

Topic 1: Introduction to Causal Inference and Potential Outcomes Framework

ECON 5783 – University of Arkansas

Prof. Kyle Butts

Fall 2024

Roadmap

Difficulties of Establishing Causality

Counterfactual Thinking and Potential Outcomes

Randomized Experiments

Estimation via Regression

LaLonde Dataset

A/B Testing

Causal Inference

The goal of the class is to think about establishing *causal* relationships:

- Given some outcome, Y , that we care about, we want to know if changing X causes a change in Y

Causal Inference

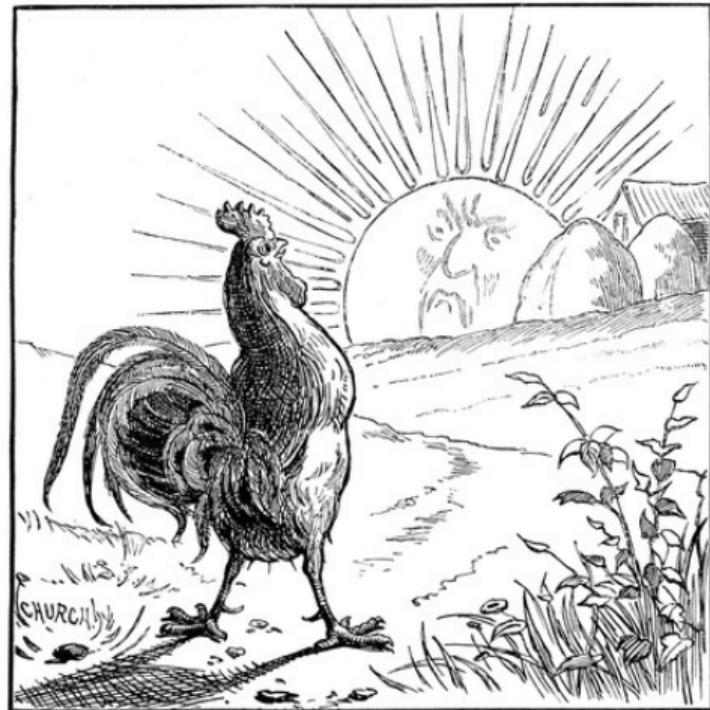
The goal of the class is to think about establishing *causal* relationships:

- Given some outcome, Y , that we care about, we want to know if changing X causes a change in Y

In the real world, we can only observe whether X and Y co-move but trying to know if X caused Y is much more difficult a problem

- We will try to tackle this problem, but be humble of the difficulties of this task

Correlation does not imply Causality



Rooster crowing does not cause sun to rise

- The rooster crows every morning when the sun comes up

Correlation does not imply causation



"By the way, the health benefits of a glass of wine a day are not retroactive."

Drinking wine is correlated with improved health

- Drinking wine is also correlated with income . . .

Causality does not imply correlation

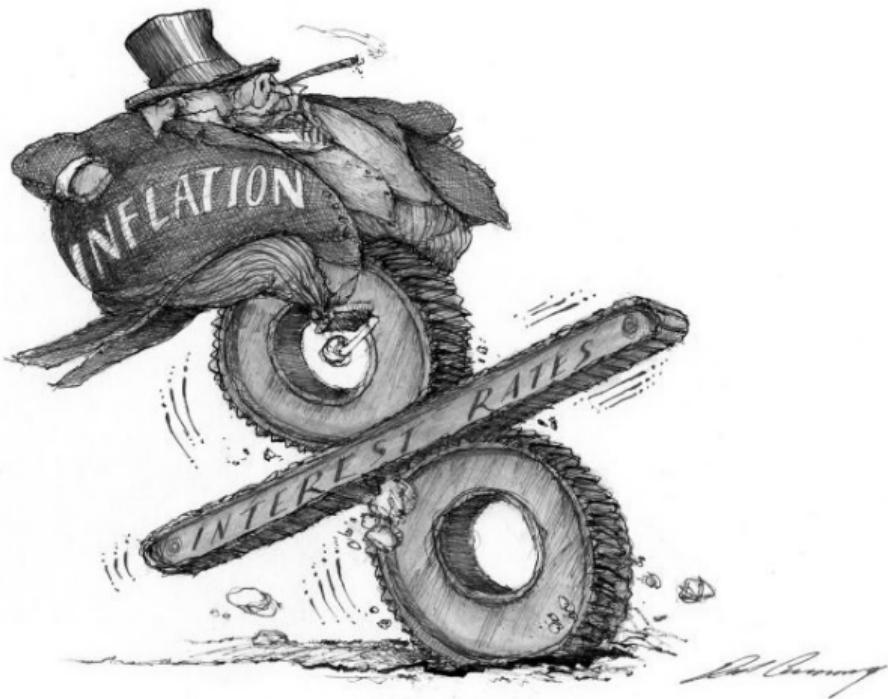


A boat is traveling on a windy day

- The sailor turns the rudder back and forth
- Boat moves in a straight line

No correlation between rudder direction and boat direction; but the rudder still *causes* the boat to change direction

Causality does not imply correlation



The federal reserve changes interest rates back and forth to try and keep inflation around target of 2%

- When things are going well, no correlation between interest rate and inflation rate

Lurking variables

Correlations can often arise due to the presence of some third underlying variable that is driving both. We call this a **lurking variable**

- Housing prices go up in places where they're building new apartments
- Employment goes up after places raise their minimum wage
- Schools with more arts spending have higher test scores
- Correlates of food consumption and health

Lurking variables

In all of these examples, there seems to be other explanations **lurking around the corner**:

- New apartments are built in neighborhoods with growing demand
- The minimum wage is passed in places with a growing economy
- Schools with more arts spending are in wealthier school districts
- Food consumption and health are both correlated with income

Reverse Causality

Sometimes we have the order of causation *reversed*:

- People who drink more coffee have more anxiety
- Places with more police officers have more crime
- People who try new medicines have higher mortality rates

We want to think carefully if X causes Y or Y causes X

Roadmap

Difficulties of Establishing Causality

Counterfactual Thinking and Potential Outcomes

Randomized Experiments

Estimation via Regression

LaLonde Dataset

A/B Testing

What is causality?

The goal of **causality** is to try and figure out how if I were to change some variable X at *some point in time*, how would the value of Y change for that person

- We have some notion of 'changing' X in some 'experimental' or 'random' way

What is causality?

The goal of **causality** is to try and figure out how if I were to change some variable X at *some point in time*, how would the value of Y change for that person

- We have some notion of 'changing' X in some 'experimental' or 'random' way

PROBLEM: we can *never* observe both states of the world *at the same point in time*.

Observational Studies

When we look out into the world, people have different X s. Why can we not just compare Y s for people with different values of X ?

- As economists, we know that people are optimizers. They chose their 'optimal' X for different reasons.
- People's background characteristics and environment shape what is the optimal X and these factors likely shape their value of Y too.

Observational Studies

When we look out into the world, people have different X s. Why can we not just compare Y s for people with different values of X ?

- As economists, we know that people are optimizers. They chose their 'optimal' X for different reasons.
- People's background characteristics and environment shape what is the optimal X and these factors likely shape their value of Y too.

So comparing people with different X s often involves comparing people with different lurking variables, Z , too

- Is it X causing the change in Y or is it the multitude of lurking variables?

Notions of Counterfactual

Causal inference is looking for some notion of a **counterfactual** world:

- What would have happened to home prices if the new apartments were not built?
- What would have happened to employment if the minimum wage was not increased?
- What would have happened to a patient had they taken a different medicine or no medicine at all?

To be clear, these counterfactual worlds are **made up**. They do not exist. This notion is straight out of sci-fi parallel universe kind of thinking.

Notions of Counterfactual

The job of causal inference is to *make assumptions* about the counterfactual world to take our best guess at that world

- Bad assumption in \implies bad answer out

Causal effects are a hard thing to identify; researchers do their best. I think the best researchers:

1. Articulate the assumptions clearly
2. Worry deeply about lurking variables (not sweep them under)

Notions of Counterfactual

Trying to determine the counterfactual

At the end of the day, our job will be to find a way to ‘impute’ the missing counterfactual.
Finding a reasonable way to *predict the missing counterfactual* is the hard part

- Look at similar neighborhoods that did not have new apartments built?
 - Pennington (2023) look at San Francisco neighborhoods where a fire burns down a single family home

Notions of Counterfactual

Trying to determine the counterfactual

At the end of the day, our job will be to find a way to ‘impute’ the missing counterfactual.
Finding a reasonable way to *predict the missing counterfactual* is the hard part

- Look at similar neighborhoods that did not have new apartments built?
 - Pennington (2023) look at San Francisco neighborhoods where a fire burns down a single family home
- What would have happened to employment if the minimum wage was not increased?
 - Dube et. al. (2010) use counties on the other side of a state border

Notions of Counterfactual

Trying to determine the counterfactual

At the end of the day, our job will be to find a way to ‘impute’ the missing counterfactual.
Finding a reasonable way to *predict the missing counterfactual* is the hard part

- Look at similar neighborhoods that did not have new apartments built?
 - Pennington (2023) look at San Francisco neighborhoods where a fire burns down a single family home
- What would have happened to employment if the minimum wage was not increased?
 - Dube et. al. (2010) use counties on the other side of a state border
- What would have happened to a patient had they taken a different medicine or no medicine at all?
 - Health care industry runs thousands of randomized control trials every year

Potential Outcome Framework

The **Potential Outcome Framework** was first introduced by Donald Rubin in a series of articles starting in 1974. It is *incredibly* influential and most people doing work in causal inference think and talk in terms of potential outcomes and counterfactual thinking.

- Sometimes called Neyman-Rubin framework because Neyman presented this framework in the context of randomized controlled trials in 1923 in a master's thesis

Potential Outcome Framework

The **Potential Outcome Framework** was first introduced by Donald Rubin in a series of articles starting in 1974. It is *incredibly* influential and most people doing work in causal inference think and talk in terms of potential outcomes and counterfactual thinking.

- Sometimes called Neyman-Rubin framework because Neyman presented this framework in the context of randomized controlled trials in 1923 in a master's thesis

We will first present the original framework and then I will try to zoom out a bit and tell you why I think potential outcomes thinking is so powerful

Potential Outcome Framework

We will use the example of building a new apartment building in a neighborhood. There are two parallel universes, one where the apartment is built and the other where it is not built.

Our outcome variable of interest the average rent in the neighborhood.

- In the present world, we can only prices if the apartment is built *OR* if the apartment was not built. This is the fundamental problem of causal inference (Holland, 1986)

Potential Outcome Framework

There are n neighborhoods and we observe (D_i, Y_i) for each neighborhood. D_i denotes whether a new apartment is built or not; Y_i denotes the average rent in the neighborhood.

Potential Outcome Framework

There are n neighborhoods and we observe (D_i, Y_i) for each neighborhood. D_i denotes whether a new apartment is built or not; Y_i denotes the average rent in the neighborhood.

The potential outcome framework expresses rent Y_i as a *function* of the new apartment building, $Y_i(D_i)$.

- $Y_i(1)$ is the average rent *in the counterfactual world* where the new apartment is built
- $Y_i(0)$ is the average rent *in the counterfactual world* where the new apartment is not built

Potential Outcome Framework

- There are n neighborhoods and we observe (D_i, Y_i) for each neighborhood. D_i denotes whether a new apartment is built or not; Y_i denotes the average rent in the neighborhood.
- The potential outcome framework expresses rent Y_i as a *function* of the new apartment building, $Y_i(D_i)$.
 - $Y_i(1)$ is the average rent *in the counterfactual world* where the new apartment is built
 - $Y_i(0)$ is the average rent *in the counterfactual world* where the new apartment is not built

Unit-level treatment effect

Define the **unit-level treatment effect** to be $\tau_i = Y_i(1) - Y_i(0)$

- Measures the effect of D_i from 0 to 1 on unit i

Note this is for the *same unit at the same point in time!*

Unit-level treatment effect

Define the **unit-level treatment effect** to be $\tau_i = Y_i(1) - Y_i(0)$

- Measures the effect of D_i from 0 to 1 on unit i

Note this is for the *same unit at the same point in time!*

- I am *not* saying a neighborhood before and after the apartment is built (though this could be a way of trying to estimate a causal effect)

What do we observe?

In our dataset, we observe average rent Y_i and whether or not the apartment is built, D_i .

How does this relate to potential outcomes?

The '**switching equation**' relates potential outcomes to observed outcomes:

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$$

What do we observe?

In our dataset, we observe average rent Y_i and whether or not the apartment is built, D_i .

How does this relate to potential outcomes?

The '**switching equation**' relates potential outcomes to observed outcomes:

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$$

When $D_i = 1$, we have $Y_i = Y_i(1)$ and when $D_i = 0$ we have $Y_i = Y_i(0)$. That is, we observe one of the two potential outcomes for each unit.

Missing potential outcomes

Here is the complete dataset

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	1	\$1200	\$1200	\$1225
2	0	\$1000	\$950	\$1000
3	1	\$1150	\$1150	\$1150
⋮				
n	0	\$1100	\$1090	\$1100

Missing potential outcomes

Here is the complete dataset

- However, we cannot actually observe the potential missing outcome!
- The goal is to use the **observed** observations to “fill in” the unobserved missing potential outcome

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	1	\$1200	\$1200	?
2	0	\$1000	?	\$1000
3	1	\$1150	\$1150	?
⋮				
n	0	\$1100	?	\$1100

Causal Estimands

Recall our **unit-level treatment effect** to be $\tau_i = Y_i(1) - Y_i(0)$

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	\$1200	\$1200	\$1225	-\$25
2	0	\$1000	\$950	\$1000	\$50
3	1	\$1150	\$1150	\$1150	\$0
⋮					
n	0	\$1100	\$1090	\$1100	-\$10

Averaging unit-level treatment effects

Recall our **unit-level treatment effect** to be $\tau_i = Y_i(1) - Y_i(0)$

- There is no reasonable way to estimate individual-level treatment effects (too much noise!)

Averaging unit-level treatment effects

Recall our **unit-level treatment effect** to be $\tau_i = Y_i(1) - Y_i(0)$

- There is no reasonable way to estimate individual-level treatment effects (too much noise!), so we will aim to estimate averages of them

Averaging unit-level treatment effects

Recall our **unit-level treatment effect** to be $\tau_i = Y_i(1) - Y_i(0)$

- There is no reasonable way to estimate individual-level treatment effects (too much noise!), so we will aim to estimate averages of them

The **Average Treatment Effect (ATE)** is given by

$$\tau_{\text{ATE}} = \mathbb{E}(\tau_i) = \mathbb{E}(Y_i(1) - Y_i(0))$$

- This averages over every unit in your population (with equal weights), including those who never receive treatment

Causal Estimands

The ATE is useful if you are trying to understand the **average** effect of treatment on the entire population you are sampling from.

- E.g. you are launching a pilot that you intend to scale up to the school level

Causal Estimands

The ATE is useful if you are trying to understand the **average** effect of treatment on the entire population you are sampling from.

- E.g. you are launching a pilot that you intend to scale up to the school level
- Be careful; when scaling up a treatment, general equilibrium can change the impact of treatment!
 - One person getting resume tips is more useful than everyone getting resume tips

Causal Estimands

The ATE is useful if you are trying to understand the **average** effect of treatment on the entire population you are sampling from.

- E.g. you are launching a pilot that you intend to scale up to the school level
- Be careful; when scaling up a treatment, general equilibrium can change the impact of treatment!
 - One person getting resume tips is more useful than everyone getting resume tips

This is less useful if you are trying to estimate the effect on a specific sub-population

- Testing the effects of medicine on outcomes *for the population with a disease*

Causal Estimands

There is the **Average Treatment Effect on the Treated** (ATT) that averages over only units that receive treatment:

$$\tau_{\text{ATT}} = \mathbb{E}(\tau_i \mid D_i = 1) = \mathbb{E}(Y_i(1) - Y_i(0) \mid D_i = 1)$$

- In some cases, this will be the only estimate we can credibly identify
- If the treated population looks very different from the population as a whole, you face the risk of **external validity**

Causal Estimands

Likewise the **Average Treatment Effect on the Control (ATC)** averages over only units that do not receive treatment:

$$\tau_{\text{ATC}} = \mathbb{E}(\tau_i \mid D_i = 0) = \mathbb{E}(Y_i(1) - Y_i(0) \mid D_i = 0)$$

Causal Estimands

All three estimands relate via the law of conditional expectations. Define $\pi = \mathbb{P}(D_i = 1)$ be the probability of a unit being in treatment.

$$\tau_{\text{ATE}} = \mathbb{E}(\tau_i) = \mathbb{E}(\mathbb{E}(\tau_i | D_i))$$

Causal Estimands

All three estimands relate via the law of conditional expectations. Define $\pi = \mathbb{P}(D_i = 1)$ be the probability of a unit being in treatment.

$$\begin{aligned}\tau_{\text{ATE}} &= \mathbb{E}(\tau_i) = \mathbb{E}(\mathbb{E}(\tau_i | D_i)) \\ &= \mathbb{P}(D_i = 1) \mathbb{E}(\tau_i | D_i = 1) + \mathbb{P}(D_i = 0) \mathbb{E}(\tau_i | D_i = 0)\end{aligned}$$

Causal Estimands

All three estimands relate via the law of conditional expectations. Define $\pi = \mathbb{P}(D_i = 1)$ be the probability of a unit being in treatment.

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}(\tau_i) = \mathbb{E}(\mathbb{E}(\tau_i | D_i)) \\ &= \mathbb{P}(D_i = 1) \mathbb{E}(\tau_i | D_i = 1) + \mathbb{P}(D_i = 0) \mathbb{E}(\tau_i | D_i = 0) \\ &= \pi\tau_{ATT} + (1 - \pi)\tau_{ATC}\end{aligned}$$

Conditional ATE

While an overall treatment effect is a useful summary measure, we often want to summarize treatment effects for *groups of units with the same characteristics*, e.g. gender, race, income, age, etc.

We define the **Conditional Average Treatment Effect** (CATE) as:

$$\tau_{\text{CATE}}(x) = \mathbb{E}(\tau_i \mid X_i = x) = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)$$

- X_i is some characteristic of the population

Conditional ATE versus ATE

There is a strong relationship between the CATE and the ATE/ATT:

$$\tau_{\text{ATE}} = \mathbb{E}(\tau_i) = \mathbb{E}(\mathbb{E}(\tau_i | X_i))$$

- The latter averages over CATE with weights proportional to the distribution of X_i in the population

Conditional ATE versus ATE

There is a strong relationship between the CATE and the ATE/ATT:

$$\tau_{\text{ATE}} = \mathbb{E}(\tau_i) = \mathbb{E}(\mathbb{E}(\tau_i | X_i))$$

- The latter averages over CATE with weights proportional to the distribution of X_i in the population

Likewise, we could average over the distribution of X_i for the treated units

$$\tau_{\text{ATT}} = \mathbb{E}(\tau_i) = \mathbb{E}(\tau_{\text{CATE}}(x)) = \mathbb{E}(\mathbb{E}(\tau_i | X_i) | D_i = 1)$$

Problems with observational data

Our treatment effect parameter is given by

$$\tau_{ATE} = \mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0))$$

What if we replace expectations with sample averages for the treated and control groups?

$$\hat{\tau}_{DIM} = \mathbb{E}(Y_i | D_i = 1) - \mathbb{E}(Y_i | D_i = 0)$$

- This is called the **difference-in-means estimator**

Difference-in-means estimator

What does the difference-in-means estimator identify? We can use the common “add and subtract” trick:

$$\begin{aligned}\hat{\tau}_{\text{DIM}} &= \mathbb{E}(Y_i | D_i = 1) - \mathbb{E}(Y_i | D_i = 0) \\ &= \mathbb{E}(Y_i(1) | D_i = 1) - \mathbb{E}(Y_i(0) | D_i = 0) \\ &= \mathbb{E}(Y_i(1) | D_i = 1) - \mathbb{E}(Y_i(0) | D_i = 0) - \mathbb{E}(Y_i(0) | D_i = 1) + \mathbb{E}(Y_i(0) | D_i = 1)\end{aligned}$$

Difference-in-means estimator

What does the difference-in-means estimator identify? We can use the common “add and subtract” trick:

$$\begin{aligned}\hat{\tau}_{\text{DIM}} &= \mathbb{E}(Y_i | D_i = 1) - \mathbb{E}(Y_i | D_i = 0) \\ &= \mathbb{E}(Y_i(1) | D_i = 1) - \mathbb{E}(Y_i(0) | D_i = 0) \\ &= \mathbb{E}(Y_i(1) | D_i = 1) - \mathbb{E}(Y_i(0) | D_i = 0) - \mathbb{E}(Y_i(0) | D_i = 1) + \mathbb{E}(Y_i(0) | D_i = 1) \\ &= \tau_{\text{ATT}} + (\mathbb{E}(Y_i(0) | D_i = 1) - \mathbb{E}(Y_i(0) | D_i = 0))\end{aligned}$$

The difference-in-means estimator equals the ATT plus an additional term.

'Selection Bias'

We refer to the last term as **selection bias**

$$\hat{\tau}_{\text{DIM}} = \tau_{\text{ATT}} + (\mathbb{E}(Y_i(0) \mid D_i = 1) - \mathbb{E}(Y_i(0) \mid D_i = 0))$$

- If the treated group has a different mean untreated potential outcome than the control group, our estimator is biased.

'Selection Bias'

We refer to the last term as **selection bias**

$$\hat{\tau}_{\text{DIM}} = \tau_{\text{ATT}} + (\mathbb{E}(Y_i(0) | D_i = 1) - \mathbb{E}(Y_i(0) | D_i = 0))$$

- If the treated group has a different mean untreated potential outcome than the control group, our estimator is biased.

For example, in our apartment example, the neighborhoods where apartments are built would have a higher *counterfactual rent* (in the absence of a new apartment) than the untreated neighborhoods

- We would mistakenly claim this is due to new apartments being built!

Roadmap

Difficulties of Establishing Causality

Counterfactual Thinking and Potential Outcomes

Randomized Experiments

Estimation via Regression

LaLonde Dataset

A/B Testing

When is selection bias not present?

$$\hat{\tau}_{\text{DIM}} = \tau_{\text{ATT}} + (\mathbb{E}(Y_i(0) \mid D_i = 1) - \mathbb{E}(Y_i(0) \mid D_i = 0))$$

The leading example of when selection bias is not present is when D_i is randomly assigned.

When is selection bias not present?

$$\hat{\tau}_{\text{DIM}} = \tau_{\text{ATT}} + (\mathbb{E}(Y_i(0) | D_i = 1) - \mathbb{E}(Y_i(0) | D_i = 0))$$

The leading example of when selection bias is not present is when D_i is randomly assigned.

When D_i is randomly assigned, we have that the untreated potential outcome and treatment are independent: $(Y_i(0) \perp\!\!\!\perp D_i)$

- If $(Y_i(0) \perp\!\!\!\perp D_i)$, then $\mathbb{E}(Y_i(0) | D_i) = \mathbb{E}(Y_i(0))$ and selection bias is 0.

When is selection bias not present?

$$\hat{\tau}_{\text{DIM}} = \tau_{\text{ATT}} + (\mathbb{E}(Y_i(0) | D_i = 1) - \mathbb{E}(Y_i(0) | D_i = 0))$$

The leading example of when selection bias is not present is when D_i is randomly assigned.

When D_i is randomly assigned, we have that the untreated potential outcome and treatment are independent: $(Y_i(0) \perp\!\!\!\perp D_i)$

- If $(Y_i(0) \perp\!\!\!\perp D_i)$, then $\mathbb{E}(Y_i(0) | D_i) = \mathbb{E}(Y_i(0))$ and selection bias is 0.
- Technically, we only need ‘mean-independence’ which is the latter term; but in experiments, we have full independence.

Experiments and causal estimands

In fact, experiments help us estimate more than the ATT. Note that if we randomly assign treatment, then $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i$. This means:

$$\mathbb{E}(Y_i(1) - Y_i(0) | D_i) = \mathbb{E}(Y_i(1) - Y_i(0))$$

and hence ATE = ATT = ATC

- Since we randomly assign treatment, the treated group looks like the control group looks like the population as a whole.

Quasi- and Natural Experiments

Writing about Empirical Work

Often you will hear the terms 'natural experiment' and 'quasi-experimental design' in applied work

- There are appropriate and inappropriate usages of this term (in my opinion)

If there are features of a research question in which we can think of D_i as being randomly assigned (for a subgroup), then we can use the term 'natural experiment' (in that it is naturally occurring)

- Angist (1990) uses the randomly assigned Vietnam-war draft
- Imbens, Rubin, and Sacerdote (2001) uses lottery winners which is randomly assigned among people who play the lottery

Observational Studies and “Exogenous Variation”

Writing about Empirical Work

In most contexts, you will not have a natural experiment. Instead, you will leverage variation in policies that we think is ‘exogenous’ (e.g. the roll-out of policies or using details of the policy)

- In general, these papers require more work to do a good job at causally identifying effects
- E.g. states roll-out policies in response to their outcomes causing selection bias

Observational Studies and “Exogenous Variation”

Writing about Empirical Work

In most contexts, you will not have a natural experiment. Instead, you will leverage variation in policies that we think is ‘exogenous’ (e.g. the roll-out of policies or using details of the policy)

- In general, these papers require more work to do a good job at causally identifying effects
- E.g. states roll-out policies in response to their outcomes causing selection bias

All hope is not lost, though. The rest of the course will teach you strategies for estimating a causal effect

Roadmap

Difficulties of Establishing Causality

Counterfactual Thinking and Potential Outcomes

Randomized Experiments

Estimation via Regression

LaLonde Dataset

A/B Testing

Regression estimates for experiments

The difference-in-means estimate is $\mathbb{E}(Y_i | D_i = 1) - \mathbb{E}(Y_i | D_i = 0)$

- We could, of course, estimate these means by hand (and you will be asked to do so on your assignment).

However it is typically far simpler to run a regression because it will also give you (robust) standard errors.

Regression

The regression we will run is as follows:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$

Since D_i is a dummy variable and α is the intercept, regression mechanics (which we'll talk about more in the next topic) tells us that

- $\hat{\alpha} = \hat{\mathbb{E}}(Y_i | D_i = 0)$ and $\hat{\tau} = \hat{\mathbb{E}}(Y_i | D_i = 1) - \hat{\mathbb{E}}(Y_i | D_i = 0)$

$\hat{\tau}$ is equal to our difference-in-means estimates!

Inference

Of course, we want to provide some measure of noise around our treatment effect estimates and be able to test the null that the ATE is equal to 0.

Our difference-in-means estimator is given by

$$\hat{\tau}_{\text{DIM}} = \frac{1}{n_1} \sum_{D_i=1} Y_i - \frac{1}{n_0} \sum_{D_i=0} Y_i$$

We will use the ‘weak null’ (more on this in a second) that $\tau_{\text{ATT}} = 0$.

- We will return in a few slides to discuss weak versus sharp null

Inference

The variance, under the null that $\tau = 0$, is given by

$$\text{var}(\hat{\tau}_{\text{DIM}}) = \mathbb{E}\left(\left(\frac{1}{n_1} \sum_{D_i=1} Y_i - \frac{1}{n_0} \sum_{D_i=0} Y_i\right)^2\right)$$

With a little algebra we can find

$$\text{var}(\hat{\tau}_{\text{DIM}}) = \frac{1}{n_1} \text{var}(Y_i(1)) + \frac{1}{n_0} \text{var}(Y_i(0))$$

- These variances can be estimated based on the variance within $D_i = 1$ and $D_i = 0$ respectively. Randomization ensures consistency

Inference via Regression

Recall the regression we run is:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$

Using robust standard errors, the variance estimate for $\hat{\tau}$ is approximately equal to the one on the previous slide.

- Using HC2 standard errors give exact variance from above

Roadmap

Difficulties of Establishing Causality

Counterfactual Thinking and Potential Outcomes

Randomized Experiments

Estimation via Regression

LaLonde Dataset

A/B Testing

The start of the credibility revolution

LaLonde (1986) is a seminal work in the onset of the 'credibility revolution' of causal inference to economics. The paper uses an experiment called the National Supported Work Demonstration (NSW) program

The program targeted randomly assigned a job training program to participants. The participants were drawn from 'ex-drug addicts, ex-criminal offenders, and high-school dropouts'.

Experimental effects

Since treatment was randomly assigned, we can use the difference-in-means estimator to estimate the $ATE = ATT = ATC$

- Note that in the context of this experiment, the ‘population’ we are thinking about is among the pool of participants
- It is not the ATE of the entire US population

LaLonde found experimental evidence of about \$886 for male participants and \$851 for female participants

LaLonde's test

LaLonde wanted to see what would happen if instead of using the experimental control group, he used observation data

- LaLonde brought in data from the Community Population Survey (CPS) and the Panel Study on Income Dynamics (PSID)

The 'observational dataset', as he called it, consists of the treated units from the experiment and the control group from the survey data

LaLonde's test

LaLonde wanted to see what would happen if instead of using the experimental control group, he used observation data

- LaLonde brought in data from the Community Population Survey (CPS) and the Panel Study on Income Dynamics (PSID)

The 'observational dataset', as he called it, consists of the treated units from the experiment and the control group from the survey data

- Control group looks very different from the treated group (higher income, better job market attachments, higher education, etc.)

LaLonde's test

He used a simple regression:

$$\text{wages}_i = \alpha + \tau D_i + X_i \beta + \varepsilon_i$$

- X_i were a vector of controls (age, education, and race)

LaLonde's test

He used a simple regression:

$$\text{wages}_i = \alpha + \tau D_i + X_i \beta + \varepsilon_i$$

- X_i were a vector of controls (age, education, and race)

The estimates from this regression using the nonexperimental sample yielded significantly different results with many estimates coming out as negative!

LaLonde's test

This was a very clever thing to do and led to many response articles trying to defend the then current methods

- Imbens and Xu (2024) present a really riviting account of the paper, responses, and modern solutions aimed to improve observational treatment effect estimates

LaLonde's test

This was a very clever thing to do and led to many response articles trying to defend the then current methods

- Imbens and Xu (2024) present a really riviting account of the paper, responses, and modern solutions aimed to improve observational treatment effect estimates

Robert Lalonde, Orley Ashenfelter, David Card, and others in the Princeton Industrial Relations Section in the 1980s were drivers of the 'credibility revolution'

Roadmap

Difficulties of Establishing Causality

Counterfactual Thinking and Potential Outcomes

Randomized Experiments

Estimation via Regression

LaLonde Dataset

A/B Testing

A/B Testing

It is *incredibly* common for firms to run many sets of experiments, which they call A/B testing

- Typically, it asks “which performs better, version A or version B”

For example, Youtubers will try to change their thumbnails randomly to different viewers and compare how often people click on the video

Difficulties with A/B Testing

Since this is effectively an experiment, understanding A/B testing should be simple; but it comes with some challenges:

- There can be many, many different ‘arms’ in the experiment so you have many comparisons to make
- As the experiment goes on, it is common to subset experimentation to the best-performing arms while dropping the worst performing ('adaptive bandit')

Multi-arm treatment effect estimates in A/B Testing

Our previous logic of randomization extends to an experiment where $D_i \in \{0, 1, \dots, K\}$ where $K + 1$ is the number of arms in the experiment.

Potential outcomes are given by $Y_i(D_i)$ and treatment effects are given by

$$\tau_i(d) = Y_i(d) - Y_i(0)$$

Multi-arm treatment effect estimates in A/B Testing

Our previous logic of randomization extends to an experiment where $D_i \in \{0, 1, \dots, K\}$ where $K + 1$ is the number of arms in the experiment.

Potential outcomes are given by $Y_i(D_i)$ and treatment effects are given by

$$\tau_i(d) = Y_i(d) - Y_i(0)$$

- If there is no natural ‘control group’, then pick one randomly to normalize. Treatment effects are then ‘relative’ to treatment 0

Randomization ensures that $\{Y_i(d)\}_{d=0}^K \perp\!\!\!\perp D_i$

Multi-arm treatment effect estimates in A/B Testing

Treatment effects can be estimated using regression:

$$Y_i = \alpha + \sum_{d=1}^K \mathbb{1}[D_i = d] \tau^d + \varepsilon_i$$

$\hat{\tau}^d$ are our treatment effect estimates

- Consistency comes from randomization (using similar math to the single treatment case)

Finding the 'best' treatment

Say you want to chose treatment d that has the largest treatment effect:

$$\hat{d} = \operatorname{argmax}_{d \in \{1, \dots, K\}} \hat{\tau}(d)$$

- As the number of observations in each arm grows, this procedure asymptotically will select the best treatment arm

The Winner's Curse

For smaller sample sizes, this will face the 'winner's curse' (Andrews, Kitagawa, and McCloskey (2024))

- The one with the highest effect is biased upwards by selecting on the error term

The Winner's Curse

For smaller sample sizes, this will face the 'winner's curse' (Andrews, Kitagawa, and McCloskey (2024))

- The one with the highest effect is biased upwards by selecting on the error term
- I've heard stories of tech companies running hundreds of arms with only a few observations per arm before picking a winner

The Winner's Curse

For smaller sample sizes, this will face the 'winner's curse' (Andrews, Kitagawa, and McCloskey (2024))

- The one with the highest effect is biased upwards by selecting on the error term
- I've heard stories of tech companies running hundreds of arms with only a few observations per arm before picking a winner

The term comes from auction theory where K bidders bid for the same product

- Even if each bidder can consistently estimate the value, the highest bidder will have the highest valuation and overvalue the good

Finding the 'best' treatment

Solutions to winner's curse:

- Use adaptive inference from Andrews, Kitagawa, and McCloskey (2024)

What about doing more experimentation on the best arm's to disentangle?

Multi-arm Bandit

A **Multi-arm bandit** (sometimes called adaptive experiments) is a method of experimentation where the probability of being assigned to any arm changes over the course of the experiment

- It is costly to experiment, so lower the probability of treatment for arms that seem to not be good

Multi-arm Bandit

A **Multi-arm bandit** (sometimes called adaptive experiments) is a method of experimentation where the probability of being assigned to any arm changes over the course of the experiment

- It is costly to experiment, so lower the probability of treatment for arms that seem to not be good

This creates issues with inference since each observation depends on all previous observations (inducing dependence)

- Solutions exist, but we will skip the details in this class. See Hadad, Hirshberg, Zhan, Wager, and Athey (PNAS, 2021) for an example