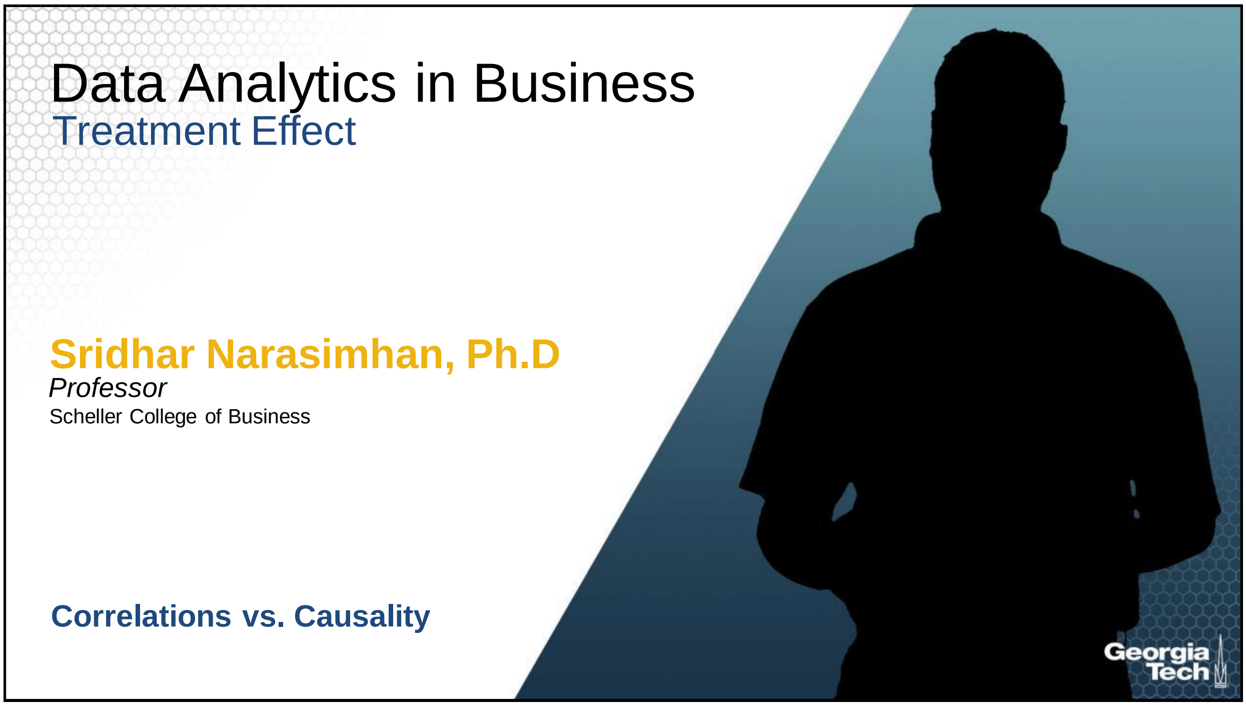


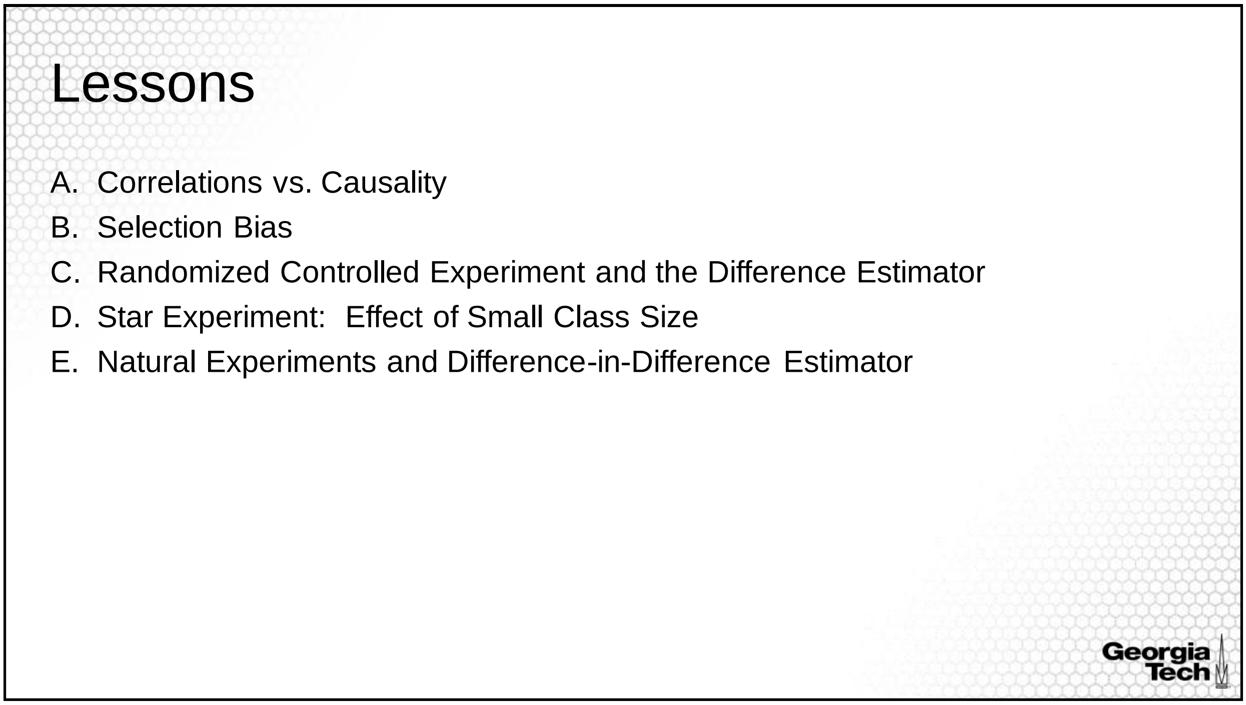
Data Analytics in Business

Treatment Effect

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

Correlations vs. Causality

Georgia Tech



Lessons

A. Correlations vs. Causality
B. Selection Bias
C. Randomized Controlled Experiment and the Difference Estimator
D. Star Experiment: Effect of Small Class Size
E. Natural Experiments and Difference-in-Difference Estimator

Georgia Tech

Correlation

- Correlation refers to any of a broad class of statistical relationships involving two variables
- The (sample) correlation between two variables *X* and *Y* is defined as:

$$Corr\ (X, Y) = \frac{\sum_{i=1}^{n}(x_i\ -\ \bar{x})(y_i\ -\ \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i\ -\ \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i\ -\ \bar{y})^2}}$$

- ***Correlation*** is a measure of the ***linear relationship*** between *X* and *Y*

  -1 <= $Corr\ (X, Y)$ <= 1
- If Y = X^2, and if X ranges from, say, -10 to 10, then the correlation between X and Y is 0; in other words, they are uncorrelated even though they are perfectly related

Strong Correlation Between A and B

If A and B are strongly correlated, there could be several possible relationships:

- A causes B
- B causes A
- If you think A causes B, be careful of reverse causality (B causes A)
  - e.g., The faster that a windmill rotates, the more the wind speed is observed to be!
- A and B are a consequence of some other common cause C
  - e.g., As ice cream sales increase, the rate of drowning deaths increases sharply. Therefore, ice cream consumption causes drowning. Here, C could be summer season!
- A causes C which in turn causes B
- A and B are correlated by chance!

Post Hoc Ergo Propter Hoc

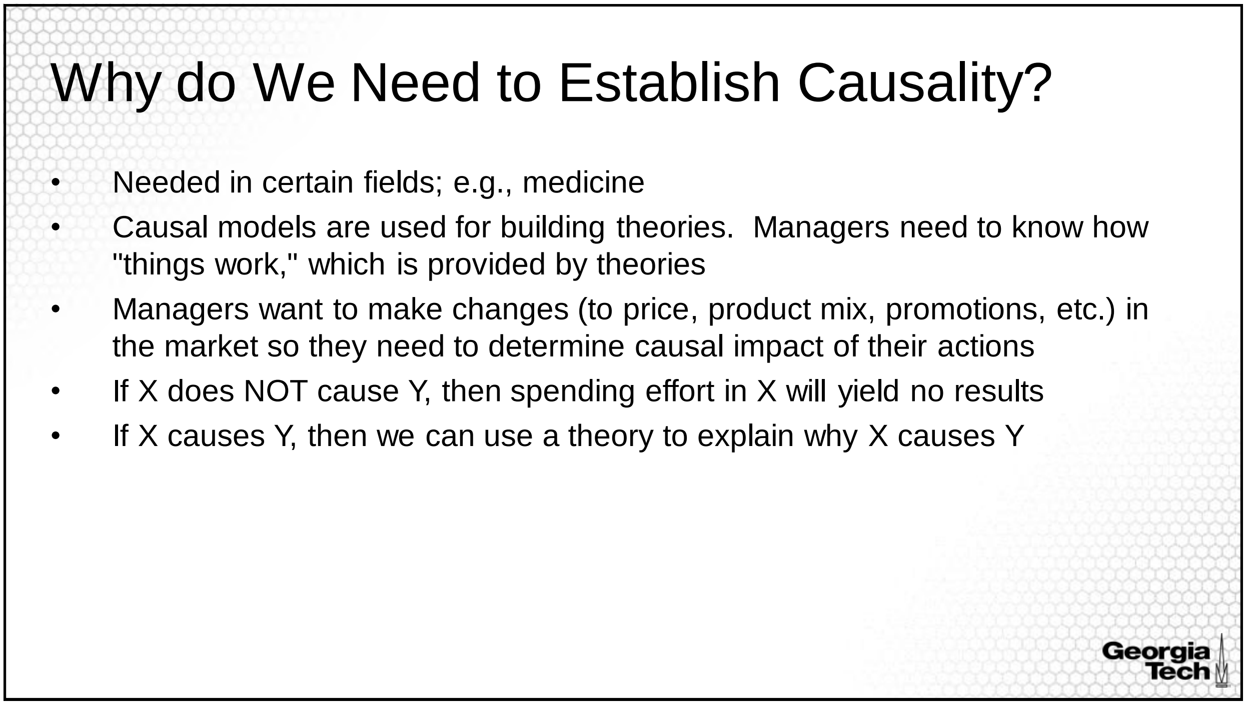
- Beware of faulty line of reasoning called **post hoc ergo propter hoc**:
  - From the Latin “after this, therefore because of this.”
  - It is the logical fallacy that, if A happened and then B happened, then A must have caused B to happen
- Examples:
  - The rooster crows just before sunrise; thus the rooster causes the sun to rise
  - Prof. Sri Narasimhan moves into a new office and the college’s air-conditioning breaks down; thus Prof. Sri caused the A/C to break
- Be very careful when you make a conclusion based *solely* on the order of events; rather, you should also consider other factors that could be responsible for the result that might rule out the order-of-events connection

Causation

To establish causation:

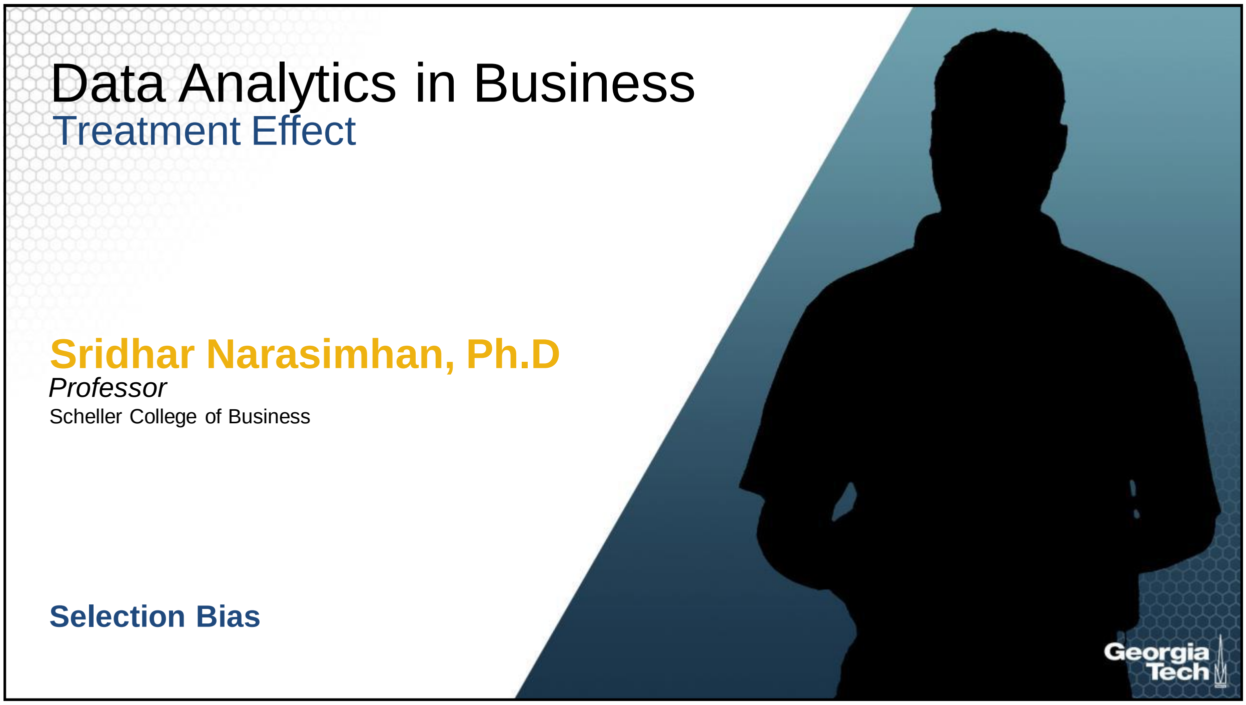
1. In time, the hypothesized cause must precede its anticipated effect
2. The change in cause must lead to a change in effect
3. Must discount all other plausible explanations, other than the one proposed, that could explain the relationship

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford, England: Oxford University Press.



Why do We Need to Establish Causality?

- Needed in certain fields; e.g., medicine
- Causal models are used for building theories. Managers need to know how "things work," which is provided by theories
- Managers want to make changes (to price, product mix, promotions, etc.) in the market so they need to determine causal impact of their actions
- If X does NOT cause Y, then spending effort in X will yield no results
- If X causes Y, then we can use a theory to explain why X causes Y

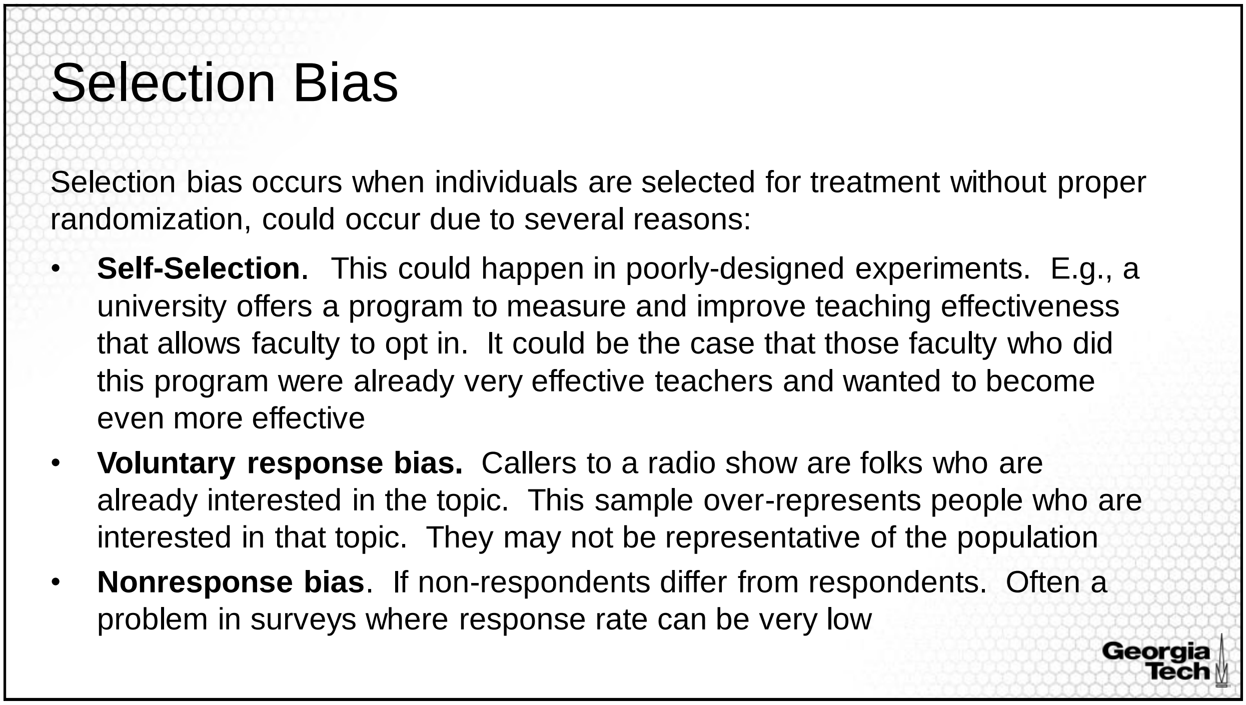


Data Analytics in Business

Treatment Effect

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

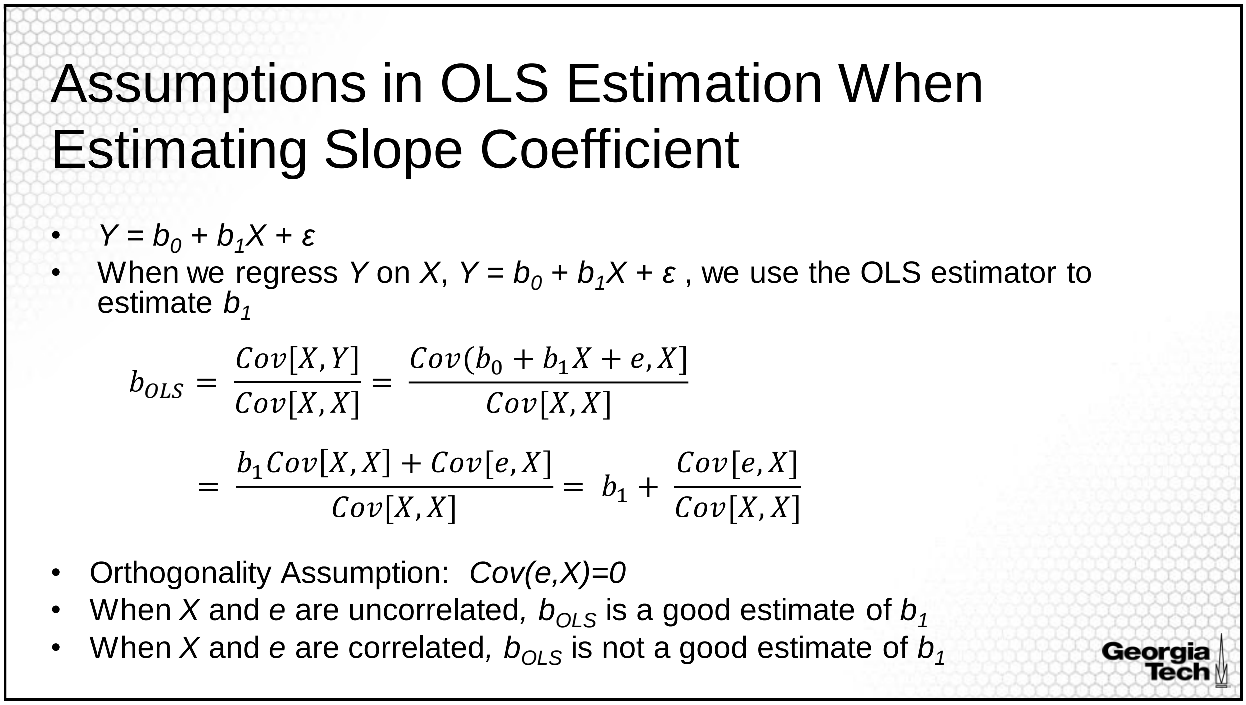
Selection Bias



Selection Bias

Selection bias occurs when individuals are selected for treatment without proper randomization, could occur due to several reasons:

- **Self-Selection**. This could happen in poorly-designed experiments. E.g., a university offers a program to measure and improve teaching effectiveness that allows faculty to opt in. It could be the case that those faculty who did this program were already very effective teachers and wanted to become even more effective
- **Voluntary response bias.** Callers to a radio show are folks who are already interested in the topic. This sample over-represents people who are interested in that topic. They may not be representative of the population
- **Nonresponse bias**. If non-respondents differ from respondents. Often a problem in surveys where response rate can be very low



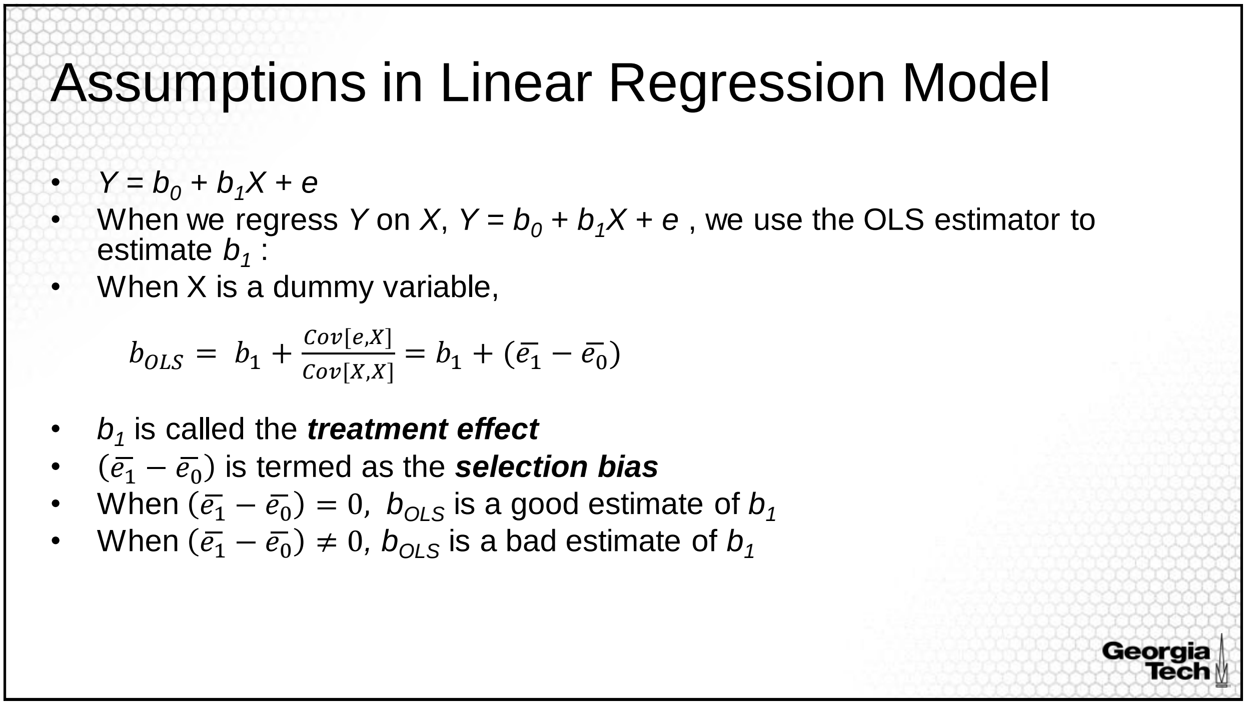
Assumptions in OLS Estimation When Estimating Slope Coefficient

- $Y = b_0 + b_1X + \varepsilon$
- When we regress Y on X, $Y = b_0 + b_1X + \varepsilon$, we use the OLS estimator to estimate b_1

$$b_{OLS} = \frac{Cov[X,Y]}{Cov[X,X]} = \frac{Cov(b_0 + b_1X + e, X]}{Cov[X,X]}$$

$$= \frac{b_1 Cov[X,X] + Cov[e,X]}{Cov[X,X]} = b_1 + \frac{Cov[e,X]}{Cov[X,X]}$$

- Orthogonality Assumption: $Cov(e,X)=0$
- When X and e are uncorrelated, b_{OLS} is a good estimate of b_1
- When X and e are correlated, b_{OLS} is not a good estimate of b_1

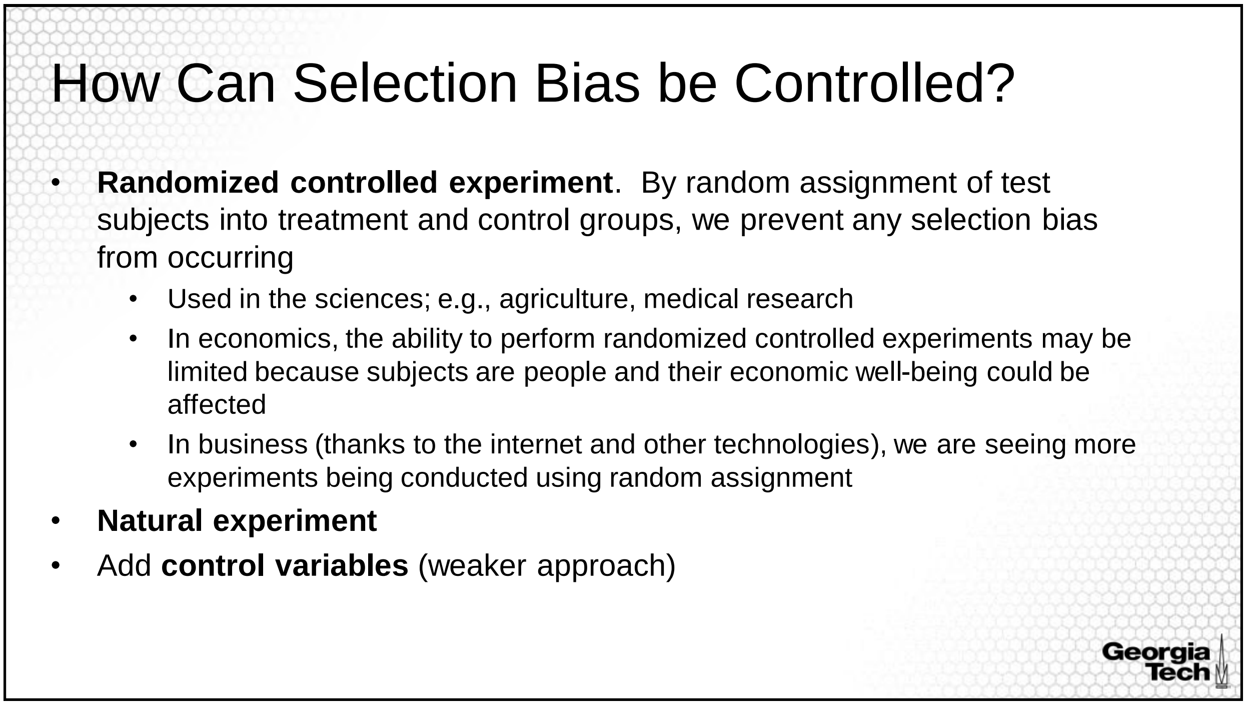


Assumptions in Linear Regression Model

- *Y* = $b_0 + b_1X + e$
- When we regress *Y* on *X*, $Y = b_0 + b_1X + e$, we use the OLS estimator to estimate b_1 :
- When X is a dummy variable,

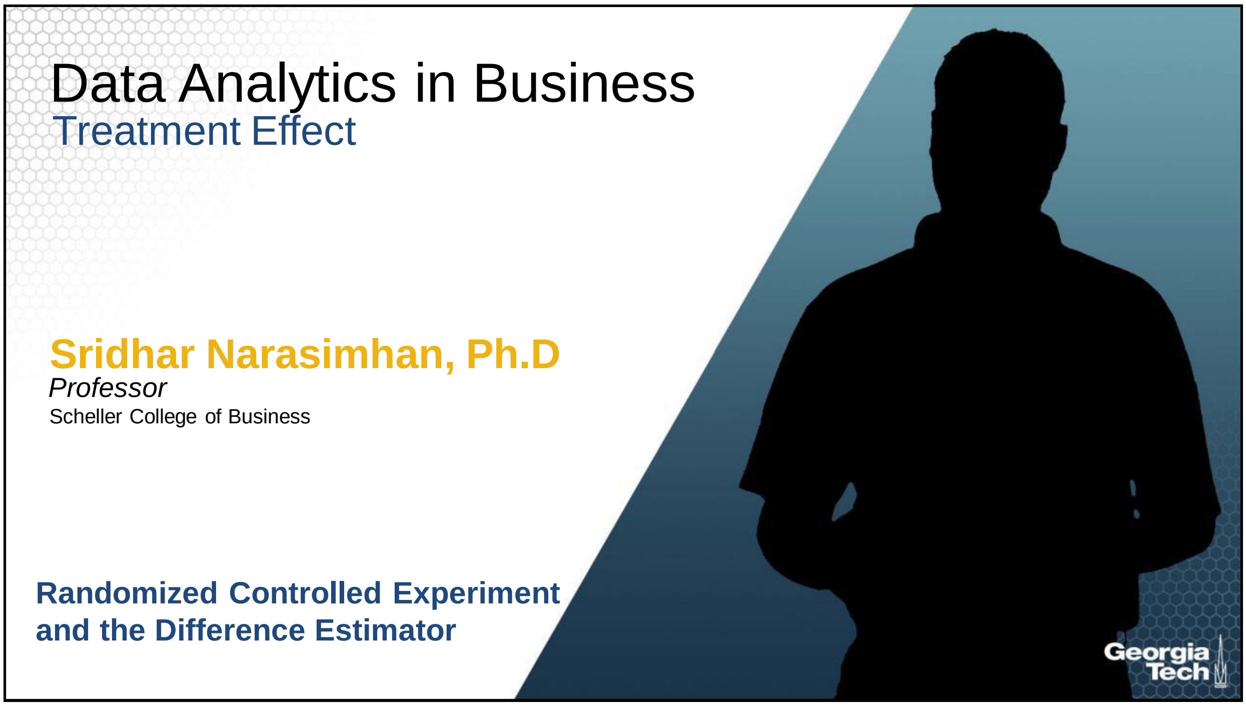
$$b_{OLS} = b_1 + \frac{Cov[e,X]}{Cov[X,X]} = b_1 + (\bar{e_1} - \bar{e_0})$$

- b_1 is called the ***treatment effect***
- $(\bar{e_1} - \bar{e_0})$ is termed as the ***selection bias***
- When $(\bar{e_1} - \bar{e_0}) = 0,$ b_{OLS} is a good estimate of b_1
- When $(\bar{e_1} - \bar{e_0}) \neq 0,$ b_{OLS} is a bad estimate of b_1



How Can Selection Bias be Controlled?

- **Randomized controlled experiment**. By random assignment of test subjects into treatment and control groups, we prevent any selection bias from occurring
  - Used in the sciences; e.g., agriculture, medical research
  - In economics, the ability to perform randomized controlled experiments may be limited because subjects are people and their economic well-being could be affected
  - In business (thanks to the internet and other technologies), we are seeing more experiments being conducted using random assignment
- **Natural experiment**
- Add **control variables** (weaker approach)



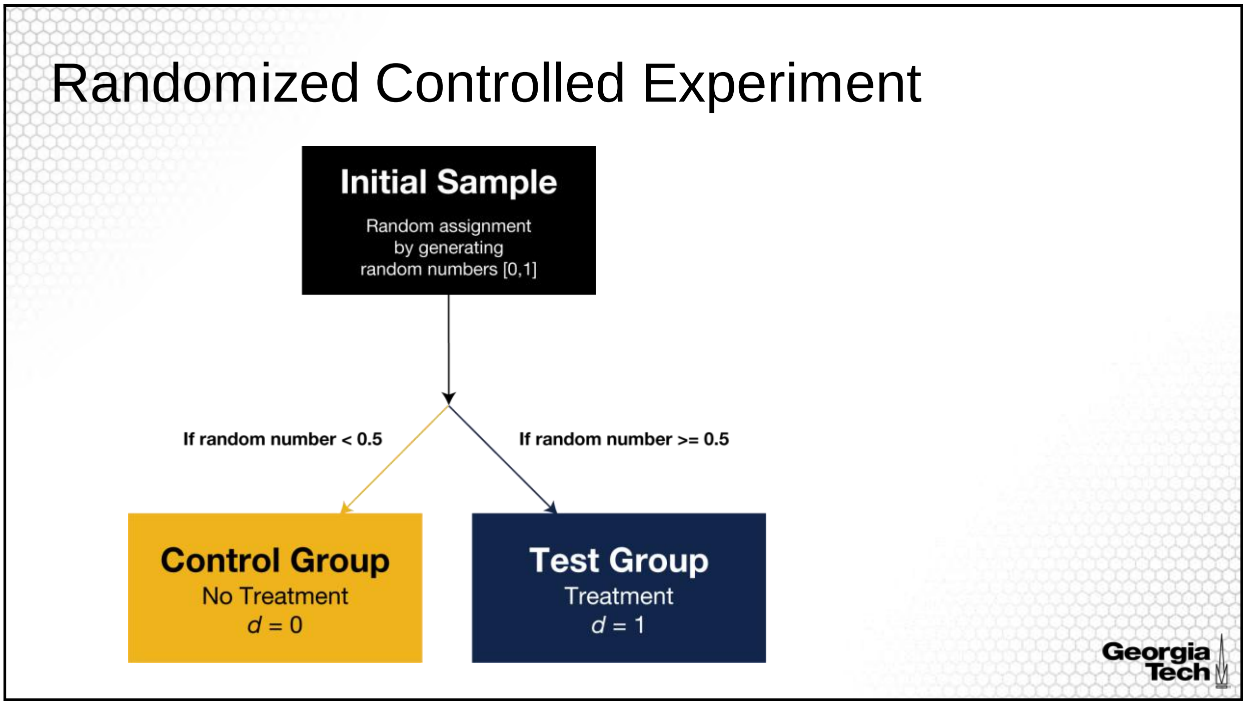
Data Analytics in Business

Treatment Effect

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

Randomized Controlled Experiment and the Difference Estimator

Georgia Tech



Randomized Controlled Experiment

Initial Sample

Random assignment by generating random numbers [0,1]

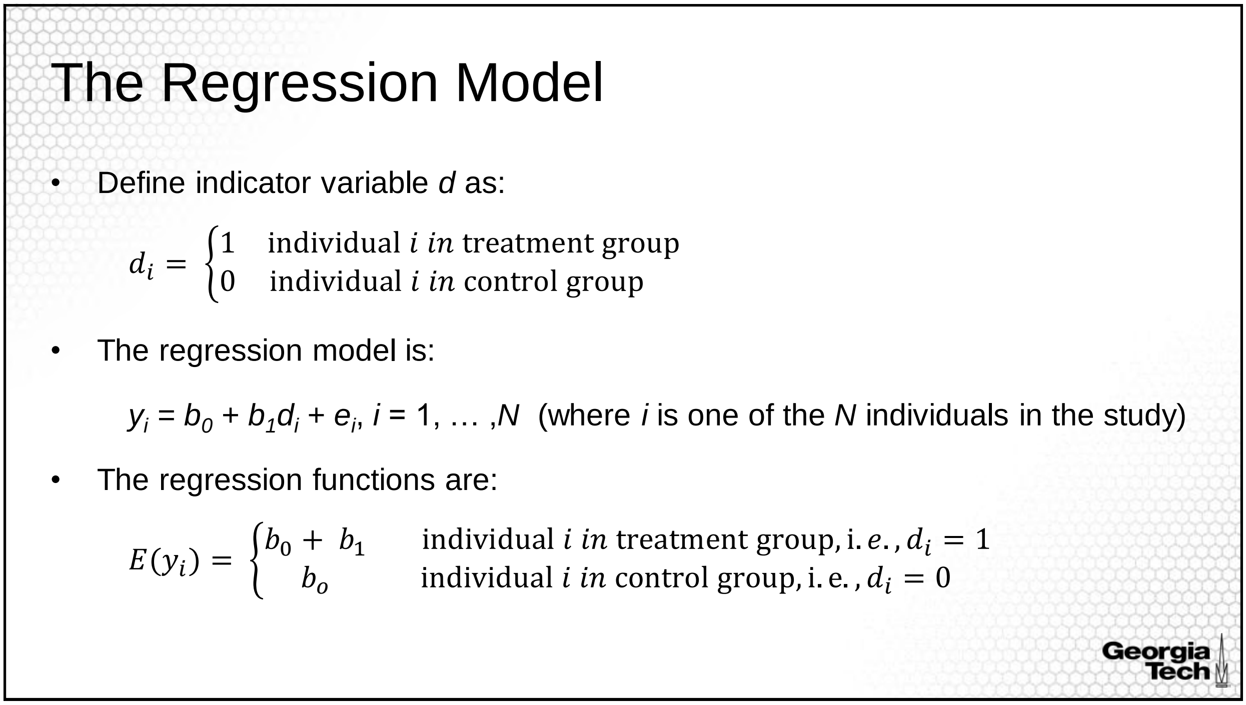
If random number < 0.5

If random number >= 0.5

Control Group
No Treatment
d = 0

Test Group
Treatment
d = 1

Georgia Tech



The Regression Model

- Define indicator variable *d* as:

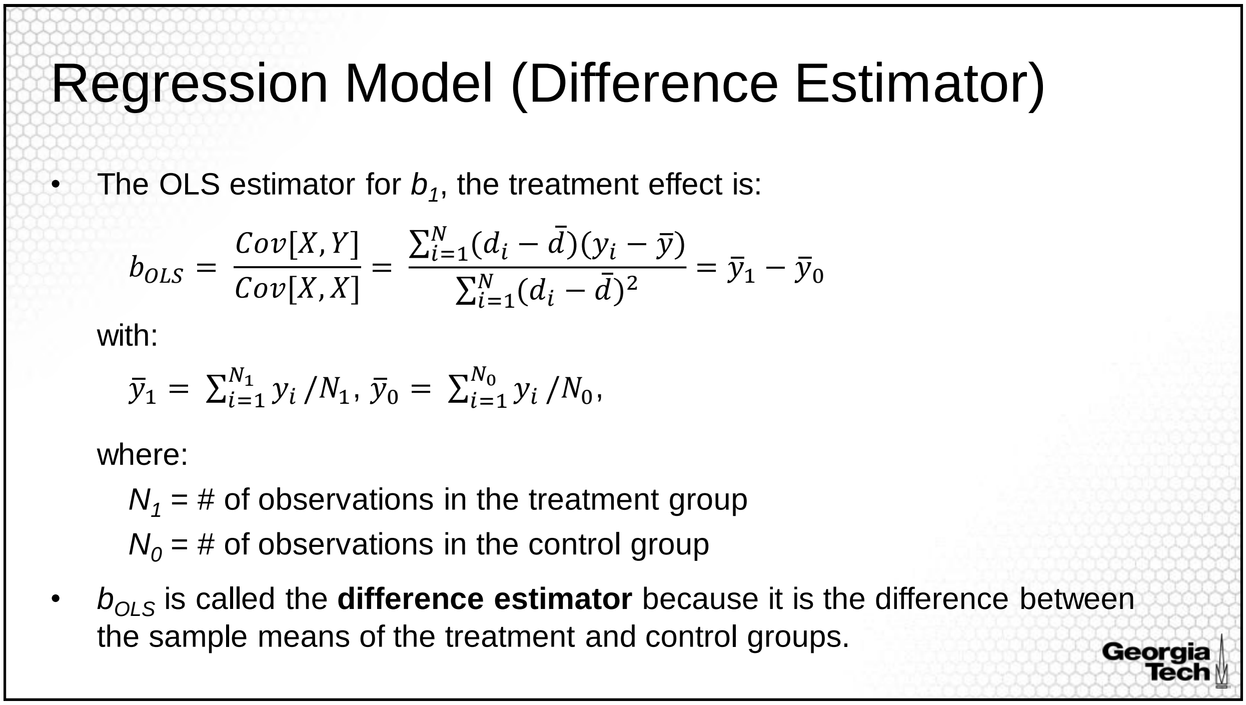
$$d_i = \begin{cases} 1 & \text{individual } i \; in \text{ treatment group} \\ 0 & \text{individual } i \; in \text{ control group} \end{cases}$$

- The regression model is:

  $y_i = b_0 + b_1 d_i + e_i$, $i = 1, \dots, N$ (where *i* is one of the *N* individuals in the study)

- The regression functions are:

$$E(y_i) = \begin{cases} b_0 + \; b_1 & \text{individual } i \; in \text{ treatment group, i.} e.\text{, } d_i = 1 \\ \quad b_o & \text{individual } i \; in \text{ control group, i.e., } d_i = 0 \end{cases}$$



Regression Model (Difference Estimator)

- The OLS estimator for b_1, the treatment effect is:

$$b_{OLS} = \frac{Cov[X,Y]}{Cov[X,X]} = \frac{\sum_{i=1}^{N}(d_i - \bar{d})(y_i - \bar{y})}{\sum_{i=1}^{N}(d_i - \bar{d})^2} = \bar{y}_1 - \bar{y}_0$$

  with:

$$\bar{y}_1 = \sum_{i=1}^{N_1} y_i / N_1, \; \bar{y}_0 = \sum_{i=1}^{N_0} y_i / N_0,$$

  where:

  N_1 = # of observations in the treatment group

  N_0 = # of observations in the control group

- b_{OLS} is called the **difference estimator** because it is the difference between the sample means of the treatment and control groups.

Regression Model (Difference Estimator)

- The difference estimator can we rewritten as:

$$b_{OLS} = \frac{\sum_{i=1}^{N}(d_i - \bar{d})(e_i - \bar{e})}{\sum_{i=1}^{N}(d_i - \bar{d})^2} = b_1 + (\bar{e}_1 - \bar{e}_0)$$

- If we allow individuals to self-select into treatment and control then $E(\bar{e}_1) - E(\bar{e}_0)$ is the selection bias in the estimation of the treatment effect
- By using random assignment of individuals to treatment and control groups, we have no systematic differences between the two groups, except for the treatment itself
- By using random assignment, we aim to have:

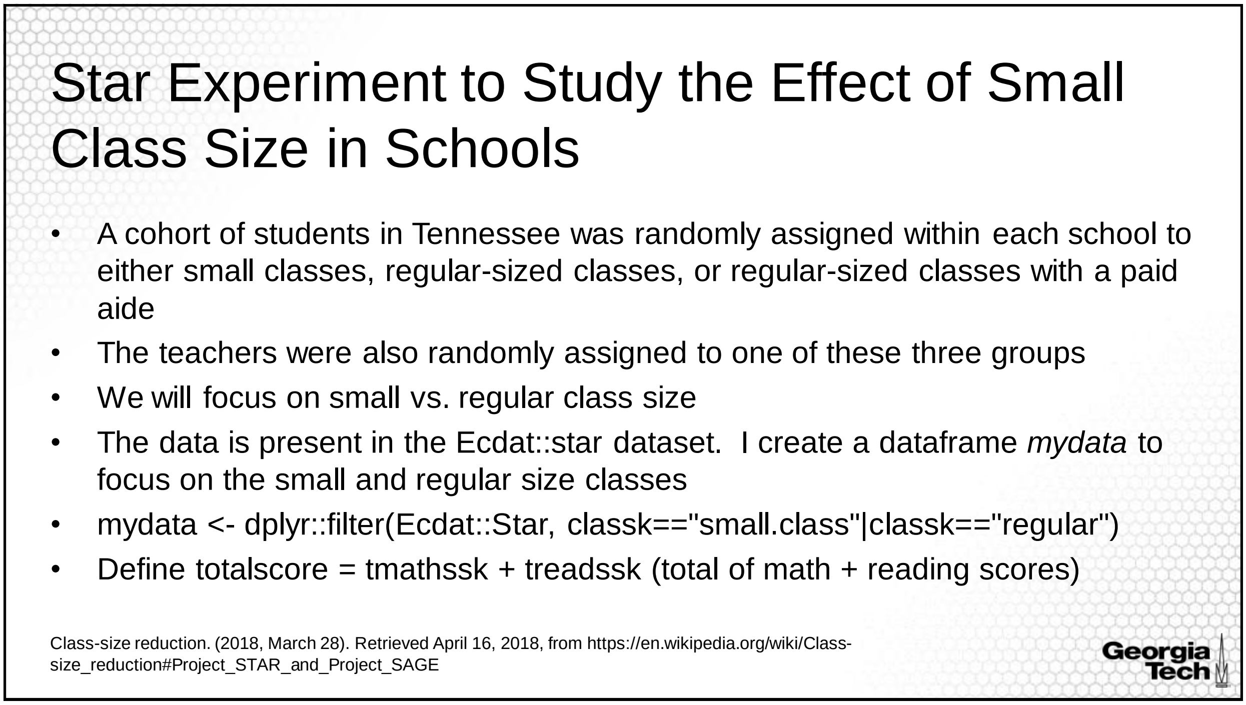
$E(\bar{e}_1 - \bar{e}_0) = E(\bar{e}_1) - E(\bar{e}_0)$ = 0, so that the OLS estimator is unbiased

Data Analytics in Business

Treatment Effect

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

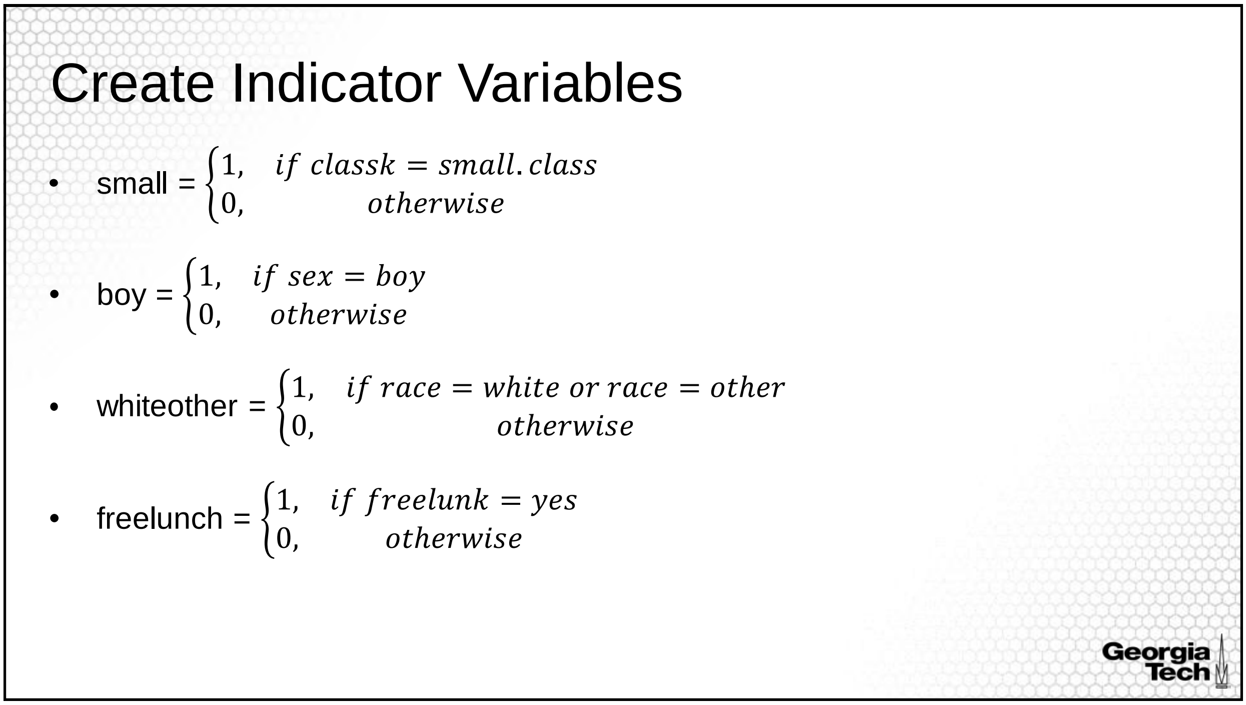
Star Experiment: Effect of Small Class Size



Star Experiment to Study the Effect of Small Class Size in Schools

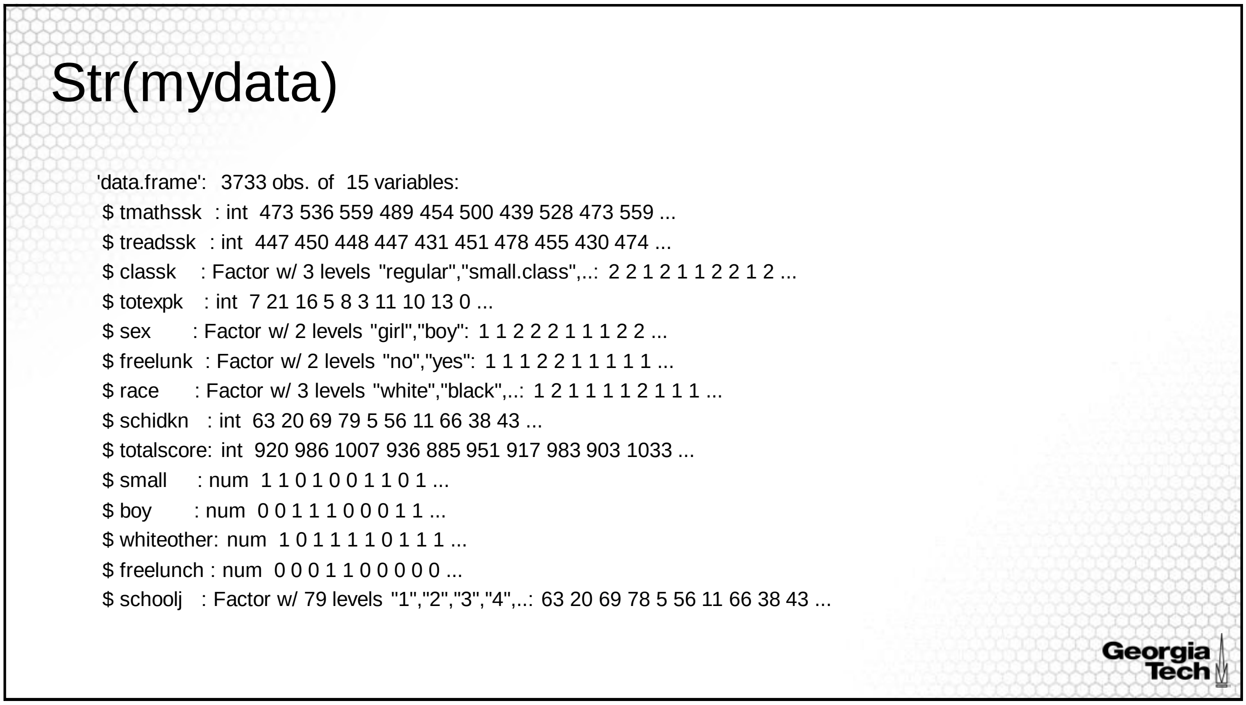
- A cohort of students in Tennessee was randomly assigned within each school to either small classes, regular-sized classes, or regular-sized classes with a paid aide
- The teachers were also randomly assigned to one of these three groups
- We will focus on small vs. regular class size
- The data is present in the Ecdat::star dataset. I create a dataframe *mydata* to focus on the small and regular size classes
- mydata <- dplyr::filter(Ecdat::Star, classk=="small.class"|classk=="regular")
- Define totalscore = tmathssk + treadssk (total of math + reading scores)

Class-size reduction. (2018, March 28). Retrieved April 16, 2018, from https://en.wikipedia.org/wiki/Class-size_reduction#Project_STAR_and_Project_SAGE



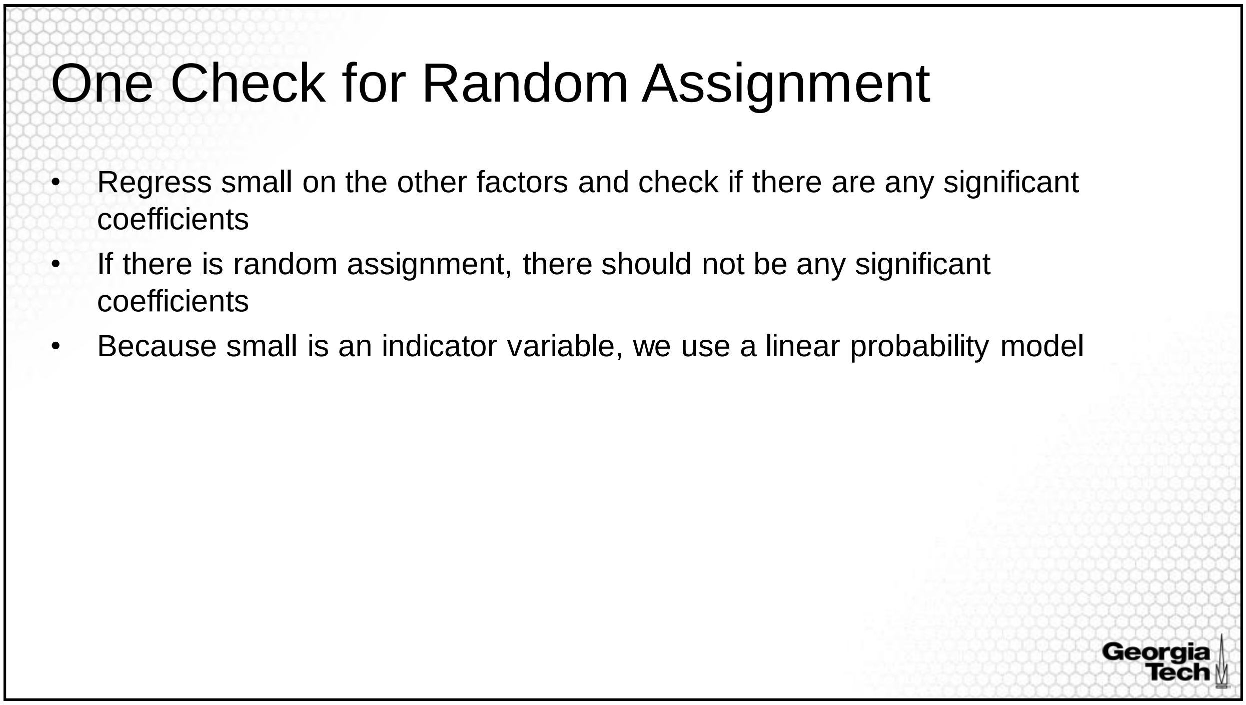
Create Indicator Variables

- small = $\begin{cases} 1, & if\ classk = small.class \\ 0, & otherwise \end{cases}$
- boy = $\begin{cases} 1, & if\ sex = boy \\ 0, & otherwise \end{cases}$
- whiteother = $\begin{cases} 1, & if\ race = white\ or\ race = other \\ 0, & otherwise \end{cases}$
- freelunch = $\begin{cases} 1, & if\ freelunk = yes \\ 0, & otherwise \end{cases}$



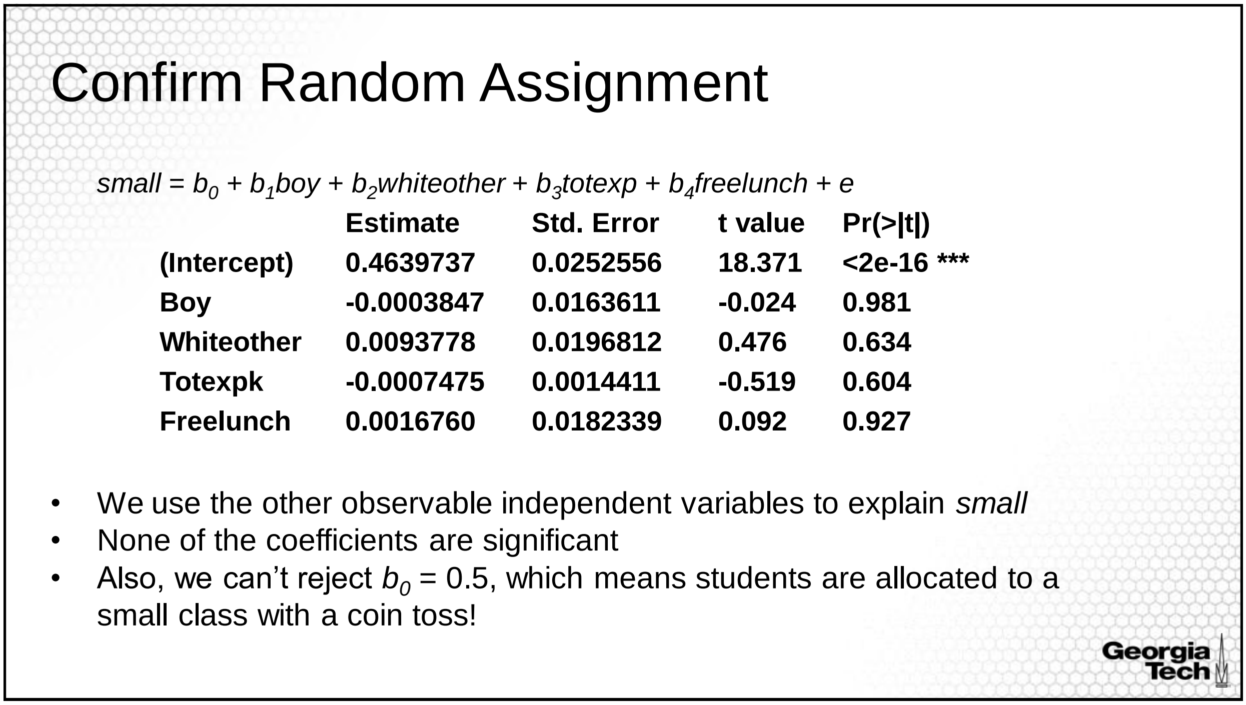
Str(mydata)

```
'data.frame':   3733 obs. of  15 variables:
 $ tmathssk  : int  473 536 559 489 454 500 439 528 473 559 ...
 $ treadssk  : int  447 450 448 447 431 451 478 455 430 474 ...
 $ classk    : Factor w/ 3 levels "regular","small.class",..: 2 2 1 2 1 1 2 2 1 2 ...
 $ totexpk   : int  7 21 16 5 8 3 11 10 13 0 ...
 $ sex       : Factor w/ 2 levels "girl","boy": 1 1 2 2 2 1 1 1 2 2 ...
 $ freelunk  : Factor w/ 2 levels "no","yes": 1 1 1 2 2 1 1 1 1 1 ...
 $ race      : Factor w/ 3 levels "white","black",..: 1 2 1 1 1 1 2 1 1 1 ...
 $ schidkn   : int  63 20 69 79 5 56 11 66 38 43 ...
 $ totalscore: int  920 986 1007 936 885 951 917 983 903 1033 ...
 $ small     : num  1 1 0 1 0 0 1 1 0 1 ...
 $ boy       : num  0 0 1 1 1 0 0 0 1 1 ...
 $ whiteother: num  1 0 1 1 1 1 0 1 1 1 ...
 $ freelunch : num  0 0 0 1 1 0 0 0 0 0 ...
 $ schoolj   : Factor w/ 79 levels "1","2","3","4",..: 63 20 69 78 5 56 11 66 38 43 ...
```



One Check for Random Assignment

- Regress small on the other factors and check if there are any significant coefficients
- If there is random assignment, there should not be any significant coefficients
- Because small is an indicator variable, we use a linear probability model

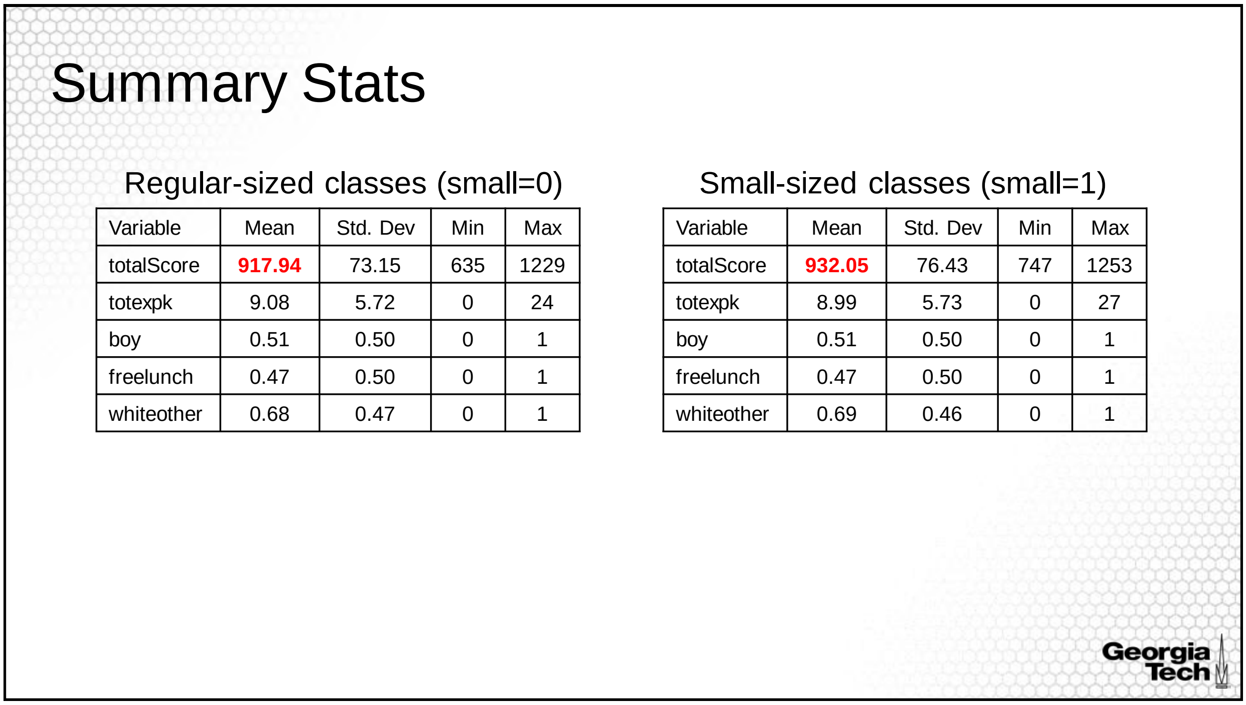


Confirm Random Assignment

$small = b_0 + b_1 boy + b_2 whiteother + b_3 totexp + b_4 freelunch + e$

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	**0.4639737**	**0.0252556**	**18.371**	**<2e-16 *****
Boy	**-0.0003847**	**0.0163611**	**-0.024**	**0.981**
Whiteother	**0.0093778**	**0.0196812**	**0.476**	**0.634**
Totexpk	**-0.0007475**	**0.0014411**	**-0.519**	**0.604**
Freelunch	**0.0016760**	**0.0182339**	**0.092**	**0.927**

- We use the other observable independent variables to explain *small*
- None of the coefficients are significant
- Also, we can’t reject $b_0 = 0.5$, which means students are allocated to a small class with a coin toss!



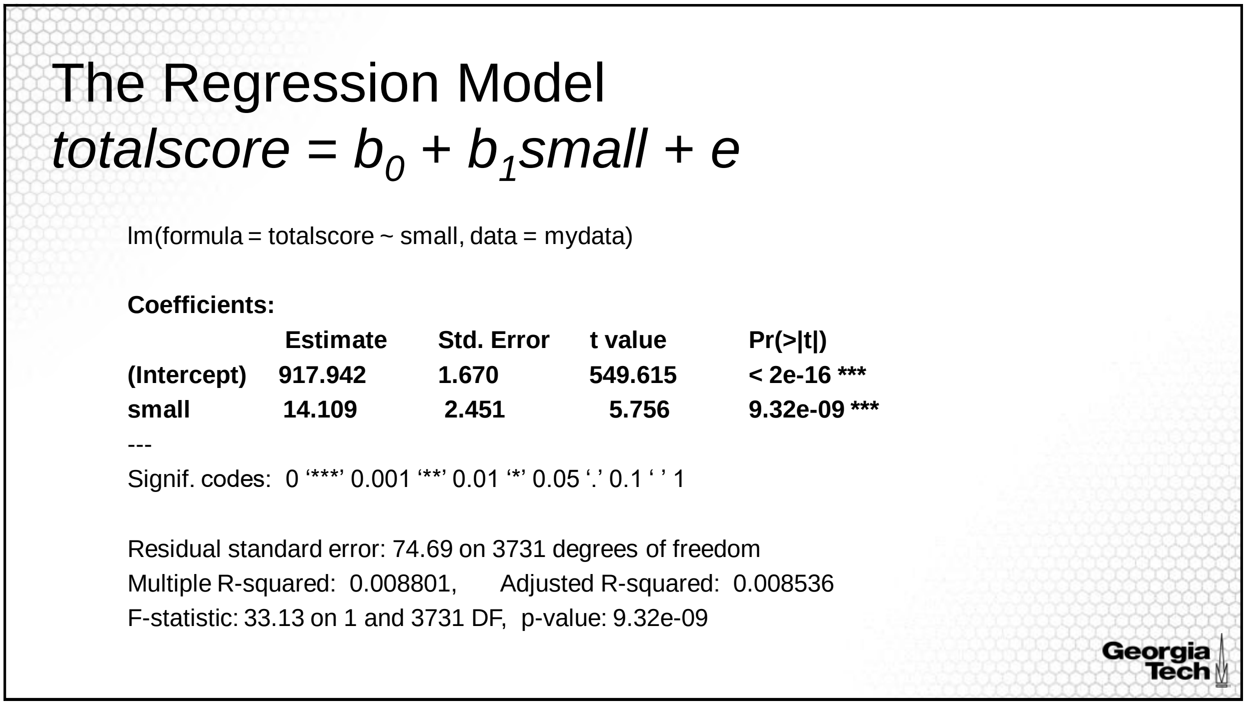
Summary Stats

Regular-sized classes (small=0)

Variable	Mean	Std. Dev	Min	Max
totalScore	917.94	73.15	635	1229
totexpk	9.08	5.72	0	24
boy	0.51	0.50	0	1
freelunch	0.47	0.50	0	1
whiteother	0.68	0.47	0	1

Small-sized classes (small=1)

Variable	Mean	Std. Dev	Min	Max
totalScore	932.05	76.43	747	1253
totexpk	8.99	5.73	0	27
boy	0.51	0.50	0	1
freelunch	0.47	0.50	0	1
whiteother	0.69	0.46	0	1

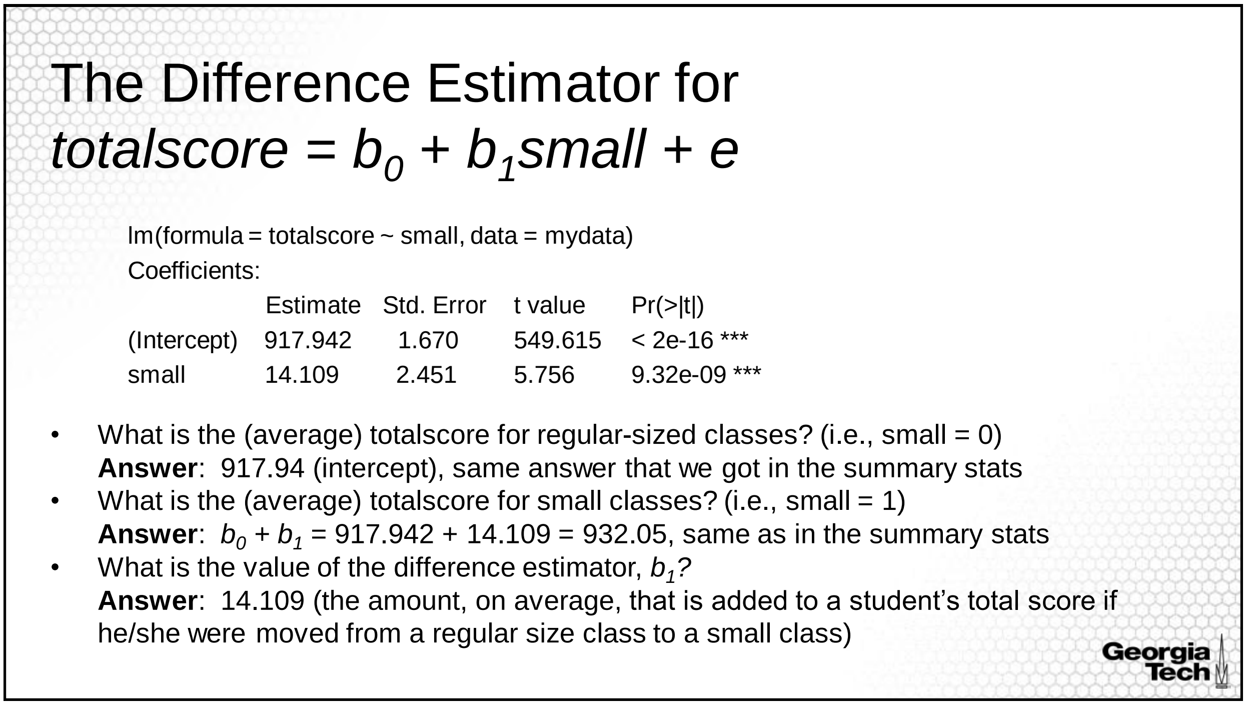


The Regression Model *totalscore* = $b_0 + b_1$*small* + *e*

```
lm(formula = totalscore ~ small, data = mydata)

Coefficients:
             Estimate    Std. Error    t value      Pr(>|t|)
(Intercept)  917.942     1.670         549.615      < 2e-16 ***
small         14.109      2.451          5.756      9.32e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.69 on 3731 degrees of freedom
Multiple R-squared:  0.008801,     Adjusted R-squared:  0.008536
F-statistic: 33.13 on 1 and 3731 DF,  p-value: 9.32e-09
```



The Difference Estimator for *totalscore* = $b_0 + b_1$*small* + *e*

```
lm(formula = totalscore ~ small, data = mydata)
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  917.942     1.670     549.615   < 2e-16 ***
small        14.109      2.451     5.756     9.32e-09 ***
```

- What is the (average) totalscore for regular-sized classes? (i.e., small = 0)
  Answer: 917.94 (intercept), same answer that we got in the summary stats
- What is the (average) totalscore for small classes? (i.e., small = 1)
  Answer: $b_0 + b_1$ = 917.942 + 14.109 = 932.05, same as in the summary stats
- What is the value of the difference estimator, b_1?
  Answer: 14.109 (the amount, on average, that is added to a student’s total score if he/she were moved from a regular size class to a small class)

Add Teacher Experience to the Model: $totalscore = b_0 + b_1 small + b_2 totexpk + e$

```
lm(formula = totalscore ~ small + totexpk, data = mydata)
                 Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)      907.4293   2.5487       356.034     < 2e-16 ***
small             14.2098   2.4420         5.819     6.42e-09 ***
totexpk            1.1580   0.2127         5.445     5.52e-08 ***
```

- Each additional year of a teacher's experience (on average) adds 1.16 points to the total score
- The difference estimator in this model = b_1 = 14.21
- The effect of small class size is the same as having a teacher with $\frac{14.21}{1.16} \approx$ 12 additional years of teaching experience!

Data Analytics in Business

Treatment Effect

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

Natural Experiments and Difference-in-Difference Estimator

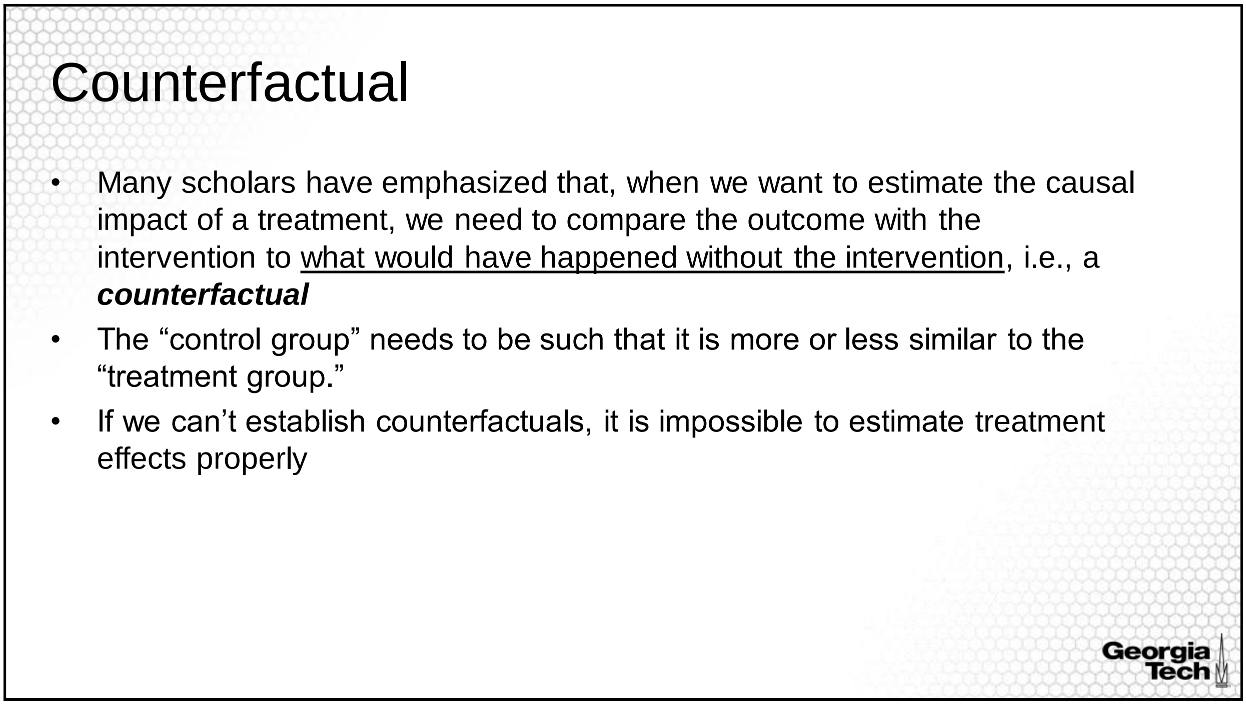
Natural Experiment

- Observational study data from real-world conditions
- Used to approximate what would happen in a Randomized Controlled Experiment
- In a natural experiment, the subjects who might be undergoing treatment are not able to choose if they are in the treatment or control group. This choice is made by an external agent or factor (weather event, policy change, etc.)
- We need to have some subjects in a group that are treated and others who are not treated in order for this natural experiment to work
- Researchers compare the average change over time of the Y variable for the treatment group to the average change over time of the Y variable for the control group. This comparison is called ***difference-in-difference***
- Panel data is used to measure these differences

Examples of Natural Experiments

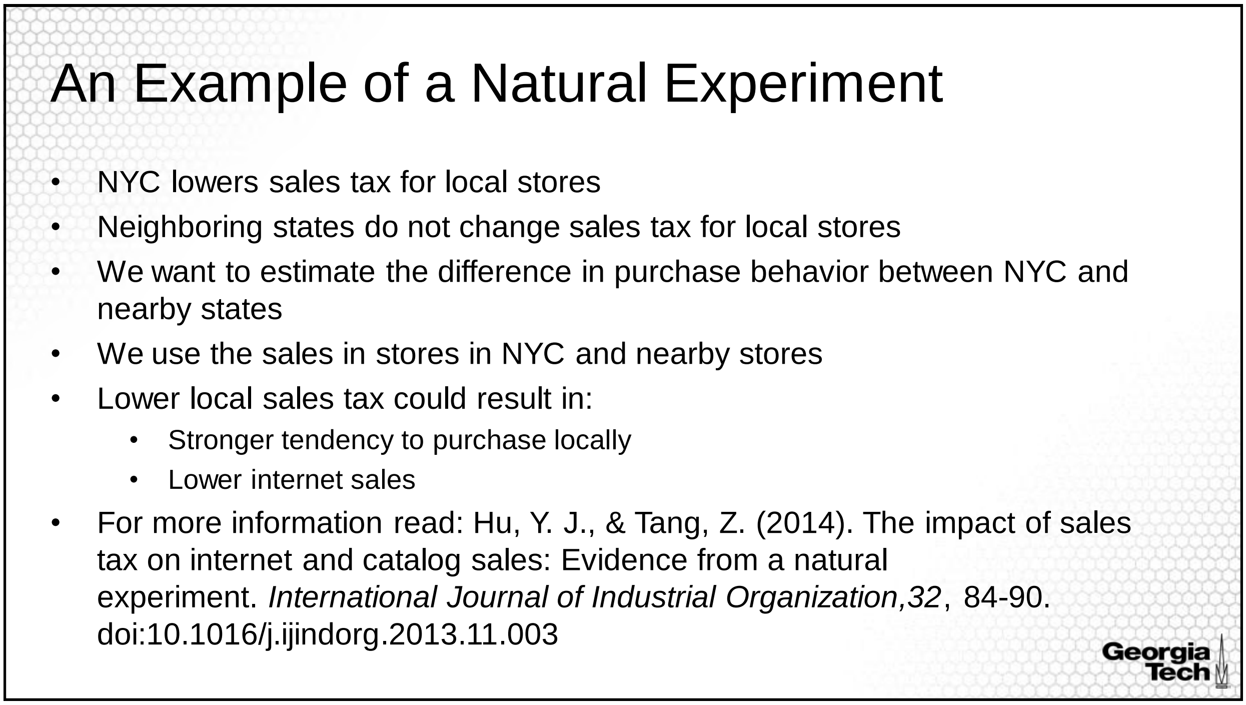
A treatment (manipulation/event) that just happened; not intentionally designed as an experiment:

- A law that changed the tax rate for some subjects, but not others
- Installing an IT-system that allows online orders to be picked in some local stores, but not others
- A hurricane that hits a few stores among a large sample of stores
- A mobile carrier implements an unlimited data plan in some cities but not others
- Minimum wage is changed in one state but not another
- State Inclusionary Zoning laws are enacted in some cities but not in others



Counterfactual

- Many scholars have emphasized that, when we want to estimate the causal impact of a treatment, we need to compare the outcome with the intervention to what would have happened without the intervention, i.e., a ***counterfactual***
- The “control group” needs to be such that it is more or less similar to the “treatment group.”
- If we can’t establish counterfactuals, it is impossible to estimate treatment effects properly



An Example of a Natural Experiment

- NYC lowers sales tax for local stores
- Neighboring states do not change sales tax for local stores
- We want to estimate the difference in purchase behavior between NYC and nearby states
- We use the sales in stores in NYC and nearby stores
- Lower local sales tax could result in:
  - Stronger tendency to purchase locally
  - Lower internet sales
- For more information read: Hu, Y. J., & Tang, Z. (2014). The impact of sales tax on internet and catalog sales: Evidence from a natural experiment. *International Journal of Industrial Organization,32*, 84-90. doi:10.1016/j.ijindorg.2013.11.003

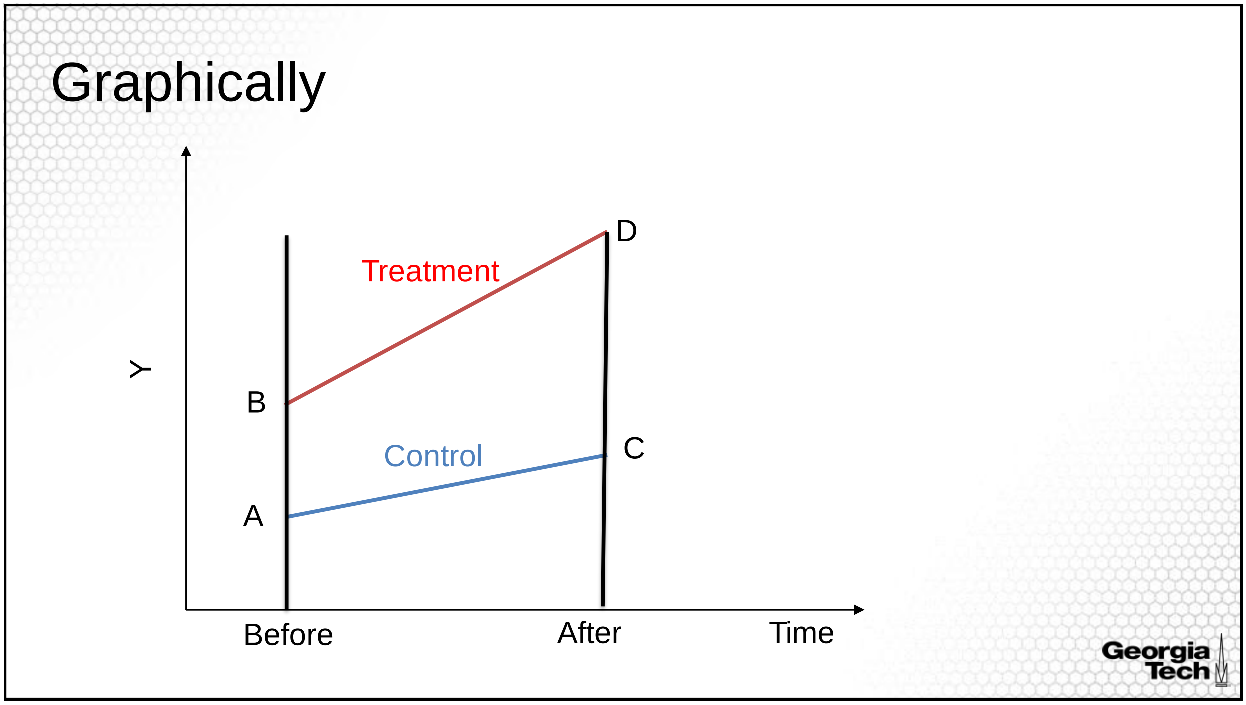
Difference-in-Difference (D-in-D)

- Consider time t_1 and t_2, where t_1 occurs before the “treatment” and t_2 occurs after the “treatment”
- Let’s call t_1 “Before” and t_2 “After”
- We are measuring the average value of the dependent variable (Y)
  - Let A be the average value of the dependent variable for the control group measured at time t_1
  - Let B be the average value of the dependent variable for the treatment group measured at time t_1
  - Let C be the average value of the dependent variable for the control group measured at time t_2
  - Let D be the average value of the dependent variable for the treatment group measured at time t_2

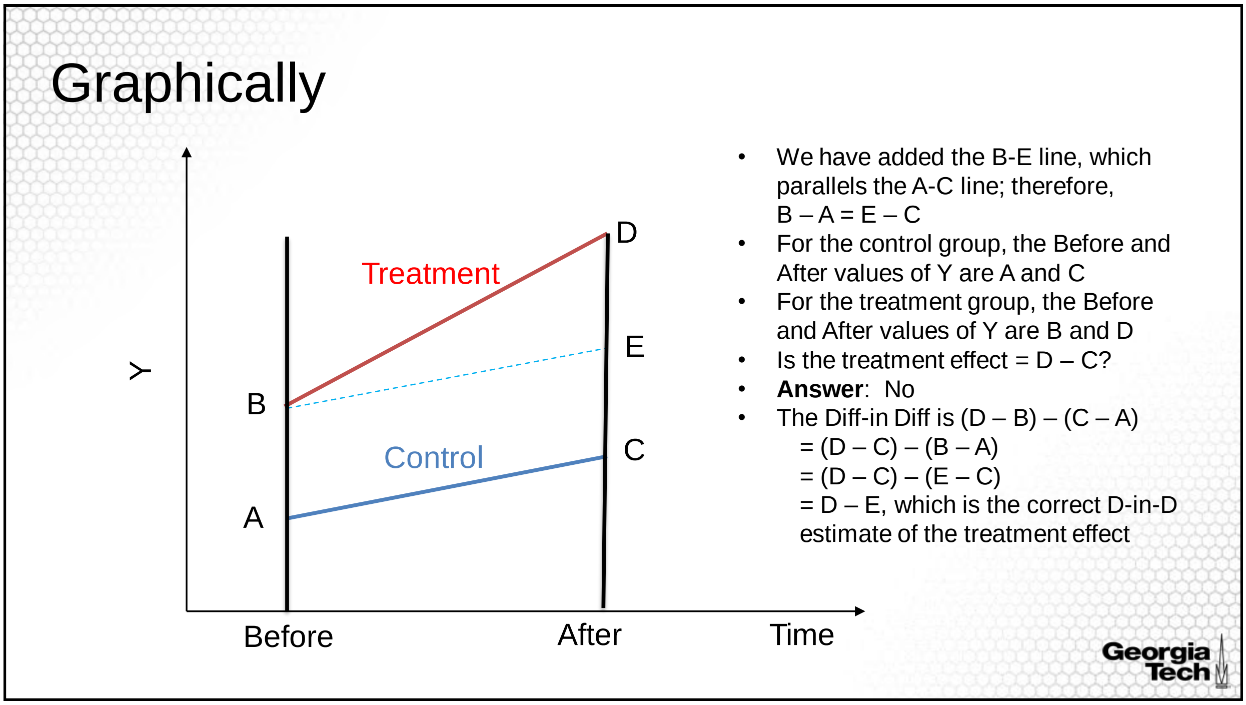
Difference-in-Difference Calculation

	Before	After	Difference
Control	A	C	C – A
Treated	B	D	D – B

- For the **control group**, the difference of the average Y values at time t_2 (After) and time t_1 (Before) = C – A
- For the **treatment group**, the difference of the average Y values at time t_2 (After) and time t_1 (Before) = D – B
- The difference between these values is called ***difference-in-difference*** (diff-in-diff)
- **Diff-in-Diff = (D – B) – (C – A)**



Graphically



Graphically

- We have added the B-E line, which parallels the A-C line; therefore, B – A = E – C
- For the control group, the Before and After values of Y are A and C
- For the treatment group, the Before and After values of Y are B and D
- Is the treatment effect = D – C?
- **Answer**: No
- The Diff-in Diff is (D – B) – (C – A)
  = (D – C) – (B – A)
  = (D – C) – (E – C)
  = D – E, which is the correct D-in-D estimate of the treatment effect

How Do We Estimate the D-in-D Using Regression?

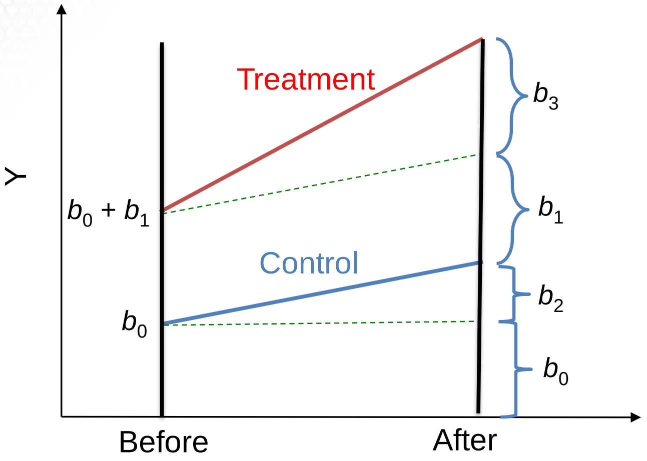
- For the NYC example, create two dummy variables:

$$\text{NYC} = \begin{cases} 1, & if\ observation\ is\ from\ store\ in\ NYC \\ 0, & otherwise \end{cases}$$

$$\text{After} = \begin{cases} 1, & if\ observation\ is\ in\ the\ After\ time\ period \\ 0, & otherwise \end{cases}$$

- Define NYCAfter = NYC*After, which is an interaction variable
- We observe sales at stores in NYC (treatment) and sales at stores in surrounding areas (control) before and after NYC lowered its sales tax
- The model is: **sales = b_0 + b_1 NYC + b_2After + b_3NYCAfter**
- We can show these coefficients graphically

Interpreting the Regression Model
Sales = b_0 + b_1 NYC + b_2After + b_3NYCAfter



- Sales for the control group at time Before = b_0 since After = 0 and NYC = 0
- Sales for the control group at time After = b_0 + b_2 since After = 1 and NYC = 0
- Sales for the treatment group at time Before = b_0 + b_1 since After = 0 and NYC = 1
- Sales for the treatment group at time After = b_0 + b_1 + b_2 + b_3 since After = 1 and NYC = 1

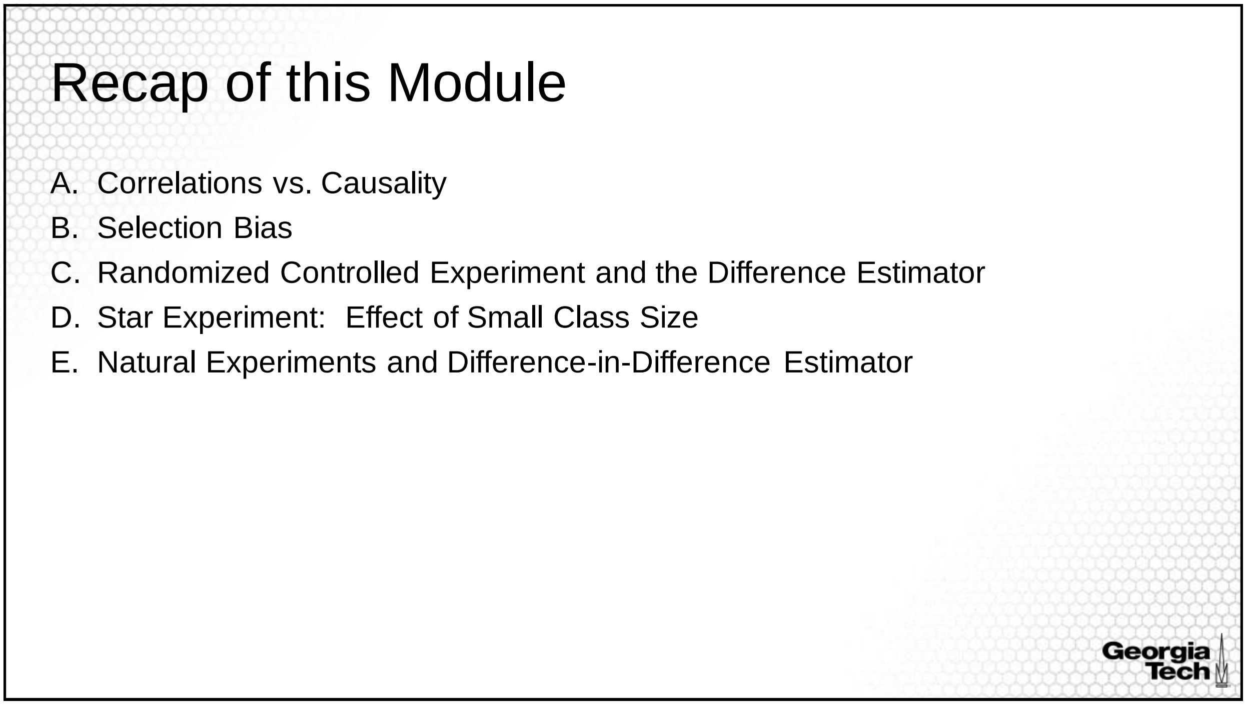
Sales = b_0 + b_1 NYC + b_2After + b_3NYCAfter

	Before	After	Difference (Before – After)
Control	b_0	$b_0 + b_2$	b_2
Treated	$b_0 + b_1$	$b_0 + b_1 + b_2 + b_3$	$b_2 + b_3$

- The diff-in-diff estimator
  = difference of the two differences, and is
  = $b_2 + b_3 - b_2 = b_3$
- b_3 is the coefficient of the interaction term, NYCAfter

Steps in Natural Experiment

1. Understand the treatment (manipulation/event) that just happened
2. Check if we can theoretically argue this treatment appears as if it were randomly assigned (i.e., assignment orthogonal to unobservable factors, X orthogonal to ε)
3. Check if there is a control group and a treatment group
4. Check if the empirical evidence shows that these two groups are roughly the same before the experiment
5. Analyze the treatment effect using the difference-in-difference estimator



Recap of this Module

A. Correlations vs. Causality
B. Selection Bias
C. Randomized Controlled Experiment and the Difference Estimator
D. Star Experiment: Effect of Small Class Size
E. Natural Experiments and Difference-in-Difference Estimator