Do this at beginning of class

- Download lecture 2 Stata "do-file" link
  - Save to a folder you can easily find
- Download lecture 2 datasets
  - Save to folder you can easily find; do not change file names
  - California school dataset link
  - Tennessee STAR Experiment dataset link
- Try doing this in Stata:
  1. Open Stata
  2. Open "do-file" editor
  3. Open lecture 2 do-file
  4. "change directories" in do-file so that file-path in "cd" command points to where you saved the data
     - You will change file-paths in two places in do-file: once for CA school data; once for Tennesee STAR data

EDUC 263: Introduction to Econometrics, Lecture 2

Experimental and observational designs

Ozan Jaquette
ozanj@ucla.edu

University of California, Los Angeles
Higher Education & Organizational Change

What we will do today

Introduction to Stata

Understanding syntax of Stata commands

Let's open a dataset in Stata. The "auto" dataset is a sample dataset saved on your computer when you install Stata.

Type the following text in the Stata commandline and then press "Enter" on your keyboard:

```
sysuse auto, clear
```

Type the following text in the Stata commandline and then press "Enter" on your keyboard:

```
describe
```

Understanding syntax of Stata commands

Stata commands often follow the following format:

```
commandname [varlist] [if] [in] [, options]
```

- only need to type underlined part of command name
- anything in brackets doesn't need to be included for command to run
- text after the comma are options

For example, the summarize command:

```
summarize [varlist] [if] [in] [weight] [,
options]
```

Try typing these commands in Stata command line:

```
summarize
sum price
sum price, detail
```

Executing Stata commands

Three ways to execute Stata commands:

- Point-and-click (Ugh!)
- Stata command line
  - Will use this to run individual commands
- Stata do-file
  - Best way to run Stata commands; required for all homework assignments
  - Will review working with do-files later in this lecture

Stata help files

In Stata, type following syntax:

```
help commandname
help generate
help gen
help reg
```

Reading Stata help files

Help files may feel too technical at first, but you'll become more comfortable with practice

Help files follow a standard outline (e.g., `help reg`):

- Syntax (command syntax and list of options)
- Menu (how to execute command using point-and-click)
- Description (text overview of what command does)
- Options (detailed description of command options)
- Examples (examples of how to use command)
- Video example (some commands have this)
- Stored results (stored results created by command)

Top right corner of help file:

- "Dialog" (run command using point and click)
- "Also see" (link to PDF documentation; related Stata commands)

Working with Stata "do-files"

A Stata do-file is just a text-file that contains Stata commands

Opening a do-file:

- Open do-files from **within** Stata rather than from Windows Explorer (or Mac equivalent)
- In Stata, click on the "New do-file editor" button; this opens a new do-file
- In the do-file, click on "file" then "open" and the find the do-file you want to open

Executing Stata commands within a do-file

Within a do-file you can run commands several different ways (try doing this within the do-file):

1. One command at a time by highlighting only that command and clicking Execute(do) button in do-file
2. Several commands at a time by highlighting several commands and clicking Execute(do) button in the do-file
3. can run the entire do-file by not highlighting any commands and clicking Execute(do) button in do-file

Comments: text that Stata will ignore when executing do-file

- Different ways to start a comment:
  - `* COMMENT`
  - `// COMMENT`
  - `/* COMMENT */`

Changing directories within a Stata do-file

"Working directory" is the directory (i.e., filepath) where Stata looks to find files (e.g., datasets)

The `cd` command changes the filepath of the working directory. Syntax:

```
cd "filepath"
```

Note: PC uses backslash "\" to separate folders in filepath
Note: Mac uses forward-slash "/" to separate folders in filepath

Essential that you become comfortable changing directories within do-files

Let's practice in the do-file

Goal of statistics

Goal of statistics

Goal of statistics (including econometrics):

- use sample data to make statement about population

Population parameter: some measure of the population

- e.g., mean income across all U.S. households
- Usually don't know this; need data on entire population

Estimator

- A formula or procedure used to calculate an educated guess of the value of the population parameter
- e.g., calculating difference between treated and untreated mean in experiments; ordinary least squares (OLS) estimator

Point estimate (or estimate):

- Numeric value calculated when you apply an estimator to a specific sample of data.

Notation: parameters vs. estimates

Population parameters

- Described using lowercase Greek letters (e.g., μ, σ, β)
  - e.g., μ ("mu") refers to population mean; σ ("sigma") refers to population standard deviation
- Subscripts usually denote variables (e.g., μ_Y, σ_X, β_X)
  - σ_X refers to population standard deviation of variable X

Estimates of population parameters

- Described using Greek letters with "hat"
  - e.g., $\hat{\mu}_Y$ is the estimate of μ_Y
  - $\hat{\sigma}_X$ is estimate of σ_X based on sample data
- Also described using Arabic letters
  - e.g., $\bar{Y}$ is estimate of μ_Y based on sample data
  - s_X is estimate of σ_X based on sample data

Desirable properties of estimators: Efficiency and unbiasedness

- Desirable properties of your point estimates (e.g., $\hat{\beta}$ or $\overline{Y}$)
  - Desire point estimates to be "unbiased"
  - Desire point estimates to be "efficient"
- Efficiency
  - Definition:
    - Efficiency refers to how close your point estimate is to the population parameter
  - Standard error:
    - On average, how far away is a point estimate from one random sample from the value of the population parameter
  - Therefore, an efficient point estimate is one with low standard error

Bias

- An "unbiased" point estimate
  - A point estimate is "unbiased" if value of point estimate gets closer to value of true population parameter as sample size increases
- Bias
  - Bias occurs when point estimate does not get closer to population parameter as sample size increases
  - A biased estimate consistently overestimates or underestimates population parameter in repeated random samples
  - There are many different types of bias
- Sampling bias:
  - The estimate of population parameter is biased because you fail to take a random sample
  - Example: goal is to estimate high school graduation rate
    - You take random sample of 10th grades and see if they graduate within three years
- Omitted variable bias:
  - Bias in estimate of β due to omitting necessary "control" variables from your regression model

Unbiased estimates of causal relationships more difficult than descriptive relationships

- Unbiased estimate of a population mean
  - e.g., what is mean hours worked per week in U.S.
  - Primary threat is sampling bias; is your sample representative of the population
- Unbiased estimate of a correlational relationship
  - e.g., how much longer (shorter) do married men live compared to unmarried men?
  - Primary threat is sampling bias
- Unbiased estimate of a causal relationship
  - Effect of marriage (X) on life expectency (Y) for men?
  - Threats to unbiased estimate
    - Sample unrepresentative of population
    - Mistaking a correlational relationship for a causal one; even if you had data on **entire population**, your estimate could be biased

Experiments

Potential outcomes

Example: what is effect of having an internship in college (X) on earning after college (Y)?

- let $i = 1 \ldots N$ be units (e.g., people) in sample
- d_i indicates receipt of treatment (e.g., internship)
  - $d_i = 1$ for treated units; $d_i = 0$ for untreated units
- "Potential outcomes", $Y_i(1)$ and $Y_i(0)$
  - $Y_i(1)$: outcome if i if i receives treatment $d_i = 1$
  - $Y_i(0)$: outcome if i if i doesn't receives treatment $d_i = 0$
- "Observed outcome," Y_i
  - For each person, we observe $Y_i(1)$ or $Y_i(0)$ but never both
  - $Y_i = Y_i(1)d_i + Y_i(0)(1 - d_i)$
  - if $d_i = 1$ (treated):
    - $Y_i = Y_i(1) * 1 + Y_i(0)(1 - 1) = Y_i(1)$
  - if $d_i = 0$ (untreated):
    - $Y_i = Y_i(1) * 0 + Y_i(0)(1 - 0) = Y_i(0)$

Table of potential outcomes

Example: what is effect of having an internship in college (X) on annual income after college ($000s) (Y)?

i	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Treatment effect**
1	65	60	5
2	30	35	-5
3	55	60	-5
4	25	30	-5
5	50	50	0
6	80	70	10
7	45	45	0
Average	**50**	**50**	**0**

How to think about relationship between potential outcomes and observed outcome

i	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Treatment effect**
1	65	60	5
2	30	35	-5
3	55	60	-5
4	25	30	-5
5	50	50	0
6	80	70	10
7	45	45	0
Average	**50**	**50**	**0**

How to think about potential vs. observed outcomes:

- for each person i, the treated potential outcome $Y_i(1)$ and the untreated potential outcome $Y_i(0)$ already exist
- Treatment d_i just determines which of the two potential outcomes we get to observe
- for each paerson, the only difference between $Y_i(1)$ and $Y_i(0)$ is the treatment
- Value of potential outcomes is driven by the treatment and by characteristics that affect Y_i (e.g., parental income)

Average treatment effect (ATE)

i	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Treatment effect**
1	65	60	5
2	30	35	-5
3	55	60	-5
4	25	30	-5
5	50	50	0
6	80	70	10
7	45	45	0
Average	**50**	**50**	**0**

$$ATE \equiv \frac{1}{N}\sum_{i=1}^{N}\tau_i \qquad (1)$$

$$\frac{1}{N}\sum_{i=1}^{N}Y_i(1) - \frac{1}{N}\sum_{i=1}^{N}Y_i(0) = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) = \frac{1}{N}\sum_{i=1}^{N}\tau_i \qquad (2)$$

Repeated random sampling and expected values

Repeated random sampling

- Imagine we take an infinite number of samples of size N from the population

Expected value

- Expected value of a variable is the average value of a random variable based on an infinite number of samples
- expected value of discrete random variable X: $E[X] = \sum xPr[X = x]$
  - $Pr[X = x]$ is probability that X takes on the value x, where summation is taken over all possible values of X
- Example of expected value of dice role, X:
  - $E[X] = (1)(\frac{1}{6}) + (2)(\frac{1}{6}) + (3)(\frac{1}{6}) + (4)(\frac{1}{6}) + (5)(\frac{1}{6}) + (6)(\frac{1}{6}) = 3.5$

Conditional expectations

Conditional expectations refer to subgroup averages

Example: Y_i=income; d_i=internship (0,1); Z_i=GPA

- $E[Y_i|d_i = 1]$
  - Expected value of (observed) income, given that student got internship
- $E[Y_i|Z_i > 3.5]$
  - Expected value of (observed) income, given that college GPA was greater than 3.5
- $E[Y_i(1)|d_i = 1]$
  - Expected value of of treated potential outcome, given that treatment student received treatment
- $E[Y_i(0)|d_i = 1]$
  - Expected value of untreated potential outcome, given that student did receive internship

Conditional expectations, potential outcomes, and random assignment

Random variable D_i

- D_i is a variable whose value is randomly assigned (e.g., coin flip, or some random number generator)
- Value of D_i determines whether person i receives treatment

Potential outcomes

- $E[Y_i(1)|D_i = 1]$
  - Treated potential outcome, given i assigned to treatment
  - Observed?: Yes
- $E[Y_i(1)|D_i = 0]$
  - Treated potential outcome, given i not assigned to treatment
  - Observed?: No
- $E[Y_i(0)|D_i = 1]$
  - Untreated potential outcome, given i assigned to treatment
  - Observed?: No
- $E[Y_i(0)|D_i = 0]$
  - Untreated potential outcome, given i not assigned to treatment
  - Observed?: Yes

Random assignment and unbiased inference: why random assignment works

Imagine 3 people [in our sample of 7] randomly assigned to internship

i	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Treatment effect**
1	65	**60**	?
2	30	**35**	?
3	**55**	60	?
4	**25**	30	?
5	50	**50**	?
6	80	**70**	?
7	**45**	45	?
Avg. (observed)	**45**	**53.75**	**-8.75**
Avg. (potential)	**50**	**50**	**0**

Random assignment and unbiased inference: why random assignment works

Now imagine that start over (re-sample): randomly assign 3 people to receive internship

i	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Treatment effect**
1	**65**	60	5
2	30	**35**	-5
3	55	**60**	-5
4	**25**	30	-5
5	50	**50**	0
6	**80**	70	10
7	45	**45**	0
Avg. (observed)	**56.67**	**47.5**	**9.17**
Avg. (potential)	**50**	**50**	**0**

If we re-sampled an infinite number of times, and calculated the average of the "average observed treatment effect" it would equal the "average potential treatment effect"

Why random assignment works

- Every person has same probability of getting treatment ($D_i = 1$); therefore, expected treated potential outcome among treated people is same as expected outcome for all people in sample
  - $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
- Every person has same probability of getting control ($D_i = 0$); therefore, expected untreated potential outcome among untreated people is same as expected outcome for all people in sample
  - $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$
- Because assignment to treatment is random:
  - Assignment to treatment has no effect on value of the potential outcomes; it just affects which potential outcome is observed for each person
  - Assignment to treatment has no relationship to characteristics (e.g., parental income) that affect value of potential outcomes
- $ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$
  - recall: for each person, only difference between $Y_i(1)$ and $Y_i(0)$ is the treatment

Observed outcomes, self-select into internship

Imagine if people self-selected into the internship; do you think assignment to treatment would be unrelated to value of potential outcomes $Y_i(1)$ and $Y_i(0)$

i	$Y_i(1)$ **Treated**	$Y_i(0)$ **Untreated**	τ_i **Treatment effect**
1	**65**	60	?
2	30	**35**	?
3	**55**	60	?
4	25	**30**	?
5	**50**	50	?
6	**80**	70	?
7	45	**45**	?
Average	**59.75**	**36.67**	**23.08**

The same characteristics (e.g., parental income, GPA) that determine the value of the dependent variable (income) also drive selection into the treatment (internship)

Observational design

Population linear regression model

- *Population* Linear Regression Model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
- Where:
  - Y_i = income for person i
  - X_i = hours worked for person i
  - β_0 (called "population intercept") = average income for someone with X=0 (i.e., works zero hours)
  - β_1 (called "population regression coefficient") = average effect of a one-unit increase in X on value of Y
  - u_i (called "error terms") = all other variables not included in your model that affect value of Y
- Draw scatterplot and population regression line
  - label components (e.g., residual=actual-predicted)

Population regression coefficient, β_1

RQ: What is the effect of hours worked per week (X) on income (Y)?

- Answer: population regression coefficient, β_1
- Estimating β_1 is the fundamental goal of program evaluation research

What is the population regression coefficient, β_1?

- β_1 measures the average change in Y for a one-unit increase in X
- Think of β_1 as measuring the slope of a line
  - $\beta_1 = \frac{\Delta Y}{\Delta X} = \frac{\Delta(income)}{\Delta(hoursworked)}$
  - $Example = \frac{\$5{,}000 \Delta \text{in income}}{1 \text{ hour} \Delta \text{in hours worked per week}} = \$5,000 = \beta_1$

Interpretation

- General interpretation:
  - On average, a one-unit increase in X is associated with a β_1 increase in the value of Y
- Interpretation for our research question:
  - On average, a one-hour increase in hours worked per week (X) is associated with a $\$\beta_1$ increase in annual income
- Imagine that β_1=2,000; How do we interpret this? β_1=4,000?

Population Intercept, β_0

- RQ: What is the effect of hours worked per week (X) on annual income (Y)
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
- β_0 (called "population intercept")
  - β_0 represents average value of Y when X=0
  - In our example, β_0 is average income for someone who works zero hours per week (X=0)
  - Draw in scatterplot
- Note:
  - Usually we are substantively interested in β_0
  - Also, do not believe β_0 if there are few observations where X=0 (e.g., effect of height on income)

Thinking about u_i as the "error term"

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y=income ; X_i = hours worked
- Thinking about u_i as the "error term"
  - In econometrics:
    - Error term, u_i, consists of all other variables not included in your model that affect the dependent variable
    - This interpretation of u_i is very important for program evaluation research
    - What variables besides hours worked (X) affect income (Y)?
  - In "conventional" statistics textbooks:
    - Overall error in prediction of Y (due to random variation)

Thinking about u_i as the "residual"

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y=income ; X_i = hours worked
- Draw scatterplot
  - Population regression line represents the predicted value of Y (income) for each value of X (hours worked)
- Thinking about u_i in terms of each observation, i
  - Y_i = actual value of income for person i
  - $(\beta_0 + \beta_1 X_i)$ = population regression line
    - Equals the predicted value of income for person i with hours worked = X_i
  - Residual, u_i
    - Residual, u_i, is the difference between actual value, Y_i, and predicted value from the population regression model for observation i
    - $u_i = Y_i - (\beta_0 + \beta_1 X_i)$

General things we do in regression analysis

1. Estimation
   - Choose estimates of β_0 and β_1 based on sample data
   - $\hat{\beta}_0$ is an estimate of β_0; $\hat{\beta}_1$ is an estimate of β_1
2. Prediction
   - What is the predicted value of Y for someone with a particular value of X
     - e.g., what is the predicted income for someone w/ an undergraduate degree in chemistry?
3. Hypothesis testing [focus of causal inference]
   - Causal interpretation of β_1:
     - the effect of a one-unit increase in X is a β_1 increase in Y
   - Hypothesis testing and confidence intervals about β_1
     - Use $\hat{\beta}_1$ to test hypotheses about β_1
     - If we knew β_1, we would not need hypothesis testing

Estimation (regression) [SKIP]

- Problem in regression:
  - Need to develop a method for choosing values of $\hat{\beta}_0$ and $\hat{\beta}_1$
  - Solution: similar to what we did for population mean
- First, some terminology (draw scatterplot):
  - Y_i is the actual value of Y for individual i
  - $\hat{Y}_i$ is the predicted value Y_i, based on sample data
    - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - Residual, $\hat{u}_i$ = difference between actual, Y_i, and predicted, $\hat{Y}_i$
    - $Y_i - \hat{Y}_i = \hat{u}_i$
    - $Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{u}_i$
    - Note: residuals, $\hat{u}_i$, are often referred to as "errors"

Estimation (regression) [SKIP]

- Criteria for choosing $\hat{\beta}_0$ and $\hat{\beta}_1$
  - Choose values for that minimize "sum of squared residuals"
- Residuals, $\hat{u}_i$
  - $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$
- Sum of squared residuals [or "sum of squared errors"]:
  - $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
  - $\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$
  - $\sum_{i=1}^{n}(\hat{u}_i)^2$
- Draw on scatterplot with two different regression lines

Ordinary least squares estimates

- The OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values that minimize the sum of squared residuals:
  - $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 = \sum_{i=1}^{n}(\hat{u}_i)^2$
- This minimization problem is solved using calculus
  - Stata does this for you
- Important point:
  - Any other choice of $\hat{\beta}_0$ and $\hat{\beta}_1$ will result in higher sum of squared errors
    - Same idea as when we found estimate of population mean, μ_Y
  - Draw two scatterplots: one with OLS estimates; one with non-OLS estimates (e.g., mean)

OLS prediction line

- *Population* linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Where Y = income ($000); X= hours worked
- OLS prediction line (based on OLS estimates)
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - Note: OLS prediction line also called "OLS regression line"
- Imagine OLS estimates are $\hat{\beta}_0$=5 and $\hat{\beta}_1$=2
  - What is the predicted income for someone who works 0 hours per week?
  - What is the predicted income for someone who works 10 hours per week?

OLS Assumption 1 (mathematically)

- Role of assumptions in statistics and causal inference
- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
    - Y= HS test score; X=0/1 MAS; u_i = all other variables that affect Y but were not included in regression model
- Assumption 1 mathematically
  - $E(u_i|X_i) = 0$
  - "expected value of u_i, given any value of X, equals 0"
- OLS Assumption 1 in words
  - the independent variable X_i is unrelated to the "other variables", u_i, not included in model
  - Pretend that u_i consists of only one variable (e.g., "grit")
  - OLS assumption 1 states that the mean value of omitted variable is equal to zero no matter what the value of variable X is

OLS Assumption 1

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
    - Y= HS test score; X=0/1 MAS; u_i = all other variables that affect Y but were not included in regression model
- Assumption 1: $E(u_i|X_i) = 0$
  - In words: the independent variable X_i is unrelated to the "other variables", u_i, not included in model
- Assumption is *always* satisfied in random assignment experiment
  - Example: effect of MAS participation (X) on graduation (Y)
    - X=0 (non participant); X=1 (MAS participant)
  - We randomly assign students to MAS participation (X)
  - Other factors, u_i (includes "grit", parental involvement, "aptitude", etc.) are *by construction* unrelated to values of X because we randomly assigned students to values of X
- In observational studies, this assumption is usually violated
  - E.g., MAS participation (X) is likely correlated with omitted variables u_i, (e.g. motivation) that affect Y

OLS Assumption 1 in practice

- Assumption 1: $E(u_i|X_i) = 0$
  - In words: the independent variable X_i is unrelated to the "other factors", u_i, not included in model
- How to think about it in practice:
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Are there any variables that are not in your model that affect Y and have a relationship (positive or negative) with X? If so, Assumption 1 is violated
- Any omitted variables that violate assumption 1?
  - Effect of participating in fraternity (X) on GPA (Y)?
  - Effect of years of education (X) on income (Y)?
  - Effect of participating in "summer bridge program" (X) on first-year retention (Y)?
  - Effect of participating in Think Tank (X) on first-year retention?

Omitted Variable Bias

- $Y_i = \beta_0 + \beta_1 X_i + u_i$; Y=test score; X=class size
  - We want to know the *causal effect* of X on Y
- Omitted Variable Bias
  - Bias in estimate $\hat{\beta}_1$ due to variables being omitted from the model (part of u_i rather than included in model)
  - Omitted variable bias is really about OLS assumption #1
- For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:
  1. Z affects value of Y (i.e. Z is part of u); ***and***
  2. Z has a relationship with X (e.g., correlation; $corr(Z, X) \neq 0$)

OLS Assumption 1 & Omitted Variable Bias

- OLS Assumption 1: $E(u_i|X_i) = 0$
  - the independent variable X_i is unrelated to the "other factors," u_i, that affect Y and are not included in model
  - assumption violated if there are omitted variables that affect Y that also have relationship with X
- Omitted Variable Bias
  - omitted variable bias: bias in estimate $\hat{\beta}_1$ due to variables being omitted from the model
  - For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:
    1. Z affects value of Y (i.e. Z is part of u); ***and***
    2. Z has a relationship with X (e.g., correlation; $corr(Z, X) \neq 0$)

Omitted variable bias in practice

- $Y_i = \beta_0 + \beta_1 X_i + u_i$;
  - Y=test score; X=class size
  - Z = % of students in district with English as a second language (ESL) [omitted from model]
- For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:
  1. Z affects value of Y (i.e. Z is part of u); ***and***
     - Would ESL affect standardized test scores? Why?
  2. Z has a relationship with X (e.g., correlation; $corr(Z, X) \neq 0$)
     - Is ESL likely to be correlated with student-teacher ratio? Why?
- If *both* conditions satisfied, then omission from ESL from model results in $\hat{\beta}_1$ having omitted variable bias

Omitted variable bias in practice

- $Y_i = \beta_0 + \beta_1 X_i + u_i$;
  - Y=test score; X=class size; Z= variable omitted from model
- For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:
  1. Z affects [causal] value of Y (i.e. Z is part of u); ***and***
  2. Z has a relationship with X (e.g., correlation; $corr(Z, X) \neq 0$)
- Would omitting Z = "time of day test administered" result in omitted variable bias?
  - Does test-time affect Y? Is test-time correlated with student-teacher ratio?
- Would omitting Z = "number of desks in class" result in omitted variable bias?
  - Does parking space per pupil affect Y? Is parking space per pupil correlated with student-teacher ratio?

How to check for omitted variable bias

- Ask yourself:
  - Could omitted variable Z affect Y?
  - Could omitted variable Z have some relationship with X?
- How researchers think about omitted variable bias in practice
  - Rely on logical argument
  - Rely on theory
  - Rely on prior research
    - e.g., past studies show that Z affects Y
    - past studies of "effect of X on Y" control for Z
- descriptive statistics (e.g., correlations)
  - only works if you have a good measure of Z

When is omitted variable bias big/small

- $Y_i = \beta_0 + \beta_1 X_i + u_i$
- Imagine there is only one omitted variable, Z
- Omitted variable bias is likely big when:
  - The omitted variable, Z, has a big causal effect on Y
  - The correlation between X and the omitted variable, Z, is strong
- Example:
  - Y= earnings ; X= participation in internship; Z= parental income

Omitted Variable Bias Formula [SKIP]

- $Y_i = \beta_0 + \beta_1 X_i + u_i$
- Omitted variable bias formula [assume u is one variable]
    - $\hat{\beta}_1 \underset{\rightarrow}{p} \beta_1 + corr(X_i, u_i) * \frac{\sigma_u}{\sigma_x}$
- What do different components of formula mean?
    - $\underset{\rightarrow}{p} =$ "approaches this value as sample size increases"
    - $\sigma_u =$ standard deviation of omitted variable(s), u?
    - $\sigma_X =$ standard deviation of X
- Formula in words
    - As sample size increases, the OLS estimate, $\hat{\beta}_1$, approaches the population regression coefficient β_1 + the correlation between X and u times the standard deviation of u divided by the standard deviation of X
- What value do we want $\hat{\beta}_1$ to approach as sample size increases?
- Omitted variable bias is this part of formula: $corr(X_i, u_i) * \frac{\sigma_u}{\sigma_x}$
    - Omitted variable bias is high when:
        - strong correlation between X and u [can see from formula]
        - the omitted variable, u, has a big effect on Y

Omitted Variable Bias Formula [SKIP]

- $\hat{\beta}_1 \underset{\rightarrow}{p} \beta_1 + corr(X_i, u_i) * \frac{\sigma_u}{\sigma_x}$
- Consistency
  - The point estimate approaches the population parameter as sample size increases (e.g. imagine if your sample is the entire population)
  - $\hat{\beta}_1$ is inconsistent when there is omitted variable bias; $\hat{\beta}_1$ does not approach β_1 as sample size increases
- $corr(X_i, u_i)$
  - If $corr(X_i, u_i) = 0$ then there is no bias; $\hat{\beta}_1 \underset{\rightarrow}{p} \beta_1$
  - The size of the omitted variable bias depends on the strength of the correlation between X and u

Upwards/Downwards Bias [SKIP]

- $Y_i = \beta_0 + \beta_1 * MAS_i + u_i$
  - Y= graduation; X=0/1 in MAS; X2= motivation; X3= household income
  - β_1 = true causal effect of participation in MAS on graduation
- Upwards bias: $\hat{\beta}_1 > \beta_1$
  - Estimate of the causal effect $\hat{\beta}_1$, is greater than true causal effect, β_1
  - Example: Omit Z1, student motivation
    - motivation positively affects graduation; positive correlation w/ participation in MAS
    - If we omit student motivation from model, our estimate $\hat{\beta}_1$ is partially picking up positive effect of motivation on graduation
- Downwards bias: $\hat{\beta}_1 < \beta_1$
  - Estimate of the causal effect $\hat{\beta}_1$, is less than true causal effect, β_1
  - Example: Omit Z2, household income
    - household income positively affects graduation; negative correlation w/ participation in MAS
    - If we omit household income from model, our estimate $\hat{\beta}_1$ is partially picking up negative effect of being low-income on graduation (because low income students are more likely to participate in MAS)

How to deal with Omitted Variable Bias

- Random assignment experiments
  - The "gold standard"
- Attempt to "recreate" experimental conditions
  - Multiple regression, matching
    - Include omitted variables in your model, so they are no longer omitted
    - This is the purpose of regression; otherwise you can use ANOVA
  - "quasi-experimental" techniques
    - More advanced methods for recreating experimental conditions
    - e.g., regression discontinuity; instrumental variables

How to deal with Omitted Variable Bias

- Research question: What is the effect of student teacher ration (X) on district average test scores (Y)?
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Imagine that we have two omitted variables
    - Z1= pct English as a Second Language (ESL)
    - Z2= average income in the district
- Multiple regression:
  - Attempt to recreate experimental conditions by including "omitted variables" in your model, so they are no longer omitted
    - Once you include Z1 and Z2 in your model, they are called "control" variables because they control for omitted variable bias; also called "covariates"
- Do in Stata

Why no control variables in experiments

- $Y_i = \beta_0 + \beta_1 X_i + u_i$; Y=graduation; X: 0=no MAS, 1=MAS
  - Omitted vars: Z1= student motivation; Z2= household income
- Omitted variable bias conditions:
  1. Z affects value of Y (i.e. Z is part of u); ***and***
  2. Z has a relationship with X (e.g., correlation)
- Imagine students randomly assigned to MAS
  - Z1 = student motivation
    - (1) does student motivation affect HS graduation(Y)?
    - (2) could student motivation be related (e.g., correlation) with value of X (MAS)?
  - Z2 = household income
    - (1) does household income affect HS graduation(Y)?
    - (2) could household income be related to value of X (MAS)?
- Randomization in treatment vs. control group
  - Any differences between treatment and control group on factors that affect Y have no relationship w/ value of X (treatment)

Why no control variables in experiments: Tennessee STAR Experiment

- Where to get data
  - https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766
- Overview of experiment
  - "The Student/Teacher Achievement Ratio (STAR) was a four- year longitudinal class-size study funded by the Tennessee General Assembly and conducted by the State Department of Education. Over 7,000 students in 79 schools were randomly assigned into one of three interventions: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide)
- RQ: what is effect of class-size treatment (X) on first-grade math scores (Y)
  - Do in Stata

Conditional Independence Assumption

- Assume students choose to participate in MAS
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
  - Y= graduation; X=0/1 in MAS; X2= motivation; X3= income
- Conditional independence assumption
  - Once we include control variables, there are no omitted variables, Z, that satisfy *both* of these two conditions
    1. Z affects value of Y (i.e. Z is part of u); ***and***
    2. Z has a relationship with X (e.g., correlation)
- If the conditional independence assumption is true:
  - Once we include relevant control variables, there are no omitted variables that affect Y and have a systemic relationship with X
  - Main point: if we satisfy the conditional independence assumption through control variables, then multiple regression is just as good as random assignment experiment!
    - In random assignment experiments, there are omitted variables that affect Y, but none of these omitted variables have a systemic relationship with X because X is randomly assigned

Population Multiple Regression Model

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ .... \ + \beta_k X_{ki} + u_i$
- Where:
  - Y_i = observation i of dependent variable
  - X_{1i} = observation i for the first regressor, X_1
  - X_{2i} = observation i for the second regressor, X_2
  - X_{ki} = observation i for the kth regressor, X_2
  - β_1 = population average effect on Y for a one-unit increase in X_1
  - β_2 = population average effect on Y for a one-unit increase in X_2
  - β_k = population average effect on Y for a one-unit increase in X_k
  - β_0 = average value of Y when the value of all independent variables, $X_1, X_2, ...X_k$, are equal to zero
  - u_i = all other variables that *affect* the value of Y_i but are not included in the model (i.e., not X_1 or X_2)
  - k = refers to the number of independent variables in your model
    - e.g., model where independent variables are age, education level, and income has k=3

Multivariate Regression & Program Evaluation

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
- Program evaluation research [econometrics]
  - We are only interested in estimating β_1 [causal effect of X1 on Y]
  - The only reason we include other variables in the model beside X1 is to eliminate omitted variable bias
  - Therefore, we include all control variables that satisfy *both* conditions of omitted variable bias:
  - Once we include control variables, there are no omitted variables, Z, that satisfy *both* of these two conditions
    1. Z affects value of Y (i.e. Z is part of u); ***and***
    2. Z has a relationship with X (e.g., correlation)
- Traditional social science statistics
  - Purpose of multiple regression is to add new variable to your model (e.g. X_3) to see effect of variable X_3 on Y
  - Can lead to sloppy research! If you don't get an "interesting" result for $\hat{\beta}_1$, then focus on a variable with a more interesting coefficient (e.g. X_3)

What does "holding constant" mean?

- RQ: What is the relationship between years of education(X1) on income(Y), after controlling for years of work experience (X2)?
- "Holding the value of X2 constant"
  - Means to estimate the relationship between X1 and Y when we don't allow value of X2 to vary [partial derivative]
  - Said different: relationship between education (X1) and income (Y) for applicants that have same years of experience (X2)
- General interpretation of β_1 (assuming causal relationship):
  - The average effect of a one-unit increase in X_1 is a β_1 unit increase in Y, holding the value of X_2 constant
- Interpretation of β_1, applied to example
  - The effect of having one additional year of education (X1) on income (Y), when we don't allow value "years of experience" (X2) to change
  - Said different: the effect of increasing years of education on income for people who have same years of experience