## Introduction: OLS Review

- Derivation of the OLS estimator
- Algebraic properties of OLS
- Statistical Properties of OLS
- Variance of OLS and standard errors

## Foundations of scientific knowledge

Scientific methodologies are the epistemological foundation of scientific knowledge

- Science does not collect evidence in order to "prove" what people already believe or want others to believe
- Science accepts unexpected and even undesirable answers
- Science is process oriented, not outcome oriented

Terminology

| $y$ | $x$ |
|---|---|
| Dependent Variable | Independent Variable |
| Explained Variable | Explanatory Variable |
| Response Variable | Control Variable |
| Predicted Variable | Predictor Variable |
| Regressand | Regressor |
| LHS | RHS |

The terms "explained" and "explanatory" are probably best, as they are the most descriptive and widely applicable. But "dependent" and "independent" are used often. (The "independence" here is not really statistical independence.)

We said we must confront three issues:

1. How do we allow factors other than $x$ to affect $y$?
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 $x$?

We will argue that the simple regression model

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

addresses each of them.

## Simple linear regression model

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

## The error term

We make a simplifying assumption (without loss of generality): the average, or expected, value of $u$ is zero in the population:

$$E(u) = 0 \tag{2}$$

where $E(\cdot)$ is the expected value operator.

## The intercept

The presence of $\beta_0$ in

$$y = \beta_0 + \beta_1 x + u \tag{3}$$

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

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0) \tag{4}$$

where $\alpha_0 = E(u)$. The new error is $u - \alpha_0$ and the new intercept is $\beta_0 + \alpha_0$. The important point is that the slope, $\beta_1$, has not changed.

## Mean independence of the error term

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 $x$:

$$E(u|x) = E(u), \text{ all values } x \tag{5}$$

where $E(u|x)$ means "the expected value of $u$ given $x$".
Then, we say $u$ is **mean independent** of $x$.

## Distribution of ability across education

- Suppose $u$ is "ability" and $x$ is years of education. We need, for example,

$$E(ability|x = 8) = E(ability|x = 12) = E(ability|x = 16)$$

  so that the average ability is the same in the different portions of the population with an $8^{th}$ grade education, a $12^{th}$ grade education, and a four-year college education.
- Because people choose education levels partly based on ability, this assumption is almost certainly false.

## Zero conditional mean assumption

Combining $E(u|x) = E(u)$ (the substantive assumption) with $E(u) = 0$ (a normalization) gives the **zero conditional mean assumption**.

$$E(u|x) = 0, \text{ all values } x \tag{6}$$

## Population regression function

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

$$E(y|x) = \beta_0 + \beta_1 x \tag{7}$$

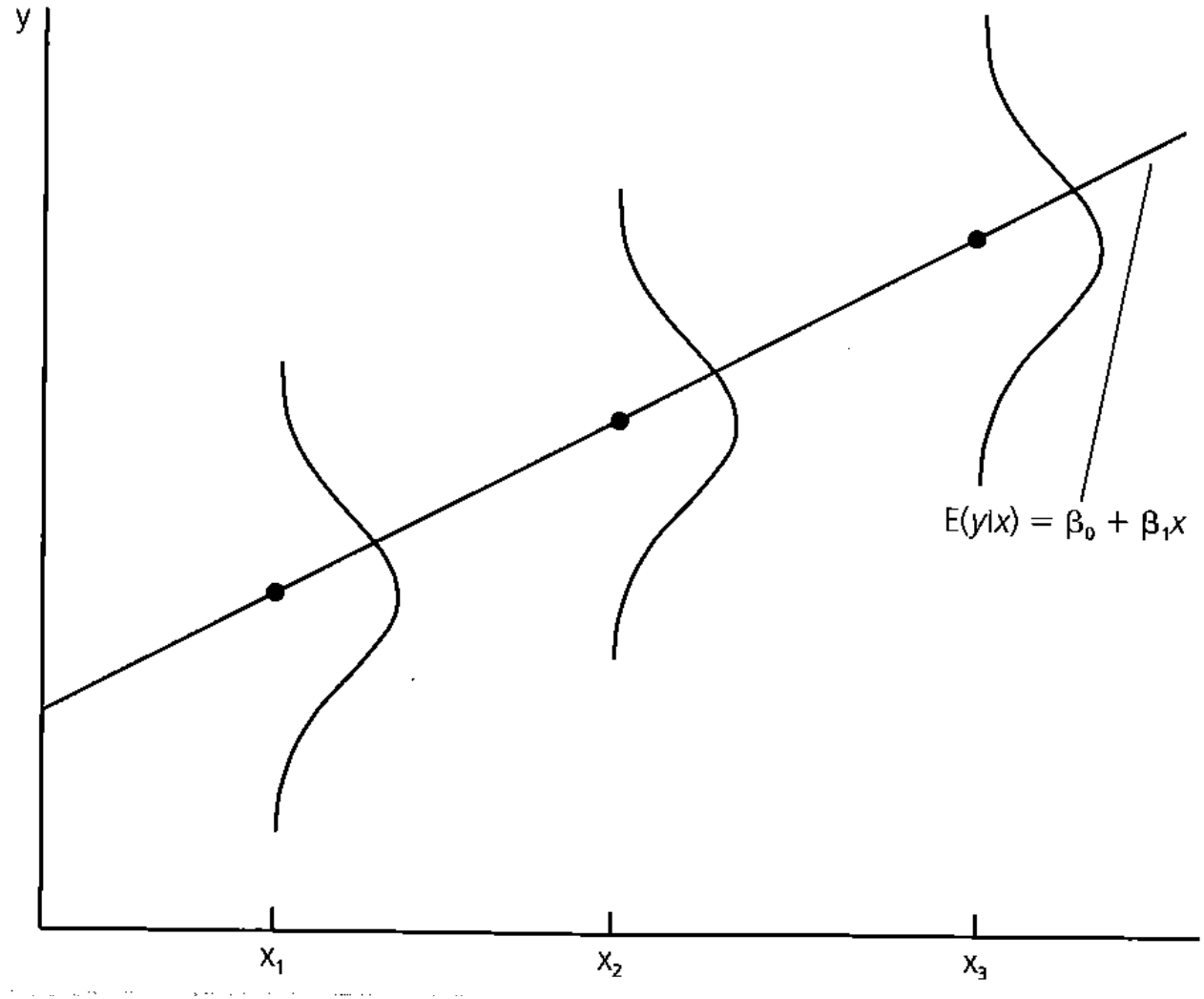
which shows the **population regression function** is a linear function of $x$.

- The straight line in the graph on the next page is what Wooldridge calls the **population regression function**, and what Angrist and Pischke call the **conditional expectation function**

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

- The conditional distribution of $y$ at three different values of $x$ are superimposed. for a given value of $x$, we see a range of $y$ values: remember, $y = \beta_0 + \beta_1 x + u$, and $u$ has a distribution in the population.

$E(y|x)$ as a linear function of $x$.



## Deriving the Ordinary Least Squares Estimates

- Given data on $x$ and $y$, how can we estimate the population parameters, $\beta_0$ and $\beta_1$?
- Let $\{(x_i, y_i) : i = 1, 2, ..., n\}$ be a **random** sample of size $n$ (the number of observations) from the population.
- Plug any observation into the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{8}$$

  where the $i$ subscript indicates a particular observation.
- We observe $y_i$ and $x_i$, but not $u_i$ (but we know it is there).

We use the two population restrictions:

$$E(u) = 0$$
$$Cov(x, u) = 0$$

to obtain estimating equations for $\beta_0$ and $\beta_1$. We talked about the first condition. The second condition means that $x$ and $u$ are uncorrelated. Both conditions are implied by $E(u|x) = 0$

With $E(u) = 0$, $Cov(x, u) = 0$ is the same as $E(xu) = 0$. Next we plug in for $u$:

$$E(y - \beta_0 - \beta_1 x) = 0$$
$$E[x(y - \beta_0 - \beta_1 x)] = 0$$

These are the two conditions in the **population** that effectively determine $\beta_0$ and $\beta_1$.

So we use their sample counterparts (which is a method of moments approach to estimation):

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

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

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

Pass the summation operator through the first equation:

$$n^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{9}$$

$$= n^{-1}\sum_{i=1}^{n} y_i - n^{-1}\sum_{i=1}^{n}\hat{\beta}_0 - n^{-1}\sum_{i=1}^{n}\hat{\beta}_1 x_i \tag{10}$$

$$= n^{-1}\sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1\left(n^{-1}\sum_{i=1}^{n} x_i\right) \tag{11}$$

$$= \overline{y} - \hat{\beta}_0 - \hat{\beta}_1\overline{x} \tag{12}$$

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

We have shown that the first equation,

$$n^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{13}$$

implies

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \tag{14}$$

Now, use this equation to write the intercept in terms of the slope:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{15}$$

Plug this into the second equation (but where we take away the division by $n$):

$$\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{16}$$

so

$$\sum_{i=1}^{n} x_i[y_i - (\overline{y} - \hat{\beta}_1 \overline{x}) - \hat{\beta}_1 x_i] = 0 \tag{17}$$

Simple algebra gives

$$\sum_{i=1}^{n} x_i(y_i - \overline{y}) = \hat{\beta}_1 \left[ \sum_{i=1}^{n} x_i(x_i - \overline{x}) \right] \tag{18}$$

So, the equation to solve is

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \hat{\beta}_1 \left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right] \tag{19}$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\text{Sample Covariance}(x_i, y_i)}{\text{Sample Variance}(x_i)} \tag{20}$$

# OLS

- The previous formula for $\hat{\beta}_1$ is important. It shows us how to take the data we have and compute the slope estimate.
- $\hat{\beta}_1$ is called the **ordinary least squares (OLS)** 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 $x_i$ has the same value.
- The intuition is that the variation in $x$ is what permits us to identify its impact on $y$.

## Solving for $\widehat{\beta}$

- Once we have $\hat{\beta}_1$, we compute $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$. This is the OLS intercept estimate.
- These days, we let the computer do the calculations, which are tedious even if $n$ is small.

## Predicting y

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

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{21}$$

  We have $n$ of these.
- $\hat{y}_i$ is the value we predict for $y_i$ given that $x = x_i$ and $\beta = \hat{\beta}$.

## The residual

- The "mistake" from our *prediction* is called the **residual**:

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \end{aligned}$$

- Suppose we measure the size of the mistake, for each $i$, by squaring it. Then we add them all up to get the **sum of squared residuals**

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

- Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to *minimize* the sum of squared residuals which gives us the same solutions we obtained before.

## Algebraic Properties of OLS Statistics

Remembering how the **first moment** condition allows us to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$, we have:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{22}$$

Notice the logic here: this means the OLS residuals *always* add up to zero, by *construction*,

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{23}$$

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

$$n^{-1}\sum_{i=1}^{n} y_i = n^{-1}\sum_{i=1}^{n} \hat{y}_i + n^{-1}\sum_{i=1}^{n} \hat{u}_i \tag{24}$$

and so $\overline{y} = \overline{\hat{y}}$.

## Second moment

Similarly the way we obtained our estimates,

$$n^{-1} \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{25}$$

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

$$n^{-1} \sum_{i=1}^{n} x_i \hat{u}_i = 0 \tag{26}$$

## Bringing things together

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

$$n^{-1} \sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \tag{27}$$

## Averages

A third property is that the point $(\overline{x}, \overline{y})$ is always on the OLS regression line. That is, if we plug in the average for $x$, we predict the sample average for $y$:

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \tag{28}$$

Again, we chose the estimates to make this true.

## Expected Value of OLS

- 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.
- Leads to the notion of **unbiasedness**, which is a "desirable" characteristic for estimators.

$$E(\hat{\beta}) = \beta \tag{29}$$

## Don't forget why we're here

- Plato's allegory of the cave - reality is outside the cave, the reflections on the wall are our estimates of that reality.
- The **population** parameter that describes the relationship between $y$ and $x$ is $\beta_1$
- For this class, $\beta_1$ is a causal parameter, and our sole objective is to estimate $\beta_1$ with a sample of data
- But never forget that $\hat{\beta}_1$ is an **estimator** of that causal parameter obtained with a *specific* sample from the population.

## Uncertainty and sampling variance

- Different samples will generate different estimates ($\hat{\beta}_1$) for the "true" $\beta_1$ which makes $\hat{\beta}_1$ a random variable.
- Unbiasedness is the idea that if we could take as many random samples on $Y$ as we want from the population, and compute an estimate each time, the average of these estimates would be equal to $\beta_1$.
- But, this also implies that $\hat{\beta}_1$ has spread and therefore variance

# Assumptions

**Assumption SLR.1 (Linear in Parameters)**

- The population model can be written as

$$y = \beta_0 + \beta_1 x + u \tag{30}$$

  where $\beta_0$ and $\beta_1$ are the (unknown) population parameters.
- We view $x$ and $u$ as outcomes of random variables; thus, $y$ is random.
- Stating this assumption formally shows that our goal is to estimate $\beta_0$ and $\beta_1$.

**Assumption SLR.2 (Random Sampling)**

- We have a random sample of size $n$, $\{(x_i, y_i) : i = 1, ..., n\}$, following the population model.
- We know how to use this data to estimate $\beta_0$ and $\beta_1$ by OLS.
- Because each $i$ is a draw from the population, we can write, for each $i$,

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{31}$$

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

**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 $\{x_i : i = 1, ..., n\}$ is not zero.
- In practice, this is no assumption at all. If the $x_i$ are all the same value, we cannot learn how $x$ affects $y$ in the population.

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

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

$$E(u|x) = E(u) = 0. \tag{32}$$

- 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.

## Showing OLS is unbiased

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

$$E(\hat{\beta}_1) = \beta_1 \tag{33}$$

where the expected value means averaging across random samples.
**Step 1:** Write down a formula for $\hat{\beta}_1$. It is convenient to use

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{34}$$

which is one of several equivalent forms.

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{SST_x} \tag{35}$$

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

**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

$$\sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i + u_i) \tag{36}$$

$$= \beta_0 \sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1 \sum_{i=1}^{n}(x_i - \overline{x})x_i + \sum_{i=1}^{n}(x_i - \overline{x})u_i \tag{37}$$

$$= 0 + \beta_1 \sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(x_i - \overline{x})u_i \tag{38}$$

$$= \beta_1 SST_x + \sum_{i=1}^{n}(x_i - \overline{x})u_i \tag{39}$$

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

We have shown

$$\hat{\beta}_1 = \frac{\beta_1 SST_x + \sum_{i=1}^{n}(x_i - \overline{x})u_i}{SST_x} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})u_i}{SST_x} \quad (40)$$

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

Now define

$$w_i = \frac{(x_i - \overline{x})}{SST_x} \tag{41}$$

so we have

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i \tag{42}$$

- $\hat{\beta}_1$ is a linear function of the unobserved errors, $u_i$. The $w_i$ are all functions of $\{x_1, x_2, ..., x_n\}$.
- The (random) difference between $\hat{\beta}_1$ and $\beta_1$ is due to this linear function of the unobservables.

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

- Under Assumptions SLR.2 and SLR.4, $E(u_i|x_1, x_2, ..., x_n) = 0$. That means, *conditional* on $\{x_1, x_2, ..., x_n\}$,

$$E(w_i u_i|x_1, x_2, ..., x_n) = w_i E(u_i|x_1, x_2, ..., x_n) = 0$$

  because $w_i$ is a function of $\{x_1, x_2, ..., x_n\}$. (In the next slides I omit the conditioning in the expectations)
- This would not be true if, in the population, $u$ and $x$ are correlated.

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

$$E(\hat{\beta}_1) = E\left(\beta_1 + \sum_{i=1}^{n} w_i u_i\right) \tag{43}$$

$$= \beta_1 + \sum_{i=1}^{n} E(w_i u_i) = \beta_1 + \sum_{i=1}^{n} w_i E(u_i) \tag{44}$$

$$= \beta_1 \tag{45}$$

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.

**THEOREM (Unbiasedness of OLS)**
Under Assumptions SLR.1 through SLR.4

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1. \tag{46}$$

- Omit the proof for $\hat{\beta}_0$.

- Each sample leads to a different estimate, $\hat{\beta}_0$ and $\hat{\beta}_1$. Some will be very close to the true values $\beta_0 = 3$ and $\beta_1 = 2$. Nevertheless, some could be very far from those values.
- If we repeat the experiment again and again, and average the estimates, we would get very close to 2.
- The problem is, we do not know which kind of sample we have. We can never know whether we are close to the population value.
- We hope that our sample is "typical" and produces a slope estimate close to $\beta_1$ but we can never know.

## Reminder

- **Errors** are the vertical distances between observations and the **unknown** Conditional Expectation Function. Therefore, they are unknown.
- **Residuals** are the vertical distances between observations and the **estimated** regression function. Therefore, they are known.

## SE and the data

The correct SE estimation procedure is given by the underlying structure of the data

- It is very unlikely that all observations in a dataset are unrelated, but drawn from identical distributions (**homoskedasticity**)
- For instance, the variance of income is often greater in families belonging to top deciles than among poorer families (**heteroskedasticity**)
- Some phenomena do not affect observations individually, but they do affect groups of observations uniformly within each group (**clustered data**)

## 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 the variance (and, ultimately, the standard deviation).
- 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 $x$:

$$Var(u|x) = \sigma^2 > 0 \tag{47}$$

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

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

$$E(u^2|x) = \sigma^2 = E(u^2) \tag{48}$$

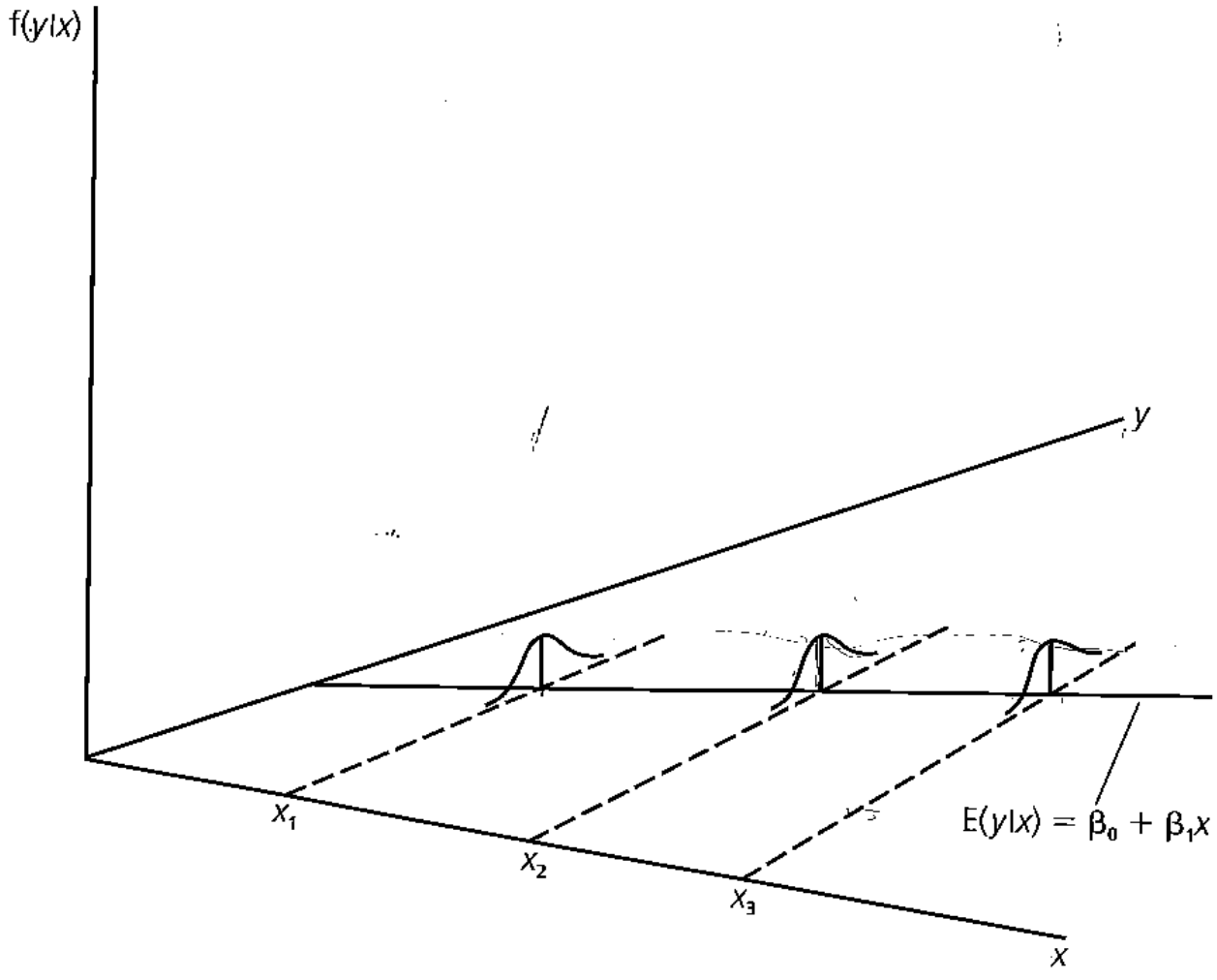
Under the population Assumptions SLR.1 ($y = \beta_0 + \beta_1 x + u$), SRL.4 ($E(u|x) = 0$) and SLR.5 ($Var(u|x) = \sigma^2$),

$$E(y|x) = \beta_0 + \beta_1 x$$
$$Var(y|x) = \sigma^2$$

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 $x$. (See Graphs on next two slides)

**Figure 2.8**

The simple regression model under homoskedasticity.



**THEOREM (Sampling Variances of OLS)**
Under Assumptions SLR.1 to SLR.2,

$$Var(\hat{\beta}_1|x) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma^2}{SST_x}$$

$$Var(\hat{\beta}_0|x) = \frac{\sigma^2 \left(n^{-1}\sum_{i=1}^{n} x_i^2\right)}{SST_x}$$

(conditional on the outcomes $\{x_1, x_2, ..., x_n\}$).

To show this, write, as before,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i \tag{49}$$

where $w_i = (x_i - \overline{x})/SST_x$. We are treating this 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.

The $\{u_i : i = 1, 2, ..., n\}$ are actually independent across $i$, and so they are uncorrelated. So (remember that if we know $x$, we know $w$)

$$Var(\hat{\beta}_1|x) = Var\left(\sum_{i=1}^{n} w_i u_i|x\right)$$

$$= \sum_{i=1}^{n} Var(w_i u_i|x) = \sum_{i=1}^{n} w_i^2 Var(u_i|x)$$

$$= \sum_{i=1}^{n} w_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} w_i^2$$

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

Now we have

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{(SST_x)^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{(SST_x)^2}$$
$$= \frac{SST_x}{(SST_x)^2} = \frac{1}{SST_x}$$

We have shown

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \tag{50}$$

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

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \tag{51}$$

1. As the error variance increases, i.e, as $\sigma^2$ increases, so does $Var(\hat{\beta}_1)$. 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:

$$SST_x \uparrow \text{ implies } Var(\hat{\beta}_1) \downarrow \tag{52}$$

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

$$SST_x \approx n\sigma_x^2 \tag{53}$$

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

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

$$sd(\hat{\beta}_1) = \frac{\sigma}{\sqrt{SST_x}} \tag{54}$$

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

**Estimating the Error Variance**

In the formula

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \tag{55}$$

we can compute $SST_x$ from $\{x_i : i = 1, ..., n\}$. But we need to estimate $\sigma^2$.

Recall that

$$\sigma^2 = E(u^2). \tag{56}$$

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

$$n^{-1} \sum_{i=1}^{n} u_i^2 \tag{57}$$

But this not an estimator because we cannot compute it from the data we observe, since $u_i$ are unobserved.
How about replacing each $u_i$ with its "estimate", the OLS residual $\hat{u}_i$?

$$u_i = y_i - \beta_0 - \beta_1 x_i$$
$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

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

$$\hat{u}_i \neq u_i \tag{58}$$

for any $i$.

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

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

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

$$n^{-1} \sum_{i=1}^{n} \hat{u}_i^2 = SSR/n \tag{59}$$

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 a little 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.

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

$$\hat{\sigma}^2 = \frac{SSR}{(n-2)} \tag{60}$$

**THEOREM: Unbiased Estimator of $\sigma^2$**
Under Assumptions SLR.1 to SLR.5,

$$E(\hat{\sigma}^2) = \sigma^2 \tag{61}$$

In regression output, it is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{SSR}{(n-2)}} \tag{62}$$

that is usually reported. This is an estimator of $sd(u)$, the standard deviation of the population error. And $SSR = \sum_{i=1}^{n} \widehat{u}^2$.

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

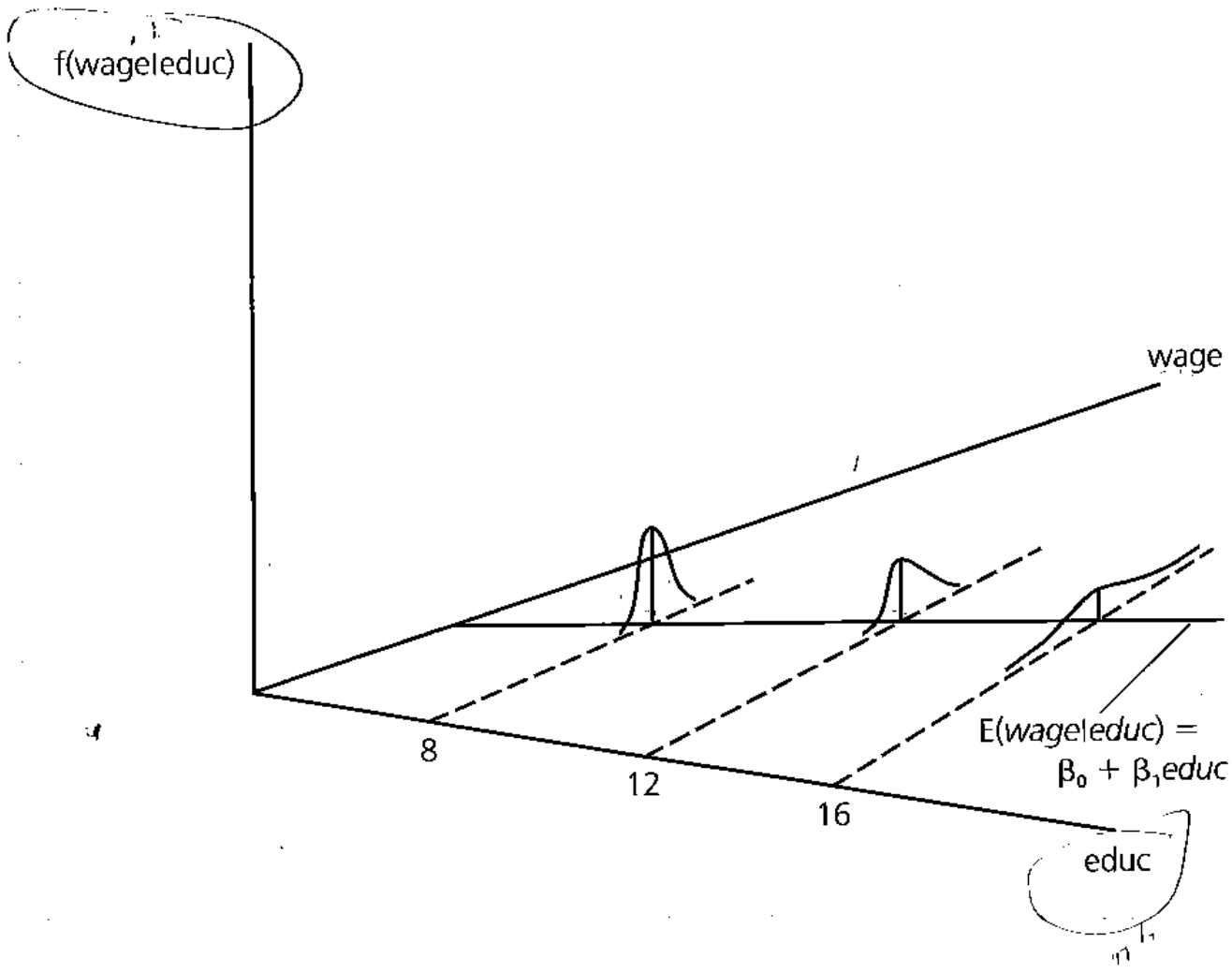
$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}} \tag{63}$$

  where both the numerator and denominator are computed from the data.
- For reasons we will see, it is useful to report the standard errors below the corresponding coefficient, usually in parentheses.

- OLS inference is generally faulty in the presence of heteroskedasticity

**Figure 2.9**

Var (*wage*|*educ*) increasing with *educ*.



- Fortunately, OLS is still useful
- Assume SLR.1-4 hold, but not SLR.5. Therefore

$$Var(u_i|x_i) = \sigma_i^2$$

- The variance of our estimator, $\widehat{\beta_1}$ equals:

$$Var(\widehat{\beta_1}) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2\sigma_i^2}{SST_x^2}$$

- When $\sigma_i^2 = \sigma^2$ for all $i$, this formula reduces to the usual form, $\frac{\sigma^2}{SST_x^2}$

- A valid estimator of Var($\widehat{\beta}_1$) for heteroskedasticity of any form (including homoskedasticity) is

$$Var(\widehat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \widehat{u}_i^2}{SST_x^2}$$

  which is easily computed from the data after the OLS regression
- As a rule, you should always use the `, robust` command in STATA.

## Clustered data

- But what if errors are not iid?
- For instance, maybe observations between units in a group are related to each other
  - You want to regress kids' grades on class size to determine the effect of class size on grades
  - The **unobservables** of kids belonging to the same classroom will be correlated (e.g., teacher quality, recess routines) while will not be correlated with kids in far away classrooms
- Then i.i.d. is violated. But maybe i.i.d. holds across clusters, just not within clusters

## Simulations

- Let's first try to understand what's going on with a few simulations
- We will begin with a baseline of non-clustered data
- We'll show the distribution of estimates in Monte Carlo simulation for 1000 draws and iid errors
- We'll then show the number of times you reject the null incorrectly at $\alpha = 0.05$.

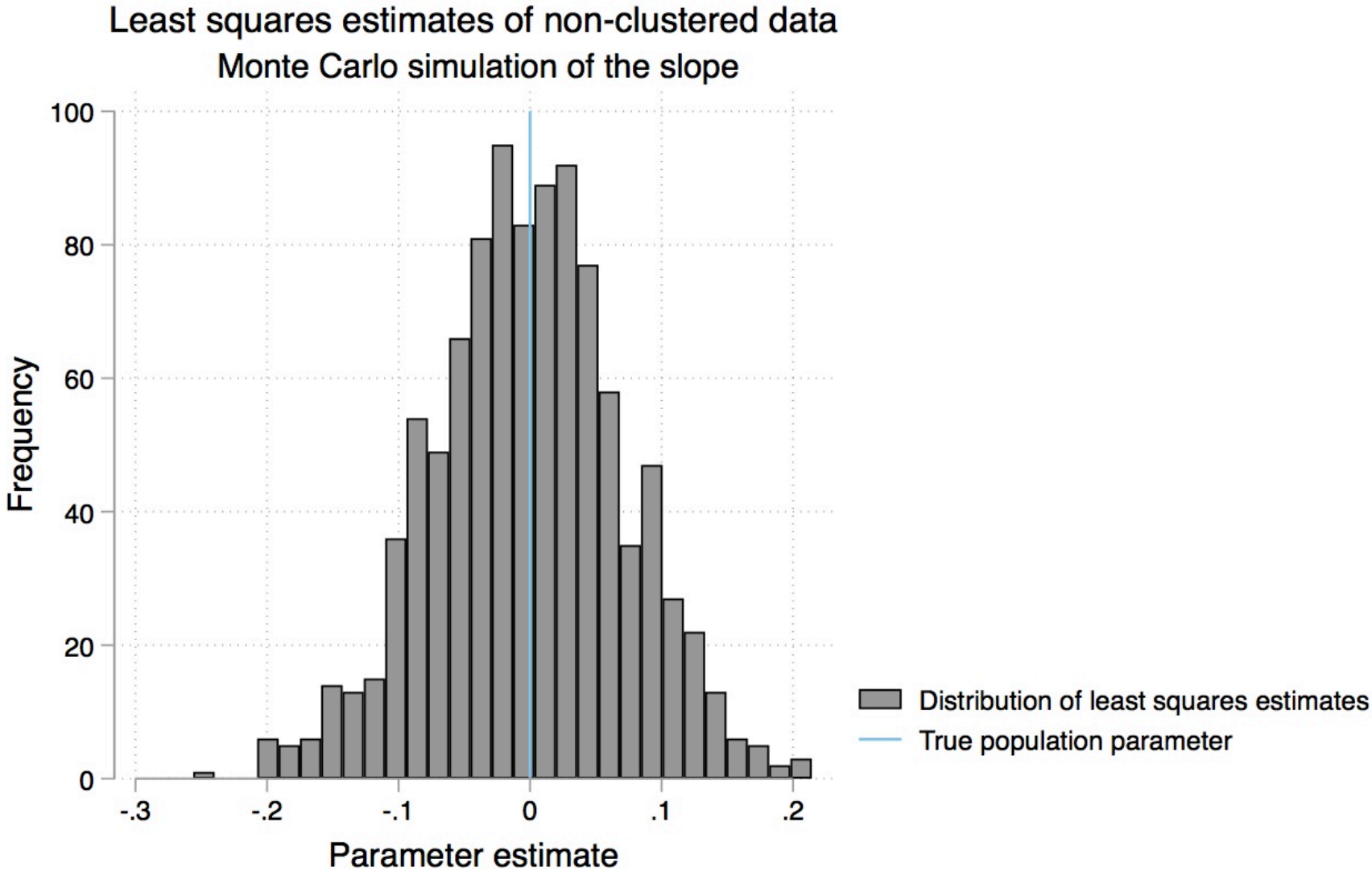



Figure: Distribution of the least squares estimator over 1,000 random draws.

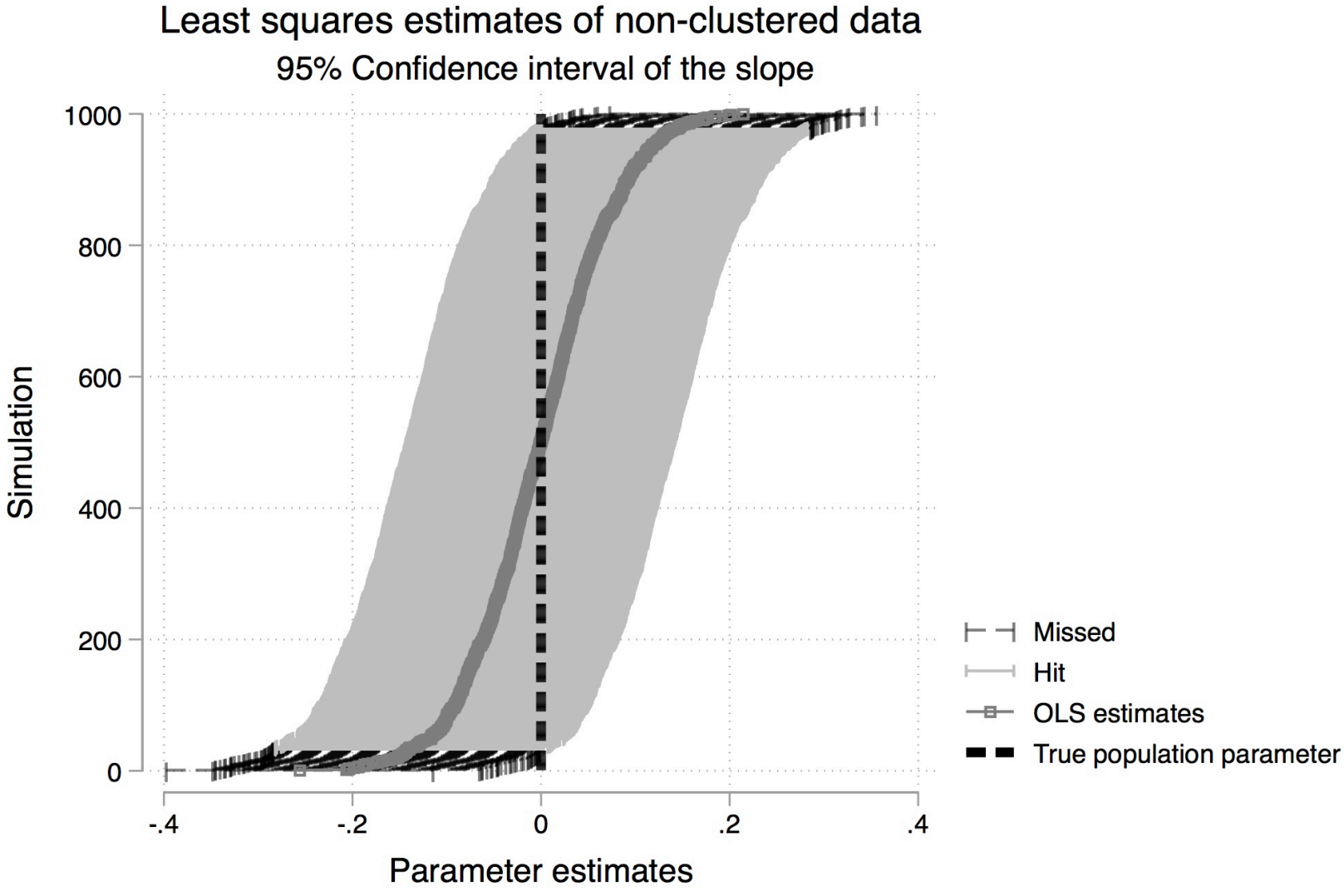



Figure: Distribution of the 95% confidence intervals with coloring showing those which are incorrectly rejecting the null.

## Clustered data and heteroskedastic robust

- Now let's look at clustered data
- But this time we will estimate the model using heteroskedastic robust standard errors
- Earlier we saw mass all the way to -2.5 to 2; what do we get when we incorrectly estimate the standard errors?

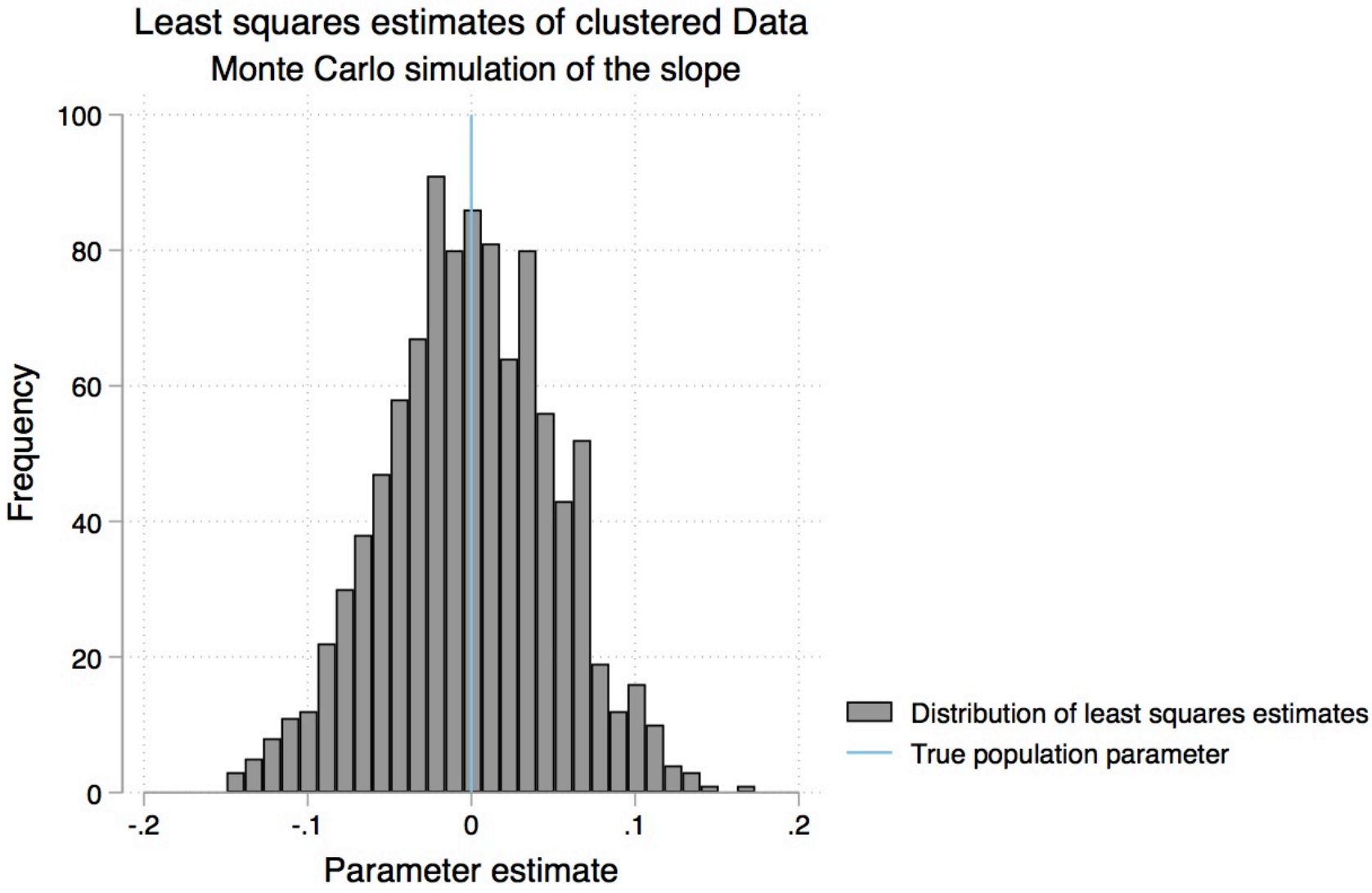



Figure: Distribution of the least squares estimator over 1,000 random draws. Clustered data without correcting for clustering

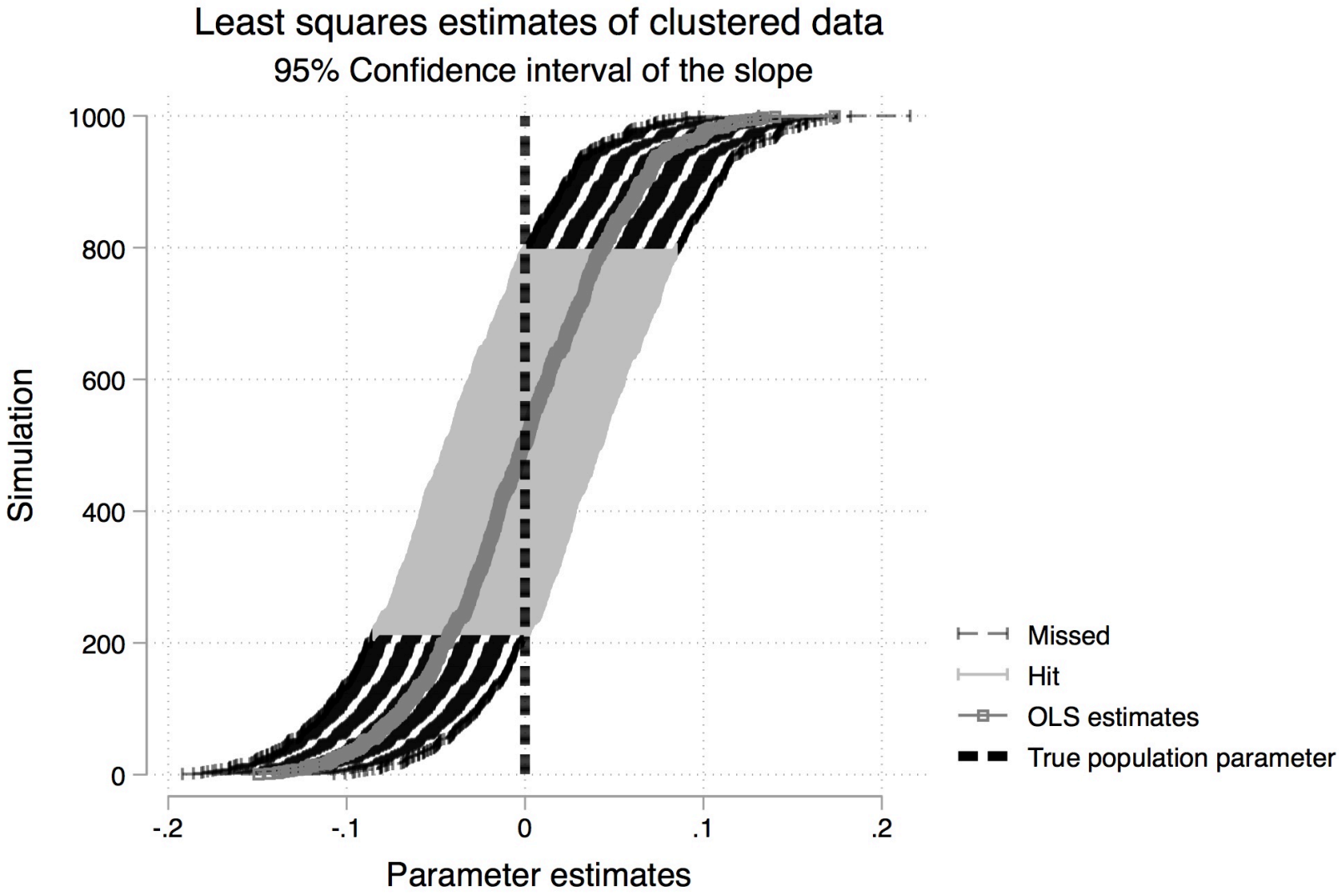



Figure: Distribution of 1,000 95% confidence intervals with dashed region representing those estimates that incorrectly reject the null.

## Over-rejecting the null

- Those 95 percent confidence intervals are based on an $\alpha = 0.05$.
- Look how many parameter estimates are different from zero; that's what we mean by "over-rejecting the null"
- You saw signs of it though in the variance of the estimated effect, bc the spread only went from -.15 to .15 (whereas earlier it had gone from -.25 to .2)
- Now let's correct for arbitrary within group correlations using the cluster robust option in Stata/R

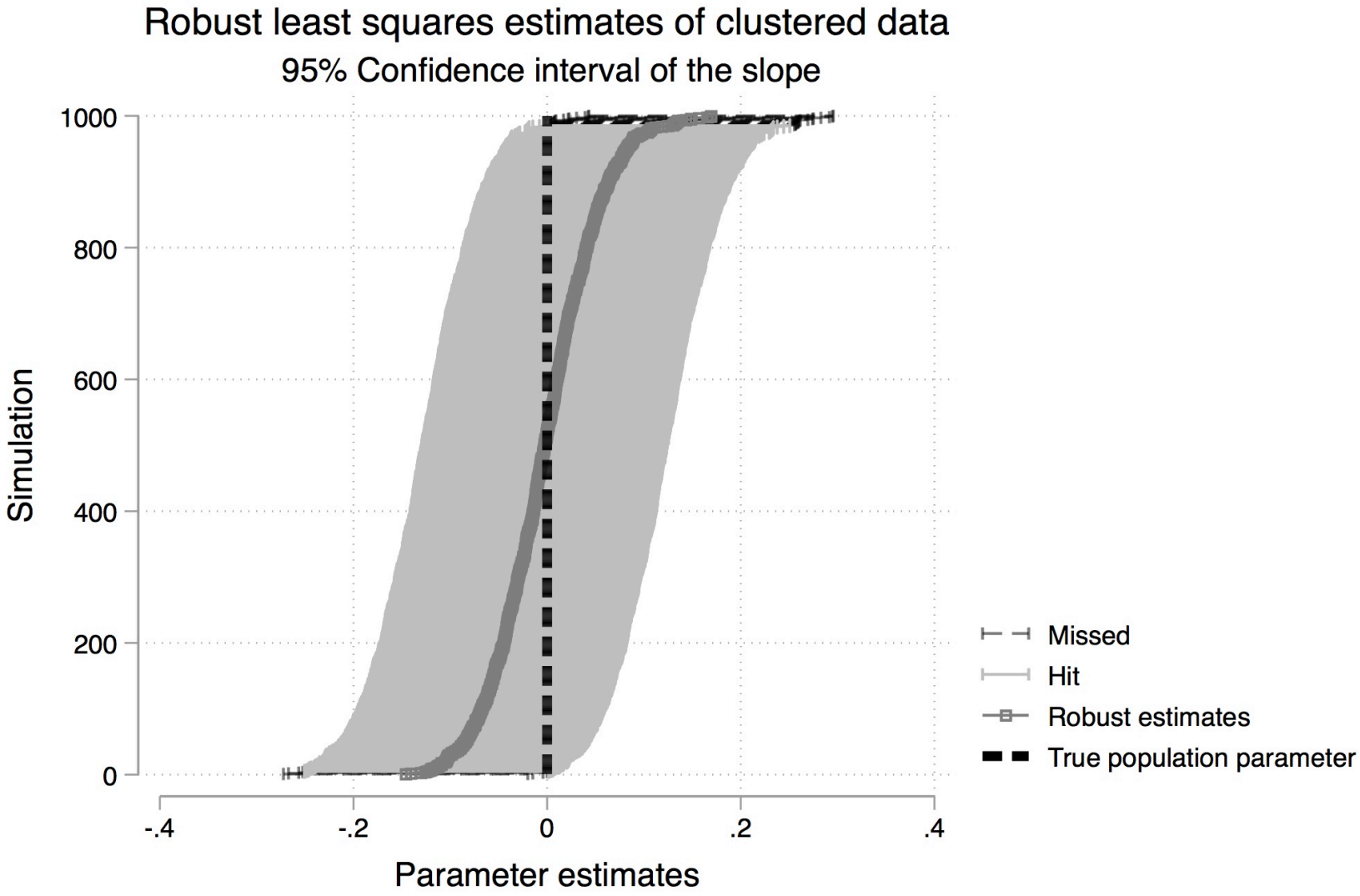



Figure: Distribution of 1,000 95% confidence intervals from a cluster robust least squares regression with dashed region representing those estimates that incorrectly reject the null.

## Cluster robust standard errors

- Better. We don't have the same over-rejection problem as before. If anything it's more conservative.
- The formula for estimating standard errors changes when allowing for arbitrary serial correlation within group.
- Instead of summing over each individual, we first sum over groups
- I'll use matrix notation as it's easier for me to explain by stacking the data.

## Clustered data

- Let's stack the observations by cluster

$$y_g = x_g\beta + u_g$$

- The OLS estimator of $\beta$ is:

$$\widehat{\beta} = [X'X]^{-1}X'y$$

- The variance is given by:

$$Var(\beta) = E[[X'X]^{-1}X'\Omega X[X'X]^{-1}]$$

## Clustered data

With this in mind, we can now write the variance-covariance matrix for clustered data

$$Var(\widehat{\beta}) = [X'X]^{-1}[\sum_{i=1}^{G} x_g' \widehat{u}_g \widehat{u}_g' x_g][X'X]^{-1}$$

where $\hat{u}_g$ are residuals from the stacked regression

- In STATA: `vce(cluster clustervar)`. Where `clustervar` is a variable that identifies the groups in which unobservables are allowed to correlate

## The importance of knowing your data

- In real world you should never go with the "independent and identically distributed" (i.e., homoskedasticity) case. Life is not that simple.
- You need to know your data in order to choose the correct error structure and then infer the required SE calculation
- If you have aggregate variables, like class size, clustering at that level is *required*

## Foundations of scientific knowledge

- Scientific methodologies are the epistemological foundation of **scientific knowledge**, which is a particular kind of knowledge
- Science **does not** collect evidence in order to "prove" what people already believe or want others to believe.
- Science is **process oriented**, not **outcome oriented**.
- Therefore science allows us to accept unexpected and sometimes even undesirable answers.

## My strong pragmatic claim

- "Credible" causal inference is essential to scientific discovery, publishing and **your career**
- Non-credibly identified empirical micro papers, even ones with ingenious theory, will have trouble getting published and won't be taken seriously
- Causal inference in 2019 is a necessary, not a sufficient, condition

## Outline

- Properties of the conditional expectation function (CEF)
- Reasons for using linear regression
- Regression anatomy theorem
- Omitted variable bias

## Properties of the conditional expectation function

- Assume we are interested in the returns to schooling in a wage regression.
- We can summarize the predictive power of schooling's effect on wages with the **conditional expectation function**

$$E(y_i|x_i) \tag{64}$$

- The CEF for a dependent variable, $y_i$, given covariates $X_i$, is the expectation, or population average, of $y_i$ with $x_i$ held constant.

- $E(y_i|x_i)$ gives the expected value of $y$ for given values of $x$
- It provides a reasonable representation of how $y$ changes with $x$
- If $x$ is random, then $E(y_i|x_i)$ is a random function

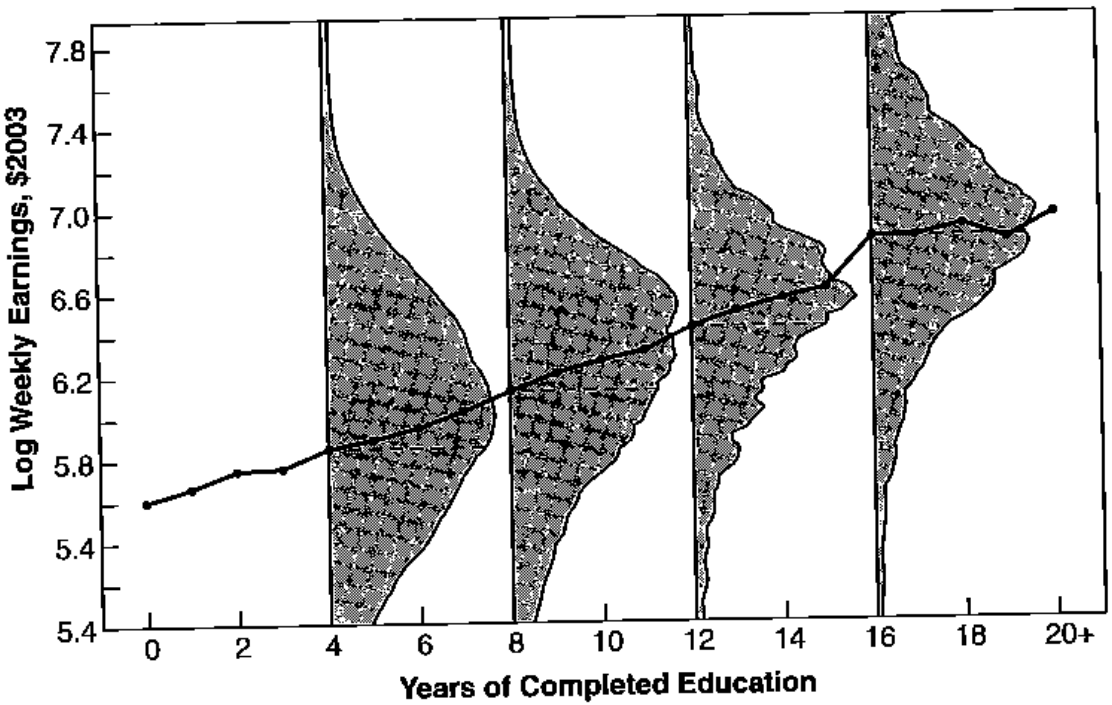



Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

- When there are only two values that $x_i$ can take on, then there are only two values the CEF can take on – but the dummy variable is a special case
- We're often interested in CEFs that are functions of many variables, conveniently subsumed in the vector $x_i$, and for a specific value of $x_i$, we will write

$$E(y_i|x_i = x)$$

## Helpful result: Law of Iterated Expectations

**Definition of Law of Iterated Expectations (LIE)**

The unconditional expectation of a random variable is equal to the expectation of the conditional expectation of the random variable conditional on some other random variable

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

.

We use LIE for a lot of stuff, and it's actually quite intuitive. You may even know it and not know you know it!

## Simple example of LIE

- Say you want to know average IQ but only know average IQ by gender.
- LIE says we get the former by taking conditional expectations by gender and combining them (properly weighted)

$$\begin{aligned} E[IQ] &= E(E[IQ|Sex]) \\ &= \sum_{Sex_i} Pr(Sex_i) \cdot E[IQ|Sex_i] \\ &= Pr(Male) \cdot E[IQ|Male] \\ &\quad + Pr(Female) \cdot E[IQ|Female] \end{aligned}$$

- In words: the weighted average of the conditional averages is the unconditional average.

| Person | Gender | IQ |
|---|---|---|
| 1 | M | 120 |
| 2 | M | 115 |
| 3 | M | 110 |
| 4 | F | 130 |
| 5 | F | 125 |
| 6 | F | 120 |

- $E[IQ] = 120$
- $E[IQ \mid Male] = 115$; $E[IQ \mid Female] = 125$
- LIE: $E\ (\ E\ [\ IQ \mid Sex\ ]\ ) = (0.5)\times 115 + (0.5)\times 125 = 120$

**Proof.**

For the continuous case:

$$
\begin{aligned}
E[E(Y|X)] &= \int E(Y|X=u) g_x(u) du \\
&= \int \left[ \int t f_{y|x}(t|X=u) dt \right] g_x(u) du \\
&= \int \int t f_{y|x}(t|X=u) g_x(u) du dt \\
&= \int t \left[ \int f_{y|x}(t|X=u) g_x(u) du \right] dt \\
&= \int t \left[ f_{x,y} du \right] dt \\
&= \int t g_y(t) dt \\
&= E(y)
\end{aligned}
$$

**Proof.**

For the discrete case,

$$\begin{aligned}
E(E[Y|X]) &= \sum_x E[Y|X=x]p(x) \\
&= \sum_x \left( \sum_y y p(y|x) \right) p(x) \\
&= \sum_x \sum_y y p(x,y) \\
&= \sum_y y \sum_x p(x,y) \\
&= \sum_y y p(y) \\
&= E(Y)
\end{aligned}$$

## Property 1: CEF Decomposition Property

### The CEF Decomposition Property

$$y_i = E(y_i|x_i) + u_i$$

where

1. $u_i$ is mean independent of $x_i$; that is

$$E(u_i|x_i) = 0$$

2. $u_i$ is uncorrelated with any function of $x_i$

In words: Any random variable, $y_i$, can be decomposed into two parts: the part that can be explained by $x_i$ and the part left over that can't be explained by $x_i$. Proof is in Angrist and Pischke (ch. 3)

## Property 2: CEF Prediction Property

**The CEF Prediction Property**

Let $m(x_i)$ be any function of $x_i$. The CEF solves

$$E(y_i|x_i) = \arg\min_{m(x_i)} E[(y_i - m(x_i))^2].$$

In words: The CEF is the minimum mean squared error predictor of $y_i$ given $x_i$. Proof is in Angrist and Pischke (ch. 3)

## 3 reasons why linear regression may be of interest

Linear regression may be interesting even if the underlying CEF is not linear. We review some of the linear theorems now. These are merely to justify the use of linear models to approximate the CEF.

## The Linear CEF Theorem

Suppose the CEF is linear. Then the population regression is it.

Comment: Trivial theorem imho because if the population CEF is linear, then it makes the most sense to use linear regression to estimate it. Proof in Angrist and Pischke (ch. 3). Proof uses the CEF Decomposition Property from earlier.

## The Best Linear Predictor Theorem

1. The CEF, $E(y_i|x_i)$, is the minimum mean squared error (MMSE) predictor of $y_i$ given $x_i$ in the class of all functions $x_i$ by the CEF prediction property
2. The population regression function, $E(x_i y_i)E(x_i x_i')^{-1}$, is the best we can do in the class of all linear functions

Proof is in Angrist and Pischke (ch. 3).

## The Regression CEF Theorem

The function $x_i\beta$ provides the minimum mean squared error (MMSE) linear approximation to $E(y_i|x_i)$, that is

$$\beta = \arg\min_b E\{(E(y_i|x_i) - x_i'b)^2\}$$

Again, proof in Angrist and Pischke (ch. 3).

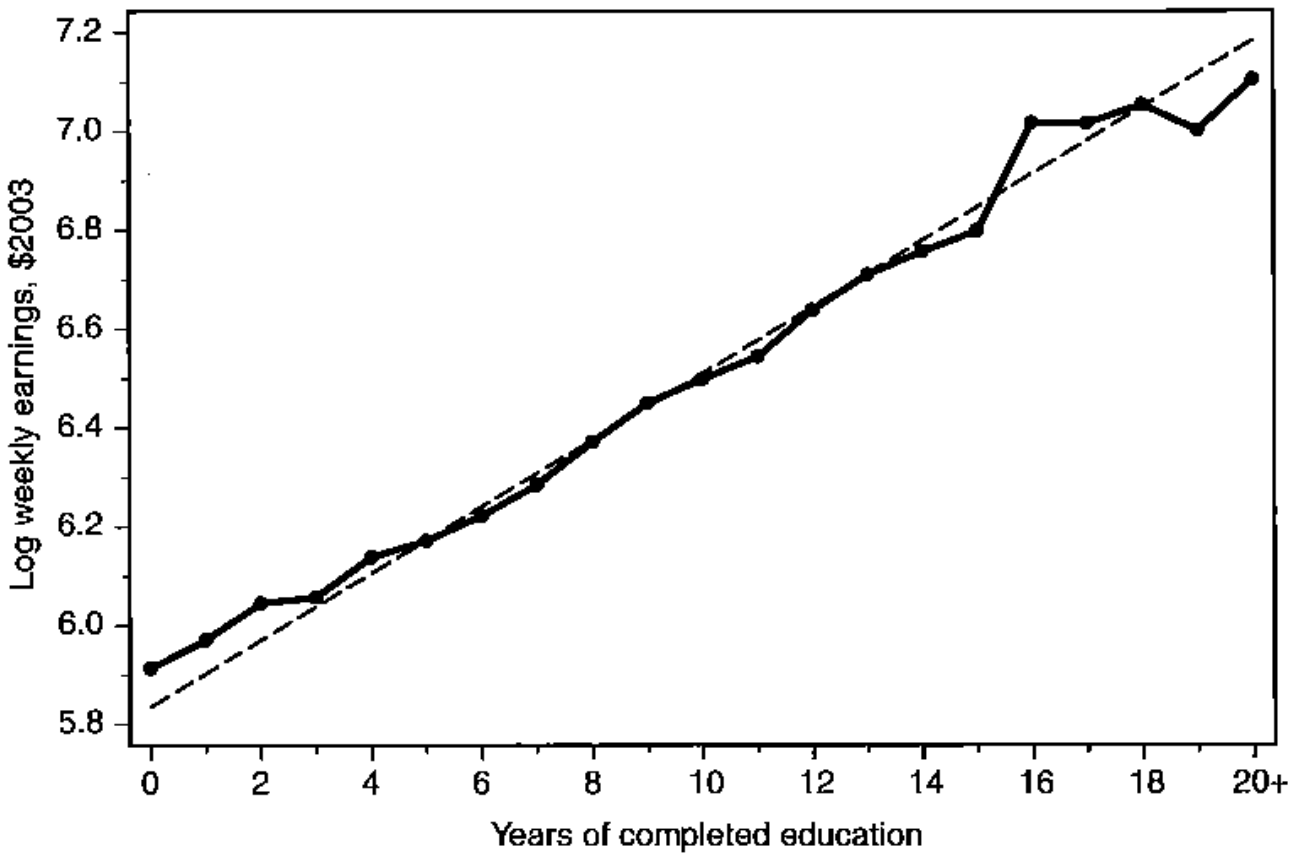



**Figure 3.1.2** Regression threads the CEF of average weekly wages given schooling (dots = CEF; dashes = regression line).

## Random families

- We are interested in the causal effect of family size on labor supply so we regress labor supply onto family size

$$labor\_supply_i = \beta_0 + \beta_1 numkids_i + \varepsilon_i$$

- If couples had kids by flipping coins, then $numkids_i$ independent of $\varepsilon_i$, then estimation is simple - just compare families with different sizes to get the causal effect of *numkids* on *labor_supply*
- But how do we interpret $\widehat{\beta}_1$ if families don't flip coins?

## Non-random families

- If family size is random, you could visualize the causal effect with a scatter plot and the regression line
- If family size is non-random, then we can't do this because we need to control for multiple variables just to remove the factors causing family size to be correlated with $\varepsilon$

## Non-random families

- Assume that family size is random once we condition on race, age, marital status and employment.

$$\begin{aligned}\texttt{labor\_supply}_i &= \beta_0 + \beta_1 \texttt{Numkids}_i + \gamma_1 \texttt{White}_i + \gamma_2 \texttt{Married}_i \\ &\quad + \gamma_3 \texttt{Age}_i + \gamma_4 \texttt{Employed}_i + \varepsilon_i\end{aligned}$$

- To estimate this model, we need:
  1. a data set with all 6 variables;
  2. `Numkids` must be randomly assigned conditional on the other 4 variables
- Now how do we interpret $\widehat{\beta}_1$? And can we visualize $\widehat{\beta}_1$ when there's multiple dimensions to the data? Yes, using the regression anatomy theorem, we can.

## Regression Anatomy Theorem

Assume your main multiple regression model of interest:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i$$

and an auxiliary regression in which the variable $x_{1i}$ is regressed on all the remaining independent variables

$$x_{1i} = \gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i$$

and $\tilde{x}_{1i} = x_{1i} - \widehat{x}_{1i}$ being the residual from the auxiliary regression. The parameter $\beta_1$ can be rewritten as:

$$\beta_1 = \frac{Cov(y_i, \tilde{x}_{1i})}{Var(\tilde{x}_{1i})}$$

In words: The regression anatomy theorem says that $\widehat{\beta}_1$ is a scaled covariance with the $\tilde{x_1}$ residual used instead of the actual data $x$.

## Regression Anatomy Proof

To prove the theorem, note $E[\tilde{x}_{ki}] = E[x_{ki}] - E[\widehat{x}_{ki}] = E[f_i]$, and plug $y_i$ and residual $\tilde{x}_{ki}$ from $x_{ki}$ auxiliary regression into the covariance $cov(y_i, \tilde{x}_{ki})$

$$\begin{aligned}
\beta_k &= \frac{cov(y_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \\
&= \frac{cov(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \\
&= \frac{cov(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, f_i)}{var(f_i)}
\end{aligned}$$

1. Since by construction $E[f_i] = 0$, it follows that the term $\beta_0 E[f_i] = 0$.
2. Since $f_i$ is a linear combination of all the independent variables with the exception of $x_{ki}$, it must be that

$$\beta_1 E[f_i x_{1i}] = \cdots = \beta_{k-1} E[f_i x_{k-1i}] = \beta_{k+1} E[f_i x_{k+1i}] = \cdots = \beta_K E[f_i x_{KI}] = 0$$

## Regression Anatomy Proof (cont.)

3. Consider now the term $E[e_i f_i]$. This can be written as:

$$\begin{aligned} E[e_i f_i] &= E[e_i f_i] \\ &= E[e_i \tilde{x}_{ki}] \\ &= E[e_i(x_{ki} - \widehat{x}_{ki})] \\ &= E[e_i x_{ki}] - E[e_i \tilde{x}_{ki}] \end{aligned}$$

Since $e_i$ is uncorrelated with any independent variable, it is also uncorrelated with $x_{ki}$: accordingly, we have $E[e_i x_{ki}] = 0$. With regard to the second term of the subtraction, substituting the predicted value from the $x_{ki}$ auxiliary regression, we get

$$E[e_i \tilde{x}_{ki}] = E[e_i(\widehat{\gamma}_0 + \widehat{\gamma}_1 x_{1i} + \cdots + \widehat{\gamma}_{k-1} x_{k-1} i + \widehat{\gamma}_{k+1} x_{k+1i} + \cdots + \widehat{\gamma}_K x_{Ki})]$$

Once again, since $e_i$ is uncorrelated with any independent variable, the expected value of the terms is equal to zero. Then, it follows $E[e_i f_i] = 0$.

## Regression Anatomy Proof (cont.)

4. The only remaining term is $E[\beta_k x_{ki} f_i]$ which equals $E[\beta_k x_{ki} \tilde{x}_{ki}]$ since $f_i = \tilde{x}_{ki}$. The term $x_{ki}$ can be substituted using a rewriting of the auxiliary regression model, $x_{ki}$, such that

$$x_{ki} = E[x_{ki}|X_{-k}] + \tilde{x}_{ki}$$

This gives

$$\begin{aligned} E[\beta_k x_{ki} \tilde{x}_{ki}] &= E[\beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})]] \\ &= \beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})] \\ &= \beta_k \{E[\tilde{x}_{ki}^2] + E[(E[x_{ki}|X_{-k}]\tilde{x}_{ki})]\} \\ &= \beta_k var(\tilde{x}_{ki}) \end{aligned}$$

which follows directly from the orthogonoality between $E[x_{ki}|X_{-k}]$ and $\tilde{x}_{ki}$. From previous derivations we finally get

$$cov(y_i, \tilde{x}_{ki}) = \beta_k var(\tilde{x}_{ki})$$

which completes the proof. □

## Stata command: `reganat` (i.e., regression anatomy)

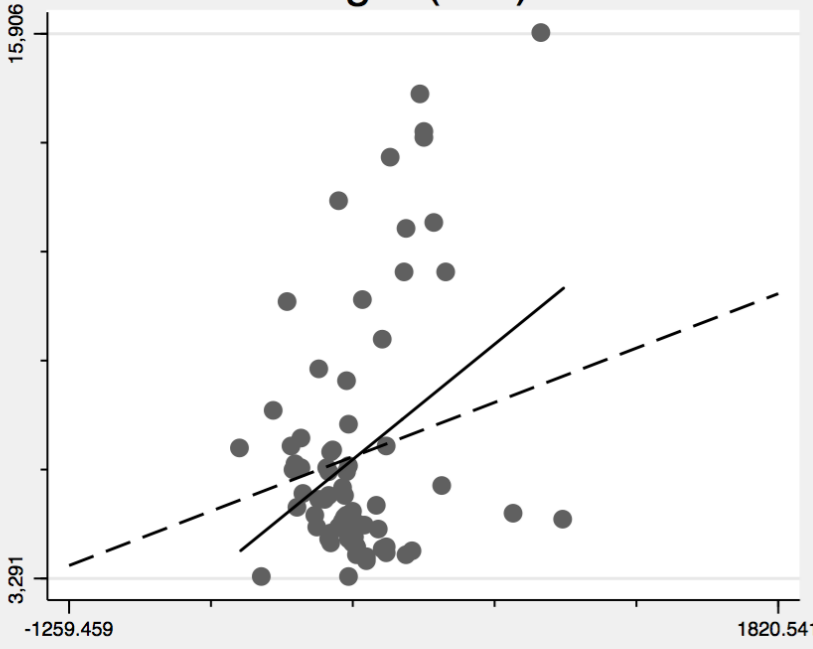
```
. ssc install reganat, replace
. sysuse auto
. regress price length weight headroom mpg
. reganat price length weight headroom mpg, dis(weight length) biline
```

```
. regress price length weight headroom mpg

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  4,    69) =   10.21
       Model |  236190226      4  59047556.6            Prob > F      =  0.0000
    Residual |  398875170     69  5780799.56            R-squared     =  0.3719
-------------+------------------------------           Adj R-squared =  0.3355
       Total |  635065396     73  8699525.97            Root MSE      =  2404.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |  -94.49651   40.39563    -2.34   0.022    -175.0836   -13.90944
      weight |   4.335045   1.162745     3.73   0.000     2.015432    6.654657
    headroom |  -490.9667   388.4892    -1.26   0.211    -1265.981    284.048
         mpg |  -87.95838    83.5927    -1.05   0.296    -254.7213    78.80449
       _cons |   14177.58   5872.766     2.41   0.018     2461.735    25893.43
------------------------------------------------------------------------------
```

# Regression Anatomy

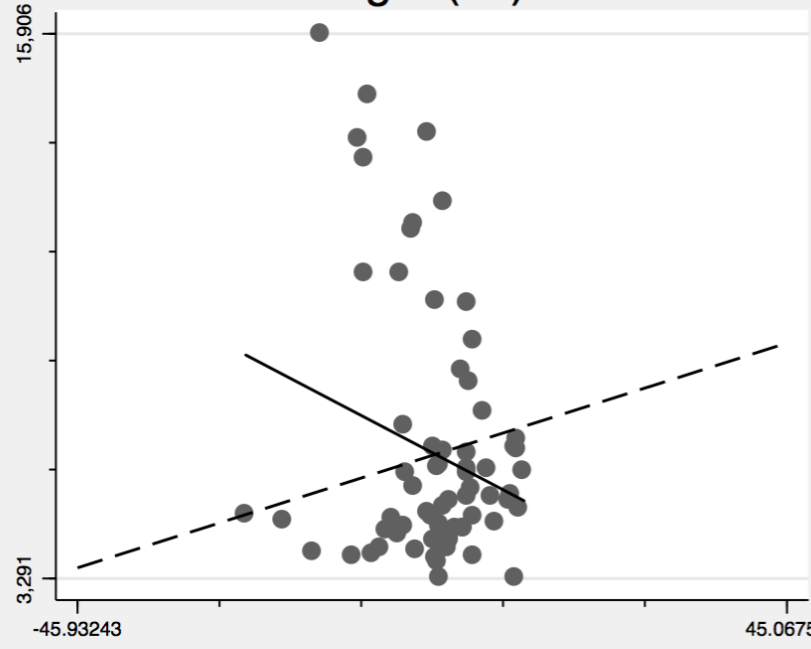
## Dependent variable: Price

### Weight (lbs.)




Multivariate slope: 4.335 (1.163)
Bivariate slope: 2.044 (0.377)

### Length (in.)




Multivariate slope: -94.497 (40.396)
Bivariate slope: 57.202 (14.080)

Covariates: Length (in.), Weight (lbs.), Headroom (in.), Mileage (mpg).

Regression lines: Solid = Multivariate, Dashed = Bivariate.

## Big picture

1. Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable
2. If we prefer to think of approximating $E(y_i|x_i)$ as opposed to predicting $y_i$, the regression CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.
3. Regression anatomy theorem helps us interpret a single slope coefficient in a multiple regression model by the aforementioned decomposition.

## Omitted Variable Bias

- A typical problem is when a key variable is omitted. Assume schooling causes earnings to rise:

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 A_i + u_i$$

$Y_i$ = log of earnings

$S_i$ = schooling measured in years

$A_i$ = individual ability

- Typically the econometrician cannot observe $A_i$; for instance, the Current Population Survey doesn't present adult respondents' family background, intelligence, or motivation.

## Shorter regression

- What are the consequences of leaving ability out of the regression? Suppose you estimated this shorter regression instead:

$$Y_i = \beta_0 + \beta_1 S_i + \eta_i$$

where $\eta_i = \beta_2 A_i + u_i$; $\beta_0$, $\beta_1$, and $\beta_2$ are population regression coefficients; $S_i$ is correlated with $\eta_i$ through $A_i$ only; and $u_i$ is a regression residual uncorrelated with all regressors by definition.

## Derivation of Ability Bias

- Suppressing the $i$ subscripts, the OLS estimator for $\beta_1$ is:

$$\widehat{\beta_1} = \frac{Cov(Y,S)}{Var(S)} = \frac{E[YS] - E[Y]E[S]}{Var(S)}$$

- Plugging in the true model for $Y$, we get:

$$\begin{aligned}
\widehat{\beta_1} &= \frac{Cov[(\beta_0 + \beta_1 S + \beta_2 A + u), S]}{Var(S)} \\
&= \frac{E[(\beta_0 S + \beta_1 S^2 + \beta_2 SA + uS)] - E(S)E[\beta_0 + \beta_1 S + \beta_2 A + u]}{Var(S)} \\
&= \frac{\beta_1 E(S^2) - \beta_1 E(S)^2 + \beta_2 E(AS) - \beta_2 E(S)E(A) + E(uS) - E(S)E(u)}{Var(S)} \\
&= \beta_1 + \beta_2 \frac{Cov(A,S)}{Var(S)}
\end{aligned}$$

- If $\beta_2 > 0$ and Cov$(A,S) > 0$ the coefficient on schooling in the shortened regression (without controlling for $A$) would be upward biased

## Summary

- When $Cov(A, S) > 0$ then ability and schooling are correlated.
- When ability is unobserved, then not even multiple regression will identify the causal effect of schooling on wages.
- Here we see one of the main justifications for this workshop – what will we do when the treatment variable is endogenous?
- We will need an *identification strategy* to recover the causal effect