Roadmap

Overview

Causality vs Prediction

Potential Outcomes

Randomized experiments

Introduction!

- Welcome to SGPE 2024!! This is going to be a fun week of learning about causal inference methods!
- I'm Scott Cunningham, Ben H. Williams professor of economics at Baylor University, author of Causal Inference: the Mixtape
- I work on a variety of eclectic topics including sex work, drug policy, abortion policy, and more recently, detecting and treating self harm in prisons and jails

What is my teaching like

- I focus on intuition, mechanics, calculations, meaning, assumptions, and coding.
- I advocate for data visualization – blending the art and science of causal inference.
- I teach with passion, enthusiasm, and deep joy.
- I'm not an econometrician, so I sometimes take the scenic route to get to the point.

This week

- This is a week long course on three core research designs in causal inference:
    1. Unconfoundedness: Matching, Propensity scores, Regression adjustment
    2. Regression discontinuity design
    3. Instrumental Variables

Class goals

1. **Confidence**: You will feel like you have a good understanding of causal inference so that by the end it doesn't feel all that mysterious or intimidating
2. **Comprehension**: You will have learned a lot both conceptually and in the specifics, particularly with regards to issues around identification and estimation
3. **Competency**: You will have more knowledge of programming syntax in Stata and R (and python!) so that later you can apply this in your own work

Roadmap

Overview

Causality vs Prediction

Potential Outcomes

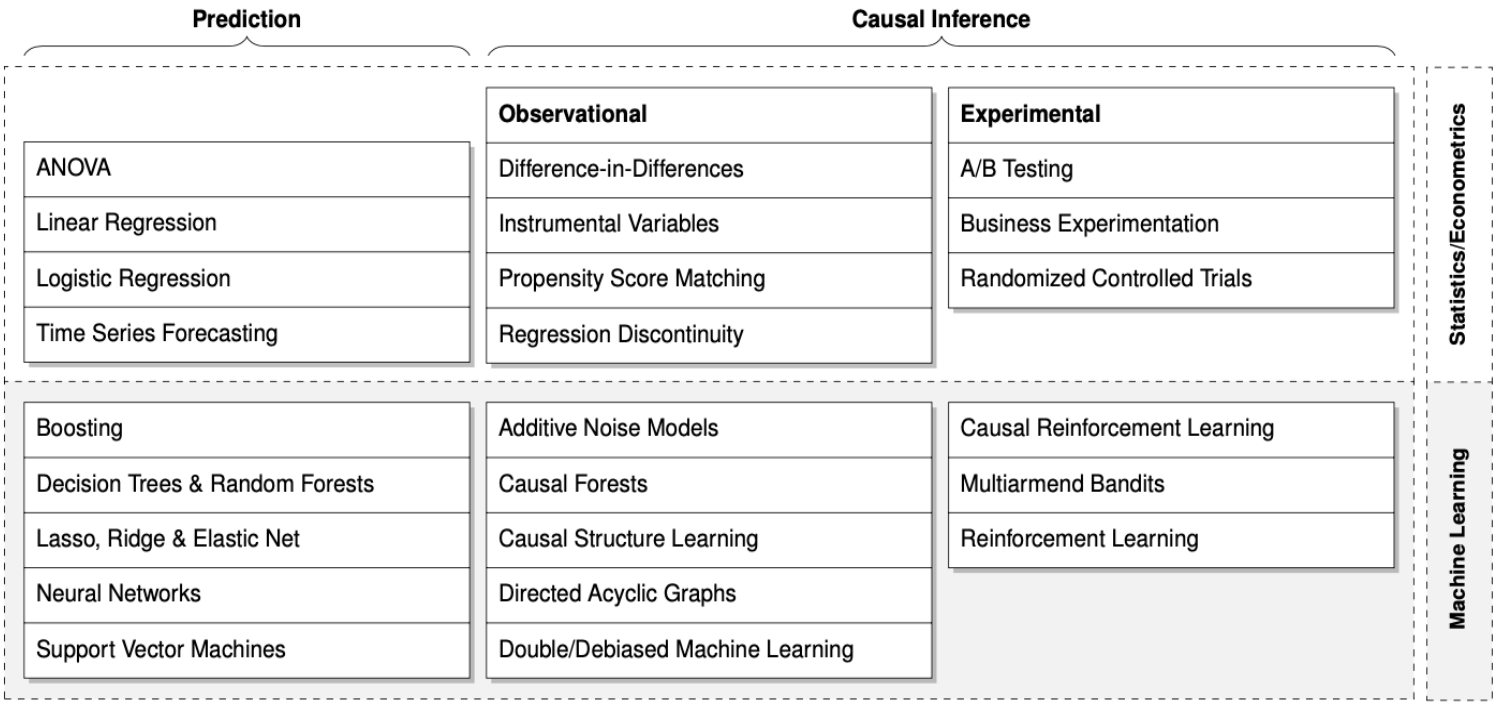
Randomized experiments

What is Causality?

- Causal inference is about beliefs – when is your belief that one thing caused another justified when is it not?
- We will be framing the question always with reference to the experimental design, and you'll hear me say "design" a lot
- We will be learning new notation for some of you and new ways of thinking in addition to econometrics
- I have some assignments for those who are wanting to go deeper called "Crits" which I'll explain

Causal Inference vs Prediction

Figure 1: Examples of popular data analysis algorithms in statistics and econometrics, as well as machine learning and artificial intelligence, classified according to prediction and causal inference methods. Causal inference methods are further differentiated according to observational (based on ex-post observed data) and experimental approaches.



Causal Inference vs Prediction

Traditional prediction

- Traditional prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy
- “Does this person have heart disease?”, “How many books will I sell?”
- It is not predictions of what effect a choice will have, though

Causal inference

- Causal inference is also a type of prediction, but it’s a prediction of a *counterfactual* associated with a particular *choice taken*
- Causal inference takes that predicted (or imputed) counterfactual and constructs a causal effect that we hope tells us about a future in the event of a similar choice taken

Roadmap

Overview

Causality vs Prediction

Potential Outcomes

Randomized experiments

Three New Ideas

1. **Counterfactual**: Philosophers come to it first and its central role in causal inference makes causality *unknowable* that the project is nearly derailed
2. **Treatment assignment mechanism**: Neyman and Fisher solve the counterfactual problem in statistics and lay the foundation of the modern randomized controlled trial (RCT) with their focus on the selection process
3. **No One Causal Effect**: There is no such thing as “the causal effect”; there’s many and your first step is to pick a parameter (not as easy as it sounds)

Modern Philosophers Introduce Counterfactual Comparisons

"If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten it, people would be apt to say that eating of that dish was the source of his death." – John Stuart Mill (19th century moral philosopher and economist)

"Causation is something that makes a difference, and the difference it makes must be a difference from what would have happened without it." – David Lewis (20th century philosopher)

Counterfactuals Almost Derailed Causal Inference

Mill's counterfactuals were immensely valuable for the clarity of the definition as well as its intuitive validity of causality, but it came at a huge price

If I have to know what would have happened had I not eaten the dish, but I did eat the dish, then isn't it actually impossible to know the causal effect of eating the dish?

Statisticians surprisingly resolve this tension in the early 20th century with the introduction of notation and the principles of treatment assignment

Statistical origins

"Yet, although the seeds of the idea that [causal effects are comparisons of potential outcomes] can be traced back at least to the 18th century, the formal notation for potential outcomes was not introduced until 1923 by Neyman." – Don Rubin (1990)

Jerzy Neyman's Notation

- Jerzy Neyman's 1923 masters thesis describes a field experiment with differing plots of land (imagine hundreds of square gardens) and many different "varieties" of fertilizer that farmers could apply to the land
- "U_{ik} is the yield of the ith variety on the kth plot..." (Neyman 1923)
- He calls U_{ik} "potential yield", as opposed to the realized yield because i (the fertilizer type) described all possible fertilizers that could be assigned to each k square garden
- For each fertilizer there was an associated "potential yield" which he called U and even though not all of them could exist, to him they were still real – just unknown

Jerzy Neyman's Notation

- Farmers draw fertilizer from an urn, like a bingo ball from a bingo ball machine, with replacement and apply it to each square garden
- Neyman's urn model was a classic thought experiment, but he was also describing the randomized experiment which until then had not been formally analyzed
- Similar ideas can be found in Ronald Fisher, but not formalized the way Neyman had

Treatment assignment mechanism

"Before the 20th century, there appears to have been only limited awareness of the concept of the assignment mechanism. Although by the 1930s, randomized experiments were firmly established in some areas of scientific investigation, notably in agricultural experiments, there was no formal statement for a general assignment mechanism and, moreover, not even formal arguments in favor of randomization until Fisher (1925)." (Imbens and Rubin 2015)

Potential outcomes notation

Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if placed on ventilator at time } t \\ 0 \text{ if not placed on ventilator at time } t \end{cases}$$

where i indexes an individual observation, such as a person

Potential outcomes notation

Potential outcomes:

$$Y_{i,t}^{j} = \begin{cases} 1 \text{ health if placed on ventilator at time } t \\ 0 \text{ health if not placed on ventilator at time } t \end{cases}$$

where j indexes a potential treatment status for the same i person at the same t point in time

Realized vs potential outcomes

- Potential outcome Y^1 and Y^0 are potential outcomes under different states of the world but they don't exist – not yet – until a choice is made
- Realized outcomes Y are what actually happens when a choice gets made – it's going to either by Y^1 or Y^0, but only one of them will ever exist
- Listen Guido Imbens explain potential outcomes versus realized outcomes: https://www.youtube.com/watch?v=drGkRy53bB4

Selection into Treatment

- Treatment assignment *mechanisms* are the precise reasons that certain individuals get placed into treatment categories
- This week we will consider three of them:
  1. Selection on covariates
  2. Selection on running variables
  3. Selection on instrumental variables
- These aren't assumptions – if selection into treatment does not happen for that reason, then the methods we discuss are incorrectly applied
- You don't start with a model – you start with an understanding of selection and the mechanism that caused units to get selected

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i, associated with a ventilator is equal to $Y_i^1 - Y_i^0$.

Important definitions

Definition 2: Switching equation

An individual's realized health outcome, Y_i, is determined by treatment assignment, D_i which selects one of the potential outcomes:

$$
\begin{aligned}
Y_i &= D_i Y_i^1 + (1 - D_i) Y_i^0 \\
Y_i &= \begin{cases} Y_i^1 \text{ if } D_i = 1 \\ Y_i^0 \text{ if } D_i = 0 \end{cases}
\end{aligned}
$$

Not the same as treatment assignment mechanism. Treatment assignment mechanism describes how D was assigned, not whether it was assigned.

Missing data problem

Definition 3: Fundamental problem of causal inference

If you need both potential outcomes to know causality with certainty, then since it is impossible to observe both Y_i^1 and Y_i^0 for the same individual, δ_i, is *unknowable*.

This is my reason from saying Mill's counterfactual framework derailed the quest for causal effects given counterfactuals are fictional

Missing data problem

- Fundamental problem of causal inference is deep and impossible to overcome – not even with more data (you will always with more data be missing one of the potential outcomes)
- Causal inference is a missing data problem
- All of causal inference involves imputing missing counterfactuals and not all imputations are equal

Average Treatment Effects

Definition 4: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$\begin{aligned} E[\delta] &= E[Y^1 - Y^0] \\ &= E[Y^1] - E[Y^0] \end{aligned}$$

Aggregate parameters based on individual treatment effects are *summaries* of individual treatment effects

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1] \end{aligned}$$

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation.

Conditional Average Treatment Effects

Definition 6: Average Treatment Effect on the Untreated (ATU)

The average treatment effect on the untreated group is equal to the average treatment effect conditional on being untreated:

$$\begin{aligned} E[\delta|D=0] &= E[Y^1 - Y^0|D=0] \\ &= E[Y^1|D=0] - E[Y^0|D=0] \end{aligned}$$

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

Average Treatment Effects are Simple Summaries

- Notice how in all three of these, all we did was take the defined treatment effect at the individual and aggregate
- Because aggregate causal parameters are *summaries* of individual treatment effects, each of which cannot be calculated, the aggregates cannot be calculated either
- Missing data in this context isn't missing your car keys – it's missing unicorns and fire breathing dragons (fictional vs real data)
- While we cannot measure average causal effects, we can estimate them, but only in situations and we review one – randomization

Simple Comparisons

Definition 7: Simple difference in mean outcomes (SDO)

A simple difference in mean outcomes (SDO) can be approximated by comparing the sample average outcome for the treatment group ($D = 1$) with a comparison group ($D = 0$)

$$SDO = E[Y^1|D = 1] - E[Y^0|D = 0]$$

SDO is not a causal parameter because it's comparing Y^1 and Y^0 for different units, not the same units, so what is it measuring?

Decomposition of the SDO

Decomposition of the SDO

The SDO is made up of three things:

$$\begin{aligned} E[Y^1|D=1] - E[Y^0|D=0] &= ATE \\ &\quad + E[Y^0|D=1] - E[Y^0|D=0] \\ &\quad + (1-\pi)(ATT - ATU) \end{aligned}$$

where π is the share of units in the treatment group. Now let's see how we get here.

Two ways to rewrite the ATE

- Before we get started, let's look closely at the definition of the ATE
  - → We can express the ATE as the weighted average of the ATT and the ATU, and ...
  - → We can express the ATE as the sum of four conditional means multiplied by corresponding weights (Law of iterated expectations)
- They are in fact the exact same formula once you write down the definition of the ATT and the ATU
- Let's do it together before we get started:
  https://docs.google.com/spreadsheets/d/10DuQqGtH_Ewea7zQoLTFYHbnvqaTVDhn2GDzq3Oa6EQ/edit?usp=sharing

Begin with ATE definition

Rewrite the definition of the ATE

$$
\begin{aligned}
\text{ATE} &= E[Y^1] - E[Y^0] \\
&= \pi ATT + (1-\pi) ATU \\
&= \pi E[Y^1|D=1] - \pi E[Y^0|D=1] \\
&\quad +(1-\pi)E[Y^1|D=0] - (1-\pi)E[Y^0|D=0] \\
\text{ATE} &= \{\pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0]\} \\
&\quad -\{\pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]\}
\end{aligned}
$$

Let's make this easier to read by replacing the last row with letters

Change notation

Substitute letters for expectations

$$
\begin{aligned}
E[Y^1|D=1] &= a \\
E[Y^1|D=0] &= b \\
E[Y^0|D=1] &= c \\
E[Y^0|D=0] &= d \\
\text{ATE} &= e
\end{aligned}
$$

Rewrite ATE definition

Rewrite ATE

$$e = \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\}$$

Simple manipulation of ATE definition

$$
\begin{aligned}
e &= \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\} \\
e &= \pi a + b - \pi b - \pi c - d + \pi d \\
e &= \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a} - \mathbf{a}) + (\mathbf{c} - \mathbf{c}) + (\mathbf{d} - \mathbf{d}) \\
0 &= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)a - (1-\pi)b + (1-\pi)d - (1-\pi)c \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)(a-c) - (1-\pi)(b-d)
\end{aligned}
$$

Carry forward from previous slide

$$\mathbf{a} - \mathbf{d} \quad = \quad e + (\mathbf{c} - \mathbf{d}) + (1 - \pi)(a - c) - (1 - \pi)(b - d)$$

Replace letters with original terms

$$\begin{aligned} E[Y^1|D=1] - E[Y^0|D=0] \quad = \quad & \text{ATE} \\ & + (E[Y^0|D=1] - E[Y^0|D=0]) \\ & + (1-\pi)(\underbrace{\{E[Y^1|D=1] - E[Y^0|D=1]\}}_{\text{ATT}} \\ & - (1-\pi)(\underbrace{\{E[Y^1|D=0] - E[Y^0|D=0]\}}_{\text{ATU}} \end{aligned}$$

Purple terms are based on missing counterfactuals and therefore cannot be calculated.

This is an *identity*

Decomposition of the SDO

Decomposition of the SDO

$$
\begin{aligned}
E[Y^1|D=1] - E[Y^0|D=0] \quad &= \quad ATE \\
&\quad +(E[Y^0|D=1] - E[Y^0|D=0]) \\
&\quad +(1-\pi)(ATT - ATU)
\end{aligned}
$$

Although we started with π (the share of units in treatment), note we have weighted the heterogeneity bias term by $1-\pi$ (the share of units in control)

Estimate SDO with sample averages

$$\underbrace{E_N[Y|D=1] - E_N[Y|D=0]}_{\text{Estimate of SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}}$$
$$+ \underbrace{E[Y^0|D=1] - E[Y^0|D=0]}_{\text{Selection bias}}$$
$$+ \underbrace{(1-\pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

Using the switching equation and sample averages, we can calculate $E_N[Y|D=1] \rightarrow E[Y^1|D=1]$, $E_N[Y|D=0] \rightarrow E[Y^0|D=0]$ and $(1-\pi)$ is the share of the population in the control group.

Illustrating selection bias with spreadsheets

- Let's do an exercise together that illustrates this decomposition: https://docs.google.com/spreadsheets/d/10DuQqGtH_Ewea7zQoLTFYHbnvqaTVDhn2GDzq3Oa6EQ/edit?usp=sharing
- We will assume that these patients go see a doctor, but the doctor is "perfect" – the Perfect Doctor – in that they know what's best for each patient *and then does what is best for them*
- We will show that this action – a particular treatment assignment mechanism – generates biased estimates of causal effects according to that formula

Roadmap

Overview

Causality vs Prediction

Potential Outcomes

Randomized experiments

Steps that define most projects

1. Define our target parameters, usually an average treatment effect of some kind
2. Who chose the treatment? What did they know? What assumption does that imply?
3. Use estimators that are unbiased in the data you have defined by step 2
4. Then we estimate standard errors to quantify the uncertainty

Independence

Independence assumption

Treatment is assigned to a population independent of that population's potential outcomes

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

When a binary treatment is assigned to units independent of any variable, then those variables have the same average in treatment and control *even if they aren't observed*

$$\begin{aligned} E[Y^0|D=1] &= E[Y^0|D=0] \\ E[Y^1|D=1] &= E[Y^1|D=0] \end{aligned}$$

Random Assignment Solves the Selection Problem

$$\underbrace{E[Y|D=1]-E[Y|D=0]}_{\text{SDO}} = \underbrace{E[Y^1]-E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{E[Y^0|D=1]-E[Y^0|D=0]}_{\text{Selection bias}} + \underbrace{(1-\pi)(ATT-ATU)}_{\text{Heterogenous treatment effect bias}}$$

- If treatment is independent of potential outcomes, then swap out equations and **selection bias** zeroes out:

$$E[Y^0|D=1]-E[Y^0|D=0] = 0$$

Random Assignment Solves the Heterogenous Treatment Effects

- How does randomization affect heterogeneity treatment effects bias from the third line? Rewrite definitions for ATT and ATU:

$$\text{ATT} = E[Y^1|D=1] - E[Y^0|D=1]$$
$$\text{ATU} = E[Y^1|D=0] - E[Y^0|D=0]$$

- Rewrite the third row bias after $1-\pi$:

$$\begin{aligned} ATT - ATU &= \textbf{E[Y}^1 \mid \textbf{D=1]} - E[Y^0|D=1] \\ &\quad -\textbf{E[Y}^1 \mid \textbf{D=0]} + E[Y^0|D=0] \\ &= 0 \end{aligned}$$

- If treatment is independent of potential outcomes, then:

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= E[Y^1] - E[Y^0] \\ SDO &= ATE \end{aligned}$$

Identification with Full Independence

$$\underbrace{E[Y|D=1] - E[Y|D=0]}_{\text{Estimate of SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{0}_{\text{Selection bias}} + \underbrace{0}_{\text{Heterogenous treatment effect bias}}$$

SDO is unbiased estimate of ATE with randomized treatment assignment because it sets selection bias to zero and $ATT = ATU$.

Interference when aggregating units

- While treatment effects are defined at individual level, aggregate parameters combine units
- This therefore means that for the aggregate parameters to be stable, one unit's treatment choice cannot "interfere" with another unit's potential outcomes
- Placing limits on those possibilities creates challenges for definitions and estimation that are probably huge headaches, even in the RCT
- Violations are an active area of scholarship and important for social networks, peer effects and various platforms (e.g., Twitter)

SUTVA

- SUTVA stands for “stable unit-treatment value assumption”
  1. **S**: ***s****table*
  2. **U**: across all ***u****nits*, or the population
  3. **TV**: ***t****reatment-value* (“treatment effect”, “causal effect”)
  4. **A**: ***a****ssumption*
- Largely about interference when aggregating but also poorly defined treatments and scale

SUTVA: No spillovers to other units

- What if we impose a treatment at one neighborhood but not a contiguous one?
- Treatment may spill over causing $Y = Y^1$ even for the control units because of spillovers from treatment group
- Can be mitigated with careful delineation of treatment and control units so that interference is impossible, may even require aggregation (e.g., classroom becomes the unit, not students)

SUTVA: No Hidden Variation in Treatment

- SUTVA requires each unit receive the same treatment dosage; this is what it means by “stable” (i.e., notice that the super scripts contain either 0 or 1, not 0.55, 0.27)
- If we are estimating the effect of aspirin on headaches, we assume treatment is 200mg per person in the treatment
- Easy to imagine violations if hospital quality, staffing or even the vents themselves vary across treatment group
- Be careful what we are and are not defining as *the treatment*; you may have to think of it as multiple arms

SUTVA: Scale can affect stability of treatment effects

Easier to imagine this with a different example.

- Let's say we estimate a causal effect of early childhood intervention in Texas
- Now President Biden wants to roll it out for the whole United States – will it have the same effect as we found?
- Scaling up a policy can be challenging to predict if there are rising costs of production
- What if expansion requires hiring lower quality teachers just to make classes?
- That's a general equilibrium effect; we only estimated a partial equilibrium effect (external versus internal validity)

Conclusion

- Potential outcomes are the foundation of causal inference in the "design" tradition
- They have no theoretical apparatus beneath them and thus were a distinctly different approach than ones historically taken in economics using models
- We saw that randomization has a special place because it distributes the mean potential outcomes (and their variance) the same for both groups
- Known and unknown confounders are equally distributed making simple comparisons causal
- Next we look at what we relax this partially with unconfoundedness