Methods of Economic Investigation I
Michaelmas Term

Matthew Chesnes
The London School of Economics

December 14, 2001

1 Week 1

- Data Points: (y_i, x_{i2}) for $i = 1...N$.
- First look for a relationship using scatterplots.
- Multivariable or Multifactor Model: $(y_i, x_{i2}, x_{i3}, ..., x_{iK})$ for $i = 1...N$ and $K < N$.
- To fit a line, Gauss looked at the vertical distance between each observation and the corresponding point on the regression line.

$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

$$\hat{\epsilon}_i = y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i).$$

- Or in the multivariate case,

$$\hat{\epsilon}_i = y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + ... + \hat{\beta}_K x_{iK}).$$

- $x_{i1} = 1$ for all i. This acts as the intercept in the regression.
- Minimum Distance Estimation (MDE): Several possibilities such as minimizing the absolute deviations from the predicted and actual values or minimizing the sum of the squares of these deviations. Criteria is that the function of the errors that is used should be an even function so as to capture the negative and positive deviations.
  - Use Least Absolute Deviations (LAD) : Minimize $\sum_i^N |\hat{\epsilon}_i|$.
  - Use Ordinary Least Squared Errors (OLS): Minimize $\sum_i^N \hat{\epsilon}_i^2$.
  - Use Least Quartics (LQ): Minimize $\sum_i^N \hat{\epsilon}_i^4$.
  - Minimize over $\hat{\beta}_1 ... \hat{\beta}_K$.
  - All these methods above will, in general, yield different estimates of the parameters
  - NOTE: If $K = 1$, meaning there is only an intercept in the regression, MDE_{LAD} = median(y).
  - NOTE: If $K = 1$, MDE_{OLS} = mean(y).
  - MDE_{OLS} is much easier to work with because MDE_{LAD} is NOT twice continuously differentiable.
- Solving the Least Squares Problem
  -

$$Min_{\hat{\beta}} \sum_{i=1}^{N} \hat{\epsilon}_i^2 = \underbrace{\sum_{i=1}^{N} [y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - ... - \hat{\beta}_K x_{iK}]^2}_{RSS}.$$

- First Order Conditions.

$$\frac{\partial}{\partial \hat{\beta}_1} = 0 \Rightarrow \sum_i 2(\hat{\epsilon}_i)(-1) = 0.$$

$$\frac{\partial}{\partial \hat{\beta}_2} = 0 \Rightarrow \sum_i 2(\hat{\epsilon}_i)(-x_{i2}) = 0.$$

$$...$$

$$\frac{\partial}{\partial \hat{\beta}_K} = 0 \Rightarrow \sum_i 2(\hat{\epsilon}_i)(-x_{iK}) = 0.$$

- Second Order Conditions: A KxK matrix of second order differentials. Since we are minimizing, this matrix must be postive definite.

- Consider the case of $K = 1$. FOC: $\frac{\partial RSS}{\partial \hat{\beta}_1} = 0 \Rightarrow -2\sum_i(\hat{\epsilon}_i) = 0$. Thus $\sum_i(y_i - \hat{\beta}_1) = 0$. Thus $\sum_i y_i = \sum_i \hat{\beta}_1$. Thus $\sum_i y_i = N\hat{\beta}_1$ or $\hat{\beta}_1 = \frac{\sum_i y_i}{N} = \bar{y}$. (As was stated above).

- SOC: $\frac{\partial^2 RSS}{\partial \hat{\beta}_1^2} = -2\sum_i(-1) = 2n > 0 \Rightarrow$ Postive Definite $\Rightarrow$ Minimum.

- Consider the General Case in Matrix Form:

  - Define the vector of errors,

$$\hat{\epsilon} = \begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \hat{\epsilon}_3 \\ ... \\ \hat{\epsilon}_N \end{bmatrix}. \tag{1}$$

  - Thus,

$$\sum_{i=1}^{N} \hat{\epsilon}_i^2 = \sum_{i=1}^{N} [y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_K x_{iK}]^2 = \hat{\epsilon}'\hat{\epsilon}.$$

  - Also define y and x in matrix notation as,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ y_N \end{bmatrix}. \tag{2}$$

$$x = \begin{bmatrix} 1 & x_{12} & x_{13} & ... & x_{1K} \\ 1 & x_{22} & x_{23} & ... & x_{2K} \\ 1 & x_{32} & x_{33} & ... & x_{3K} \\ ... & ... & ... & ... & ... \\ 1 & x_{N2} & x_{N3} & ... & x_{NK} \end{bmatrix}. \tag{3}$$

– Thus, the OLS problem now reduces to:

$$Min_{\hat{\beta}} \ \hat{\epsilon}'\hat{\epsilon}.$$

$$Min_{\hat{\beta}} \ (y - x\hat{\beta})'(y - x\hat{\beta}).$$

1.1 Extra Notes

– In a simple linear regression model, $y = \beta_1 + \beta_2 x + \epsilon$,

$$\hat{\beta}_2 = \frac{\sum_t (x_t - \bar{x}) y_t}{\sum_t (x_t - \bar{x})^2}.$$

And,

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_t (x_t - \bar{x})^2}.$$

– Residuals: $\hat{\epsilon}_t = y_t - \hat{y}_t$.

2 Week 2

- Second order condition in matrix Notation:

$$\frac{\partial^2 RSS}{\partial\hat{\beta}\partial\hat{\beta}'} = 0.$$

- This gives us the hessian matrix which for a minimum, should be positive definite.

- First order conditions in matrix notation:

$$FOC \equiv \begin{bmatrix} \sum_i 1\hat{\epsilon}_i = 0 \\ \sum_i x_{i2}\hat{\epsilon}_i = 0 \\ \cdots \\ \sum_i x_{ij}\hat{\epsilon}_i = 0 \\ \cdots \\ \sum_i x_{iK}\hat{\epsilon}_i = 0 \end{bmatrix} \equiv x'\hat{\epsilon} = 0_K. \tag{4}$$

- Second order condition in matrix notation: Since FOC: $-2x'\hat{\epsilon} = -2x'(y - x\hat{\beta}) = -2x'y + 2x'x\hat{\beta}$. Thus SOC: $2x'x$. This matrix should be positive definite. Note the inclusion of the "2" when computing the SOC. It was dropped in the FOC above.

- Continuing the OLS problem using the first order conditions in matrix notation.

$$FOC \equiv x'\hat{\epsilon} = 0_K.$$

  Thus,

$$x'(y - x\hat{\beta}) = 0_K.$$
$$x'y - x'x\hat{\beta} = 0_K.$$
$$x'y = x'x\hat{\beta}.$$
$$(x'x)^{-1}(x'y) = \hat{\beta}.$$

- NOTE: This last equality is only assuming $(x'x)^{-1}$ exists! This is equivalent to saying that $(x'x)$ is nonsingular or of full rank. Thus this is true whenever the SOC condition is satisfied. Thus the only necessary condition is now that $N \geq K$.

- We can therefore write,

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

$$\hat{\epsilon} = [I_N - x(x'x)^{-1}x']y.$$
$$\hat{\epsilon} = M_x y.$$

- M_x is called the "projection" matrix. It is symmetric and idempotent. ($(M_x)^p = M_x$ for $p \in Z^+$).

- Given any data on y's and x's, then by premultiplying y by the projection matrix defined above will give the vector of distances from the y's to the regression plane. (As long as $(x'x)$ has full rank, K).

- NOTE: $Rank(M_x) = \rho(M_x) = K$. Since M_x is NxN and $N > K$, M_x is singular. In fact, M_x is positive semidefinite (nonnegative definite).

- Now, define the fitted values,
$$\hat{y} = x\hat{\beta} = y - \hat{\epsilon}.$$
Therefore,
$$\hat{\epsilon}'\hat{y} = \hat{y}'\hat{\epsilon} = (x\hat{\beta})'\hat{\epsilon} = \hat{\beta}'x'\hat{\epsilon} = \hat{\beta}'x'(M_x y) = \hat{\beta}'x'M_x'y = \hat{\beta}'(M_x x)'y.$$
But by construction, $M_x x = 0$ (Basically this is just the residuals of x being regressed on x). Thus,
$$\hat{\epsilon}'\hat{y} = 0.$$
Thus the residuals and the fitted values are orthogonal to one and other. This guarantees that we have found the best fit.

- Now consider,
$$\sum_{i=1}^{N} y_i^2 = y'y = (\hat{y} + \hat{\epsilon})'y = (\hat{y} + \hat{\epsilon})'(\hat{y} + \hat{\epsilon}) = \hat{y}'\hat{y}' + \hat{\epsilon}'\hat{\epsilon} + \underbrace{2\hat{y}'\hat{\epsilon}}_{0}.$$
We have just shown that the residuals and fitted values are orthogonal, so,
$$\sum_{i=1}^{N} y_i^2 = \sum_i \hat{y}_i^2 + \sum_i \hat{\epsilon}_i^2.$$
$$y'y = \hat{y}'\hat{y} + \hat{\epsilon}'\hat{\epsilon}.$$
Which we write as: Total Sum of Squares (TSS) = Explained Sum of Squares (ESS) + Residual Sum of Squares (RSS).

- Additional Result: if the intercept is included in the regression, $\frac{1}{N}\sum_i \hat{\epsilon}_i = 0$. Also, $\bar{\hat{y}} = \bar{y}$, or the sample average of the fitted y's equals the average of the y's.

- Thus $(\bar{x}, \bar{y})$ is always on the OLS regression line iff there is an intercept in the regression equation.

- Another Feature of this approach involves the sample correlation coefficient.
$$r_{y_i,\hat{y}_i} = 1 - \frac{\sum_i \hat{\epsilon}_i^2}{\sum_i (y_i - \bar{y})^2}.$$

- Now define,

$$TSS^* = \sum_i (y_i - \bar{y})^2.$$

$$ESS^* = \sum_i (\hat{y}_i - \bar{\hat{y}})^2.$$

$$RSS^* = \sum_i (\hat{\epsilon}_i - \bar{\epsilon})^2.$$

  Thus the correlation can also be written as,

$$r_{y_i,\hat{y}_i} = \frac{TSS^* - ESS^*}{TSS^*} = 1 - \frac{ESS^*}{TSS^*}.$$

  Thus in the OLS method, we minimized ESS^* which implies that $r_{y_i,\hat{y}_i}$ is **maximized**. Which is a good result because this justifies our method as having the best possible fit.

2.1 Extra Notes

- Key facts involving M_x and $RSS = \hat{\epsilon}'\hat{\epsilon}$.

$$M_x = (I_N - X(X'X)^{-1}X').$$

$$M_x * X = (I_N - X(X'X)^{-1}X')X = 0.$$

  Since the meaning of M_xX is to regress X on X and obtain the residuals. This will be a perfect fit so the residuals are zero.

$$M_x * \epsilon = \hat{\epsilon}.$$

- Key point, key point ...
  - Regression Equation:

$$y = X\hat{\beta} + \hat{\epsilon}.$$

    Rearranging,

$$y - X\hat{\beta} = \hat{\epsilon}.$$

    Substituting in for $\hat{\beta}$,

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

    Pulling out y,

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

    Substituting in M_x,

$$(M_x)y = \hat{\epsilon}.$$

    So regressing y on X gives us the residuals $\hat{\epsilon}$.

- So note:

$$\hat{\epsilon} = M_x y = M_x \epsilon.$$

- If $E[\epsilon\epsilon'|X] = Var(\epsilon|X) = \sigma^2 I_N$, then assume i.i.d.
- Estimate of σ^2, s^2:

$$s^2 = \frac{RSS}{N-K} = \frac{\hat{\epsilon}'\hat{\epsilon}}{N-K} = \frac{(M_x y)'(M_x y)}{N-K} = \frac{y' M_x y}{N-K}$$

$$= \frac{\epsilon' M_x \epsilon}{N-K} = \frac{y'y - y'X\hat{\beta}}{N-K}$$

3 Week 3

- Method of Minimum Distance. Method I.1 is minimizing the squares of the vertical residuals. It leads to unbiased estimators of β only under the assumption that $N \geq K$. ($X'X$ invertible) We could also maximize the correlation between the fitted and actual y's (Method I.1b) but we have just shown that this will yield the same result. There is also the least absolute deviation (LAD) approach. Call it Method I.2. And Method I.3 could be the method of minimizing the quartics.

- In Econometrics, we take data and try to find the true underlying relationships. Aim 1 = Understand economic relationships and test statistically, theories against real world data. Aim 2 = Forecast the future path of economic data. Aim 3 = Policy Analysis / Sensitivity Analysis.

- Assumption 1 was that $\rho(x) = K$ where $N \geq K$ so that $\rho(x'x) = K$ (full rank).

- Now define Assumption 2: There exists a true linear relationship between the y vector and the x matrix. Thus $y = x\beta^* + \mu^*$. Where β^* and μ^* are parameters (the true values of the relationship).

- μ^* is a stochastic component while β^* is non-stochastic. Let $E[\mu^*] = 0$. Since μ^* is stochastic, so is y. μ^* is often referred to as the "stochastic carpet" as one can say, "if we can't explain it, we'll just sweep it under the stochastic carpet and include it in μ^*!?"

- x could be either non-stochastic or stochastic depending on the assumptions in the model. If x is non-stochastic, we say that it is "fixed in repeated samples." Thus $E[y] = x\beta^*$.

- If x is stochastic, the story becomes much more complicated so we will now introduce a new set of assumptions.

- Assumption 3:

  - $A3F = A3_{fixed}$ = "No relationship of a certain type between μ^* and x."
  - $A3Rfi = A3_{random\ fully\ independent\ x's}$ = "(x_{ij}, μ_l^*) are mutually statistically independent for all $i = 1...N$, $j = 1...K$, $l = 1...N$".
  - $A3Rmi = A3_{random\ mean\ independent\ x's}$ = "$E[\mu^*|x] = 0$".
  - $A3Rcu = A3_{random\ contemporaneously\ uncorrelated\ x's}$ = "$Corr(x_{ij}, \mu_i^*) = 0$ or there is no linear relationship." This condition is on individual observations.
  - Note that $A3Rfi \Rightarrow A3Rmi \Rightarrow A3Rcu$.

- NOTE: The Law of Iterated projections / Expectations: $E_x[E[\mu^*|x]] = E[\mu^*]$.

- Conclusions:

- Under assumptions 1 and 2,

$$\hat{\beta}_{OLS} = (x'x)^{-1}(x'y) = (x'x)^{-1}x'[x\beta^* + \mu^*] = \beta^* + (x'x)^{-1}x'\mu^*.$$

- Adding $A3F$.

$$E[\hat{\beta}] = E[\beta^* + (x'x)^{-1}x'\mu^*] = \beta^* + (x'x)^{-1}x'E[\mu^*] = \beta^*.$$

  This is because $(x'x)^{-1}x'$ is fixed and $E[\mu^*] = 0$.

- Adding $A3Rfi$.

$$E[\hat{\beta}] = E[\beta^* + (x'x)^{-1}x'\mu^*] = \beta^* + E[(x'x)^{-1}x']E[\mu^*] = \beta^*.$$

  This is because $(x'x)^{-1}x'$ is independent of μ^* by A2.

- Adding $A3Fmi$.

$$E[\hat{\beta}|x] = E[\beta^* + (x'x)^{-1}x'\mu^*|x] = \beta^* + (x'x)^{-1}x'E[\mu^*|x] = \beta^*.$$

  This is because $E[(x'x)^{-1}x']$ is fixed given an x and $E[\mu^*|x] = 0$ by $A3Rmi$. Thus, since $E[\hat{\beta}|x] = \beta^*$, by the law of iterated expectations, $E[\hat{\beta}] = \beta^*$.

- Adding $A3Fcu$.

$$E[\hat{\beta}] = E[\beta^* + (x'x)^{-1}x'\mu^*] = \beta^* + E[(x'x)^{-1}x'\mu^*] \neq \beta^*.$$

  So under this assumption, the weakest subclass of assumption 3, the estimator is biased.

- A further conclusion under $A1 + A2 + A3Rcu$ (or stronger). Since $Corr(x_{ij}, \mu_i^*) = 0$, $Cov(x_{ij}, \mu_i^*) = 0$ or,

$$E[(x_{ij} - E[x_{ij}])(\mu_i^* - E[\mu_i^*])] = 0.$$

  Thus, since $E[\mu_i^*] = 0$,

$$E[(x_{ij} - E[x_{ij}])(\mu_i^*)] = E[x_{ij}\mu_i^* - E[x_{ij}]\mu_i^*] = 0.$$

$$E[x_{ij}\mu_i^*] - E[E[x_{ij}]\mu_i^*] = E[x_{ij}\mu_i^*] - E[x_{ij}]E[\mu_i^*] = 0.$$

$$E[x_{ij}\mu_i^*] = 0.$$

- This leads us to the method of moments approach (MME). Note that $\mu_i^* = y_i - x_i'\beta^*$.

- Population Moment Conditions:

$$E[x_{i1}\mu_i^*] = 0.$$

$$\ldots$$

$$E[x_{ij}\mu_i^*] = 0.$$

$$\ldots$$

$$E[x_{iK}\mu_i^*] = 0.$$

- Sample Analogues:

$$\frac{1}{N}\sum_i x_{i1}\tilde{\tilde{\mu}}_i = 0.$$
$$\ldots$$
$$\frac{1}{N}\sum_i x_{ij}\tilde{\tilde{\mu}}_i = 0.$$
$$\ldots$$
$$\frac{1}{N}\sum_i x_{iK}\tilde{\tilde{\mu}}_i = 0.$$

  Where $\tilde{\tilde{\mu}}_i = y_i - x_i'\tilde{\tilde{\beta}}_{MME}$.

- But these K equations are exactly the same as the FOCs from the OLS estimation. (The Normal Equations)

$$x'\tilde{\tilde{\mu}} = 0.$$

  Thus,

$$\tilde{\tilde{\beta}} = \hat{\beta}_{OLS}.$$

- So, now consider a fourth assumption, A4, defined as:

$$E[\mu^*\mu^{*'}|x] = Var[\mu^*|x] = \sigma^2_{\mu^*} * I_N.$$

  Assuming (!!) that $E[\mu^*|x] = 0$ which means that at least A3Rmi must be satisfied.

- NOTE: $Var[\mu^*|x] = Var[y|x]$.

- The variance / covariance Matrix is:

$$\sigma^2 = \begin{bmatrix} \sigma^2_{\mu^*_1} & \sigma_{\mu^*_1\mu^*_2} & \sigma_{\mu^*_1\mu^*_3} & \ldots & \sigma_{\mu^*_1\mu_N} \\ \sigma_{\mu^*_2\mu^*_1} & \sigma^2_{\mu^*_2} & \sigma_{\mu^*_2\mu^*_3} & \ldots & \sigma_{\mu^*_2\mu_N} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_{\mu^*_N\mu^*_1} & \sigma_{\mu^*_N\mu^*_2} & \sigma_{\mu^*_N\mu^*_3} & \ldots & \sigma^2_{\mu_N} \end{bmatrix}. \tag{5}$$

- Thus, there are two fundamental properties of the variance / covariance matrix. 1) Symmetric (it has up to $\frac{n(n+1)}{2}$ distinct elements.) 2) Must be positive definite for a random vector with a non-degenerate distribution.

- Thus $Var(\mu^*|x) = \sigma^2_{\mu^*} * I_N =$

$$\begin{bmatrix} \sigma^2_{\mu^*} & & 0 \\ & \ldots & \\ 0 & & \sigma^2_{\mu^*} \end{bmatrix}. \tag{6}$$

- This matrix clearly has only ONE distinct element.

- So, there are two parts to this assumption.
  - $E[\mu_i^{*2}|x] = \sigma_{\mu^*}^2 \; \forall \; i$. In other words, we have homoskedasticity.
  - $E[\mu_i^* \mu_j^*]|x] = 0 \; \forall \; i \neq j$. In other words, we have no autocorrelation.
- One final conclusion of Assumption 4.
- Recall $\hat{\beta}_{OLS} = \hat{\beta}_{MME} = (x'x)^{-1}x'y$ and $y = x\beta^* + \mu^*$.
- Thus,
$$\hat{\beta} = (x'x)^{-1}x'(x\beta^* + \mu^*) = \beta^* + (x'x)^{-1}x'\mu^*.$$
- Now take A1 + A2 + A3Rmi (at least) + A4.
- $\hat{\beta} - \beta^* = (x'x)^{-1}x'\mu^*$. Thus,

$$\begin{aligned}
Var(\hat{\beta}|x) &= E[(\hat{\beta} - E[\hat{\beta}|x])(\hat{\beta} - E[\hat{\beta}|x])'|x] \\
&= E[(\hat{\beta} - \beta^*)(\hat{\beta} - \beta^*)'|x] \\
&= E[((x'x)^{-1}x'\mu^*)((x'x)^{-1}x'\mu^*)'|x] \\
&= E[((x'x)^{-1}x'\mu^*)(\mu^{*'}x(x'x)^{-1})|x] \\
&= (x'x)^{-1}x'E[(\mu^*)(\mu^{*'})|x]x(x'x)^{-1} \\
&= (x'x)^{-1}x'\sigma_{\mu^*}I_N x(x'x)^{-1} \\
&= (x'x)^{-1}x'x\sigma_{\mu^*}I_N(x'x)^{-1} \\
&= \sigma_{\mu^*}I_N(x'x)^{-1} \\
&= \sigma_{\mu^*}(x'x)^{-1}
\end{aligned}$$

4 Week 4

- From last week's lecture, $\hat{\beta}_{OLS} = \hat{\beta}_{MME} \sim (E[\hat{\beta}] = \beta^*, Var(\hat{\beta}|x) = \sigma^2_{\mu^*}(x'x)^{-1})$.
- Without Assumption 4, we would have, $E[\mu^*\mu^{*'}|x] = V(\mu^*|x) =$ an NxN positive definite symmetric matrix, Ω.
- Thus
$$Var(\hat{\beta}|x) = (x'x)^{-1}x'E[\mu^*\mu^{*'}|x]x(x'x)^{-1}.$$
$$Var(\hat{\beta}|x) = (x'x)^{-1}x'\Omega x(x'x)^{-1}.$$
But this is as far as we can go.
- However, if Assumption 4 is valid,
$$Var(\hat{\beta}|x) = \sigma^2_{\mu^*}(x'x)^{-1}.$$
- For reasons of statistical inference, we now aim to find out the statistical distribution of $\hat{\beta}|x$.
- We know already: $E[\hat{\beta}] = \beta^*$ and $Var(\hat{\beta}|x) = \sigma^2_{\mu^*}(x'x)^{-1}$.
- But how is it distributed? Hence Assumption 5:
$$\mu^*|x \sim N(0, \sigma^2_{\mu^*}I_N).$$
And thus,
$$\hat{\beta}|x \sim N(\beta^*, \sigma^2_{\mu^*}(x'x)^{-1}).$$
This is due to the fact that the normal distribution is a "closed family" under linear transforms. A linear combination of normally distributed random variables is itself, normal.

4.1 Conclusions of Assumptions 1 - 5

- Consider
$$\hat{\beta}_j \sim N(\beta^*_j, [\sigma^2_{\mu^*}(x'x)^{-1}]_{jj}).$$
(ie, the variance of $\hat{\beta}_j$ is the jj^{th} element of the matrix.)
- Thus suppose $\sigma^2_{\mu^*}$ is known (unrealistic).
- By subtracting off the mean and dividing by the standard error, we get:
$$\frac{\hat{\beta}_j - \beta^*_j}{\sqrt{[\sigma^2_{\mu^*}(x'x)^{-1}]_{jj}}} \sim N(0,1).$$
- Now assume that $\sigma^2_{\mu^*}$ is unknown. (Likely)

- Idea would be to estimate it with,

$$\frac{1}{N-K}\sum_i \hat{\mu}_i^2 = \frac{RSS}{N-K} = \frac{\hat{\mu}'\hat{\mu}}{N-K}.$$

- To show this is an unbiased estimator, consider RSS,

$$\begin{aligned}
RSS &= \hat{\mu}'\hat{\mu} = (M_x y)'(M_x y) = y' M_x' M_x y. \\
&\quad \text{(By Assumption 1: For } M_x \text{ to exist, } (x'x)^{-1} \text{ must exist).} \\
&= (x\beta^* + \mu^*)' M_x (x\beta^* + \mu^*). \\
&\quad \text{(By Assumption 2: } y = x\beta^* + \mu^*). \\
&= \mu^{*'} M_x \mu^*. \\
&\quad \text{(Because } M_x x = 0 \text{ by construction.) Thus,} \\
E[RSS|x] &= E[\mu^{*'} M_x \mu^* | x]. \\
&= E[Trace(\mu^{*'} M_x \mu^*)|x]. \\
&= E[Trace(M_x \mu^* \mu^{*'}|x]. \\
&\quad \text{(Because Trace}(AB) = \text{Trace}(BA).) \\
&= Trace(E[M_x \mu^* \mu^{*'}|x]). \\
&= Trace(M_x E[\mu^* \mu^{*'}|x]). \\
&= Trace(M_x \sigma^2_{\mu^*} I_N). \\
&= \sigma^2_{\mu^*} I_N * Trace(M_x). \\
&= \sigma^2_{\mu^*} * Trace(I_N - x(x'x)^{-1}x'). \\
&= \sigma^2_{\mu^*} * (N - Trace(x(x'x)^{-1}x')). \\
&= \sigma^2_{\mu^*} * (N - Trace(x'x(x'x)^{-1})). \\
&= \sigma^2_{\mu^*} * (N - Trace(I_K)). \\
&= \sigma^2_{\mu^*} * (N - K).
\end{aligned}$$

  Thus, $s^2 = \frac{RSS}{N-K}$ and

$$E[s^2] = E[\frac{RSS}{N-K}|x] = \sigma^2_{\mu^*}.$$

  Which proves that is unbiased.

- So for inference we can now write,

$$\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{[s^2_{\mu^*}(x'x)^{-1}]_{jj}}}.$$

  However this quantity is NOT distributed $N(0,1)$.

- To determine this distribution, consider,

$$Var(\hat{\beta}|x) = \sigma_{\mu}^2 (x'x)^{-1}.$$

- And thus the estimated variance,

$$\hat{Var}(\hat{\beta}|x) = s^2 (x'x)^{-1}.$$

- Notice that in the expression,

$$\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{[s_{\mu^*}^2 (x'x)^{-1}]_{jj}}},$$

  The numerator is clearly a function of μ^*.

- But the denominator is as well due to the definition of s^2:

$$s^2 = \frac{RSS}{N-K} = \frac{y' M_x y}{N-K} = \frac{\mu^{*'} M_x \mu^*}{N-K}.$$

  So thus, we have a ratio of two functions that both depend on μ^*.

- We have assumed,

$$\mu^* \sim N(0, \sigma_{\mu^*}^2 I_N).$$

- Thus,

$$\frac{\mu^*}{\sigma_{\mu^*}} \sim N(0, I_N).$$

  Given that if you have a normal random variable, Z, then $AZ \sim N(\cdot)$ and $Z'BZ \sim \chi^2(\cdot)$, for a matrix, B. Then AZ and $Z'BZ$ are statistically independent iff $A \cdot B = 0$.

- Thus,

$$\frac{RSS}{\sigma_{\mu^*}^2} = \frac{\mu^{*'}}{\sigma_{\mu^*}} M_x \frac{\mu^*}{\sigma_{\mu^*}} \sim \chi^2(N-K).$$

- We have a further result that says that the ratio of a normal divided by a statistically independent χ^2 is distributed as a t.

- Therefore

$$\frac{\frac{\hat{\beta}_j - \beta_j^*}{\sigma_{\mu^*}}}{\sqrt{\frac{RSS * V}{(N-K)\sigma_{\mu^*}^2}}} = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{s^2 * V}} \sim t(N-K).$$

  Where V = the jj^{th} element of $[(x'x)^{-1}]$. **Note** that assumption 5 (normality) is essential for this result.

4.2 Maximum Likelihood Estimation

- We search for another method to back up our results thus far. If another approach confirms our findings, we can conclude robustness.
- Thus far we have found that,

$$y = x\beta^* + \mu^*.$$

$$y|x \sim N(x\beta^*, \sigma^2_{\mu^*} I_N).$$

  Which accounts for all assumptions through $A5N$.
- Consider a pdf function $g(y|x; \beta^*, \sigma^2_{\mu^*}) = g(y|x; \theta^*)$ where θ^* is the convention used for those factors in the function that are known.
- A sensible estimate of θ is,

$$\hat{\theta} = \underbrace{argmax}_{\theta}[g(y|x; \theta)].$$

- We can reverse this argument because it is actually the y's and x's that we know and we would like to estimate θ.
- Thus,

$$\hat{\theta} = \underbrace{argmax}_{\theta} L^*(\theta; y, x).$$

  Were L^* is the "Likelihood function."
- A sensible way to find $\hat{\theta}$ is the find the value of θ^* that makes the y's and x's to be as likely as possible to have come from the pdf, g.
- Call the estimator, $\tilde{\hat{\theta}}_{MLE}$.
- Key results of Maximum likelihood estimation: 1) If the problem is twice continuously differentiable, solution will be characterised by the first and second order conditions. 2) In general, the first order conditions are non-linear functions of data and θ^*. 3) MLE will result in the best (linear or nonlinear) unbiased estimator for θ^*.
- Next week we will find that the MLE estimator for β^* is the same as the OLS estimator under assumptions 1-5.

5 Week 5

- More on Likelihood functions. We are trying to find the estimator for θ^* by,

$$\hat{\theta}_{MLE} = \underbrace{argmax}_{\theta} L^*(\theta|y, x).$$

- So in the case of our linear regression, we assume $A1 - A5N$. The last being the most important in that since the $y's$ are distributed normally, we know:

$$f(y_i|x, \theta^*) = \frac{1}{\sqrt{2\pi\sigma^2_{\mu^*}}} exp[-\frac{1}{2}\frac{1}{\sigma^2_{\mu^*}}(y_i - x_i'\beta^*)^2].$$

- Since the y_i's are independent, the joint pdf is just the product of the individual pdfs:

$$f(y|x, \theta^*) = \prod_{i=1}^{N} f(y_i|x, \theta^*).$$

  Thus taking the product we find,

$$L^* = f(y|x, \theta^*) = \left(\frac{1}{\sqrt{2\pi\sigma^2_{\mu^*}}}\right)^N * exp[-\frac{1}{2\sigma^2_{\mu^*}}\sum_{i=1}^{N}(y_i - x_i'\beta^*)^2].$$

- The solution to this problem will be characterized by the first and second order conditions. Do to so would be fairly difficult with the above equation but we can use the property that the solution to this problem will be the same as when you take a monotonically increasing function of the above equation. Because the expression involves an exponential, the logical monotonic transformation is the natural log.

- Taking the natural log,

$$ln(L^*) = -\frac{N}{2}ln(2\pi\sigma^2_{\mu^*}) - \frac{1}{2\sigma^2_{\mu^*}}\underbrace{\sum_{i=1}^{N}(y_i - x_i'\beta^*)^2}_{RSS}.$$

- So obtaining $\hat{\beta}$ by MLE yields the same estimator as OLS because maximizing the log likelihood function involves minimizing RSS (because it enters the function negatively).

- Thus we have the very robust result:

$$\hat{\beta}_{OLS} = \hat{\beta}_{MME} = \hat{\beta}_{MLE}.$$

- However, maximizing the log likelihood function with respect to $\sigma^2_{\mu^*}$ yields,

$$\hat{\sigma}^2_{\mu^*} = s^2 = \frac{RSS}{N}.$$

Where as in OLS,

$$s^2_{OLS} = \frac{RSS}{N-K}.$$

So MLE yields a biased estimator of the variance/covariance matrix, but it is obvious that this bias disappears when N is large. Thus $\hat{\sigma}^2_{\mu^*}(MLE)$ is asymptotically unbiased.

5.1 Matrix Partitioning

- So starting with the characterization of $\hat{\beta}$:

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

We have shown via assumption 1,

$$y = x\hat{\beta} + \hat{\mu}.$$

Where,

$$\mu = (I_N - x(x'x)^{-1}x')y = M_x y.$$

- Now suppose that we are running a regression with many explanatory factors but we are only interested in one or several of them. The others are important but we really don't want to waste computer time running a large regression on the whole model just to analyze one coefficient. We can use an interesting and useful property of the definition of $\hat{\beta}$ to help us. Consider partitioning in the x matrix into two blocks. $x = [x_1|x_2]$, where the columns (explanatory factors) in x_1 are the ones we are interested in and everything is thrown into x_2. Thus,

$$y = x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \hat{\mu}.$$

Which is just an algebraic feature.

- We will now show that,

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (x'x)^{-1}x'y = \begin{bmatrix} (x_1'M_{x_2}x_1)^{-1}x_1'M_{x_2}y \\ (x_2'M_{x_1}x_2)^{-1}x_2'M_{x_1}y \end{bmatrix} \tag{7}$$

Where,

$$M_{x_1} = I_N - x_1(x_1'x_1)^{-1}x_1'.$$
$$M_{x_2} = I_N - x_2(x_2'x_2)^{-1}x_2'.$$

- This can also be written as a series of equalities from the Firsch Waugh Theorem:

$$\hat{\beta} = \begin{bmatrix} (x_1'M_{x_2}x_1)^{-1}x_1'M_{x_2}y \\ (x_2'M_{x_1}x_2)^{-1}x_2'M_{x_1}y \end{bmatrix} = \begin{bmatrix} (x_1'M_{x_2}'M_{x_2}x_1)^{-1}x_1'M_{x_2}'M_{x_2}y \\ (x_2'M_{x_1}'M_{x_1}x_2)^{-1}x_2'M_{x_1}'M_{x_1}y \end{bmatrix} = \begin{bmatrix} (x_1'M_{x_2}'M_{x_2}x_1)^{-1}x_1'M_{x_2}'y \\ (x_2'M_{x_1}'M_{x_1}x_2)^{-1}x_2'M_{x_1}'y \end{bmatrix} \tag{8}$$

- Now just focus on $\hat{\beta}_1$ because the formula for $\hat{\beta}_2$ is symmetric.
- Let $M_{x_2}x_1 = x_1^+$ and $M_{x_2}y = y^+$. Thus y^+ is the residual from regressing y on x_2.
- Thus,

$$\hat{\beta}_1 = (x_1'M_{x_2}'M_{x_2}x_1)^{-1}x_1'M_{x_2}'y = (x_1^{+\prime}x_1^+)^{-1}x_1^{+\prime}y^+ = (x_1^{+\prime}x_1^+)^{-1}x_1^{+\prime}y.$$

- Example: Suppose $y = x_1'\beta_1 + i'\beta_2 + \mu^*$, where i is an $nx1$ vector of 1's.
- Now suppose we only want to find the estimate for β_1. Thus,

$$\hat{\beta}_1 = (x_1^{+\prime}x_1^+)^{-1}x_1^{+\prime}y^+.$$

  Where "+" means premultiplied by $M_{x_2} = I_N - i(i'i)^{-1}i'$.
- Thus a typical element of $y^+ = \{y_i - \bar{y}\}$.
- Thus the cross(+) can be calculated without running a regression. Thus for a large matrix were computation is difficult, we will not have to calculate the inverse of a large matrix.
- Proof.

$$\hat{\beta} = (x'x)^{-1}x'y.$$
$$(x'x)\hat{\beta} = x'y.$$
$$(x'x)\hat{\beta} = x'y \iff \begin{bmatrix} x_1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} x_1'y \\ x_2'y \end{bmatrix}. \tag{9}$$

- Thus,

$$x_1'x_1\hat{\beta}_1 + x_1'x_2\hat{\beta}_2 = x_1'y.$$
$$x_2'x_1\hat{\beta}_1 + x_2'x_2\hat{\beta}_2 = x_2'y.$$

- Note from the first equation,

$$\hat{\beta}_2 = (x_1'x_2)^{-1}\big[x_1'y - x_1'x_1\hat{\beta}_1\big].$$

  Premultiply the second equation by $(x_2'x_2)^{-1}$,

$$(x_2'x_2)^{-1}[x_2'x_1\hat{\beta}_1 + x_2'x_2\hat{\beta}_2] = (x_2'x_2)^{-1}[x_2'y].$$

Simplify,

$$(x_2'x_2)^{-1}x_2'x_1\hat{\beta}_1 + \hat{\beta}_2 = (x_2'x_2)^{-1}x_2'y.$$

Substitute in $\hat{\beta}_2$ from above,

$$(x_2'x_2)^{-1}x_2'x_1\hat{\beta}_1 + (x_1'x_2)^{-1}\left[x_1'y - x_1'x_1\hat{\beta}_1\right] = (x_2'x_2)^{-1}x_2'y.$$

Premultiply everything by $(x_1'x_2)$,

$$(x_1'x_2)\left[(x_2'x_2)^{-1}x_2'x_1\hat{\beta}_1 + (x_1'x_2)^{-1}\left[x_1'y - x_1'x_1\hat{\beta}_1\right]\right] = (x_1'x_2)\left[(x_2'x_2)^{-1}x_2'y\right].$$

Simplify,

$$(x_1'x_2)\left[(x_2'x_2)^{-1}x_2'x_1\hat{\beta}_1\right] + x_1'y - x_1'x_1\hat{\beta}_1 = (x_1'x_2)(x_2'x_2)^{-1}x_2'y.$$

Move $x_1'y$ to the right,

$$(x_1'x_2)\left[(x_2'x_2)^{-1}x_2'x_1\hat{\beta}_1\right] - x_1'x_1\hat{\beta}_1 = (x_1'x_2)(x_2'x_2)^{-1}x_2'y - x_1'y.$$

On the left side, pull x_1' out to the left and $x_1\hat{\beta}_1$ to the right. On the right hand side, pull x_1' out to the left and y out to the right.

$$x_1'\Big[x_2(x_2'x_2)^{-1}x_2' - I_N\Big]x_1\hat{\beta}_1 = x_1'\Big[x_2(x_2'x_2)^{-1}x_2' - I_N\Big]y.$$

Multiply through by -1,

$$x_1'\Big[I_N - x_2(x_2'x_2)^{-1}x_2'\Big]x_1\hat{\beta}_1 = x_1'\Big[I_N - x_2(x_2'x_2)^{-1}x_2'\Big]y.$$

$$x_1'[M_{x_2}]x_1\hat{\beta}_1 = x_1'[M_{x_2}]y.$$
$$\hat{\beta}_1 = (x_1'M_{x_2}x_1)^{-1}(x_1'M_{x_2}y).$$

As we were expecting.

5.2 Further Statistical Properties of $\hat{\beta}$

- Recall,
$$\hat{\beta}_{OLS} = \hat{\beta}_{MME} = \hat{\beta}_{MLE} = (x'x)^{-1}x'y.$$
And under $A1 + A2 + A3R_{mi}$ (or stronger),
$$E[\hat{\beta}_{all}] = \beta^*.$$

- Notice that the equation for $\hat{\beta}$ is linear in y. If we add on assumption 4,
$$V(\beta_{OLS}|x) = \sigma^2_{\mu^*}(x'x)^{-1}.$$

- Suppose a skeptic comes along and disagrees with our estimator of β^*. Suppose he suggests another estimator,
$$\tilde{\tilde{\beta}} = A * y$$
Such that $E[\tilde{\tilde{\beta}}] = \beta^*$, where A is a general KxN matrix so again the equation is linear in y. Also, $V(\tilde{\tilde{\beta}}|x) \leq V(\beta_{OLS}|x)$. Note that when comparing matrices, this statements is equivalent to saying $V(\tilde{\tilde{\beta}}|x) - V(\beta_{OLS}|x)$ is negative definite.

- Thus, if $\tilde{\tilde{\beta}}$ exists, it would be unbiased and more efficient than $\hat{\beta}_{OLS}$.

- Must now introduce the Gauss Markov Theorem: Under $A1 - A4$, $\beta_{OLS} = (x'x)^{-1}x'y$ is B.L.U.E.
Proof:
$$E[\tilde{\tilde{\beta}}|x] = E[Ay] = AE[y] = AE[x\beta^* + \mu^*] = Ax\beta^*.$$
Therefore,
$$Ax = I_K \text{ if } \tilde{\tilde{\beta}} \text{ is unbiased.}$$
Recall, by $A4$,
$$V(y|x) = \sigma^2_{\mu^*} I_N.$$
It follows that,
$$V(\tilde{\tilde{\beta}}|x) = V(Ay|x) = AA'V(y|x) = AA'\sigma^2_{\mu^*} I_N = AA'\sigma^2_{\mu^*}.$$
Thus,
$$\begin{aligned} V(\tilde{\tilde{\beta}}|x) - V(\beta_{OLS}|x) &= \sigma^2_{\mu^*}AA' - \sigma^2_{\mu^*}(x'x)^{-1}. \\ &= \sigma^2_{\mu^*}[AA' - (x'x)^{-1}]. \\ &= \sigma^2_{\mu^*}[AA' - I_K(x'x)^{-1}I_K]. \\ &= \sigma^2_{\mu^*}[AA' - Ax(x'x)^{-1}x'A']. \\ &= \sigma^2_{\mu^*}A[I - x(x'x)^{-1}x']A'. \\ &= \sigma^2_{\mu^*}A[M_x]A'. \end{aligned}$$

But M_x is postive semidefinite, and therefore $\sigma^2_{\mu^*} A[M_x]A'$ is as well. Thus,

$$V(\tilde{\tilde{\beta}}|x) \geq V(\beta_{OLS}|x).$$

- So under $A1 - A4$, $\hat{\beta}_{OLS}$ is BLUE (Best Linear Unbiased Estimator) of β^*.
- Another result which we will not prove: Under $A1 - A5N$,

  $$\hat{\beta}_{OLS} = \hat{\beta}_{MLE}$$

  is BUE (Best Unbiased Estimator), linear or nonlinear, of β^*.

6 Week 6

6.1 Main type of Data Generation

- Cross Sectional data: Data the comes from all the same point in time.
- Time series data: Data on the same economic unit at different points in time.
- Panel Data or Longitudinal Data: Observations of different economic units over a length of time which may be different for each unit that you are looking at. In the form $\{y_{ht}\}$ indexed for the economic unit, h, and the time index, t.

6.2 Alternative Notation

- Individual Observations: $i = 1...N$ or $t = 1...T$.
- Explanatory Variables: Usually X, but sometimes Z or W.
- True Parameter Vector: β^*, β, or β^0.
- Estimated Parameter Vector: $\hat{\beta}$ or b.
- True Residual: μ^* or ϵ.
- Estimated / fitted residuals: $\hat{\mu}$, $\hat{\epsilon}$, or e.

6.3 The Linearity of the Regression Equation

- $y = x\beta^* + \mu^*$ is a linear function of the x's.
- If the regression equation is not in this simple linear form, it is either intrinsically or non-intrinsically nonlinear.
- Example of a non-intrinsically nonlinear equation:

$$y_i = x_{i1}\beta_1^* + \frac{1}{x_{i2}}\beta_2^* + ln(x_{i3})\beta_3^* + \mu_i^*.$$

  Simply let $x'_{i2} = \frac{1}{x_{i2}}$ and let $x'_{i3} = ln(x_{i3})$. This is a suitable redefinition of the variables so the equation is now linear.

- Another example:

$$y_i = Ax_{i1}^{\beta_1^*} x_{i2}^{\beta_2^*} \mu_i^*.$$

  Just redefine as:

$$ln(y_i) = ln(A) + \beta_1^* ln(x_{i1}) + \beta_2^* ln(x_{i2}) + ln(\mu_1^*).$$

  This is simply a suitable application of a monotonic transformation.

- Regression equations that cannot be transformed into a linear form are intrinsically nonlinear.

6.4 Hypothesis Testing in Linear Regression

- General key steps in all "classical" hypothesis testing (NOT Baysian).
- Step 1: Develop hypothesis.
  - a.) Identify the null, H_0, the alternative, H_1, and the maintained set of hypotheses. The maintained set of hypotheses can be thought of as "everything else" outside the null and alternative. Basically it could be things like model assumptions and specifications which would fall under different tests but lie in the background of what you're testing.
  - b.) Verify that H_0 and H_1 are "Nested" hypothesis or that that they are not logically parallel. For instance a parallel test would be $H_0 : \beta_1 \neq 0$ and $\beta_i = 0$ for all $i \neq 1$, versus $H_1 : \beta_i \neq 0$ for all $i \neq 1$. I'm not extremely clear on this but later examples will probably help.
  - c.) Eliminate any redundant and contradictory restrictions, or make sure hypotheses are "Clean". For example $H_0 : \beta_1 = 3, \beta_1 * 3 = 9$, and $\beta_1 = 0$ has both redundant and contradictory hypotheses.
- Step 2: Choose a test statistic.
  - The statistic should only be a function of the data and it should do what you want it to do in that it should get as close as possible to testing H_0.
  - The test statistic should also have certain optimality properties. All test statistics will have two types of error, Type I error = P(Rejecting H_0 when it is true) = α and Type II error = P(Failing to reject H_0 when it is false) = β.
  - Thus for a given significance level, α, the object is the minimize the probability of a type II error, β.
  - Distribution under H_0 must be known.
  - Distribution of test statistic under H_1, though frequently not fully known, must be known to be far from the distribution of the statistic under H_0. This provides us with the ability to discriminate between the null and alternative hypothesis.
- Step 3: Construct a "Critical Region" of the range of possible values of the test statistic defining evidence against H_0.
  - Determine if the test should be one or two tailed.
  - Optimality properties (to be discussed).

- Since we have assumed that the distribution of the test statistic under H_0 is known, let t^* = the test statistic and define:

$$P(t^* \in C_\alpha | H_0 \text{ is true}) = \alpha,$$

Where C_α is the critical region associated with α. Thus,

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ is true}) = \text{ Type I Error.}$$

- (Ideally) figure out the "power function" of this test. Power $= 1-\beta = P(\text{Reject } H_0|H_0 \text{ is false})$. However, H_0 being false can happen in many ways. All it says is that some H_1 is true, though we don't know which one. Thus power functions are difficult to compute usually because the problem must be very well defined to calculate power at different possible truths.

- Step 4: Define hypthesis testing decision rule.
  - "Reject H_0 iff $t^* \in C_\alpha$, do not reject otherwise."
  - Note that in classical hypothesis testing, you can never accept a hypothesis. You can only reject.
- Step 5: Based on data, calculate the test statistic and make conclusions based on the decision rule.

6.5 Hypothesis Testing in the Linear Regression Model

- Set of hypotheses may be linear or non-linear.
- Maintained Hypothesis: $A1$, $A2$, $A3R_{mi}$ or stronger, $A4$, and $A5N$.
- Linear Case
  - Example - $H_0 : \beta_1 = 3; \beta_2 + 2\beta_3 = -1$. These hypotheses are linear.
  - General hypotheses: $R\beta = q$ where R is an rXK matrix with the restrictions being tested, β is the $KX1$ vector of parameter estimates, and q is an $rX1$ vector of coefficients.
  - Even more General: $R\theta = q$ where R is now an rXp matrix with the restrictions begin tested, θ is now a $pX1$ vector of parameter estimates being tested and q again is $rX1$.
  - Thus in our example, $H_0 : \beta_1 = 3; \beta_2 + 2\beta_3 = -1$, suppose there are only 3 parameters in the model, β_1, β_2, and β_3. Then,

$$H_0 : \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}. \tag{10}$$

- Note that R and q are known under H_0. Also $r < p$ which means that you cannot have more restrictions than parameters.

- Non-linear Case
  - $g(\beta) = 0$ or $g(\theta) = 0$.
  - Consider an example - $H_0 : \beta_1\beta_2 = -\beta_3$.
  - So we just have to write $H_0 : \beta_1\beta_2 + \beta_3 = 0$.

6.6 The Trinity of Classical Test Procedures

- The "Wald Test."
- Consider an estimator, $\hat{\theta}$ which ignores whether or not H_0 is valid. Only formed under the maintained hypothesis. $\hat{\theta}$ is called the unrestricted estimator.
- Now apply restrictions, $R\hat{\theta} - q$ and test to see if it is statistically significantly far from 0.
- Under our assumptions in the maintained hypothesis,

$$\hat{\theta} \sim N(\theta^*, V(\hat{\theta})).$$

  Or in the case that $\hat{\theta} = \hat{\beta}_{OLS}$,

$$\hat{\beta}_{OLS} \sim N(\beta^*, \sigma^2_{\mu^*}(x'x)^{-1}).$$

  Thus,

$$R\hat{\theta} - q \sim N(0, RV(\hat{\theta})R').$$

  Because a linear transformation of a normal random variable is itself normal.

- Thus define the Wald statistic as:

$$W = (R\hat{\theta} - q)'\Big[RV(\hat{\theta})R'\Big]^{-1}(R\hat{\theta} - q).$$

  Thus W is a $1X1$ scaler as it is a quadratic form.

- Thus, under H_0,

$$W \sim \chi^2(r).$$

7 Week 7

7.1 The Trinity of Classical Test Procedures : Wald

- Step 1: Estimate unknown parameter vector θ^* with $\hat{\theta}$, ignoring the stipulations of H_0.
- Step 2: Use $\hat{\theta}$ to test statistically whether or not the H_0 stipulations are violated. The general linear hypothesis is of the form: $R\theta^* = q$. (R is rXp, θ^* is $pX1$, and q is $rX1$). So do the calcuation of $R\hat{\theta}$ and see how far the result is from q.
- If we can show,
$$\hat{\theta} \sim N(\theta^*, V(\hat{\theta})),$$
then we know,
$$R\hat{\theta} \sim N(R\theta^*, RV(\hat{\theta})R'),$$
provided that the rank of R is r with $r < p$. (Otherwise R would be singular.)
- Define our test statistic as,
$$W = (R\hat{\theta} - R\theta^*)'(RV(\hat{\theta})R')^{-1}(R\hat{\theta} - R\theta^*).$$
- If H_0 is correct, then $W \sim \chi^2(r)$. If not, W will follow some other distribution like a noncentral χ^2.
- Of course, θ^* is unobservable, so in this final step, substitute q for $R\theta^*$.
- Wald in the Linear Regression Model. Again, under H_0,
$$W = (R\hat{\theta} - q)'(RV(\hat{\theta})R')^{-1}(R\hat{\theta} - q) \sim \chi^2(r).$$
Let $\hat{\theta} = \hat{\beta} = (x'x)^{-1}x'y$. Thus,
$$W = (R\hat{\beta} - q)'(RV(\hat{\beta})R')^{-1}(R\hat{\beta} - q).$$
$$W = (R\hat{\beta} - q)'(R\sigma^2_{\mu^*}(x'x)^{-1}R')^{-1}(R\hat{\beta} - q).$$
However, we usually do not know σ^2 so estimate with s^2. Note that $RSS = (N-K)s^2$ and,
$$\frac{(N-K)s^2}{\sigma^2_{\mu^*}} \sim \chi^2(N-K).$$
Since this last expression is fully independent of W and both are distributed χ^2, we can divide both by their degrees of freedom and the ratio of the results will be distributed under an F. Consider the messy algebra:
$$\frac{\Big[(R\hat{\beta} - q)'(RV(\hat{\beta})R')^{-1}(R\hat{\beta} - q)\Big]/r}{\Big[\frac{(N-K)s^2}{\sigma^2_{\mu^*}}\Big]/(N-K)}.$$

$$\frac{\left[(R\hat{\beta}-q)'(RV(\hat{\beta})R')^{-1}(R\hat{\beta}-q)\right]/r}{s^2/\sigma^2_{\mu^*}}.$$

$$\frac{\left[(R\hat{\beta}-q)'(R\sigma^2_{\mu^*}(X'X)^{-1}R')^{-1}(R\hat{\beta}-q)\right]/r}{s^2/\sigma^2_{\mu^*}}.$$

$$\frac{\left[(R\hat{\beta}-q)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta}-q)\right]/r}{s^2}.$$

$$\frac{(R\hat{\beta}-q)'(Rs^2(x'x)^{-1}R')^{-1}(R\hat{\beta}-q)}{r} \sim F(r, N-K).$$

7.2 The Trinity of Classical Test Procedures : Likelihood Ratio

- Step 1: Estimate the model disregarding the stipulations of H_0. Note some goodness of fit measure of this estimation such as the maximum of the log-likelihood function of $\hat{\theta}$. (Other possibilities would be the minimum RSS or maximum R^2).

- Step 2: Estimate the model again, but this time incorporating H_0. Note the corresponding goodness of fit measure for this estimation.

- Ignoring the R^2 measure for now, call goodness of fit measures from the first estimation, LLF_1 and RSS_1. And from the regression including the H_0 restrictions, LLF_0 and RSS_0.

- Step 3: Evaluate whether or not H_0 is violated by comparing the alternate goodness of fit measures. If $LLF_1 \gg LLF_0$, this is evidence against H_0. If $RSS_1 \ll RSS_0$, this is also evidence against H_0.

- Two, unproved, key results,

  $$2[LLF_1 - LLF_0] \sim \chi^2(r),$$

  for large N where r is the number of restrictions (under H_0).

  $$\frac{(RSS_0 - RSS_1)/r}{RSS_1/(N-K)} \sim F(r, N-K).$$

  In fact, this second result can be shown to be equivalent to the Wald test.

7.3 The Trinity of Classical Test Procedures : Lagrange Multiplier

- Only requires estimating the restricted model (model assuming H_0 is correct).
- Step 1: Define $\bar{\theta}_R$ (or $\bar{\theta}_0$) as follows:

$$\bar{\theta}_R = \underbrace{argmax}_{\theta} \text{ LLF s.t. } R\theta = q.$$

- Step 2: Statistically investigate whether or not H_0 restrictions are violated by the data by examining the lagrangian vector, λ (the shadow prices of the contraints).

$$LLF + \lambda(R\theta - q).$$

  Values of λ far from 0 constitute evidence against H_0.

7.4 Summary of the Trinity

- We have now defined three test procedures but deciding which will be optimal to use will depend on the type of problem under consideration. All three have the property that they are the most powerful test available. This means that, for a given significance level, the result from each test will be the best possible. (For a given α, the power of the test, $1 - \beta$, will be a high as possible. Or the probability of a type II error, β, will be as low as possible.)
- Note that each of the trinity of tests will not in general yield the same results!

8 Week 8

8.1 More on the Lagrange Multiplier (LM) Test Procedure

- We examine the statistical properties of the Lagrange Multiplier, indicating how strongly the H_0 constraints bind. The higher are the λ's (far from 0), the more evidence we have against H_0. In other words, a high shadow price of the constraints implies that the "data wants" these constraint to be relaxed.

- Practical Implementation of the LM procedure. For the linear regression model, an equivalent way for implementing the LM test is:

  - Step 1: Estimate the R or 0 model (Restricted) and save the residuals and call them $\hat{u}_R$.
  - Step 2: Run the auxiliary regression with $\hat{u}_R$ as the dependent variable (or a suitable tranformation of it – we'll see this later) and a suitable set of regressors reflecting H_0 and H_1.
  - The Result: $N * R^2_{aux} \sim \chi^2(r)$ under H_0 for large N. Since the χ^2 distribution is a positive distribution, then values far from 0 will be in the rejection region. This means that $N * R^2_{aux}$ is high which means that the residuals contain information what would be better explained with something other than H_0.

8.2 Example with the Trinity of Solutions

- Example: Consider the linear regression model: $y = X\beta^* + u^*$. Also assume $A1 - A5N$ hold and that there is an intercept included in the regression.

- Consider the null hypothesis:

$$H_0 : \beta_2^* = \beta_3^* = \cdots = \beta_K^* = 0.$$

- Using the LM procedure:

  - Thus the restricted (R) model is:

$$y = i\beta_1^* + u^*.$$

    Basically just y regressed on a vectors of ones, i.

  - Therefore,

$$\hat{\beta}_1^*(OLS) = \bar{y}.$$

    And a typical element of the residuals from this regression,

$$\hat{u}_R = \{y_i - \bar{y}\}.$$

– Now consider the auxillary regression:

$$\hat{u}_R = \{i \; x_{i2} \; x_{i3} \; \ldots \; x_{iK}\}.$$

We regress the saved residuals, $\hat{u}_R$, on H_0 and H_1 combined (In this case, the full model).

– If $N * R^2_{aux}$ is large, it means that at least one of the regressors was significant: Regression had explanatory power. If H_0 was to hold: auxillary regression should have no explanatory power.

- Now consider the same setup but we'll solve using the Wald Test procedure.

– In the unrestricted model,

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

– Under H_0 however, $\hat{\beta}_1$ could be anything and the other β's are all close to 0. Let,

$$\hat{\beta}^*_2$$

represent the set of restrictions that all the β's except the first are equal to 0. Thus, in the Wald setting,

$$\hat{\beta}^*_2 = R\hat{\beta} - q.$$

– Thus the Wald statistic is defined as,

$$W = \hat{\beta}^{*'}_2 \left[Var(\hat{\beta}^*_2) \right]^{-1} \hat{\beta}^*_2.$$

But by the work we have covered on partitioning,

$$Var(\hat{\beta}^*_2) = \sigma^2_{\mu^*} \left(X_2' M_1 X_2 \right)^{-1}.$$

Thus,

$$W = \hat{\beta}^{*'}_2 \left[\sigma^2_{\mu^*} \left(X_2' M_1 X_2 \right)^{-1} \right]^{-1} \hat{\beta}^*_2.$$

Now we have shown that we can divide W by $K-1$ and divide the whole expression by $\frac{s^2}{\sigma^2_{\mu^*}}$ and get an F distribution. Thus,

$$\frac{\hat{\beta}^{*'}_2 \left[\sigma^2_{\mu^*} \left(X_2' M_1 X_2 \right)^{-1} \right]^{-1} \hat{\beta}^*_2 / K - 1}{s^2 / \sigma^2_{\mu^*}} \sim F(K-1, N-K).$$

Or, simplifying,

$$\frac{\hat{\beta}_2^{*\prime}(X_2'M_1X_2)\hat{\beta}_2^*/K-1}{s^2} \sim F(K-1, N-K).$$

But we also know from partitioning,

$$\hat{\beta}_2^* = (X_2'M_1X_2)^{-1}X_2'M_1y.$$

Thus,

$$\frac{\big((X_2'M_1X_2)^{-1}X_2'M_1y\big)'(X_2'M_1X_2)^{-1}\big((X_2'M_1X_2)^{-1}X_2'M_1y\big)/K-1}{s^2} \sim F(K-1, N-K).$$

But I'm not sure why this step is necessary because now I want to look back to the previous equation and rewrite the equation using the idempotency of M_1. So,

$$\frac{\hat{\beta}_2^{*\prime}(X_2'M_1X_2)\hat{\beta}_2^*/K-1}{s^2} = \frac{\hat{\beta}_2^{*\prime}X_2'M_1'M_1X_2\hat{\beta}_2^*/K-1}{s^2}.$$

But $M_1X_2\hat{\beta}_2^*$ is just the fitted y's from the full regression, $\hat{y}$. And $\hat{\beta}_2^{*\prime}X_2'M_1' = \hat{y}'$. Thus,

$$\frac{\hat{\beta}_2^{*\prime}X_2'M_1'M_1X_2\hat{\beta}_2^*/K-1}{s^2} = \frac{\hat{y}'\hat{y}/K-1}{s^2}.$$

But

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

Thus,

$$\frac{ESS/K-1}{s^2} = \frac{(TSS-RSS)/K-1}{RSS/N-K} = \frac{(\frac{TSS-RSS}{TSS})/K-1}{(\frac{RSS}{TSS})/N-K}.$$

But since $1-\frac{RSS}{TSS} = R^2$,

$$\frac{(\frac{TSS-RSS}{TSS})/K-1}{(\frac{RSS}{TSS})/N-K} = \frac{R^2/K-1}{(1-R^2)/N-K} \sim F(K-1, N-K).$$

- Again, high values of the test statistic will lead to rejecting H_0. This occurs if R^2 is high which means that more than just the intercept in this regression is important.

- Now consider the same setup but we'll solve using the Likelihood Ratio approach.
  - Version 1 of this test is to compare RSS_U with RSS_R or LLF_U with LLF_R.
  - In Version 2, however,

$$\frac{(RSS_R-RSS_U)/K-1}{RSS_U/N-K}.$$

– But we defined the residuals under the restricted model as $\hat{u}_R$, thus,

$$RSS_R = \sum_{i=1}^{N} \hat{u}_{iR}^2.$$

– But because of the restricted regression only having the intercept as a regressor,

$$RSS_R = \sum_{i=1}^{N} \hat{u}_{iR}^2 = \{(y_i - \bar{y})^2\} = TSS.$$

– Thus, version 2 becomes,

$$\frac{(TSS - RSS_U)/K - 1}{RSS_U/N - K} = \frac{(TSS - RSS_U)/K - 1}{s^2}.$$

And this result is identical to the Wald test procedure. Note:

$$\frac{(TSS - RSS_U)/K - 1}{s^2} = \frac{R^2/K - 1}{(1 - R^2)/N - K} \sim F(K - 1, N - K).$$

8.3 One Final Issue in Hypothesis Testing

- When considering the restricted models (R or 0 models), for instance in the second of the trinity of tests, we are either maximizing LLF s.t. H_0 constraints or minimizing RSS s.t. H_0 constraints.
- This, however, could prove to be fairly difficult or even impossible.
- There are two cases we will now consider, those restricted models that are "especially intractable" and those that are "always tractable."
- Consider the following case of an especially intractable restricted model.
  - The null hypothesis as follows:

$$H_0 : g(\theta^*) = 0,$$

where θ^* is a $px1$ vector of constraints, 0 is the $rx1$ zero vector, and the inclusion of g refers to these restrictions being nonlinear.

  - If it is possible to solve the constraint and plug into the objective, the optimisation becomes unconstrainted and would be solvable. For instance,

$$H_0 : \beta_2\beta_3 = \beta_4.$$

  - So an otherwise linear model with K explanatory variables could be estimated as follows:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_2\beta_3 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_K x_{iK} + u_i^*.$$

– And then in both the likelihood ratio approach (number 2) or the Lagrange approach (number 3), one would just minimize RSS of this regression to solve and the optimization which, of course, is unconstrained. However! The unknowns in this regression, (ie, those parameters that we will be optimizing over) are:

$$\beta_1,\ \beta_2,\ \beta_3,\ \beta_5,\ \ldots,\ \beta_K.$$

– This amounts to $K-1$ free parameters so most typical statistical software can solve this type of problem using a technique called non-linear least squares. However, this is much more difficult. (for the computer?)

- Consider the following case of a model that is always tractable.

  – The null hypothesis as follows:

  $$H_0 : R\beta^* = q.$$

  Where R is a rxp matrix of linear restrictions and q is $rx1$. Also $\rho(R) = r$ and $r < p$.

  – So, again we would like to substitute the restictions into our objective function and do the unconstrained optimization. The result is that this procedure is always feasible and corresponds to a particular OLS.

  – Proof of this result. Since the R matrix has full row rank, we can partition R as follows:

  $$R\beta = \left[\underbrace{R_1}_{rxr} \quad \underbrace{R_2}_{rx(K-r)} \right] \left[\begin{array}{c} \underbrace{\beta_1^*}_{rx1} \\ \underbrace{\beta_2^*}_{(K-r)x1} \end{array} \right] = q. \tag{11}$$

  Now it must be that $\rho(R_1) = r \equiv$ Full Rank. Recall the general form our model (in vector notation) and the resulting partition,

  $$y = X\beta^* + u^* = X_1\beta_1^* + X_2\beta_2^* + u^*.$$

  Now consider the restriction and the resulting partition,

  $$R\beta^* = R_1\beta_1^* + R_2\beta_2^* = q.$$

  Solving the restriction for β_1*,

  $$\beta_1^* = R_1^{-1}[q - R_2\beta_2^*].$$

  We can do this rearranging of the restrictions because we have shown that R_1 has full rank and therefore is invertible. Substituting β_1^* back into our model,

  $$y = X_1\left(R_1^{-1}[q - R_2\beta_2^*] \right) + X_2\beta_2^* + u^*.$$

And thus we have a regression equation that is only a function of β_2^*'s and u^*'s while everything else is in the data. So moving all the data terms over to left,

$$\underbrace{y - X_1 R_1^{-1} q}_{\tilde{y}} = \underbrace{(X_2 - X_1 R_1^{-1} R_2)}_{\tilde{X}_2} \beta_2^* + u^*.$$

Or rewriting after substitution,

$$\tilde{y} = \tilde{X}_2 \beta_2^* + u^*.$$

This is an unconstrained linear regression. Since we did this in the general case, any null hypothesis of the form,

$$H_0 : R\beta^* = q,$$

is ALWAYS tractable. QED.

8.4 A Comment on Prediction in the Linear Regression Problem

- Consider the model under the maintained set of hypotheses,

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

Where assumptions $A1 - A5N$ hold and X is a NxK matrix.

- Now consider what we'll call the future equation:

$$y_f = x_f' \beta^* + u_f^*.$$

Note that the same β^* governs both equations and we expect the error terms to behave the same such that,

$$\begin{pmatrix} u^* \\ u_f^* \end{pmatrix} | X, x_f' \sim N \left(\begin{pmatrix} \tilde{0} \\ 0 \end{pmatrix}, \ \sigma_{\mu^*}^2 \begin{pmatrix} I_n & \tilde{0} \\ \tilde{0}' & 1 \end{pmatrix} \right) \tag{12}$$

So, basically, what is true about the first N observations is also true about the $N+1^{st}$ observation.

- If we estimate the model based on the full data, we would get the usual results:

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

$$V(\hat{\beta}) = \sigma_{\mu^*}^2 (x'x)^{-1}.$$

$$\hat{V}(\hat{\beta}) = s^2 (x'x)^{-1}.$$

- For the future period the only data we have is the data point x_f. We would like to consider the prediction of y_f or features of it.

- The best we can do is $x_f'\hat{\beta}$ because we could either be estimating y_f or $E[y_f|X, x_f]$. Since,

$$y_f = x_f'\beta^* + u_f^*.$$
$$E[y_f|X, x_f] = x_f'\beta^*.$$

Since our best estimate of u_f^* is zero because that is its expected value, and the best estimate for β^* is $\hat{\beta}$, $x_f'\hat{\beta}$ is our best estimate for y_f. HOWEVER, there is one difference in the variance of these two estimates. With $\sigma^2_{\mu^*}$ estimated by s^2,

$$\hat{V}(x_f\hat{\beta}) = x_f's^2(x'x)^{-1}x_f : \text{If predicting } E[y_f|X, x_f].$$

$$\hat{V}(x_f\hat{\beta}) = x_f's^2(x'x)^{-1}x_f + s^2 : \text{If predicting } y_f.$$

This is because there is the extra variability when estimating y_f because we are unsure if u_f^* is close to u^*. If the future equation is much more variable, then the prediction interval must be wider.

8.5 Introduction to Dummy Variables

- Dummy variables are often referred to as qualitative, catagorical, or discrete variables.
- Define the "Indicator Function:"

$$1(\text{event}) = \begin{cases} 1 & \text{If Event is True} \\ 0 & \text{Otherwise} \end{cases} \tag{13}$$

- For instance we could have a dummy for war periods or gender. Also seasonal dummies for monthly data (12 different) or quarterly data (4 different).
- One complication which will be examined later probably in the context of probit models, is if the dummy should be treated as a explanatory or dependent variable. If it is explanatory, nothing changes. But if we are trying to predict the probability that a person marries given a set of predictors, things get messy. The estimation becomes nonlinear.

8.6 More on Dummy Variables

- Consider the seasonable dummies where $S_{i1} \equiv$ Spring, $S_{i2} \equiv$ Summer, $S_{i3} \equiv$ Fall, and $S_{i4} \equiv$ Winter.
- Suppose the dependent variable is heating costs for a home or something that would be dependent on those seasonable dummies.
- Consider the model in matrix notation,

$$y = X_1\beta_1 + X_2\beta_2 + u^*.$$

Where X_1 are those explanatory variables such as income, house temperature, house area, etc. and X_2 is an $Nx4$ matrix as follows:

$$X_2 = \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & S_{1,4} \\ S_{2,1} & S_{2,2} & S_{2,3} & S_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ S_{N,1} & S_{N,2} & S_{N,3} & S_{N,4} \end{bmatrix} \tag{14}$$

- Note that $E[y|X_1, X_2] = X_1\beta_1 + X_2\beta_2$. However, since X_2 is a matrix of dummy variables and dummy variables are discrete, it makes more sense to quote expectations at specified values of the dummies. For instance,

$$E[y|X_1 \text{ and Winter}] = X_1\beta_1 + \beta_2^4.$$

  Where the superscript 4 is refers to the fourth element of β_2.

- Basically, since the dummy terms are additive, the inclusion of them allows for different intercepts for each of the 4 seasons.

- Now consider a slightly different model where not only do we have dummy terms that affect the intercepts of the regression lines, but also "interaction terms" that change the slope of the relevant variables. Consider the following model:

$$\begin{aligned} y_i = &\text{Area}_i\beta_{1,1} + \text{Income}_i\beta_{1,2} + S_{i,1}\beta_{2,1} + S_{i,2}\beta_{2,2} + S_{i,3}\beta_{2,3} + S_{i,4}\beta_{2,4} \\ &+\text{Area}_i * S_{i1}\beta_3 + \text{Area}_i * S_{i2}\beta_4 + \text{Area}_i * S_{i3}\beta_5 + \text{Area}_i * S_{i4}\beta_6. \end{aligned}$$

  Where the last 4 terms are interaction terms.

  This of course can be written in matrix notation as,

$$y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5 + X_6\beta_6 + u^*.$$

  Where X_1 include normal individual variables, X_2 include intercept dummies, and X_3 thru X_6 are interaction dummies.

- In this case,

$$\begin{aligned} E[y_i|X_1; \text{spring}] &= \text{Area}_i\beta_{1,1} + \text{Income}_i\beta_{1,2} + \beta_{2,1} + \text{Area}_i\beta_3 \\ &= \beta_{2,1} + \text{Area}_i(\beta_{1,1} + \beta_3) + \text{Income}_i\beta_{1,2}. \end{aligned}$$

8.7 The Dummy Trap

- Recall $A1 : \rho(X'X) = K \equiv$ Full Rank. Thus there are no collinearities between columns (explanatory variables) of the X matrix.

- Consider the case where we have both an intercept in the model and all 4 seasonal dummies. Consider the i^{th} row of the X matrix:

$$x_i = (1 \dots S_{i1}\ S_{i2}\ S_{i3}\ S_{i4}).$$

  Where the $\dots$ represents all other explanatory variables in the model. Now consdier,

$$\sum_{j=1}^{4} S_{ij} = 1 = x_{i1}.$$

  Thus we have shown that in this model specification we can take a linear combination of the dummy variables, namely, $S_{i1} + S_{i2} + S_{i3} + S_{i4}$, and get the first variable in the X matrix, the intercept. Thus X is singular and therefore not of full rank. So the rule is to make sure you don't put in too many dummy variables so you cause collinearities among the explanatory variables.

- There are two ways to get around this problem:
  - Drop 1 of the 4 dummies. The intercepts for each of the sub models corresponding to the dummies in the model are the same as those described above (adding the regular intercept to the dummy intercept), and the intercept for the dummy that is left out is the value of the regular intercept itself.
  - Drop the intercept and include the 4 dummies. The intercepts for the 4 submodels now become the values of each dummy coefficient.

- Both of these solutions are equally good and completely equivalent. But there clearly is a way out of Dummy traps so there is no reason for making them.

8.8 Chow or Stability Test

- Suppose we have two periods or two groupings of dependent variables and the same of independent variables. Therefore we have two regressions,

$$y_I = X_I\beta_I + u_I^*.$$

$$y_{II} = X_{II}\beta_{II} + u_{II}^*.$$

  With N_I and N_{II} observations respectively.

- We would like to test,

$$H_0 : \beta_I = \beta_{II}.$$

- In this test there are $r = K$ restrictions because the hypothesis is stating that every one of the explanatory variable coefficients is the same in the two regressions.

- Thus, we can do this test with an F test comparing the restricted and unrestricted model (Trinity Test #2).

- Consider the unrestricted model written in matrix notation,

$$\begin{pmatrix} y_I \\ y_{II} \end{pmatrix} = \begin{pmatrix} X_I & 0 \\ 0 & X_{II} \end{pmatrix} \begin{pmatrix} \beta_I \\ \beta_{II} \end{pmatrix} + \begin{pmatrix} u_I^* \\ u_{II}^* \end{pmatrix}. \tag{15}$$

- And the restricted model:

$$y_{I+II} = X_{I+II}\beta_I + u_{I+II}^*.$$

  Where $\beta_I = \beta_{II}$.

- Test statistic:

$$\frac{(RSS_R - RSS_U)/r}{RSS_U/(N_I + N_{II} - 2r)} \sim F(r, N_I + N_{II} - 2r).$$

  Where $RSS_U = RSS_I + RSS_{II}$.

- Note that the unrestricted model which requires estimating two regressions, can be done with dummy variables for each of the two periods, but requires the use of Hadamard Products because matrix multiplication would not work.

8.9 Analysis of Assumptions

- We will now consider the 5 assumptions and consider what happens when one or more of the assumptions fails.

- Consider $A1$.

  - $A1 : \rho(X'X) = K$. If $A1$ is violated, this could be the result of the dummy variable trap because there would be multicollinearity among the explanatory variables. But we have shown that the dummy variable trap is easily avoided and the problem is only in the design. So if $A1$ is violated due to dummy variables, then the solution is simply to drop enough (unnecessary / redundant) variables from the model to raise the rank of $(X'X)$.
  - But there are also two further extreme cases of violations against $A1$:
  - Almost perfect collinearity
    - This still is not a problem, because, sort of like being pregnant, either you are or your aren't: there's no inbetween.
    - If there is perfect collinearity, ie, $\rho(X'X) < K$, $(X'X)$ is not invertible or equivalently $det(X'X) = 0$. If there is almost perfect collinearity, $det(X'X)$ would be very close to zero. So the closer $det(X'X)$ is to zero, the more numerically unstable the coefficients of the regression will be because $(X'X)^{-1}$ would be highly volatile as you have to divide by the determinant to compute it.
  - There is also the case of Perfect Absence of collinearity. This implies that the $X'X$ matrix is diagonal such that,

*

$$(X'X) = \begin{bmatrix} \sum x_{i1}^2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sum x_{i1}^2 \end{bmatrix}. \tag{16}$$

Which implies, because of diagonality,

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{\sum x_{i1}^2} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sum x_{i1}^2} \end{bmatrix}. \tag{17}$$

* But this would only happen in the special case where the explanatory variables are uncorrelated (which might happen with dummy variables, but otherwise is very unlikely).
* Therefore, you have to be "Thick" to fall into the Dummy variable trap and the case of perfect absense of collinearity is so rare that the $A1$ assumption is almost always assumed to be valid.

- Now consider $A2$ and $A3$:

  –

$$A2 : y = X\beta^* + u^* \text{ and } E[u^*] = 0.$$
$$A3R_{mi} : E[u^*|X] = 0.$$
$$A3R_{cu} : Corr(u_i^*, x_{ij}^*) = 0 \; \forall \; j.$$

  – Note that $A3R_{cu} \Rightarrow E[u_i^*, x_{ij}] = 0$ and if $A2$ holds, $Cov(u_i^*, x_{ij}) = 0$.
  – Consider the case where $E[u_i^*] = c \neq 0$. (Violation of $A2$.) Provided that $A3R_{mi}$ is still valid, $\hat{\beta}(OLS)$ will still be unbiased except for the intercept. (If there is no intercept in the problem, we would have problems). Thus, be sure to include an intercept in the regression to avoid these types of problems because the intercept soaks up the bias. NOTE: Sometimes, the reasons that $A2$ is violated would also necessarily imply that $A3R_{mi}$ is also violated!

- Consider violations of $A3$: this is the worst possibly scenerio.

  – Particularly, lets consider violations of $A3R_{mi}$ which implies $A3R_{fi}$ and $A3F$ are also both violated. So,

$$E[u^*|X] = q \neq 0.$$

- Starting from the OLS definition of $\hat{\beta}$,

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

we derived,

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

Thus,

$$E[\hat{\beta} - \beta^*|x] = (X'X)^{-1}X'E[u^*|X] = (X'X)^{-1}X'q \neq 0.$$

- Therefore under this violation, every estimated coefficent will be biased.

- Suppose $A4$ is violated but $A1, A2, A3R_{mi}$ are all ok.
  - Thus,

$$E[\beta^*|X] = E[\hat{\beta}] = \beta^* \Rightarrow \text{Unbiased.}$$

  - Since $A4$ is violated however,

$$V(\hat{\beta}) = (X'X)^{-1}X'E[u^*u^{*'}|X]X(X'X)^{-1}.$$

  - Note that under $A4$, this simplifies to $V(\hat{\beta}) = \sigma^2_{\mu^*}(X'X)^{-1}$.
  - But when $A4$ is not assumed, we can't simplify any more and we write,

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

Where Ω is a positive definite and symmetric matrix. So in this case, even though our estimators are unbiased, the variances and covariances would be wrong if we used the OLS estimates so all significance tests would also be incorrect.

  - Therefore $\hat{\beta}$ is NOT B.L.U.E.
  - By the Gauss Markov theorem however, there must exist a "Best" estimator, so define,

$$\hat{\beta}(\text{Generalized Least Squares}) \equiv \hat{\beta}(GLS).$$

  - Assuming $A1$, $A2$, $A3R_{mi}$, and $A4G$ (for generalized), and also $E[u^*u^{*'}|X] = \Omega$. Define,

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

Which is B.L.U.E. because,

$$V(\hat{\beta}(GLS)) = (X'\Omega^{-1}X)^{-1} \leq (X'X)^{-1}X'\Omega X(X'X)^{-1} = \hat{\beta}(OLS).$$

  - But again, we run into the problem of comparing two matrices: What does "less than or equal to" mean in this sense? When comparing computer output of OLS versus GLS, it may seem like the OLS numbers are better (lower variances etc), but since the OLS esimates are wrong to begin with, they do not provide any sort of basis for comparison.
  - A final note: If we have a lot of observations, Ω being an NxN matrix, might be difficult to invert as required for the estimator. Even super computers would have trouble inverting an NxN matrix for $N = 10000$.

9 Week 9

9.1 More on Generalized Least Squares

- Recall,
$$\hat{\beta}(GLS) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.$$
$$V(\hat{\beta}(GLS)) = (X'\Omega^{-1}X)^{-1}.$$

- We mentioned that Ω^{-1} would be fairly difficult to compute with large N, but there is another problem. We don't even know what it is! It is definitely a function of θ, an unknown parameter vector but that's all we can say.

- Typically, some suitable method will exist to esimate θ. Lets call our estimate $\hat{\theta}$. In general, it will be $\hat{\theta} = f(X, y)$ because X and y are all we know. Note that $f(X, y)$ may or may not be linear in y.

- So now consider (rewriting the β equation),
$$\hat{\beta}(GLS) = (X'\Omega^{-1}(\hat{\theta})X)^{-1}X'\Omega^{-1}(\hat{\theta})y.$$
When we write it in this way, we call it the Feasible Generalised Least Squares Estimator or,
$$\tilde{\hat{\beta}}_{FGLS}.$$

- But $\tilde{\hat{\beta}}_{FGLS}$ is not linear so it cannot be BLUE because of the inverses. It is also not unbiased in general. So it's all around shitty.

- The second issue related to Ω^{-1} is the following. Consider,
$$\Omega^{-1} = \Omega^{-1/2}\Omega^{-1/2'}.$$
We can do this because Ω is a positive definite matrix so Ω^{-1} is also positive definite and by some fundamental theorem of linear algebra, we can write the above equality.

- But there exists at least two different ways of defining the square root matrix, $\Omega^{-1/2}$.
  - Eigenvalue Diagonalization.
  - Cholesky Method where Ω^{-1} = a lower triangular matrix times an upper triangular matrix.

- Either way, we can substitute this into our expression for $\tilde{\hat{\beta}}_{FGLS}$,
$$\hat{\beta}_{FGLS} = (X'\Omega^{-1/2}\Omega^{-1/2'}X)^{-1}X'\Omega^{-1/2}\Omega^{-1/2'}y.$$

- Now define $\tilde{X}$ and $\tilde{y}$ as follows,
$$\hat{\beta}_{FGLS} = (X'\Omega^{-1/2}\underbrace{\Omega^{-1/2'}X}_{\tilde{X}})^{-1}X'\Omega^{-1/2}\underbrace{\Omega^{-1/2'}y}_{\tilde{y}}.$$
Thus, $\hat{\beta}_{FGLS}$ becomes the OLS coefficient of $\tilde{y}$ regressed on $\tilde{X}$.

- However, it still hard to calculate $\Omega^{-1/2}$ but we may be able to find $\tilde{X}$ and $\tilde{y}$ without inverting. This means we need to find out what $\Omega^{-1/2}$ does when it premultiplies y for example. We'll get to this later, but it may be that $\Omega^{-1/2}y$ gives us $y_i - Py_{i-1}$ or $\frac{y_i}{\sigma^2_{\mu^*}}$ for example.

9.2 More Failures of the Assumptions

- What if $A3R_{cu}$ is all that can be assumed and nothing stronger? Therefore,

$$E[\hat{\beta}|X] <=> \beta^*.$$

  Also,

$$E[\tilde{\tilde{\beta}}_{GLS}|X] <=> \beta^*.$$

- Note that $A5N$ gave us,

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

  And as a result we were able to do all sorts of hypothesis tests and inference techniques based on t, F, and χ^2 distributions.

- Without $A5N$ though, we're screwed. "It-goes-out-of-the-window."

- Thus if we only have $A3R_{cu}$ and/or $A5N$ fails, we lose all inference techniques.

- Note that if N is large enough, all of these results become asymptotically true anyways. In linear models, large is about 25 while in nonlinear models, we need samples of around 2000 or more. (More on this next week)

- A Key Complication: What if all versions of $A3$ fail? Then this is the worst possible scenerio. Both in large and small samples, $\hat{\beta}_{OLS}$ and $\tilde{\tilde{\beta}}_{GLS}$ are biased.

10 Week 10

10.1 Leading Causes of Violations of $A3R_{cu}$

- 1. Misspecifications of the Regression function. The regression function is defined as $E[y|X] = X\beta^*$ if the relationship is linear. This is equivalent to saying $y = X\beta^* + u^*$ with $E[u^*|X] = 0$.
  - a. Omission of Important Explanatory Variables.
    - Suppose the true model is:

      $$y = X\beta^* + Z\gamma^* + u^*.$$

      Though we run the regression only on the X's and neglect the Z's. Thus,

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

    - The appropriate estimators would be (via the partitioning results),

      $$\hat{\beta}_{X,Z} = (X'M_ZX)^{-1}X'M_Zy.$$

      $$\hat{\gamma}_{X,Z} = (Z'M_XZ)^{-1}Z'M_Xy.$$

    - But when we run the regression only on the X's, we are assuming the true model is:

      $$y = X\beta^* + v,$$

      where,

      $$v = Z\gamma^* + u^*.$$

    - Taking expectations,

      $$E[\hat{\beta}_X|X] = (X'X)^{-1}X'E[y] = (X'X)^{-1}X'XE[\beta^* + Z\gamma^* + u^*].$$

      Note we know that the true model involves both X's and Z's so we must include that when taking expectations. Thus,

      $$E[\hat{\beta}_X|X] = \beta^* + (X'X)^{-1}X'Z\gamma^*.$$

      This is of course biased because of the second term. The bia is equal to 0 iff $\gamma^* = 0$ (ie, the Z variables weren't important in the regression equation anyways), or $X'Z = 0$, which implies that X and Z would have to be uncorrelated. Correlation would be a violation of $A3R_{cu}$.
    - Thus $E[v|X] \neq 0$ as long as X and Z are correlated!!
  - b. Subcases - Frequent Omissions
    - Frequent omissions in explanatory variables include things like seasonality, non-linearities and other important infrequent events.
    - If these are ignored, you will get biases as shown in case a.

    - One way to determine if there are non-linearities missing from the model is to plot the residuals against the explanatory variables to determine if the nonlinearity in the model has been dumped in the stochastic component.
  - c. Non-Linearities because of "Selectivity."
    - Suppose we run the model: $y = X\hat{\beta} + \hat{u}$, where y is an individual's labor supply and the X's include a set of characteristics of the worker. This is based on the true model: $y = X\beta^* + u^*$.
    - But suppose we only include those observations such that,

$$W = Z\gamma^* + v^* > 0.$$

      Where this equation is called the "participation equation." W is a measure of the wage that the worker could achieve if he worked while the Z variables reflect the worker's reservation wage. If W is greater than the reservation wage, the worker is selected into being in the work force.
    - Thus there are two types of incompleteness of the data: 1) Truncation where we have no data on the W's or Z's and we only observe those that are actually in the labor market and we can't distinguish anything else about them. 2) Censoring: we only have data on the Z characteristics which we use to estimate W. In any case,

$$E[y|X] \neq X\beta^*,$$

      but rather,

$$E[y|X] = X\beta^* + [\text{linear and nonlinear terms involving } Z \text{ and } X].$$

      The other terms depend on the type of incompleteness we have in the data and also on the distributions of u^* and v^*.

- 2. Measurment errors in one more more of the X variables.
  - Next Lecture.
- 3. Endogeneity / Simultaneity.
  - Next Lecture.
- 4. Lagged y's among the X's.
  - Next Lecture.

10.2 Vassilis's Last Lecture - More violations for $A3R_{cu}$

- 1. Misspecification of the regression function (see above).
- 2. Mismeasured explanatory variables.
  - Suppose the true model is of the following form:

    $$y = x_1^*\beta_1^* + u^*.$$

    Where x_1 is a $Nx1$ vector representing the only explanatory variable in the model. However, suppose we don't know x_1^* exactly, but rather we know x_1 such that,

    $$x_1 = x_1^* + v_1.$$

    Here, v_1 is called the measurement error associated with x_1^*. Thus x_1 is a sort of proxy for x_1^*.
  - Thus when we run the regression we are running,

    $$y = x_1^*\beta_1^* + u^*.$$

    $$y = (x_1 - v_1)\beta_1^* + u^*.$$

    $$y = x_1\beta_1^* + u^* - \beta_1^* v_1.$$
  - But this implies that $Corr(x_1, u^* - \beta_1^* v_1) \neq 0$. Also,

    $$Cov(x_1, u^* - \beta_1^* v_1) = \beta_1^*\sigma_v^2 + \text{other terms possibly} \neq 0.$$

  - Thus the OLS estimate would be biased. Note that this type of error is VERY common. Most explanatory variables that we use in regressions are measured with a certain degree of error.
  - So can we say anything else about this bias that has been introduced? Yes, we can determine the direction of the bias. Consider,

    $$\hat{\beta}_1 - \beta_1^* = (x_1'x_1)^{-1}x_1'[Error].$$

    $$\hat{\beta}_1 - \beta_1^* = (x_1'x_1)^{-1}x_1'[u^* - \beta_1^* v_1].$$

    Note that we cannot even take expectations of this quantity because the x's are completely stochastic in that they contain the measurement error which is by definition, random. Thus they are beyond our control.
  - We can say, that because of the negative sign in front of the β_1^*, we know the the bias will be in the opposite direction of the sign of the coefficient on the mismeasured variable. Thus, if β_1^* is positive, the bias is negative and makes the coefficient smaller, ie, closer to zero. If β_1^* is negative, the bias is positive and makes the coefficient larger, ie, closer to zero. Thus in both cases, the coefficient is pushed towards zero when there is bias in the model.

- This type of bias is called "attenuation bias" because the coefficient shrinks to zero and might lead to hypothesis tests rejecting their significance.
- This is known as the "Ivon Law of Econometrics."
- Though Ivon's law applies when there is only one biased coefficient in the model (case 1) it is not so lawful when you consider the following two cases.
- Case 2. 1 biased estimator and 1 or more other non-biased estimators.
  * Suppose the true model is:

$$y = x_1^*\beta_1^* + X_2\beta_2^* + u^*.$$

    Where $x_1 = x_1^* + v_1$ and the other variables in the X_2 matrix are all perfectly measured.
  * The result is that $\hat{\beta}_1$ will suffer from attenuation bias and is pushed towards zero, but we cannot say anything about the other parameter estimates. They will be biased is any direction depending on correlations between x_1 and X_2.
- Case 3. 2 or more biased estimators and 1 or more other non-biased estimators.
  * Suppose the true model is:

$$y = x_1^*\beta_1^* + x_2^*\beta_2^* + X_3\beta_3^* + u^*.$$

    Where $x_1 = x_1^* + v_1$ and $x_2 = x_2^* + v_2$.
  * The result if that $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ may be biased in any direction depending on the variances of measurement error and correlations between the x's. We cannot therefore draw attenuation conclusions.

• 3. Simultaneity Problems.

- This occurs when some of the explanatory variables are simultaneously determined together with the left hand side variable in the same econometric system.
- Consider the following illustration. Consider a supply and demand model were we are estimating the demand side using price and income as explanatory variables. Thus,

$$y_t^d = \alpha_1 + \beta_1 P_t + \gamma_1 I_t + u_{1t}^*.$$

  It is quite likely that p_t and u_{1t}^* are correlated because the demand works with the supply to determine the price level in a market clearing environment. Define the supply as a function of price and weather and write it inversely so we will be able to substitute it into the demand equation presently. Thus,

$$P_t = \alpha_2 + \beta_2 y_t^s + \gamma_2 W_t + u_{2t}^*.$$

  Substituting under $y_t^d = y_t^s = y_t$,

$$y_t = \alpha_1 + \beta_1 \left[\alpha_2 + \beta_2 y_t^s + \gamma_2 W_t + u_{2t}^* \right] + \gamma_1 I_t + u_{1t}^*.$$

Note that the demand equation $y = y(P, I)$ is a function of the price level. Also, the supply equation in terms of price, $P = P(y, W)$ is a function of the demand level. So P_t is correlated with u^*_{1t} and y_t is correlated with u^*_{2t} which is a violation of $A3R_{cu}$. P and y are simultaneously determined and thus they are endogenous in the model. Note that the intercept, W_t, and I_t are all exogeneous variables.

- 4. Lagged y's on the right hand side combined with violations of $A4$.
  - Suppose our model is as follows:

$$y_t = \beta_1 X_t + \beta^*_2 y_{t-1} + \beta^*_3 y_{t-2} + u^*_t.$$

    Thus taking this equation back one and two periods,

$$y_{t-1} = \beta_1 X_{t-1} + \beta^*_2 y_{t-2} + \beta^*_3 y_{t-3} + u^*_{t-1}.$$

$$y_{t-2} = \beta_1 X_{t-2} + \beta^*_2 y_{t-3} + \beta^*_3 y_{t-4} + u^*_{t-2}.$$

  - This alone is ok because in the original model above, the explanatory variables are not correlated with the contemporaneous error terms, only with the lagged errors.
  - But suppose now that $A4$ is violated and all we have is $A4G$:

$$E(u^*_t \cdot u^*_s) \neq 0 \text{ for some } t \neq s.$$

    If this occurs and we have lags in the model, we get a violation of $A3R_{cu}$ because the error terms in different periods are correlated which makes the lagged variables correlated with u^*_t. This is very very bad.

10.3 One final note on the some Asymptotic Results

- The Law of Large Numbers (LLN): Sample moments become true population moments and the parameter coefficients become asymptotically unbiased as N gets large.
- The Central Limit Theorem (CLT): As N gets large, everything becomes normal and we can apply all inference procedures relying on the t, F, and χ^2 distributions.