Impact Evaluation Methods

Topic 2: Potential Outcomes Framework

Alex Alekseev

May 2, 2025

University of Regensburg, Department of Economics

Previously on *Impact Evaluation Methods*...

- What is causal inference
- Correlation is not causation
- Types of data
- Research design

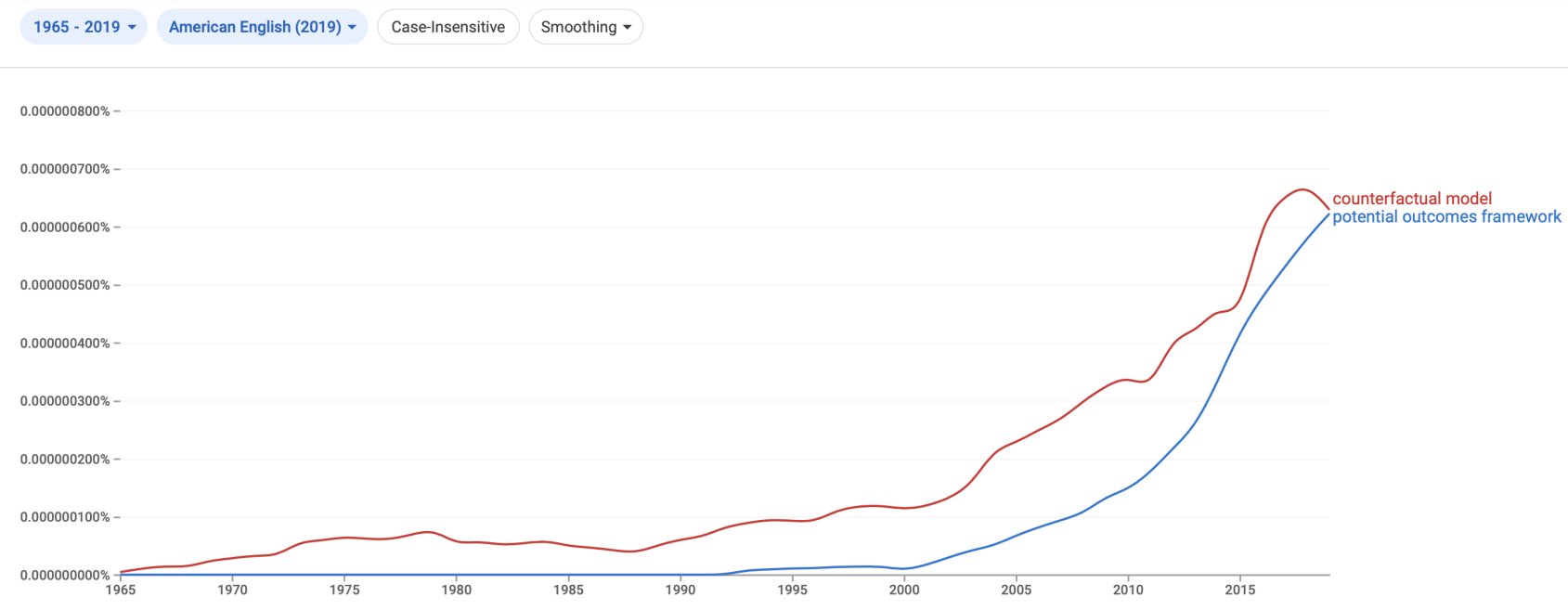
Coding Time: The Sorting Hat

Potential Outcomes Framework

Background

- Origins in early work on experimental design by Neyman (1923), Fisher (1935), Cochran and Cox (1950), Kempthorne (1952), and Cox (1958)
- Donald Rubin (1974) formalized the counterfactual model for causal analysis of observational data
- Also has roots in the economics literature (Roy 1951; Quandt 1972) (the so-called Roy Model), with important subsequent contributions by James Heckman, Charles Manski, and others
- Dominant in both statistics and economics, and it is being used in sociology, psychology, and political science

The Rise of Potential Outcomes Framework



Causal States

- Binary case: two well-defined **causal states** to which all members of the population of interest could be exposed
- Usually called **treatment** and **control**
- In a non-binary case, one can refer to the alternative states as **alternative treatments**

Examples of Causal States

Example	Treatment	Control
job-training	received the job training	did not receive the job training
hospital	went to a hospital	did not go to a hospital
class size	assigned to a small class	assigned to a regular class
human capital	received a college degree	did not receive a college degree

Note

We have to be careful in assuming that the meaning of what constitutes a treatment (job training, hospital, college) is the same for everyone. Suppose, e.g., there are good and bad colleges. Do we really want to mix them together? We will come back to that point later.

Less Clearly Defined Causal States

- The relationship between socioeconomic status and political participation
- Many well-defined causal effects
- E.g., the effect of having obtained at least a college degree on the frequency of voting in local elections
- The effect of having a family income greater than some cutoff value on the amount of money donated to political campaigns
- Not clear that well-defined causal states exist for the general concepts of "socioeconomic status" and "political participation"

Ceteris Paribus Assumption

- When we are defining causal states, we are making a **ceteris paribus** assumption
- In some cases, the assumption might be unrealistic if it rules out other changes that **must** occur at the same time
- Notice a relationship to the ideal experiment from the previous lecture
- If your ceteris paribus assumption cannot hold even in the ideal experiment, your causal relationship of interest is not identified

School and Starting Age

- Whether starting school later (say, at age 7 vs. at age 6) improves a student's academic performance
- We randomize students into the treatment (start at 7) and control groups (start at 6)
- Performance in the first grade: students who start at 7 are likely to get better grades than students who start at 6
- How about we look at the kids of the same age but in different grades?
- Test the treatment group in the first grade and the control group in the second grade (both groups are 7yo)
- But now the control group spent more time in school
- No way to identify the start-age effect separately from the maturation and time-in-school effects

Potential Outcomes

- We can define **potential outcome random variables** over all individuals in the population of interest
- For a binary case: Y^1 (potential outcome in the **treatment** state) and Y^0 (potential outcome in the **control** state)
- y_i^1 is the potential outcome in the treatment state for individual i, and y_i^0 is the potential outcome in the control state for individual i
- The individual-level **causal effect** of the treatment is

$$\delta_i = y_i^1 - y_i^0$$

Note

Individual-level causal effects can be defined in different ways, for example, as the ratio y_i^1/y_i^0. However, the above definition of the individual-level effect can easily accommodate ratios by redefining the outcomes as logarithms.

Causal Exposure Variable

- A **causal exposure** variable, D: D is equal to 1 for members exposed to the treatment state (the treatment group) and equal to 0 for members exposed to the control state (the control group)
- What determines exposure?
  - An individuals's decision to enter one state or another
  - an outside actor's decision
  - a planned random allocation carried out by an investigator
  - some combination of these alternatives
- The random variable D takes on values of $d_i = 1$ for each individual i who is an observed member of the treatment group and $d_i = 0$ for each individual i who is an observed member of the control group

Observed Outcome Variable

- The **observed outcome** variable Y is then

$$Y = \begin{cases} Y^1, D = 1, \\ Y^0, D = 0 \end{cases}$$

- or (this is called the **switching equation**)

$$Y = DY^1 + (1 - D)Y^0$$

The Fundamental Problem of Causal Inference

- One can **never** observe the potential outcome under the treatment state for those observed in the control state
- One can **never** observe the potential outcome under the control state for those observed in the treatment state
- Holland (1986): the **fundamental problem of causal inference**

Group/Potential Outcome	Y^1	Y^0
Treatment $D = 1$	Observable as Y	Counterfactual
Control $D = 0$	Counterfactual	Observable as Y

Treatment Effects

Average Treatment Effect

- The fundamental problem of causal inference does not mean we cannot estimate some average effect
- We will often talk about the **average treatment effect** (ATE) defined as

$$E[\delta] \equiv E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$$

- Equal to the expected value of the what-if difference in outcomes for a randomly selected individual from the population

Note

The expectation is defined with reference to the population of interest. In the training example, the population would be defined as "all adults eligible for training," and eligibility would need to be defined carefully. Thus, to define average causal effects and then interpret estimates of them, one must clearly define the characteristics of the individuals in the assumed population of interest.

Conditional Treatment Effects

- Sometimes we are interested in other treatment effects
- The **average treatment effect for the treated** (ATT) is

$$E[\delta \mid D = 1] = E[Y^1 - Y^0 \mid D = 1] = E[Y^1 \mid D = 1] - E[Y^0 \mid D = 1]$$

- The **average treatment effect for the untreated** (ATU) is

$$E[\delta \mid D = 0] = E[Y^1 - Y^0 \mid D = 0] = E[Y^1 \mid D = 0] - E[Y^0 \mid D = 0]$$

Connection Between Treatment Effects

- If we denote the probability of receiving the treatment as $\pi \equiv P[D = 1]$, then the relationship between the ATE, ATT, and ATU is

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

- For the estimation of the ATE we need five quantities, out of which two are **not observed** (counterfactual outcomes)

The Difference

- What is the conceptual difference between all these effects?
- Recall our opening example
- The ATT was the Gryffindor effect for those who were assigned to Gryffindor by the Sorting Hat
- The ATU was the Gryffindor effect for those who were assigned to Slytherin by the Sorting Hat
- Since half of the students went to Gryffindor, the ATE was exactly zero

College Example

- The ATT is the average effect of college on income of those who **decided to go to college** rather than across all students who could potentially attend college
- The ATE is the expected what-if difference in income that would be observed if we could assign a **randomly selected individual** to both college and no-college treatments
- The ATT is the expected what-if difference in income that would be observed if we could assign a **randomly selected college student** to both college and no-college treatments

Which TE is the Right One?

- For the college example, the ATT is a theoretically important quantity
- If no college effect for college students, unlikely be a college effect for students who typically do not go to college
- If policy interest were focused on whether or not college is beneficial for college students (whether public support of colleges is a reasonable, etc.), then the college effect for college students is the only quantity we would want to estimate
- ATU: if the goal to determine the effect of a potential policy intervention designed to move more students to college

Naive Treatment Effect

Naive Treatment Effect

- What happens if go ahead and try to estimate the **naive treatment effect** (NTE) using the data that we observe?
- The NTE is

$$NTE \equiv E[Y^1 \mid D = 1] - E[Y^0 \mid D = 0].$$

- This is the effect that we can estimate from the data, since it does not involve any counterfactuals
- Does it correspond to any of the treatment effects of interest that we introduced?

Math Time: Decomposition of the NTE

$$
\begin{aligned}
NTE &= E[Y^1 \mid D = 1] - E[Y^0 \mid D = 0] \\
&= \underbrace{E[Y^1 \mid D = 1] - E[Y^0 \mid D = 1]}_{ATT} + E[Y^0 \mid D = 1] - E[Y^0 \mid D = 0]
\end{aligned}
$$

Now let's rewrite the ATT as

$$
\begin{aligned}
ATT &= \pi ATT + (1 - \pi) ATT \\
&= \underbrace{\pi ATT + (1 - \pi) ATU}_{ATE} - (1 - \pi) ATU + (1 - \pi) ATT \\
&= ATE + (1 - \pi)(ATT - ATU).
\end{aligned}
$$

Finally, we can re-write the NTE as

$$
NTE = ATE + \underbrace{E[Y^0 \mid D = 1] - E[Y^0 \mid D = 0]}_{\text{selection bias}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{diff treatment eff bias}}
$$

Math Time: Decomposition of the NTE

- There are two equivalent formulations
- The first starts with the ATT

$$NTE = ATE + \underbrace{E[Y^0 \mid D = 1] - E[Y^0 \mid D = 0]}_{\text{selection bias}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{diff TE bias}}$$

- The second starts with the ATU

$$NTE = ATE + \underbrace{E[Y^1 \mid D = 1] - E[Y^1 \mid D = 0]}_{\text{selection bias}} + \underbrace{\pi(ATU - ATT)}_{\text{diff TE bias}}$$

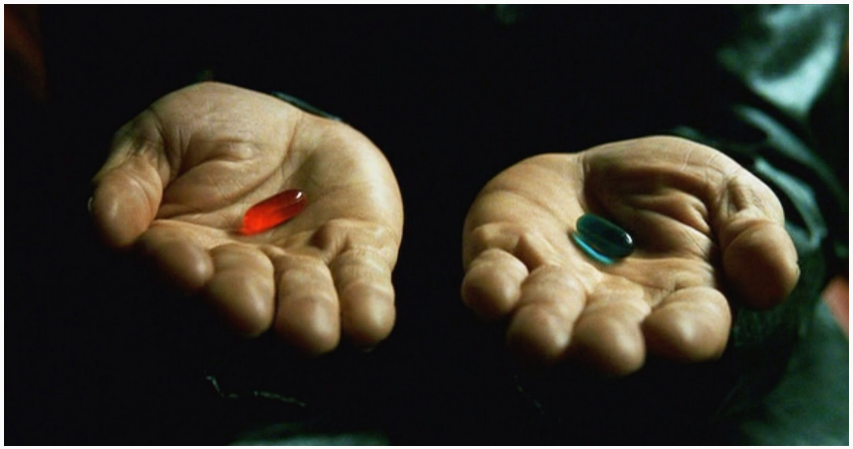
- Unless no biases are present, the NTE will not be equal to the ATE

Homework

Prove that

$$NTE = ATE + E[Y^1 \mid D = 1] - E[Y^1 \mid D = 0] + \pi(ATU - ATT)$$

Example: The Matrix



Group	$E[Y^1 \mid D]$	$E[Y^0 \mid D]$	$E[\delta \mid D]$	$P(D)$
Red Pill ($D = 1$)	8	2	6	0.2
Blue Pill ($D = 0$)	3	5	-2	0.8

- The NTE is $8 - 5 = 3$
- The ATE is $0.2 \times 6 - 0.8 \times 2 = 1.2 - 1.6 = -0.4$
- Then we can decompose the NTE as

$$NTE = -0.4 + (2 - 5) + 0.8(6 + 2) = -0.4 - 3 + 6.4 = 3$$

$$NTE = -0.4 + (8 - 3) + 0.2(-2 - 6) = -0.4 + 5 - 1.6 = 3$$

Example: The Sorting Hat

Group	$E[Y^1 \mid D]$	$E[Y^0 \mid D]$	$E[\delta \mid D]$	$P(D)$
Gryffindor ($D = 1$)	38.6	21	17.6	0.5
Slytherin ($D = 0$)	13.4	31	-17.6	0.5

- The NTE is $38.6 - 31 = 7.6$
- The ATE is $0.5 \times 17.6 - 0.5 \times 17.6 = 0$
- Then we can decompose the NTE as

$$NTE = 0 + (21 - 31) + 0.5(17.6 + 17.6) = -10 + 17.6 = 7.6$$

$$NTE = 0 + (38.6 - 13.4) + 0.5(-17.6 - 17.6) = 25.2 - 17.6 = 7.6$$

Treatment Assignment and Randomization

Randomization Magic

- In the observational studies the naive way of estimating the ATE does not in general give us the ATE
- The NTE is **biased**
- Let's go back to the experimental ideal and see where the magic happens

Independence

- The main feature of a randomized experiment: treatment assignment is done randomly, **independent of potential outcomes**
- The formal way of writing this independence feature is

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

- D is jointly independent of all functions of the potential outcomes
- This assumption implies that

$$E[Y^1 \mid D = 1] = E[Y^1 \mid D = 0]$$
$$E[Y^0 \mid D = 1] = E[Y^0 \mid D = 0]$$

- These equalities automatically kill the **selection bias**

Differential Treatment Effect Bias

- How about the **differential treatment effect bias**?
- Notice that

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

- This, in turn, implies that

$$ATE = \pi ATT + (1 - \pi)ATU = ATT = ATU$$

- In a randomized experiment a **simple difference in mean outcomes** is an unbiased estimator of the ATE

Note

The independence assumption does **not** imply that $E[Y^1 \mid D = 1] = E[Y^0 \mid D = 1]$ or that $E[Y^1 \mid D = 0] = E[Y^0 \mid D = 0]$. In other words, exposure is independent of potential outcomes but observed outcomes and exposure can be related

Back to Real Life

- In observational data, the independence assumption typically does not hold

Rosenbaum (2002)

An observational study is an empiric investigation of treatments, policies, or exposures and the effects they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects

Treatment Selection Mechanism

- The first step in the analysis of observational data is to investigate the **treatment selection mechanism**
- Notice the switch in language from **assignment** to **selection**
- Some of the process by which individuals select alternative treatments can be examined empirically
- A full accounting of treatment selection is sometimes impossible
- E.g., subjects are motivated to select on the causal effect itself and a researcher does not have a valid measure of their expectations

Weaker Assumptions

- While the independence assumptions will almost surely not hold in the observational data, some **weaker** assumptions can be more defendable
- For example, consider the following two assumptions:

$$\text{Assumption 1: } E[Y^1 \mid D = 1] = E[Y^1 \mid D = 0]$$
$$\text{Assumption 2: } E[Y^0 \mid D = 1] = E[Y^0 \mid D = 0]$$

- Can the naive estimate say anything useful about our treatment effects of interest if at least of these assumptions holds?

Math Time: Assumptions

- If only Assumption 1 holds, then the decomposition implies that

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

- If only Assumption 2 holds, then the decomposition implies that

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

Assumptions in Practice

- We might have a good theoretical reason to believe that
  - Assumption 2 is valid because those in the treatment group would, on average, do no better or no worse under the control than those in the control group
  - Assumption 1 is invalid because those in the control group would not do nearly as well under the treatment as those in the treatment group.
- Under this scenario, the naive estimator will deliver an unbiased and consistent estimate of the ATT
- But it is still biased and inconsistent for **both** the ATU and the ATE
- With observation data, we need to try to **condition** or **stratify** data or find **some features of the data** that allow us to defend either of these assumptions

Regression Interpretation

Switching Equation as a Regression

- Let's write the **switching equation** as

$$\begin{aligned} Y &= Y^0 + (Y^1 - Y^0)D \\ &= E[Y^0] + (Y^1 - Y^0)D + Y^0 - E[Y^0] \end{aligned}$$

- Suppose the individual treatment effects are the same for everyone, $\delta = Y^1 - Y^0$ is a constant ($ATE = ATT = ATU = \delta$)
- Now denote $\alpha \equiv E[Y^0]$ and $\epsilon \equiv Y^0 - E[Y^0]$
- The switching equation becomes

$$Y = \alpha + \delta D + \epsilon$$

- Looks like a **regression equation**, where α is the constant, δ is the coefficient of interest, D is the independent variable, and ϵ is the error term

Selection Bias

- Consider the conditional expectations of Y

$$E[Y \mid D = 1] = \alpha + \delta + E[\epsilon \mid D = 1]$$

- and

$$E[Y \mid D = 0] = \alpha + E[\epsilon \mid D = 0].$$

- Then the difference between the two is

$$E[Y \mid D = 1] - E[Y \mid D = 0] = \delta + E[\epsilon \mid D = 1] - E[\epsilon \mid D = 0].$$

- As long as D is **independent of the error term**, the difference on the left identifies the ATE (=ATT=ATU)
- If the error term and D are correlated, the estimate of δ will be biased

SUTVA

SUTVA

- The stable unit treatment value assumption or **SUTVA** (Rubin 1980b, 1986)
- In economics: no-macro-effect or partial equilibrium assumption (Heckman 2000, 2005)
- SUTVA requires that the potential outcomes of individuals be unaffected by the treatment exposures of other individuals

Rubin (1986)

SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive.

General Formulation

- Generally, suppose that d is an $N \times 1$ vector of treatment indicator variables for N individuals
- The outcome for individual i under the treatment is $y_i^1(d)$, and the outcome for individual i under the control is $y_i^0(d)$
- The individual-level causal effect for individual i is $\delta(d)$
- SUTVA: $y_i^1(d) = y_i^1$, $y_i^0(d) = y_i^0$, and $\delta_i = \delta_i(d)$

Implications

SUTVA implies three things

- each individual within a treatment group receives the **same treatment**
- **no spillovers** to other individuals' potential outcomes when an individual is exposed to some treatment
- **no general equilibrium** effects

Violations of SUTVA

- When does SUTVA break down?
- A typical example is **peer effects**
- Your peer's exposure to the treatment might affect your potential outcomes
- Suppose we conduct a randomized experiment on the effect of a new textbook on academic performance in high school
- If students from the treatment group interact with the students from the control group and share the new textbook, then we might find a zero treatment effect of the new textbook

College Example

- For SUTVA to hold, the college effect cannot be a function of the number (and/or composition) of students who enroll into college
- College effect would change if large numbers of non-college educated students entered college
- It may be that we can estimate the causal effect of college only for those who would typically choose to attend college, but also subject to the constraint that the proportion of students in college remain relatively constant
- It may be impossible to determine from any data what the college effect on income would be under a new distribution of students that would result from a large and effective policy intervention

What To Do If SUTVA Is Violated

- Certain types of marginal effect estimates can usually still be defended
- Estimates of average causal effects hold only for what-if movements of a very small number of individuals from one hypothetical treatment state to another
- If more extensive what-if contrasts are of interest (widespread intervention), then SUTVA would need to be dropped
- Variation of the causal effect as a function of treatment assignment patterns would need to be modeled explicitly

Next Time on *Impact Evaluation Methods*...

Causal diagrams (aka directed acyclical graphs)