3.3 — Omitted Variable Bias

ECON 480 • Econometrics • Fall 2022

Dr. Ryan Safner
Associate Professor of Economics
safner@hood.edu
ryansafner/metricsF22
metricsF22.classes.ryansafner.com

Contents

Omitted Variables and Bias

The Error Term

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

- 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

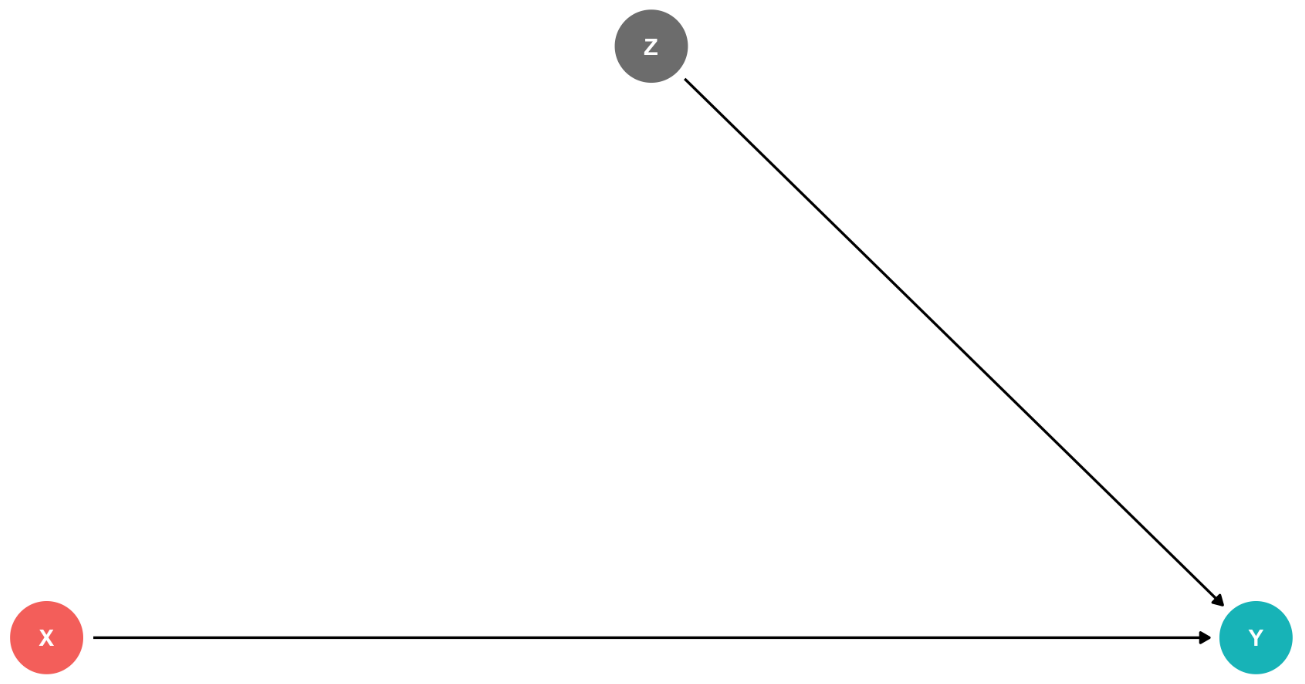
- **Omitted variable bias (OVB)** for some omitted variable $\mathbf{Z}$ exists if two conditions are met:

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

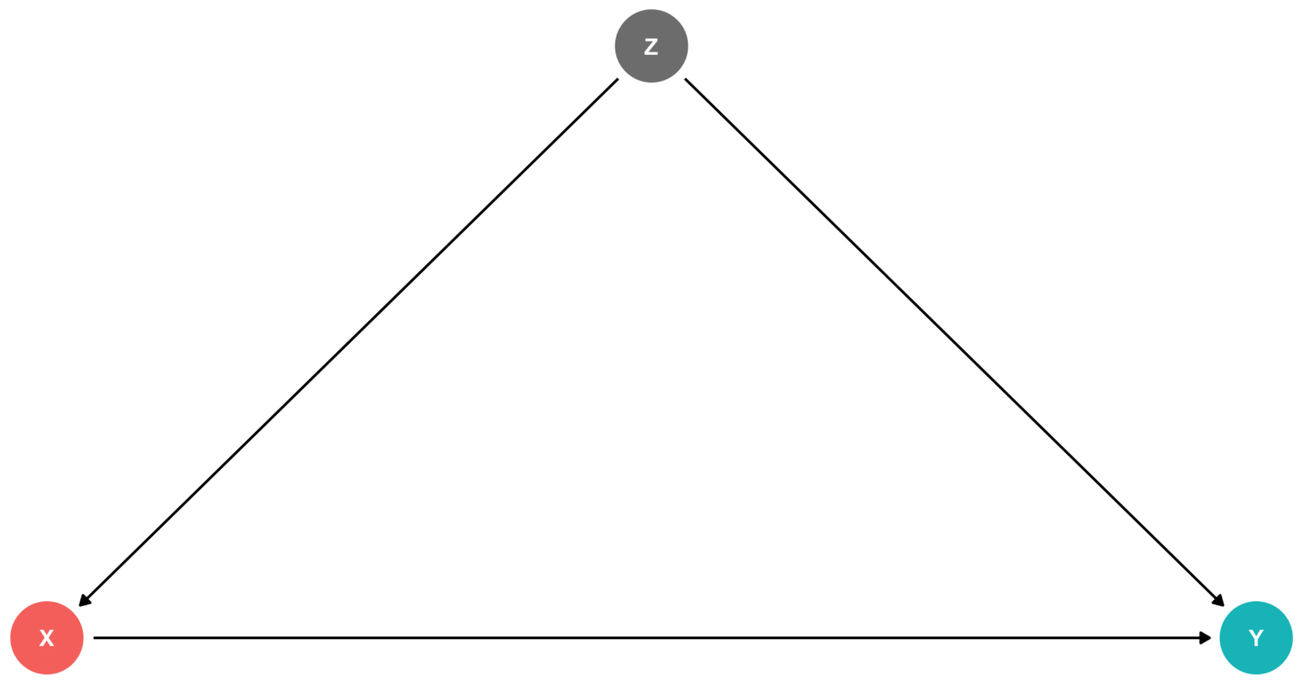
- **Omitted variable bias (OVB)** for some omitted variable $\mathbf{Z}$ exists if two conditions are met:

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**

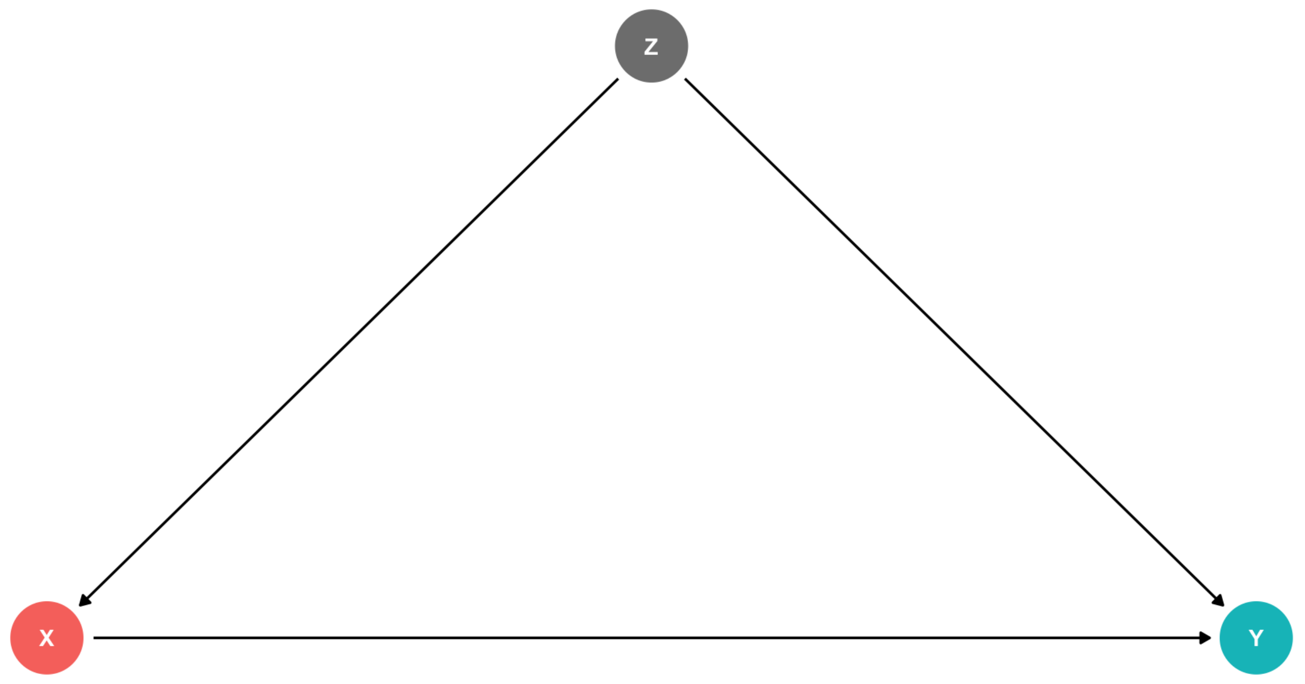


Omitted Variable Bias II

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

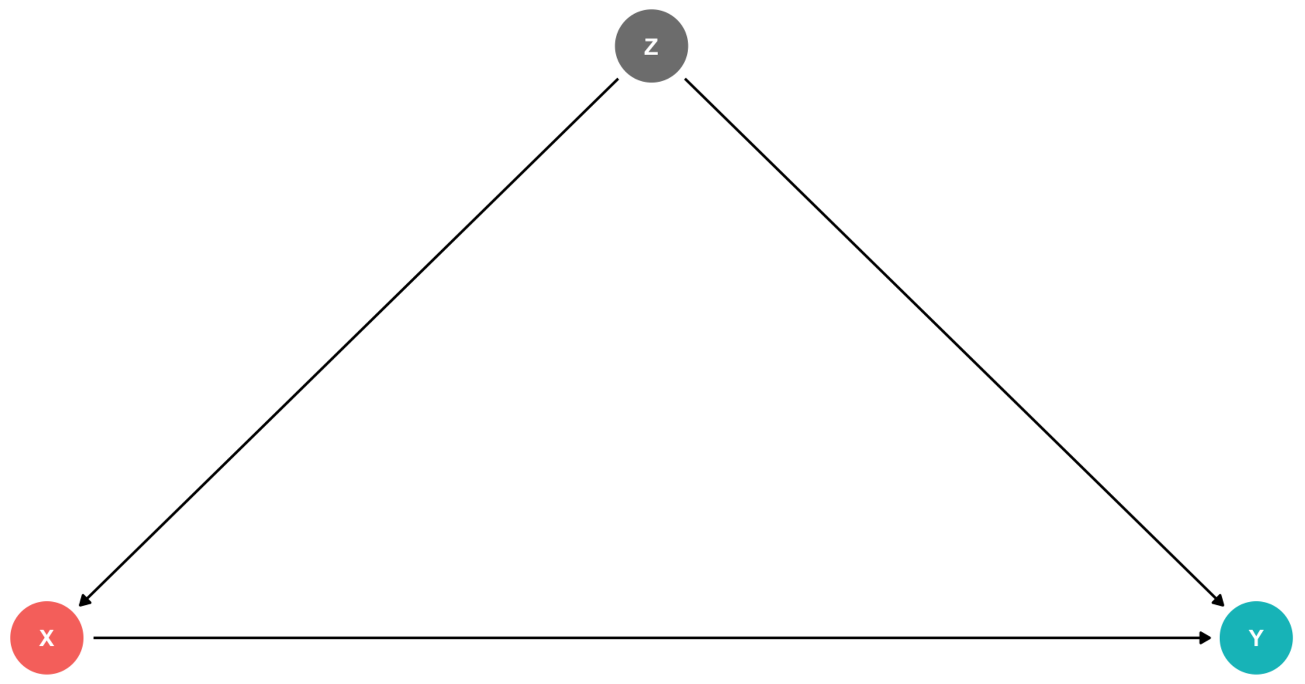
$$\mathbb{E}(u_i|X_i) \neq 0 \implies$$

- 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**: $\mathbb{E}[\hat{\beta}_1] \neq \beta_1$
- $\hat{\beta}_1$ 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

- The true expected value of $\hat{\beta_1}$ is actually: [See class 2.4 for proof.]

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

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 high)
- **Negative** $cor(X,u)$ underestimates the true β_1 ($\hat{\beta_1}$ is too low)

Endogeneity and Bias: Correlations I

- Here is where checking correlations between variables can help us:

```
ca_school %>%
  # Select only the three variables we want (there are many)
  select(str, testscr, el_pct) %>%
  # make a correlation table (all variables must be numeric)
  cor()
```

```
               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)

Look at Conditional Distributions I

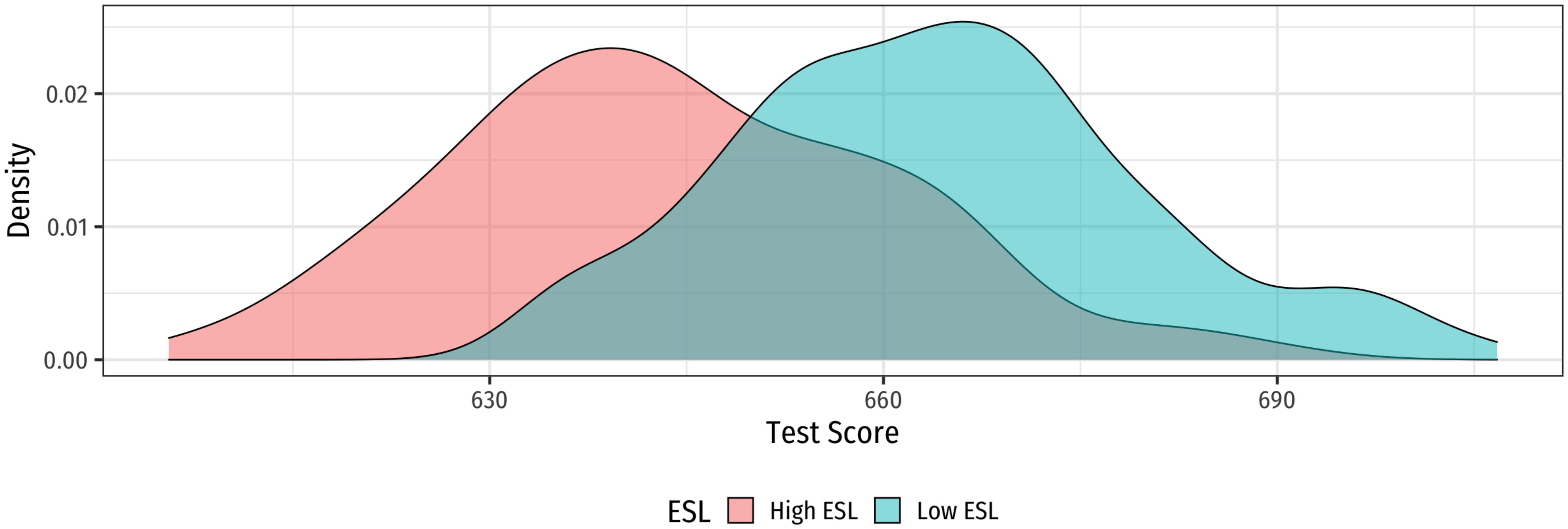
```r
# make a new variable called EL
# = high (if el_pct is above median) or = low (if below median)
ca_school <- ca_school %>% # next we create a new dummy variable called ESL
  mutate(ESL = ifelse(el_pct > median(el_pct), # test if ESL is above median
                      yes = "High ESL", # if yes,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

# get average test score by high/low EL
ca_school %>%
  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

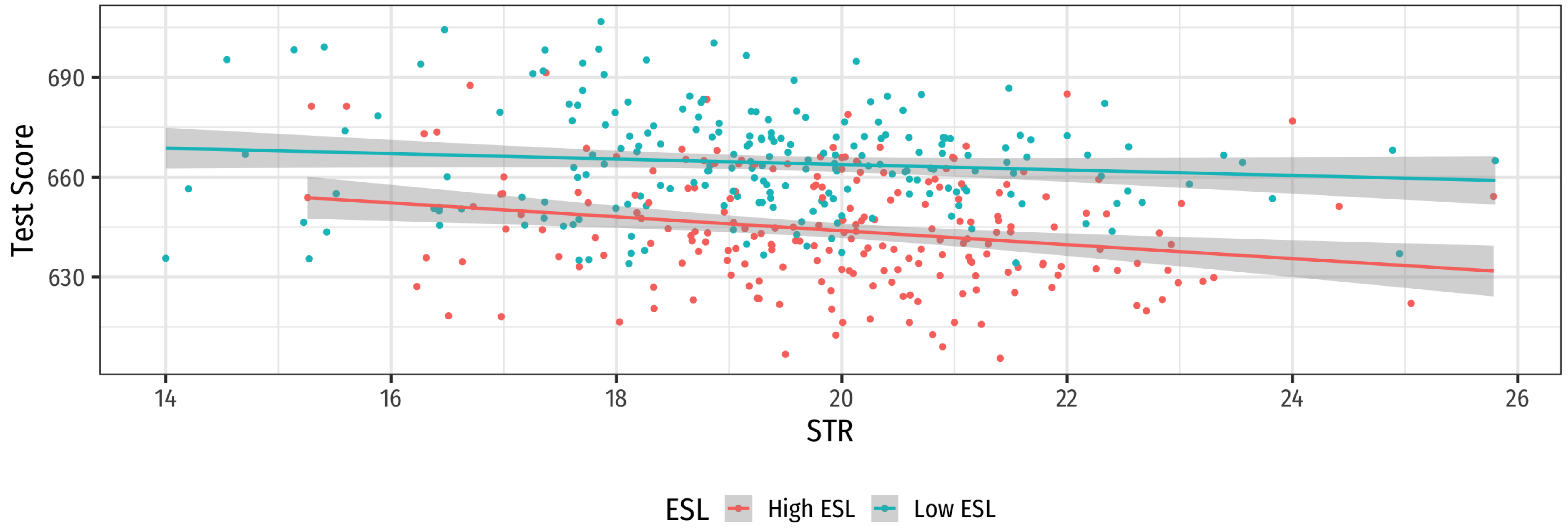
Look at Conditional Distributions II

Plot | Code



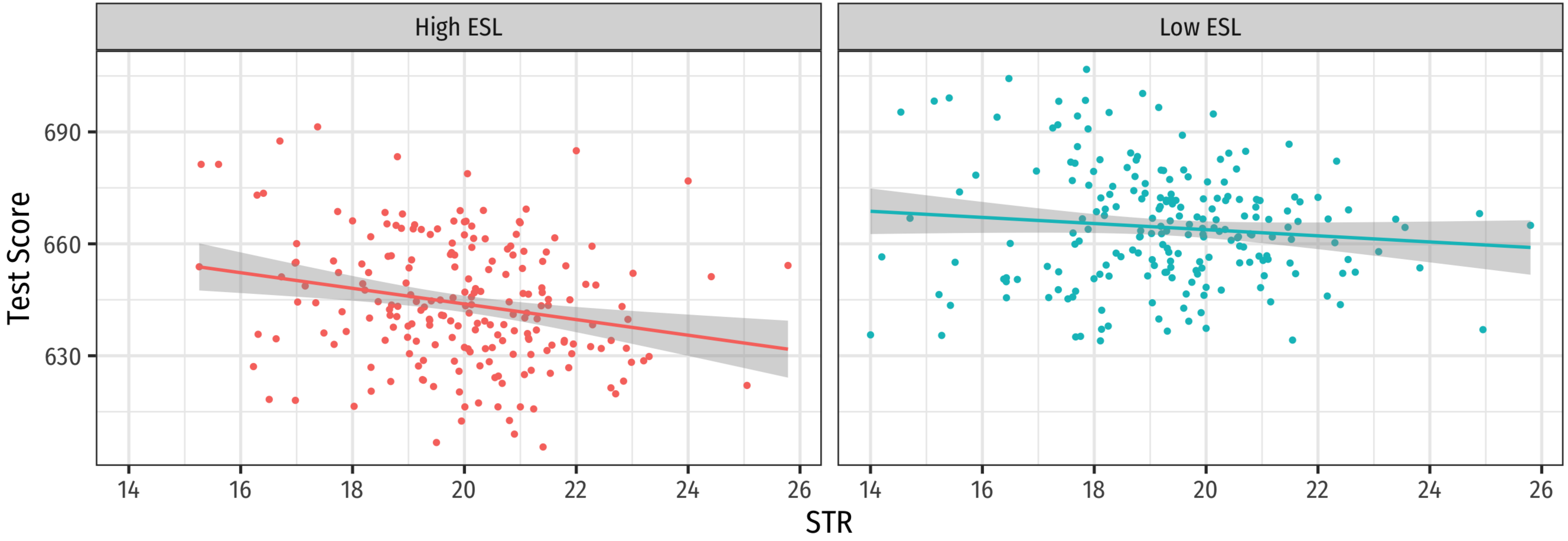
Look at Conditional Distributions III

Plot | Code



Look at Conditional Distributions IV

Plot | Code



Omitted Variable Bias in the Class Size Example

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

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

- $cor(STR, u)$ is positive (via $\%EL$)
- $cor(u, \text{Test score})$ is negative (via $\%EL$)
- β_1 is negative (between test score and str)
- Bias from $\%EL$ is positive
  - Since β_1 is negative, it's made to be a *larger* negative number than it truly is
  - Implies that our $\hat{\beta}_1$ ***overstates*** the effect of reducing STR on improving Test Scores

Omitted Variable Bias: Messing with Causality I

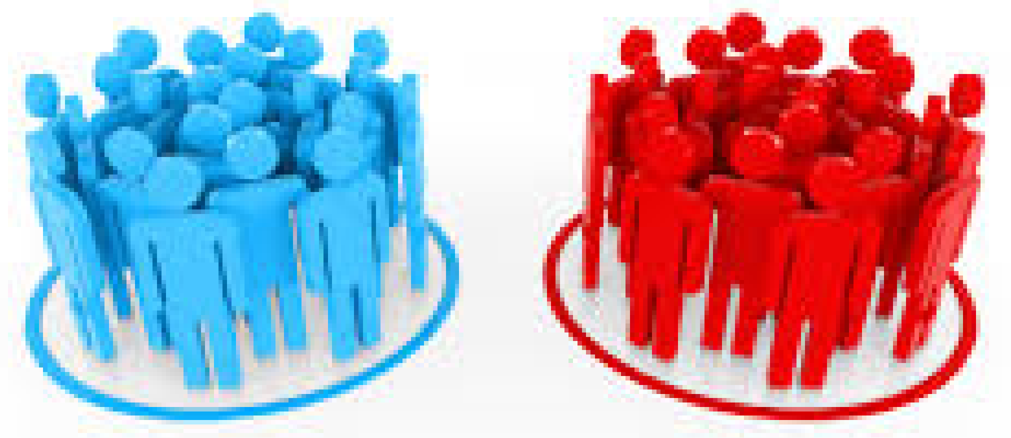
- If school districts with higher Test Scores happen to have both lower STR **AND** districts with smaller STR sizes tend to have less $\%EL$ ...

- 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 → Test Score

But We Rarely, if Ever, Can Do RCTs

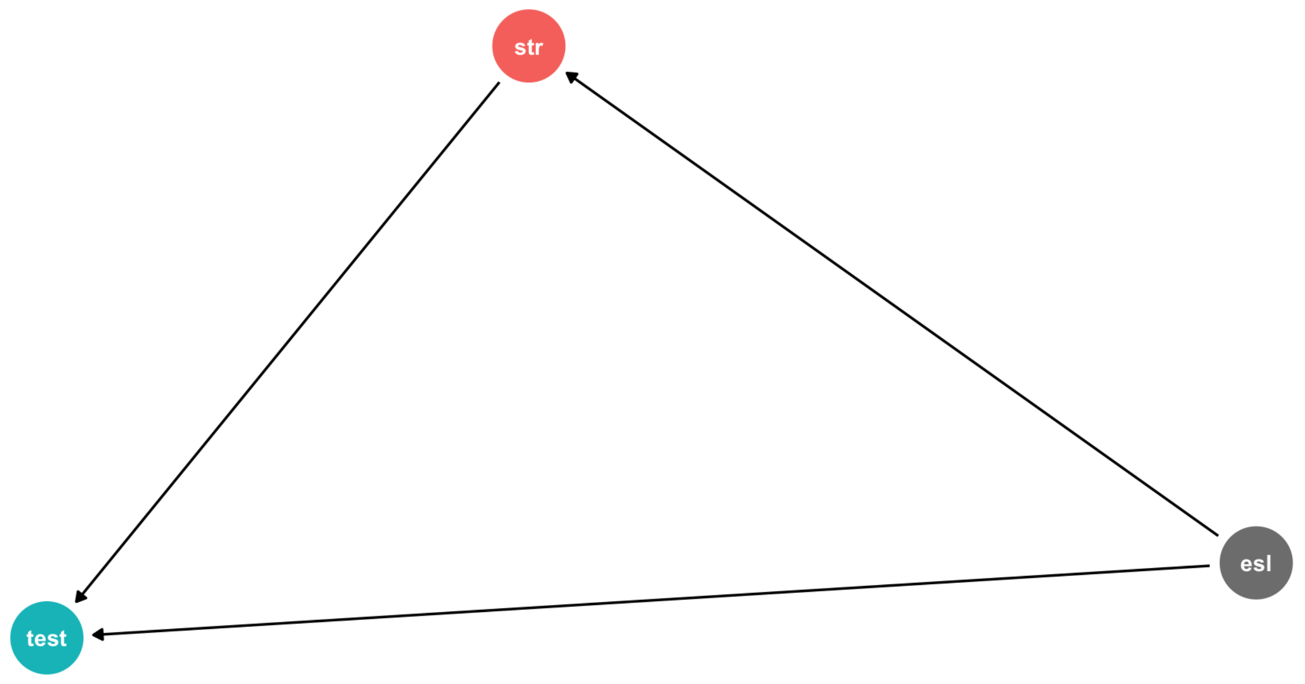
- But we **didn't** run an RCT, we have observational data!
- "Treatment" of having a large or small class size is **NOT** randomly assigned!
- $\%EL$: plausibly fits criteria of O.V. bias!
  1. $\%EL$ is a determinant of Test Score
  2. $\%EL$ is correlated with STR
- Thus, "control" group and "treatment" 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

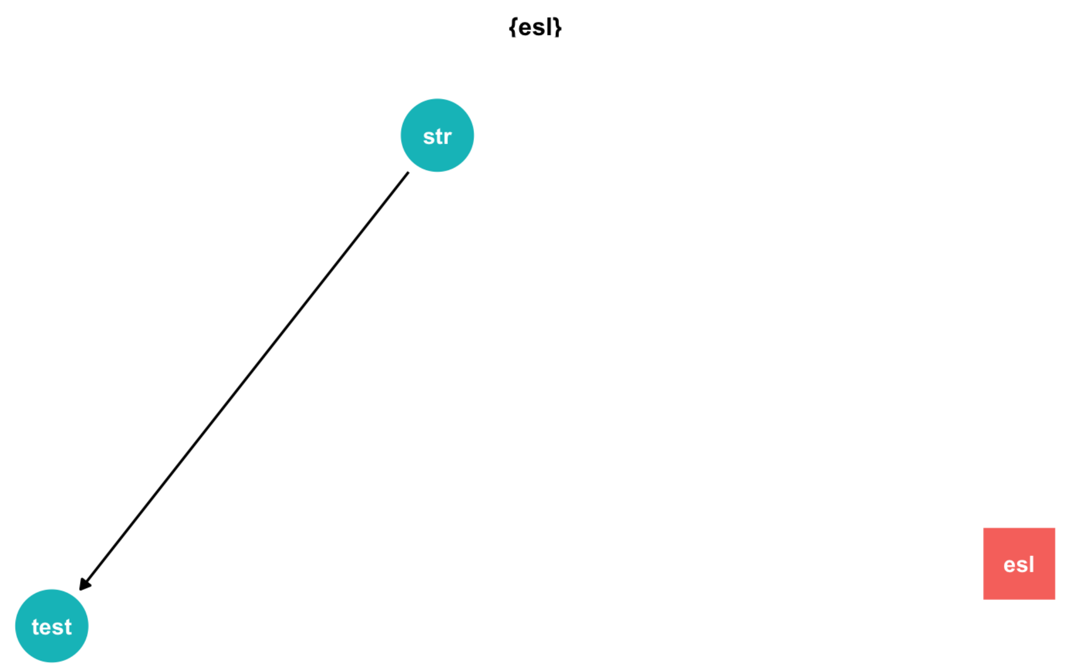
Another Way to Control for Variables I

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



Another Way to Control for Variables II

- Pathways connecting str and test score:
  1. str → test score
  2. str ← ESL → testscore
- DAG rules tell us we need to **control for ESL** in order to identify the causal effect of str → test score
- So now, **how *do* we control for a variable**?



Controlling for Variables

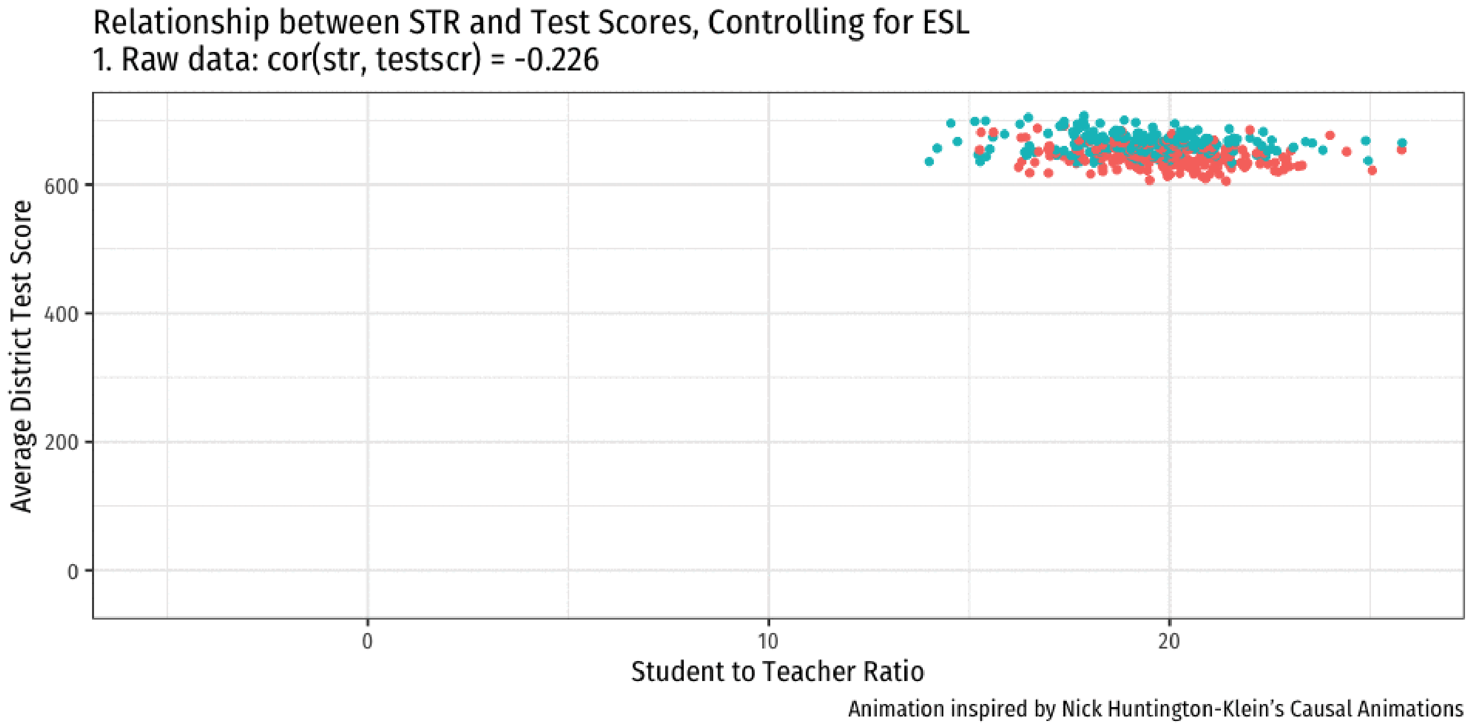
- Look at effect of STR on Test Score by comparing districts with the **same** %EL
  - Eliminates differences in %EL between high and low STR classes
  - "As if" we had a control group! Hold %EL constant
- The simple fix is just to **not omit %EL**!
  - Make it *another* independent variable on the righthand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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

- Y is the **dependent variable** of interest
  - AKA "response variable," "regressand," "Left-hand side (LHS) variable"
- $X_1, X_2, \cdots, X_k$ are **independent variables**
  - AKA "explanatory variables", "regressors," "Right-hand side (RHS) variables", "covariates"
- Our data consists of a spreadsheet of observed values of $(Y_i, X_{1i}, X_{2i}, \cdots, X_{ki})$

Multivariate Econometric Models: Overview II

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

- To model, we **"regress Y on $X_1, X_2, \cdots, X_k$"**
- $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ are **parameters** that describe the population relationships between the variables
  - unknown! to be estimated
  - we estimate $k+1$ parameters ("betas") on k variables[1]
- u is a random **error term**
  - **'U'nobservable**, we can't measure it, and must model with assumptions about it

1. Note Bailey defines k to include both the number of variables plus the constant

Marginal Effects I

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

- 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

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

- 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

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

- 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

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

- 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:

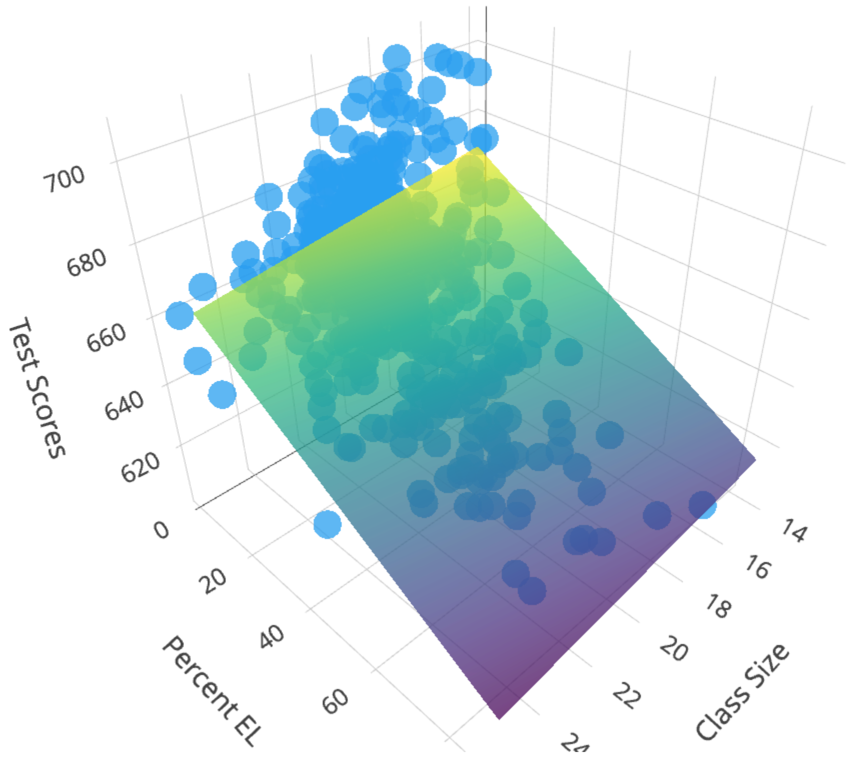
$$\beta_0 = \text{predicted value of Y when } X_1 = 0, \; X_2 = 0$$

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...

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

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

- 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 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 II

Example

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

- Interpret each $\hat{\beta}$

The Multivariate Regression Model

Multivariate Regression in R

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

- 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 variables
- `df` is the dataframe where the data is stored

Multivariate Regression in R

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

```
Call:
lm(formula = testscr ~ str + el_pct, data = ca_school)

Coefficients:
(Intercept)          str       el_pct  
   686.0322      -1.1013      -0.6498  
```

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

Multivariate Regression in R

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

```
Call:
lm(formula = testscr ~ str + el_pct, data = ca_school)

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 ***
---
```

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

Multivariate Regression with Broom

- The `tidy()` function creates a *tidy* `tibble` of regression output

```
# load packages
library(broom)

# tidy regression output
school_reg_2 %>%
  tidy()
```

term <chr>	estimate <dbl>
(Intercept)	686.0322487
str	-1.1012959
el_pct	-0.6497768

3 rows | 1-2 of 5 columns

Multivariate Regression Output Table

```
# load package
library(modelsummary)

modelsummary(models = list("Test Score" = school_r
                           "Test Score" = school_r
             fmt = 2, # round to 2 decimals
             output = "html",
             coef_rename = c("(Intercept)" = "Cons
                             "str" = "STR"),
             gof_map = list(
               list("raw" = "nobs", "clean" = "n",
               list("raw" = "r.squared", "clean" =
               list("raw" = "rmse", "clean" = "SER
             ),
             escape = FALSE,
             stars = c('*' = .1, '**' = .05, '***'
)
```

	Test Score	Test Score
Constant	698.93***	686.03***
	(9.47)	(7.41)
STR	-2.28***	-1.10***
	(0.48)	(0.38)
el_pct		-0.65***
		(0.04)
n	420	420
R^2	0.05	0.43
SER	18.54	14.41

* p < 0.1, ** p < 0.05, *** p < 0.01