Lecture 6

Administrative Miscellanea

- No homework this week
- Quiz next Wednesday
- Problem set 4 posted along with the paper for pset 3
- Will have a second grade snapshot by end of the week
- Introduction to causal inference and quasiexperimental designs

Introduction to Potential Outcomes

- We have already motivated our goal in this class as finding the causal effect of various relationships using regression analysis
- We discussed that we can interpret a causal effect if we have randomization, either explicitly by using a controlled trial or by using quasi-experimental designs
- Before continuing to quasi-experimental designs, we will discuss more about causality and experimental designs
- We will use the Rubins causal model and potential outcomes framework to motivate this

Rubins Causal Model

- Let Y_{i1} denote the outcomes for individual i under treatment, and Y_{i0} denote outcomes for individual i who are not treated
- The causal effect of treatment on individual i is $Y_{i1} - Y_{i0}$
  - e.g. if Jeff does not take blood pressure medication his systolic blood pressure is 140 mm Hg, while if he does it's 125 mm Hg
  - The causal effect of the blood pressure medication is a 15 mm Hg decline for Jeff
- Note that we never observe $Y_{i1} - Y_{i0}$.

Rubins Causal Model Example Setup

- Consider the following setup:
- An elementary school initially has an after-school program that focuses on developing kids' reading skills. Due to budget cuts, the program is dropped. You wish to study the effect of removing this program on childrens' outcomes as measured by their standardized reading test scores

Rubins Causal Model Example Setup

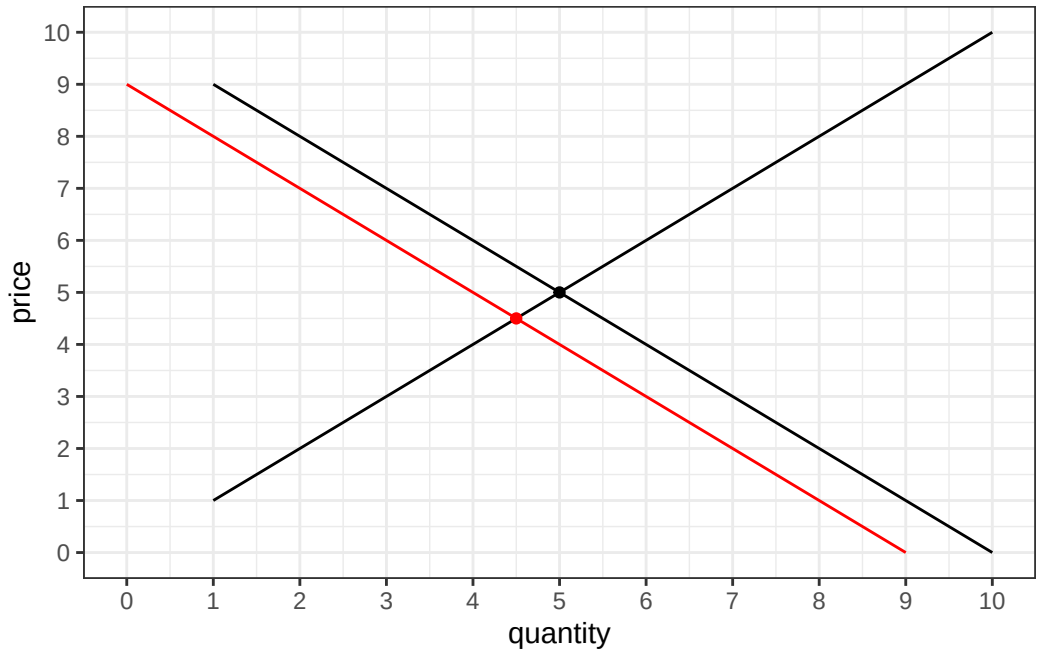
- In response to the school dropping the program, parents decide to spend more time at home reading to their kids. As a result, there is no change in test scores after the program change. If the parents did not spend more time at home, test scores would have dropped by 5 points. What is the casual effect of dropping the program on test scores?

Negative Feedback Cycles

- While the previous example seems extreme, it is actually fairly common due to negative feedback systems
- Consider the following question: does consuming salt (or sodium in general) raise blood pressure?
  - Sodium changes water concentrations which mechanically raises blood pressure. This is well established
  - But the kidneys are very good at regulation of sodium in the body
  - Meta-analyses of healthy individuals finds no relationship between sodium intake and blood pressure!

Negative Feedback Cycles: supply and demand

- Supply decreases by 2 units, but output and price only fall by 1 since they are in a feedback cycle



Positive Feedback Cycles Are Rare

- Consider the scenario where a student does poorly on their SAT. They want to know the effect of a low SAT score on future income.
- A common thought process would be that the low SAT score means that they miss out on going to a good school
  - This means they get a worse job out of college
  - They then find it more difficult to get promotions
  - By age 45 the earnings difference is massive
- (I call this a "college essay analysis")

Positive Feedback Cycles Are Rare

- In reality, the difference is actually tiny. Being a top student in a mediocre college gives many opportunities which make it comparable to being a poor student at a top college
- Often when difficult choices are made, it's because the alternatives tend to be similar, so we expect the effects to be small
  - Or it's because there are tradeoffs. A harder college means more time spent studying that could have been spent on finding an internship

Potential Outcomes Framework

- Treatment is a random variable D_i for individual i
  - $D_i = 1$ means that individual i was treated, and $D_i = 0$ means untreated
- Y_{1i} is the outcome variable for individual i if they were treated
- Y_{0i} is the outcome variable for individual i if they were not treated
- We can decompose this: $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$
  - Interpretation?

Potential Outcomes Framework

- We can't compare Y_{1i} to Y_{0i}. What if we just compare Y_i for individuals with $D_i = 1$ vs $D_i = 0$?
  - $E[Y_i|D_i = 1] - E[Y_i|D_i = 0] =$
  - $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] =$
  - $(E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1])+$
  - $(E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0])$
- Interpretation?

Counterfactuals

- In the potential outcomes framework, we compare Y_{i1} to Y_{i0}
  - Y_{i0} is the counterfactual for Y_{i1}. It's what would have happened in the absence of treatment
- In a randomized trial, what is the counterfactual for the treatment group $\{D_i = 1\}$?

Counterfactuals In regression

- We want to know the effect of having a bachelor's degree on wages
- Run the regression $wage_i = \beta_0 + \beta_1 bachelors_i + \varepsilon_i$ where bachelors is 1 if individual i has a bachelor's degree
- Jeff is a 27 year old with a bachelor's degree from Harvard. What is his counterfactual?

Counterfactuals In regression

- We want to know the effect of having a bachelor's degree on wages
- Run the regression $wage_i = \beta_0 + \beta_1 bachelors_i + age_i + \varepsilon_i$ where bachelors is 1 if individual i has a bachelor's degree
- Jeff is a 27 year old with a bachelor's degree from Harvard. What is his counterfactual?

Counterfactuals In regression

- We want to know the effect of education on wages
- Run the regression $wage_i = \beta_0 + \beta_1 education_i + \varepsilon_i$ where education is the number of years of education obtained for individual i
- Jeff is a 27 year old with a bachelor's degree (education=16) from Harvard. What is his counterfactual?

Non-causal example

- Suppose we want to study the effect of twitter sentiment of a company on stock returns. We don't have a way to isolate sentiment in a causal framework, so we just want to know whether twitter stocks can predict earnings so that we can use it to make money. What controls should we include?
- A None
- B Only controls that we currently use to predict firm returns
- C Controls that predict both firm performance and company sentiment
- D Controls that predict firm returns and time fixed effects
- E Time fixed effects and firm fixed effects

Do High Temperatures Increase Violent Murders in Mexico?

- Suppose we have $homicide_{nmst} = \beta_0 + \beta_1 temp_{nmst} + \varepsilon_{nmst}$
- Where n is municipality, m is month, s is state, and t is year
- What concerns would you have over this specification?
  - What if we control for precipitation?

Do High Temperatures Increase Violent Murders in Mexico?

- What if we now do $homicide_{nmst} + \beta_0 + \beta_1 temp_{nmst} +$
- $\delta precip_{nmst} + \xi_m + \lambda_t + \zeta_n + \varepsilon_{nmst}$?
- What comparison is being made?
- What additional data would you like to see to confirm this is a real effect?

Two Way Fixed Effects: Productivity in Ghana

- In Ghana, some tribes pass land on through their sons (patrilineal) while others do so through their daughters (matrilineal)
- Farming is low productivity and requires low education compared to other potential occupations
- Do families decrease human capital investment (education) for women who will inherit land?

Productivity in Ghana: Naive Estimation Equation

- Consider $education_i = \beta_0 + \beta female_i +$
- $\gamma matrilineal_i + \delta female_i * matrilineal_i + \varepsilon_i$
- What is our comparison being made? What would bias this?
- What if we had multiple years of data (survey waves) and information on geography (district)?

Table 1.4: Human Capital, by Gender, Descent

	(1)
	Years, Education
Female	-2.108***
	(0.1530)
Matrilineal Tribe	1.911***
	(0.2470)
Female * Matrilineal	-0.378*
	(0.2190)
Sample Mean	5.796
Patrilineal Male Mean	5.863
Observations	6,137
F-Test: Female * Matri=0	2.96
p-value	0.085
F-Test: Matri + Fem*Matri=0	41.70
p-value	0.000
F-Test: Female + Fem*Matri=0	247.80
p-value	0.000

This table compares years of education by gender and descent, among individuals who ever appear as a prime-age adult (aged 25-54), born in a rural districts, in the three waves of the panel. Regressions include district of birth fixed effects (and take the mean in cases where an individual's education is reported differently across waves)

- Any confounding variables still present?

Designing an Experiment from a research design

- We want to study the effect of babies exposure to excessive prenatal heat on future income as adults.
- We have administrative data that has the region of birth of the child (at the zip-5 level) along with the exact date of birth and future earnings
- We also have data on the parents income records. We don't observe their geographic location.
- We also have information on daily temperature by day at the zip5 level

Designing an Experiment from a research design

- What regression should we run to extract a causal effect?

The Basic bivariate regression

- $salary_i = \beta_0 + \beta_1 temperature_i + \varepsilon_i$
  - Any issues with this?
  - What are the counterfactuals for babies born in high temperature areas?
- Ignoring regression, what is the ideal comparison we should be making?
  - Given our data, what is the best comparison we can make?

Some control options

- Saturated Model: $salary_{ismt} = \beta_0 + \beta_1 temperature_{ismt} + \mu_s + \lambda_t + \delta_m +$
- $\nu_{st} + \eta_{sm} + \theta_{mt} + \rho_{smt} + \varepsilon_{ismt}$
- If we run the above equation, what variation is left?
- What if we run $salary_{ismt} = \beta_0 + \beta_1 temperature_{ismt} + \psi_i + \varepsilon_{ismt}$?

What comparison is being made?

- $salary_{ismt} = \beta_0 + \beta_1 temperature_{ismt} + \mu_s + \lambda_t + \nu_{st} + \varepsilon_{ismt}$
- A For babies born in July 1990, the difference between Texas and Wisconsin
- B For babies born in July in Texas and Wisconsin, the difference between 1990 and 1991
- C For babies born in Texas and Wisconsin in 1990, the difference between January and July
- D Babies born in July vs January in the same location and year

- E A, B, and C

What comparison is being made?

- $salary_{ismt} = \beta_0 + \beta_1 temperature_{ismt} + \mu_s + \lambda_t + \delta_m + \nu_{st} + \varepsilon_{ismt}$
- A For babies born in July 1990, the difference between Texas and Wisconsin
- B For babies born in July in Texas and Wisconsin, the difference between 1990 and 1991
- C For babies born in Texas and Wisconsin in 1990, the difference between January and July
- D Babies born in July vs January in the same location and year
- E A, B, and C

What comparison is being made?

- $salary_{ismt} = \beta_0 + \beta_1 temperature_{ismt} + \mu_s + \lambda_t + \delta_m + \theta_{mt} + \varepsilon_{ismt}$
- A For babies born in July 1990, the difference between Texas and Wisconsin
- B For babies born in July in Texas and Wisconsin, the difference between 1990 and 1991
- C For babies born in Texas and Wisconsin in 1990, the difference between January and July
- D Babies born in July vs January in the same location and year
- E A, B, and C

What comparison is being made?

- $salary_{ismt} = \beta_0 + \beta_1 temperature_{ismt} + \mu_s + \lambda_t + \delta_m + \varepsilon_{ismt}$
- A For babies born in July 1990, the difference between Texas and Wisconsin
- B For babies born in July in Texas and Wisconsin, the difference between 1990 and 1991
- C For babies born in Texas and Wisconsin in 1990, the difference between January and July
- D Babies born in July vs January in the same location and year
- E A, B, and C