_0) \xrightarrow{d} \xi$. That is, the “Normalized Score” converges to some real valued random vector.

- **Theorem** If assumptions 5.2 hold, then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow^d -[G_0' \Xi_0 G_0]^{-1} G_0' \Xi_0 \xi.$$

And if $\xi \sim N(0, \Lambda_0)$, then,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1} B_0 A_0^{-1}),$$

with $A_0 = G_0' \Xi_0 G_0$ and $B_0 = G_0' \Xi_0 \Lambda_0 \Xi_0 G_0$.

- So if assumptions 5.2 are satisfied, then the GMM estimator is consistent and asymptotically normal.

24.2 Asymptotic Normality - Catalogues of Assumptions

General Method of Moments Estimators - General

- So consider the case of GMM where our objective function is of the form:

$$Q_n(z_1, \dots, z_n, \beta) = \left[\frac{1}{n} \sum_{i=1}^n q(z_i, \beta) \right]' \hat{\Xi}_n \left[\frac{1}{n} \sum_{i=1}^n q(z_i, \beta) \right].$$

- Suppose $\hat{\beta}_n$ is a unique, consistent estimator that minimizes the above objective function. Thus all assumptions (4.4) of a consistent GMM estimator must be satisfied. If we add that $q(z_i, \beta)$ is differentiable and other certain domination conditions on the gradient, we get the following theorem.
- **Theorem** Given the assumptions of a consistent estimator and nice properties of the gradient, we have the following for the GMM estimator, $\hat{\beta}_n$:

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1} B_0 A_0^{-1}),$$

with:

$$A_0 = G_0' \Xi_0 G_0, \quad \text{and} \quad B_0 = G_0' \Xi_0 \Lambda_0 \Xi_0 G_0,$$

and:

$$G_0 = E[\nabla_{\beta} q(z_i, \beta_0)], \quad \text{and} \quad \Lambda_0 = E[q(z_i, \beta_0) q(z_i, \beta_0)'].$$

Furthermore, let:

$$\hat{A}_n = \hat{G}'_n \hat{\Xi}_n \hat{G}_n, \quad \text{and} \quad \hat{B}_n = \hat{G}'_n \hat{\Xi}_n \hat{\Lambda}_n \hat{\Xi}_n \hat{G}_n,$$

with:

$$\hat{G}_n = n^{-1} \sum \nabla_{\beta} q(z_i, \hat{\beta}_n), \quad \text{and} \quad \hat{\Lambda}_n = n^{-1} \sum q(z_i, \hat{\beta}_n) q(z_i, \hat{\beta}_n)'$$

Then:

$$\hat{A}_n \rightarrow^p A_0, \quad \hat{B}_n \rightarrow^p B_0, \quad \hat{G}_n \rightarrow^p G_0, \quad \hat{\Lambda}_n \rightarrow^p \Lambda_0.$$

And finally:

$$(\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1}) \rightarrow^p (A_0^{-1} B_0 A_0^{-1}).$$

Brilliant.

25 Lecture 25: May 5, 2005

25.1 Review: Asymptotic Normality and Consistency of GMM Estimators

- Consider the GMM objective function:

$$Q(z_1, \dots, z_n, \beta) = \left[n^{-1} \sum q(z_i, \beta) \right]' \hat{\Xi}_n \left[n^{-1} \sum q(z_i, \beta) \right].$$

- We get consistency with the following key conditions:
 - (1) $\hat{\Xi}_n \rightarrow \Xi_0$. Note this could be equal to the identity in some cases.
 - (2) $E[q(z_i, \beta)] = 0$ at $\beta = \beta_0$. The nonstochastic analogue to the moment condition is uniquely defined.
 - (3) q is continuous in β .
 - (4) z_i are iid.
 - (5) Dominance condition.
- To get asymptotic normality, we add:
 - (6) β_0 is in the interior of the space, B .
 - (7) q is C^1 . So $\nabla_{\beta} q(z_i, \beta)$ exists.
 - (8) Dominance on the derivatives.

- Recall:

$$E[\nabla_{\beta} q(z_i, \beta_0)] = G_0,$$

and,

$$E[q(z_i, \beta_0)q'(z_i, \beta_0)] = \Lambda_0,$$

the variance/covariance matrix of the moments.

- We had the result that if $\hat{\Xi}_n \rightarrow \Xi_0$, then:

$$n^{-1/2} \sum q(z_i, \beta_0) \rightarrow^d N(0, \Lambda_0).$$

Which means:

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}B_0A_0^{-1}).$$

with $A_0 = G_0'\Xi_0G_0$ and $B_0 = G_0'\Xi_0\Lambda_0\Xi_0G_0$.

- So consider the special case when $\hat{\Xi}_n = I_n$. Then the variance/covariance matrix becomes:

$$\begin{aligned} A_0^{-1} B_0 A_0^{-1} &= (G_0' \Xi_0 G_0)^{-1} G_0' \Xi_0 \Lambda_0 \Xi_0 G_0 (G_0' \Xi_0 G_0)^{-1} \\ &= (G_0' G_0)^{-1} G_0' \Lambda_0 G_0 (G_0' G_0)^{-1} \end{aligned}$$

This simplified things a bit, but we will show that using the identity matrix is not efficient.

- How about the special case where $\hat{\Xi}_n = \hat{\Lambda}_n^{-1} \rightarrow \Lambda_0^{-1}$. Then the variance/covariance matrix becomes:

$$\begin{aligned} A_0^{-1} B_0 A_0^{-1} &= (G_0' \Xi_0 G_0)^{-1} G_0' \Xi_0 \Lambda_0 \Xi_0 G_0 (G_0' \Xi_0 G_0)^{-1} \\ &= (G_0' \Lambda_0^{-1} G_0)^{-1} G_0' \Lambda_0^{-1} \Lambda_0 \Lambda_0^{-1} G_0 (G_0' \Lambda_0^{-1} G_0)^{-1} \\ &= (G_0' \Lambda_0^{-1} G_0)^{-1} G_0' \Lambda_0^{-1} G_0 (G_0' \Lambda_0^{-1} G_0)^{-1} \\ &= (G_0' \Lambda_0^{-1} G_0)^{-1} \end{aligned}$$

It can be shown that this indeed is the MOST efficient weighting matrix we can put in there. The inverse of the variance/covariance matrix of the moments weights the moments in such a way that the variance of the resulting estimator is minimized.

- So now what do we do to estimate things like G_0 and Λ_0 ? Logical estimators would be $\hat{A}_n = \hat{G}_n' \hat{\Xi}_n \hat{G}_n$ and $\hat{B}_n = \hat{G}_n' \hat{\Xi}_n \hat{\Lambda}_n \hat{\Xi}_n \hat{G}_n$, with:

$$\hat{G}_n = n^{-1} \sum \nabla_{\beta} q(z_i, \hat{\beta}_n),$$

and,

$$\hat{\Lambda}_n = n^{-1} \sum q(z_i, \hat{\beta}_n) q'(z_i, \hat{\beta}_n).$$

Note that $\hat{\beta}_n$ is ANY consistent estimator. It doesn't have to be the one that we end up determining to be the best, it just has to be consistent. Using this result we have the following algorithm.

Estimating $\hat{\Lambda}_n$

- Since we don't have consistent estimates of β to start out with, we can't do GMM unless we can form some estimate of the variance/covariance matrix of the moments. Since we just need a consistent estimate of β , consider following a 2 step process:
- Step 1. Using $\hat{\Xi}_n = I_n$, our objective function results in:

$$\tilde{\beta} = \arg \min \left[n^{-1} \sum q(z_i, \beta) \right]' \left[n^{-1} \sum q(z_i, \beta) \right].$$

This gives us a consistent estimator: $\tilde{\beta}$. Now use this estimator to calculate:

$$\hat{\Lambda}_n = n^{-1} \sum q(z_i, \tilde{\beta}_n) q'(z_i, \tilde{\beta}_n).$$

- Step 2. Solve for our (consistent and most efficient) estimator:

$$\hat{\beta} = \arg \min \left[n^{-1} \sum q(z_i, \beta) \right]' \hat{\Lambda}_n^{-1} \left[n^{-1} \sum q(z_i, \beta) \right].$$

- Note we could continue this process and iterate again and this process is called Continuously Updated GMM. Sounds exciting.

26 Lecture 26: May 10, 2005

26.1 Asymptotic Normality - Catalogues of Assumptions

Least Mean Distance Estimators - General

- Consider the objective function of the LMD estimator:

$$Q(z_1, \dots, z_n; \beta) = \frac{1}{n} \sum_{i=1}^n q(z_i, \beta),$$

where q is now a scalar real valued function.

- **Assumptions 6.1** for Consistency:

- (a) Compact parameter space.
- (b) q is continuous in β (not necessarily in z).
- (c) z_i are iid.
- (d) Dominance on q .
- (e) β_0 is a unique minimizer of the non-stochastic analogue, $E[q(z_i, \beta)]$.

- **Assumptions 6.1*** for Asymptotic Normality:

- (f) β_0 is an interior point of B .
- (g) q is C^2 .
- (h) Dominance conditions on the gradient and hessian of q .
- (i) The expected value of the hessian is non-singular.

- **Theorem 6.1** If the assumptions above hold, we have:

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}B_0A_0^{-1}),$$

with:

$$A_0 = E[\nabla_{\beta\beta}q(z_i, \beta_0)],$$

and,

$$B_0 = E[\nabla_{\beta'}q(z_i, \beta_0)\nabla_{\beta}q(z_i, \beta_0)].$$

Estimating as usual with:

$$\hat{A}_n = E[\nabla_{\beta\beta}q(z_i, \hat{\beta}_n)],$$

and,

$$\hat{B}_n = E[\nabla_{\beta'}q(z_i, \hat{\beta}_n)\nabla_{\beta}q(z_i, \hat{\beta}_n)].$$

Yields:

$$\hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1} \rightarrow^p A_0^{-1}B_0A_0^{-1}.$$

Note we call this variance term, $A_0^{-1}B_0A_0^{-1}$, a “sandwich” variance.

Least Mean Distance Estimators - Nonlinear Least Squares

- Consider the objective function of the NLLS (LMD) estimator:

$$Q(z_1, \dots, z_n; \beta) = \frac{1}{n} \sum_{i=1}^n (y_i - g(x_i, \beta))^2.$$

- This comes from the model: $y_i = g(x_i, \beta_0) + \epsilon_i$, where we assume:

$$E[\epsilon_i | x_i] = 0 \implies E[y_i | x_i] = g(x_i, \beta_0).$$

Note that $z_i = [y_i, x_i]$ is iid but this does not rule out the possibility of conditional heteroskedasticity:

$$E[\epsilon_i^2 | x_i] = \sigma_i^2 \neq \sigma^2.$$

- **Assumptions 6.2** for Consistency:

- (a) Compact parameter space.
- (b) g is continuous in β (not necessarily in z).
- (c) z_i are iid.
- (d) $E[y_i | x_i] = g(x_i, \beta_0)$.
- (e) Dominance on g .
- (f) β_0 is a unique minimizer of the non-stochastic analogue, $E[(g(z_i, \beta_0) - g(x_i, \beta))^2]$.

- **Assumptions 6.2*** for Asymptotic Normality:

- (g) β_0 is an interior point of B .
- (h) g is C^2 .
- (i) Dominance conditions on the gradient and hessian of g .
- (j) The expected value of the hessian is non-singular.

- **Theorem 6.2** If the assumptions above hold, we have:

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1} B_0 A_0^{-1}),$$

with:

$$A_0 = E[\nabla_{\beta'} g(x_i, \beta_0) \nabla_{\beta} g(x_i, \beta_0)],$$

and,

$$B_0 = E[(y_i - g(x_i, \beta_0))^2 \nabla_{\beta'} g(x_i, \beta_0) \nabla_{\beta} g(x_i, \beta_0)].$$

Estimating as usual with \hat{A}_n and \hat{B}_n , yields:

$$\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} \rightarrow^p A_0^{-1} B_0 A_0^{-1}.$$

- Note that if there is no Heteroskedasticity:

$$E[(y_i - g(x_i, \beta_0))^2 | x_i] = E[\epsilon_i^2 | x_i] = \sigma^2,$$

then:

$$B_0 = E[(y_i - g(x_i, \beta_0))^2 \underbrace{\nabla_{\beta'} g(x_i, \beta_0) \nabla_{\beta} g(x_i, \beta_0)}_{A_0}] = \sigma^2 A_0.$$

So,

$$A_0^{-1} B_0 A_0^{-1} = \sigma^2 A_0^{-1}.$$

- For the case of a linear model: $y_i = x_i \beta + \epsilon_i$, (with homoskedasticity):

$$A_0 = E[\nabla_{\beta'} g(x_i, \beta_0) \nabla_{\beta} g(x_i, \beta_0)] = E[x_i x_i'],$$

so,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, \sigma^2 (X'X)^{-1}),$$

as usual.

- For the case of a linear model with heteroskedasticity:

$$\hat{B}_n = \frac{1}{n} \sum_{i=1}^n (y_i - g(x_i, \beta_0))^2 x_i x_i' = \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 x_i x_i'.$$

So,

$$\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} = \left(\sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 x_i x_i' \left(\sum_{i=1}^n x_i x_i' \right)^{-1}.$$

Least Mean Distance Estimators - Maximum Likelihood

- Consider the objective function of the ML (LMD) estimator:

$$Q(z_1, \dots, z_n; \beta) = -\frac{1}{n} \sum_{i=1}^n \ln f(z_i, \beta),$$

where f is the likelihood function.

- If the true distribution is the same as the one you base your likelihood function on, then you have a true maximum likelihood estimator, but if the true distribution is say t , and you use a normal, then you have estimated a pseudo-maximum likelihood estimator. In the former case, the sandwich will simplify, but in the latter, you have to hold on your sandwich.

- So we are actually interested in the conditional likelihood function, so we write the objective function:

$$Q(z_1, \dots, z_n; \beta) = \frac{1}{n} \sum_{i=1}^n q(z_i, \beta) = -\frac{1}{n} \sum_{i=1}^n \ln f(y_i | x_i; \beta).$$

- **Assumptions 6.3** for Consistency:

- (a) Compact parameter space.
- (b) q is continuous in β (not necessarily in z).
- (c) z_i are iid.
- (d) Dominance on q .
- (e) β_0 is a unique minimizer of the non-stochastic analogue, $E[q(z_i, \beta)]$.

- **Assumptions 6.3*** for Asymptotic Normality:

- (f) β_0 is an interior point of B .
- (g) q is C^2 .
- (h) Dominance conditions on the gradient and hessian of q .
- (i) The expected value of the hessian is non-singular.

- **Theorem 6.3** If the assumptions above hold, we have:

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1} B_0 A_0^{-1}),$$

with:

$$A_0 = E[\nabla_{\beta\beta} q(z_i, \beta_0)] = -E[\nabla_{\beta\beta} \ln f(y_i | x_i; \beta_0)],$$

and,

$$B_0 = E[\nabla_{\beta\beta} q(z_i, \beta_0)] = E[\nabla_{\beta'} \ln f(y_i | x_i; \beta_0) \nabla_{\beta} \ln f(y_i | x_i; \beta_0)].$$

Estimating as usual with \hat{A}_n and \hat{B}_n , yields:

$$\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} \rightarrow^p A_0^{-1} B_0 A_0^{-1}.$$

HOWEVER, since we have a likelihood as our objective function, we can interchange integration and differentiation, which implies $A_0 = B_0$. Thus,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}).$$

Is this only true when you have the correct distribution ??? When can we assume this condition ???

27 Lecture 27: May 12, 2005

27.1 Numerical Methods - General

- Suppose we want to minimize an objective function:

$$Q(z_1, \dots, z_n, \beta) = Q(\beta),$$

following the convention in the notes that we are seeking a valley not a peak.

- Suppose β^0 is our initial parameter value. How do we pick a direction to go to reach the min? We would like to have:

$$Q(\beta^0) > Q(\beta^1).$$

- So consider a step, ξ_g , such that:

$$\beta_g = \beta_{g-1} + \xi_g.$$

- Formulating ξ_g will be addressed next, but once we have it and we start iterating, how do we know when to stop? There are several criteria:

- (1) Euclidean distance criteria:

$$(\beta_g - \beta_{g-1})'(\beta_g - \beta_{g-1}) < \epsilon,$$

for some $\epsilon > 0$.

- (2) Objective function criteria:

$$Q(\beta_{g-1}) - Q(\beta_g) < \epsilon.$$

- (3) Score vector criteria:

$$\left[\frac{\partial Q(\beta_g)}{\partial \beta} \right] \left[\frac{\partial Q(\beta_g)}{\partial \beta} \right]' < \epsilon.$$

- (4) Maximum number of iterations is reached.

- If any or all of the criteria above are not satisfied, then try a new initial parameter vector.
- So how does the program determine ξ_g ? Denote:

$\delta_g \equiv$ The direction of the step.

$\lambda_g \equiv$ The magnitude of the step.

- Define the directional derivative:

$$\left. \frac{\partial Q(\beta_{g-1} + \lambda \delta_g)}{\partial \lambda} \right|_{\lambda=0} = \frac{\partial Q(\beta_{g-1})}{\partial \beta} \delta_g.$$

- So consider a choice of δ_g :

$$\delta_g = -P_g \frac{\partial Q(\beta_{g-1})}{\partial \beta'}.$$

where P_g is a POSITIVE definite matrix. Then, our directional derivative becomes:

$$\left. \frac{\partial Q(\beta_{g-1} + \lambda \delta_g)}{\partial \lambda} \right|_{\lambda=0} = -\frac{\partial Q(\beta_{g-1})}{\partial \beta} P_g \frac{\partial Q(\beta_{g-1})}{\partial \beta'}.$$

Or,

$$\left. \frac{\partial Q(\beta_{g-1} + \lambda \delta_g)}{\partial \lambda} \right|_{\lambda=0} = -X' P_g X < 0,$$

since we have a quadratic form with a positive definite matrix in the middle. So we are going in the right direction - Down!

- Thus, the updating formula for β becomes:

$$\beta_g = \beta_{g-1} - \lambda_g P_g \frac{\partial Q(\beta_{g-1})}{\partial \beta'}.$$

27.2 Numerical Methods - Gradient Methods

Method of Steepest Descent

- Let $P_g = I_K$. This may take a long time to converge and is NOT recommended.

Newton-Raphson Method

- Let:

$$P_g = \left[\frac{\partial^2 Q(\beta_{g-1})}{\partial \beta \partial \beta'} \right]^{-1}.$$

- This term comes from a first order Taylor series expansion of the gradient (which should be zero at an optimum).
- If the objective function is linear-quadratic, then we would have an exact Taylor expansion (the remainder terms are all zero). In this case, we would converge to our optimum in one step and the only thing we would have to choose would be λ_g .
- However, there are major problems with this method including:

- (1) P_g needs to be positive definite (we're minimizing) but this will only be true, in general, in the neighborhood of the min. If this isn't the case at our starting parameter vector, we may not converge.
- (2) We need second order derivatives which are often hard to calculate. Analytic equations are usually not available (even for the first derivative!) so numerical differentiation is computationally taxing.

Gauss-Newton Method Method

- See Prucha notes.

Berndt, Hall, Hall, and Hausman (BHHH) Method

- Consider our objective function:

$$Q(\beta) = n^{-1} \sum_{i=1}^n q_i(\beta),$$

where $q_i(\beta)$ is the (negative of) a log likelihood function. Then:

$$E \left[\frac{\partial^2 Q(\beta^0)}{\partial \beta \partial \beta'} \right] = \sum_{i=1}^n E \left[\frac{\partial q_i(\beta^0)}{\partial \beta'} \frac{\partial q_i(\beta^0)}{\partial \beta} \right].$$

- Then the term on the RHS of the equation above is ALWAYS positive definite, even if we start far from the optimum. It also doesn't involve second order derivatives which helps with computation time. Thus, our matrix is:

$$P_g = \left[\sum \frac{\partial q_i(\beta_{g-1})}{\partial \beta'} \frac{\partial q_i(\beta_{g-1})}{\partial \beta} \right].$$

Remarks on Numerical Optimization

- There is a certain art to choose good starting values and it depends crucially on the economic model you are working on. If possible, make the problem smaller initially by imposing a fixed value for one parameter (one that you are more confident in) and then optimizing the other. Then use the optimized value for the starting value of a further optimization of both parameters jointly.
- You can also linearize a model (eg, take a 3rd or 4th order Taylor series expansion of an exponential function) to get starting values.
- Finally, be sure to check multiple starting values to make sure you always converge to the same optimum. You don't want to end up at a local optimum.

28 Final Review

28.1 Notes from Problem Sets - Part I

- $tr(ABC) = tr(BCA) = tr(CAB)$.
- $X'X$ is symmetric, psd. $c'X'Xc = d'd \geq 0$. pd if X is full rank.
- If A is real symmetric, then $A = c\Lambda c'$ with $cc' = c'c = I$ so $c' = c^{-1}$ and:

- (1) $|A| = \prod \lambda_i$, $tr(A) = \sum \lambda_i$, $r(A) = r(\Lambda)$.
- (2) A p.d. $\Rightarrow A^{-1}$ p.d.
- (3) A p.d. $\Leftrightarrow \lambda_i < 0 \forall i$.
- (4) A p.d. $\Rightarrow A = pp'$ with $p = c\Lambda^{1/2}$.

- A idempotent, $AA = A$. Then:

- (1) $\lambda_i \in \{0, 1\}$.
- (2) $r(A) = tr(A)$.

- Variance decomp:

$$\sigma_x^2 = E[x^2] - (E[x])^2.$$

$$\Sigma_y = E[yy'] - E[y]E[y]'$$

- $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$ for $\theta \in \mathfrak{R}$. In general:

$$MSE = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'] = E[(\hat{\theta} - E[\hat{\theta}])(\hat{\theta} - E[\hat{\theta}])'] + (E[\hat{\theta}] - \theta)(E[\hat{\theta}] - \theta)' = Var(\hat{\theta}) + Bias^2.$$

- If $u \sim N(0, \sigma^2 I_n)$ and A is symmetric, idempotent with rank p , then:

$$\frac{u' Au}{\sigma^2} \sim \chi^2(p).$$

- If $u \sim N(0, \Sigma)$ where Σ is nonsingular and u is $nx1$, then:

$$u' \Sigma^{-1} u \sim \chi^2(n).$$

- Consider $y = X\beta + u$. Then (if X includes an intercept):

$$R^2 = 1 - \frac{y' M_x y}{y' M_{e_T} y}.$$

Or,

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} = 1 - \frac{\frac{1}{T} \hat{u}' \hat{u} - \bar{\hat{u}}^2}{\frac{1}{T} y' y - \bar{y}^2} = 1 - \frac{\hat{u}' \hat{u}}{y' y - T \bar{y}^2}.$$

- Sums of Squares: $TSS = ESS + RSS$:

$$TSS = \frac{1}{T}y'y - \bar{y}^2.$$

$$RSS = \frac{1}{T}\hat{y}'\hat{y} - \bar{\hat{y}}^2.$$

$$ESS = \frac{1}{T}\hat{u}'\hat{u} - \bar{\hat{u}}^2 = \frac{1}{T}\hat{u}'\hat{u}.$$

- Fisher information matrix for $y = X\beta + u$:

$$J(\theta)^{-1} = -E \left[\frac{\partial^2 \ln(L)}{\partial \theta \partial \theta'} \right]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}.$$

- Wald Test: Note $\hat{\beta} \sim N(\beta, \Sigma)$. To test a linear hypothesis on β , construct:

$$(R\hat{\beta} - q)'[R\Sigma R']^{-1}(R\hat{\beta} - q) \sim \chi^2(r(\Sigma)).$$

Note (CHECK THIS!):

$$(R\hat{\beta} - q)'[R\Sigma R']^{-1}(R\hat{\beta} - q)/G \sim F(G, T - K).$$

But I think the Wald is just the numerator of this statistic which is distributed χ^2 .

28.2 Notes from Problem Sets - Part II

- Unbiased but not consistent: $X_i \sim iid(\mu, \sigma^2)$, with:

$$\hat{\mu} = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} X_i + \frac{1}{2} X_n.$$

- Given $X_1 \dots X_n$ iid Bernoulli(p) and estimator \bar{X}_n . Since $E[\bar{X}_n] = p$, $Var(\bar{X}_n) = p(1-p) < \infty$, and X_i is iid, by Lindeberg Levy CLT,

$$\sqrt{n}(\bar{X}_n - p) \rightarrow^d N(0, p(1-p)).$$

28.3 Key Lectures Notes - Part I

- OLS:

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

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

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

- Note $MX = X'M' = X'M = 0$. $MM = 0$, $MX = 0$.

- $X'\hat{u} = \hat{y}'\hat{u} = 0$.
- Given $y = X\beta + u$, $\hat{u} = My = Mu$.
- $y = \hat{y} + \hat{u}$
- Adjusted R^2 :

$$R_a^2 = R^2 - \left(\frac{K-1}{T-K} \right) (1 - R^2).$$

- $VC(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(X'X)^{-1}X'uu'X(X'X)^{-1}] = \sigma^2(X'X)^{-1}$.
- Gauss Markov: Given A1-A3, OLS for β is BLUE.
- $E[s^2] = \sigma^2$. Use the fact that $\hat{u} = Mu$.

28.4 Laws of Large Numbers

- Kolmogorov Strong: $Z_t \sim iid$, $E[Z_t] = \mu < \infty$. Then $\frac{1}{n} \sum Z_t \xrightarrow{as} \mu$.
- Chebychev: Z_t rvs with $E[Z_t] = \mu_t$, $Var[Z_t] = \sigma_t^2$, and $\frac{1}{n} \sum \sigma_t^2 \rightarrow \sigma^2$. Then $\frac{1}{n} \sum Z_t \xrightarrow{p} \mu_t$.

28.5 Central Limit Theorems

- Lindeberg-Levy: $Z_t \sim iid$ with $E[Z_t] = 0$, $Var(Z_t) = \sigma^2 < \infty$. Then $\frac{1}{\sqrt{n}} \sum Z_t \xrightarrow{d} N(0, \sigma^2)$. **Regardless of the distribution of Z_t . Just needs to be iid.
- Lindeberg-Feller: Z_t independent random variable (not necessarily iid) with $E[Z_t] = 0$ and $Var(Z_t) = \sigma_t^2 < \infty$. If the lindeberg condition holds:

$$\frac{1}{\sigma_{(n)}} \sum Z_t \xrightarrow{d} N(0, 1),$$

where $\sigma_{(n)} = \sqrt{\sum \sigma_t^2}$. Lindeberg condition states that the tails are not too heavy.

- Regression analysis. Suppose X is $n \times k$, non-stochastic, and $u_t \sim iid(0, \sigma^2)$.

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

Scaling:

$$\sqrt{n}(\hat{\beta} - \beta) = (1/nX'X)^{-1} \frac{1}{\sqrt{n}}X'u.$$

Assume:

$$\frac{1}{n}X'X \rightarrow Q, \text{ positive definite.}$$

Then, the theorem says:

$$\sqrt{n}(\hat{\beta} - \beta) = \underbrace{(1/nX'X)^{-1}}_{\rightarrow Q^{-1}} \underbrace{\frac{1}{\sqrt{n}}X'u}_{\rightarrow N(0, \sigma^2 Q)} \rightarrow N(0, Q^{-1}\sigma^2 Q Q^{-1}) \equiv N(0, \sigma^2 Q^{-1}).$$

Denote:

$$\eta_n = \sqrt{n}(\hat{\beta} - \beta).$$

$$\hat{\beta}_n = \beta + \underbrace{\frac{1}{\sqrt{n}} \eta_n}_{\substack{\rightarrow R.V. \\ \rightarrow 0}} \xrightarrow{p} \beta.$$

Thus $\hat{\beta}_n$ is consistent for β .

28.6 Key Lectures Notes - Part II

- In the limit, the ML estimator attains the R-C lower bound.
- Distributions of OLS estimators. Assume:

$$\frac{1}{T}X'X \rightarrow Q.$$

Consider $E[\frac{1}{T}X'u] = 0$ and $Var[\frac{1}{T}X'u] = \frac{1}{T^2}X'XE[uu'] \rightarrow 0$. Thus,

$$\frac{1}{T}X'u \xrightarrow{p} 0.$$

By CLT 4.7,

$$\frac{1}{\sqrt{T}}X'u \xrightarrow{d} N(0, \sigma^2 Q).$$

So if we have these last two equations we can get the limiting distributions of the OLS estimators.

- Limiting distributions (β):

$$\sqrt{T}(\hat{\beta}_T - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1}).$$

Or,

$$\hat{\beta}_T \approx N(\beta, s^2(X'X)^{-1}).$$

- Limiting distributions (σ^2):

$$s^2 = \frac{T}{T-K} \left[\frac{u'u}{T} - \frac{1}{T}u'X \left(\frac{1}{T}X'X \right)^{-1} \frac{1}{T}X'u \right] \rightarrow \sigma^2,$$

because $u'u/T \xrightarrow{as} \sigma^2$ by Khinchines LLN. Thus the limiting distribution:

$$\sqrt{T}(s^2 - \sigma^2) \rightarrow^d N(0, \mu_4 - \sigma^4).$$

- Limiting distributions (t):

$$t \rightarrow N(0, 1).$$

- Limiting distributions (F):

$$F(G, T - K) \rightarrow^d \chi^2(G)/G.$$

- Nonlinear objective functions:

$$\text{Least Mean Distance - General: } Q_n = n^{-1} \sum_{i=1}^n q(z_i, \beta).$$

$$\text{Least Mean Distance - NLS: } Q_n = n^{-1} \sum_{i=1}^n [y_i - g(x_i, \alpha)]^2.$$

$$\text{Least Mean Distance - ML: } Q_n = n^{-1} \sum_{i=1}^n -f(y_i|x_i; \alpha_0).$$

$$\text{GMM - General: } Q_n = [n^{-1} \sum_{i=1}^n q(z_i, \beta)]' \hat{\Xi}_n [n^{-1} \sum_{i=1}^n q(z_i, \beta)].$$

- **Remark** A sufficient set of conditions for an estimator to be identifiably unique is:

- (1) \bar{Q} continuous.
- (2) B compact (the parameter space).
- (3) β_0 must be a unique minimizer.

- **Theorem**

$$R_n(\omega, \beta) = Q_n(z_1, \dots, z_n, \beta),$$

$$\bar{R}(\beta) = \bar{Q}(\beta).$$

Now assume:

$$\text{Sup}_{\beta \in B} |R_n(\omega, \beta) - \bar{R}(\beta)| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty.$$

Then for our minimizer, $\hat{\beta}_n$, such that:

$$R(\omega, \hat{\beta}_n) = \text{Inf}_{\beta \in B} R(\omega, \beta),$$

we have:

$$\hat{\beta}_n \xrightarrow{p} \beta_0 \text{ as } n \rightarrow \infty.$$

- We really need a ULLN result to show consistency.

- **Theorem** Uniform Law of Large Numbers. Assume the following holds:

- (1) z_i are iid.
- (2) q can be either scalar or vector valued.
- (3) $q(\cdot, \beta)$ is measurable for each $\beta \in B$.
- (4) $q(z, \cdot)$ is continuous for each $z \in Z$.
- (5) B is compact.
- (6) Domination Condition: $|q(z, \beta)| \leq h(z)$ with $E[h(z)] < \infty$.

Then we have:

$$\text{Uniform Convergence: } \text{Sup}_{\beta \in B} \left| \frac{1}{n} \sum q(z_i, \beta) - E[q(z_i, \beta)] \right| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty,$$

and,

$$E[q(z_i, \beta)] \text{ if finite and continuous in } \beta.$$

28.7 Consistency - Catalogues of Assumptions

Least Mean Distance Estimators - General

- **Assumptions 4.1**

- (a) $q : Z \times B \mapsto \mathfrak{R}$. q is real valued.
- (b) B is compact.
- (c) $q(\cdot, \beta)$ is measurable for each $\beta \in B$. $q(z, \cdot)$ is continuous for each $z \in Z$.
- (d) z_i is iid.
- (e) Domination: $\text{Sup}_{\beta \in B} |q(z_i, \beta)| < \infty$.

- If assumptions 4.1 are satisfied then we have the following ULLN:

$$\text{Sup}_{\beta \in B} \left| \frac{1}{n} \sum q(z_i, \beta) - E[q(z_i, \beta)] \right| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty,$$

and,

$$E[q(z_i, \beta)] \text{ is continuous.}$$

- If we add to assumptions 4.1 the assumption that β_0 is a UNIQUE minimizer of \bar{R} , then we get a consistent estimator:

$$\hat{\beta}_n \xrightarrow{as} \beta_0 \text{ as } n \rightarrow \infty.$$

Least Mean Distance Estimators - Nonlinear Least Squares

- **Assumptions 4.2**

- (a) g is real valued.
- (b) B is compact.
- (c) $g(\cdot, \beta)$ is measurable for each $\beta \in B$. $g(x, \cdot)$ is continuous for each $x \in p_x$.
- (d) $z_i = [y_i, x_i]$ is iid.
- (e) $E[y_i|x_i] = g(x_i, \beta_0)$.
- (f) Domination: $E[(y_i - g(x_i, \beta))^2] < \infty$ and $Sup_{\beta \in B} g(x_i, \beta)^2 < \infty$.

- **Theorem** If assumptions 4.2 are satisfied and $E[(g(x_i, \beta_0) - g(x_i, \beta))^2] > 0 \forall \beta \neq \beta_0$, then:

$$\hat{\beta}_n \rightarrow^{as} \beta_0 \text{ as } n \rightarrow \infty.$$

Least Mean Distance Estimators - Maximum Likelihood

- **Assumptions 4.3**

- (a) q is real valued.
- (b) B is compact.
- (c) $q(\cdot, \beta)$ is measurable for each $\beta \in B$. $q(z, \cdot)$ is continuous for each $z \in Z$.
- (d) $z_i = [y_i, x_i]$ is iid.
- (e) Domination: $E[Sup_{\beta \in B} |q(z_i, \beta)|] < \infty$.

- **Theorem** If assumptions 4.3 are satisfied and β_0 is a unique minimizer for $E[q(z_i, \beta)]$, then:

$$\hat{\beta}_n \rightarrow^{as} \beta_0 \text{ as } n \rightarrow \infty.$$

General Method of Moments Estimators - General

- **Assumptions 4.4**

- (a) $q : Z \times B \mapsto \mathfrak{R}^{p_q}$ is a real vector valued function.
- (b) B is compact.
- (c) $q(\cdot, \beta)$ is measurable for each $\beta \in B$. $q(z, \cdot)$ is continuous for each $z \in Z$.
- (d) z_i is iid.
- (e) Domination: $E[Sup_{\beta \in B} \|q(z_i, \beta)\|] < \infty$. where $\|\cdot\|$ is the euclidean norm.

- **Theorem** If assumptions 4.4 are satisfied and,

$$\bar{R}(\beta) = [E[q(z_i, \beta)]]' \Xi_0 E[q(z_i, \beta)] \begin{cases} \neq 0 & \text{if } \beta \neq \beta_0 \\ = 0 & \text{if } \beta = \beta_0 \end{cases}$$

Then:

$$\hat{\beta}_n \xrightarrow{as} \beta_0 \text{ as } n \rightarrow \infty.$$

28.8 Asymptotic Normality - Catalogues of Assumptions

Least Mean Distance Estimators - General

- **Assumptions 5.1**

- (a) Parameter Spaces are compact
- (b) Q_n is C^2 .
- (c) $\hat{\beta}_n \xrightarrow{p} \beta_0$ as $n \rightarrow \infty$ and $\beta_0 \in B$. We need the true parameter to be in the strict interior so we can do a Taylor expansion around it. Also, $n^{1/2}(\hat{\tau}_n - \tau_0) = O_p(1)$, a random variable.
- (d) $n^{1/2} \nabla_{\beta'} Q(\hat{\beta}_n) = o_p(1)$. This means that our estimator, $\hat{\beta}_n$ satisfies our FOC up to an error of magnitude $o_p(1)$.
- (e) $\nabla_{\beta\beta} Q_n(\tilde{\beta}_n) \xrightarrow{p} A_0$. This we have shown above.
- (f) $\nabla_{\beta\tau} Q_n(\tilde{\tau}_n, \tilde{\beta}_n) \xrightarrow{p} 0$, so the cross derivative terms are 0. τ is truly a nuisance. This was clear before we started ...
- (g) There exists a real matrix D_0 , such that:

$$-n^{1/2} \nabla_{\beta'} Q_n(z_1, \dots, z_n, \tau_0, \beta_0) = D_0 \xi_n + o_p(1).$$

- **Theorem** If assumptions 5.1 hold, then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} A_0^{-1} D_0 \xi.$$

And if $\xi \sim N(0, \Lambda_0)$, then as above,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, A_0^{-1} D_0 \Lambda_0 D_0' A_0^{-1}) \equiv N(0, A_0^{-1} B_0 A_0^{-1}),$$

with $B_0 = D_0 \Lambda_0 D_0'$.

GMM Estimators - General

- **Assumptions 5.2**

- (a) Parameter Spaces are compact
- (b) S_n is C^1 .

- (c) $\hat{\beta}_n \xrightarrow{p} \beta_0$ as $n \rightarrow \infty$ and $\beta_0 \in B$. We need the true parameter to be in the strict interior so we can do a Taylor expansion around it. Also, $\hat{\Xi}_n \xrightarrow{p} \Xi_0$.
- (d) $\nabla_{\beta'} S_n(\hat{\beta}_n) \hat{\Xi}_n S_n(\hat{\beta}_n) = o_p(1)$. This means that our estimator, $\hat{\beta}_n$ satisfies our FOC up to an error of magnitude $o_p(1)$.
- (e) $\nabla_{\beta} S_n(\tilde{\beta}) \xrightarrow{p} G_0$, for any consistent estimator, $\tilde{\beta}$. Note $\hat{\beta}_n$ is also consistent.
- (f) $n^{1/2} S_n(\beta_0) \rightarrow^d \xi$. That is, the “Normalized Score” converges to some real valued random vector.

- **Theorem** If assumptions 5.2 hold, then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow^d -[G_0' \Xi_0 G_0]^{-1} G_0' \Xi_0 \xi.$$

And if $\xi \sim N(0, \Lambda_0)$, then,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1} B_0 A_0^{-1}),$$

with $A_0 = G_0' \Xi_0 G_0$ and $B_0 = G_0' \Xi_0 \Lambda_0 \Xi_0 G_0$. Furthermore, let:

$$\hat{A}_n = \hat{G}_n' \hat{\Xi}_n \hat{G}_n, \quad \text{and} \quad \hat{B}_n = \hat{G}_n' \hat{\Xi}_n \hat{\Lambda}_n \hat{\Xi}_n \hat{G}_n,$$

with:

$$\hat{G}_n = n^{-1} \sum \nabla_{\beta} q(z_i, \hat{\beta}_n), \quad \text{and} \quad \hat{\Lambda}_n = n^{-1} \sum q(z_i, \hat{\beta}_n) q(z_i, \hat{\beta}_n)'$$

Then:

$$\hat{A}_n \xrightarrow{p} A_0, \quad \hat{B}_n \xrightarrow{p} B_0, \quad \hat{G}_n \xrightarrow{p} G_0, \quad \hat{\Lambda}_n \xrightarrow{p} \Lambda_0.$$

And finally:

$$(\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1}) \xrightarrow{p} (A_0^{-1} B_0 A_0^{-1}).$$

Brilliant.

- How about the special case where $\hat{\Xi}_n = \hat{\Lambda}_n^{-1} \rightarrow \Lambda_0^{-1}$. Then the variance/covariance matrix becomes:

$$\begin{aligned} A_0^{-1} B_0 A_0^{-1} &= (G_0' \Xi_0 G_0)^{-1} G_0' \Xi_0 \Lambda_0 \Xi_0 G_0 (G_0' \Xi_0 G_0)^{-1} \\ &= (G_0' \Lambda_0^{-1} G_0)^{-1} G_0' \Lambda_0^{-1} \Lambda_0 \Lambda_0^{-1} G_0 (G_0' \Lambda_0^{-1} G_0)^{-1} \\ &= (G_0' \Lambda_0^{-1} G_0)^{-1} G_0' \Lambda_0^{-1} G_0 (G_0' \Lambda_0^{-1} G_0)^{-1} \\ &= (G_0' \Lambda_0^{-1} G_0)^{-1} \end{aligned}$$

It can be shown that this indeed is the MOST efficient weighting matrix we can put in there.

Least Mean Distance Estimators - General

- **Assumptions 6.1** for Consistency:

- (a) Compact parameter space.
- (b) q is continuous in β (not necessarily in z).
- (c) z_i are iid.
- (d) Dominance on q .
- (e) β_0 is a unique minimizer of the non-stochastic analogue, $E[q(z_i, \beta)]$.

• **Assumptions 6.1*** for Asymptotic Normality:

- (f) β_0 is an interior point of B .
- (g) q is C^2 .
- (h) Dominance conditions on the gradient and hessian of q .
- (i) The expected value of the hessian is non-singular.

• **Theorem 6.1** If the assumptions above hold, we have:

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}B_0A_0^{-1}),$$

with:

$$A_0 = E[\nabla_{\beta\beta}q(z_i, \beta_0)],$$

and,

$$B_0 = E[\nabla_{\beta'}q(z_i, \beta_0)\nabla_{\beta}q(z_i, \beta_0)].$$

Least Mean Distance Estimators - Nonlinear Least Squares

• **Assumptions 6.2** for Consistency:

- (a) Compact parameter space.
- (b) g is continuous in β (not necessarily in z).
- (c) z_i are iid.
- (d) $E[y_i|x_i] = g(x_i, \beta_0)$.
- (e) Dominance on g .
- (f) β_0 is a unique minimizer of the non-stochastic analogue, $E[(g(z_i, \beta_0) - g(x_i, \beta))^2]$.

• **Assumptions 6.2*** for Asymptotic Normality:

- (g) β_0 is an interior point of B .
- (h) g is C^2 .
- (i) Dominance conditions on the gradient and hessian of g .
- (j) The expected value of the hessian is non-singular.

- **Theorem 6.2** If the assumptions above hold, we have:

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}B_0A_0^{-1}),$$

with:

$$A_0 = E[\nabla_{\beta'}g(x_i, \beta_0)\nabla_{\beta}g(x_i, \beta_0)],$$

and,

$$B_0 = E[(y_i - g(x_i, \beta_0))^2\nabla_{\beta'}g(x_i, \beta_0)\nabla_{\beta}g(x_i, \beta_0)].$$

- Note that if there is no Heteroskedasticity:

$$E[(y_i - g(x_i, \beta_0))^2|x_i] = E[\epsilon_i^2|x_i] = \sigma^2,$$

then:

$$B_0 = E[(y_i - g(x_i, \beta_0))^2 \underbrace{\nabla_{\beta'}g(x_i, \beta_0)\nabla_{\beta}g(x_i, \beta_0)}_{A_0}] = \sigma^2 A_0.$$

So,

$$A_0^{-1}B_0A_0^{-1} = \sigma^2 A_0^{-1}.$$

- For the case of a linear model: $y_i = x_i\beta + \epsilon_i$, (with homoskedasticity):

$$A_0 = E[\nabla_{\beta'}g(x_i, \beta_0)\nabla_{\beta}g(x_i, \beta_0)] = E[x_i x_i'],$$

so,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, \sigma^2(X'X)^{-1}),$$

as usual.

- For the case of a linear model with heteroskedasticity:

$$\hat{B}_n = \frac{1}{n} \sum_{i=1}^n (y_i - g(x_i, \beta_0))^2 x_i x_i' = \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 x_i x_i'.$$

So,

$$\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} = \left(\sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 x_i x_i' \left(\sum_{i=1}^n x_i x_i' \right)^{-1}.$$

Least Mean Distance Estimators - Maximum Likelihood

- **Assumptions 6.3** for Consistency:

- (a) Compact parameter space.
- (b) q is continuous in β (not necessarily in z).
- (c) z_i are iid.
- (d) Dominance on q .

– (e) β_0 is a unique minimizer of the non-stochastic analogue, $E[q(z_i, \beta)]$.

• **Assumptions 6.3*** for Asymptotic Normality:

– (f) β_0 is an interior point of B .

– (g) q is C^2 .

– (h) Dominance conditions on the gradient and hessian of q .

– (i) The expected value of the hessian is non-singular.

• **Theorem 6.3** If the assumptions above hold, we have:

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}B_0A_0^{-1}),$$

with:

$$A_0 = E[\nabla_{\beta\beta}q(z_i, \beta_0)] = -E[\nabla_{\beta\beta}\ln f(y_i|x_i; \beta_0)],$$

and,

$$B_0 = E[\nabla_{\beta\beta}q(z_i, \beta_0)] = E[\nabla_{\beta'}\ln f(y_i|x_i; \beta_0)\nabla_{\beta}\ln f(y_i|x_i; \beta_0)].$$

HOWEVER, since we have a likelihood as our objective function, we can interchange integration and differentiation, which implies $A_0 = B_0$. Thus,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow^d N(0, A_0^{-1}).$$

28.9 Numerical Methods

• Stopping criteria:

– (1) Euclidean distance criteria:

$$(\beta_g - \beta_{g-1})'(\beta_g - \beta_{g-1}) < \epsilon,$$

for some $\epsilon > 0$.

– (2) Objective function criteria:

$$Q(\beta_{g-1}) - Q(\beta_g) < \epsilon.$$

– (3) Score vector criteria:

$$\left[\frac{\partial Q(\beta_g)}{\partial \beta} \right] \left[\frac{\partial Q(\beta_g)}{\partial \beta} \right]' < \epsilon.$$

– (4) Maximum number of iterations is reached.

- Define the directional derivative:

$$\left. \frac{\partial Q(\beta_{g-1} + \lambda \delta_g)}{\partial \lambda} \right|_{\lambda=0} = \frac{\partial Q(\beta_{g-1})}{\partial \beta} \delta_g.$$

- So consider a choice of δ_g :

$$\delta_g = -P_g \frac{\partial Q(\beta_{g-1})}{\partial \beta'}.$$

- Thus, the updating formula for β becomes:

$$\beta_g = \beta_{g-1} - \lambda_g P_g \frac{\partial Q(\beta_{g-1})}{\partial \beta}.$$

- Method of Steepest Descent

$$P_g = I_K$$

- Newton-Raphson Method

$$P_g = \left[\frac{\partial^2 Q(\beta_{g-1})}{\partial \beta \partial \beta'} \right]^{-1}.$$

- Berndt, Hall, Hall, and Hausman (BHHH) Method

$$P_g = \left[\sum \frac{\partial q_i(\beta_{g-1})}{\partial \beta'} \frac{\partial q_i(\beta_{g-1})}{\partial \beta} \right].$$

28.10 Notes from Exams - Part I

– $\frac{1}{N}(\sum_i X_i)^2 = X' e_T (e_T' e_T)^{-1} e_T' X.$

- More on R^2 (first expression always, others with intercept):

$$R^2 = \frac{S_{y\hat{y}}^2}{S_{yy} S_{\hat{y}\hat{y}}} = \frac{S_{\hat{y}\hat{y}}^2}{S_{yy} S_{\hat{y}\hat{y}}} = \frac{S_{\hat{y}\hat{y}}}{S_{yy}} = \frac{S_{yy} - S_{\hat{u}\hat{u}}}{S_{yy}} = 1 - \frac{S_{\hat{u}\hat{u}}}{S_{yy}}.$$

- Test of a total regression relationship (besides an intercept):

$$F = \frac{R^2/G}{(1 - R^2)/(T - K)} \sim F(G, T - K).$$

- General F test of a linear restriction:

$$F = \frac{(ESS_R - ESS_U)/G}{ESS_U/(T - K)}.$$

But $ESS_R = ESS_U + (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$, $ESS_U = \hat{u}'\hat{u}$ and $s^2 = \hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{T - K}$, so:

$$F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/G}{s^2} = (R\hat{\beta} - r)'[RVC(\hat{\beta})R']^{-1}(R\hat{\beta} - r)/G.$$

– Chow Test for a Structural Break. Split the regression equation into different sets of observations. Then:

$$F = \frac{(ESS_R - (ESS_1 + ESS_2))/G}{(ESS_1 + ESS_2)/(T - 2K)} \sim F(G, T - 2K).$$

– $A^{-1} = \frac{1}{|A|}(Adj(A))'$. Be sure to check for block diagonal.

– Prediction. Given out of sample observation x_p where the model still holds:

$$y_p = \alpha + \gamma x_p + u_p.$$

Estimated with:

$$\hat{y}_p = \hat{\alpha} + \hat{\gamma}x_p.$$

So,

$$E[\hat{y}_p] = \alpha + \gamma x_p.$$

So,

$$\begin{aligned} VC(\hat{y}_p) &= E[(\hat{y}_p - E[\hat{y}_p])(\hat{y}_p - E[\hat{y}_p])'] \\ &= E[(X_p\hat{\beta} - X_p\beta)(X_p\hat{\beta} - X_p\beta)'] \\ &= E[X_p(\hat{\beta} - \beta)(\hat{\beta} - \beta)'X_p'] \\ &= X_p\sigma^2(X'X)^{-1}X_p' \end{aligned}$$

Where $\beta = [\alpha \ \gamma]'$ and $X_p = [1 \ x_p]$. Then if the prediction error is:

$$v_p = y_p - \hat{y}_p = X_p\beta + u_p - X_p\hat{\beta} = u_p - X_p[\hat{\beta} - \beta].$$

So $E[v_p] = 0$ and:

$$VC[v_p] = VC(u_p) + VC(X_p(\hat{\beta} - \beta)) = \sigma^2 I_s + X_p\sigma^2(X'X)^{-1}X_p' = \sigma^2(I_s + X_p(X'X)^{-1}X_p').$$

– Restricted OLS. Model: $y = X\beta + u$, restriction: $R\beta = r$. Thus,

$$\text{Min } (y - X\beta)'(y - X\beta),$$

s.t.

$$R\beta = r.$$

Lagrangian:

$$\mathcal{L} = y'y - 2\beta'X'y + \beta'X'X\beta - 2\lambda'[R\beta - r].$$

FOC(β):

$$\begin{aligned} -2X'y + 2(X'X)\beta - 2R'\lambda &= 0. \\ -X'y + X'X\beta - R'\lambda &= 0. \quad (*) \end{aligned}$$

– Premultiply (*) by $R(X'X)^{-1}$:

$$\begin{aligned} -R(X'X)^{-1}X'y + R(X'X)^{-1}X'X\beta - R(X'X)^{-1}R'\lambda &= 0. \\ -R\hat{\beta} + R\beta - R(X'X)^{-1}R'\lambda &= 0. \\ R(\beta - \hat{\beta}) &= R(X'X)^{-1}R'\lambda. \\ \lambda &= [R(X'X)^{-1}R']^{-1}R(\beta - \hat{\beta}). \end{aligned}$$

– Premultiply (*) by $(X'X)^{-1}$ (Note under $H_0 : R\beta = r$):

$$\begin{aligned} -(X'X)^{-1}X'y + (X'X)^{-1}X'X\beta - (X'X)^{-1}R'\lambda &= 0. \\ \beta &= \hat{\beta} + (X'X)^{-1}R'\lambda. \\ \beta &= \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(\beta - \hat{\beta}). \\ \beta_{restricted} &= \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}). \end{aligned}$$

28.11 Notes from Exams - Part II

– Proof that $\hat{\beta}$ is blue. Conditions: $E[u_t] = 0$, $E[u_t^2] = \sigma^2 \forall t$, $E[u_t u_s] = 0$ for $t \neq s$, X full rank and nonstoch. Consider estimator:

$$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'u, \quad E[\hat{\beta}] = \beta, \quad Var\hat{\beta} = \sigma^2(X'X)^{-1}.$$

Denote:

$$\begin{aligned} \tilde{\beta} &= (D = (X'X)^{-1}X')y = Dy + \hat{\beta}. \\ E[\tilde{\beta}] &= \beta \implies DX = 0. \end{aligned}$$

$$Var[\tilde{\beta}] = Var(Dy) + \sigma^2(X'X)^{-1} = DVar(u)D' + \sigma^2(X'X)^{-1} = D\sigma^2D' + \sigma^2(X'X)^{-1} > Var(\hat{\beta})$$

– To show consistence always consider chebychev's theorem with expected value and variance.

– Kinchine. For $\epsilon_t \sim iid(0, \sigma^2)$ with $E[\epsilon_t^2] = \sigma^2$, then:

$$plim \frac{1}{T} \sum \epsilon_t^2 = \sigma^2.$$

- Lindeberg Levy CLT: $u_t \sim iid(0, \sigma^2)$ then:

$$\frac{1}{\sqrt{T}} \sum u_t \rightarrow^d N(0, \sigma^2).$$

- Theorem 4.7. If $u_t \sim iid(0, \sigma^2 < \infty)$ with $\lim n^{-1} X'X = Q$, finite, then:

$$\frac{1}{\sqrt{n}} X'u \rightarrow^d N(0, \sigma^2 Q).$$

- CLT for growing regressors. $u_t \sim iid(0, \sigma^2)$. If:

$$\frac{\text{Max}\{x_t^2\}}{\sum x_t^2} \rightarrow 0,$$

then by the Modified Lindberg-Levy,

$$\frac{\sum x_t u_t}{\sqrt{\sigma^2 \sum x_t^2}} \rightarrow^d N(0, 1),$$

Or,

$$\frac{\sum x_t u_t}{\sqrt{\sum x_t^2}} \rightarrow^d N(0, \sigma^2).$$

- Prediction: future model: $y_f = a + bx_f + u_f = X_f \beta + u_f$. A BLUE estimator is :

$$y_f^p = X_f \hat{\beta}.$$

Prediction error:

$$y_f - y_f^p = X_f \beta + u_f - X_f \hat{\beta} = u_f - X_f (\hat{\beta} - \beta).$$

Variance of prediction error:

$$\sigma^2 - X_f \sigma^2 (X'X)^{-1} X_f'.$$

- ESS = Error sum of squares or residual sum of squares.
- Assumptions for asymptotic distribution of $\hat{\beta}$: $E[u] = 0$, $E[uu'] = \sigma^2 I$, u iid, X full rank and nonstoch, $T^{-1} X'X \rightarrow Q$, finite nonsingular.
- Lemma 1: If $Z \sim N(\mu, \Sigma)$, then $c + bZ \sim N(c + b\mu, b\Sigma b')$. So with $\hat{\beta} - \beta = (X'X)^{-1} X'u$ and $u \sim N(0, \sigma^2)$,

$$\hat{\beta} - \beta \sim N(0, \sigma^2 (X'X)^{-1}), \quad \hat{\beta}_K - \beta_K \sim N(0, \sigma^2 (X'X)^{KK}).$$

- Lemma 2. If $Z \sim N(0, \sigma^2 I)$, A is a $T \times T$ symmetric, idempotent matrix with rank T , then:

$$\frac{Z'AZ}{\sigma^2} \sim \chi^2(r(A)).$$

So:

$$\frac{u'Mu}{\sigma^2} \sim \chi^2(T - K).$$

- Lemma 3. If $Z \sim N(0, \sigma^2 I)$ and A is $T \times T$ symmetric, idempotent and B is $K \times T$ with $BA = 0$, then BZ and $Z'AZ$ are independent. So for :

$$B = (X'X)^{-1}X', \quad A = M, \quad Z = u \sim N(0, \sigma^2 I),$$

then $BA = 0$ and $(X'X)^{-1}X'u = \hat{\beta} - \beta$ are $u'Mu$ are independent!

- Derive the t :

$$\begin{aligned} H &= \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}_K}} = \frac{\hat{\beta} - \beta}{\sqrt{s^2(x'x)^{kk}}} = \\ &= \frac{\hat{\beta} - \beta}{\sqrt{(T - K)^{-1}u'Mu(x'x)^{kk}}} = \frac{[\hat{\beta} - \beta]/[\sqrt{\sigma^2(x'x)^{kk}}]}{\sqrt{(T - K)^{-1}\sigma^2u'Mu}} = \frac{N(0, 1)}{\sqrt{\chi^2(T - K)}} \sim t(T - K). \end{aligned}$$

- Cheychev: For a random variable, Z , with $E[Z] = \mu$ and $E[(Z - \mu)^2] = \sigma^2$, then $\forall \epsilon > 0$,

$$Pr\{|Z - \mu| \geq \epsilon\} \leq \frac{E[(Z - \mu)^2]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}.$$