Topic 1: Introduction to Causal Inference and Potential Outcomes Framework

ECON 5783 — University of Arkansas

Prof. Kyle Butts

September 2025

Potential Outcomes Framework

We learn:

→ How to think of Causal Effects in terms of ”All Else Equal”

→ What are “Potential Outcomes”?

→ Identifying causal effects via randomized experiments

Readings to complement this lecture are:

→ Cunningham (2021) *Causal Inference: The Mixtape*, Chapters 4

→ Good summary videos: https://www.youtube.com/watch?v=lFKYfCeXiI4 and https://www.youtube.com/watch?v=jVkQPnARPQ0

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:

$\rightarrow$ 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 (statistical inference) but trying to know if X caused Y (causal inference) is much more difficult a problem

$\rightarrow$ We will try to tackle this problem, but be humble of the difficulties of this task

The Enlightenment Era and The Scientific Method

In some sense, the methods of establish causal links date back to the Enlightenment Era

→ Empiricism, Experimentation, and 'Reasoning' led to many scientific breakthroughs and, some scholars argue, was essential for the Industrial Revolution (e.g. Joel Mokyr)

The Enlightenment Era and The Scientific Method

In some sense, the methods of establish causal links date back to the Enlightenment Era

→ Empiricism, Experimentation, and 'Reasoning' led to many scientific breakthroughs and, some scholars argue, was essential for the Industrial Revolution (e.g. Joel Mokyr)

They would have "Scientific Societies" that were basically large dinner parties where they would dicuss how to learn about the world ('experimentation')

Royal (Scientific) Society of London

A MEETING OF THE ROYAL SOCIETY IN CRANE COURT (*see p.* 106).

Most famously, was the Scientific Society of London

→ Ben Franklin's key and kite was written in a letter to the Society

Scientific Society of London

'Nullius in verba'

The Royal Society's motto 'Nullius in verba' was adopted in its First Charter in 1662. is taken to mean 'take nobody's word for it'. It is an expression of the determination of Fellows to withstand the domination of authority and to verify all statements by an appeal to facts determined by experiment.

Source: https://royalsociety.org

Randomized Controlled Trial

The first randomized controlled trial (RCT) was conducted by James Lind in 1753

→ Randomly assigned 12 sailors into 6 groups to have different potential ‘treatments’

Citrus ended up winning out

What makes an RCT work well?

In groups, let's think about what makes an RCT effective at identifying what the true effects of a treatment is

What makes an RCT work well?

In groups, let's think about what makes an RCT effective at identifying what the true effects of a treatment is

By *randomly* giving people different treatments, we can be sure the groups are similar

$\rightarrow$ 'Controlled' trial == making sure the units are comparable

“Observational” version

Say, instead, that James Lind went out and surveyed a bunch of people

→ Asked whether or not they had scurvy and whether or not they consumed citrus

He documents that people that have a lot of citrus rarely have scurvy

→ Why is this evidence much weaker?

Random Allocation ensures comparability

Having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity.

—Austin Bradford Hill, 1952

The ‘Social Scientific Method’

What then are we to do when we try to establish causality in the messy messy world?

Option 1: Run experiments. Get out in the world and run RCTs!

→ E.g. guaranteed income experiments, assigning a free lawyer for eviction court, micro-credit loans in developing countries

The ‘Social Scientific Method’

But, in many cases, we can not run experiments

→ e.g. we care about the impacts of kids on parents, but can not randomly assign people to have kids

Option 2: Look for ‘natural experiments’ in the world

→ e.g. draft lottery; discrete cut-offs in policies; random cost shocks to firms

The ‘Social Scientific Method’

As our ability to approximate a random experiment worsens, the credibility of our estimates worsens as well

$\rightarrow$ The RCT is often called the ‘gold standard’ for this reason

Option 3: Control for important variables and do as good as we can

The 'Social Scientific Method'

Sometimes we need to answer questions even if a natural experiment does not fall in our laps

→ But, this requries the *most caution*; we will discuss strategies to approximate an experiment as best as we can

Point of Caution: Just because a question is important, does not mean we should throw away careful scientific rigor

The Social Scientist' obligation

A bit soap-boxy, but it's important

> *The first principle is that you must not fool yourself — and you are the easiest person to fool.*
>
> *— Richard Feynman, 1974*

The key insight of the scientific method is to *rule out potential alternative explanations*

→ As much as possible, you want to *prove your theory wrong*.

A good social scientist is an expert on the topic they are researching

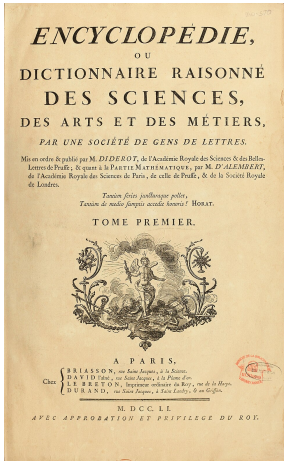
→ Just because you do not *know* an alternative explanation, does not mean it does not exist

In that spirit, let us go back to the original claim in this introduction: “scientific breakthroughs and, some scholars argue, was essential for the Industrial Revolution”

Is this true?

Social Scientific Method in Action

The Enlightenment and the Industrial Revolution


ENCYCLOPÉDIE,
OU
DICTIONNAIRE RAISONNÉ
DES SCIENCES,
DES ARTS ET DES MÉTIERS,
PAR UNE SOCIÉTÉ DE GENS DE LETTRES.
Mis en ordre & publié par M. DIDEROT, de l'Académie Royale des Sciences & des Belles-Lettres de Prusse; & quant à la PARTIE MATHÉMATIQUE, par M. D'ALEMBERT, de l'Académie Royale des Sciences de Paris, de celle de Prusse, & de la Société Royale de Londres.
Tantùm series juncturaque pollet,
Tantùm de medio sumptis accedit honoris! HORAT.
TOME PREMIER.
A PARIS,
Chez BRIASSON, rue Saint Jacques, à la Science.
DAVID l'aîné, rue Saint Jacques, à la Plume d'or.
LE BRETON, Imprimeur ordinaire du Roy, rue de la Harpe.
DURAND, rue Saint Jacques, à Saint Landry, & au Griffon.
M. DCC. LI.
AVEC APPROBATION ET PRIVILEGE DU ROY.

Squicciarini and Voigtländer (2015, QJE) try to answer this using *the Encylopedié*

→ Published in France in 1751 and contained a bunch of evidence from Enlightenment scientists

A single person in the town might buy this expensive book and share it with other people

Social Scientific Method in Action

The Enlightenment and the Industrial Revolution

Squicciarini and Voigtländer (2015, QJE) digitized records of who purchased the Encyclopedié (7081 in total)

→ In their sample of French cities, 85 had purchased one and 108 did not

Comparing cities of similar size and geographic location, they find that cities with this scientific knowledge grew at faster rates than other cities

→ With your groups, think about whether or not this is a good 'experiment'

To bolster their evidence, they do:

1. Subscriber rate is uncorrelated with growth from 1400 to 1750 (potentially alleviating some concerns about which kinds of places subscribed)
2. They include controls for port access, universities, and city density

Social Scientific Method in Action

Jon Snow and Cholera

Three major waves of cholera in the early to mid 1800s in London, largely thought to be spread by miasma ("dirty air")

→ No theory of the germ yet

Social Scientific Method in Action

Jon Snow and Cholera

John Snow believed cholera was spread through the Thames water supply through an 'invisible creature' that entered the body through food and drink

London passes ordinance requiring water utility companies to move inlet pipe further up the Thames, above the city center, but not everyone complies

→ Natural experiment: Lambeth water company moves its pipe between 1849 and 1854; Southwark and Vauxhall water company delayed

"Staggered rollout" of policies is a common empirical method

Social Scientific Method in Action

Computers and Wage Structure

Alan Krueger (1993, QJE) use a regression strategy and US survey data and documents that workers who use a computer earn 10-15% more than those that do not

$\rightarrow$ Use a bunch of different regression methods and find 'robust' results

Social Scientific Method in Action

Computers and Wage Structure

Alan Krueger (1993, QJE) use a regression strategy and US survey data and documents that workers who use a computer earn 10-15% more than those that do not

→ Use a bunch of different regression methods and find 'robust' results

DiNardo and Pischke (1997, QJE) revisit this analysis using German data

→ Find similar estimates for computers

→ But, find remarkably similar results for the use of pencils at work‼

Don't forget to rule out potential explanations!

Difficulties of Establishing Causality

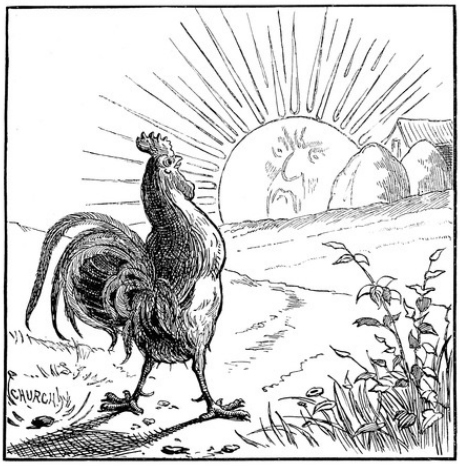
Counterfactual Thinking and Potential Outcomes

Randomized Experiments

- Estimation via Regression
- LaLonde Dataset

A/B Testing

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

Drinking wine is correlated with improved health

→ Drinking wine is also correlated with income . . .

Causality does not imply correlation

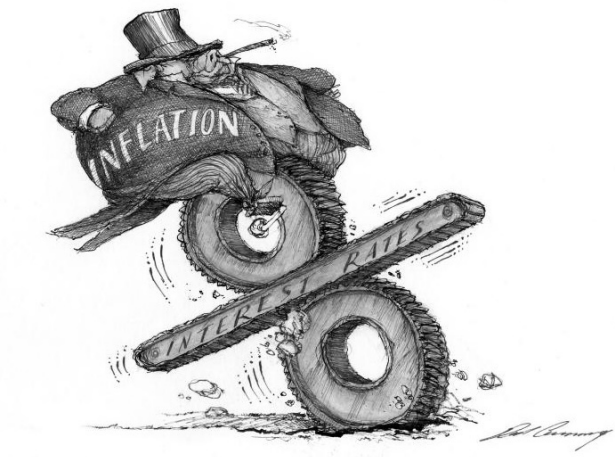
A boat is traveling on a windy day

→ 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

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

The Encyclopedié

One main concern with Squicciarini and Voigtlände's approach is reverse causality

→ Places with more innovations might seek out more scientific knowledge (growth causing subscriptions)

To address this, they use the presence of historic scientific society to 'predict' which places would buy an Encyclopedié

→ This is called an 'instrument' which we will discuss later in this class

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

$\rightarrow$ 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

$\rightarrow$ 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 Xs. Why can we not just compare Ys for people with different values of X?

$\rightarrow$ As economists, we know that people are optimizers. They chose their 'optimal' X for different reasons.

$\rightarrow$ 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 Xs. Why can we not just compare Ys for people with different values of X?

$\rightarrow$ As economists, we know that people are optimizers. They chose their 'optimal' X for different reasons.

$\rightarrow$ 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 Xs often involves comparing people with different lurking variables, $\boldsymbol{Z}$, too

$\rightarrow$ Is it X causing the change in Y or is it the multiude 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

$\rightarrow$ Bad assumption in $\Longrightarrow$ 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)

The Fundamental Problem of Causal Inference

We want to understand: what is the causal effect of some treatment D on outcome Y?

The fundamental problem is that we can only observe one potential outcome for each unit:

$\rightarrow$ If a unit receives treatment ($D_i = 1$), we observe $Y_i(1)$ but not $Y_i(0)$

$\rightarrow$ If a unit doesn't receive treatment ($D_i = 0$), we observe $Y_i(0)$ but not $Y_i(1)$

To estimate causal effects, we need to predict the *missing counterfactual outcome*

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

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

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 neigborhood. 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, $Y_i(D_i)$.

$\rightarrow$ $Y_i(1)$ is the average rent *in the counterfactual world* where the new apartment is built

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

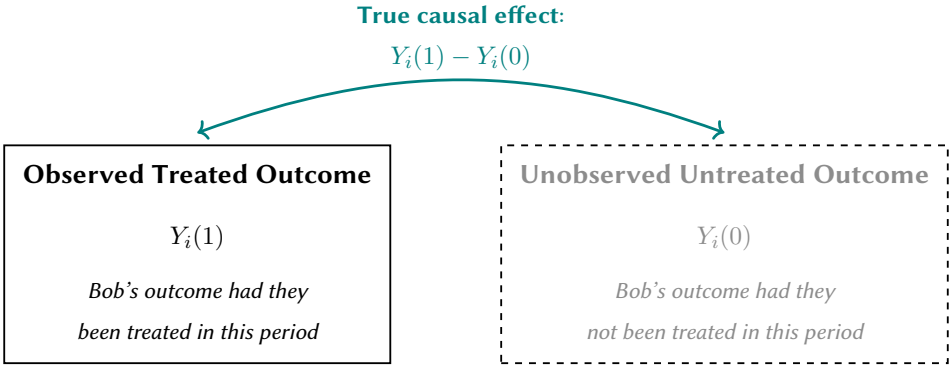
Potential Outcome Framework

→ 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

The Problem: We Only Observe One Potential Outcome



For any individual, we can only observe *one* of these potential outcomes, never both simultaneously

Unit-level treatment effect

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

$\rightarrow$ 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 $\tau_i = Y_i(1) - Y_i(0)$

$\rightarrow$ 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*!

$\rightarrow$ 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 oberve?

In our dataset, we observe average rent Y_i and whether or not the aparmtent 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 oberve?

In our dataset, we observe average rent Y_i and whether or not the aparmtent 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 $\tau_i = Y_i(1) - Y_i(0)$

$\rightarrow$ 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 $\tau_i = Y_i(1) - Y_i(0)$

$\rightarrow$ 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 $\tau_i = Y_i(1) - Y_i(0)$

$\rightarrow$ 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

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

$\rightarrow$ 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.

$\rightarrow$ 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 then 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 then 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

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.

$$\tau_{\mathsf{ATE}} = \mathbb{E}[\tau_i] = \mathbb{E}[\,\mathbb{E}[\tau_i \mid D_i]]$$

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.

$$\begin{aligned}\tau_{\mathsf{ATE}} &= \mathbb{E}[\tau_i] = \mathbb{E}[\,\mathbb{E}[\tau_i \mid D_i]] \\ &= \mathbb{P}(D_i = 1)\,\mathbb{E}[\tau_i \mid D_i = 1] + \mathbb{P}(D_i = 0)\,\mathbb{E}[\tau_i \mid D_i = 0]\end{aligned}$$

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.

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

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 charactersitics*, e.g. gender, race, income, age, etc.

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

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

$\rightarrow$ X_i is some characteristic of the population

Conditional ATE versus ATE

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

$$\tau_{\text{ATE}} = \mathbb{E}[\tau_i] = \mathbb{E}[\,\mathbb{E}[\tau_i \mid X_i]]$$

$\rightarrow$ 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:

$$\tau_{\mathsf{ATE}} = \mathbb{E}[\tau_i] = \mathbb{E}[\,\mathbb{E}[\tau_i \mid X_i]]$$

$\rightarrow$ 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

$$\tau_{\mathsf{ATT}} = \mathbb{E}[\tau_i] = \mathbb{E}[\tau_{\mathsf{CATE}}(x)] = \mathbb{E}[\,\mathbb{E}[\tau_i \mid X_i] \mid D_i = 1]$$

Problems with observational data

Our treatment effect parameter is given by

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

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

$$\hat{\tau}_{\text{DIM}} = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

$\rightarrow$ 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:

$$
\begin{aligned}
\hat{\tau}_{\mathsf{DIM}} &= \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \\
&= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\
&= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] - \mathbb{E}[Y_i(0) \mid D_i = 1] + \mathbb{E}[Y_i(0) \mid D_i = 1]
\end{aligned}
$$

Difference-in-means estimator

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

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

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

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

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

For example, in our appartment 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 mistakingly claim this is due to new apartments being built!

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?

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

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

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

$\rightarrow$ If $(Y_i(0) \perp\!\!\!\perp D_i)$, then $\mathbb{E}[Y_i(0) \mid D_i] = \mathbb{E}[Y_i(0)]$ and selection bias is 0.

When is selection bias not present?

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

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

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

$\rightarrow$ If $(Y_i(0) \perp\!\!\!\perp D_i)$, then $\mathbb{E}[Y_i(0) \mid D_i] = \mathbb{E}[Y_i(0)]$ and selection bias is 0.

$\rightarrow$ 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:

$$\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$$

and hence ATE $=$ ATT $=$ ATC

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

‘Balance Tables’

When running an experiment, it is common to collect information about the participants, call this information a vector X_i.

It is common to show that the randomization “worked” by showing that the distribution of X_i is similar between treatment and control groups. This is often done with a balance table that shows means of X_i by treatment status

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 occuring)

→ 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

Regression estimates for experiments

The difference-in-means estimate is $\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$

$\rightarrow$ We could, of course, estimate these means by hands (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

$\rightarrow \hat{\alpha} = \hat{\mathbb{E}}[Y_i \mid D_i = 0]$ and $\hat{\tau} = \hat{\mathbb{E}}[Y_i \mid D_i = 1] - \hat{\mathbb{E}}[Y_i \mid D_i = 0]$

$\hat{\tau}$ 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 $\tau_{\text{ATT}} = 0$.

$\rightarrow$ 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

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

With a little algebra we can find

$$\text{Var}\left(\hat{\tau}_{\text{DIM}}\right) = \frac{1}{n_1}\,\text{Var}\left(Y_i(1)\right) + \frac{1}{n_0}\,\text{Var}\left(Y_i(0)\right)$$

$\rightarrow$ 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.

$\rightarrow$ Using HC2 standard errors give exact variance from above

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

$\rightarrow$ Note that in the context of this experiment, the 'population' we are thinking about is among the pool of participants

$\rightarrow$ 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:

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

$\rightarrow$ X_i were a vector of controls (age, education, and race)

LaLonde's test

He used a simple regression:

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

$\rightarrow$ 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'

A/B Testing

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

$\rightarrow$ 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 $D_i \in \{0, 1, \ldots, K\}$ 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 $D_i \in \{0, 1, \ldots, K\}$ 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 $\tau_i(d) = Y_i(d) - Y_i(0)$

$\rightarrow$ 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

$\rightarrow$ 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:

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

$\rightarrow$ 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))

$\rightarrow$ 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 consistenly estimate the value, the highest bidder will have the highest valuation and overvalue the good

Finding the ‘best’ treatment

Solutions to winner’s curse:

$\rightarrow$ 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