3.3 — Omitted Variable Bias

ECON 480 • Econometrics • Fall 2021

Ryan Safner

Assistant Professor of Economics

safner@hood.edu

ryansafner/metricsF21

metricsF21.classes.ryansafner.com

Review: u

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- u_i includes **all other variables that affect Y**
- Every regression model always has **omitted variables** assumed in the error
    - Most are unobservable (hence "u")
    - **Examples**: innate ability, weather at the time, etc
- Again, we *assume* u is random, with $E[u|X] = 0$ and $var(u) = \sigma_u^2$
- *Sometimes*, omission of variables can **bias** OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$)

Omitted Variable Bias I

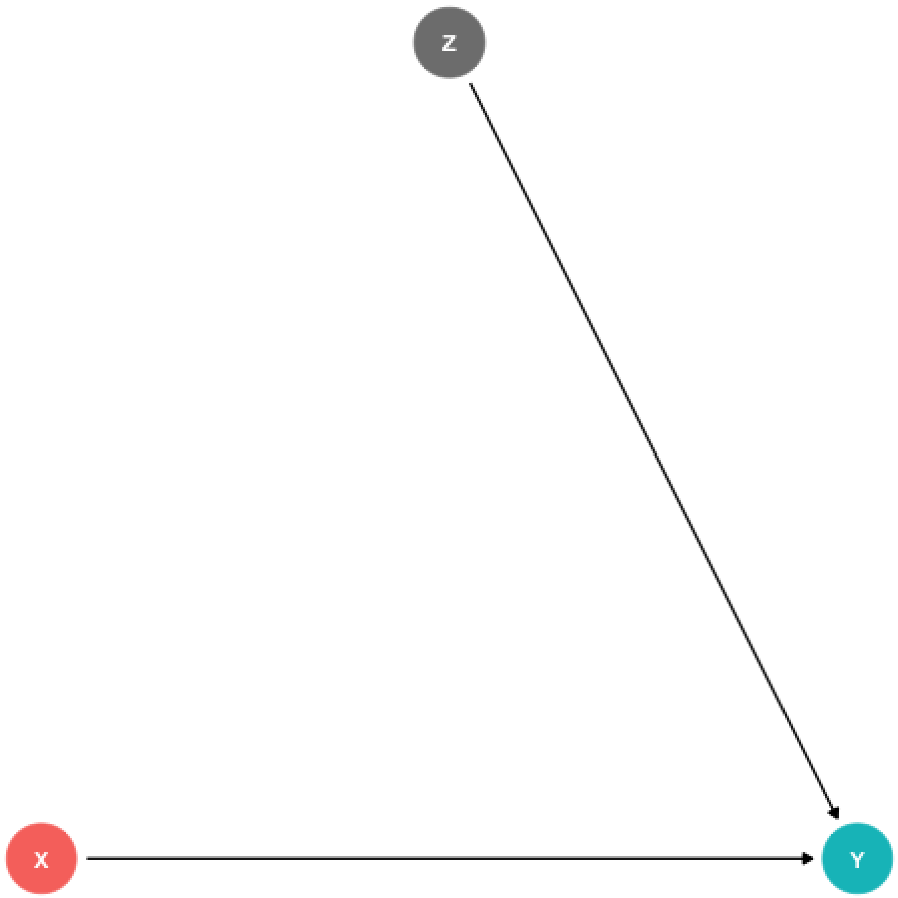
- **Omitted variable bias (OVB)** for some omitted variable $\mathbf{Z}$ exists if two conditions are met:

Omitted Variable Bias I

- **Omitted variable bias (OVB)** for some omitted variable $\mathbf{Z}$ exists if two conditions are met:

1. Z is a determinant of Y

- i.e. Z is in the error term, u_i



Omitted Variable Bias I

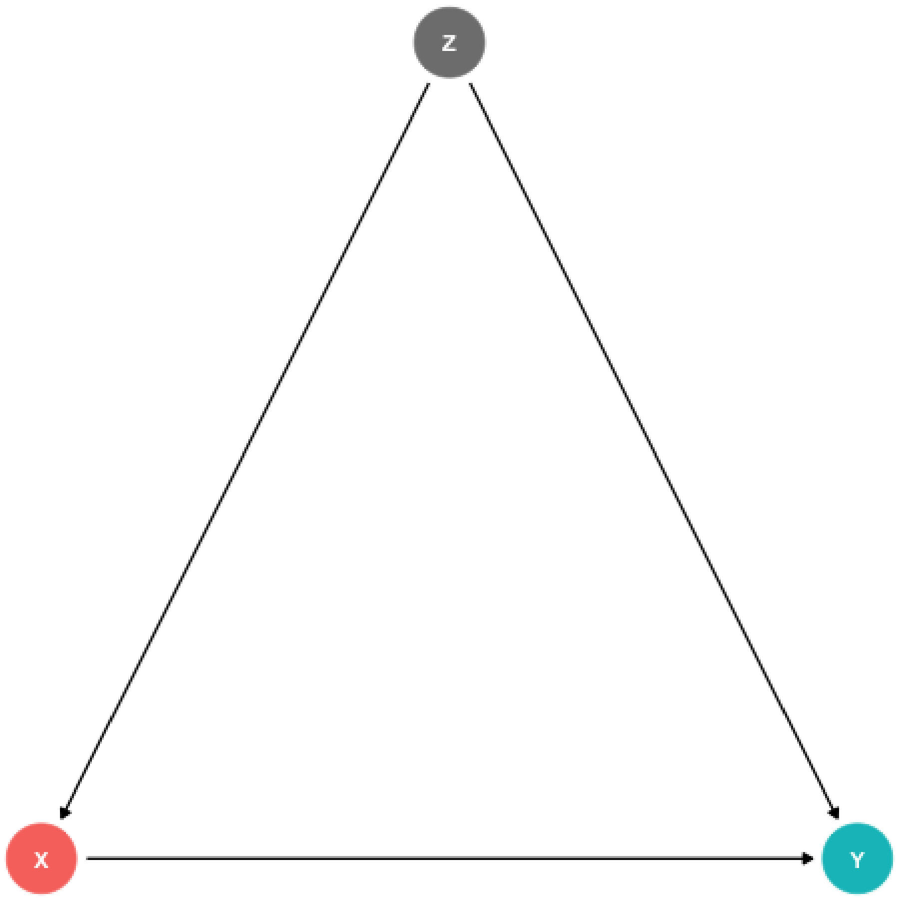
- **Omitted variable bias (OVB)** for some omitted variable $\mathbf{Z}$ exists if two conditions are met:

1. Z is a determinant of Y

- i.e. Z is in the error term, u_i

2. Z is correlated with the regressor X

- i.e. $cor(X, Z) \neq 0$
- implies $cor(X, u) \neq 0$
- implies **X is endogenous**

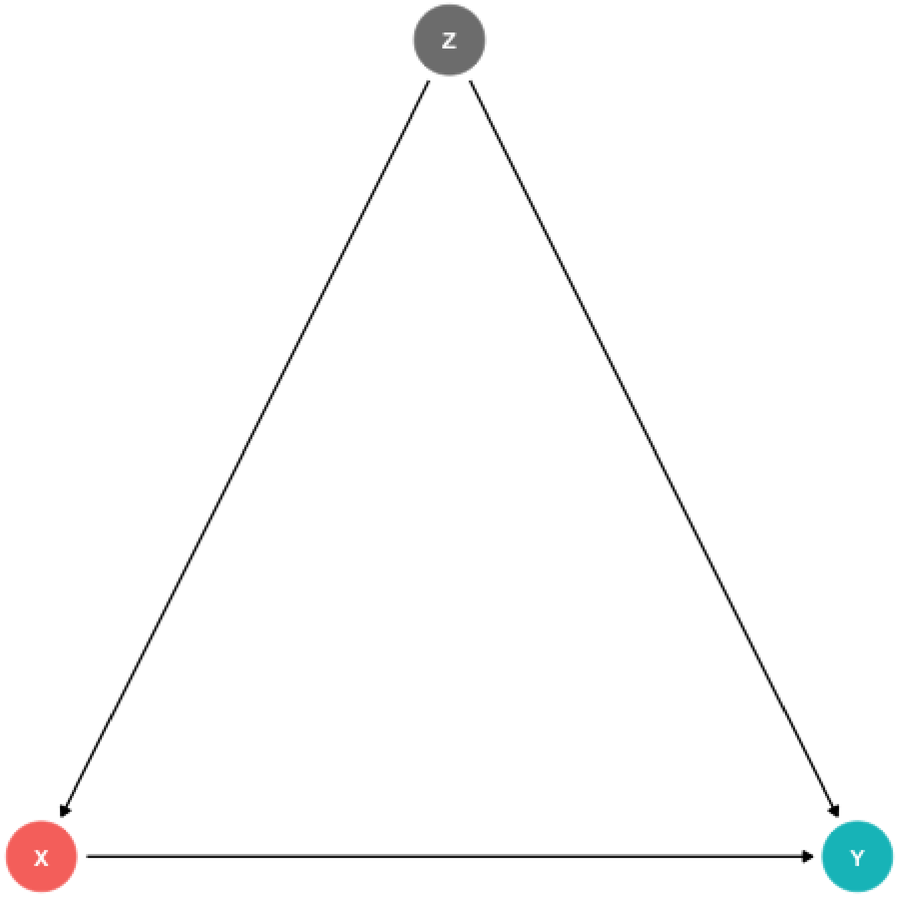


Omitted Variable Bias II

- Omitted variable bias makes X **endogenous**
- Violates **zero conditional mean assumption**

$$E(u_i|X_i) \neq 0 \implies$$

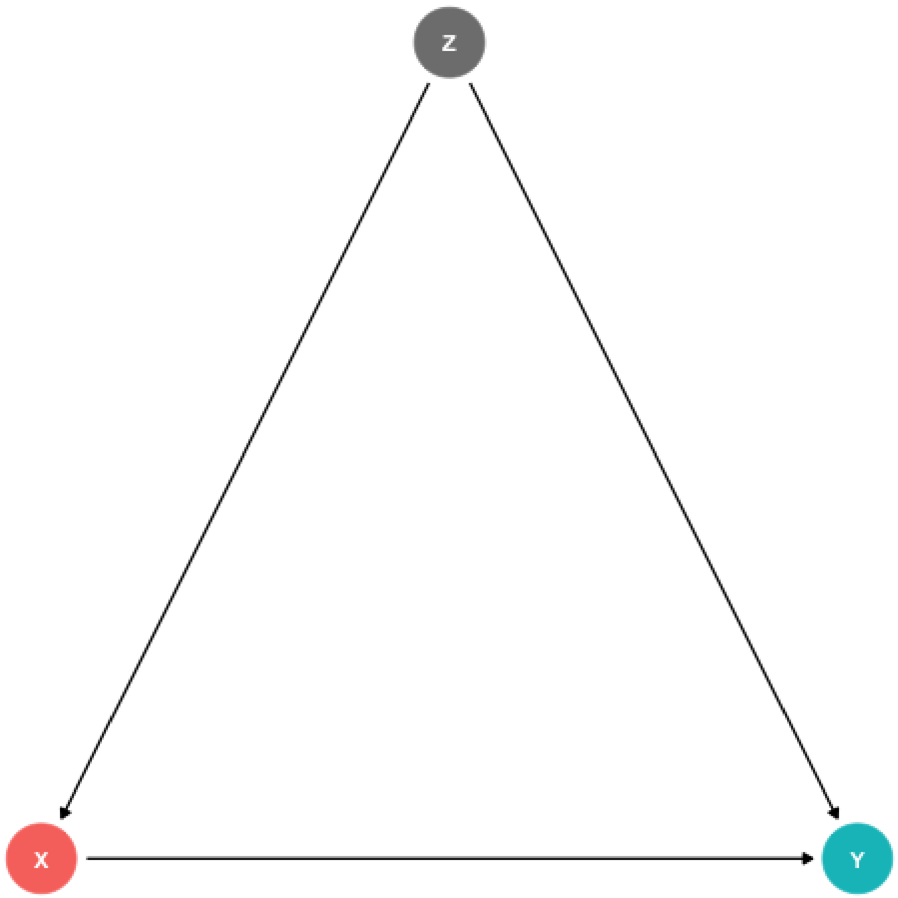
  - knowing X_i tells you something about u_i (i.e. something about Y *not* by way of X)!



Omitted Variable Bias III

- $\hat{\beta}_1$ is **biased**: $E[\hat{\beta}_1] \neq \beta_1$
- $\hat{\beta}_1$ systematically over- or under-estimates the true relationship (β_1)
- $\hat{\beta}_1$ "picks up" *both* pathways:

  1. $X \rightarrow Y$
  2. $X \leftarrow Z \rightarrow Y$



Omited Variable Bias: Class Size Example

Example: Consider our recurring class size and test score example:

$$\text{Test score}_i = \beta_0 + \beta_1 \text{STR}_i + u_i$$

- Which of the following possible variables would cause a bias if omitted?

1. Z_i: time of day of the test

2. Z_i: parking space per student

3. Z_i: percent of ESL students

Recall: Endogeneity and Bias

- (Recall): the true expected value of $\hat{\beta_1}$ is actually:[†]

$$E[\hat{\beta_1}] = \beta_1 + cor(X,u)\frac{\sigma_u}{\sigma_X}$$

1) If X is exogenous: $cor(X,u) = 0$, we're just left with β_1

2) The larger $cor(X,u)$ is, larger **bias**: $\left(E[\hat{\beta_1}] - \beta_1\right)$

3) We can **"sign"** the direction of the bias based on $cor(X,u)$

- **Positive** $cor(X,u)$ overestimates the true β_1 ($\hat{\beta_1}$ is too large)
- **Negative** $cor(X,u)$ underestimates the true β_1 ($\hat{\beta_1}$ is too small)

[†] See 2.4 class notes for proof.

Endogeneity and Bias: Correlations I

- Here is where checking correlations between variables helps:

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

# Make a correlation table
cor_table <- cor(CAcorr)
cor_table # look at it
```

```
##                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
```

- `el_pct` is strongly (negatively) correlated with `testscr` (Condition 1)

- `el_pct` is reasonably (positively) correlated with `str` (Condition 2)

Endogeneity and Bias: Correlations II

- Here is where checking correlations between variables helps:

```r
# Make a correlation plot
library(corrplot)

corrplot(cor_table, type="upper",
         method = "circle",
         order="original")
```

Look at Conditional Distributions I

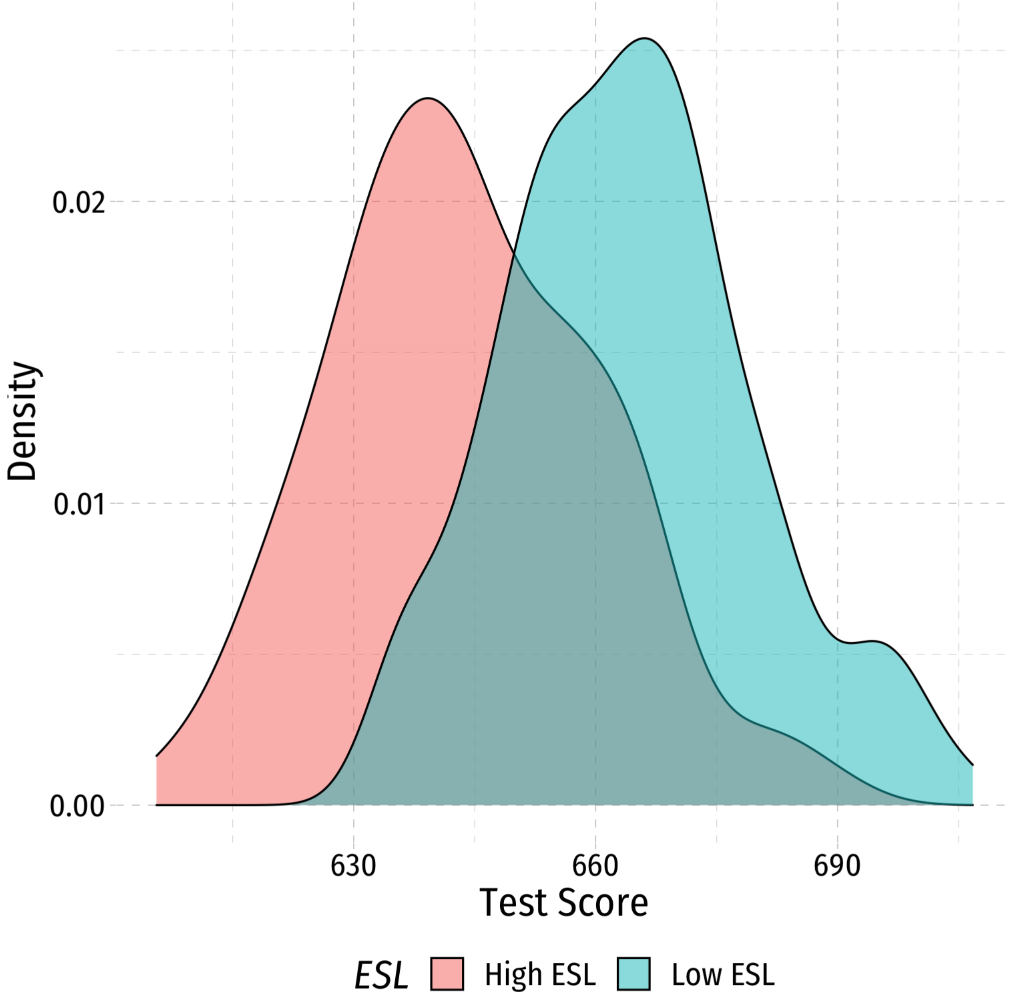
```r
# make a new variable called EL
# = high (if el_pct is above median) or = low (if below median)
CASchool <- CASchool %>% # next we create a new dummy variable called ESL
  mutate(ESL = ifelse(el_pct > median(el_pct), # test if ESL is above median
                      yes = "High ESL", # if yes, call this variable "High ESL"
                      no = "Low ESL")) # if no, call this variable "Low ESL"

# get average test score by high/low EL
CASchool %>%
  group_by(ESL) %>%
  summarize(Average_test_score = mean(testscr))
```

ESL <chr>	Average_test_score <dbl>
High ESL	643.9591
Low ESL	664.3540

2 rows

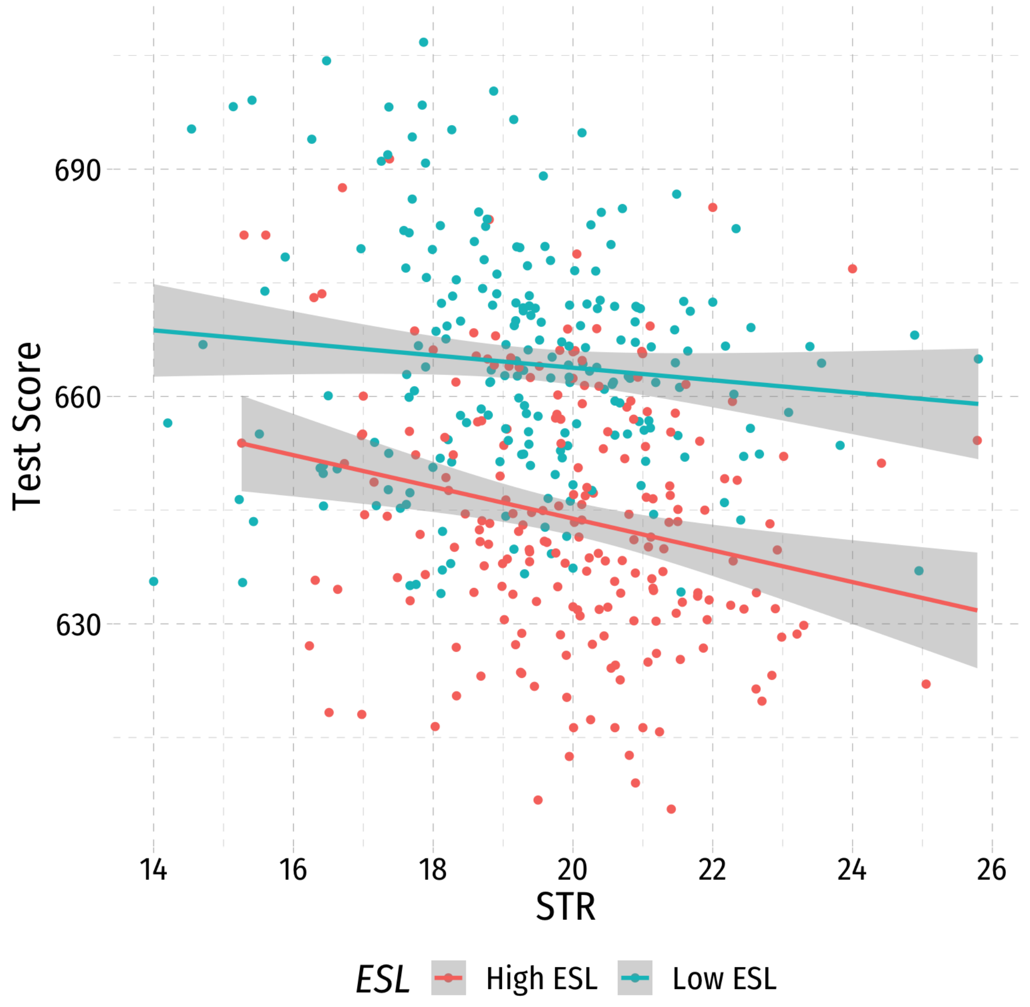
Look at Conditional Distributions II

```r
ggplot(data = CASchool)+
  aes(x = testscr,
      fill = ESL)+
  geom_density(alpha=0.5)+
  labs(x = "Test Score",
       y = "Density")+
  ggthemes::theme_pander(
    base_family = "Fira Sans Condensed",
    base_size=20
    )+
  theme(legend.position = "bottom")
```



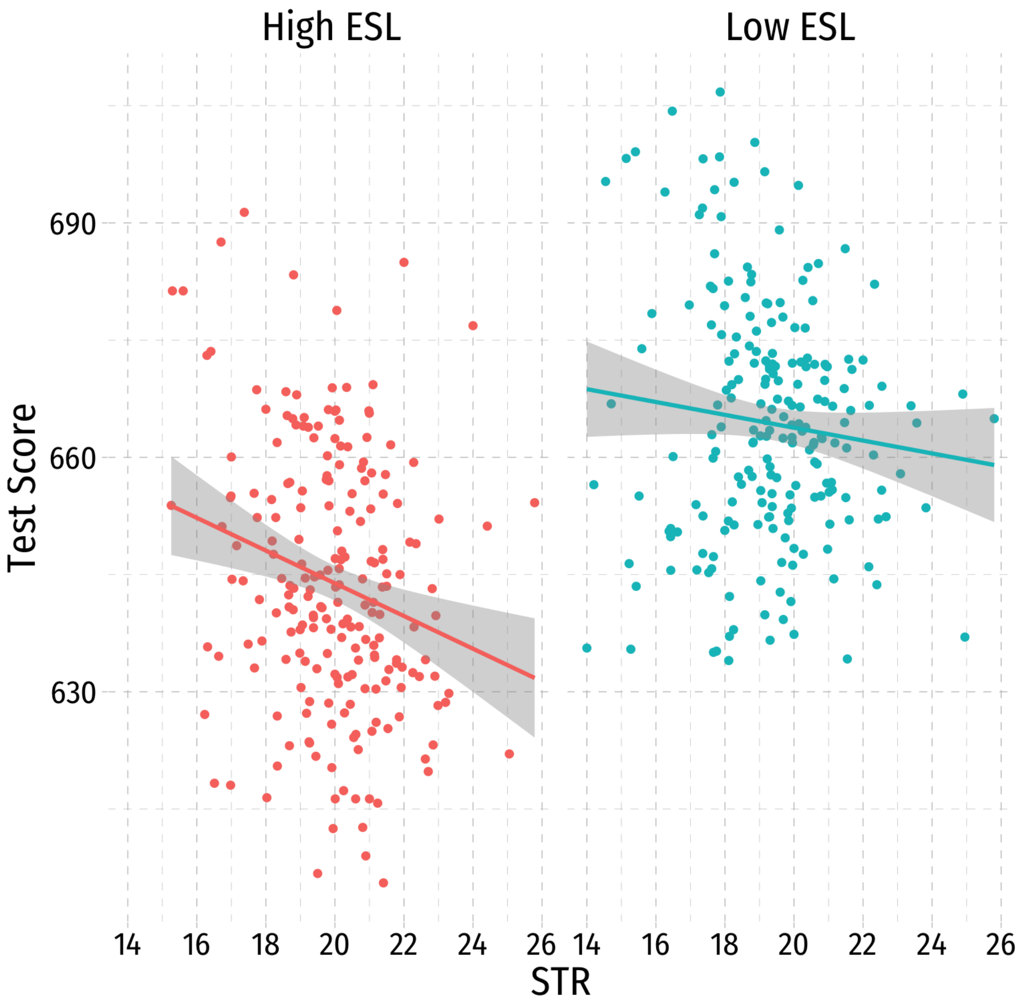
Look at Conditional Distributions III

```
esl_scatter <- ggplot(data = CASchool)+
  aes(x = str,
      y = testscr,
      color = ESL)+
  geom_point()+
  geom_smooth(method = "lm")+
  labs(x = "STR",
       y = "Test Score")+
  ggthemes::theme_pander(
    base_family = "Fira Sans Condensed",
    base_size=20
    )+
  theme(legend.position = "bottom")

esl_scatter
```



Look at Conditional Distributions III

```r
esl_scatter+
  facet_grid(~ESL)+
  guides(color = F)
```



Omitted Variable Bias in the Class Size Example

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

$$E[\hat{\beta}_1] = \beta_1 + cor(X, u)\,\frac{\sigma_u}{\sigma_X}$$

- $cor(STR, u)$ is positive (via $\%EL$)
- $cor(u, \text{Test score})$ is negative (via $\%EL$)
- β_1 is negative (between Test score and STR)
- Bias is positive
    - But since β_1 is negative, it's made to be a *larger* negative number than it truly is
    - Implies that β_1 *over*states the effect of reducing STR on improving Test Scores

Omitted Variable Bias: Messing with Causality I

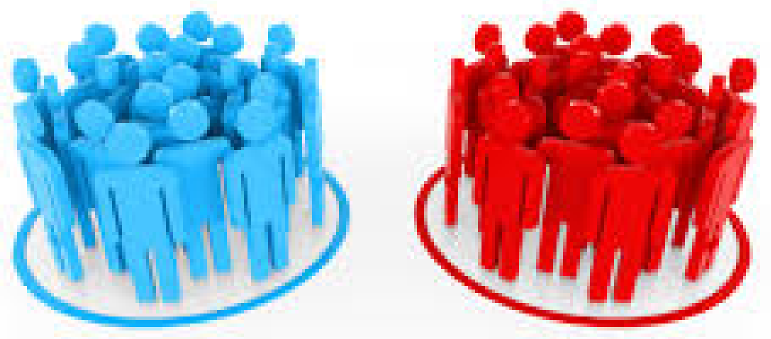
- If school districts with higher Test Scores happen to have both lower STR **AND** districts with smaller STR sizes tend to have less $\%EL$ ...

- How can we say $\hat{\beta}_1$ estimates the **marginal effect** of $\Delta STR \rightarrow \Delta \text{Test Score}$?

- (We can't.)

Omitted Variable Bias: Messing with Causality II

- Consider an ideal **random controlled trial (RCT)**
- **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(s)**: gets *no* treatment(s)
- Compare results of two groups to get **average treatment effect**



RCTs Neutralize Omitted Variable Bias I

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

- School districts would be **randomly assigned** a student-teacher ratio

- With random assignment, all factors in u (%ESL students, family size, parental income, years in the district, day of the week of the test, climate, etc) are distributed *independently* of class size

RCTs Neutralize Omitted Variable Bias II

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

- Thus, $cor(STR, u) = 0$ and $E[u|STR] = 0$, i.e. **exogeneity**

- Our $\hat{\beta}_1$ would be an **unbiased estimate** of β_1, measuring the **true causal effect** of STR $\rightarrow$ Test Score

But We Rarely, if Ever, Have RCTs

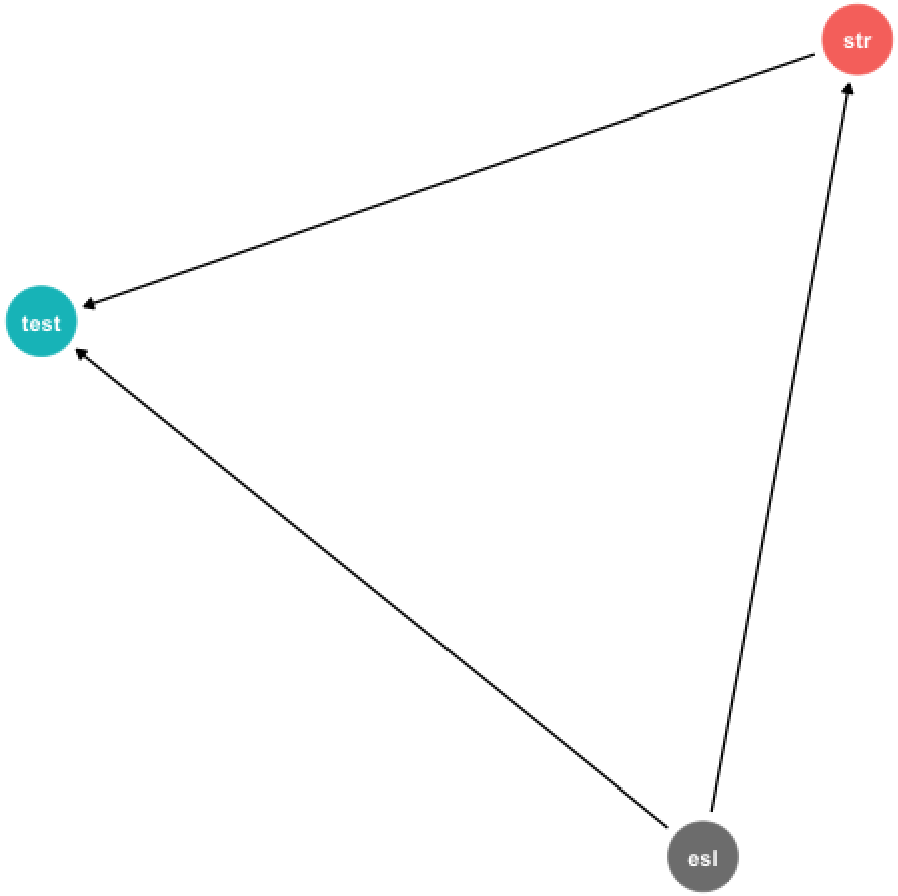
- But we **didn't** run an RCT, it's observational data!
- "Treatment" of having a large or small class size is **NOT** randomly assigned!
- $\%EL$: plausibly fits criteria of O.V. bias!
  1. $\%EL$ is a determinant of Test Score
  2. $\%EL$ is correlated with STR
- Thus, "control" group and "treatment" group differ systematically!
  - Small STR also tend to have lower $\%EL$; large STR also tend to have higher $\%EL$
  - **Selection bias**: $cor(STR, \%EL) \neq 0$, $E[u_i|STR_i] \neq 0$

Treatment Group

Control Group

Another Way to Control for Variables

- Pathways connecting str and test score:
  1. str → test score
  2. str ← ESL → testscore



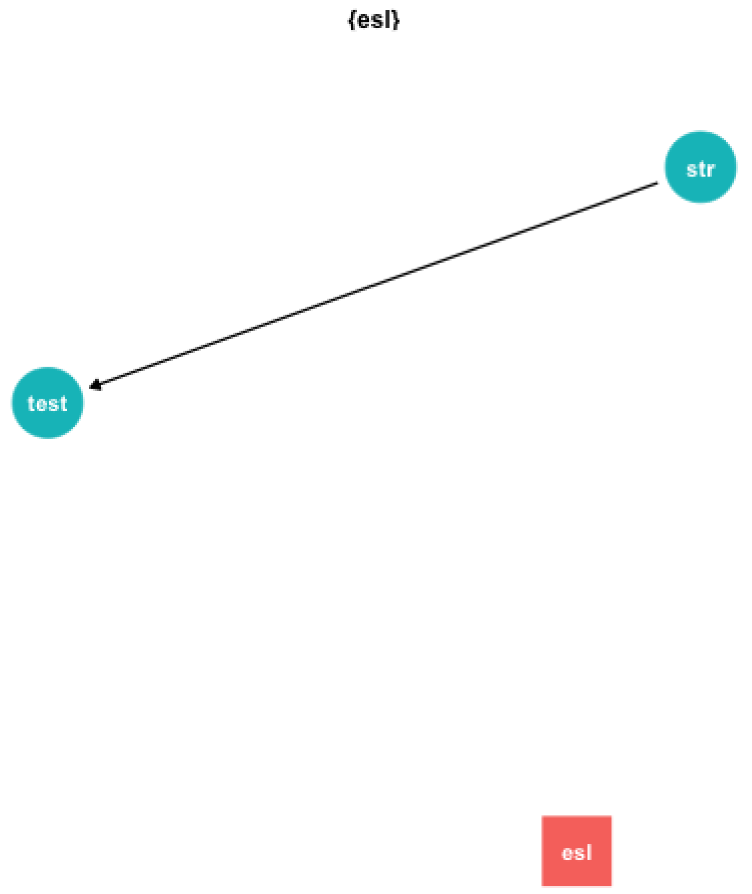
Another Way to Control for Variables

- Pathways connecting str and test score:

    1. str $\rightarrow$ test score
    2. str $\leftarrow$ ESL $\rightarrow$ testscore

- DAG rules tell us we need to **control for ESL** in order to identify the causal effect of str $\rightarrow$ test score

- So now, **how *do* we control for a variable**?



Controlling for Variables

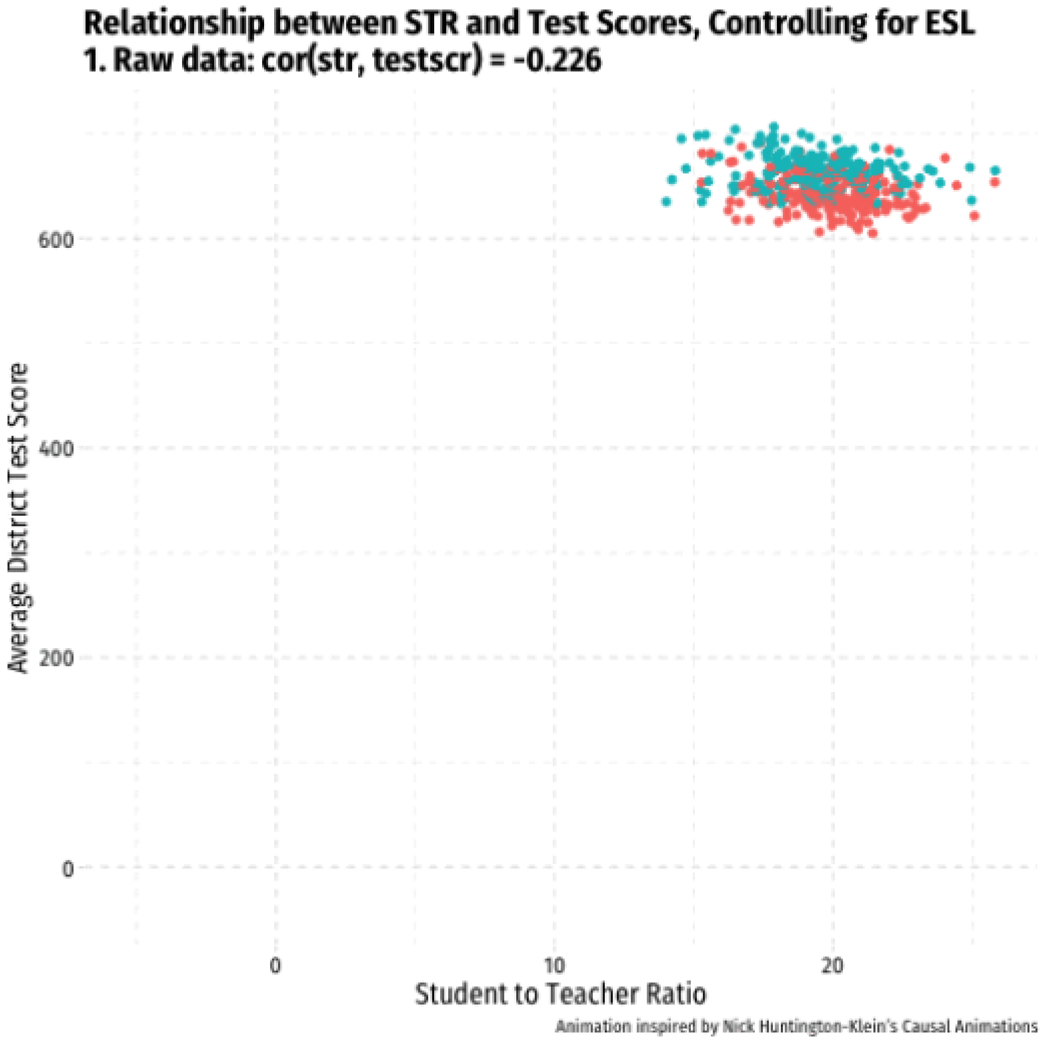
- Look at effect of STR on Test Score by comparing districts with the **same** %EL
    - Eliminates differences in %EL between high and low STR classes
    - "As if" we had a control group! Hold %EL constant
- The simple fix is just to **not omit %EL**!
    - Make it *another* independent variable on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

Treatment Group

Control Group

Controlling for Variables

- 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

Multivariate Econometric Models Overview

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

- Y is the **dependent variable** of interest
  - AKA "response variable," "regressand," "Left-hand side (LHS) variable"

- X_1 and X_2 are **independent variables**
  - AKA "explanatory variables", "regressors," "Right-hand side (RHS) variables", "covariates"

- Our data consists of a spreadsheet of observed values of (X_{1i}, X_{2i}, Y_i)

- To model, we **"regress Y on X_1 and X_2"**

- $\beta_0, \beta_1, \cdots, \beta_k$ are **parameters** that describe the population relationships between the variables
  - We estimate $k+1$ parameters ("betas")[†]

[†] Note Bailey defines k to include both the number of variables plus the constant.

Marginal Effects I

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

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad \text{Before the change}$$

Marginal Effects I

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

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 && \text{Before the change} \\ Y + \Delta Y &= \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 && \text{After the change} \end{aligned}$$

Marginal Effects I

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

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 && \text{Before the change} \\ Y + \Delta Y &= \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 && \text{After the change} \\ \Delta Y &= \beta_1 \Delta X_1 && \text{The difference} \end{aligned}$$

Marginal Effects I

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

- Consider changing X_1 by ΔX_1 while holding X_2 constant:

$$\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 && \text{Before the change} \\
Y + \Delta Y &= \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 && \text{After the change} \\
\Delta Y &= \beta_1 \Delta X_1 && \text{The difference} \\
\frac{\Delta Y}{\Delta X_1} &= \beta_1 && \text{Solving for } \beta_1
\end{aligned}$$

Marginal Effects II

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

Similarly, for β_2:

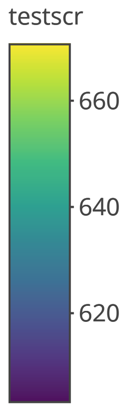
$$\beta_2 = \frac{\Delta Y}{\Delta X_2} \text{ holding } X_1 \text{ constant}$$

And for the constant, β_0:

$$\beta_0 = \text{predicted value of Y when } X_1 = 0, \; X_2 = 0$$

You Can Keep Your Intuitions...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_0: "intercept"
  - β_1: "slope"
- With 3+ variables, OLS regression is no longer a "line" for us to estimate...



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 useful next class!

The Population Regression Model: Example I

Example:

$$\text{Beer Consumption}_i = \beta_0 + \beta_1 Price_i + \beta_2 Income_i + \beta_3 \text{Nachos Price}_i + \beta_4 \text{Wine Price}$$

- 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 sign should it have? What should inferior or normal (necessities & luxury) goods look like?
- What measures the **cross-price effect(s)**? What sign should substitutes and complements have?

The Population Regression Model: Example I

Example:

$$\widehat{\text{Beer Consumption}}_i = 20 - 1.5Price_i + 1.25Income_i - 0.75\text{Nachos Price}_i + 1.3\text{Wine}$$

- Interpret each $\hat{\beta}$

Multivariate OLS in R

```r
# run regression of testscr on str and el_pct
school_reg_2 <- lm(testscr ~ str + el_pct,
                   data = CASchool)
```

- Format for regression is

```r
lm(y ~ x1 + x2, data = df)
```

- `y` is dependent variable (listed first!)
- `~` means “is modeled by” or “is explained by”
- `x1` and `x2` are the independent variable
- `df` is the dataframe where the data is stored

Multivariate OLS in R II

```r
# look at reg object
school_reg_2
```

```
##
## Call:
## lm(formula = testscr ~ str + el_pct, data = CASchool)
##
## Coefficients:
## (Intercept)          str       el_pct
##    686.0322      -1.1013      -0.6498
```

- Stored as an `lm` object called `school_reg_2`, a `list` object

Multivariate OLS in R III

```r
summary(school_reg_2) # get full summary
```

```
##
## 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
```

Multivariate OLS in R IV: broom

```r
# load packages
library(broom)

# tidy regression output
tidy(school_reg_2)
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	686.0322487	7.41131248	92.565554	3.871501e-280
str	-1.1012959	0.38027832	-2.896026	3.978056e-03
el_pct	-0.6497768	0.03934255	-16.515879	1.657506e-47

3 rows

Multivariate Regression Output Table

```r
library(huxtable)
huxreg("Model 1" = school_reg,
       "Model 2" = school_reg_2,
       coefs = c("Intercept" = "(Intercept)",
                 "Class Size" = "str",
                 "%ESL Students" = "el_pct"),
       statistics = c("N" = "nobs",
                      "R-Squared" = "r.squared",
                      "SER" = "sigma"),
       number_format = 2)
```

	Model 1	Model 2
Intercept	698.93 ***	686.03 ***
	(9.47)	(7.41)
Class Size	-2.28 ***	-1.10 **
	(0.48)	(0.38)
%ESL Students		-0.65 ***
		(0.04)
N	420	420
R-Squared	0.05	0.43
SER	18.58	14.46

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.