ECON 257D2 Guide and Review

Nikhil Raman

Winter 2024

Information

Professor Dufour's Website

1 Point Estimation

Would like to link estimated values with unobservable parameters. For example, μ, σ^2, β (regression coefficient), can be estimated with observable data, through $\bar{X}, s^2, \hat{\beta}$.

1.1 Estimators

For any parameter of data or a distribution θ, there is an estimated value $\hat{\theta}$ that takes the form $\hat{\theta} = f(x_1, x_2, ..., x_n)$, doesn't depend on θ. Estimator is consistent if $\hat{\theta} \xrightarrow{p} \theta$ as sample size $N \to \infty$

Bias

The *bias* of $\hat{\theta}$ is defined as $\mathbb{E}(\hat{\theta}) - \theta$. An estimator is said to be unbiased if its bias is 0

Efficiency

An estimator is more *efficient* if its variance is lower. If $\mathbb{V}(\theta_1) < \mathbb{V}(\theta_2)$, θ_1 is more efficient than θ_2. For a covariance matrix, we show an estimator is more efficient than another if the their difference is positive semidefinite.

Loss Function

Loss associated of estimate with true value θ . denoted generally as $l(\hat{\theta}, \theta)$. A risk function is the expectation of the loss function, but the two are typically used interchangeably. Examples can be the Mean Squared Error, Linex, or Mean Absolute Error

1.2 Least Squares

The Mean Squared Error (MSE) Loss Function has the form:

$$\mathbb{E}\Big[\Big(\hat{\theta} - \theta\Big)\Big]^2$$

Therefore our estimator seeks to minimize the sum of these "residuals" The MSE formula can be decomposed into the following:

$$= \mathbb{E}[(\hat{\theta} - \theta)]^2 - \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})^2]$$

Where the first term is the bias squared and the second term is $\mathbb{V}(\hat{\theta})$. We would like to minimize both, but they go in opposite directions. Later the Gauss-Markov Theorem will show that minimizing squared residuals leads to the most efficient (best) linear unbiased estimator

2 Covariance and Correlation

2.1 Notation and Preliminaries

(a) a.s. means "almost surely," with probability 1

(b) ":=" means defined as

(c) $\iff$ means an equivalent condition, both imply each other

(d) Expectation/mean: $\mathbb{E}(X) = \mu_X$

  (i) $\mathbb{E}(X^2) = 0 \iff X = 0$ a.s.

  (ii) $\mathbb{E}(a + bX) = a + b\mathbb{E}(X)$

(e) Variance: $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \sigma_X^2$

  (i) $\mathbb{V}(X) = 0 \iff X = \mathbb{E}(X)$ a.s. (degenerate random variable)

  (ii) $\mathbb{V}(a + bX) = b^2\mathbb{V}(X)$

(f) Standard Deviation: $\sigma(X) = \sigma_X = \sqrt{\sigma_X^2}$

  (i) $\sigma(a + bX) = |b|\sigma_X$

2.2 Definitions

Covariance and Correlation provide a measure of association between random variables.

Covariance

$$C(X, Y) := \sigma_{XY} := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

$C(X, Y) = 0 \iff$ X and Y orthogonal

Correlation

$$\rho(X, Y) := \rho_{XY} := \frac{C(X,Y)}{\sigma_X \sigma_Y}$$

if $C(X, Y) = 0$ and/or $\sigma_X\sigma_Y = 0$, then $\rho(X, Y)$ is set to 0

2.3 Properties

(a) $C(a + bX, c + dY) = bdC(X, Y)$

(b) $\rho(a + bX, c + dY) = \rho(X, Y)$ *(Covariance is unscaled, but correlation is)*

(c) $C(X, Y) = C(Y, X)$, $\rho(X, Y) = \rho(Y, X)$

(d) $C(X, X) = \mathbb{V}(X)$

(e) $\rho(X, X) = 1$

(f) $C(X, Y)^2 \leq \mathbb{V}(X)\mathbb{V}(Y)$ *(Cauchy-Schwarz Inequality)*

(g) $|\rho(X, Y)| \leq 1$ *(From part f)*

(h) X and Y independent $\implies C(X, Y) = 0$, $\rho(X, Y) = 0$

(i) $\rho(X, Y)^2 = 1 \iff Y = a + bX$ a.s. *(X and Y can be expressed with linear combination)*

2.4 Regression Coefficients

From property (i), we can be more precise about what parameters will give a linear relationship between two variables. In particular, b is the regression coefficient of Y on X

Linear Regression Coefficient of Y on X

$$\beta(Y \rightharpoonup X) := \frac{C(X,Y)}{\mathbb{V}(X)}$$

$\beta(Y \rightharpoonup X)$ is set to 0 when $\mathbb{V}(X)$ is 0, $\beta(Y \leftharpoonup X) = \beta(X \rightharpoonup Y)$. Harpoons represent statistical dependence or predictability.

3 Random Vectors and Covariance Matrices

Suppose instead of one random variable, we had multiple we wished to speak of, $X_1, X_2, ...X_k$. To make it easier, we compact them into a random vector $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}, \mathbb{E}(\mathbf{X}) = \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_k) \end{bmatrix}$$

3.1 Covariance Matrix

For random vector $\mathbf{X}$, there is no longer one variance value but a k×k covariance matrix $\mathbf{\Sigma}$. Each entry in the matrix is the covariance between the random variables X_i and X_j, σ_{ij}

$$\mathbf{\Sigma}(X) := \mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))']$$

3.2 Properties

Given $\mathbf{X}$ as defined above, $a, b \in \mathbb{R}^k, \mathbf{A} \in \mathbb{R}^{g \times k}$, and $\alpha \in \mathbb{R}$

(a) $\mathbb{E}(\mathbf{X} + a) = \mathbb{E}(\mathbf{X}) + a$

(b) $\mathbb{E}(\alpha\mathbf{X}) = \alpha\mathbb{E}(\mathbf{X})$

(c) $\mathbb{E}(a'\mathbf{X}) = a'\mathbb{E}(\mathbf{X})$, $\mathbb{E}(\mathbf{AX}) = \mathbf{A}\mathbb{E}(\mathbf{X})$

(d) $\mathbb{V}(\mathbf{X} + a) = \mathbb{V}(\mathbf{X})$

(e) $\mathbb{V}(\alpha\mathbf{X}) = \alpha^2\mathbb{V}(\mathbf{X})$

(f) $\mathbb{V}(a'\mathbf{X}) = a'\mathbb{V}(\mathbf{X})a$, $\mathbb{V}(\mathbf{AX}) = \mathbf{A'}\mathbb{V}(\mathbf{X})\mathbf{A}$

(g) $C(a'\mathbf{X}, b'\mathbf{X}) = a'\mathbb{V}(\mathbf{X})b = b'\mathbb{V}(\mathbf{X})a$

Generalized Variance

$|\mathbb{V}(\mathbf{X})|$ is the generalized variance of X

3.3 Theorem on Covariance Matrices

(a) $\mathbf{\Sigma}' = \mathbf{\Sigma}$

(b) $\mathbf{\Sigma}$ is p.s.d.

(c) $\mathbf{\Sigma}$ is p.d. $\iff$ $\mathbf{\Sigma}$ is nonsingular (invertible)

(d) $0 \leq |\mathbf{\Sigma}| \leq \prod_{i=1}^{k} \sigma_i^2$

(e) $|\mathbf{\Sigma}| = 0 \iff$ at least one random variable X_i in $\mathbf{X}$ is linearly dependent

Note:

- Positive Semi-Definite (p.s.d.) $\iff v'\mathbf{A}v \geq 0 \; \forall v$
- Positive Definite (p.d.) $\iff v'\mathbf{A}v > 0 \; \forall v \neq 0$

Covariance between Two Random Vectors

Let $\mathbf{Y}$ be another random vector but of dimension $j \times 1$.

$$C(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))']$$

if $k = j$, then the covariance matrix is square and we can say $|C(\mathbf{X}, \mathbf{Y})|$ is the generalized covariance between $\mathbf{X}$ and $\mathbf{Y}$

3.4 More Properties

$a \in \mathbb{R}^k$, $b \in \mathbb{R}^j$, $\alpha, \beta \in \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{g_1 \times k}$, $\mathbf{B} \in \mathbb{R}^{g_1 \times j}$

(a) $C(\mathbf{X}, \mathbf{Y}) = \mathbb{E}(\mathbf{X}\mathbf{Y}') - \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{Y})'$

(b) $C(\mathbf{X}, \mathbf{Y}) = C(\mathbf{Y}, \mathbf{X})'$

(c) $C(\mathbf{X}, \mathbf{X}) = \mathbb{V}(\mathbf{X})$, $C(\mathbf{Y}, \mathbf{Y}) = \mathbb{V}(\mathbf{Y})$

(d) $C(a + \mathbf{X}, b + \mathbf{Y}) = C(\mathbf{X}, \mathbf{Y})$

(e) $C(\alpha\mathbf{X}, \beta\mathbf{Y}) = \alpha\beta C(\mathbf{X}, \mathbf{Y})$

(f) $C(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) = \mathbf{A}'C(\mathbf{X}, \mathbf{Y})\mathbf{B}$

(g) if k = j, and $\mathbf{Z}$ is $h \times 1$, $C(\mathbf{X} + \mathbf{Y}, \mathbf{Z}) = C(\mathbf{X}, \mathbf{Z}) + C(\mathbf{Y}, \mathbf{Z})$

(h) if k=j, $\mathbb{V}(\mathbf{X} \pm \mathbf{Y}) = \mathbb{V}(\mathbf{X}) + \mathbb{V}(\mathbf{Y}) \pm C(\mathbf{X}, \mathbf{Y}) \pm C(\mathbf{Y}, \mathbf{X})$

Partitioned Covariance Matrix

Let $\mathbf{X}$ take the form:

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

Where $\mathbf{X_1}, \mathbf{X_2}$ are $k_1 \times 1$ and $k_2 \times 1$ respectively. Its covariance matrix is resultingly:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

$\Sigma_{11} = \mathbb{V}(\mathbf{X_1})$, $\Sigma_{22} = \mathbb{V}(\mathbf{X_2})$, and $\mathbf{\Sigma}_{12} = C(\mathbf{X_1}, \mathbf{X_2}) = \mathbf{\Sigma_{21}}'$

Equivalent Conditions for Paritioned Covariance Matrix

(a) $\mathbf{\Sigma}$ is nonsingular

(b) $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}$ are nonsingular

(c) $|\mathbf{\Sigma}_{11}| > 0$ and $|\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}| > 0$

(d) $|\mathbf{\Sigma}| = |\mathbf{\Sigma}_{11}| \cdot |\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}|$

4 Multinormal Distribution

Given $\mathbf{X}$ from above:

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

We know if $\mathbf{X_1}$ and $\mathbf{X_2}$ are independent $\implies C(\mathbf{X_1}, \mathbf{X_2}) = 0$. The converse is generally not true, except if $\mathbf{X}$ is *multinormal*

4.1 Multinormal Density

If $\mathbf{X} \sim N_k[\mu, \mathbf{\Sigma}]$, its density is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

4.2 Theorems on Multinormally Distributed Variables

If $\mathbf{X} \sim N_k[\mu, \mathbf{\Sigma}]$, $a \in \mathbb{R}^k, A \in \mathbb{R}^{g \times k}$:

(a) $\mathbf{X} + a \sim N_k[\mu + a, \mathbf{\Sigma}]$

(b) $a'\mathbf{X} \sim N_1[a'\mu, a'\mathbf{\Sigma}a]$

(c) $A\mathbf{X} \sim N_g[A\mu, A\mathbf{\Sigma}A']$

When $\mathbf{X}$ is as such:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} \sim N_k\left(\begin{bmatrix} \mu_\mathbf{1} \\ \mu_\mathbf{1} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}\right)$$

(d) $\mathbf{X_1} \sim N_{k_1}[\mu_1, \mathbf{\Sigma}_{11}], \mathbf{X_2} \sim N_{k_2}[\mu_2, \mathbf{\Sigma}_{22}]$,

(e) $\mathbf{X_1}$ and $\mathbf{X_2}$ independent $\iff \mathbf{\Sigma}_{12} = 0$

(f) Conditional Distribution: $\mathbf{X_2}|\mathbf{X_1} \sim N_k\left[\mu_2 + \mathbf{\Sigma_{21}\Sigma_{11}^{-1}}(\mathbf{X_1} - \mu_\mathbf{1}), \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\right]$

Chi-Squared Multinormal Theorem

If $\mathbf{X} \sim N_k[\mu, \mathbf{\Sigma}]$, with $|\mathbf{\Sigma}| \neq 0$, then we can create a multivariate chi-squared distribution:

$$(\mathbf{X} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{X} - \mu) \sim \boldsymbol{\chi}^2(k)$$

5 Multiple Linear Regression

5.1 Least Squares Revisited

To predict a random variable Y with a vector of random variables $\mathbf{X}$, lets reformulate the least squares problem to one compatible with the notion of Random Vectors

$$\min_b S(b) := \mathbb{E}[Y - \mathbf{X}'b]^2$$

Let β be a vector such that the uncentered covariance between it and the error is zero:

$$\mathbb{E}[\mathbf{X}(Y - \mathbf{X}'\beta)] = 0$$

After some matrix algebra, we find the minimizing choice of $S(b)$ to be β. This makes sense since a covariance of 0 means random variables are orthogonal, which implies its the shortest distance and thus the best approximation.

5.2 Fitted Values, Residuals, R squared

For each β, there is a unique approximation or "fitted value":

$$P(Y; \mathbf{X}) := \mathbf{X}'\beta$$

Subsequently, the residuals for each β, are denoted:

$$U(Y; \mathbf{X}) := Y - P(Y; \mathbf{X})$$

Thus,

(a) $Y = P(Y; \mathbf{X}) + U(Y; \mathbf{X})$

(b) $\mathbb{E}[\mathbf{X}U(Y; \mathbf{X})] = 0$

(c) $\mathbb{E}[P(Y; \mathbf{X})U(Y; \mathbf{X})] = 0$

Finally, we define the uncentered r-squared as the ratio of the uncentered variances of Y and its predicted values

$$R_0^2 := \frac{P(Y; \mathbf{X})^2}{Y^2}$$

6 Ordinary Least Squares (OLS)

6.1 Model-Free Linear Regression

As mentioned earlier, we have a variable y (a $T \times 1$ vector of observations), and a set of variables $X = (X_1, \ldots, X_k)'$ (a $T \times k$ matrix of variables and their observations). Let's call y the dependent variable and X the set of explanatory variables. "Model-free" simply means assumption-free.

In order to explain y through X, we need a vector of parameters β, and a vector of unexplained residuals ε such that:

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

Or for each individual variable:

$$y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \ldots + X_{ki}\beta_k + \varepsilon_i$$

β and ε are the unobserved, "true" values, so we need to do a least squares process to estimate them.

6.2 Minimization Problem and Solution

Least squares operates by finding the β that minimizes the sum of squares function S, which we use as a function of β:

$$\min_{\beta} S(\beta) \equiv (y - X\beta)(y - X\beta)'$$

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'y + 2X'X\beta$$

Let $\hat{\beta}$ be the value of β that minimizes $S(\beta)$, and we get the *normal equations*:

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

Solving for $\hat{\beta}$ gives:

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

Now that we have our estimated parameters, we can created fitted values $\hat{y}$, being equal to $X\hat{\beta}$

6.3 Properties of OLS Objects

Projection Matrix P

(a) $P = X(X'X)^{-1}X'$

(b) $\hat{y} = X\hat{\beta} \implies \hat{y} = Py$

(c) $P' = P$, $PP = P$ (P is symmetric and idempotent)

(d) $PX = X$

Annihilator Matrix M

(a) $M = I - P$

(b) $\hat{\varepsilon} = My$

(c) $MX = 0$

(d) $PM = 0$, $MP = 0$

Orthogonality

(a) $X'\hat{\varepsilon} = 0$

(b) $\hat{y}'\hat{\varepsilon} = 0$

(c) $y = \hat{y} + \hat{\varepsilon}$

7 Classical Linear Model

You may have noticed that in OLS we had to make an assumptions in order for our solution to be valid. Namely, if $\text{rank}(X) = k$, $\text{rank}(X'X)$ must also be k so that it is invertible.

We will make more basic assumptions to create the *Classical Linear Model* (CLM)

7.1 Assumptions

Note: y, β, and ε are $T \times 1$ vectors, X is a $T \times k$ matrix.

(a) **Specification**: $y = X\beta + \varepsilon$

(b) **Errors have mean** 0: $\mathbb{E}(\varepsilon) = 0$

(c) **Homoskedasticity**: $\mathbb{E}(\varepsilon'\varepsilon) = \sigma^2 I_T$

(d) **Non-Stochasticity**: X is a fixed matrix

(e) **Full-rank**: $\text{rank}(X) = k < T$

(f) **NCLM**: ε is distributed multinormally

7.2 Linear Unbiased Estimation and Observations

Following from the assumptions, we can make observations on certain properties and show that OLS is unbiased.

(a) $\mathbb{E}(y) = X\beta$

(b) $\mathbb{V}(y) = \sigma^2 I_T = \mathbb{V}(\varepsilon)$

(c) $\hat{\beta}$ is linear with respect to y

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

(e) $\hat{\beta}$ is an unbiased estimator of β, $\mathbb{E}(\hat{\beta}) = \beta$

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

7.3 Gauss-Markov Theorem

We can show that the OLS estimator $\hat{\beta}$ is the *Best Linear Unbiased Estimator* (BLUE) in its class, by proving its variance is the lowest. Given another linear unbiased estimator $\tilde{\beta} = Cy$, Gauss-Markov shows that $V(\tilde{\beta}) - \mathbb{V}(\hat{\beta})$ is p.s.d. If $D = C - (X'X)^{-1}X'$, then:

$$\mathbb{V}(\tilde{\beta}) = \mathbb{V}(\hat{\beta}) + \sigma^2 DD'$$

Where $\sigma^2 DD'$ is p.s.d.

Corollary to Linear Combinations of Estimators

If w is a $k \times 1$ vector of constants, then:

$$\mathbb{V}(w'\tilde{\beta}) \geq \mathbb{V}(w'\hat{\beta})$$

Generalized Gauss-Markov

Let L be an $r \times k$ fixed matrix such that $\gamma = L\beta$. $\hat{\gamma}$ is the BLUE of γ, because:

$$\mathbb{V}(\tilde{\gamma}) = \mathbb{V}(\hat{\gamma}) + \sigma^2 DD'$$

Further, we can say the the inefficient estimator does not provide any new information, just noise:

$$C(\tilde{\gamma} - \hat{\gamma}, \hat{\gamma}) = 0$$

Quadratic Gauss-Markov Optimality

If Q is an $r \times r$ fixed matrix, then:

(a) $\mathbb{E}\left[(\tilde{\gamma}-\gamma)'Q(\tilde{\gamma}-\gamma)\right] \geq \mathbb{E}\left[(\hat{\gamma}-\gamma)'Q(\hat{\gamma}-\gamma)\right]$

(b) $\text{tr}[\mathbb{V}(\tilde{\gamma})] \geq \text{tr}[\mathbb{V}(\hat{\gamma})]$

(c) $|\mathbb{V}(\tilde{\gamma}| \geq |\mathbb{V}(\hat{\gamma})|$

7.4 Other Conclusions about OLS Estimators

Based on everything above, there are numerous observations to be made about the remaining least squares estimators, $\hat{y}$ and $\hat{\varepsilon}$

(a) $\hat{y} = X\beta + P\varepsilon$, $\mathbb{E}(\hat{y}) = X\beta$

(b) $\hat{\varepsilon} = My = M\varepsilon$, $\mathbb{E}(\hat{\varepsilon}) = 0$

(c) $\mathbb{V}(\hat{y}) = \sigma^2 P$

(d) $\mathbb{V}(\hat{\varepsilon}) = \sigma^2 M$

(e) $\hat{y}$ is the BLUE of $X\beta$, and $\hat{\varepsilon}$ is the BLUE of ε

(f) $C(\hat{\beta}, \hat{\varepsilon}) = 0$

(g) $C(\hat{y}, \hat{\varepsilon}) = 0$

Estimation of variance

We know $\mathbb{E}(\varepsilon'\varepsilon) = \sigma^2 I_T$, but since this is unobservable, we would like to create an estimate of σ^2 using $\hat{\varepsilon}$.

$\mathbb{E}(\hat{\varepsilon}'\hat{\varepsilon})$ can be shown as equal to $\sigma^2(T-k)$ since $\hat{\varepsilon} = M\varepsilon$, and $\text{tr}(M)$ is equal to $T-k$, so:

$$\mathbb{E}\left[\frac{\hat{\varepsilon}'\hat{\varepsilon}}{T-k}\right] = \sigma^2$$

Therefore we define the statistic s^2, an unbiased estimator of σ^2, and $s^2(X'X)^{-1}$ as an unbiased estimator of $\mathbb{V}(\hat{\beta})$:

$$s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{T-k}$$

8 Prediction

We used least squares to predict a relationship between y and X using $\hat{\beta}$, but there is another unobservable y_0 of m future observations that we'd like to predict. The problem can be formulated into:

$$y_0 = X_0\beta + \varepsilon_0$$

ε_0 follows the same assumptions as the classical linear model, except $\mathbb{E}(\varepsilon\varepsilon_0') = 0$ such that it is uncorrelated with ε.

The natural predictor to use in this case is $\hat{\beta}$, where the predicted values of y_0 take the form:

$$\hat{y_0} = X_0\hat{\beta} = X_0X'(X'X)^{-1}y$$

8.1 Properties of Predictors

(a) $\hat{y_0}$ is an unbiased estimator for $X_0\beta$

(b) $\mathbb{V}(\hat{y_0}) = \sigma^2 X_0'(X'X)^{-1}X_0$

(c) $C(y_0, \hat{y_0}) = 0$

(d) $\hat{y_0}$ is the BLUE for $X_0\beta$

Prediction Errors

The prediction errors of the model $\hat{e_0}$, different from the general error terms, is defined as:

$$\hat{e_0} = \hat{y_0} - y_0 = \varepsilon_0 - X_0(\beta - \hat{\beta})$$

(e) $\hat{y_0}$ is a linear unbiased predictor (LUP) of y_0

(f) $\mathbb{E}(\hat{e_0}) = 0$

(g) $\mathbb{V}(\hat{e_0}) = \sigma^2[I_m - X_0(X'X)^{-1}X_0']$

(h) $\hat{y_0}$ is the best linear unbiased predictor (BLUP) of y_0

9 Estimation with Guassian Errors

Earlier, we made the assumption that $\varepsilon \sim N_T[0, \sigma^2 I_T]$ to form the normal classical linear model (NCLM).

This implies each ε_t is i.i.d. $N[0, \sigma^2]$, allowing us to establish the distribution of all the other OLS estimators.

(a) $y \sim N_T[X\beta, \sigma^2 I_T]$ since $y = X\beta + \varepsilon$

(b) $\hat{\beta} \sim N_T[\beta, \sigma^2(X'X)^{-1}]$

(c) $\hat{y} \sim N_T[X\beta, \sigma^2 P]$

(d) $\hat{\varepsilon} \sim N_T[0, \sigma^2 M]$

(e) $\hat{\varepsilon}$ and $\hat{\beta}$ are independent

(f) $\hat{\varepsilon}$ and $\hat{y}$ are independent

9.1 Maximum Likelihood Estimation (MLE)

We now have a distribution for y, where $\mu = X\beta$ and $\mathbf{\Sigma} = \sigma^2 I_T$. We now define the likelihood function, separate from the probability density f, which describes the probability of y occurring *given* any values for parameters:

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

To use the maximum likelihood method to estimate parameters, we can maximize the log of L.

$$\frac{\partial \ln(L)}{\partial \beta} = -\frac{1}{2\sigma^2}[-2X'y + 2(X'X)^{-1}\beta]$$

$$\frac{\partial \ln(L)}{\partial \sigma^2} = -\frac{T}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta)$$

Applying the first order conditions, we find that MLE leads to $\hat{\beta}$ being the same BLUE from least squares, however we must separately defined the biased estimator $\hat{\sigma}^2$:

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

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T}$$

9.2 Distribution of Idempotent Quadratic Form

If Q is a $T \times T$ symmetric idempotent matrix, with $\text{rank}(Q) = q \leq T$, then:

$$\frac{\varepsilon' Q \varepsilon}{\sigma^2} \sim \chi^2(q)$$

Other Properties

$$\frac{S(\hat{\beta})}{\sigma^2} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} \sim \chi^2(T-k)$$

Let R be a $q \times k$ fixed matrix. Then $R\hat{\beta}$ has the distribution below and is independent with s^2:

$$R\hat{\beta} \sim N_q[R\beta, \sigma^2 R(X'X)^{-1}R']$$

If R is $q \times k$ with rank q, and $r = R\beta$, then:

$$\frac{S(R, \hat{\beta})}{\sigma^2} = \frac{}{\sigma^2} \sim \chi^2(q)$$

10 Confidence and Prediction Intervals

From the NCLM and above, we established the estimated sum of squares divided by σ^2 follows a $\chi^2(T-k)$ distribution. Therefore there exists values of a and b in the distrubution such that:

$$\Pr\left[\chi^2(T-k) > a\right] = \frac{\alpha}{2}$$

$$\Pr\left[\chi^2(T-k) < b\right] = \frac{\alpha}{2}$$

$$\Pr\left[a < \chi^2(T-k) < b\right] = 1 - \alpha$$

We can then generate a confidence interval for σ^2 using the statistic s^2:

$$\Pr\left[\frac{(T-k)s^2}{b} \leq \sigma^2 \leq \frac{(T-k)s^2}{a}\right] = 1 - \alpha, \text{ since } \hat{\varepsilon}'\varepsilon = \frac{s^2}{T-k}$$

This isn't the smallest confidence interval but it is a decent one.

10.1 Linear Combination of Regression Coefficients

Consider a linear combination $w'\beta$. We can establish the distribution of $w'\hat{\beta} - w'\beta$ as normal:

$$w'\hat{\beta} - w'\beta \sim N\left[0, \sigma^2 w'(X'X)^{-1}w\right]$$

If we let $\Delta = \sqrt{w'(X'X)^{-1}w}$, then:

$$\frac{w'\hat{\beta} - w'\beta}{\Delta\sigma} \sim N[0,1]$$

Therefore if we define the variable t as such, where $Y \sim N[0,1]$ and $X \sim \chi^2(T-k)$:

$$\begin{aligned} t &= \frac{w'\hat{\beta}-w'\beta}{\Delta s} \\ &= \frac{w'\hat{\beta}-w'\beta}{\Delta\sigma} / \sqrt{\frac{(T-k)s^2}{(T-k)\sigma^2}} \\ &= Y/\sqrt{\frac{X}{T-k}} \end{aligned}$$

t follows the symmetric t-distribution with $T-k$ degrees of freedom. So to generate a confidence interval for $w'\beta$:

$$\Pr\left[w'\hat{\beta} - t_{\alpha/2}s\Delta \leq w'\beta \leq w'\hat{\beta} + t_{\alpha/2}s\Delta\right] = 1-\alpha$$

Where $t_{\alpha/2}$ is the $\frac{\alpha}{2}$ quantile of the distribution.

10.2 Confidence Region for Vector of Coefficients

Instead of a scalar linear combination of β, we now make a confidence region of many dimensions for a vector of linear combinations $R\beta$.

We know that $S(R,\hat{\beta})/\sigma^2 \sim \chi^2(k)$, but since σ^2 is unknown, let us make the following proposition:

$$F = \frac{S(R,\hat{\beta})}{qs^2} = \frac{S(R,\hat{\beta})/q\sigma^2}{(T-k)s^2/(T-k)\sigma^2} = \frac{X_1/q}{X_2/(T-k)}$$

where X_1 and X_2 are independent chi-squared variables. Therefore F follows the Fisher distribution with $q, T-k$ degrees of freedom. If we let F_α be defined as below, we can make a confidence region for $R\beta$:

$$\Pr\left[F(q,T-k) > F_\alpha\right] = \alpha$$

$$\Pr\left[\frac{S(R,\hat{\beta})}{qs^2} \leq F_\alpha\right] = 1-\alpha$$

The set is an ellipsoid.

10.3 Prediction Intervals

If we have a y_0 of one additional period, we can easily extend the t-distribution confidence interval from before to suit a prediction problem:

$$y_0 = x_0'\beta + \varepsilon_0$$

$$\begin{bmatrix}\varepsilon \\ \varepsilon_0\end{bmatrix} \sim N[0, \sigma^2 I_{T+1}]$$

Recalling the definition of $\hat{y_0}$, we can establish the distribution of the prediction errors:

$$y_0 - \hat{y_0} \sim N\left[0, \sigma^2(1 + x_0'(X'X)^{-1}x_0)\right]$$

Further, if we define Δ_1 as $\sqrt{1 + x_0'(X'X)^{-1}x_0}$:

$$\frac{y_0 - \hat{y_0}}{\sigma\Delta_1} \sim N[0,1]$$

$$\frac{y_0 - \hat{y_0}}{s\Delta_1} \sim t(T-k)$$

$$\Pr\left[\hat{y_0} - t_{\alpha/2}s\Delta_1 \leq y_0 \leq \hat{y_0} + t_{\alpha/2}s\Delta_1\right] = 1-\alpha$$

10.4 Confidence Region for Several Predictions

Combining what we did in the previous two sections, we finalize by extending our prediction interval to multiple periods.

$$y_0 = X_0\beta + \varepsilon_0$$

$$\begin{bmatrix} \varepsilon \\ \varepsilon_0 \end{bmatrix} \sim N[0, \sigma^2 I_{T+m}]$$

Establishing the distribution of the prediction errors $\hat{e_0}$:

$$\hat{e_0} \sim N\left[0, \sigma^2(I_m + X_0(X'X)^{-1}X_0')\right] \text{ ,where } I_m + X_0(X'X)^{-1}X_0' = D_0$$

Using the Chi-Squared Multinormal Theorem, we can make the following claim, since D_0 is the covariance matrix of $\hat{e_0}$.

$$\frac{\hat{e_0}D_0^{-1}\hat{e_0}}{\sigma^2} \sim \chi^2(m)$$

Finally, we replace σ^2 with $s^2(T-k)$, the ratio of which is also chi-squared, to get an F-distributed variable:

$$F = \frac{\hat{e_0}D_0^{-1}\hat{e_0}/\sigma^2 m}{(T-k)s^2/(T-k)\sigma^2} \sim F(m, T-k)$$

$$(y_0 - \hat{y_0})'[I_m + X_0(X'X)^{-1}X_0']^{-1}(y_0 - \hat{y_0}/ms^2 \sim F(m, T-k)$$

Thus, y_0 that yields $F \leq F_\alpha(m, T-k)$ is a confidence region of level $1-\alpha$.

11 Hypothesis Tests

With techniques for creating confidence intervals, we can now start to test hypotheses surrounding β.

11.1 Linear Combination of Coefficients, Two-Sided and One-Sided Tests

We would like to test if some linear combination of β is significantly different from a value w_0. The null takes the form:

$$H_0 : w'\beta = w_0$$

Looking at the difference $w'\hat{\beta} - w_0$, we can easily establish it as normally distributed with mean 0 and variance $\sigma^2 w'(X'X)^{-1}w$ from section 10.1. Therefore, under the null,

$$t = \frac{w'\hat{\beta} - w_0}{\Delta s} \sim t(T-k)$$

with s and Δ as defined above. We can then perform the following tests to reject or accept:

1. **Two-Sided Test** to reject H_0 as $w'\beta - w_0 \neq 0$ at level α if $|t| \geq t_{\alpha/2}$
2. **One-Sided Test** to reject H_0 at level α as $w'\beta > w_0$ when $t \geq t_{\alpha/2}$
3. **One-Sided Test** to reject H_0 at level α as $w'\beta < w_0$ when $t \leq t_{\alpha/2}$

11.2 Wald Test

We can go further and test the individual values of β's components through a vector linear transformation, where R is a full-rank $q \times k$ matrix:

$$H_0 : R\beta = r = \begin{bmatrix} w_1'\beta \\ w_2'\beta \\ \vdots \\ w_q'\beta \end{bmatrix}$$

Under H_0:

$$R\hat{\beta} \sim N\left[r, \Sigma_R\right], \text{ where } \Sigma_R = \sigma^2 R(X'X)^{-1}R'$$

$$W = (R\hat{\beta} - r)'\Sigma_R^{-1}(R\hat{\beta} - r) \sim \chi^2(q)$$

To make this testable, we need to replace σ^2 with s^2, so we get $\hat{\Sigma}_R = s^2 R(X'X)^{-1}R'$. This yields a Wald-Type Criterion, and a F variable we can test:

$$\hat{W} = (R\hat{\beta} - r)'\hat{\Sigma}_R^{-1}(R\hat{\beta} - r) = \frac{S(R, \hat{\beta})}{s^2}$$

$$F = \frac{\hat{W}}{q} = \frac{S(R, \hat{\beta})}{qs^2} \sim F(q, T - k)$$

Let F_α be such that $\Pr[F > F_\alpha] = \alpha$. Finally, we reject H_0 at level α when $F > F_\alpha$.

11.3 Likelihood Ratio Test

Another method of testing the hypothesis above, similar to MLE, is finding the β and σ^2 that maximize the probability of the observed sample while satisfying H_0. This is based on the likelihood function L we saw before:

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

Let Ω be the set of unrestricted values for β and σ^2, and ω be the restricted set of those that satisfy H_0.

$$\Omega = \left\{(\beta, \sigma^2) : \beta \in \mathbb{R}, \sigma^2 \in \mathbb{R}_0^+\right\}$$

$$\omega = \left\{\beta, \sigma^2) \in \Omega : R\beta = r\right\}$$

Further, the likelihood of these two is defined:

$$L(\hat{\Omega}) = \max_{\beta, \sigma^2 \in \Omega} L,\ L(\hat{\omega}) = \max_{\beta, \sigma^2 \in \omega} L$$

The likelihood of the unrestricted values is obviously much higher than the restricted ones, so the likelihood ratio $LR(y)$ is a ratio that is always greater than or equal to 1:

$$LR(y) = \frac{L(\hat{\Omega})}{L(\hat{\omega})} \geq 1$$

Let's denote the rejection point for H_0 as λ_α such that $\Pr[LR(y) \geq \lambda_\alpha] = \alpha$. The parameters for $L(\hat{\Omega})$ are simply the maximum likelihood estimators we derived earlier, $\hat{\beta}$ and $\hat{\sigma}^2$. Inputting those yields:

$$L(\hat{\Omega}) = \frac{T^{T/2}e^{-T/2}}{(2\pi)^{T/2}S_\Omega^{T/2}}, \text{ where } S_\Omega = (y - X\hat{\beta})'(y - X\hat{\beta})$$

To find $L(\hat{\omega})$, we must solve a constrained optimization problem using Lagrange multipliers, to maximize the log of L such that $R\beta = r$. First, with respect to β, letting $\tilde{\beta}$ be the likelihood-maximizing parameter, then using that to find $\tilde{\sigma}^2$.

$$\mathcal{L}(\beta, \lambda) = (y - X\beta)'(y - X\beta) - \lambda'(r - R\beta)$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = -2X'y + 2X'X\tilde{\beta} - R'\lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = r - R\tilde{\beta} = 0$$

Solving for $\tilde{\beta}$ substituting back to find $\tilde{\sigma}$ gives

$$L(\hat{\omega}) = \frac{T^{T/2}e^{-T/2}}{(2\pi)^{T/2}S_\omega^{T/2}}, \text{ where } S_\omega = (y - X\tilde{\beta})'(y - X\tilde{\beta})$$

All in all, the critical region of the likelihood ratio test becomes:

$$LR(y) = \left(\frac{S_\omega}{S_\Omega}\right)^{T/2} \geq \lambda_\alpha$$

$$\frac{S_\omega}{S_\Omega} \geq (\lambda_\alpha)^{2/T}$$

It can be shown that $S_\omega - S_\Omega = S(R, \hat{\beta}) = qs^2F$. Therefore, doing some more algebra, the critical region for $H_0 : R\beta = r$ is defined by:

$$F = \frac{(S_\omega - S_\Omega)/q}{S_\Omega/(T-k)} \geq F_\alpha$$

Importantly, under the CLM, the likelihood ratio test yields the same Wald-Type criterion as the F-test, providing an easy method to test the null.

12 Coefficient of Determination (R-Squared)

Suppose we want to characterize how well the explanatory variables explain y. The solution is to calculate the R^2 coefficient.

12.1 Definitions

Let $y = X\beta + \varepsilon$ be a model abiding by the CLM assumptions. We know:

(a) $\hat{y} = X\hat{\beta}$

(b) $\hat{\varepsilon} = y - \hat{y}$

We can further make some definitions which will be useful:

(c) $i = (1, 1, \ldots, 1)'$

(d) $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t = i'y/T$

(e) SST (total sum of squares) $= \sum_{t=1}^{T}(y_t - \bar{y})^2 = (y - i\bar{y})'(y - i\bar{y})$

(f) SSR (regression sum of squares) $= \sum_{t=1}^{T}(\hat{y}_t - \bar{y}_t)^2 = (\hat{y} - i\bar{y})'(\hat{y} - i\bar{y})$

(g) SSE (error sum of squares) $= \sum_{t=1}^{T}(\hat{y}_t - \hat{y}_t)^2 = \hat{\varepsilon}'\hat{\varepsilon}$

Further, we can use these as "variance estimators":

$$\hat{\mathbb{V}}(y) = \frac{SST}{T}$$

$$\hat{\mathbb{V}}(\hat{y}) = \frac{SSR}{T}$$

$$\hat{\mathbb{V}}(\varepsilon) = \frac{SSE}{T}$$

Finally, there is the definition of R^2:

$$R^2 = 1 - \frac{\hat{\mathbb{V}}(\varepsilon)}{\hat{\mathbb{V}}(y)} = 1 - \frac{SSE}{SST}$$

The formula is 1 minus the ratio of unexplained residuals to total, giving a value that shows how effective the fitted values are at explaining y. There is also $R = \sqrt{R^2}$, called the coefficient of multiple correlation.

12.2 Properties of R-Squared

The following properties can be shown with some proof:

(a) $R^2 \leq 1$

(b) $y'y = \hat{y}'\hat{y} + \hat{\varepsilon}'\hat{\varepsilon}$

(c) If one of the regressors is a constant (β_0), $SST = SSR + SSE$ and $\hat{\mathbb{V}}(y) = \hat{\mathbb{V}}(\hat{y}) + \hat{\mathbb{V}}(\varepsilon)$

(d) If one of the regressors is a constant, $R^2 = SSR/SST = \hat{\mathbb{V}}(\hat{y})/\hat{\mathbb{V}}(\varepsilon)$

Empirical Correlation

$$\hat{\rho}(y, \hat{y}) = \frac{\hat{C}(y, \hat{y})}{\sqrt{\hat{\mathbb{V}}(y)\hat{\mathbb{V}}(\hat{y})}}, \text{ where } \hat{C}(y, \hat{y}) = (y - i\bar{y})'(\hat{y} - i\bar{y})$$

(e) If one of the regressors is a constant, $\hat{\rho}(y, \hat{y}) = \sqrt{R^2}$ and is non-negative

12.3 Significance Tests

Consider the model and hypothesis:

$$y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k + \varepsilon$$

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$$

To do an F-test using a restricted and unrestricted sum of squares, we have:

$$F = \frac{S_\omega - S_\Omega/(k-1)}{S_\Omega/(T-k)}, \text{ where } \Omega : y = X\beta + \varepsilon, \text{ and } \omega : y = i\beta_0 + \varepsilon$$

We can show:

$$S_\Omega = (y - X\hat{\beta})'(y - X\hat{\beta}) = SSE$$

$$\hat{\beta}_0 = \bar{y} \text{ (under } \omega)$$

$$S_\omega = (y - i\bar{y})'(y - i\bar{y}) = SST$$

Therefore we can change our F-statistic to:

$$F = \frac{R^2/(k-1)}{(1-R^2)(T-k)} \sim F(k-1, T-k)$$

As R^2 increases, F increases.

12.4 General Relation between R-Squared and F-Tests

The linear hypothesis of the form:

$$H_0 : R\beta = r$$

Where $\text{rank}(R) = q$ can also provide a transformation of the F-statistic. We can compute the restricted and unrestricted R^2 values by:

$$R_0^2 = 1 - \frac{S_\omega}{SST}, R_1^2 = 1 - \frac{S_\Omega}{SST}$$

Therefore, F becomes:

$$F = \left(\frac{T-k}{q}\right)\frac{R_1^2 - R_0^2}{1 - R_1^2}$$

Large values of $R_1^2 - R_0^2$ allow us to reject the null. $q = k - 1$, $S_\omega = SST$, and $R_0^2 = 0$ is a special case of the above detailed in the preceding section.

12.5 Uncentered R-Squared

In models without a constant regressor, R^2 can take negative values and has little meaning. Therefore we define the uncentered coefficient of determination $\tilde{R}^2$.

$$\tilde{R}^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y}$$

It's easy to show $0 \leq \tilde{R}^2 \leq 1$.

12.6 Adjusted R-Squared and Properties

One may be quick to note that the previous definition of R^2 will keep increasing as we add more explanatory variables, regardless of their importance. It has been suggested to "penalize" models by adjusting R^2 by degrees of freedom. We can replace $\hat{\mathbb{V}}(y)$ and $\hat{\mathbb{V}}(\varepsilon)$ by "unbiased estimators":

$$s^2 = \frac{SSE}{T-k}$$

$$s_y^2 = \frac{SST}{T-1}$$

R^2 adjusted by degrees of freedom is defined by:

$$\bar{R}^2 = 1 - \frac{s^2}{s_y^2} = 1 - \frac{T-1}{T-k}\left(\frac{SSE}{SST}\right)$$

We can derive the following properties:

(a) $\bar{R}^2 = R^2 - \frac{k-1}{T-k}(1-R^2)$

(b) $\bar{R}^2 \leq R^2 \leq 1$

(c) $\bar{R}^2 = R^2 \iff k = 1$ or $R^2 = 1$

(d) $\bar{R}^2 \leq 0 \iff R^2 \leq \frac{k-1}{T-1}$

(e) When comparing models on the basis of R^2 or $\bar{R}^2$, the dependent variable must be the same. If so, maximizing R^2 is the same as minimizing the standard error of the regression, s.

13 Dummy Variables

In a regression, it may be useful to describe certain conditions as "binary," either having an effect or having no effect. These are called dummy variables, and their value is either 0 or 1 depending on if the condition is met. For example, a consumption function can be estimated by:

$$C_t = \alpha + \beta Y_t + \varepsilon_t$$

But suppose the constant term is not the same during a normal year and a war year, say α_1 and α_2. Then we can introduce the variable D and change the regression:

$$D_t = \begin{cases} 1, \text{ if } t \text{ is a war year,} \\ 0, \text{ otherwise.} \end{cases}$$

$$C_t = \alpha_1 + (\alpha_2 - \alpha_1)D_t + \beta Y_t + \varepsilon_t$$

13.1 Seasonal Dummy Variables

We can add even more dummy variables to the model to take into account seasonal variation in consumption. The new consumption function is:

$$C_t = \alpha + \beta Y_t + \lambda_1 D_{1t} + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \varepsilon_t$$

Where D_1, D_2, D_3 are 1 when it is their corresponding quarter and 0 otherwise. We could also look at the model

$$C_t = \beta Y_t + \lambda_1 D_{1t} + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \lambda_4 D_{4t} + \varepsilon_t, \text{ where } D_{4t} = 1 - D_{1t} - D_{2t} - D_{3t}$$

However, it important to note we cannot use the model described below, because of multicollinearity in the X matrix. Suppose we were looking at one year, or four quarters worth of data:

$$C_t = \alpha + \beta Y_t + \lambda_1 D_{1t} + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \lambda_4 D_{4t} + \varepsilon_t, t = 1, 2, 3, 4$$

$$X = \begin{bmatrix} 1 & Y_1 & 1 & 0 & 0 & 0 \\ 1 & Y_2 & 0 & 1 & 0 & 0 \\ 1 & Y_3 & 0 & 0 & 1 & 0 \\ 1 & Y_4 & 0 & 0 & 0 & 1 \end{bmatrix}$$

We can see that the column of ones for the constant is exactly a linear combination of the 4 dummy variables, meaning $X'X$ is not invertible.

14 Specification Errors

14.1 Classification of Errors

Recall the 6 assumptions that defined the classical linear model. It is often the case that one or more of these assumptions fails when doing a regression with real data. Some of these failures include:

(a) Incorrect Regression: $y = Z\gamma + \varepsilon$ instead of $y = X\beta + \varepsilon$

(b) Errors do not have mean 0: $\mathbb{E}(\varepsilon) \neq 0$

(c) Incorrect Covariance Matrix: $\mathbb{E}(\varepsilon\varepsilon') = \Omega \neq \sigma^2 I_T$, where Ω is any p.s.d. matrix

(d) Non-normality: ε is not multinormal

(e) Multicollinearity: X is $T \times k$, but $\text{rank}(X) < k$

(f) Stochastic regressors: X is not fixed

14.2 Incorrect Regression Function

Suppose the "true model" for the regression is $y = X\beta + \varepsilon$, which satisfies the CLM assumptions, but instead we estimate the model:

$$y = Z\gamma + \varepsilon$$

where Z is a $T \times G$ fixed matrix. Running least squares on γ, we get:

$$\begin{aligned} \hat{\gamma} &= (Z'Z)^{-1}Z'y \\ &= (Z'Z)^{-1}Z'X\beta + (Z'Z)^{-1}Z'\varepsilon \\ \mathbb{E}(\hat{\gamma}) &= (Z'Z)^{-1}ZX\beta = P\beta \end{aligned}$$

It is clear that for most values of Z and X, the expected value of $\hat{\gamma}$ is not equal to β, and is thus a biased estimator. The confidence intervals and tests we described earlier are no longer valid.

One Missing Explanatory Variable

We can apply a partitioned regression to separate one of the explanatory variables:

$$X = [X_1|x_k], y = X_1\beta_1 + x_k\beta_k + \varepsilon$$

Suppose we omit x_k and estimate the model by:

$$y = X_1\gamma + \varepsilon$$

This corresponds to the case where $Z = X_1$, and $P = [I_{k-1}|\hat{\delta_k}]'$ where $\hat{\delta_k} = (X_1'X_1)^{-1}X_1'x_k$. $\hat{\delta_k}$ is equivalent to the resulting coefficient vector from solving the problem below, regressing x_k on X_1.

$$\min_{\delta_k}(x_k - \delta_k)'(x_k - \delta_k)$$

Therefore, the bias of $\hat{\delta_k}$ can be shown:

$$\mathbb{E}(\hat{\gamma}) = \begin{pmatrix} I_{k-1} & \hat{\delta_k} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_k \end{pmatrix} = \beta_1 + \hat{\delta_k}\beta_k$$

$$\implies \mathbb{E}(\hat{\gamma}) - \beta_1 = \hat{\delta_k}\beta_k$$

There are two cases in which this is an unbiased estimator:

1. x_k was not part of the regression at all ($\beta_k = 0$)
2. x_k is orthogonal to all the other regressors ($X_1'x_k = 0$)

Estimation of Mean from Misspecified Model

Estimation of Error Variance

As we know, the unbiased estimator for error variance, s^2, is equal to the estimated sum of squared errors divided by degrees of freedom. In our misspecified model, the "unbiased" estimator would be:

$$s_Z^2 = \frac{y'M_Zy}{T-G}$$

Substituting $y = X\beta + \varepsilon$, we get:

$$y'M_zy = \beta'X'M_ZX\beta + \varepsilon'M_z\varepsilon + 2\beta'X'M_Z\varepsilon$$

$$\mathbb{E}(y'M_Zy) = \beta'X'M_ZX\beta + \sigma^2(T-G)$$

Therefore, working out the bias of the estimator:

$$\mathbb{E}(s_Z^2) = \sigma^2 + \frac{1}{T-G}\beta'X'M_ZX\beta \geq \sigma^2$$

Comparing the two models, the correct specification never has the largest variance, which justifies choosing the one with the lowest (in other words the lowest $\bar{R}^2$).

14.3 Non-zero Mean Errors

Let's examine the model where $\mathbb{E}(\varepsilon) = \xi \neq 0$. We still assume homoeskedasticity, so the covariance matrix for the errors is:

$$\mathbb{V}(\varepsilon) = \sigma^2 I_T = \mathbb{E}[(\varepsilon - \xi)(\varepsilon - \xi)']$$

Further, in general, $\hat{\beta}$ is not unbiased anymore:

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

This can be easily remedied by having a constant term however, and making the following substitutions. Supposing all the errors have the same non-zero mean:

$$\xi = \mu i$$

$$\nu = \varepsilon - \mu i, \mathbb{E}(\nu) = 0$$

Rewrite the model:

$$y = i\beta_0 + X_1\beta_1 + \ldots + X_k\beta_k + \varepsilon$$

$$y = i(\beta_0 + \mu) + X_1\beta_1 + \ldots + X_k\beta_k + \nu$$

So the new constant and error terms become $\beta_0 + \mu$ and ν respectively, and we are back to business as usual. The assumption of ε having mean 0 is not a restrictive one, but note that this workaround is not possible without a constant term, suggesting that in the vast majority of cases we should have one.

14.4 Incorrect Covariance Matrix

Unlike the errors, the assumption $\mathbb{E}(\varepsilon\varepsilon') = \sigma^2 I_T$ is far more problematic, and not satisfied in real regressions. Suppose that $\mathbb{V}(\varepsilon) = \Omega$ instead. We can show that $\hat{\beta}$ is still unbiased:

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

But the covariance matrix of $\hat{\beta}$ is altered, with grave implications:

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

This suggests that $\hat{\beta}$ is no longer BLUE for β, and the usual formulas for confidence intervals and tests no longer apply. Consequently, dealing with non-homoskedasticity is of great importance, which will be mainly achieved through generalized least squares.

14.5 Stochastic Regressors

Assuming that X is a fixed matrix is also implausible in many regressions. If ε is independent from X, however, most of our results still hold. We must treat the mean and variance of the errors as thought they are conditional on X. In the independent case:

$$\mathbb{E}(\varepsilon|X) = \mathbb{E}(\varepsilon) = 0, \text{ and } \mathbb{V}(\varepsilon|X) = \mathbb{V}(\varepsilon) = \sigma^2 I_T$$

Then $\hat{\beta}$ is still unbiased, and usual tests and confidence intervals (assuming Gaussianity) hold. However, if $\mathbb{E}(\varepsilon|X) \neq 0$, then we're out of luck, as not much can be said about the bias of $\hat{\beta}$, and tests and confidence intervals fail.

15 Analysis of Residuals

After the model has been estimated, it is usually valuable to examine its residuals, $\hat{\varepsilon}_i$, $i = 1, 2, \ldots, T$, which are an estimator of ε_i. Based on our assumptions, the residuals $\hat{\varepsilon}_i$ should be i.i.d. random variables

15.1 Graphical Analysis

By graphing the residuals, we may be able to discern:

(a) "very large" residuals

(b) relationships between residuals and certain variables

(c) heteroskedasticity

(d) autocorrelation

15.2 Standardization and Properties

Recall the following facts about residuals:

$$\varepsilon \sim N[0, \sigma^2 I_T]$$

$$\hat{\varepsilon} = y - X\hat{\beta} = M\varepsilon$$

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

$$\mathbb{E}(\hat{\varepsilon}) = 0, \mathbb{V}(\hat{\varepsilon}) = \sigma^2 M_X$$

Note that each $\hat{\varepsilon}_i$ does not have the same variance and are not independent. Examining their variances more closely reveals the following relationship:

$$\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2(1 - p_i) \leq \sigma^2, p_i = X_i'(X'X)^{-1}X_i$$

$$C(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \sigma^2(-p_{ij}), p_{ij} = X_i'(X'X)^{-1}X_j$$

15.3 Standardized and Studentized Residuals

Internally Studentized Residuals

To get residuals with the same variance, we can do a simple transformation:

$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{(1 - p_i)^{1/2}}$$

$$\mathbb{V}(\tilde{\varepsilon}_i) = \sigma^2$$

And to make them more interpretable, we divide by s:

$$s = \left[\frac{\hat{\varepsilon}'\hat{\varepsilon}}{T - k}\right]^{1/2}$$

$$r_i = \frac{\tilde{\varepsilon}_i}{s} = \frac{\hat{\varepsilon}_i}{s(1 - p_i)^{1/2}}$$

These are known as "internally studentized residuals." r_i does not follow a t-distribution, since the residuals are not independent of each other. It is hard to determine whether or not certain residuals are "large."

Externally Studentized Residuals

Consider the model estimated for each variable without the i-th entry:

$$y_{(i)} = \begin{bmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_T \end{bmatrix}, X_{(i)} = \begin{bmatrix} X_1 \\ \vdots \\ X_{i-1} \\ X_{i+1} \\ \vdots \\ X_T \end{bmatrix}$$

Running least squares on this, we get:

$$\hat{\beta}_{(i)} = \left[X_{(i)}'X_{(i)}\right]^{-1} X_{(i)}'y_{(i)}$$

$$\varepsilon_{(i)} = y_{(i)} - X_{(i)}\hat{\beta}_{(i)}, s^2_{(i)} = \frac{\varepsilon_{(i)}'\varepsilon_{(i)}}{T - k - 1}$$

$$d_i = X_i'\left[X_{(i)}'X_{(i)}\right]X_i, v_i = y_i - X_i'\hat{\beta}_{(i)}$$

With our new residuals, we can see:

$$\mathbb{V}(v_i) = \sigma^2(1 + d_i)$$

$$t_i = \frac{v_i}{s_{(i)}(1 + d_i)^{1/2}} \sim t(T - k - 1)$$

Doing some algebra, there is also a convenient transformation between internally and externally studentized residuals:

$$t = (T - k - 1)^{1/2} \frac{r_i}{(T - k - r_i^2)^{1/2}}$$

So, to compute whether a given residual $\hat{\varepsilon}_i$ is large, we need only to internally studentize them, apply the transformation, then see if $|t_i| \geq t_{\alpha/2}(T - k - 1)$ for a given level α. This test only works for one residual at a time however.

15.4 Test for an Outlier

Declaring a residuals "outliers" as opposed to "large" comes with added difficulty: for a level α, we may say $\hat{\varepsilon}_i$ is an outlier if:

$$|t_i| \geq t_{\alpha/2}(T - k - 1)$$

However, if we were to run this test T times, the probability of at least one being an "outlier" is larger than α even if there truly are none, stemming from the fact that:

$$\Pr\left(\bigcup_{i=1}^{T} A_i\right) \geq \Pr(A_i)$$

Thus, we adopt the following rules:

$$\max_{1 \leq i \leq T} |t_i| \geq c_\alpha$$

$$\max_{1 \leq i \leq T} |t_i'| \geq c_\alpha^2$$

The observations that are declared outliers are:

$$|t_i| \geq c_\alpha \text{ or } t_i^2 \geq c_\alpha^2$$

Using the Boole-Bonferroni inequality, we can show the following the distribution of $\max |t_i|$, so that when we declare an observation outlying, the following holds:

$$\max t_i^2 \geq F_{\alpha/T}(1, T - k - 1)$$

$$\max |t_i| \geq t_{\alpha/2T}(1, T - k - 1)$$

$$c_\alpha^2 \leq F_{\alpha/T}(1, T - k - 1) = \left[t_{\alpha/2T}(1, T - k - 1)\right]^2$$

15.5 Tests for Heteroskedasticity

Given the model:

$$y_t = x_t'\beta + \varepsilon_t \text{ , such that } \sigma_t^2 = \mathbb{V}(\varepsilon_t) = \mathbb{E}(\varepsilon_t^2)$$

Showing heteroskedasticity entails rejecting a null hypothesis of homoskedasticity:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_T^2 = \sigma^2$$

Further suppose we have reason to believe the variance increases with time: $\mathbb{V}(\varepsilon_t) > \mathbb{V}(\varepsilon_{t-1})$ This could be informally checked by plotting residuals against time, but a more rigorous method involves dividing the sample into two even parts:

$$\underbrace{t = 1, 2, \ldots, T_1}_{T_1 \text{ observations}}, \quad \underbrace{t = T_1 + 1, \ldots, T}_{T_2 \text{ observations}}$$

Under the hypothesis of increasing variance, we know:

$$\mathbb{E}\left[\frac{1}{T_1}\sum_{t=1}^{T_1}\varepsilon_t^2\right] < \mathbb{E}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T}\varepsilon_t^2\right]$$

We get an F distributed test statistic and can perform the following tests:

$$F = \frac{T_2}{T_1}\frac{\sum_{t=1}^{T_1}\varepsilon_t^2}{\sum_{t=T_1+1}^{T}} \sim F(T_2, T_1)$$

(a) One-sided Tests

- Against σ_t^2 increasing, reject when $F > F_\alpha(T_2, T_1)$
- Against σ_t^2 decreasing, reject when $F > F_{1-\alpha}(T_\alpha, T_1)$

(b) Two-sided Test

- Reject when $F > F_{\alpha/2}(T_2, T_1)$ or $F < F_{1-\alpha/2}(T_2, T_1)$

Since the ε_i are unknown, it is tempting to replace them with the $\hat{\varepsilon}_i$. However, these are not independent, so we don't get a valid test statistic. The solution, suggested by Goldfeld and Quandt, lies in again splitting the model into two groups but estimating errors via least squares separately and then testing.

The Goldfeld-Quandt Solution

$$\underset{T_1\times 1}{y_A} = X_A\beta + \varepsilon_A \ , \ \hat{\varepsilon}_A = y_A - X_A\hat{\beta}_A$$

$$\underset{T_2\times 1}{y_B} = X_B\beta + \varepsilon_B \ , \ \hat{\varepsilon}_B = y_B - X_B\hat{\beta}_B$$

$$F = \frac{\hat{\varepsilon}_B'\hat{\varepsilon}_B/(T_2-k)}{\hat{\varepsilon}_A'\hat{\varepsilon}_A/(T_1-k)} \sim F(T_2-k, T_1-k)$$

The same one-sided and two-sided tests from above apply.