Topic 5: Statistical Properties of the OLS Estimator

Statistical Properties of the OLS Estimator

1 Introduction
2 Unbiased estimators
3 The law of iterated expectations
4 Unbiasedness of the OLS estimator
5 Variance-Covariance matrices
6 The Variance-Covariance matrix of the OLS estimator
7 The Gauss-Markov Theorem
8 Appendix 1: Proof of the unbiasedness of the OLS estimator
9 Appendix 2: Derivation of the Variance-Covariance matrix for a 2x1 random vector

- In econometrics we wish to use the information in our sample to make inferences about the underlying **population** from which the **sample** was drawn.
- For example, in the wage-education example, we wish to use the data in our sample to make inferences about the relationship between wages and education in the general population, not just in our sample.
- The conceptual framework which enables us to use the sample data to make inferences about the population is to treat the collection of the sample as a random experiment, and both the dependent variable and each of the explanatory variables as random variables from an underlying population.

- For example, in the bivariate linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i, \;\; i = 1, 2, ...n,$$

we treat the sample

$$(y_{i,} x_i), i = 1, 2, ..., n,$$

as just one of many samples we might have drawn from the underlying population.

- Mathematically, this entails treating the variables y_i and x_i , $i = 1, 2, ..., n$, as random variables whose values are unknown before we conduct the random experiment of collecting the sample.

- A very important implication of viewing y_i and x_i , $i = 1, 2, ..., n$, as random variables is that the OLS estimators of β_0 and β_1 in the linear regression equation

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i, \;\; i = 1, 2, ...n,$$

  are also random variables.

- Recall that the OLS estimators of β_0 and β_1 are given by

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{1}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}. \tag{2}$$

- It is clear from (1) and (2) respectively that $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are functions of the random variables y_i and x_i, $i = 1, 2, ..., n$,

- This means that both $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are also random variables (since they depend on the random variables y_i and x_i, $i = 1, 2, ..., n$) with associated pdfs.
- This reasoning extends to the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i, \;\; i = 1, 2, ...n, \qquad (3)$$

in which each of the OLS estimators $(\widehat{\beta}_0, \widehat{\beta}_{1,} \widehat{\beta}_2, ..., \widehat{\beta}_k)$ is a random variable whose value varies from sample to sample.

- We can express this idea more compactly by stating that the vector

$$\underset{(k+1x1)}{\widehat{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ . \\ . \\ \widehat{\beta}_k \end{bmatrix}$$

is a **random vector** (each element is a random variable).

- In this topic we will explore the **statistical properties** of the random variables $(\widehat{\beta}_0, \widehat{\beta}_{1,} \widehat{\beta}_2, ..., \widehat{\beta}_k)$, or equivalently, the statistical properties of the random vector $\widehat{\beta}$.
- By the statistical properties of $\widehat{\beta}_j$, $j = 0, 1, ..., k$, we mean the properties of its pdf, in particular the mean and variance of the distribution.

- The first property of an estimator that we wish to consider is the property of **unbiasedness**.
- Assume that we wish to estimate the value of an unknown parameter, θ.
- Let $\widehat{\theta}$ be an estimator of θ.
- We say that $\widehat{\theta}$ is an **unbiased estimator** of θ if

$$E(\widehat{\theta}) = \theta.$$

That is, $\widehat{\theta}$ is an unbiased estimator of θ if the mean of $\widehat{\theta}$ is equal to θ.

- For example, let $(y_1, y_2, ..., y_n)$ be a set of random variables with the property that

$$E(y_i) = \mu, i = 1, 2, ..., n.$$

  That is, each of the random variables $(y_1, y_2, ..., y_n)$ has the same **population mean** μ.

- By definition, the **sample mean** is

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- Therefore, using the properties of the expectation operator discussed in Topic 2,

$$
\begin{aligned}
E(\overline{y}) &= E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n}(n\mu) \\
&= \mu.
\end{aligned}
$$

- Since
$$E(\overline{y}) = \mu.$$
it follows that $\overline{y}$ is an unbiased estimator of μ.
- Intuitively,
$$E(\overline{y}) = \mu$$
means that if we were to collect a very large number of samples of size n, $(y_1, y_2, ..., y_n)$, and for each sample compute the sample mean, $\overline{y}$, the average value of $\overline{y}$ across all these sample would be equal to μ.
- In this sense, an unbiased estimator of μ gives us a very accurate estimate "on average".
- Of course, even if $\overline{y}$ is an unbiased estimator of μ, if we were to take an very large number of samples $(y_1, y_2, ..., y_n)$ from the population and for each sample compute $\overline{y}$, for some samples $\overline{y}$ would overestimate μ and for some samples $\overline{y}$ would underestimate μ.

- However, because $\overline{y}$ is an unbiased estimator of μ, the average value of $\overline{y}$ over all the samples would be equal to μ.
- It turns out that, **under certain assumptions**, the OLS estimator of β, $\widehat{\beta}$, in the linear regression model

$$y = X\beta + u \tag{4}$$

is an unbiased estimator of β. That is,

$$E(\widehat{\beta}) = \beta.$$

- In order to prove this, we need to introduce a very useful mathematical tool called the **Law of Iterated Expectations** (LIE).

- Let X and Y denote two random variables.
- Let $E(Y)$ denote the unconditional mean of Y and let $E(Y|X)$ denote the mean of Y conditional on X assuming a particular value.
- The LIE states that

$$E(Y) = E_X[E(Y|X)]. \tag{5}$$

- Recall from Topic 2 that $E(Y|X)$ is a function of X and its value changes as the value of X changes.
- Equation (5) states that "the unconditional mean of Y is equal to the mean (with respect to X) of the conditional mean of Y".

- The LIE is usually written as

$$E(Y) = E[E(Y|X)].$$

  We have written it in the form of (5) to emphasize that while the "inside expectation" is being taken with respect to the random variable Y, the "outside expectation" is being taken with respect to the random variable X.

- The LIE states that the unconditional mean of Y can be computed in two steps:

  S1: Compute $E(Y|X)$. When we view $E(Y|X)$ as a function of X, it is a random variable.

  S2: Compute the mean of the random variable $E(Y|X)$ over all possible values of X.

- Note that an immediate implication of

$$E(Y) = E_X[E(Y|X)] \tag{5}$$

is that if

$$E(Y|X) = 0$$

then

$$E(Y) = 0,$$

since

$$E_X[0] = 0.$$

- The LIE can be quite confusing when one first encounters it, so let's look at a simple example of the LIE in action.

- Recall the example in Topic 2 in which the random variable Y denotes the number bathrooms and the random variable X denotes the number of bedrooms in a randomly selected apartment in Melbourne.
- In Topic 2 we derived the following pdfs:

Table 1: $f_X(x)$	
X	$P(X = x)$
1	0.40
2	0.40
3	0.20

,

Table 2: $f_Y(y)$	
Y	$P(Y = y)$
1	0.68
2	0.32

,

Table 3:	
X	$E(Y\|X = x)$
1	1.00
2	1.40
3	1.80

.

  - Table 1 is the marginal pdf of X.
  - Table 2 is the marginal pdf of Y.
  - Table 3 is the conditional pdf of Y.

- From Table 2

$$E(Y) = 1(0.68) + 2(0.32) = 1.32. \tag{6}$$

- We next show how we can arrive at the same result using the LIE. That is, using the fact that

$$E(Y) = E_X[E(Y|X)].$$

S1: Derive $E(Y|X)$. This is given in Table 3.

S2: Compute

$$E_X[E(Y|X)].$$

To do this, think of $E(Y|X)$ as a random variable which can take on the values 1, 1.4 and 1.8. Then, from Table 3,

$$\begin{aligned} E_X[E(Y|X)] &= E(Y|X=1)P(X=1) + E(Y|X=2)P(X=2) \\ +E(Y|X &= 3)P(X=3). \end{aligned} \tag{7}$$

Substituting the information in Table 3 and Table 1 into (7) we obtain

$$\begin{aligned} E_X[E(Y|X)] &= 1(0.4) + 1.4(0.40) + 1.8(0.20) \\ &= 0.40 + 0.56 + 0.36 \\ &= 1.32. \end{aligned} \tag{8}$$

Comparing

$$E(Y) = 1(0.68) + 2(0.32) = 1.32 \tag{6}$$

and (8) we see that

$$E(Y) = E_X[E(Y|X)] = 1.32.$$

- The LIE is a very useful tool in situations in which it is difficult to compute $E(Y)$ directly, but it is easy to compute $E_X[E(Y|X)]$ (this happens whenever $E(Y|X)$ has a simple form).

- Under the following assumptions, it can be shown that the OLS estimator of β,

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

in the linear regression model

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

is an unbiased estimator of β. That is,

$$E(\widehat{\beta}) = \beta.$$

A1: **The model is linear in the parameters.** That is,

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \\ &= X\beta + u. \end{aligned} \tag{9}$$

A2: **The columns of X are linearly independent.** That is, no column of the X matrix can be written as an exact linear function of one or more of the remaining columns.

A3: **The the "zero conditional mean" assumption holds.** That is,

$$E(u|X) = 0.$$

- For ease of reference we state this important result as a theorem.

Theorem (1)

When assumptions A1,A2 and A3 hold

$$E(\widehat{\beta}) = \beta,$$

where

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

is the OLS estimator of β in the linear regression model

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

Proof: See Appendix 1.

- Theorem 1 states that when A1-A3 hold, $\widehat{\beta}$ is an unbiased estimator of β. That is,

$$\begin{bmatrix} E(\widehat{\beta}_0) = \beta_0 \\ E(\widehat{\beta}_1) = \beta_1 \\ . \\ . \\ E(\widehat{\beta}_k) = \beta_k \end{bmatrix}. \tag{10}$$

- Let's consider whether or not A1, A2 and A3 are reasonable.

A1 **The model is linear in the parameters:**

- This assumption is not as restrictive as it may first seem. Since A1 only requires that the model is linear in the parameters, it permits nonlinear transformations of both the dependent variable and the regressors. As we will see later in the unit, models such as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + u_i, \;\; i = 1, 2, ...n$$

and

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_{i1}) + u_i, \;\; i = 1, 2, ...n$$

satisfy A1, since they are linear in the parameters.

A2 **The columns of X are linearly independent** (no perfect multicollinearity):

- When this assumption is violated, the X matrix is said to display **perfect multicollinearity,** in which case $(X'X)$ is a singular matrix (the matrix $(X'X)^{-1}$ does not exist) and

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

is not defined.

- The following is an example of an X matrix whose columns are perfectly collinear.

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}. \tag{11}$$

Let c_j denote the jth column of X.Clearly, in the matrix given by (11)

$$c_1 = c_2 + c_3.$$

Since there is an exact linear relationship between the columns of X, the matrix X displays perfect multicollinearity.

- Perfect multicollinearity rarely occurs by chance. When it does occur, it is usually the result of an error on the part of the researcher when attempting to include in the regression equation a special type of regressor, called a **dummy variable**.
- A dummy variable is a regressor which is used to incorporate qualitative information, such as information about gender or race, into the regression model.
- We will discuss dummy variables in detail later in the unit.

A3 **The zero conditional mean assumption:**

- To get an intuitive understanding of what this assumption entails, consider the bivariate linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i, i = 1, 2, ...n. \tag{12}$$

In the context of this model A3 requires that

$$E(u_i | x_1, x_2, ...x_i, ..., x_n) = 0, i = 1, 2, ...n. \tag{13}$$

An implication of (13), which is easier to interpret than (13) itself, is that

$$corr(u_i, x_1) = corr(u_i, x_2) = ...corr(u_i, x_i)... = corr(u_i, x_n) = 0. \tag{14}$$

- Equation (14) states that the error term associated with observation i is uncorrelated with each of the random variables $(x_1, x_2, ...x_i.., x_n)$, and that this must hold for every u_i.
- The condition in (14) is often plausible for cross-sectional data but, as we will see later in the unit, it is seldom holds for time series data.
- In summary, the assumptions required for the OLS estimator to be unbiased are often reasonable for cross-sectional data, but A3 rarely holds for time series data.

- Consider the 3x1 vector of random variables given by

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}. \tag{15}$$

- Given the vector z, we define the 3x3 $Var - Covariance$ matrix of z (which is usually abbreviated to $Var(z)$) as

$$\underset{(3x3)}{Var(z)} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}, \tag{16}$$

where

$$\begin{aligned} \sigma_i^2 &= Var(z_i), i = 1, 2, 3, \\ \sigma_{ij} &= Cov(z_i, z_j). \end{aligned} \tag{17}$$

- See Appendix 2 for the derivation of $Var(z)$ for the case in which z is a 2x1 vector of random variables.
- From (16) it is evident that:
  - The value of each element on the principal diagonal of $Var(z)$ is the variance of one of the elements of z.
  - The value of each off-diagonal element is the covariance between two of the elements of z.

- Since, by definition,

$$Cov(z_i, z_j) = Cov(z_j, z_i),$$

  it follows that

$$Var(z) = [Var(z)]'.$$

  That is, $Var(z)$ is a **symmetric** matrix.

- Note that when z is a 3x1 vector, $Var(z)$ is a 3x3 matrix.

- By extension, in the general case in which z is an (nx1) vector of random variables given by

$$z = \left[\begin{array}{ccccc} z_1 & z_2 & . & . & z_n \end{array} \right]',$$

the variance-covariance matrix of z is given by

$$\underset{(nxn)}{Var(z)} = \left[\begin{array}{ccccc} \sigma_1^2 & \sigma_{12} & . & . & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & . & . & \sigma_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ \sigma_{n1} & & & & \sigma_n^2 \end{array} \right]. \tag{18}$$

- There are two important special cases of (18) which sometimes arise.

C1 In the case in which $(z_1, z_2, ..., z_n)$ are **uncorrelated random variables,** all the off-diagonal elements in (18) are equal to zero and (18) reduces to

$$\underset{(nxn)}{Var(z)} = \begin{bmatrix} \sigma_1^2 & 0 & . & . & 0 \\ 0 & \sigma_2^2 & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & \sigma_n^2 \end{bmatrix}, \tag{19}$$

which we can write more compactly as

$$\underset{(nxn)}{Var(z)} = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2).$$

C2 If, in addition to being uncorrelated random variables, $(z_1, z_2, ..., z_n)$ also have the same variance σ^2, (19) reduces to

$$\underset{(nxn)}{Var(z)} = \begin{bmatrix} \sigma^2 & 0 & . & . & 0 \\ 0 & \sigma^2 & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & . & . & 0 \\ 0 & 1 & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & 1 \end{bmatrix} = \sigma^2 I_n. \tag{20}$$

When (20) holds, we say that $Var(z)$ is a **scalar-identity matrix** (since it is an identity matrix I_n multiplied by the scalar σ^2).

- Writing

$$Var(z) = \sigma^2 I_n$$

is a compact way of stating that the random variables $(z_1, z_2, ..., z_n)$ are both **homoskedastic** (have a common variance,σ^2) and pair-wise uncorrelated.

- The analysis in the preceding section can be applied to any vector of random variables, including the vector $(k+1)x1$ vector

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

which is the OLS estimator of β in the linear regression model

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

- For example, when

$$k = 2$$

the linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, \;\; i = 1, 2, \ldots n,$$

and the OLS estimator of the 3x1 vector β is given by

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix}.$$

The statistical properties of the OLS estimator III

6. The Variance-Covariance matrix of the OLS estimator

- Since $\widehat{\beta}$ is a 3x1 vector of random variables, using the results in section 5 above we may write the $Var-Covariance$ matrix of $\widehat{\beta}$ as

$$\underset{(3x3)}{Var(\widehat{\beta})} = \begin{bmatrix} Var(\widehat{\beta}_0) & Cov(\widehat{\beta}_0, \widehat{\beta}_1) & Cov(\widehat{\beta}_0, \widehat{\beta}_2) \\ Cov(\widehat{\beta}_1, \widehat{\beta}_0) & Var(\widehat{\beta}_1) & Cov(\widehat{\beta}_1, \widehat{\beta}_2) \\ Cov(\widehat{\beta}_2, \widehat{\beta}_0) & Cov(\widehat{\beta}_2, \widehat{\beta}_1) & Var(\widehat{\beta}_2) \end{bmatrix}. \quad (21)$$

- The diagonal elements of (21) are of particular interest, since they determine the precision of the OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\beta}_2$.
- Note that the elements of $Var(\widehat{\beta})$ in (21) denote the **true** variances and covariances of $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\beta}_2$. They are not estimates of these variances and covariances.

- As we will see later in the unit, in order to test hypotheses involving restrictions on β_0, β_1 and β_2, and to construct confidence intervals for β_0, β_1 and β_2, we need to estimate the elements in the matrix given by (21).
- In the general linear regression model given by

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

there are $k + 1$ regressors and $Var(\widehat{\beta})$ is a (k+1)x(k+1) matrix.
- Using the properties of the variance discussed in Topic 2, it follows that

$$\begin{aligned} Var(y|X) &= Var(X\beta + u|X) \\ &= Var(u|X) \end{aligned}$$

since, when we condition on X, the vector $X\beta$ is a vector of constants.

- Note that, because

$$Var(y|X) = Var(u|X), \tag{22}$$

any assumptions we make about $Var(y|X)$ are equally applicable to $Var(u|X)$ and vice versa.

- Deriving the mathematical formula for $Var(\widehat{\beta})$ is beyond the scope of this unit. However, under the assumption that

$$Var(y|X) = Var(u|X) = \sigma^2 I_n, \tag{23}$$

where

$$\sigma^2 = var(y_i|X) = var(u_i|X), i = 1, 2, ..., n,$$

it can be shown that the mathematical formula for $Var(\widehat{\beta})$ in the general linear regression model is given by

$$Var(\widehat{\beta}) = \sigma^2 E(X'X)^{-1}. \tag{24}$$

- Since we don't observe either σ^2 or $E(X'X)^{-1}$, we don't observe $Var(\widehat{\beta})$.
- We estimate $Var(\widehat{\beta})$ using the formula

$$\widehat{Var}(\widehat{\beta}) = \widehat{\sigma}^2 (X'X)^{-1}, \tag{25}$$

where, with k explanatory variables,

$$\widehat{\sigma}^2 = \frac{1}{(n-k-1)} \sum_{i=1}^{n} \widehat{u}_i^2 = \frac{SSR}{(n-k-1)}. \tag{26}$$

- For example, when $k = 2$ the linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, \;\; i = 1, 2, ...n,$$

and

$$\underset{(3x3)}{Var(\widehat{\beta})} = \begin{bmatrix} Var(\widehat{\beta}_0) & Cov(\widehat{\beta}_0, \widehat{\beta}_1) & Cov(\widehat{\beta}_0, \widehat{\beta}_2) \\ Cov(\widehat{\beta}_1, \widehat{\beta}_0) & Var(\widehat{\beta}_1) & Cov(\widehat{\beta}_1, \widehat{\beta}_2) \\ Cov(\widehat{\beta}_2, \widehat{\beta}_0) & Cov(\widehat{\beta}_2, \widehat{\beta}_1) & Var(\widehat{\beta}_2) \end{bmatrix}. \qquad (21)$$

- When we replace the elements in the matrix given by (21) with their estimated values given by the formula

$$\widehat{Var}(\widehat{\beta}) = \widehat{\sigma}^2 (X'X)^{-1},$$

we obtain the estimated variance matrix of $\widehat{\beta}$ given by

$$\widehat{Var}(\widehat{\beta}) = \begin{bmatrix} \widehat{Var}(\widehat{\beta}_0) & \widehat{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) & \widehat{Cov}(\widehat{\beta}_0, \widehat{\beta}_2) \\ \widehat{Cov}(\widehat{\beta}_1, \widehat{\beta}_0) & \widehat{Var}(\widehat{\beta}_1) & \widehat{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) \\ \widehat{Cov}(\widehat{\beta}_2, \widehat{\beta}_0) & \widehat{Cov}(\widehat{\beta}_2, \widehat{\beta}_1) & \widehat{Var}(\widehat{\beta}_2) \end{bmatrix}. \quad (27)$$

- When we estimate a linear regression model in Eviews, the reported output includes

$$se(\widehat{\beta}_0) = \sqrt{\widehat{Var}(\widehat{\beta}_0)}, se(\widehat{\beta}_1) = \sqrt{\widehat{Var}(\widehat{\beta}_1)}, se(\widehat{\beta}_2) = \sqrt{\widehat{Var}(\widehat{\beta}_2)}. \quad (28)$$

The statistical properties of the OLS estimator IX

- The quantities in (28) are used for three purposes:
  - To measure of the precision with which $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ estimate β_0, β_1 and β_2 respectively.
  - To construct confidence intervals for β_0, β_1 and β_2.
  - To conduct hypothesis tests involving restrictions on β_0, β_1 and β_2.
- We saw in Topic 4 that when we estimate the linear regression equation

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 IQ_i + u_i, i = 1, 2, ...935, \qquad (29)$$

we obtain the results reported in Figure 1 below.

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-128.8899	92.18232	-1.398206	0.1624
EDUC	42.05762	6.549836	6.421171	0.0000
IQ	5.137958	0.955827	5.375403	0.0000

R-squared	0.133853	Mean dependent var	957.9455
Adjusted R-squared	0.131995	S.D. dependent var	404.3608
S.E. of regression	376.7300	Akaike info criterion	14.70414
Sum squared resid	1.32E+08	Schwarz criterion	14.71967

Figure: 1

- The **standard errors** (se) of the estimated regression coefficients are reported in column 3 of Figure 1.

- In econometrics we refer to the estimated standard deviation of a variable as the standard error (se) of the variable.
- In this example,

$$se(\widehat{\beta}_0) = 92.18, se(\widehat{\beta}_1) = 6.55, se(\widehat{\beta}_2) = 0.96.$$

- That is,

$$\begin{aligned} se(\widehat{\beta}_0) &= \sqrt{\widehat{Var}(\widehat{\beta}_0)} = 92.18, \\ se(\widehat{\beta}_1) &= \sqrt{\widehat{Var}(\widehat{\beta}_1)} = 6.55, \\ se(\widehat{\beta}_2) &= \sqrt{\widehat{Var}(\widehat{\beta}_2)} = 0.96. \end{aligned}$$

- The variable "S.E. of regression" reported in Figure 1, which we denote by $\widehat{\sigma}$, is called the **standard error of the regression** and is defined as

$$\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} = \sqrt{\frac{1}{(n-k-1)} \sum_{i=1}^{n} \widehat{u}_i^2} = \sqrt{\frac{SSR}{(n-k-1)}}. \tag{30}$$

- By definition,

$$\widehat{\sigma} = se(y_i|X) = se(u_i|X), i = 1, 2, ..., n. \tag{31}$$

- From Figure 1,

$$\widehat{\sigma} = 376.73 \Rightarrow \widehat{\sigma}^2 = (376.73)^2.$$

- Therefore, the estimated standard deviation of weekly wages **in the population of wage earners under consideration** is approximately $377.
- Since, from Figure 1, the mean weekly wage in our sample is approximately $958, the estimated standard deviation of weekly wages for the population is quite large.

- The fact an estimator of some unknown parameter of interest is unbiased, does not necessarily imply that we will obtain an accurate estimate of the parameter if we use the estimator in question.
- To see this, let $\widetilde{\theta}$ and $\widehat{\theta}$ denote two **unbiased estimators** of some unknown population parameter of interest, θ.
- Recall that because the values of $\widetilde{\theta}$ and $\widehat{\theta}$ depend on the particular sample we collect, they are both random variables.
- The pdf of each estimator is shown in Figure 2 below.

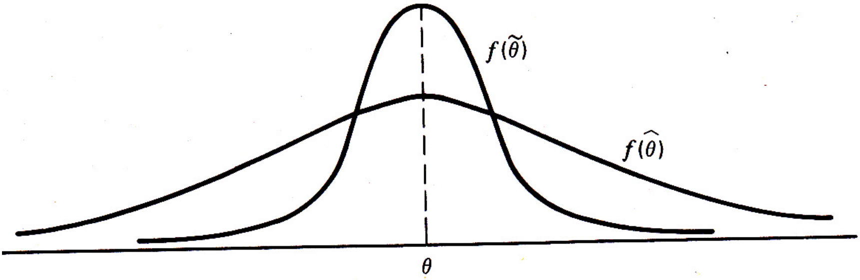

Figure: 2

- Notice that even though $\widehat{\theta}$ is an unbiased estimator of θ, it has a very large variance.

- When we collect a sample of data and we use $\widehat{\theta}$ to estimate θ, we are effectively making a random draw from $f(\widehat{\theta})$.
- Even though
$$E(\widehat{\theta}) = \theta,$$
because $\widehat{\theta}$ has such a large variance, there is a high probability that we will get of value of $\widehat{\theta}$ which is very different from θ.
- In contrast, because the estimator $\widetilde{\theta}$ has a much smaller variance than $\widehat{\theta}$, there is a much higher probability of getting a value of $\widetilde{\theta}$ which is close to θ.
- When two estimators of θ, $\widetilde{\theta}$ and $\widehat{\theta}$, are unbiased and
$$Var\left(\widetilde{\theta}\right) < Var\left(\widehat{\theta}\right),$$
say that $\widetilde{\theta}$ is a more **efficient** estimator of θ than is $\widehat{\theta}$.

- That is, the unbiased estimator $\widetilde{\theta}$ is more efficient than the unbiased estimator $\widehat{\theta}$ if $\widetilde{\theta}$ has a smaller variance than $\widehat{\theta}$.
- Given a choice between two unbiased estimators, we would always choose the more efficient estimator.
- Notice that when we compare estimators in terms of their efficiency, we restrict our attention to unbiased estimators.

- The OLS estimator is a called a linear estimator since

$$
\begin{aligned}
\widehat{\beta} &= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X' \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \\
&= (X'X)^{-1} \begin{bmatrix} x_{.1} & x_{.2} & . & . & x_{.n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \\
&= (X'X)^{-1}x_{.1}y_1 + (X'X)^{-1}x_{.2}y_2 + ... + (X'X)^{-1}x_{.n}y_n,
\end{aligned}
$$

where $x_{.i}$ is the ith column of X'. (Check that the conformability conditions for matrix multiplication and matrix addition hold).

- That is, $\widehat{\beta}$ can be written as a linear function of the elements of the vector y.
- There is a very famous theorem in econometrics, called the **Gauss-Markov Theorem (GMT)**, which states that **under certain assumptions** the OLS estimator, $\widehat{\beta}$, is at least as efficient as any other linear, unbiased estimator of β in the linear regression model

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

- This proposition is summarized by stating that the OLS estimator is the **BLUE (best linear unbiased estimator)** of β.
- The proof of the GMT requires the following assumptions:

A1 The model is linear in the parameters. That is,

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

A2 There is no perfect multicollinearity. (No column of the X matrix can be written as an exact linear function of the remaining columns of X).

A3

$$E(u|X) = 0.$$

A4

$$Var(y|X) = \sigma^2 I_n,$$

or equivalently,

$$Var(u|X) = \sigma^2 I_n.$$

- We next state the GMT. The proof of the theorem is beyond the scope of this unit.

Theorem (2: The Gauss-Markov Theorem)

Let

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

denote the OLS estimator of β *in the linear regression model*

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

When assumptions A1-A4 hold, $\widehat{\beta}$ *is the best linear unbiased estimator of* β.

- Let

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ . \\ . \\ \widehat{\beta}_k \end{bmatrix} \text{ and } \widetilde{\beta} = \begin{bmatrix} \widetilde{\beta}_0 \\ \widetilde{\beta}_1 \\ \widetilde{\beta}_2 \\ . \\ . \\ \widetilde{\beta}_k \end{bmatrix},$$

where $\widehat{\beta}$ denotes the OLS estimator of β and $\widetilde{\beta}$ denotes **any other linear unbiased estimator** of β in the linear regression model

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

- The GMT states that, when assumptions A1-A4 hold, $\widehat{\beta}$ is at least as efficient as $\widetilde{\beta}$ in the sense that

$$\begin{bmatrix} Var(\widehat{\beta}_0) \\ Var(\widehat{\beta}_1) \\ Var(\widehat{\beta}_2) \\ . \\ . \\ Var(\widehat{\beta}_k) \end{bmatrix} \leq \begin{bmatrix} Var(\widetilde{\beta}_0) \\ Var(\widetilde{\beta}_1) \\ Var(\widetilde{\beta}_2) \\ . \\ . \\ Var(\widetilde{\beta}_k) \end{bmatrix}.$$

- In other words, when A1-A4 hold, we can't do better then OLS!
- Note carefully that in order to prove that the OLS estimator of β is unbiased, we only need A1, A2 and A3.
- However, in order to prove the GMT, in addition to A1, A2 and A3 we also need A4.

- A4, which states that

$$Var(y|X) = \sigma^2 I_n, \tag{33}$$

is a very strong assumption that often fails to hold in practise.

- A4 can be decomposed into two parts:

A4 (a)

$$Var(y_1|X) = Var(y_2|X) = ... = Var(y_n|X) = \sigma^2.$$

This assumption states that the conditional variance of the dependent variable is the same for all observations. This assumption, which is known as the **homoskedasticity** assumption, is often unrealistic when working with cross-sectional data.

- For example, in the wage equation the assumption of homoskedasticity implies that the variance of wages for the population of individuals with high levels of education is the same as the variance of wages for the population of individuals with low levels of education. This is clearly unrealistic.

A4 (b)

$$Cov(y_i, y_j|X) = 0 \text{ for } i \neq j.$$

For example,

$$Cov(y_1, y_2|X) = 0.$$

As we will see later in the unit, the assumption that different observations on y are uncorrelated seldom holds when working with time series data.

- In summary, while the GMT demonstrates the superiority of the OLS estimator of β relative to any other linear unbiased estimator of β, the assumptions required to prove the theorem seldom hold. Specifically:

- A4 (a) is usually unrealistic when working with cross-sectional data.
- A4 (b) is almost always unrealistic when working with time series data.

8. Appendix 1: Proof of the unbiasedness of the OLS estimator I

Theorem (1)

When assumptions A1,A2 and A3 hold

$$E(\widehat{\beta}) = \beta,$$

where

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

is the OLS estimator of β in the linear regression model

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

8. Appendix 1: Proof of the unbiasedness of the OLS estimator II

Proof:

$$\begin{aligned}
\widehat{\beta} &= (X'X)^{-1}X'y \text{ (using A2)} \\
&= (X'X)^{-1}X'(X\beta + u) \text{ (using A1)} \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\
&= I_{k+1}\beta + (X'X)^{-1}X'u \\
&= \beta + (X'X)^{-1}X'u \qquad \text{(A1.1)}
\end{aligned}$$

8. Appendix 1: Proof of the unbiasedness of the OLS estimator III

Taking expectations on both sides of A1.1

$$
\begin{aligned}
E(\widehat{\beta}) &= E\left[\beta + (X'X)^{-1}X'u\right] \\
&= E(\beta) + E[(X'X)^{-1}X'u] \\
&= \beta + E[(X'X)^{-1}X'u] \\
&= \beta + E_X\{E[(X'X)^{-1}X'u)|X\} \text{ (using the LIE)} \\
&= \beta + E_X[(X'X)^{-1}X'E(u|X)] \\
&= \beta + E_X[(X'X)^{-1}X'\mathbf{0}] \text{ (using A3)} \\
&= \beta + E_X(\mathbf{0}) \\
&= \beta + \mathbf{0} \\
&= \beta.
\end{aligned}
$$

9. Appendix 2: Derivation of the Variance-Covariance matrix for a 2x1 random vector I

- Let

$$\underset{(2x1)}{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \Rightarrow \underset{(1x2)}{z'} = \begin{bmatrix} z_1 & z_2 \end{bmatrix}.$$

- By definition

$$Var(z) = E\{[z - E(z)][z - E(z)]'\}. \tag{A2.1}$$

- By definition

$$[z - E(z)] = \begin{bmatrix} z_1 - E(z_1) \\ z_2 - E(z_2) \end{bmatrix}$$

and

$$[z - E(z)]' = \begin{bmatrix} z_1 - E(z_1) & z_2 - E(z_2) \end{bmatrix}.$$

9. Appendix 2: Derivation of the Variance-Covariance matrix for a 2x1 random vector II

- Therefore,

$$[z - E(z)][z - E(z)]' = \begin{bmatrix} z_1 - E(z_1) \\ z_2 - E(z_2) \end{bmatrix} \begin{bmatrix} z_1 - E(z_1) & z_2 - E(z_2) \end{bmatrix}$$

$$= \begin{bmatrix} [z_1 - E(z_1)]^2 & [z_1 - E(z_1)][z_2 - E(z_2)] \\ [z_2 - E(z_2)][z_1 - E(z_1)] & [z_2 - E(z_2)]^2 \end{bmatrix} \quad \text{(A2.2)}$$

9. Appendix 2: Derivation of the Variance-Covariance matrix for a 2x1 random vector III

- Applying the expectation operator to A2.2 we obtain

$$
\begin{aligned}
& E\begin{bmatrix} [z_1 - E(z_1)]^2 & [z_1 - E(z_1)][z_2 - E(z_2)] \\ [z_2 - E(z_2)][z_1 - E(z_1)] & [z_2 - E(z_2)]^2 \end{bmatrix} \\
= & \begin{bmatrix} E[z_1 - E(z_1)]^2 & E[z_1 - E(z_1)][z_2 - E(z_2)] \\ E[z_2 - E(z_2)][z_1 - E(z_1)] & E[z_2 - E(z_2)]^2 \end{bmatrix} \\
= & \begin{bmatrix} Var(z_1) & Cov(z_1, z_2) \\ Cov(z_2, z_1) & Var(z_2) \end{bmatrix}. \qquad \text{(A2.3)}
\end{aligned}
$$

9. Appendix 2: Derivation of the Variance-Covariance matrix for a 2x1 random vector IV

- Finally, substituting (A2.3) into

$$Var(z) = E\{[z - E(z)][z - E(z)]'\}. \qquad \text{(A2.1)}$$

we obtain

$$\underset{(2x2)}{Var(z)} = \begin{bmatrix} Var(z_1) & Cov(z_1, z_2) \\ Cov(z_2, z_1) & Var(z_2) \end{bmatrix},$$

which we may write more compactly as

$$\underset{(2x2)}{Var(z)} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix},$$

where

$$\sigma_1^2 = Var(z_1), \sigma_2^2 = Var(z_2), \sigma_{12} = Cov(z_1, z_2), \sigma_{21} = Cov(z_2, z_1).$$

9. Appendix 2: Derivation of the Variance-Covariance matrix for a 2x1 random vector V

- In summary, we have shown that, given a 2x1 vector of random variables,

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

there is an associated 2x2 variance-covariance matrix given by

$$\underset{(2x2)}{Var(z)} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}.$$