3. OLS Part III

In this section we derive some finite-sample properties of the OLS estimator.

3.1 The Sampling Distribution of the OLS Estimator

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim N[0\,, \sigma^2 I_n]$$

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = f(\boldsymbol{y})$$

$\boldsymbol{\varepsilon}$ is random $\longrightarrow$ $\boldsymbol{y}$ is random $\longrightarrow$ $\boldsymbol{b}$ is random

- ***b*** is an *estimator* of ***β*.** It is a function of the *random* sample data.
- ***b*** is a “statistic”.
- ***b*** has a probability distribution – called its *Sampling Distribution*.
- Interpretation of *sampling distribution* –

  Repeatedly draw all possible samples of size *n*.

  Calculate values of ***b*** each time.

  Construct relative frequency distribution for the ***b*** values and probability of occurrence.

  It is a *hypothetical* construct. Why?

- Sampling distribution offers *one* basis for answering the question:

“How good is *b* as an estimator of *β* ?”

Note:

Quality of estimator is being assessed in terms of performance in *repeated samples*. Tells us nothing about quality of estimator for *one particular sample*.

- Let's explore some of the properties of the LS estimator, $\boldsymbol{b}$, and build up its sampling distribution.
- Introduce some general results, and apply them to our problem.

Definition: An estimator, $\hat{\boldsymbol{\theta}}$ is an *unbiased* estimator of the parameter vector, $\boldsymbol{\theta}$, if $E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$.

That is, $E[\hat{\boldsymbol{\theta}}(\boldsymbol{y})] = \boldsymbol{\theta}$.

That is, $\int \hat{\theta}(\boldsymbol{y})p(\boldsymbol{y} \mid \boldsymbol{\theta})d\boldsymbol{y} = \boldsymbol{\theta}$.

The quantity, $\boldsymbol{B}(\boldsymbol{\theta}, \boldsymbol{y}) = E[\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}]$, is called the "Bias" of $\hat{\boldsymbol{\theta}}$.

Example: $\{y_1, y_2, \ldots \ldots, y_n\}$ is a random sample from population with a finite mean, μ, and a finite variance, σ^2 .

Consider the *statistic* $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Then, $E[\bar{y}] = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(y_i)$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \left(\frac{1}{n} n\mu\right) = \mu \ .$$

So, $\bar{y}$ is an *unbiased estimator* of the parameter, μ.

- Here, there are lots of possible unbiased estimators of μ.
- So, need to consider additional characteristics of estimators to help choose.

Return to our LS problem –

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y}$$

- Recall – either assume that X is *non-random*, or condition on X.
- We'll assume X is non-random – get same result if we condition on X.

Then: $E(\boldsymbol{b}) = E[(X'X)^{-1}X'\boldsymbol{y}] = (X'X)^{-1}X'E(\boldsymbol{y})$

So,

$$E(\boldsymbol{b}) = (X'X)^{-1}X'E[X\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = (X'X)^{-1}X'[X\boldsymbol{\beta} + E(\boldsymbol{\varepsilon})]$$

$$= (X'X)^{-1}X'[X\boldsymbol{\beta} + \boldsymbol{0}] = (X'X)^{-1}X'X\boldsymbol{\beta}$$

$$= \boldsymbol{\beta} \ .$$

The LS estimator of β is Unbiased

Definition: Any estimator that is a *linear function* of the random sample data is called a *Linear Estimator*.

Example: $\{y_1, y_2, \ldots \ldots, y_n\}$ is a random sample from population with a finite mean, μ, and a finite variance, σ^2 .

Consider the *statistic* $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}[y_1 + y_2 + \cdots + y_n]$.

This statistic is a *linear estimator* of μ.

(Note that the "weights" are non-random.)

Return to our LS problem –

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} = A\boldsymbol{y}$$

$(k\times1)$ $(k\times n)(n\times1)$

Note that, under our assumptions, A is a *non-random* matrix.

So,

$$\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \cdots & \vdots \\ a_{k1} & \cdots & a_{kn} \end{bmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} .$$

For example, $b_1 = [a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n]$; *etc.*

The LS estimator, $\boldsymbol{b}$, is a linear (& unbiased) estimator of $\boldsymbol{\beta}$

Now let's consider the dispersion (variability) of $\boldsymbol{b}$, as an estimator of $\boldsymbol{\beta}$.

Definition: Suppose we have an $(n\times 1)$ random vector, $\boldsymbol{x}$. Then the *Covariance Matrix* of $\boldsymbol{x}$ is defined as the $(n\times n)$ matrix:

$$V(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))'].$$

- Diagonal elements of $V(\boldsymbol{x})$ are $var.(x_1), \ldots\ldots, var.(x_n)$.
- Off-diagonal elements are $covar.(x_i, x_j)$; $i, j = 1, \ldots, n$; $i \neq j$.

Return to our LS problem –

We have a $(k\times 1)$ random vector, $\boldsymbol{b}$, and we know that $E(\boldsymbol{b}) = \boldsymbol{\beta}$.

$$V(\boldsymbol{b}) = E[(\boldsymbol{b} - E(\boldsymbol{b}))(\boldsymbol{b} - E(\boldsymbol{b}))']$$

Now,

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} = (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon})$$

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

$$= I\boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon}.$$

So,

$$(\boldsymbol{b} - \boldsymbol{\beta}) = (X'X)^{-1}X'\boldsymbol{\varepsilon}\ . \qquad \textbf{[*]}$$

Using the result, **[*]**, in $V(\boldsymbol{b})$, we have:

$$V(\boldsymbol{b}) = E\{[(X'X)^{-1}X'\boldsymbol{\varepsilon}][(X'X)^{-1}X'\boldsymbol{\varepsilon}]'\}$$

$$= (X'X)^{-1}X'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']X(X'X)^{-1}\ .$$

We showed, earlier, that because $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $V(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 I_n$.

(*What other assumptions did we use to get this result?*)

So, we have:

$$V(\boldsymbol{b}) = (X'X)^{-1}X'E[\boldsymbol{\varepsilon\varepsilon}']X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} = \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}.$$

$$\boxed{V(\boldsymbol{b}) = \sigma^2(X'X)^{-1} \quad (k\times k)}$$

Interpret diagonal and off-diagonal elements of this matrix.

Finally, because the error term, $\boldsymbol{\varepsilon}$ is assumed to be Normally distributed,

1. $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$: this implies that $\boldsymbol{y}$ is also Normally distributed. (Why?)
2. $\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} = A\boldsymbol{y}$: this implies that $\boldsymbol{b}$ is also Normally distributed.

So, we now have the full **Sampling Distribution** of the LS estimator, $\boldsymbol{b}$:

$$\boxed{\boldsymbol{b} \sim N[\boldsymbol{\beta}\,, \sigma^2(X'X)^{-1}]}$$

Note:

- This result depends on our various, *rigid*, assumptions about the various components of the regression model.
- The Normal distribution here is a "*multivariate* Normal" distribution.
  (*See handout on "Spherical Distributions"*.)
- As with estimation of population mean, $\boldsymbol{\mu}$, in previous example, there are lots of other *unbiased* estimators of $\boldsymbol{\beta}$ in the model $= X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- How might we choose between these possibilities? Is *linearity* desirable?
- We need to consider other *desirable* properties that these unbiased estimators may have.
- *One option* is to take account of estimators' *precisions*.

3.2 The Efficiency of OLS

Definition: Suppose we have two *unbiased* estimators, $\widehat{\theta_1}$ and $\widehat{\theta_2}$, of the (scalar) parameter, θ. Then we say that $\widehat{\theta_1}$ is **at least as efficient** as $\widehat{\theta_2}$ if $var.\left(\widehat{\theta_1}\right) \leq var.\left(\widehat{\theta_2}\right)$.

Note:

1. The variance of an estimator is just the variance of its sampling distribution.
2. "Efficiency" is a *relative* concept.
3. What if there are 3 or more unbiased estimators being compared?

- What if one or more of the estimators being compared is *biased* ?
- In this case we can take account of both variance, and any bias, at the same time by using "*mean squared error*" (MSE) of the estimators.

Definition: Suppose we have two *unbiased* estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\hat{\theta}_1$ is **at least as efficient** as $\hat{\theta}_2$ if $\Delta = V\left(\hat{\theta}_2\right) - V(\hat{\theta}_1)$ is *at least positive semi-definite*.

Taking account of its *linearity*, *unbiasedness*, and its *precision*, in what sense is the LS estimator, ***b***, of β *optimal*?

Theorem (Gauss-Markhov):

In the "standard" linear regression model, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the LS estimator, ***b***, of $\boldsymbol{\beta}$ is **Best Linear Unbiased** (BLU). That is, it is **Efficient** in the class of all linear and unbiased estimators of β.

1. Is this an *interesting* result?
2. What *assumptions* about the "standard" model are we going to exploit?

Proof

Let $\boldsymbol{b_0}$ be any other *linear* estimator of $\boldsymbol{\beta}$:

$$\boldsymbol{b_0} = C\boldsymbol{y} \qquad ; \qquad \text{for } \textit{some} \text{ non-random } C .$$

$$(k\times 1) \quad (k\times n)(n\times 1)$$

Now, $V(\boldsymbol{b_0}) = CV(\boldsymbol{y})C' = C(\sigma^2 I_n)C' = \sigma^2 CC'$

$$(k\times k)$$

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

so that $D\boldsymbol{y} = C\boldsymbol{y} - (X'X)^{-1}X'\boldsymbol{y} = \boldsymbol{b_0} - \boldsymbol{b}$.

Now restrict $\boldsymbol{b_0}$ to be *unbiased*, so that $E(\boldsymbol{b_0}) = E(C\boldsymbol{y}) = CX\boldsymbol{\beta} = \boldsymbol{\beta}$.

This requires that $CX = I$, which in turn implies that

$$DX = [C - (X'X)^{-1}X']X = CX - I = 0 \qquad (and\ D'X' = 0)$$

(*What assumptions have we used so far?*)

Now, focus on covariance matrix of $\boldsymbol{b_0}$:

$$V(\boldsymbol{b_0}) = \sigma^2[D + (X'X)^{-1}X'][D + (X'X)^{-1}X']'$$

$$= \sigma^2[DD' + (X'X)^{-1}X'X(X'X)^{-1}] \qquad ; \qquad DX = 0$$

$$= \sigma^2 DD' + \sigma^2(X'X)^{-1}$$

$$= \sigma^2 DD' + V(\boldsymbol{b}),$$

or, $[V(\boldsymbol{b_0}) - V(\boldsymbol{b})] = \sigma^2 DD' \qquad ; \qquad \sigma^2 > 0$

Now we just have to "sign" this (matrix) difference:

$$\boldsymbol{\eta}'(DD')\boldsymbol{\eta} = (D'\boldsymbol{\eta})'(D'\boldsymbol{\eta}) = v'v = \sum_{i=1}^{n} v_i^2 \geq 0 .$$

So, $\Delta = [V(\boldsymbol{b_0}) - V(\boldsymbol{b})]$ is a p.s.d. matrix, implying that $\boldsymbol{b_0}$ is *relatively less efficient* than $\boldsymbol{b}$.

Result:

The LS estimator is the Best Linear Unbiased estimator of $\boldsymbol{\beta}$.

- What assumptions did we use, and where?
- Were there any standard assumptions that we *didn't* use?
- What does this suggest?