Module 5: Treatment Effects

Correlations vs. Causality

I. Correlation
   a. Any broad class of statistical relationships involving 2 variables
   b. Measure of linear relationship between X and Y
   c. Always lies between -1 and 1

- The (sample) correlation between two variables X and Y is defined as:

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

   d. If Y = X2 and X between -10 and 10 the correlation between X and Y is 0; they are uncorrelated even though they are perfectly related

II. Strong Correlation
   a. If A and B strongly correlated there could be several possible relationships:
      i. A causes B
      ii. B causes A

III. Reverse causality
   a. Thinking A causes B but B actually causes A

IV. Post Hoc Ergo Propter Hoc
   a. Translates to “after this, therefore because of this”
   b. Faulty logic
   c. i.e. if A happened and then B happened so A must have caused B to happen

V. Causation
   a. Change in cause must lead to change in effect
   b. Hypothesized cause must precede its anticipated effect
   c. Must discount all other plausible explanations, other than the one proposed, that can explain relationship

VI. Causal models used to build theories which tell you how things work

Selection Bias

I. Selection Bias occurs when individuals selected for treatment w/o proper randomization

II. Selection Bias can occur due to several reasons:
   a. Self-selection Bias: participants allowed to opt in
   b. Voluntary response Bias: sample over represents people interested in the topic i.e. people calling into radio show to discuss topic they are already interested in
   c. Nonresponse Bias: often occurs when survey response rate is really low

III. Assumptions when estimating OLS slope coefficient by regressing Y on X to find b1, slope, plus covariance of the error term
   a. Orthogonality Assumption: Cov(e,X) = Cov(X,e) = 0; the error terms and predictors are not related; when X and error are uncorrelated the OLS estimator is a good estimate of b1, slope
   b. Treatment Effect: b1 is treatment effect when

- $Y = b_0 + b_1X + e$
- 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,

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

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

IV. Controlling Selection Bias
   a. Random assignment of test subjects into treatment and control groups
      i. Random assignment has no significant coefficients
   b. Use natural experiment
   c. Add control variables

<u>Randomized Controlled Experiment and Difference Estimator</u>

I. Set-up Randomized controlled experiment by drawing random number for each observation
   a. Value < 0.5 goes to control group (placebo) others get treatment
   b. Set each dummy variable to 0, control, or 1, test group
   c. Regression model where Y is function of d

The Regression Model

- Define indicator variable d as:

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

- The regression model is:

$y_i = b_0 + b_1d_i + e_i$, $i = 1, \ldots, N$ (where i is one of the N individuals in the study)

- The regression functions are:

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

II. Difference estimator
   a. Used to calculate treatment effect (b1/slope)

b. bOLS is difference estimator because it is difference between sample means of treatment and control groups
c. y1 bar is average value of y for observations in treated group
d. y0 bar is average value of y for observations in control group
e. N1 and N2 defined similarly

- 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:

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

f. Difference estimator can be rewritten as

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

g. Using random assignment of individuals through treatment and control groups gives no systemic difference between the 2 groups except 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

Natural Experiments and Difference in Difference Estimator

I. Natural Experiments are not intentional randomized control experiments
   a. Studies from real-world conditions used to approximate what would happen in Randomized Controlled Experiment
   b. Subjects can't choose what group they are in (control or test)
      i. Choice made by external agent like weather, policy changes etc.
      ii. Compare average change in Y over time in test and control groups (difference-in-difference) and panel data used to measure 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

II. Difference-in-Difference estimator gets the treatment effects

Difference-in-Difference Calculation

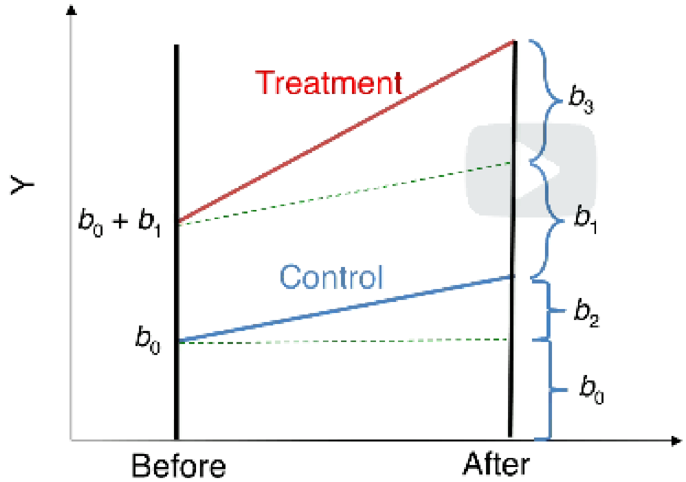
	Before	After	Difference
Control	A	C	C – A
Treated	B	D	D – B

- 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)**

Georgia

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

Georgia

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

III. Counterfactual: comparison of outcome with the intervention to the outcome w/o the intervention
   a. Can't estimate treatment effects properly w/o them

IV. Control group needs to be more or less similar to treatment group