

LECTURE 10: OMITTED VARIABLE BIAS

ECON 480 - ECONOMETRICS - FALL 2018

Ryan Safner

October 23, 2018





HOOD
COLLEGE

OMITTED VARIABLE BIAS

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i$$

- Error term, ϵ_i , includes all other variables that affect Y_i

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i$$

- Error term, ϵ_i , includes all other variables that affect Y_i
- Every regression has always omitted variables assumed into the error term (ϵ_i)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i$$

- Error term, ϵ_i , includes **all other variables that affect Y_i**
- Every regression has always **omitted variables** assumed into the error term (ϵ_i)
 - Often unobservable or hard to measure (e.g. innate ability, the weather at the time, etc.)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i$$

- Error term, ϵ_i , includes **all other variables that affect Y_i**
- Every regression has always **omitted variables** assumed into the error term (ϵ_i)
 - Often unobservable or hard to measure (e.g. innate ability, the weather at the time, etc.)
- Again, we assume ϵ_i is **random** with $E[\epsilon|X] = 0$ and $\text{var}(\epsilon) = \sigma_\epsilon^2$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i$$

- Error term, ϵ_i , includes **all other variables that affect Y_i**
- Every regression has always **omitted variables** assumed into the error term (ϵ_i)
 - Often unobservable or hard to measure (e.g. innate ability, the weather at the time, etc.)
- Again, we assume ϵ_i is **random** with $E[\epsilon|X] = 0$ and $\text{var}(\epsilon) = \sigma_\epsilon^2$
- *Sometimes* the omission of these variables can **bias** OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$)

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y
 - i.e. Z is included in the error term, ϵ_i

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y
 - i.e. Z is included in the error term, ϵ_i
 2. Z is correlated with the regressor X

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y
 - i.e. Z is included in the error term, ϵ_i
 2. Z is correlated with the regressor X
 - i.e. $\text{corr}(X, Z) \neq 0$

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y
 - i.e. Z is included in the error term, ϵ_i
 2. Z is correlated with the regressor X
 - i.e. $\text{corr}(X, Z) \neq 0$
- Omitted variable bias makes X endogenous

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y
 - i.e. Z is included in the error term, ϵ_i
 2. Z is correlated with the regressor X
 - i.e. $\text{corr}(X, Z) \neq 0$
- Omitted variable bias makes X endogenous
 - $E(\epsilon_i | X_i) \neq 0 \implies$ knowing X tells you something about ϵ

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y
 - i.e. Z is included in the error term, ϵ_i
 2. Z is correlated with the regressor X
 - i.e. $\text{corr}(X, Z) \neq 0$
- Omitted variable bias makes X endogenous
 - $E(\epsilon_i | X_i) \neq 0 \implies$ knowing X tells you something about ϵ
 - Thus, X tells you something about Y not by way of X !

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

OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y
 - i.e. Z is included in the error term, ϵ_i
 2. Z is correlated with the regressor X
 - i.e. $\text{corr}(X, Z) \neq 0$
- Omitted variable bias makes X endogenous
 - $E(\epsilon_i | X_i) \neq 0 \implies$ knowing X tells you something about ϵ
 - Thus, X tells you something about Y not by way of X !
- Therefore, $\hat{\beta}_1$ is biased and systematically over- or under-estimates the true relationship β_1

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



OMITTED VARIABLE BIAS

- Omitted variable bias for some omitted variable Z exists if two conditions are met:
 1. Z is a determinant of Y
 - i.e. Z is included in the error term, ϵ_i
 2. Z is correlated with the regressor X
 - i.e. $\text{corr}(X, Z) \neq 0$
- Omitted variable bias makes X endogenous
 - $E(\epsilon_i | X_i) \neq 0 \implies$ knowing X tells you something about ϵ
 - Thus, X tells you something about Y not by way of X !
- Therefore, $\hat{\beta}_1$ is biased and systematically over- or under-estimates the true relationship β_1
 - $\hat{\beta}_1$ “picks up” both the effect of $X \rightarrow Y$ and the effect of $Z \rightarrow Y$ through X

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

Example

$$\widehat{\text{Test Score}} = \hat{\beta}_0 + \hat{\beta}_1 \text{STR}_i + \epsilon_i$$

Example

$$\widehat{\text{Test Score}} = \hat{\beta}_0 + \hat{\beta}_1 \text{STR}_i + \epsilon_i$$

- Z_i : Time of Day of the Test (?)

Example

$$\widehat{\text{Test Score}} = \hat{\beta}_0 + \hat{\beta}_1 \text{STR}_i + \epsilon_i$$

- Z_i : Time of Day of the Test (?)
- Z_i : Parking Space per Student (?)

Example

$$\widehat{\text{Test Score}} = \hat{\beta}_0 + \hat{\beta}_1 \text{STR}_i + \epsilon_i$$

- Z_i : Time of Day of the Test (?)
- Z_i : Parking Space per Student (?)
- Z_i : Percent of ESL Students (?)

RECALL: ENDOGENEITY AND BIAS

- The true expected value of $\hat{\beta}_1$ is actually¹:

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

¹See handout on unbiasedness for proof

RECALL: ENDOGENEITY AND BIAS

- The true expected value of $\hat{\beta}_1$ is actually¹:

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

- Takeaways:

¹See handout on unbiasedness for proof

RECALL: ENDOGENEITY AND BIAS

- The true expected value of $\hat{\beta}_1$ is actually¹:

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

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

¹See handout on unbiasedness for proof

RECALL: ENDOGENEITY AND BIAS

- The true expected value of $\hat{\beta}_1$ is actually¹:

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

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

¹See **handout** on unbiasedness for proof

RECALL: ENDOGENEITY AND BIAS

- The true expected value of $\hat{\beta}_1$ is actually¹:

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

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



¹See handout on unbiasedness for proof

RECALL: ENDOGENEITY AND BIAS

- The true expected value of $\hat{\beta}_1$ is actually¹:

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

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

¹See handout on unbiasedness for proof

RECALL: ENDOGENEITY AND BIAS

- The true expected value of $\hat{\beta}_1$ is actually¹:

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

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

¹See handout on unbiasedness for proof

ENDOGENEITY AND BIAS: CORRELATIONS

- Here is where checking correlations between variables helps:

```
# Select only the three variables we want (there are many)
CAcorr<-subset(CASchool, select=c("str","testscr","el_pct"))
```

```
# Make a correlation table
```

```
corr<-cor(CAcorr)
```

```
corr
```

```
##           str    testscr    el_pct
## str     1.0000000 -0.2263628  0.1876424
## testscr -0.2263628  1.0000000 -0.6441237
## el_pct   0.1876424 -0.6441237  1.0000000
```



```
library("stargazer")
stargazer(corr, type="latex", header=FALSE, float=FALSE)
```

	str	testscr	el_pct
str	1	-0.226	0.188
testscr	-0.226	1	-0.644
el_pct	0.188	-0.644	1

- %EL is strongly correlated with Test Score (Condition 1)

```
library("stargazer")
stargazer(corr, type="latex", header=FALSE, float=FALSE)
```

	str	testscr	el_pct
str	1	-0.226	0.188
testscr	-0.226	1	-0.644
el_pct	0.188	-0.644	1

- %EL is strongly correlated with Test Score (Condition 1)
- %EL is reasonably correlated with STR (Condition 2)

ENDOGENEITY AND BIAS: LOOKING AT CONDITIONAL DISTRIBUTIONS

```
summary(CASchool$testscr)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    605.5    640.0   654.5    654.2   666.7    706.8
```

```
# find the median of %EL
```

```
summary(CASchool$el_pct)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    0.000   1.941   8.778  15.768  22.970  85.540
```



ENDOGENEITY AND BIAS: LOOKING AT CONDITIONAL DISTRIBUTIONS II

```
# look at test scores for districts with more than median %EL  
summary(CASchool$testscr[CASchool$el_pct<8.7])
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.  
##    634.0    653.2   663.9    664.4    672.6    706.8
```

```
# look at test scores for districts with median or less %EL  
summary(CASchool$testscr[CASchool$el_pct>=8.7])
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.  
##    605.5    632.5   643.2    644.2    655.7    691.4
```

- Test scores are *lower* in districts with relatively *high* %EL!

ENDOGENEITY AND BIAS: LOOKING AT CONDITIONAL DISTRIBUTIONS II

```
# look at test scores for districts with more than median %EL  
summary(CASchool$testscr[CASchool$el_pct<8.7])
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.  
##    634.0    653.2   663.9    664.4    672.6    706.8
```

```
# look at test scores for districts with median or less %EL  
summary(CASchool$testscr[CASchool$el_pct>=8.7])
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.  
##    605.5    632.5   643.2    644.2    655.7    691.4
```

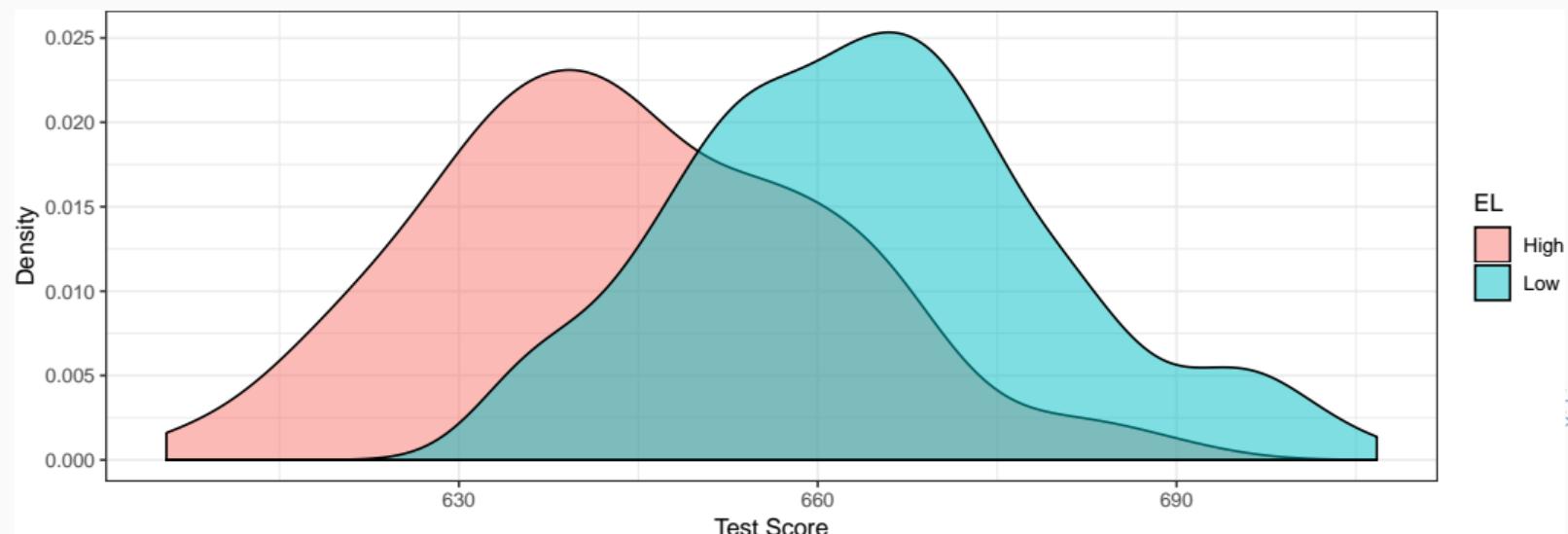
- Test scores are *lower* in districts with relatively *high* %EL!
- Test scores are *higher* in districts with relatively *low* %EL!

```
# Very useful function:  
# ifelse(conditions, do.this.if.conditions.are.met, this.if.not)  
  
CASchool$EL<-ifelse(CASchool$el_pct>8.7,"High","Low")  
  
# i.e. I am making a new variable in the dataframe called EL  
# defining it to be "High" if el_pct>8.7  
# defining it to be "Low" if el_pct is NOT >8.7
```



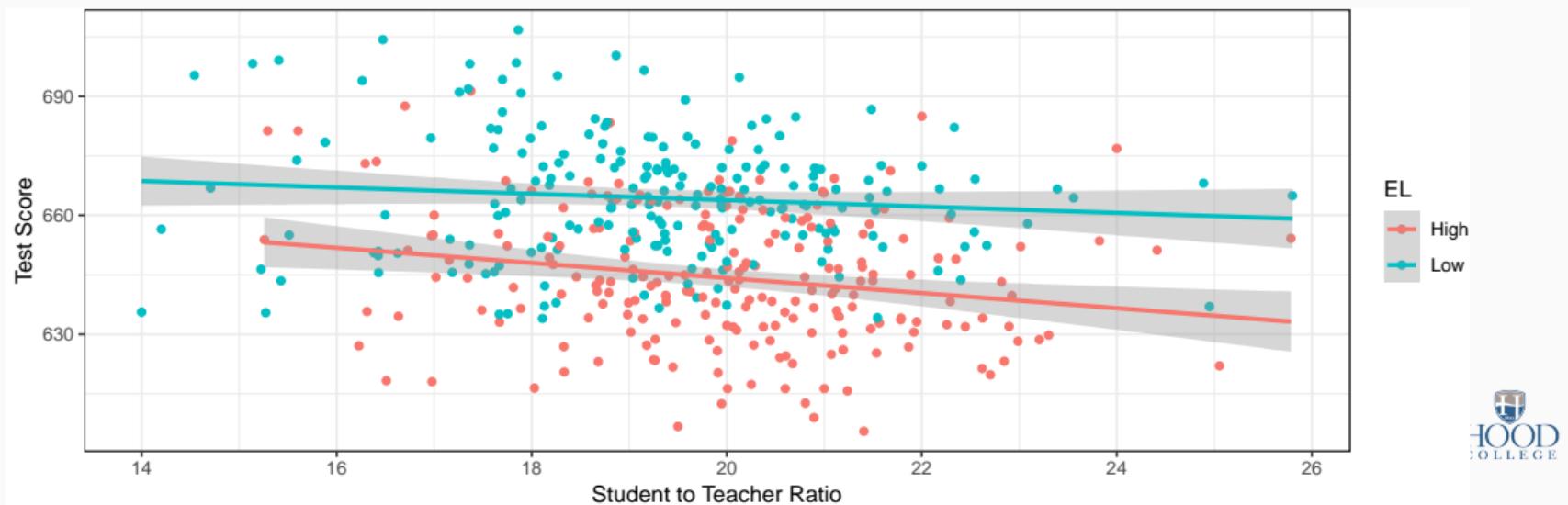
ENDOGENEITY AND BIAS: LOOKING AT CONDITIONAL DISTRIBUTIONS IV

```
library("ggplot2")  
  
ggplot(CASchool, aes(x=testscr, fill=EL))+  
  geom_density(alpha=0.5)+theme_bw() +xlab("Test Score") +ylab("Density")
```



ENDOGENEITY AND BIAS: LOOKING AT CONDITIONAL DISTRIBUTIONS V

```
ggplot(CASchool, aes(x=str,y=testscr,color=EL))+
  geom_point() + geom_smooth(method="lm") +
  theme_bw() + xlab("Student to Teacher Ratio") + ylab("Test Score")
```



OMITTED VARIABLE BIAS IN CLASS SIZE EXAMPLE

$$E[\hat{\beta}_1] = \beta_1 + \underbrace{corr(X, \epsilon) \frac{\sigma_\epsilon}{\sigma_X}}_{\text{Omitted Variable Bias}}$$

²Hard to think about...but you'll see when we run the different regressions below!

OMITTED VARIABLE BIAS IN CLASS SIZE EXAMPLE

$$E[\hat{\beta}_1] = \beta_1 + \underbrace{corr(X, \epsilon) \frac{\sigma_\epsilon}{\sigma_X}}_{\text{Omitted Variable Bias}}$$

- $corr(STR, \epsilon)$ is positive (through %EL)

²Hard to think about...but you'll see when we run the different regressions below!

OMITTED VARIABLE BIAS IN CLASS SIZE EXAMPLE

$$E[\hat{\beta}_1] = \beta_1 + \underbrace{corr(X, \epsilon) \frac{\sigma_\epsilon}{\sigma_X}}_{\text{Omitted Variable Bias}}$$

- $corr(STR, \epsilon)$ is positive (through %EL)
- $corr(\epsilon, \text{Test Score})$ is negative (through %EL)

²Hard to think about...but you'll see when we run the different regressions below!

OMITTED VARIABLE BIAS IN CLASS SIZE EXAMPLE

$$E[\hat{\beta}_1] = \beta_1 + \underbrace{corr(X, \epsilon) \frac{\sigma_\epsilon}{\sigma_X}}_{\text{Omitted Variable Bias}}$$

- $corr(STR, \epsilon)$ is positive (through %EL)
- $corr(\epsilon, \text{Test Score})$ is negative (through %EL)
- β_1 is negative (between Test Score and STR)

²Hard to think about...but you'll see when we run the different regressions below!

OMITTED VARIABLE BIAS IN CLASS SIZE EXAMPLE

$$E[\hat{\beta}_1] = \beta_1 + \underbrace{corr(X, \epsilon) \frac{\sigma_\epsilon}{\sigma_X}}_{\text{Omitted Variable Bias}}$$

- $corr(STR, \epsilon)$ is positive (through %EL)
- $corr(\epsilon, \text{Test Score})$ is negative (through %EL)
- β_1 is negative (between Test Score and STR)
- Bias is positive, but since β_1 is negative, it is made a *more* negative number than it should be²



²Hard to think about...but you'll see when we run the different regressions below!

OMITTED VARIABLE BIAS IN CLASS SIZE EXAMPLE

$$E[\hat{\beta}_1] = \beta_1 + \underbrace{corr(X, \epsilon) \frac{\sigma_\epsilon}{\sigma_X}}_{\text{Omitted Variable Bias}}$$

- $corr(STR, \epsilon)$ is positive (through %EL)
- $corr(\epsilon, \text{Test Score})$ is negative (through %EL)
- β_1 is negative (between Test Score and STR)
- Bias is positive, but since β_1 is negative, it is made a *more* negative number than it should be²
 - Implies that β_1 overstates the effect of reducing STR on improving Test Scores

²Hard to think about...but you'll see when we run the different regressions below!

OMITTED VARIABLE BIAS: MESSING WITH CAUSALITY

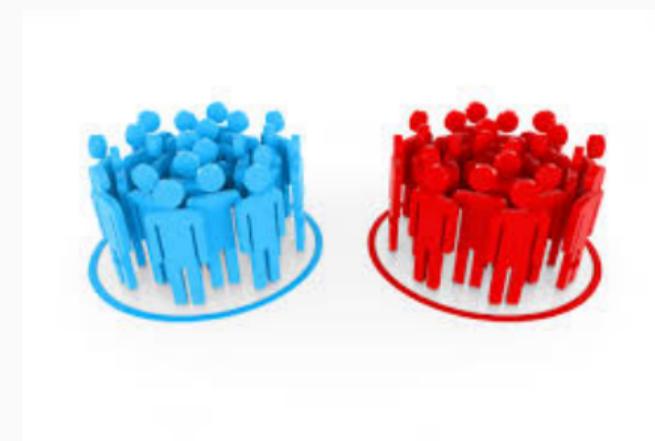
- If school districts with higher Test Scores happen to have both lower STR **AND** districts with smaller STR sizes tend to have less %EL...



- If school districts with higher Test Scores happen to have both lower STR **AND** districts with smaller STR sizes tend to have less %EL...
- How can we say $\hat{\beta}_1$ estimates the **marginal effect** of $\Delta STR \rightarrow \Delta \text{Test Score}$?

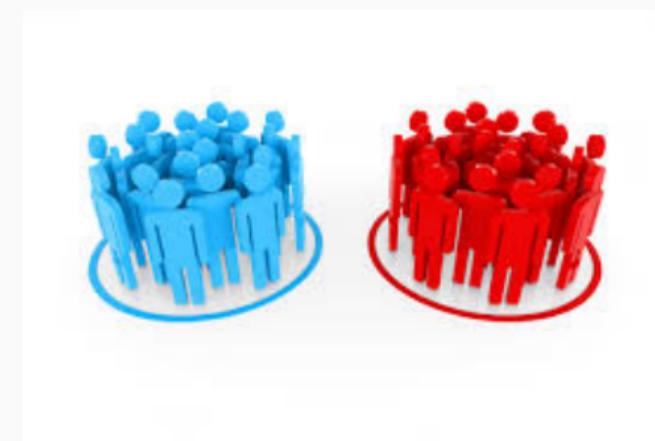
OMITTED VARIABLE BIAS: MESSING WITH CAUSALITY II

- Recall our best working definition of causality: result of ideal random controlled trials (RCTs)



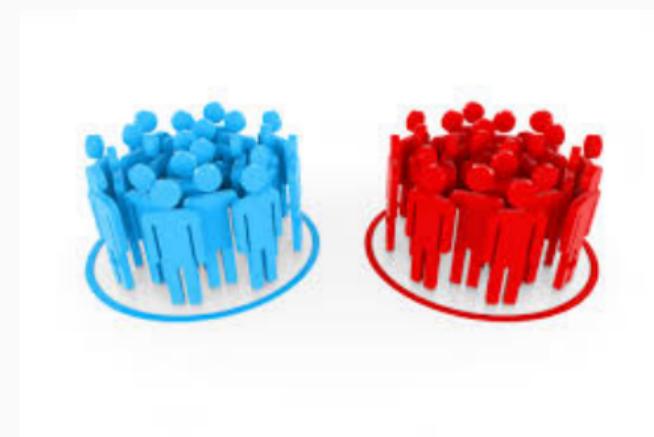
OMITTED VARIABLE BIAS: MESSING WITH CAUSALITY II

- Recall our best working definition of causality: result of ideal **random controlled trials (RCTs)**
 - Randomly assign experimental units (e.g. people, cities, etc) into two (or more) groups:



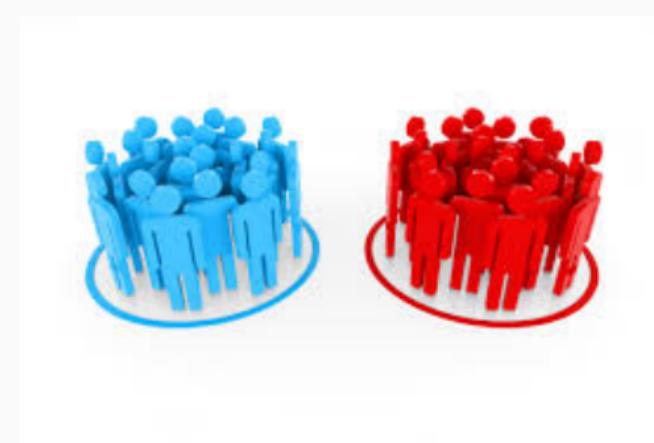
OMITTED VARIABLE BIAS: MESSING WITH CAUSALITY II

- Recall our best working definition of causality: result of ideal **random controlled trials (RCTs)**
 - Randomly assign experimental units (e.g. people, cities, etc) into two (or more) groups:
 - **Treatment group(s)**: gets a (certain type or level of) treatment



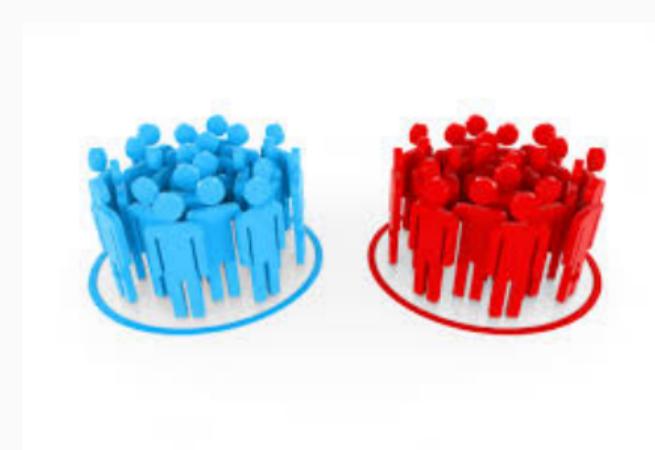
OMITTED VARIABLE BIAS: MESSING WITH CAUSALITY II

- Recall our best working definition of causality: result of ideal **random controlled trials (RCTs)**
 - Randomly assign experimental units (e.g. people, cities, etc) into two (or more) groups:
 - **Treatment group(s)**: gets a (certain type or level of) treatment
 - **Control group**: gets *no* treatment(s)



OMITTED VARIABLE BIAS: MESSING WITH CAUSALITY II

- Recall our best working definition of causality: result of ideal **random controlled trials (RCTs)**
 - Randomly assign experimental units (e.g. people, cities, etc) into two (or more) groups:
 - **Treatment group(s)**: gets a (certain type or level of) treatment
 - **Control group**: gets *no* treatment(s)
 - Compare results of two groups to get the causal effect of treatment (on average)



Example

- Imagine an ideal RCT for measuring the effect of STR on Test Score



Example

- Imagine an ideal RCT for measuring the effect of STR on Test Score
- School districts would be **randomly assigned** a STR



Example

- Imagine an ideal RCT for measuring the effect of STR on Test Score
- School districts would be **randomly assigned** a STR
- With random assignment, all factors in ϵ (parental income, family size, # of siblings, English proficiency, etc) are distributed *independently* of class size



Example

- Imagine an ideal RCT for measuring the effect of STR on Test Score
- School districts would be **randomly assigned** a STR
- With random assignment, all factors in ϵ (parental income, family size, # of siblings, English proficiency, etc) are distributed *independently* of class size
- Thus, $\text{corr}(\text{STR}, \epsilon) = 0$ and $E(\epsilon_i|\text{STR}_i) = 0$: exogeneity!



Example

- Imagine an ideal RCT for measuring the effect of STR on Test Score
- School districts would be **randomly assigned** a STR
- With random assignment, all factors in ϵ (parental income, family size, # of siblings, English proficiency, etc) are distributed *independently* of class size
- Thus, $\text{corr}(\text{STR}, \epsilon) = 0$ and $E(\epsilon_i | \text{STR}_i) = 0$: exogeneity!
- The resulting β_1 is an unbiased estimate for the marginal effect of $\Delta\text{STR} \rightarrow \Delta\text{Test Score}$



BUT WE RARELY HAVE RCTs

- But our data is *not* an RCT, it is observational data!



BUT WE RARELY HAVE RCTs

- But our data is *not* an RCT, it is observational data!
- “Treatment” of having a large or small class size is **NOT** randomly assigned!



BUT WE RARELY HAVE RCTs

- But our data is *not* an RCT, it is observational data!
- “Treatment” of having a large or small class size is **NOT** randomly assigned!
- Again consider %EL: plausibly fits the criteria of O.V. bias!



BUT WE RARELY HAVE RCTs

- But our data is *not* an RCT, it is observational data!
- “Treatment” of having a large or small class size is **NOT** randomly assigned!
- Again consider %EL: plausibly fits the criteria of O.V. bias!
 1. %EL is a determinant of Test Score



BUT WE RARELY HAVE RCTs

- But our data is *not* an RCT, it is observational data!
- “Treatment” of having a large or small class size is **NOT** randomly assigned!
- Again consider %EL: plausibly fits the criteria of O.V. bias!
 1. %EL is a determinant of Test Score
 2. %EL is correlated with STR



BUT WE RARELY HAVE RCTs

- But our data is *not* an RCT, it is observational data!
- “Treatment” of having a large or small class size is **NOT** randomly assigned!
- Again consider %EL: plausibly fits the criteria of O.V. bias!
 1. %EL is a determinant of Test Score
 2. %EL is correlated with STR
- Thus, “control” group and “treatment” group differs systematically!



BUT WE RARELY HAVE RCTs

- But our data is *not* an RCT, it is observational data!
- “Treatment” of having a large or small class size is **NOT** randomly assigned!
- Again consider %EL: plausibly fits the criteria of O.V. bias!
 1. %EL is a determinant of Test Score
 2. %EL is correlated with STR
- Thus, “control” group and “treatment” group differs systematically!
 - Small STR also tend to have lower %EL; large STR also tend to have higher %EL



BUT WE RARELY HAVE RCTs

- But our data is *not* an RCT, it is observational data!
- “Treatment” of having a large or small class size is **NOT** randomly assigned!
- Again consider %EL: plausibly fits the criteria of O.V. bias!
 1. %EL is a determinant of Test Score
 2. %EL is correlated with STR
- Thus, “control” group and “treatment” group differs systematically!
 - Small STR also tend to have lower %EL; large STR also tend to have higher %EL
 - **Selection bias:** $\text{corr}(\text{STR}, \%EL) \neq 0, E[\epsilon_i | \text{STR}_i] \neq 0$



THERE'S ANOTHER WAY TO NEUTRALIZE OVB

- Look at effect of STR on Test Score by comparing districts with the **same %EL**.



THERE'S ANOTHER WAY TO NEUTRALIZE OVB

- Look at effect of STR on Test Score by comparing districts with the **same %EL**.
 - Eliminates differences in %EL between high and low STR classes



THERE'S ANOTHER WAY TO NEUTRALIZE OVB

- Look at effect of STR on Test Score by comparing districts with the **same %EL**.
 - Eliminates differences in %EL between high and low STR classes
 - “As if” we had a control group! Hold %EL constant



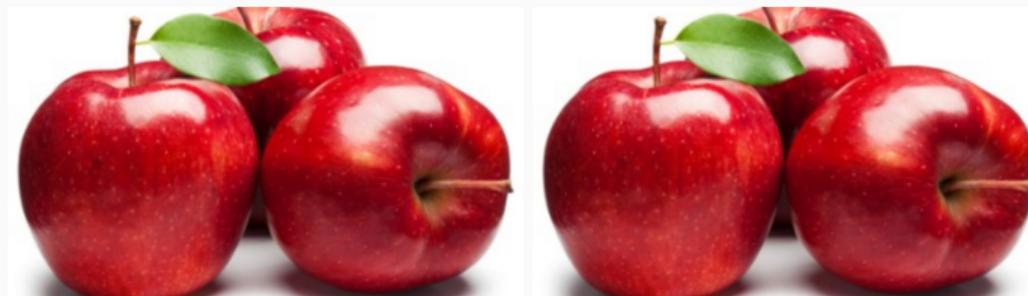
THERE'S ANOTHER WAY TO NEUTRALIZE OVB

- Look at effect of STR on Test Score by comparing districts with the **same %EL**.
 - Eliminates differences in %EL between high and low STR classes
 - “As if” we had a control group! Hold %EL constant
- The simple fix is just to **not omit %EL!**



THERE'S ANOTHER WAY TO NEUTRALIZE OVB

- Look at effect of STR on Test Score by comparing districts with the **same %EL**.
 - Eliminates differences in %EL between high and low STR classes
 - “As if” we had a control group! Hold %EL constant
- The simple fix is just to **not omit %EL!**
 - Make it *another* independent variable on the righthand side of the regression



THE MULTIVARIATE REGRESSION MODEL

THE POPULATION MULTIVARIATE REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

THE POPULATION MULTIVARIATE REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- X_{1i} and X_{2i} are two independent variables (regressors)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- X_{1i} and X_{2i} are two independent variables (regressors)
- (Y_i, X_{1i}, X_{2i}) are the values of variables Y , X_1 , and X_2 for individual i

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- X_{1i} and X_{2i} are two independent variables (regressors)
- (Y_i, X_{1i}, X_{2i}) are the values of variables Y , X_1 , and X_2 for individual i
- β_0 : a population constant

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- X_{1i} and X_{2i} are two independent variables (regressors)
- (Y_i, X_{1i}, X_{2i}) are the values of variables Y , X_1 , and X_2 for individual i
- β_0 : a population constant
- β_1 : marginal effect on Y of a change in X_1 , holding X_2 constant

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- X_{1i} and X_{2i} are two independent variables (regressors)
- (Y_i, X_{1i}, X_{2i}) are the values of variables Y , X_1 , and X_2 for individual i
- β_0 : a population constant
- β_1 : marginal effect on Y of a change in X_1 , holding X_2 constant
- β_2 : marginal effect on Y of a change in X_2 , holding X_1 constant

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- X_{1i} and X_{2i} are two independent variables (regressors)
- (Y_i, X_{1i}, X_{2i}) are the values of variables Y , X_1 , and X_2 for individual i
- β_0 : a population constant
- β_1 : marginal effect on Y of a change in X_1 , holding X_2 constant
- β_2 : marginal effect on Y of a change in X_2 , holding X_1 constant
- ϵ_i : regression error (omitted variables)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Before the change

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Before the change

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$$

After the change

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Before the change

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$$

After the change

$$\Delta Y = \beta_1 \Delta X_1$$

The difference

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Before the change

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$$

After the change

$$\Delta Y = \beta_1 \Delta X_1$$

The difference

$$\frac{\Delta Y}{\Delta X_1} = \beta_1$$

Solving for β_1

$$\beta_1 = \frac{\Delta Y}{\Delta X_1} \text{ holding } X_2 \text{ constant}$$

$$\beta_1 = \frac{\Delta Y}{\Delta X_1} \text{ holding } X_2 \text{ constant}$$

Similarly, for β_2 :

$$\beta_2 = \frac{\Delta Y}{\Delta X_2} \text{ holding } X_1 \text{ constant}$$

$$\beta_1 = \frac{\Delta Y}{\Delta X_1} \text{ holding } X_2 \text{ constant}$$

Similarly, for β_2 :

$$\beta_2 = \frac{\Delta Y}{\Delta X_2} \text{ holding } X_1 \text{ constant}$$

And for the constant, β_0 :

β_0 = predicted value of Y when $X_1 = 0, X_2 = 0$

IF YOU LIKE YOUR INTUITIONS, YOU CAN KEEP THEM...BUT THEY'RE WRONG Now

- We have been envisioning OLS regressions as the equation of a line through a scatterplot of data on two variables, X and Y



IF YOU LIKE YOUR INTUITIONS, YOU CAN KEEP THEM...BUT THEY'RE WRONG Now

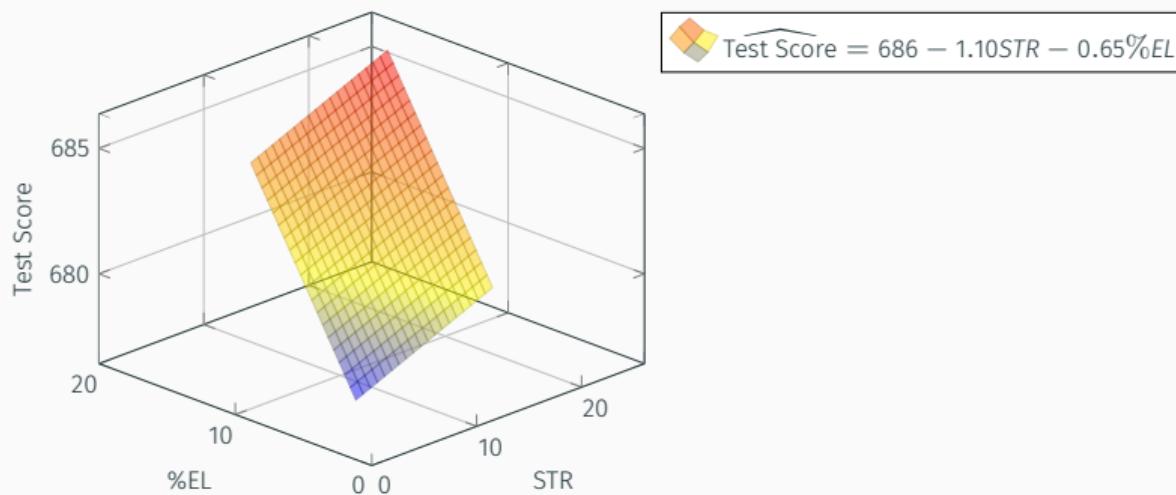
- We have been envisioning OLS regressions as the equation of a line through a scatterplot of data on two variables, X and Y
 - β_1 : "slope"

IF YOU LIKE YOUR INTUITIONS, YOU CAN KEEP THEM...BUT THEY'RE WRONG Now

- We have been envisioning OLS regressions as the equation of a line through a scatterplot of data on two variables, X and Y
 - β_1 : "slope"
 - β_0 : "y-intercept"

IF YOU LIKE YOUR INTUITIONS, YOU CAN KEEP THEM...BUT THEY'RE WRONG Now

- We have been envisioning OLS regressions as the equation of a line through a scatterplot of data on two variables, X and Y
 - β_1 : "slope"
 - β_0 : "y-intercept"
- With 3+ variables, it OLS regressions is no longer a "line" for us to estimate



- Alternatively, we can write the population regression equation as:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

THE “CONSTANT”

- Alternatively, we can write the population regression equation as:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Here, we added X_{0i} to β_0

THE “CONSTANT”

- Alternatively, we can write the population regression equation as:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Here, we added X_{0i} to β_0
- X_{0i} is a **constant regressor**, as we define $X_{0i} = 1$ for all i observations

THE “CONSTANT”

- Alternatively, we can write the population regression equation as:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Here, we added X_{0i} to β_0
- X_{0i} is a **constant regressor**, as we define $X_{0i} = 1$ for all i observations
- Likewise, β_0 is more generally called the **constant term** in the regression (instead of the “intercept”)

THE “CONSTANT”

- Alternatively, we can write the population regression equation as:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Here, we added X_{0i} to β_0
- X_{0i} is a **constant regressor**, as we define $X_{0i} = 1$ for all i observations
- Likewise, β_0 is more generally called the **constant term** in the regression (instead of the “intercept”)
- This may seem silly and trivial, but this will be important soon!

THE POPULATION MULTIPLE REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \epsilon_i$$

- In general, a multivariate model has k regressor variables

³ Note your Bailey textbook defines k to include both the number of variables plus the constant

THE POPULATION MULTIPLE REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \epsilon_i$$

- In general, a multivariate model has k regressor variables
- We estimate $k + 1$ parameters (including β_0 , the constant, when $\beta_1 = \beta_2 = \dots = \beta_k = 0$)³

³ Note your Bailey textbook defines k to include both the number of variables plus the constant

Example

$$\widehat{Consumption}_i = \beta_0 + \beta_1 Price_i + \beta_2 Income_i + CompsPrice_i + SubsPrice + \epsilon_i$$

- Let's see what you remember from micro(econ)!

Example

$$\widehat{Consumption}_i = \beta_0 + \beta_1 Price_i + \beta_2 Income_i + CompsPrice_i + SubsPrice + \epsilon_i$$

- Let's see what you remember from micro(econ)!
- What measures the **price effect**? What sign should it have?

Example

$$\widehat{Consumption}_i = \beta_0 + \beta_1 Price_i + \beta_2 Income_i + CompsPrice_i + SubsPrice + \epsilon_i$$

- Let's see what you remember from micro(econ)!
- What measures the **price effect**? What sign should it have?
- What measures the **income effect**? What should inferior, necessities, and luxuries look like?

Example

$$\widehat{Consumption}_i = \beta_0 + \beta_1 Price_i + \beta_2 Income_i + CompsPrice_i + SubsPrice + \epsilon_i$$

- Let's see what you remember from micro(econ)!
- What measures the **price effect**? What sign should it have?
- What measures the **income effect**? What should inferior, necessities, and luxuries look like?
- What measures the **cross-price effect**? What should substitutes and complements look like?

Example

$$\widehat{BeerCons}_i = 20 - 1.5Price_i + 1.25Income_i - 0.75WingsPrice_i + 1.3WinePrice_i$$

- Interpret each $\hat{\beta}$

```
# syntax: reg.object<-(y~x1+x2, data=mydf)  
# y goes first, then all of your x's after the ~, separated by +'s
```

Remember, we “regress Y on X’s”



MULTIVARIATE REGRESSION IN R II

```
multireg<-lm(testscr~str+el_pct, data=CASchool)
summary(multireg)

##
## Call:
## lm(formula = testscr ~ str + el_pct, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -48.845 -10.240  -0.308   9.815  43.461 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 686.03225    7.41131  92.566 < 2e-16 ***
## str         -1.10130    0.38028  -2.896  0.00398 **  
## el_pct      -0.64978    0.03934 -16.516 < 2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.46 on 417 degrees of freedom
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237 
## F-statistic: 155 on 2 and 417 DF,  p-value: < 2.2e-16
```

