Lecture 7: Goodness of Fit and Bias

ECON 480 - Econometrics - Fall 2018

Ryan Safner

September 24, 2018

Goodness of Fit

- How well does a line fit data? How tightly clustered around the line are the data points?

$$\underbrace{Y_i}_{Actual} = \underbrace{\hat{Y}_i}_{Model} + \underbrace{\hat{\epsilon}_i}_{Error}$$

Goodness of Fit

- How well does a line fit data? How tightly clustered around the line are the data points?
- Quantify how much variation in Y_i is "explained" by the model

$$\underbrace{Y_i}_{\text{Actual}} = \underbrace{\hat{Y}_i}_{\text{Model}} + \underbrace{\hat{\epsilon}_i}_{\text{Error}}$$

HOOD COLLEGE

- How well does a line fit data? How tightly clustered around the line are the data points?
- Quantify how much variation in Y_i is "explained" by the model
- Recall

$$\underbrace{Y_i}_{Actual} = \underbrace{\hat{Y}_i}_{Model} + \underbrace{\hat{\epsilon}_i}_{Error}$$

- How well does a line fit data? How tightly clustered around the line are the data points?
- Quantify how much variation in Y_i is "explained" by the model
- Recall

$$\underbrace{Y_i}_{Actual} = \underbrace{\hat{Y}_i}_{Model} + \underbrace{\hat{\epsilon}_i}_{Error}$$

- Recall OLS estimators are chosen specifically to minimize SSE $\left(\sum_{i=1}^{n} \hat{\epsilon}_i^{\,2}\right)$

- Primary measure[1] is regression R^2, the fraction of variation in Y explained by variation in predicted values

$$R^2 = \frac{\text{variation in } \widehat{Y_i}}{\text{variation in } Y_i}$$

[1] Sometimes called the "**coefficient of determination**"

Goodness of Fit: R^2 Formula

$$R^2 = \frac{ESS}{TSS}$$

[2] Sometimes called Model Sum of Squares (MSS) or Regression Sum of Squares (RSS)

[3] It can be shown that $\bar{\hat{Y}}_i = \bar{Y}$

$$R^2 = \frac{ESS}{TSS}$$

- Explained Sum of Squares (ESS):[2] sum of squared deviations of predicted values from their mean

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

[2] Sometimes called Model Sum of Squares (MSS) or Regression Sum of Squares (RSS)

[3] It can be shown that $\bar{\hat{Y}}_i = \bar{Y}$

Goodness of Fit: R^2 Formula

$$R^2 = \frac{ESS}{TSS}$$

- Explained Sum of Squares (ESS):[2] sum of squared deviations of predicted values from their mean

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

- Total Sum of Squares (TSS): sum of squared deviations of actual values from their mean[3]

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

[2] Sometimes called Model Sum of Squares (MSS) or Regression Sum of Squares (RSS)

[3] It can be shown that $\bar{\hat{Y}}_i = \bar{Y}$

- Equivalently, the complement of the fraction of *unexplained* variation in Y_i

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

- Equivalently, the complement of the fraction of *unexplained* variation in Y_i

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

- Equivalently, the square of the correlation coefficient between X and Y:

$$R^2 = (r_{X,Y})^2$$

- The Standard Error of the Regression[4], $\hat{\sigma}$ or $\hat{\sigma}_\epsilon$ is an estimator of the standard deviation of ϵ_i

$$\hat{\sigma}_\epsilon = \sqrt{\frac{SSE}{n-2}}$$

[4] Why standard *"error"* and not standard *"deviation"*? You'll know by the end of this lecture!

- The Standard Error of the Regression[4], $\hat{\sigma}$ or $\hat{\sigma}_\epsilon$ is an estimator of the standard deviation of ϵ_i

$$\hat{\sigma}_\epsilon = \sqrt{\frac{SSE}{n-2}}$$

- Measures the average size of the residuals (distance between a data point and the line)

[4] Why standard *"error"* and not standard *"deviation"*? You'll know by the end of this lecture!

- The Standard Error of the Regression[4], $\hat{\sigma}$ or $\hat{\sigma}_{\epsilon}$ is an estimator of the standard deviation of ϵ_i

$$\hat{\sigma}_{\epsilon} = \sqrt{\frac{SSE}{n-2}}$$

- Measures the average size of the residuals (distance between a data point and the line)
    - Degrees of Freedom correction of $n-2$: we use up 2 df to first calculate $\hat{\beta}_0$ and $\hat{\beta}_1$

[4] Why standard *"error"* and not standard *"deviation"*? You'll know by the end of this lecture!

Goodness of Fit: Standard Error of the Regression

- The Standard Error of the Regression[4], $\hat{\sigma}$ or $\hat{\sigma}_\epsilon$ is an estimator of the standard deviation of ϵ_i

$$\hat{\sigma}_\epsilon = \sqrt{\frac{SSE}{n-2}}$$

- Measures the average size of the residuals (distance between a data point and the line)
    - Degrees of Freedom correction of $n-2$: we use up 2 df to first calculate $\hat{\beta}_0$ and $\hat{\beta}_1$
- R calls this `Residual Standard Error`

[4] Why standard *"error"* and not standard *"deviation"*? You'll know by the end of this lecture!

- The Standard Error of the Regression[4], $\hat{\sigma}$ or $\hat{\sigma_\epsilon}$ is an estimator of the standard deviation of ϵ_i

$$\hat{\sigma_\epsilon} = \sqrt{\frac{SSE}{n-2}}$$

- Measures the average size of the residuals (distance between a data point and the line)
  - Degrees of Freedom correction of $n-2$: we use up 2 df to first calculate $\hat{\beta}_0$ and $\hat{\beta}_1$
- R calls this `Residual Standard Error`
  - Note R tells you it calculates this with a df of $n-2$

[4] Why standard *"error"* and not standard *"deviation"*? You'll know by the end of this lecture!

Goodness of Fit: Looking at R

```r
summary(school.regression)
```

```
##
## Call:
## lm(formula = testscr ~ str, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Goodness of Fit: Example

- `R-squared` of this regression is `0.05124`

Goodness of Fit: Example

- `R-squared` of this regression is `0.05124`
    - 5% of variation in Test Scores are explained by our model

Goodness of Fit: Example

- `R-squared` of this regression is `0.05124`
  - 5% of variation in Test Scores are explained by our model
- `Residual standard error` is `18.58`

Goodness of Fit: Example

- `R-squared` of this regression is `0.05124`
  - 5% of variation in Test Scores are explained by our model
- `Residual standard error` is `18.58`
  - Average test score is 18.58 points above/below our model's prediction

Goodness of Fit: Example

- `R-squared` of this regression is `0.05124`
  - 5% of variation in Test Scores are explained by our model
- `Residual standard error` is `18.58`
  - Average test score is 18.58 points above/below our model's prediction
- Indicates there are other important factors that also influence test scores

Goodness of Fit: Example

- `R-squared` of this regression is `0.05124`
    - 5% of variation in Test Scores are explained by our model
- `Residual standard error` is `18.58`
    - Average test score is 18.58 points above/below our model's prediction
- Indicates there are other important factors that also influence test scores
- **Note: it is very rare in econo(metr)ics that we get very high R^2 values**

Goodness of Fit: Example

- `R-squared` of this regression is `0.05124`
  - 5% of variation in Test Scores are explained by our model
- `Residual standard error` is `18.58`
  - Average test score is 18.58 points above/below our model's prediction
- Indicates there are other important factors that also influence test scores
- **Note: it is very rare in econo(metr)ics that we get very high R^2 values**
  - Lots of unobserved variables affecting economic outcomes

Goodness of Fit: Example

- `R-squared` of this regression is `0.05124`
  - 5% of variation in Test Scores are explained by our model
- `Residual standard error` is `18.58`
  - Average test score is 18.58 points above/below our model's prediction
- Indicates there are other important factors that also influence test scores
- **Note: it is very rare in econo(metr)ics that we get very high R^2 values**
  - Lots of unobserved variables affecting economic outcomes
  - **Don't get discouraged!** We care about **marginal (causal) effects**, not R^2!

- A lot of regression diagnostics have to do with exploring the residuals a bit more

Measures of Fit: Looking at Residuals

- A lot of regression diagnostics have to do with exploring the residuals a bit more

```
# Save the residuals as a vector called 'res'
CASchool$res <- residuals(school.regression) # use 'res()' function
summary(CASchool$res) # get summary stats of residuals
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -47.7267 -14.2507   0.4826   0.0000  12.8222  48.5404
```

- A lot of regression diagnostics have to do with exploring the residuals a bit more

```r
# Save the residuals as a vector called 'res'
CASchool$res <- residuals(school.regression) # use 'res()' function
summary(CASchool$res) # get summary stats of residuals
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -47.7267 -14.2507   0.4826   0.0000  12.8222  48.5404
```

```r
# Save the predicted values of the regression as a vector called 'yhat'
CASchool$yhat <- predict(school.regression) # use 'predict()' function
summary(CASchool$yhat) # get summary stats of predictions
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   640.1   651.3   654.0   654.2   656.6   667.0
```

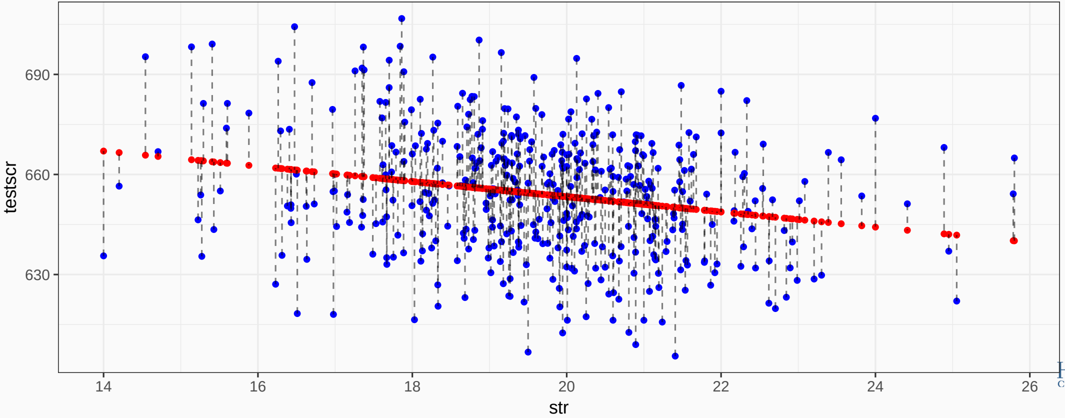
Measures of Fit: Looking at Residuals II

```r
# Remake the scatterplot and point out the residuals
scatterplot.res<-ggplot(CASchool, aes(x=str, y=testscr))+
  geom_point(color="blue")+ # plot original points blue
  geom_point(aes(y=yhat),color="red")+ # plot predicted yhat in red
  geom_segment(aes(xend=str,yend=yhat),linetype=2, alpha=0.5)
# last line connects predicted (yhat) and actual points with dashed line
```

HOOD COLLEGE

Measures of Fit: Looking at Residuals III

```
scatterplot.res
```



Measures of Fit: Looking at Residuals IV

```r
# A vector of all residuals for each observation is stored in the reg object:
head(school.regression$residuals) #look at first 6 obs residuals
```

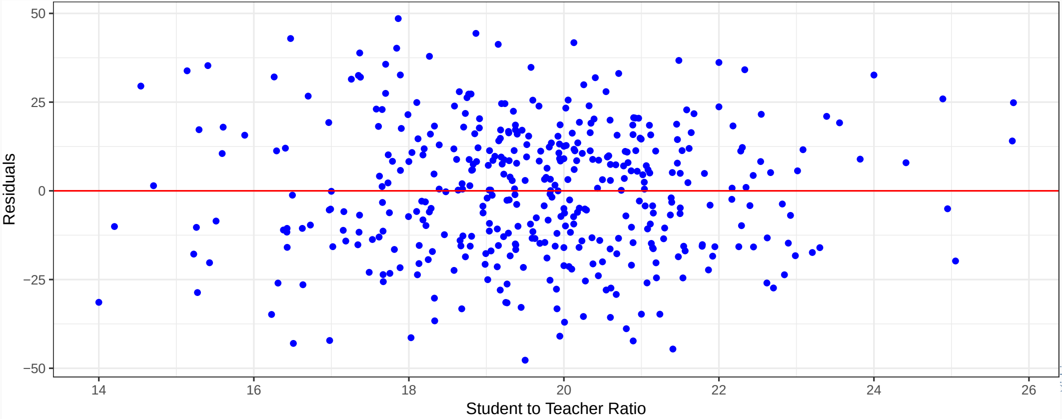
```
##         1         2         3         4         5         6
##  32.65260  11.33917 -12.70689 -11.66198 -15.51593 -44.58076
```

Measures of Fit: Residual Plot

- We often plot the residuals against X in a residual plot

```
# Create a scatterplot with the residuals.
# Same as before, but instead of testscr, we will use residuals (res)
school.resplot<-ggplot(CASchool, aes(str,school.regression$residuals))+
  geom_point(color="blue",fill="blue")+
  xlab("Student to Teacher Ratio")+
  ylab("Residuals")+theme_bw()+
  geom_hline(yintercept=0, color="red") #add horizontal line at y=0 to graph
```

```
school.resplot
```



The Sampling Distributions of the OLS Estimators

Recall: The Two Big Problems with Data

- We use econometrics to **identify** causal relationships and make **inferences** about them

Recall: The Two Big Problems with Data

- We use econometrics to **identify** causal relationships and make **inferences** about them

   1. Problem for **identification**: endogeneity

- We use econometrics to **identify** causal relationships and make **inferences** about them

  1. Problem for **identification**: endogeneity

     - X is exogenous if its variation is *unrelated* to other factors (ϵ) that affect Y

Recall: The Two Big Problems with Data

- We use econometrics to **identify** causal relationships and make **inferences** about them

  1. Problem for **identification**: endogeneity

     - X is exogenous if its variation is *unrelated* to other factors (ϵ) that affect Y
     - X is endogenous if its variation is *related* to other factors (ϵ) that affect Y

Recall: The Two Big Problems with Data

- We use econometrics to **identify** causal relationships and make **inferences** about them

1. Problem for **identification**: endogeneity
    - X is exogenous if its variation is *unrelated* to other factors (ϵ) that affect Y
    - X is endogenous if its variation is *related* to other factors (ϵ) that affect Y
2. Problem for **inference**: randomness

Recall: The Two Big Problems with Data

- We use econometrics to **identify** causal relationships and make **inferences** about them

1. Problem for **identification**: endogeneity
    - X is exogenous if its variation is *unrelated* to other factors (ϵ) that affect Y
    - X is endogenous if its variation is *related* to other factors (ϵ) that affect Y
2. Problem for **inference**: randomness
    - Data is random due to **natural sampling variation**

- We use econometrics to **identify** causal relationships and make **inferences** about them

1. Problem for **identification**: endogeneity
    - X is exogenous if its variation is *unrelated* to other factors (ϵ) that affect Y
    - X is endogenous if its variation is *related* to other factors (ϵ) that affect Y
2. Problem for **inference**: randomness
    - Data is random due to **natural sampling variation**
    - Taking one sample of a population will yield slightly different information than another sample of the same population

- OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are computed from a specific sample of data

- OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are computed from a specific sample of data
- Our OLS model contains **2 sources of randomness**:

- OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are computed from a specific sample of data
- Our OLS model contains **2 sources of randomness**:
    - Modeled randomness: ϵ includes all factors affecting Y *other* than X, different samples have different values of those other factors

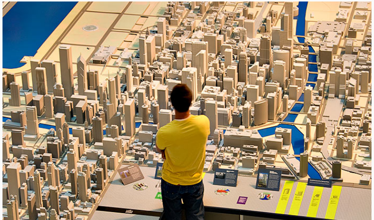
Distributions of the OLS Estimators

- OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are computed from a specific sample of data
- Our OLS model contains **2 sources of randomness**:
  - Modeled randomness: ϵ includes all factors affecting Y *other* than X, different samples have different values of those other factors
  - Sampling randomness: different samples will generate different OLS estimators

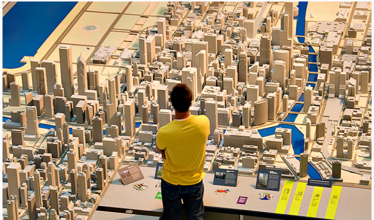
Distributions of the OLS Estimators

- OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are computed from a specific sample of data
- Our OLS model contains **2 sources of randomness**:
    - Modeled randomness: ϵ includes all factors affecting Y *other* than X, different samples have different values of those other factors
    - Sampling randomness: different samples will generate different OLS estimators
- Thus, $\hat{\beta}_0, \hat{\beta}_1$ are also random variables, with their own **sampling distribution**

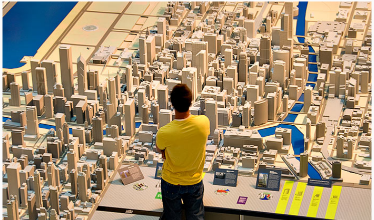
- **Inferential statistics** analyzes a **sample** to make inferences about a much larger (unobservable) **population**



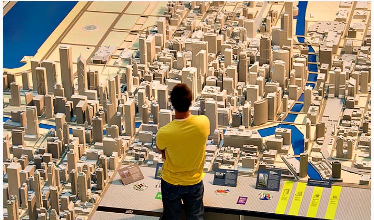
- **Inferential statistics** analyzes a **sample** to make inferences about a much larger (unobservable) **population**
    - Population: all possible individuals that match some well-defined criterion of interest (people, firms, cities, etc)



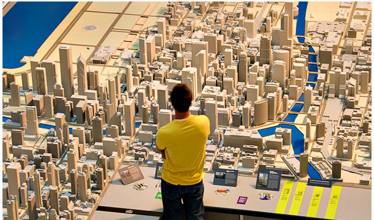
- **Inferential statistics** analyzes a **sample** to make inferences about a much larger (unobservable) **population**
  - Population: all possible individuals that match some well-defined criterion of interest (people, firms, cities, etc)
    - Characteristics about (relationships between variables describing) populations are called **parameters**



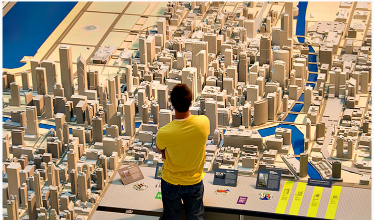
- **Inferential statistics** analyzes a **sample** to make inferences about a much larger (unobservable) **population**
  - Population: all possible individuals that match some well-defined criterion of interest (people, firms, cities, etc)
    - Characteristics about (relationships between variables describing) populations are called **parameters**
  - Sample: some portion of the population of interest to *represent the whole*



- **Inferential statistics** analyzes a **sample** to make inferences about a much larger (unobservable) **population**
  - Population: all possible individuals that match some well-defined criterion of interest (people, firms, cities, etc)
    - Characteristics about (relationships between variables describing) populations are called **parameters**
  - Sample: some portion of the population of interest to *represent the whole*
    - Samples examine part of a population to generate **statistics** used to estimate population parameters



- **Inferential statistics** analyzes a **sample** to make inferences about a much larger (unobservable) **population**
  - Population: all possible individuals that match some well-defined criterion of interest (people, firms, cities, etc)
    - Characteristics about (relationships between variables describing) populations are called **parameters**
  - Sample: some portion of the population of interest to *represent the whole*
    - Samples examine part of a population to generate **statistics** used to estimate population parameters
- We almost never can directly study the population, so we *model* it with our samples



Example

Suppose you randomly select 100 people and ask how many hours they spend on the internet each day. You take the mean of your sample, and it comes out to 5.4 hours.

Example

Suppose you randomly select 100 people and ask how many hours they spend on the internet each day. You take the mean of your sample, and it comes out to 5.4 hours.

- 5.4 hours is a **sample statistic** describing the sample

HOOD COLLEGE

Example

Suppose you randomly select 100 people and ask how many hours they spend on the internet each day. You take the mean of your sample, and it comes out to 5.4 hours.

- 5.4 hours is a **sample statistic** describing the sample
- We are more interested in the corresponding **parameter** of the population (e.g. all Americans)

Example

Suppose you randomly select 100 people and ask how many hours they spend on the internet each day. You take the mean of your sample, and it comes out to 5.4 hours.

- 5.4 hours is a **sample statistic** describing the sample
- We are more interested in the corresponding **parameter** of the population (e.g. all Americans)
- If we take another sample of $n = 100$ people, would we get the same number?

Example
Suppose you randomly select 100 people and ask how many hours they spend on the internet each day. You take the mean of your sample, and it comes out to 5.4 hours.

- 5.4 hours is a **sample statistic** describing the sample
- We are more interested in the corresponding **parameter** of the population (e.g. all Americans)
- If we take another sample of $n = 100$ people, would we get the same number?
    - **Roughly, but probably not exactly**

Example
Suppose you randomly select 100 people and ask how many hours they spend on the internet each day. You take the mean of your sample, and it comes out to 5.4 hours.

- 5.4 hours is a **sample statistic** describing the sample
- We are more interested in the corresponding **parameter** of the population (e.g. all Americans)
- If we take another sample of $n = 100$ people, would we get the same number?
    - **Roughly, but probably not exactly**
    - Sampling variability describes the effect of a statistic varying somewhat from sample to sample

Example

Suppose you randomly select 100 people and ask how many hours they spend on the internet each day. You take the mean of your sample, and it comes out to 5.4 hours.

- 5.4 hours is a **sample statistic** describing the sample
- We are more interested in the corresponding **parameter** of the population (e.g. all Americans)
- If we take another sample of $n = 100$ people, would we get the same number?
    - **Roughly, but probably not exactly**
    - Sampling variability describes the effect of a statistic varying somewhat from sample to sample
    - This is normal, not the result of any error or bias

I.I.D. Samples

- If we collect many samples, and each sample is randomly drawn from the population (and then replaced), then the distribution of samples is said to be independently and identically distributed (i.i.d.)

Hood College

- If we collect many samples, and each sample is randomly drawn from the population (and then replaced), then the distribution of samples is said to be independently and identically distributed (i.i.d.)
    - Each sample is *independent* of each other sample (due to replacement)

I.I.D. Samples

- If we collect many samples, and each sample is randomly drawn from the population (and then replaced), then the distribution of samples is said to be independently and identically distributed (i.i.d.)
    - Each sample is *independent* of each other sample (due to replacement)
    - Each sample comes from the *identical* underlying population distribution

HOOD COLLEGE

- So calculating OLS estimators for a sample of data makes the OLS estimators *themselves* random variables:

- So calculating OLS estimators for a sample of data makes the OLS estimators *themselves* random variables:
    - Draw of i is random $\implies$ value of each (X_i, Y_i) is random $\implies$ $\hat{\beta}_0, \hat{\beta}_1$ are random

- So calculating OLS estimators for a sample of data makes the OLS estimators *themselves* random variables:
    - Draw of i is random $\implies$ value of each (X_i, Y_i) is random $\implies$ $\hat{\beta}_0, \hat{\beta}_1$ are random
    - Taking different samples will create different values of $\hat{\beta}_0, \hat{\beta}_1$

- So calculating OLS estimators for a sample of data makes the OLS estimators *themselves* random variables:
  - Draw of i is random $\implies$ value of each (X_i, Y_i) is random $\implies$ $\hat{\beta}_0, \hat{\beta}_1$ are random
  - Taking different samples will create different values of $\hat{\beta}_0, \hat{\beta}_1$
  - Therefore, $\hat{\beta}_0, \hat{\beta}_1$ have sampling distributions across different samples

- Central Limit Theorem (CLT) says if we collect samples of size n from the same population and generate a sample statistic (e.g. OLS estimator), then with large enough n, the distribution of the sample statistic is approximately normal IF

- Central Limit Theorem (CLT) says if we collect samples of size n from the same population and generate a sample statistic (e.g. OLS estimator), then with large enough n, the distribution of the sample statistic is approximately normal IF
    1. $n \geq 30$

- Central Limit Theorem (CLT) says if we collect samples of size n from the same population and generate a sample statistic (e.g. OLS estimator), then with large enough n, the distribution of the sample statistic is approximately normal IF

  1. $n \geq 30$
  2. Samples come from a *known* normal distribution $\sim N(\mu, \sigma)$

- Central Limit Theorem (CLT) says if we collect samples of size n from the same population and generate a sample statistic (e.g. OLS estimator), then with large enough n, the distribution of the sample statistic is approximately normal IF
    1. $n \geq 30$
    2. Samples come from a *known* normal distribution $\sim N(\mu, \sigma)$
- If neither of these are true, we have other methods (coming shortly!)

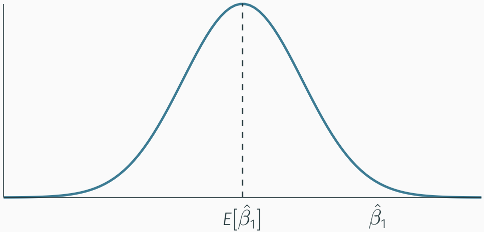
The Central Limit Theorem

- Central Limit Theorem (CLT) says if we collect samples of size n from the same population and generate a sample statistic (e.g. OLS estimator), then with large enough n, the distribution of the sample statistic is approximately normal IF
  1. $n \geq 30$
  2. Samples come from a *known* normal distribution $\sim N(\mu, \sigma)$
- If neither of these are true, we have other methods (coming shortly!)
- One of the most fundamental principles in all of statistics

- Central Limit Theorem (CLT) says if we collect samples of size n from the same population and generate a sample statistic (e.g. OLS estimator), then with large enough n, the distribution of the sample statistic is approximately normal IF
    1. $n \geq 30$
    2. Samples come from a *known* normal distribution $\sim N(\mu, \sigma)$
- If neither of these are true, we have other methods (coming shortly!)
- One of the most fundamental principles in all of statistics
    - Allows for virtually all testing of statistical hypotheses $\rightarrow$ estimating probabilities of values on a normal distribution

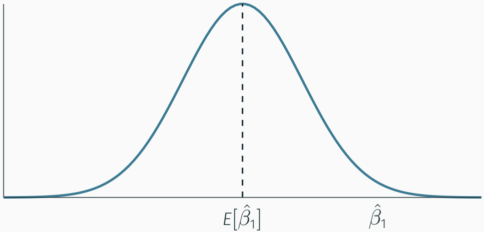
- The CLT allows us to approximate the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ as normal

$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$



- The CLT allows us to approximate the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ as normal
  - Generally agreed for $n > 100$

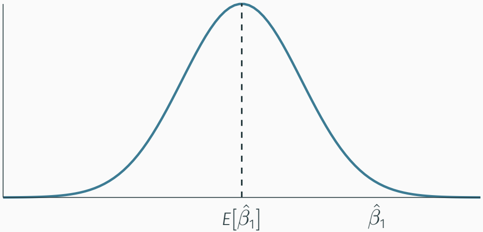
$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$



The Sampling Distribution of $\hat{\beta}_1$

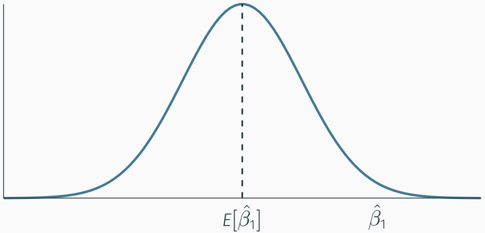
- The CLT allows us to approximate the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ as normal
  - Generally agreed for $n > 100$
- We care about $\hat{\beta}_1$ (slope) since it has economic meaning, rarely about $\hat{\beta}_0$ (intercept)

$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$



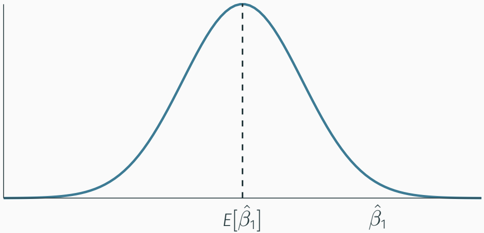
$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$

- We want to know:



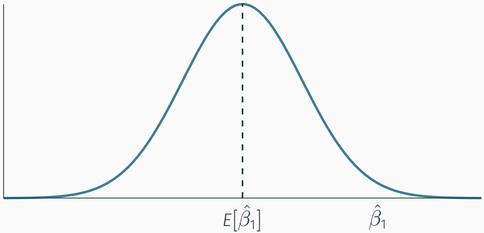
$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$

- We want to know:
    - $E[\hat{\beta}_1]$; what is the center of the distribution?



$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$

- We want to know:
  - $E[\hat{\beta}_1]$; what is the center of the distribution?
  - $\sigma_{\hat{\beta}_1}$; how precise is our estimate?



Assumptions about Errors

- In order to talk about $E[\hat{\beta}_1]$, we need to talk about ϵ

Assumptions about Errors

- In order to talk about $E[\hat{\beta}_1]$, we need to talk about ϵ
- Recall: ϵ is a random variable, and we can never measure the error term

- In order to talk about $E[\hat{\beta}_1]$, we need to talk about ϵ
- Recall: ϵ is a random variable, and we can never measure the error term
- We make four critical **assumptions about** ϵ:

- In order to talk about $E[\hat{\beta}_1]$, we need to talk about ϵ
- Recall: ϵ is a random variable, and we can never measure the error term
- We make four critical **assumptions about** ϵ:
    1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

Assumptions about Errors

- In order to talk about $E[\hat{\beta}_1]$, we need to talk about ϵ
- Recall: ϵ is a random variable, and we can never measure the error term
- We make four critical **assumptions about ϵ**:
  1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

  2. The variance of the residuals is constant, written:

$$var(\epsilon) = \sigma^2_\epsilon$$

Assumptions about Errors

- In order to talk about $E[\hat{\beta}_1]$, we need to talk about ϵ
- Recall: ϵ is a random variable, and we can never measure the error term
- We make four critical **assumptions about** ϵ:
  1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

  2. The variance of the residuals is constant, written:

$$var(\epsilon) = \sigma^2_\epsilon$$

  3. Errors are not correlated across observations:

$$Cov(\epsilon_i, \epsilon_j) = 0 \text{ or } Corr(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

- In order to talk about $E[\hat{\beta}_1]$, we need to talk about ϵ
- Recall: ϵ is a random variable, and we can never measure the error term
- We make four critical **assumptions about** ϵ:
  1. The expected value of the residuals is 0
  $$E[\epsilon] = 0$$
  2. The variance of the residuals is constant, written:
  $$var(\epsilon) = \sigma^2_\epsilon$$
  3. Errors are not correlated across observations:
  $$Cov(\epsilon_i, \epsilon_j) = 0 \text{ or } Corr(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$
  4. There is no correlation between X and the error term:
  $$Cov(X, \epsilon) = 0 \text{ or } Corr(X, \epsilon) = 0 \text{ or } E[\epsilon|X] = 0$$

Assumptions about Errors II

1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

2. The variance of the residuals is constant, written:

$$var(\epsilon) = \sigma_{\epsilon}^{2}$$

1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

2. The variance of the residuals is constant, written:

$$var(\epsilon) = \sigma_{\epsilon}^2$$

- The first two assumptions simply state that errors are **i.i.d.**, drawn from the same distribution with mean 0 and variance σ_{ϵ}^2

Assumptions about Errors II

1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

2. The variance of the residuals is constant, written:

$$var(\epsilon) = \sigma^2_\epsilon$$

- The first two assumptions simply state that errors are **i.i.d.**, drawn from the same distribution with mean 0 and variance σ^2_ϵ
- The second assumption implies that errors have the same variance across X, "homoskedastic"

HOOD COLLEGE

Assumptions about Errors II

1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

2. The variance of the residuals is constant, written:

$$var(\epsilon) = \sigma^2_\epsilon$$

- The first two assumptions simply state that errors are **i.i.d.**, drawn from the same distribution with mean 0 and variance σ^2_ϵ
- The second assumption implies that errors have the same variance across X, "homoskedastic"
    - Many times, this assumption turns out to be false, when errors are called "heteroskedastic"

Assumptions about Errors II

1. The expected value of the residuals is 0

$$E[\epsilon] = 0$$

2. The variance of the residuals is constant, written:

$$var(\epsilon) = \sigma^2_\epsilon$$

- The first two assumptions simply state that errors are **i.i.d.**, drawn from the same distribution with mean 0 and variance σ^2_ϵ
- The second assumption implies that errors have the same variance across X, "homoskedastic"
  - Many times, this assumption turns out to be false, when errors are called "heteroskedastic"
  - This *would* be a problem (for inference), but we have a simple fix for this (coming shortly)

3. Errors are not correlated across observations:

$$Cov(\epsilon_i, \epsilon_j) = 0 \text{ or } Corr(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

Assumptions about Errors III

3. Errors are not correlated across observations:

$$Cov(\epsilon_i, \epsilon_j) = 0 \text{ or } Corr(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

- For simple cross-sectional data, this assumption is rarely an issue

3. Errors are not correlated across observations:

$$Cov(\epsilon_i, \epsilon_j) = 0 \text{ or } Corr(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

- For simple cross-sectional data, this assumption is rarely an issue
- Time-series & panel data nearly always contain serial correlation in the errors, also known as autocorrelation

3. Errors are not correlated across observations:

$$Cov(\epsilon_i, \epsilon_j) = 0 \text{ or } Corr(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

- For simple cross-sectional data, this assumption is rarely an issue
- Time-series & panel data nearly always contain serial correlation in the errors, also known as autocorrelation
    - e.g. "this months sales look like last months's sales, which look like...etc"

HOOD COLLEGE

3. Errors are not correlated across observations:

$$Cov(\epsilon_i, \epsilon_j) = 0 \text{ or } Corr(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

- For simple cross-sectional data, this assumption is rarely an issue
- Time-series & panel data nearly always contain serial correlation in the errors, also known as autocorrelation
  - e.g. "this months sales look like last months's sales, which look like...etc"
  - We have fixes to deal with autocorrelation (coming much later)

4. There is no correlation between X and the error term:

$$Cov(X,\epsilon) = 0 \text{ or } Corr(X,\epsilon) = 0 \text{ or } E[\epsilon|X] = 0$$

4. There is no correlation between X and the error term:

$$Cov(X, \epsilon) = 0 \text{ or } Corr(X, \epsilon) = 0 \text{ or } E[\epsilon|X] = 0$$

- **This is the absolute killer assumption, because it assumes exogeneity**

HOOD COLLEGE

The Zero Conditional Mean Assumption

4. There is no correlation between X and the error term:

$$Cov(X, \epsilon) = 0 \text{ or } Corr(X, \epsilon) = 0 \text{ or } E[\epsilon|X] = 0$$

- **This is the absolute killer assumption, because it assumes exogeneity**
- AKA the Zero Conditional Mean assumption (third way of writing it above)

4. There is no correlation between X and the error term:

$$Cov(X, \epsilon) = 0 \text{ or } Corr(X, \epsilon) = 0 \text{ or } E[\epsilon | X] = 0$$

- **This is the absolute killer assumption, because it assumes exogeneity**
- AKA the Zero Conditional Mean assumption (third way of writing it above)
- "Does knowing X give me any useful information about ϵ?"

The Zero Conditional Mean Assumption

4. There is no correlation between X and the error term:

$$Cov(X, \epsilon) = 0 \text{ or } Corr(X, \epsilon) = 0 \text{ or } E[\epsilon | X] = 0$$

- **This is the absolute killer assumption, because it assumes exogeneity**
- AKA the Zero Conditional Mean assumption (third way of writing it above)
- "Does knowing X give me any useful information about ϵ?"
    - If yes, your model is endogenous, biased and **not-causal**!

- We want to see if $\hat{\beta}_1$ is unbiased: there is no systematic difference, on average, between sample values of $\hat{\beta}_1$ and the true population β_1, i.e.

$$E[\hat{\beta}_1] = \beta_1$$

- We want to see if $\hat{\beta}_1$ is unbiased: there is no systematic difference, on average, between sample values of $\hat{\beta}_1$ and the true population β_1, i.e.

$$E[\hat{\beta}_1] = \beta_1$$

- Does *not* mean any sample gives us $\hat{\beta}_1 = \beta_1$, only the estimation procedure will, *on average*, yield the correct value

- We want to see if $\hat{\beta}_1$ is unbiased: there is no systematic difference, on average, between sample values of $\hat{\beta}_1$ and the true population β_1, i.e.

$$E[\hat{\beta}_1] = \beta_1$$

- Does *not* mean any sample gives us $\hat{\beta}_1 = \beta_1$, only the estimation procedure will, *on average*, yield the correct value
  - Random errors above and below the true value cancel (so that on average, $E[\hat{\epsilon}|X] = 0$)

HOOD COLLEGE

Sidenote: Estimators of Statistics

- In statistics, an **estimator** is simply a rule that for calculating a statistic (often about a wider population parameter)

- In statistics, an **estimator** is simply a rule that for calculating a statistic (often about a wider population parameter)

Example

We want to estimate the average height (H) of U.S. adults (population) and have a random sample of 100 adults.

- In statistics, an **estimator** is simply a rule that for calculating a statistic (often about a wider population parameter)

Example

We want to estimate the average height (H) of U.S. adults (population) and have a random sample of 100 adults.

- Calculate the mean height of our sample ($\overline{H}$) to estimate the true mean height of the population (μ_H)

- In statistics, an **estimator** is simply a rule that for calculating a statistic (often about a wider population parameter)

Example

We want to estimate the average height (H) of U.S. adults (population) and have a random sample of 100 adults.

- Calculate the mean height of our sample ($\bar{H}$) to estimate the true mean height of the population (μ_H)
  - $\bar{H}$ is an **estimator** of μ_H

- In statistics, an **estimator** is simply a rule that for calculating a statistic (often about a wider population parameter)

Example

We want to estimate the average height (H) of U.S. adults (population) and have a random sample of 100 adults.

- Calculate the mean height of our sample ($\bar{H}$) to estimate the true mean height of the population (μ_H)
    - $\bar{H}$ is an **estimator** of μ_H
    - There are many estimators we *could* use to estimate μ_H

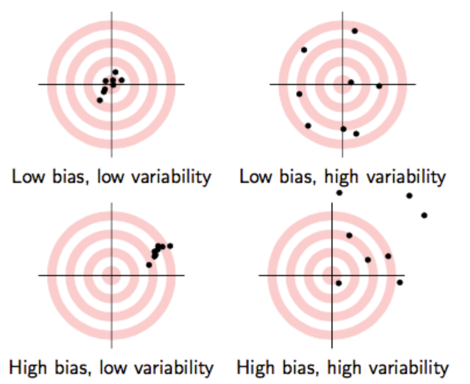
HOOD COLLEGE

- In statistics, an **estimator** is simply a rule that for calculating a statistic (often about a wider population parameter)

Example
We want to estimate the average height (H) of U.S. adults (population) and have a random sample of 100 adults.

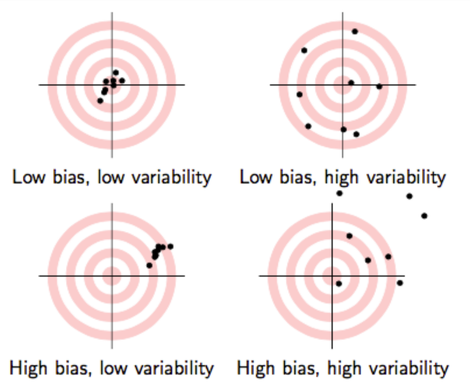
- Calculate the mean height of our sample ($\bar{H}$) to estimate the true mean height of the population (μ_H)
  - $\bar{H}$ is an **estimator** of μ_H
  - There are many estimators we *could* use to estimate μ_H
    - How about using the first value in our sample: H_1

- What makes one estimator (e.g. $\bar{H}$) better than another (e.g. H_1)?[5]

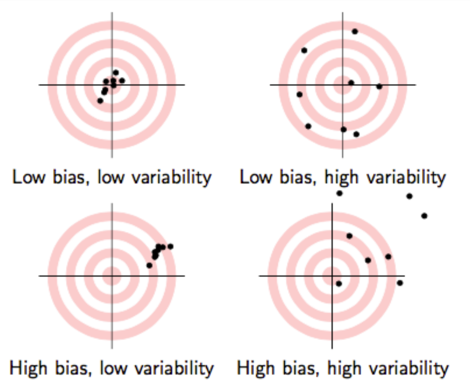


Bias & Efficiency

- What makes one estimator (e.g. $\bar{H}$) better than another (e.g. H_1)?[5]
    1. **Biasedness**: does the estimator give us the correct value *on average*?



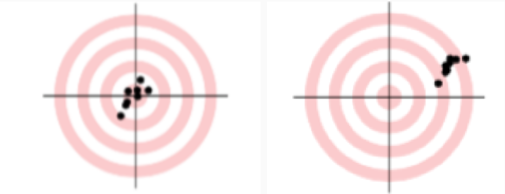
- What makes one estimator (e.g. $\bar{H}$) better than another (e.g. H_1)?[5]
    1. **Biasedness**: does the estimator give us the correct value *on average*?
    2. **Efficiency** an estimator with a smaller variance is better



Bias & Efficiency

- $\hat{\beta}_1$ **is an unbiased estimator of β_1 when X is exogenous**[6]: there is no systematic difference, on average, between sample values of $\hat{\beta}_1$ and the true population β_1

$$E[\hat{\beta}_1] = \beta_1$$

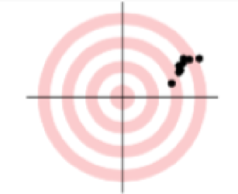


[6] See **handout** of unbiasedness on Blackboard for proofs

- $\hat{\beta}_1$ **is an** unbiased **estimator of** β_1 **when** X **is exogenous**[6]: there is no systematic difference, on average, between sample values of $\hat{\beta}_1$ and the true population β_1

$$E[\hat{\beta}_1] = \beta_1$$

- Does *not* mean that each sample gives us $\hat{\beta}_1 = \beta_1$, only the estimation *procedure* will, on average, yield the correct value



[6] See **handout** of unbiasedness on Blackboard for proofs

- Recall, an exogenous variables (X) is unrelated to other factors affecting Y, i.e.:

$$corr(X, \epsilon) = 0$$

- Recall, an exogenous variables (X) is unrelated to other factors affecting Y, i.e.:

$$corr(X, \epsilon) = 0$$

- Again, this is called the **Zero Conditional Mean Assumption**

$$E(\epsilon | X) = 0$$

- Recall, an exogenous variables (X) is unrelated to other factors affecting Y, i.e.:

$$corr(X, \epsilon) = 0$$

- Again, this is called the **Zero Conditional Mean Assumption**

$$E(\epsilon|X) = 0$$

- For any known value of X, the expected value of ϵ is 0.

- Recall, an exogenous variables (X) is unrelated to other factors affecting Y, i.e.:

$$corr(X, \epsilon) = 0$$

- Again, this is called the **Zero Conditional Mean Assumption**

$$E(\epsilon|X) = 0$$

- For any known value of X, the expected value of ϵ is 0.
- Knowing the value of X must tell us *nothing* about the value of ϵ (anything else relevant to Y other than X)

- Recall, an exogenous variables (X) is unrelated to other factors affecting Y, i.e.:

$$corr(X, \epsilon) = 0$$

- Again, this is called the **Zero Conditional Mean Assumption**

$$E(\epsilon | X) = 0$$

- For any known value of X, the expected value of ϵ is 0.
- Knowing the value of X must tell us *nothing* about the value of ϵ (anything else relevant to Y other than X)
- We can then confidently assert causation: $X \longrightarrow Y$

Endogeneity and Bias

- Nearly all independent variables are endogenous, they are related to the error term ϵ

$$corr(X, \epsilon) \neq 0$$

- Nearly all independent variables are endogenous, they are related to the error term ϵ

$$corr(X, \epsilon) \neq 0$$

Example
Suppose we estimate the following relationship:

$$\text{Violent crimes}_t = \beta_0 + \beta_1 \text{Ice cream sales}_t + \epsilon_t$$

- Nearly all independent variables are endogenous, they are related to the error term ϵ

$$corr(X, \epsilon) \neq 0$$

Example
Suppose we estimate the following relationship:

$$\text{Violent crimes}_t = \beta_0 + \beta_1 \text{Ice cream sales}_t + \epsilon_t$$

- We find $\hat{\beta}_1 > 0$

- Nearly all independent variables are endogenous, they are related to the error term ϵ

$$corr(X, \epsilon) \neq 0$$

Example
Suppose we estimate the following relationship:

$$\text{Violent crimes}_t = \beta_0 + \beta_1 \text{Ice cream sales}_t + \epsilon_t$$

- We find $\hat{\beta}_1 > 0$
- Does this mean Ice cream sales$\rightarrow$Violent crimes?

- The true expected value of $\hat{\beta}_1$ is actually[7]:

$$E[\hat{\beta}_1] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

[7] See **handout** on unbiasedness for proof

- The true expected value of $\hat{\beta}_1$ is actually[7]:

$$E[\hat{\beta}_1] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

- Takeaways:

[7] See **handout** on unbiasedness for proof

- The true expected value of $\hat{\beta}_1$ is actually[7]:

$$E[\hat{\beta}_1] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

- Takeaways:
  - If X is exogenous: $corr(X, \epsilon) = 0$, we're just left with β_1

[7] See **handout** on unbiasedness for proof

- The true expected value of $\hat{\beta}_1$ is actually[7]:

$$E[\hat{\beta}_1] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

- Takeaways:
  - If X is exogenous: $corr(X, \epsilon) = 0$, we're just left with β_1
  - The larger $corr(X, \epsilon)$ is, larger bias: $\left(E[\hat{\beta}_1] - \beta_1\right)$

[7] See **handout** on unbiasedness for proof

- The true expected value of $\hat{\beta}_1$ is actually[7]:

$$E[\hat{\beta}_1] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

- Takeaways:
    - If X is exogenous: $corr(X, \epsilon) = 0$, we're just left with β_1
    - The larger $corr(X, \epsilon)$ is, larger bias: $\left(E[\hat{\beta}_1] - \beta_1\right)$
    - We can also "sign" the direction of the bias based on $corr(X, \epsilon)$

[7] See **handout** on unbiasedness for proof

- The true expected value of $\hat{\beta}_1$ is actually[7]:

$$E[\hat{\beta}_1] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

- Takeaways:
  - If X is exogenous: $corr(X, \epsilon) = 0$, we're just left with β_1
  - The larger $corr(X, \epsilon)$ is, larger bias: $\left(E[\hat{\beta}_1] - \beta_1\right)$
  - We can also "sign" the direction of the bias based on $corr(X, \epsilon)$
    - **Positive** $corr(X, \epsilon)$ overestimates the true β_1 ($\hat{\beta}_1$ is too high)

[7]See **handout** on unbiasedness for proof

- The true expected value of $\hat{\beta}_1$ is actually[7]:

$$E[\hat{\beta}_1] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

- Takeaways:
  - If X is exogenous: $corr(X, \epsilon) = 0$, we're just left with β_1
  - The larger $corr(X, \epsilon)$ is, larger bias: $\left(E[\hat{\beta}_1] - \beta_1\right)$
  - We can also "sign" the direction of the bias based on $corr(X, \epsilon)$
    - **Positive** $corr(X, \epsilon)$ overestimates the true β_1 ($\hat{\beta}_1$ is too high)
    - **Negative** $corr(X, \epsilon)$ underestimates the true β_1 ($\hat{\beta}_1$ is too low)

[7] See **handout** on unbiasedness for proof

Example

$$wages_i = \beta_0 + \beta_1 education_i + \epsilon$$

- Is this an accurate reflection of $educ \longrightarrow wages$?

Example

$$wages_i = \beta_0 + \beta_1 education_i + \epsilon$$

- Is this an accurate reflection of $educ \longrightarrow wages$?
- Does $E\left[\epsilon \middle| education\right] = 0$?

Example

$$wages_i = \beta_0 + \beta_1 education_i + \epsilon$$

- Is this an accurate reflection of $educ \rightarrow wages$?
- Does $E[\epsilon | education] = 0$?
- What would $E[\epsilon | education] > 0$ mean?

Example

$$\text{per capita cigarette consumption} = \beta_0 + \beta_1 \text{State cig tax rate} + \epsilon$$

- Is this an accurate reflection of $tax \rightarrow cons$?

Example

$$\text{per capita cigarette consumption} = \beta_0 + \beta_1 \text{State cig tax rate} + \epsilon$$

- Is this an accurate reflection of $tax \longrightarrow cons$?
- Does $E[\epsilon | tax] = 0$?

Example

$$\text{per capita cigarette consumption} = \beta_0 + \beta_1 \text{State cig tax rate} + \epsilon$$

- Is this an accurate reflection of $tax \rightarrow cons$?
- Does $E[\epsilon | tax] = 0$?
- What would $E[\epsilon | tax] > 0$ mean?

Exogeneity and RCTs

- Think about an idealized randomized controlled experiment

- Think about an idealized randomized controlled experiment
- Subjects randomly assigned to treatment or control group

Exogeneity and RCTs

- Think about an idealized randomized controlled experiment
- Subjects randomly assigned to treatment or control group
    - Implies knowing whether someone is treated (X) tells us nothing about their personal characteristics (ϵ)

- Think about an idealized randomized controlled experiment
- Subjects randomly assigned to treatment or control group
  - Implies knowing whether someone is treated (X) tells us nothing about their personal characteristics (ϵ)
  - Random assignment makes ϵ independent of X, so

$$corr(X, \epsilon) = 0 \text{ and } E[\epsilon|X] = 0$$