

# ECO585: Applied Econometrics

## **OLS**

### Lecture 2

Prof. Mary Kaltenberg

Pace University

February 4, 2021

# Simple OLS

Suppose there are two variables,  $x$  and  $y$ , and we would like to “study how  $y$  varies with changes in  $x$ .”

For example  $x$  is years of schooling,  $y$  is hourly wage.

We must confront three issues:

1. How do we allow factors other than  $x$  to affect  $y$ ? There is never an exact relationship between two variables (in interesting cases).
2. What is the functional relationship between  $y$  and  $x$ ?
3. How can we be sure we are capturing a ceteris paribus relationship between  $y$  and  $x$  (as is so often the goal)?

# Simple OLS

We can consider the simple linear regression model:

$$y = \beta_0 + \beta_1 x + u,$$

$y$  and  $x$  are not treated symmetrically. We want to explain  $y$  in terms of  $x$ . From a causality standpoint, it makes no sense to “explain” past educational attainment in terms of future labor earnings.

As another example, we want to explain student performance ( $y$ ) in terms of class size ( $x$ ), not the other way around.

# Simple OLS

The **error term** or **disturbance**,  $u$

$$y = \beta_0 + \beta_1 x + u$$

explicitly allows for other factors, contained in  $u$ , to affect  $y$ .

This equation also addresses the functional form issue (in a simple way). Namely,  $y$  is assumed to be *linearly* related to  $x$ .

$\beta_0$  the **intercept parameter** and  $\beta_1$  the **slope parameter**.

These describe a population, and our ultimate goal is to estimate them.

# Simple OLS

The equation also addresses the ceteris paribus issue. In

$$y = \beta_0 + \beta_1 x + u,$$

all other factors that affect  $y$  are in  $u$ . We want to know how  $y$  changes when  $x$  changes, *holding  $u$  fixed*.

Let  $\Delta$  denote “change.” Then holding  $u$  fixed means  $\Delta u = 0$ .

# Simple OLS

How can we hope to generally estimate the ceteris paribus effect of  $y$  on  $x$  when we have assumed all other factors affecting  $y$  are unobserved and lumped into  $u$ ?

The key is that the simple linear regression (SLR) model is a population model. When it comes to *estimating*  $\beta_1$  (and  $\beta_0$ ) using a random sample of data, we must restrict how  $u$  and  $x$  are related to each other.

But  $x$  and  $u$  are properly viewed as having distributions in the population.

What we must do is restrict the way in which  $u$  and  $x$  relate to each other in the population.

## Simple OLS

First, we make a simplifying assumption that is without loss of generality: the average, or expected, value of  $u$  is zero in the population:

$$E(u) = 0$$

where  $E(\cdot)$  is the expected value (or averaging) operator.  
The presence of  $\beta_0$  in

$$y = \beta_0 + \beta_1 x + u$$

allows us to assume  $E(u) = 0$ . If the average of  $u$  is different from zero, we just adjust the intercept, leaving the slope the same. If  $\alpha_0 = E(u)$  then we can write

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0),$$

where the new error,  $u - \alpha_0$ , has a zero mean.

The new intercept is  $\beta_0 + \alpha_0$ . The important point is that the slope,  $\beta_1$ , has not changed.

# Simple OLS

How do we need to restrict the dependence between  $u$  and  $x$ ?

We could assume  $u$  and  $x$  **uncorrelated** in the population:

$$\text{Corr}(x, u) = 0$$

Zero correlation actually works for many purposes, but it implies only that  $u$  and  $x$  are not **linearly** related. Ruling out only linear dependence can cause problems with interpretation and makes statistical analysis more difficult.

An assumption that meshes well with our introductory treatment involves the mean of the error term for each slice of the population determined by values of  $x$ :

$$E(u|x) = E(u), \text{ all values } x,$$

where  $E(u|x)$  means “the expected value of  $u$  given  $x$ .”

We say  $u$  is **mean independent** of  $x$ .



# Simple OLS

Suppose  $u$  is “land quality” and  $x$  is fertilizer amount. Then  $E(u|x) = E(u)$  if fertilizer amounts are chosen independently of quality. This assumption is reasonable but assumes fertilizer amounts are assigned at random.

Combining  $E(u|x) = E(u)$  (the substantive assumption) with  $E(u) = 0$  (a normalization) gives

$$E(u|x) = 0, \text{ all values } x$$

Called the **zero conditional mean assumption**.

# Simple OLS

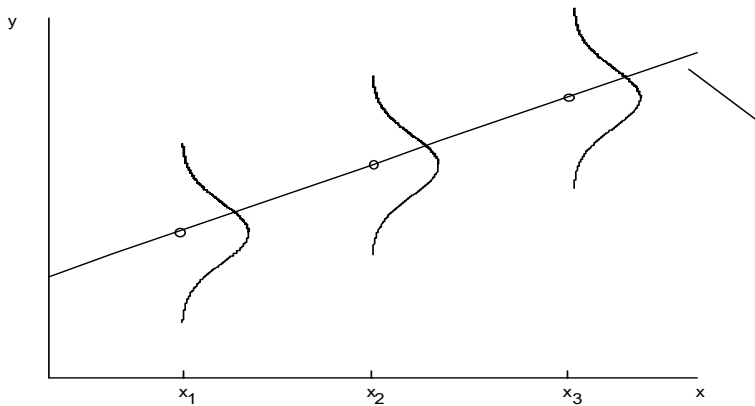
Because the expected value is a linear operator,  $E(u|x) = 0$  implies

$$E(y|x) = \beta_0 + \beta_1 x + E(u|x) = \beta_0 + \beta_1 x,$$

which shows the **population regression function** is a linear function of  $x$ .

A different approach to simple regression ignores the causality issue and just starts with a linear model for  $E(y|x)$  as a descriptive device.

# Simple OLS



# Deriving the OLS Estimates

Given data on  $x$  and  $y$ , how can we estimate the population parameters,  $\beta_0$  and  $\beta_1$ ?

Let  $\{(x_i, y_i) : i = 1, 2, \dots, n\}$  be a sample of size  $n$  (the number of observations) from the population. Think of this as a random sample.

The next graph shows  $n = 15$  families and the population regression of saving on income.

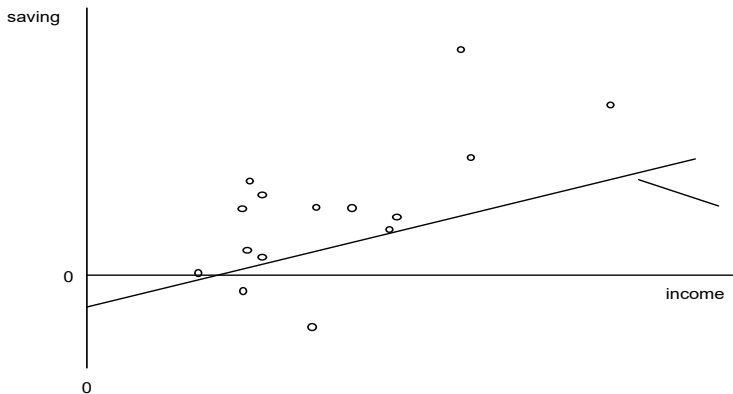
Plug any observation into the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where the  $i$  subscript indicates a particular observation.

We observe  $y_i$  and  $x_i$ , but not  $u_i$ . (However, we know  $u_i$  is there.)

# Deriving the OLS Estimates



# Deriving the OLS Estimates

We use the two restrictions

$$E(u) = 0$$

$$\text{Cov}(x, u) = 0$$

to obtain estimating equations for  $\beta_0$  and  $\beta_1$ .

Remember, the first condition essentially defines the intercept.

The second condition, stated in terms of the covariance, means that  $x$  and  $u$  are uncorrelated.

Both conditions are implied by the zero conditional mean assumption

$$E(u|x) = 0$$

# Deriving the OLS Estimates

With  $E(u) = 0$ ,  $Cov(x, u) = 0$  is the same as  $E(xu) = 0$  because  $Cov(x, u) = E(xu) - E(x)E(u)$ .

Next we plug in for  $u$  into the two equations:

$$\begin{aligned} E(y - \beta_0 - \beta_1 x) &= 0 \\ E[x(y - \beta_0 - \beta_1 x)] &= 0 \end{aligned}$$

These are the two conditions in the population that determine  $\beta_0$  and  $\beta_1$ . So we use their sample analogs, which is a method of moments approach to estimation.

# Deriving the OLS Estimates

In other words, we use

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$n^{-1} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

to determine  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , the estimates from the data.

These are two linear equations in the two unknowns  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .



# Deriving the OLS Estimates

To solve the equations, pass the summation operator through the first equation:

$$\begin{aligned}n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= n^{-1} \sum_{i=1}^n y_i - n^{-1} \sum_{i=1}^n \hat{\beta}_0 - n^{-1} \sum_{i=1}^n \hat{\beta}_1 x_i \\&= n^{-1} \sum_{i=1}^n y_i - \hat{\beta}_0 - \hat{\beta}_1 \left( n^{-1} \sum_{i=1}^n x_i \right) \\&= \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}\end{aligned}$$

# Deriving the OLS Estimates

We use the standard notation  $\bar{y} = n^{-1} \sum_{i=1}^n y_i$  for the average of the  $n$  numbers  $\{y_i : i = 1, 2, \dots, n\}$ . For emphasis, we call  $\bar{y}$  a **sample average**.

We have shown that the first equation,

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

implies

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

# Deriving the OLS Estimates

Rewrite this equation so that the intercept is terms of the slope (and the sample averages on  $y$  and  $x$ ):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and plug this into the second equation (and drop the division by  $n$ ):

$$\sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

so

$$\sum_{i=1}^n x_i [y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] = 0$$

# Deriving the OLS Estimates

Simple algebra gives

$$\sum_{i=1}^n x_i(y_i - \bar{y}) = \hat{\beta}_1 \left[ \sum_{i=1}^n x_i(x_i - \bar{x}) \right]$$

and so we have one linear equation in the one unknown  $\hat{\beta}_1$ .

# Deriving the OLS Estimates

Showing the solution for  $\hat{\beta}_1$  uses three useful facts about the summation operator:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^n x_i (y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x}) y_i$$

$$\sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})^2$$

# Deriving the OLS Estimates

So, we can write the equation to solve is

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

If  $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$ , we can write

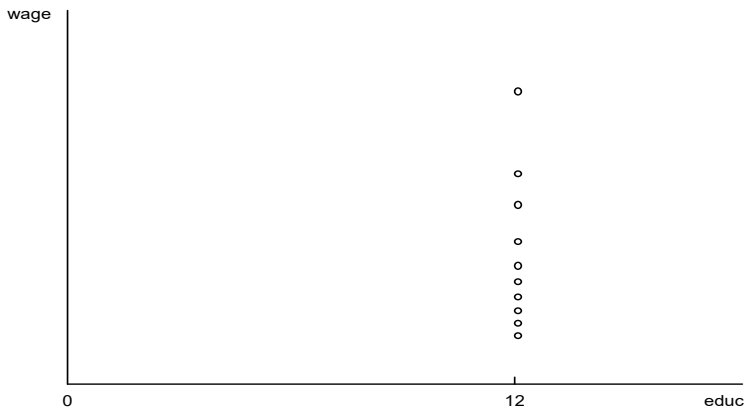
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x_i, y_i)}{\text{Sample Variance}(x_i)}$$

# Deriving the OLS Estimates

The previous formula for  $\hat{\beta}_1$  is important. It shows us how to take the data we have and compute the slope estimate. For reasons we will see,  $\hat{\beta}_1$  is called the **ordinary least squares (OLS)** slope estimate. We often refer to it as the **slope estimate**.

It can be computed whenever the sample variance of the  $x_i$  is not zero, which only rules out the case where each  $x_i$  is the same value.

# Deriving the OLS Estimates





# Deriving the OLS Estimates

Where does the name “ordinary least squares” come from?

For any candidates  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , define a **fitted value** for each data point  $i$  as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

We have  $n$  of these. It is the value we predict for  $y_i$  given that  $x$  has taken on the value  $x_i$ .

The mistake we make is the **residual**:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

and we have  $n$  residuals.

# Deriving the OLS Estimates

Suppose we measure the size of the mistake, for each  $i$ , by squaring the residual:  $\hat{u}_i^2$ . Then we add them all up:

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

This quantity is called the **sum of squared residuals**.

If we choose  $\hat{\beta}_0$  and  $\hat{\beta}_1$  to *minimize* the sum of squared residuals it can be shown (using calculus or other arguments) that the solutions are the slope and intercept estimates we obtained before.

# Deriving the OLS Estimates

Once we have the numbers  $\hat{\beta}_0$  and  $\hat{\beta}_1$  for a given data set, we write the **OLS regression line** as a function of  $x$ :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The OLS regression line allows us to predict  $y$  for any (sensible) value of  $x$ . It is also called the **sample regression function**.

The intercept,  $\hat{\beta}_0$ , is the predicted  $y$  when  $x = 0$ . (The prediction is usually meaningless if  $x = 0$  is not possible.)

The slope,  $\hat{\beta}_1$ , allows us to predict changes in  $y$  for any (reasonable) change in  $x$ :

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x$$

If  $\Delta x = 1$ , so that  $x$  increases by one unit, then  $\Delta \hat{y} = \hat{\beta}_1$ .

# Properties of OLS

Once we have

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

we get the OLS fitted values by plugging the  $x_i$  into the equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, \dots, n$$

# Properties of OLS

## **Algebraic Properties of OLS Statistics**

(1) The OLS residuals *always* add up to zero:

$$\sum_{i=1}^n \hat{u}_i = 0$$

Because  $y_i = \hat{y}_i + \hat{u}_i$  by definition,

$$n^{-1} \sum_{i=1}^n y_i = n^{-1} \sum_{i=1}^n \hat{y}_i + n^{-1} \sum_{i=1}^n \hat{u}_i$$

and so  $\bar{y} = \overline{\hat{y}}$ . In other words, the sample average of the actual  $y_i$  is the same as the sample average of the fitted values.

# Properties of OLS

(2) The sample covariance (and therefore the sample correlation) between the explanatory variables and the residuals is always zero:

$$\sum_{i=1}^n x_i \hat{u}_i = 0$$

Because the  $\hat{y}_i$  are linear functions of the  $x_i$ , the fitted values and residuals are uncorrelated, too:

$$\sum_{i=1}^n \hat{y}_i \hat{u}_i = 0$$

Both of these properties hold by construction.  $\hat{\beta}_0$  and  $\hat{\beta}_1$  were chosen to make them true.

# Properties of OLS

(3) The point  $(\bar{x}, \bar{y})$  is always on the OLS regression line. That is, if we plug in the average for  $x$ , we predict the sample average for  $y$ :

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Again, we chose the estimates to make this true.

# Properties of OLS

## Goodness-of-Fit

For each observation, write

$$y_i = \hat{y}_i + \hat{u}_i$$

Define the total sum of squares (SST), explained sum of squares (SSE) – Stata calls this the “model sum of squares” – and residual sum of squares (or sum of squared residuals) as

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^n \hat{u}_i^2$$



# Properties of OLS

Each of these is a sample variance when divided by  $n$  (or  $n - 1$ ).  $SST/n$  is the sample variance of  $y_i$ ,  $SSE/n$  is the sample variance of  $\hat{y}_i$ , and  $SSR/n$  is the sample variance of  $\hat{u}_i$ .

By writing

$$\begin{aligned} SST &= \sum_{i=1}^n (y_i - \bar{y})^2 = SST = \sum_{i=1}^n [(y_i - \hat{y}_i) - (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^n [\hat{u}_i - (\hat{y}_i - \bar{y})]^2 \end{aligned}$$

and using that the fitted values and residuals are uncorrelated, can show

$$SST = SSE + SSR$$

# Properties of OLS

Assuming  $SST > 0$ , we can define the fraction of the total variation in  $y_i$  that is explained by  $x_i$  (or the OLS regression line) as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Called the **R-squared** of the regression.

It can be shown to equal the *square* of the correlation between  $y_i$  and  $\hat{y}_i$ . Therefore,

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

# Properties of OLS

$R^2 = 0$  means no linear relationship between  $y_i$  and  $x_i$ .  $R^2 = 1$  means a perfect linear relationship.

As  $R^2$  increases, the  $y_i$  are closer and closer to falling on the OLS regression line.

Do not want to fixate on  $R^2$ . It is a useful summary measure but tells us nothing about causality. Having a “high”  $R$ -squared is neither necessary nor sufficient to infer causality.

# Expected Value of OLS

So far, our analysis so far has been purely algebraic, based on a sample of data. (Residuals always average to zero regardless of any underlying model.)

Now our job gets harder. We have to study statistical properties of the OLS estimator, referring to a population model and assuming random sampling.

Mathematical statistics: How do our estimators behave across different samples of data? On average, would we get the right answer if we could repeatedly sample?

We need to find the expected value of the OLS estimators – in effect, the average outcome across all possible random samples – and determine if we are right on average.

This leads to the notion of **unbiasedness**.

# Expected Value of OLS

## Assumption SLR.1 (Linear in Parameters)

The population model can be written as

$$y = \beta_0 + \beta_1 x + u$$

where  $\beta_0$  and  $\beta_1$  are the (unknown) population parameters.

We view  $x$  and  $u$  as outcomes of random variables; thus,  $y$  is of course random.

Stating this assumption formally shows that our goal is to estimate  $\beta_0$  and  $\beta_1$ .

# Expected Value of OLS

## Assumption SLR.2 (Random Sampling)

We have a random sample of size  $n$ ,  $\{(x_i, y_i) : i = 1, \dots, n\}$ , following the population model.

We know how to use these data to estimate  $\beta_0$  and  $\beta_1$  by OLS. Because each  $i$  is a draw from the population, we can write

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

for each  $i$ .

Notice that  $u_i$  here is the unobserved error for observation  $i$ . It is not the residual that we compute from the data!

# Expected Value of OLS

## **Assumption SLR.3 (Sample Variation in the Explanatory Variable)**

The sample outcomes on  $x_i$  are not all the same value.

This is the same as saying the sample variance of

$\{x_i : i = 1, \dots, n\}$  is not zero.

In practice, this is hardly an assumption at all. If in the population  $x$  does not change then we are not asking an interesting question.

If the  $x_i$  are all the same value in the sample, we are unlucky and cannot proceed.

# Expected Value of OLS

## **Assumption SLR.4 (Zero Conditional Mean)**

In the population, the error term has zero mean given any value of the explanatory variable:

$$E(u|x) = 0 \quad \text{for all } x.$$

This is the key assumption for showing that OLS is unbiased, with the zero value not being important once we assume  $E(u|x)$  does not change with  $x$ .

Note that we can compute the OLS estimates whether or not this assumption holds, or even if there is an underlying population model.



# Expected Value of OLS

We will focus on  $\hat{\beta}_1$ . A few approaches to showing unbiasedness. One explicitly computes the expected value of  $\hat{\beta}_1$  conditional on the sample outcomes on  $x$ ,  $\{x_i : i = 1, 2, \dots, n\}$ . But it is cumbersome.

A second approach is to treat the  $x_i$  as nonrandom in the derivation. So, the randomness in  $\hat{\beta}_1$  comes through the  $u_i$  (equivalently, the  $y_i$ ).

The nonrandomness of the  $x_i$ , also called “fixed in repeated samples,” is not realistic in most cases but gets us to the same place. See the textbook (5e, pp. 48-49).

We use it only as a simplifying device. When thinking about applying OLS to the simple regression model, it is best to think in terms of Assumptions SLR.1 to SLR.4.

## Expected Value of OLS

How do we show  $\hat{\beta}_1$  is unbiased for  $\beta_1$ ? What we need to show is

$$E(\hat{\beta}_1) = \beta_1$$

where the expected value means averaging across random samples.

**Step 1:** Write down a formula for  $\hat{\beta}_1$ . It is convenient to use

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

which is one of several equivalent forms.

It is convenient to define  $SST_x = \sum_{i=1}^n (x_i - \bar{x})^2$ , to total variation in the  $x_i$ , and write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{SST_x}$$

Remember,  $SST_x$  is just some positive number. The existence of  $\hat{\beta}_1$  follows from SLR.3.

# Expected Value of OLS

**Step 2:** Replace each  $y_i$  with  $y_i = \beta_0 + \beta_1 x_i + u_i$  (which uses SLR.1 and the fact that we have data from SLR.2).

The numerator becomes

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x}) y_i &= \sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i + u_i) \\&= \beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) u_i \\&= 0 + \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x}) u_i \\&= \beta_1 SST_x + \sum_{i=1}^n (x_i - \bar{x}) u_i\end{aligned}$$

## Expected Value of OLS

We used  $\sum_{i=1}^n (x_i - \bar{x}) = 0$  and  $\sum_{i=1}^n (x_i - \bar{x})x_i = \sum_{i=1}^n (x_i - \bar{x})^2$ .  
We have shown

$$\hat{\beta}_1 = \frac{\beta_1 SST_x + \sum_{i=1}^n (x_i - \bar{x})u_i}{SST_x} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{SST_x}$$

# Expected Value of OLS

Note how the last piece is the slope coefficient from the OLS regression of  $u_i$  on  $x_i$ ,  $i = 1, \dots, n$ . We cannot do this regression because the  $u_i$  are not observed.

Now define

$$w_i = \frac{(x_i - \bar{x})}{SST_x}$$

so we have

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i$$

$\hat{\beta}_1$  is a linear function of the unobserved errors,  $u_i$ . The  $w_i$  are all functions of  $\{x_1, x_2, \dots, x_n\}$ .

The (random) difference between  $\hat{\beta}_1$  and  $\beta_1$  is due to this linear function of the unobservables.

# Expected Value of OLS

**Step 3:** Find  $E(\hat{\beta}_1)$ .

Under Assumptions SLR.2 and SLR.4,  $E(u_i | x_1, x_2, \dots, x_n) = 0$ .

That means, conditional on  $\{x_1, x_2, \dots, x_n\}$  (and using SLR.3),

$$E(w_i u_i) = w_i E(u_i) = 0$$

because  $w_i$  is a function of  $\{x_1, x_2, \dots, x_n\}$ .

This would not be true if, in the population,  $u$  and  $x$  are correlated.

## Expected Value of OLS

Now we can complete the proof: Conditional on  $\{x_1, x_2, \dots, x_n\}$ ,

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\beta_1 + \sum_{i=1}^n w_i u_i\right) \\ &= \beta_1 + \sum_{i=1}^n E(w_i u_i) = \beta_1 + \sum_{i=1}^n w_i E(u_i) \\ &= \beta_1, \end{aligned}$$

where we used two important properties of expected values:

1. the expected value of a sum is the sum of the expected values
2. the expected value of a constant,  $\beta_1$  in this case, is just itself.

Remember,  $\beta_1$  is the fixed constant in the population. The estimator,  $\hat{\beta}_1$ , varies across samples and is the random outcome: before we collect our data, we do not know what  $\hat{\beta}_1$  will be.

# Expected Value of OLS

## **THEOREM (Unbiasedness of OLS)**

Under Assumptions SLR.1 through SLR.4 and conditional on the outcomes  $\{x_1, x_2, \dots, x_n\}$ ,

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1.$$



# Expected Value of OLS

The four assumptions:

SLR.1:  $y = \beta_0 + \beta_1 x + u$

SLR.2: random sampling from the population

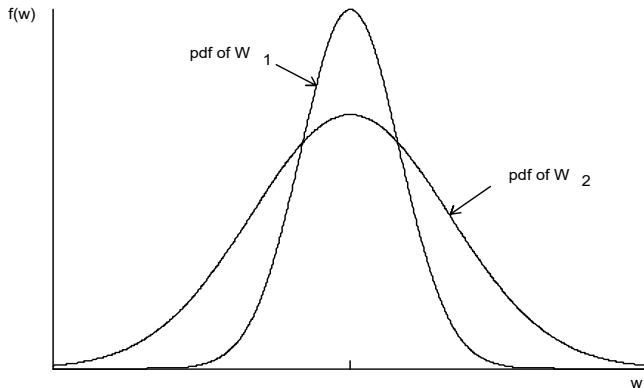
SLR.3: some sample variation in the  $x_i$

SLR.4:  $E(u|x) = 0$

The focus should mainly be on the last of these. What are the omitted factors? Are they likely to be correlated with  $x$ ? If so, SLR.4 fails and OLS will be biased.

## Variance of the OLS Estimators

Under SLR.1 to SLR.4, the OLS estimators are unbiased. This tells us that, on average, the estimates will equal the population values. But we need a measure of dispersion (spread) in the sampling distribution of the estimators. We use variance (and, ultimately, standard deviation).



# Variance of the OLS Estimators

We could characterize the variance of the OLS estimators under SLR.1 to SLR.4 (and we will later). For now, it is easiest to introduce an assumption that simplifies the calculations.

**Assumption SLR.5 (Homoskedasticity, or Constant Variance)**

The error has the same variance given any value of the explanatory variable  $x$ :

$$\text{Var}(u|x) = \sigma^2 > 0 \text{ for all } x,$$

where  $\sigma^2$  is (virtually always) unknown.

# Variance of the OLS Estimators

Because we assume SLR.4, that is,  $E(u|x) = 0$ , whenever we assume SLR.5, we can also write

$$E(u^2|x) = \sigma^2 = E(u^2)$$

Under the population Assumptions SLR.1 ( $y = \beta_0 + \beta_1 x + u$ ), SLR.4 ( $E(u|x) = 0$ ) and SLR.5 ( $Var(u|x) = \sigma^2$ ),

$$E(y|x) = \beta_0 + \beta_1 x$$

$$Var(y|x) = \sigma^2$$

So the average or expected value of  $y$  is allowed to change with  $x$  – in fact, this is what interests us – but the variance does not change with  $x$ .

# Variance of the OLS Estimators

The constant variance assumption may not be realistic; it must be determined on a case-by-case basis.

**EXAMPLE:** Suppose  $y = sav$ ,  $x = inc$  and we think

$$E(sav|inc) = \beta_0 + \beta_1 inc$$

with  $\beta_1 > 0$ . This means average family saving increases with income. If we impose SLR.5 then

$$Var(sav|inc) = \sigma^2$$

which means the variability in saving does not change with income. There are reasons to think saving would be more variable as income increases.

# Variance of the OLS Estimators

## THEOREM (Sampling Variances of OLS)

Under Assumptions SLR.1 to SLR.5, and conditional on the outcomes  $\{x_1, x_2, \dots, x_n\}$ ,

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x} \\ \text{Var}(\hat{\beta}_0) &= \frac{\sigma^2 (n^{-1} \sum_{i=1}^n x_i^2)}{SST_x} \end{aligned}$$

# Variance of the OLS Estimators

To show this result, write, as before,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i$$

where  $w_i = (x_i - \bar{x})/SST_x$ . We are treating the  $w_i$  as nonrandom in the derivation. Because  $\beta_1$  is a constant, it does not affect  $Var(\hat{\beta}_1)$ . Now, we need to use the fact that, for uncorrelated random variables, the variance of the sum is the sum of the variances.

# Variance of the OLS Estimators

The  $\{u_i : i = 1, 2, \dots, n\}$  are actually independent across  $i$ , and so they are uncorrelated. Therefore,

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{Var}\left(\sum_{i=1}^n w_i u_i\right) = \sum_{i=1}^n \text{Var}(w_i u_i) \\ &= \sum_{i=1}^n w_i^2 \text{Var}(u_i) = \sum_{i=1}^n w_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n w_i^2 \end{aligned}$$

where the second-to-last equality uses Assumption SLR.5, so that the variance of  $u_i$  does not depend on  $x_i$ .



# Variance of the OLS Estimators

Now we have

$$\begin{aligned}\sum_{i=1}^n w_i^2 &= \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(SST_x)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(SST_x)^2} \\ &= \frac{SST_x}{(SST_x)^2} = \frac{1}{SST_x}.\end{aligned}$$

We have shown

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$$

(again, conditional on  $\{x_1, x_2, \dots, x_n\}$ ).

## Variance of the OLS Estimators

This is the “standard” formula for the variance of the OLS slope estimator. It is *not* valid if Assumption SLR.5 does not hold.

The homoskedasticity assumption was *not* used to show unbiasedness of the OLS estimators. That requires only SLR.1 to SLR.4.

Usually we are interested in  $\beta_1$ . We can easily study the two factors that affect its variance.

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$$

1. As the error variance increases, that is, as  $\sigma^2$  increases, so does  $\text{Var}(\hat{\beta}_1)$ . The more “noise” in the relationship between  $y$  and  $x$  – that is, the larger variability in  $u$  – the harder it is to learn about  $\beta_1$ .
2. By contrast, more variation in  $\{x_i\}$  is a *good* thing:

$$SST_x \uparrow \text{ implies } \text{Var}(\hat{\beta}_1) \downarrow$$

# Variance of the OLS Estimators

Notice that  $SST_x/n$  is the sample variance in  $x$ . We can think of this as getting close to the population variance of  $x$ ,  $\sigma_x^2$ , as  $n$  gets large. This means

$$SST_x \approx n\sigma_x^2$$

which means, as  $n$  grows,  $Var(\hat{\beta}_1)$  shrinks at the rate  $1/n$ . This is why more data is a good thing: More data shrinks the sampling variance of our estimators.

The standard deviation of  $\hat{\beta}_1$  is the square root of the variance. So

$$sd(\hat{\beta}_1) = \frac{\sigma}{\sqrt{SST_x}}$$

This turns out to be the measure of variation that appears in confidence intervals and test statistics.

# Estimating the Error Variance

In the formula

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$$

we can compute  $SST_x$  from the observed data  $\{x_i : i = 1, \dots, n\}$ .  
We need a way to estimate  $\sigma^2$  Recall that

$$\sigma^2 = E(u^2).$$

## Variance of the OLS Estimators

Therefore, if we could observe a sample on the errors,  $\{u_i : i = 1, 2, \dots, n\}$ , an unbiased estimator of  $\sigma^2$  would be the sample average of the squared errors,

$$n^{-1} \sum_{i=1}^n u_i^2$$

But this not an estimator because we cannot compute it from the data we observe.

How about replacing each  $u_i$  with its “estimate,” the OLS residual  $\hat{u}_i$ ?

$$u_i = y_i - \beta_0 - \beta_1 x_i$$

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

# Variance of the OLS Estimators

$\hat{u}_i$  can be computed from the data because it depends on the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . Except by fluke,

$$\hat{u}_i \neq u_i$$

for any  $i$ .

In fact, simple algebra gives

$$\begin{aligned}\hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i\end{aligned}$$

$E(\hat{\beta}_0) = \beta_0$  and  $E(\hat{\beta}_1) = \beta_1$ , but the estimators almost always differ from the population values in a sample.

## Variance of the OLS Estimators

Now, what about this as an estimator of  $\sigma^2$ ?

$$n^{-1} \sum_{i=1}^n \hat{u}_i^2 = SSR/n$$

It is a true estimator and easily computed from the data after OLS. As it turns out, this estimator is slightly biased: Its expected value is less than  $\sigma^2$ .

The estimator does not account for the two restrictions on the residuals, used to obtain  $\hat{\beta}_0$  and  $\hat{\beta}_1$ :

$$\sum_{i=1}^n \hat{u}_i = 0$$

$$\sum_{i=1}^n x_i \hat{u}_i = 0$$

There is no such restriction on the unobserved errors,  $u_i$ .

# Variance of the OLS Estimators

The unbiased estimator of  $\sigma^2$  uses a **degrees-of-freedom** adjustment. The residuals have only  $n - 2$  degrees-of-freedom, not  $n$ .

The estimator used universally is

$$\hat{\sigma}^2 = SSR/(n - 2) = (n - 2)^{-1} \sum_{i=1}^n \hat{u}_i^2.$$

## **THEOREM (Unbiased Estimator of $\sigma^2$ )**

Under Assumptions SLR.1 to SLR.5, and conditional on  $\{x_1, \dots, x_n\}$ ,

$$E(\hat{\sigma}^2) = \sigma^2.$$



# Variance of the OLS Estimators

In regression output, it is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{SSR/(n-2)}$$

that is usually reported. This is an estimator of  $sd(u)$ , the standard deviation of the population error.

One small glitch is that  $\hat{\sigma}$  is not unbiased for  $\sigma$ . This will not matter for our purposes (and there is no unbiased estimator of  $\sigma$ ).  $\hat{\sigma}$  is called the **standard error of the regression**, which means it is an estimate of the standard deviation of the error in the regression. Stata calls it the **root mean squared error**.

Given  $\hat{\sigma}$ , we can now estimate  $sd(\hat{\beta}_1)$  and  $sd(\hat{\beta}_0)$ . The estimates of these are called the **standard errors** of the  $\hat{\beta}_j$ . We will use these a lot.

# Variance of the OLS Estimators

We just plug  $\hat{\sigma}$  in for  $\sigma$ :

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}}$$

where both the numerator and denominator are easily computed from the data.

# Multiple Regression

Generally, we can write a model with two independent variables as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

where  $\beta_0$  is the intercept,  $\beta_1$  measures the change in  $y$  with respect to  $x_1$ , holding other factors fixed, and  $\beta_2$  measures the change in  $y$  with respect to  $x_2$ , holding other factors fixed.

In the model with two explanatory variables, the key assumption about how  $u$  is related to  $x_1$  and  $x_2$  is

$$E(u|x_1, x_2) = 0.$$

For any values of  $x_1$  and  $x_2$  in the population, the average unobservable is equal to zero. (The value zero is not important because we have an intercept,  $\beta_0$  in the equation.)

# Multiple Regression

The **multiple linear regression model** can be written in the population as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

where  $\beta_0$  is the **intercept**,  $\beta_1$  is the parameter associated with  $x_1$ ,  $\beta_2$  is the parameter associated with  $x_2$ , and so on.

Contains  $k + 1$  (unknown) population parameters. We call  $\beta_1, \dots, \beta_k$  the **slope parameters**.

Now we have multiple explanatory or independent variables. We still have one explained or dependent variable. We still have an error term,  $u$ .

# Multiple Regression

Multiple regression allows more flexible functional forms. As a shorthand, let  $lwage = \log(wage)$ :

$$lwage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 exper^2 + u,$$

so that  $exper$  is allowed to have a quadratic effect on  $lwage$ .

Take  $x_1 = educ$ ,  $x_2 = IQ$ ,  $x_3 = exper$ , and  $x_4 = exper^2$ . Note that  $x_4$  is a *nonlinear* function of  $x_3$ .

We already know that  $100 \cdot \beta_1$  is the (ceteris paribus) percentage change in  $wage$  [not  $\log(wage)$ !] when  $educ$  increases by one year.  $100 \cdot \beta_2$  has a similar interpretation (for a one point increase in  $IQ$ ).  $\beta_3$  and  $\beta_4$  are harder to interpret, but we can use calculus to get the slope of  $lwage$  with respect to  $exper$ :

$$\frac{\partial lwage}{\partial exper} = \beta_3 + 2\beta_4 exper$$

Multiply by 100 to get the percentage effect. (More later.)

# Multiple Regression

The key assumption for the general multiple regression model is easy to state in terms of a conditional expectation:

$$E(u|x_1, \dots, x_k) = 0$$

Provided we are careful, we can make this condition closer to being true by “controlling for” more variables. In the wage example, we “control for” IQ when estimating the return to education.

# Multiple Regression

Suppose we have  $x_1$  and  $x_2$  ( $k = 2$ ) along with  $y$ . We want to fit an equation of the form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

given data  $\{(x_{i1}, x_{i2}, y_i) : i = 1, \dots, n\}$ . The sample size is again  $n$ . Now the explanatory variables have two subscripts:  $i$  is the observation number (as always) and the second subscript (1 and 2 in this case) are labels for particular variables. For example

$$x_{i1} = \text{educ}_i, i = 1, \dots, n$$

$$x_{i2} = \text{IQ}_i, i = 1, \dots, n$$

# Multiple Regression

As in the simple regression case, different ways to motivate OLS. We choose  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and  $\hat{\beta}_2$  (so three unknowns) to minimize the sum of squared residuals,

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2$$

The case with  $k$  independent variables is easy to state: choose the  $k + 1$  values  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$  to minimize

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik})^2$$

Can be solved with multivariable calculus. The **OLS first order conditions** are the  $k + 1$  linear equations in the  $k + 1$  unknowns  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ :



# Multiple Regression

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^n x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^n x_{i2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

$\vdots$

$$\sum_{i=1}^n x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

Later we discuss the condition needed for this set of equations to have a unique solution. Computers are good at finding the solution.

# Multiple Regression

This is important. The slope coefficients now explicitly have ceteris paribus interpretations.

Consider  $k = 2$ :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Then

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

allows us to compute how predicted  $y$  changes when  $x_1$  and  $x_2$  change by any amount.

# Multiple Regression

What if we “hold  $x_2$  fixed,” that is, its change is zero,  $\Delta x_2 = 0$ ?

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 \text{ if } \Delta x_2 = 0$$

In particular,

$$\hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x_1} \text{ if } \Delta x_2 = 0$$

In other words,  $\hat{\beta}_1$  is the slope of  $\hat{y}$  with respect to  $x_1$  when  $x_2$  is held fixed.

# Multiple Regression

Similarly,

$$\Delta \hat{y} = \hat{\beta}_2 \Delta x_2 \text{ if } \Delta x_1 = 0$$

and

$$\hat{\beta}_2 = \frac{\Delta \hat{y}}{\Delta x_2} \text{ if } \Delta x_1 = 0$$

We call  $\hat{\beta}_1$  and  $\hat{\beta}_2$  **partial effects** or **ceteris paribus effects**.

# Multiple Regression

## **Assumption MLR.3 (No Perfect Collinearity)**

In the sample (and, therefore, in the population), none of the explanatory variables is constant, and there are no exact linear relationships among them.

The need to rule out cases where  $\{x_{ij} : i = 1, \dots, n\}$  has no variation for each  $j$  is clear from simple regression.

There is a new part to the assumption because we have more than one explanatory variable. We must rule out the (extreme) case that one (or more) of the explanatory variables is an exact *linear* function of the others.

# Multiple Regression

If, say,  $x_{i1}$  is an exact linear function of  $x_{i2}, \dots, x_{ik}$  in the sample, we say the model suffers from **perfect collinearity**.

Under perfect collinearity, there are no unique OLS estimators.

Stata and other regression packages will indicate a problem.

Usually perfect collinearity arises from a bad specification of the population model. A small sample size or bad luck in drawing the sample can also be the culprit.

Assumption MLR.3 can only hold if  $n \geq k + 1$ , that is, we must have at least as many observations as we have parameters to estimate.

# Multiple Regression

Suppose that  $k = 2$  and  $x_1 = educ$ ,  $x_2 = exper$ . If we draw our sample so that

$$educ_i = 2exper_i$$

for every  $i$ , then Assumption MLR.3 is violated. This is very unlikely unless the sample is small. (In any realistic population there are plenty of people whose education level is not twice their years of workforce experience.)

With the samples we have looked at ( $n = 680$ ,  $n = 759$ , even  $n = 173$ ), the presence of perfect collinearity is usually a result of poor model specification, or defining variables inappropriately. Such problems can almost always be detected by remembering the *ceteris paribus* nature of multiple regression.

# Multiple Regression

**EXAMPLE:** Do not include the same variable in an equation that is measured in different units. For example, in a CEO salary equation, it would make no sense to include firm sales measured in dollars along with sales measured in millions of dollars. There is no new information once we include one of these. The return on equity should be included as a percent or proportion, but not both.



## Multiple Regression

**EXAMPLE:** Be careful with functional forms! Suppose we start with a constant elasticity model of family consumption:

$$\log(\text{cons}) = \beta_0 + \beta_1 \log(\text{inc}) + u$$

How might we allow the elasticity to be nonconstant, but include the above as a special case? The following does *not* work:

$$\log(\text{cons}) = \beta_0 + \beta_1 \log(\text{inc}) + \beta_2 \log(\text{inc}^2) + u$$

because  $\log(\text{inc}^2) = 2 \log(\text{inc})$ , that is,  $x_2 = 2x_1$ , where  $x_1 = \log(\text{inc})$ .

Instead, we probably mean something like

$$\log(\text{cons}) = \beta_0 + \beta_1 \log(\text{inc}) + \beta_2 [\log(\text{inc})]^2 + u$$

which means  $x_2 = x_1^2$ . With this choice,  $x_2$  is an exact *nonlinear* function of  $x_1$ , but this (fortunately) is allowed in MLR.3.

Tracking down perfect collinearity can be harder when it involves more than two variables.

# Multiple Regression

**EXAMPLE:** Motivated by the data in VOTE1.DTA:

$$\text{voteA} = \beta_0 + \beta_1 \text{expendA} + \beta_2 \text{expendB} + \beta_3 \text{totexpend} + u$$

where *expendA* is campaign spending by candidate A, *expendB* is spending by candidate B, and *totexpend* is total spending. All are in thousands of dollars. Mechanically, the problem is that, by definition,

$$\text{expendA} + \text{expendB} = \text{totexpend}$$

which, of course, will also be true for any sample we collect.

# Multiple Regression

**A Key Point:** Assumption MLR.3 does *not* say the explanatory variables have to be uncorrelated – in the population or sample. Nor does it say they cannot be “highly” correlated. MLR.3 rules out *perfect correlation* in the sample, that is, correlations of  $\pm 1$ . Again, in practice violations of MLR.3 are rare unless a mistake has been made in specifying the model.

Multiple regression would be useless if we had to insist  $x_1, \dots, x_k$  were uncorrelated in the sample (or population)!

If the  $x_j$  were all pairwise uncorrelated, we could just use a bunch of simple regressions.

# Multiple Regression

MLR.1:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$  (similar to SLR1)

MLR.2: random sampling from the population (similar to SLR2)

MLR.3: no perfect collinearity in the sample (new!)

The last assumption ensures that the OLS estimators are unique and can be obtained from the first order conditions (minimizing the sum of squared residuals).

We need a final assumption for unbiasedness.

# Multiple Regression

## Assumption MLR.4 (Zero Conditional Mean)

$$E(u|x_1, x_2, \dots, x_k) = 0 \text{ for all } (x_1, \dots, x_k)$$

Remember, the real assumption is  $E(u|x_1, x_2, \dots, x_k) = E(u)$ : the average value of the error does not change across different slices of the population defined by  $x_1, \dots, x_k$ . Setting  $E(u) = 0$  essentially defines  $\beta_0$ .

If  $u$  is correlated with any of the  $x_j$ , MLR.4 is violated. This is usually a good way to think about the problem.

# Multiple Regression

When Assumption MLR.4 holds, we say  $x_1, \dots, x_k$  are **exogenous explanatory variables**. If  $x_j$  is correlated with  $u$ , we often say  $x_j$  is an **endogenous explanatory variable**. (This name makes more sense in more advanced contexts, but it is used generally.)

## THEOREM (Unbiasedness of OLS)

Under Assumptions MLR.1 through MLR.4, and conditional on  $\{(x_{i1}, \dots, x_{ik}) : i = 1, \dots, n\}$ , the OLS estimators are unbiased:

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, 2, \dots, k$$

for any values of the  $\beta_j$ .

Often the hope is that if our focus is on, say,  $x_1$ , we can include enough other variables in  $x_2, \dots, x_k$  to make MLR.4 true, or close to true.

## OV B

It is important to see that the unbiasedness result allow for the  $\beta_j$  to be any value, including zero.

Suppose, then, that we specify the model

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 motheduc + u,$$

where MLR.1 through MLR.4 hold. Suppose that  $\beta_3 = 0$ , but we do not know that. We estimate the full model by OLS:

$$\widehat{lwage} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 exper + \hat{\beta}_3 motheduc$$

We automatically know from the unbiasedness result that

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, 2$$

$$E(\hat{\beta}_3) = 0$$

The result that **including an irrelevant variable**, or overspecifying the model, does not cause bias in any coefficients is often presented with an extra argument. But it follows from the general unbiasedness result

Leaving a variable out when it should be including in multiple regression is a serious problem. This is called **excluding a relevant variable** or underspecifying the model.

We can perform a **misspecification analysis** in this case. The general case is more complicated.

Consider the case where the correct model has two explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

satisfies MLR.1 through MLR.4.



# OVB

If we regress  $y$  on  $x_1$  and  $x_2$ , we know the resulting estimators will be unbiased. But suppose we leave out  $x_2$  and use simple regression of  $y$  on  $x_1$ :

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

In most cases, we omit  $x_2$  because we cannot collect data on it. We can easily derive the bias in  $\tilde{\beta}_1$  (conditional on the sample outcomes  $\{(x_{i1}, x_{i2}) : i = 1, \dots, n\}$ ).

We already have a relationship between  $\tilde{\beta}_1$  and the multiple regression estimator,  $\hat{\beta}_1$ :

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

where  $\hat{\beta}_2$  is the multiple regression estimator of  $\beta_2$  and  $\tilde{\delta}_1$  is the slope coefficient in the auxiliary regression

$$x_{i2} \text{ on } x_{i1}, i = 1, \dots, n$$

Now just use the fact that  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are unbiased (or would be if we could compute them): Conditional on the sample values of  $x_1$  and  $x_2$ ,

$$\begin{aligned} E(\tilde{\beta}_1) &= E(\hat{\beta}_1) + E(\hat{\beta}_2)\tilde{\delta}_1 \\ &= \beta_1 + \beta_2\tilde{\delta}_1 \end{aligned}$$

Therefore,

$$\text{Bias}(\tilde{\beta}_1) = \beta_2\tilde{\delta}_1$$

Recall that  $\tilde{\delta}_1$  has the same sign as the sample correlation  $\text{Corr}(x_{i1}, x_{i2})$ .

The simple regression estimator is unbiased (for the given outcomes  $\{(x_{i1}, x_{i2})\}$ ) in two cases.

1.  $\beta_2 = 0$ . But this means that  $x_2$  does not appear in the population model, so simple regression is the right thing to do.

2.  $\text{Corr}(x_{i1}, x_{i2}) = 0$  (in the sample). Then the simple and multiple regression estimators are identical because  $\tilde{\delta}_1 = 0$ .

If  $\beta_2 \neq 0$  and  $\text{Corr}(x_{i1}, x_{i2}) \neq 0$  then  $\tilde{\beta}_1$  is generally biased. We do know  $\beta_2$  and might only have a vague idea about the size of  $\tilde{\delta}_1$ . But we often can guess at the signs.

Technically, the bias computed holds for a particular “sample” on  $(x_1, x_2)$ . But acting as if what matters is correlation between  $x_1$  and  $x_2$  in the population gives us the correct answer when we turn to asymptotic analysis.

In what follows, we do not make the distinction between the sample and population correlation between  $x_1$  and  $x_2$ .

## Bias in the Simple Regression Estimator of $\beta_1$

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive Bias	Negative Bias
$\beta_2 < 0$	Negative Bias	Positive Bias

# The Variance of the OLS Estimators

So far, we have assumed

MLR.1:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$

MLR.2: random sampling from the population

MLR.3: no perfect collinearity in the sample

MLR.4:  $E(u|x_1, x_2, \dots, x_k) = 0$

Under MLR.3 we can compute the OLS estimates in our sample.

The other assumptions then ensure that OLS is unbiased (conditional on the outcomes of the explanatory variables). When we have omitted an important variable, we have derived that OLS is biased, and we have shown how to obtain the sign of the bias in simple cases.

As in the simple regression case, to obtain  $Var(\hat{\beta}_j)$  we add a simplifying assumption: homoskedasticity (constant variance).

# The Variance of the OLS Estimators

## Assumption MLR.5 (Homoskedasticity)

The variance of the error,  $u$ , does not change with any of  $x_1, x_2, \dots, x_k$ :

$$\text{Var}(u|x_1, x_2, \dots, x_k) = \text{Var}(u) = \sigma^2$$

This assumption can never be guaranteed. We make it for now to get simple formulas, and to be able to discuss efficiency of OLS. Assumptions MLR.1 and MLR.4 imply

$$E(y|x_1, x_2, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

and when we add MLR.5,

$$\text{Var}(y|x_1, x_2, \dots, x_k) = \text{Var}(u|x_1, x_2, \dots, x_k) = \sigma^2$$

Assumptions MLR.1 through MLR.5 are called the **Gauss Markov assumptions**.

# The Variance of the OLS Estimators

If we have a savings equation,

$$sav = \beta_0 + \beta_1 inc + \beta_2 famsize + \beta_3 pareduc + u$$

where *famsize* is size of the family and *pareduc* is total parents' education, MLR.5 means that the variance in *sav* cannot depend in income, family size, or parents's education.

To set up the following theorem, we focus only on the slope coefficients. (A different formula is needed for  $Var(\hat{\beta}_0)$ ).

As before, we are computing the variance conditional on the values of the explanatory variables in the sample.

We need to define two quantities associated with each  $x_j$ . The first is the total variation in  $x_j$  in the sample:

$$SST_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

( $SST_j/n$  is the sample variance of  $x_j$ .)

# The Variance of the OLS Estimators

The second is a measure of correlation between  $x_j$  and the other explanatory variables, in the sample. This is the  $R$ -squared from the regression

$$x_{ij} \text{ on } x_{i1}, x_{i2}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{ik}$$

That is, we regress  $x_j$  on all of the *other* explanatory variables. ( $y$  plays no role here). Call this  $R$ -squared  $R_j^2$ ,  $j = 1, \dots, k$ .

Important:  $R_j^2 = 1$  is ruled out by Assumption MLR.3 because  $R_j^2 = 1$  means that, in the sample,  $x_j$  is an exact linear function of the other explanatory variables.

Any value  $0 \leq R_j^2 < 1$  is permitted. As  $R_j^2$  gets closer to one,  $x_j$  is more linearly related to the other independent variables.



## The Variance of the OLS Estimators

### **THEOREM (Sampling Variances of OLS Slope Estimators)**

Under Assumptions MLR.1 to MLR.5, and condition on the values of the explanatory variables in the sample,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, j = 1, 2, \dots, k.$$

All five Gauss-Markov assumptions are needed to ensure this formula is correct.

Suppose  $k = 3$ ,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$E(u|x_1, x_2, x_3) = 0$$

$$\text{Var}(u|x_1, x_2, x_3) = \gamma_0 + \gamma_1 x_1$$

where  $x_1 \geq 0$  (as are  $\gamma_0$  and  $\gamma_1$ ). This violates MLR.5, and the standard variance formula is generally incorrect for *all* OLS estimators, not just  $\text{Var}(\hat{\beta}_1)$ .

# The Variance of the OLS Estimators

The variance

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

has three components.  $\sigma^2$  and  $SST_j$  are familiar from simple regression. The third,  $1 - R_j^2$ , is new to multiple regression.

What Factors Affect  $\text{Var}(\hat{\beta}_j)$ ?

1. As the error variance (in the population),  $\sigma^2$ , decreases,  $\text{Var}(\hat{\beta}_j)$  decreases. One way to reduce the error variance is to take more stuff out of the error. That is, add more explanatory variables.
2. As the total sample variation in  $x_j$ ,  $SST_j$ , increases,  $\text{Var}(\hat{\beta}_j)$  decreases. As in the simple regression case, it is easier to estimate how  $x_j$  affects  $y$  if we see more variation in  $x_j$ .

# The Variance of the OLS Estimators

As we mentioned earlier,  $SST_j/n$  [or  $SST_j(n-1)$  – the difference is unimportant here] is the sample variance of  $\{x_{ij} : i = 1, 2, \dots, n\}$ . So we can assume

$$SST_j \approx n\sigma_j^2$$

where  $\sigma_j^2 > 0$  is the population variance of  $x_j$ .

We can increase  $SST_j$  by increasing the sample size.  $SST_j$  is roughly a linear function of  $n$ . [Of the three components in  $\text{Var}(\hat{\beta}_j)$ , this is the only one that depends systematically on  $n$ .]

# The Variance of the OLS Estimators

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

3. As  $R_j^2 \rightarrow 1$ ,  $\text{Var}(\hat{\beta}_j) \rightarrow \infty$ .  $R_j^2$  measures how linearly related  $x_j$  is to the other explanatory variables.

We get the smallest variance for  $\hat{\beta}_j$  when  $R_j^2 = 0$ :

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j},$$

which looks just like the simple regression formula.

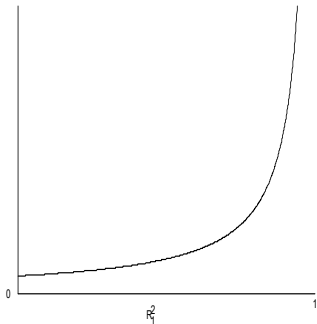
## The Variance of the OLS Estimators

If  $x_j$  is unrelated to all other independent variables, it is easier to estimate its ceteris paribus effect on  $y$ .

$R_j^2 = 0$  is very rare. Even small values are not especially common.

In fact,  $R_j^2 \approx 1$  is somewhat common, and this can cause problems for getting a sufficiently precise estimate of  $\beta_j$ .

Below is a graph of  $\text{Var}(\hat{\beta}_1)$  as a function of  $R_1^2$ :



# The Variance of the OLS Estimators

Loosely,  $R_j^2$  “close” to one is called the “problem” of **multicollinearity**. Unfortunately, we cannot define what we mean by “close” that is relevant for all situations. We have ruled out the case of perfect collinearity,  $R_j^2 = 1$ .

Here is an important point: One often hears discussions of multicollinearity as if high correlation among two or more of the  $x_j$  is a violation of an assumption we have made.

But it *does not* violate any of the Gauss-Markov assumptions, including MLR.3.

# The Variance of the OLS Estimators

We know that if the zero conditional mean assumption is violated, OLS is not unbiased. If MLR.1 through MLR.4 hold, but homoskedasticity (constant variance) does not, then

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

is not the correct formula.

But multicollinearity does not cause the OLS estimators to be biased. We still have  $E(\hat{\beta}_j) = \beta_j$ .

# The Variance of the OLS Estimators

Further, any claim that the OLS variance formula is “biased” in the presence of multicollinearity is also wrong. The formula is correct under MLR.1 through MLR.5.

In fact, the formula is doing its job: It shows that if  $R_j^2$  is “close” to one,  $\text{Var}(\hat{\beta}_j)$  might be very large. If  $R_j^2$  is “close” to one,  $x_j$  does not have much sample variation separate from the other explanatory variables. We are trying to estimate the effect of  $x_j$  on  $y$ , holding  $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k$  fixed, but the data might not be allowing us to do that very precisely.

Because multicollinearity violates none of our assumptions, it is essentially impossible to state hard rules about when it is a “problem.” This has not stopped some from trying.



## The Variance of the OLS Estimators

Other than just looking at the  $R_j^2$ , a common “measure” of multicollinearity is called the **variance inflation factor (VIF)**:

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Because


$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j} \cdot VIF_j,$$

the  $VIF_j$  tells us how many times larger the variance is than if we had the “ideal” case of no correlation of  $x_{ij}$  with

$x_{j1}, \dots, x_{j,j-1}, x_{j,j+1}, \dots, x_{jk}$ .

This sometimes leads to silly rules-of-thumb. For example, one should be “concerned” if  $VIF_j > 10$  (equivalently,  $R_j^2 > .9$ ).

Is  $R_j^2 > .9$  “large”? Yes, in the sense that it would be better to have  $R_j^2$  smaller.

But, if we want to control for, say,  $x_2, \dots, x_k$  to get a good ceteris paribus effect of  $x_1$  on  $y$ , we often have no choice. 

# The Variance of the OLS Estimators

A large  $VIF_j$  can be offset by a large  $SST_j$ :

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j} \cdot VIF_j$$

Remember,  $SST_j$  grows roughly linearly with the sample size,  $n$ . A large  $VIF_j$  can be offset by a large sample size. The value of  $VIF_j$  *per se* is irrelevant. Ultimately, it is  $\text{Var}(\hat{\beta}_j)$  that is important. Even so, at this point, we have no way of knowing whether  $\text{Var}(\hat{\beta}_j)$  is “too large” for the estimate  $\hat{\beta}_j$  to be useful. Only when we discuss confidence intervals and hypothesis testing will this be apparent.

# The Variance of the OLS Estimators

Be wary of work that reports a set of multicollinearity “diagnostics” and concludes nothing useful can be learned because multicollinearity is “too severe.” Sometimes a *VIF* of about 10 is used to make such a claim.

Other “diagnostics” are even more difficult to interpret. Using them indiscriminately is often a mistake.

Consider an example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

where  $\beta_1$  is the coefficient of interest. In fact, assume  $x_2$  and  $x_3$  act as controls so that we hope to get a good ceteris paribus estimate of  $x_1$ . Such controls are often highly correlated. (For example,  $x_2$  and  $x_3$  could be different standardized test scores.)

The key is that the correlation between  $x_2$  and  $x_3$  has nothing to do with  $\text{Var}(\hat{\beta}_1)$ . It is only correlation of  $x_1$  with  $(x_2, x_3)$  that matters.

# The Variance of the OLS Estimators

In an example to determine whether communities with larger minority populations are discriminated against in lending, we might have

$$\begin{aligned} \text{percapproved} = & \beta_0 + \beta_1 \text{percminority} \\ & + \beta_2 \text{avginc} + \beta_3 \text{avghouseval} + u, \end{aligned}$$

where  $\beta_1$  is the key coefficient. We might expect *avginc* and *avghouseval* to be highly correlated across communities. But we do not care really care whether we can precisely estimate  $\beta_2$  or  $\beta_3$ .

# The Variance of the OLS Estimators

As with bias calculations, we can study the variances of the OLS estimators in misspecified models.

Consider the same case with (at most) two explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

which we assume satisfies the Gauss-Markov assumptions.

We run the “short” regression,  $y$  on  $x_1$ , and also the “long” regression,  $y$  on  $x_1, x_2$ :

$$\begin{aligned}\tilde{y} &= \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \\ \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\end{aligned}$$

# The Variance of the OLS Estimators

We know from the previous analysis that

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_1(1 - R_1^2)}$$

conditional on the values  $x_{i1}$  and  $x_{i2}$  in the sample.

What about the simple regression estimator? Can show

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$$

which is again conditional on  $\{(x_{i1}, x_{i2}) : i = 1, \dots, n\}$ .

# The Variance of the OLS Estimators

Whenever  $x_{i1}$  and  $x_{i2}$  are correlated,  $R_1^2 > 0$ , and

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1} < \frac{\sigma^2}{SST_1(1 - R_1^2)} < \text{Var}(\hat{\beta}_1)$$

So, by omitting  $x_2$ , we can in fact get an estimator with a smaller variance, even though it is biased. When we look at bias and variance, we have a tradeoff between simple and multiple regression.

In the case  $R_1^2 > 0$ , we can draw two conclusions.

# The Variance of the OLS Estimators

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

1. If  $\beta_2 \neq 0$ ,  $\tilde{\beta}_1$  is biased,  $\hat{\beta}_1$  is unbiased, but  $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$ .
2. If  $\beta_2 = 0$ ,  $\tilde{\beta}_1$  and  $\hat{\beta}_1$  are both unbiased and  $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$ .

Case 2 is clear cut. If  $\beta_2 = 0$ ,  $x_2$  has no (partial) effect on  $y$ . When  $x_2$  is correlated with  $x_1$ , including it along with  $x_1$  in the regression makes it more difficult to estimate the partial effect of  $x_1$ . Simple regression is clearly preferred.



# The Variance of the OLS Estimators

Case 1 is more difficult, but there are reasons to prefer the unbiased estimator,  $\hat{\beta}_1$ . First, the bias in  $\tilde{\beta}_1$  does not systematically change with the sample size. We should assume the bias is as large when  $n = 1,000$  as when  $n = 10$ . By contrast, the variances  $Var(\tilde{\beta}_1)$  and  $Var(\hat{\beta}_1)$  both shrink at the rate  $1/n$ . With a large sample size, the difference between  $Var(\tilde{\beta}_1)$  and  $Var(\hat{\beta}_1)$  is less important, especially considering the bias in  $\tilde{\beta}_1$  is not shrinking.

# The Variance of the OLS Estimators

Second reason for preferring  $\hat{\beta}_1$  is more subtle. The formulas

$$\begin{aligned} \text{Var}(\tilde{\beta}_1) &= \frac{\sigma^2}{SST_1} \\ \text{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{SST_1(1 - R_1^2)} \end{aligned}$$

because they condition on the same explanatory variables, act as if the error variance does not change when we add  $x_2$ . But if  $\beta_2 \neq 0$ , the variance does shrink.

## Estimating the Error Variance

We still need to estimate  $\sigma^2$ . With  $n$  observations and  $k + 1$  parameters, we only have

$$df = n - (k + 1)$$

degrees of freedom. Recall we lose the  $k + 1$   $df$  due to  $k + 1$  restrictions on the OLS residuals:

$$\sum_{i=1}^n \hat{u}_i = 0$$

$$\sum_{i=1}^n x_{ij} \hat{u}_i = 0, j = 1, 2, \dots, k$$

### **THEOREM: (Unbiased Estimation of $\sigma^2$ )**

Under the Gauss-Markov assumptions (MLR.1 through MLR.5)

$$\hat{\sigma}^2 = (n - k - 1)^{-1} \sum_{i=1}^n \hat{u}_i^2 = SSR/df$$

## Estimating the Error Variance

This means that, if we divide by  $n$  rather than  $n - k - 1$ , the bias is

$$-\sigma^2 \left( \frac{k+1}{n} \right)$$

which means the estimated variance would be too small, on average. The bias disappears as  $n$  increases.

The square root of  $\hat{\sigma}^2$ ,  $\hat{\sigma}$ , is reported by all regression packages. (**standard error of the regression**, or **root mean squared error**). Note that  $SSR$  falls when a new explanatory variable is added, but  $df$  falls, too. So  $\hat{\sigma}$  can increase or decrease when a new variable is added in multiple regression.

## Estimating the Error Variance

The **standard error** of each  $\hat{\beta}_j$  is computed (for the slopes) as

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1 - R_j^2)}}$$

and it will be critical to report these along with the coefficient estimates.

# Efficiency of OLS: The Gauss-Markov Theorem

How come we use OLS, rather than some other estimation method?

Consider simple regression:

$$y = \beta_0 + \beta_1 x + u$$

and write, for each  $i$ ,

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

If we average across the  $n$  observations we get

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$$

# Efficiency of OLS: The Gauss-Markov Theorem

For any  $i$  with  $x_i \neq \bar{x}$ , subtract and rearrange:

$$\beta_1 = \frac{(y_i - \bar{y})}{(x_i - \bar{x})} + \frac{(u_i - \bar{u})}{(x_i - \bar{x})}$$

The last term has a zero expected value under random sampling and  $E(u|x) = 0$ . If  $x_i \neq \bar{x}$  for all  $i$ , we could use an estimator

$$\check{\beta}_1 = n^{-1} \sum_{i=1}^n \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

## Efficiency of OLS: The Gauss-Markov Theorem

- $\check{\beta}_1$  is not the same as the OLS estimator,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

How do we know OLS is better than this new estimator,  $\check{\beta}_1$ ?

Generally, we do not. Under MLR.1 to MLR.4, both estimators are unbiased.

Comparing their variances is difficult in general.

But, if we add the homoskedasticity assumption, MLR.5, we can show

$$\text{Var}(\hat{\beta}_1) < \text{Var}(\check{\beta}_1)$$



# Efficiency of OLS: The Gauss-Markov Theorem

This means  $\hat{\beta}_1$  has a sampling distribution that is less spread out around  $\beta_1$  than  $\check{\beta}_1$ . When comparing unbiased estimators, we prefer an estimator with smaller variance.

We can make very general statements for the multiple regression case, provided the 5 Gauss-Markov assumptions hold.

However, we must also limit the class of estimators that we can compare with OLS.

# Efficiency of OLS: The Gauss-Markov Theorem

## THEOREM (Gauss-Markov)

Under Assumptions MLR.1 through MLR.5, the OLS estimators  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  are the **best linear unbiased estimators (BLUEs)**

Start from the end of “BLUE” and work backwards:

**E(estimator)**: We must be able to compute an estimate from the observable data, using a fixed rule.

# Efficiency of OLS: The Gauss-Markov Theorem

**L (linear):** The estimator is a linear function of  $\{y_i : i = 1, 2, \dots, n\}$ . It can be a nonlinear function of the explanatory variables.

These estimators have the general form

$$\tilde{\beta}_j = \sum_{i=1}^n w_{ij} y_i$$

where the  $\{w_{ij} : i = 1, \dots, n\}$  are any functions of  $\{(x_{i1}, \dots, x_{ik}) : i = 1, \dots, n\}$ .

The OLS estimators can be written in this way.

# Efficiency of OLS: The Gauss-Markov Theorem

**U (unbiased):** We must impose enough restrictions on the  $w_{ij}$  – we omit those here – so that

$$E(\tilde{\beta}_j) = \beta_j, j = 0, 1, \dots, k$$

(conditional on  $\{(x_{i1}, \dots, x_{ik}) : i = 1, \dots, n\}$ ).

We know the OLS estimators are unbiased under MLR.1 through MLR.4. So are a lot of other linear estimators.

# Efficiency of OLS: The Gauss-Markov Theorem

**B (best):** This means smallest variance (which makes sense once we impose unbiasedness). In other words, what can be shown is that, under MLR.1 through MLR.5, and conditional on the explanatory variables in the sample,

$$\text{Var}(\hat{\beta}_j) \leq \text{Var}(\tilde{\beta}_j) \text{ all } j$$

(and usually the inequality is strict).

If we do not impose unbiasedness, then we can use silly rules – such as  $\tilde{\beta}_j = 1$  always – to get estimators with zero variance.

# Efficiency of OLS: The Gauss-Markov Theorem

How do we use the GM Theorem? If the Gauss-Markov assumptions hold, and we insist on unbiased estimators that are also linear functions of  $\{y_i : i = 1, 2, \dots, n\}$ , then we need look no further than OLS: it delivers the smallest possible variances.

It might be possible (but even so, not practical) to find unbiased estimators that are nonlinear functions of  $\{y_i : i = 1, 2, \dots, n\}$  that have smaller variances than OLS. The GM Theorem only allows linear estimators in the comparison group.

Appendix 3A in Wooldridge contains a proof of the GM Theorem. If MLR.5 fails, that is,  $\text{Var}(u|x_1, \dots, x_k)$  depends on one or more  $x_j$ , the GM conclusions do not hold. There may be linear, unbiased estimators of the  $\beta_j$  with smaller variance.

# Efficiency of OLS: The Gauss-Markov Theorem

Remember: Failure of MLR.5 does not cause bias in the  $\hat{\beta}_j$ , but it does have two consequences:

1. The usual formulas for  $Var(\hat{\beta}_j)$ , and therefore for  $se(\hat{\beta}_j)$ , are wrong.
2. The  $\hat{\beta}_j$  are no longer BLUE.

The first of these is more serious, as it will directly affect statistical inference (next). The second consequence means we may want to search for estimators other than OLS. This is not so easy. And with a large sample, it may not be very important.