Causal Inference

Author: Wenxiao Yang

Institute: Haas School of Business, University of California Berkeley

Date: 2025

All models are wrong, but some are useful.

Contents

Chapter 1 Causal Inference

The fundamental problem of causal inference:

(a). Never see the same person treated and untreated

(b). Missing data problem

(c). "Solve" by finding a comparison group

Definition 1.1 (Notations and Estimands)

- Treatment: $T \in \{0, 1\}$
- Potential Outcome with treatment $Y(1), Y(0)$
- Other Variable X
- Individual Treatment Effect (ITE) $= Y_i(1) - Y_i(0)$
- Conditional Average Treatment Effect (CATE) $= \mathbb{E}[Y(1) - Y(0)|X = x] := \tau(x)$
- Average Treatment Effect (ATE) $= \mathbb{E}[Y(1) - Y(0)] := \tau$
- Average Treatment Effects on Treated (ATT) $= \mathbb{E}[Y(1) - Y(0) \mid T = 1]$

Difference in Means

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - T_i)$$

By the Law of Large Numbers,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i &= \lim_{n\to\infty} \frac{n}{n_i}\frac{1}{n}\sum_{i=1}^{n} Y_i T_i \\
&= (P[T = 1])^{-1}\,\mathbb{E}[YT] \\
&= (P[T = 1])^{-1}\,\mathbb{E}[YT \mid T = 1]P[T = 1] \\
&= \mathbb{E}[YT \mid T = 1] \\
\bar{Y}_1 &\xrightarrow{P} \mathbb{E}[YT \mid T = 1]
\end{aligned}$$

Causal Effect

Assumption 1.1

(1). SUTVA: Only your treatment matters;

(2). Consistency: Observed outcome matches treatment "assignment": $Y = TY(1) + (1 - T)Y(0)$.

Only yields $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0 \xrightarrow{P} \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]$

$$\begin{aligned}
&\mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] \\
=&\underbrace{\mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]}_{\text{selection bias}}
\end{aligned}$$

To get the ATT (eliminate the selection bias), we need exclusion/independence: Randomization.

Assume $Y(t) = \mu(t) + \epsilon_t$ (SUTVA). Consider the consistency assumption:

$$\begin{aligned}
Y &= TY(1) + (1 - T)Y(0) \\
&= Y(0) + T(Y(1) - Y(0)) \\
&= \underbrace{\mu_0}_{\alpha} + T\underbrace{(\mu_1 - \mu_0)}_{\beta^T} + \underbrace{\epsilon_0 + T(\epsilon_1 - \epsilon_0)}_{\epsilon}
\end{aligned}$$

Consider the covariate between T and X. Why important?

$$\begin{aligned}
&\mathbb{E}[Y|X = 1, T = 1] - \mathbb{E}[Y|X = 1, T = 0] \\
=&\underbrace{\mathbb{E}[Y(1)|X(1) = 1] - \mathbb{E}[Y(0)|X(1) = 1]}_{\text{ATE}|X(1)=1} + \underbrace{\mathbb{E}[Y(0)|X(1) = 1] - \mathbb{E}[Y(0)|X(0) = 1]}_{\text{selection bias}}
\end{aligned}$$

$Y(t) = \mu(t, X) + \epsilon_t$. Then,

$$\begin{aligned}
Y &= Y(1)T + Y(0)(1 - T) \\
&= \underbrace{\mu(0, X)}_{\alpha(X)} + T\underbrace{(\mu_1(X) - \mu_0(X))}_{\beta(X)} + \epsilon
\end{aligned}$$

Chapter 2 STAT 256

2.1 Lecture Zero: Causal Inference vs Association

Amanda Coston

2.1.1 Causal Inference on the Effect of Red Wine on Health

2.1.2 Is red wine good for heart health?

For many years consensus held that red wine improved cardiovascular health. The evidence behind this was largely from studies on people of drinking age that compared the health outcomes of those who self-reported that they drank wine to those who self-reported that they did not. Findings under such a design showed that people who drink red wine have better cardiovascular outcomes than those who don't drink alcohol.

What are potential problems with this study design? You may be thinking that people who drink wine systematically differ from those who don't in ways that matter for health outcomes. In fact, sociologist Kaye Middleton Fillmore showed that the statistical significance of these findings hinged on the inclusion of *previous* drinkers in the "non-drinkers" category. That is, some people who said they did not drink red wine previously drank alcohol. Problematically, a common reason these people gave up alcohol was poor health. Therefore the definition of "non-drinkers" selected for people who had poorer health outcomes. Fillmore showed that redefining "non-drinkers" to be "never-drinkers" eliminated any supposed advantage of drinking wine.

The debate here was one of causal inference – looking at the cause and effect of red wine on health.

For more details, see [Fillmore et al.(2007)].

2.1.3 Syllabus

See course syllabi here: https://stat156.berkeley.edu/fall-2024/syllabus.html.

2.1.4 Association vs Causation

Association is the focus of much of statistics but in causal inference our focus is, of course, on causation! As a starting point, we will today consider common measures of association and discuss why they may not capture causation. In the next lecture we will see causal analogues of these measures. aaa We first consider the setting where the outcome Y and treatment Z are both binary.

Definition 2.1 (Risk Difference)

The associative risk difference (RD) is $E[Y \mid Z = 1] - E[Y \mid Z = 0]$.

Definition 2.2 (Risk Ratio)

The associative risk ratio (RR) is $P[Y = 1 \mid Z = 1]/P[Y = 1 \mid Z = 0]$.

Definition 2.3 (Odds Ratio)

The associative odds ratio (OR) is $\frac{P[Y=1|Z=1]}{P[Y=0|Z=1]} / \frac{P[Y=1|Z=0]}{P[Y=0|Z=0]}$.

Which measure one chooses depends on their particular setting – what question they are interested in and what data they have available (odds ratios can be estimated in with outcome-dependent sampling whereas the risk difference and risk ratio generally cannot). The measures are related to each other as follows:

1. $Z \perp Y \iff \text{RD} = 0 \iff \text{RR} = 1 \iff \text{OR} = 1$
2. $\text{RD} > 0 \iff \text{RR} > 1 \iff \text{OR} > 1$ assuming all conditional probabilities are non-zero.
3. $\text{RR} \approx \text{OR}$ when $P(Y = 1)$ is small.

Next we consider measures of association that can accommodate non-binary outcomes.

2.1.5 Correlation and linear regression

Suppose we are now interested in the outcome blood pressure as a measure of cardiovascular health. A natural starting point is to model the relationship between blood pressure (Y) and whether one drinks red wine (Z) as

$$Y = \beta Z + \alpha + \epsilon$$

where $E[\epsilon] = 0$ and $E[\epsilon Z] = 0$.

Recall that we can relate β to the Pearson correlation coefficient ρ as follows

$$\beta = \rho \frac{\text{var}(Z)}{\text{var}(Y)}.$$

The coefficient β describes the change in Y associated with whether one drinks red wine. More generally, the coefficient β describes the change in Y associated with one unit increase in Z. Sometimes people refer to β as the "effect" of Z on Y but this is generally misleading (without further assumptions). We have simply modeled an associative relationship; we can't claim anything causal yet! Next time we will introduce a new language, potential outcomes, so that we can make causal claims.

2.2 Lecture One: Potential Outcomes Framework

Aryan Shafat, Frederik Stihler, Mika Lee (Revisions)

2.2.1 Review of last lecture

Review of Risk Difference, Risk Ratio and Odds Ratio for setting where the outcome Y and treatment Z are both binary. (See lecture notes of Lecture 0 for details)

Explanation of the last equivalence in the following statement:

$$\text{RD} > 0 \iff \text{RR} > 1 \iff \text{OR} > 1 \text{ assuming all conditional probabilities are non-zero}$$

We know that:

$$\mathbb{P}[Y = 1 \mid Z = 1] + \mathbb{P}[Y = 0 \mid Z = 1] = 1 \text{ and } \mathbb{P}[Y = 1 \mid Z = 0] + \mathbb{P}[Y = 0 \mid Z = 0] = 1.$$

$\text{RR} > 1$ implies that $\mathbb{P}[Y = 1 \mid Z = 1] > \mathbb{P}[Y = 1 \mid Z = 0]$, and hence $\mathbb{P}[Y = 0 \mid Z = 0] > \mathbb{P}[Y = 0 \mid Z = 1]$. This leads us to $\text{OR} > 1$.

2.2.2 Potential Outcome Framework

2.2.2.1 Types of questions we are interested in

1. Does journaling reduce the risk of depression?
2. Do personalized AI tutors improve a student's grade?
3. Does smoking cigarettes cause cancer?

We can summarize these questions in the following table:

	Z	Y
1	Journal	Depression
2	AI tutor	Grade
3	Cigarettes	Cancer

For all the examples, we want to form a causal estimand using potential outcomes. The potential outcome framework was developed by [Neyman(1923)] in 1923 over one hundred years ago and later revitalized and repopularized by [Rubin(1980)] in 1980.

2.2.2.2 Potential Outcomes

We are interested in potential (hypothetical) outcomes when we are thinking about causal questions.

Definition 2.4 (Potential Outcome)

The potential outcome Y is a function of a particular treatment value $Y(Z = z)$.

E.g. $Y(Z = 1) = Y(1)$ (the potential outcome under the intervention that assigns treatment) vs. $Y(Z = 0) = Y(0)$ (the potential outcome under the intervention that assigns no treatment).

In our particular examples, these quantities are represented by the following hypothetical questions:

(1) Would someone have depression if they journalled $\approx Y(1)$

(2) Would a student get a particular grade if they had an AI tutor $\approx Y(1)$

(3) Would someone get cancer if they smoked cigarettes $\approx Y(1)$

2.2.2.3 Causal Estimands

Next we define the causal versions of our measures of association (called causal estimands).

Definition 2.5 (Causal Risk Difference)

The causal risk difference (on a population level) is $E[Y(1) - Y(0)]$.

Definition 2.6 (Causal Risk Ratio)

The causal risk ratio is $\frac{E[Y(1)]}{E[Y(0)]} = \frac{P(Y(1)=1)}{P(Y(0)=1)}$.

Definition 2.7 (Causal Odds Ratio)

The causal odds ratio is $\frac{P(Y(1)=1)}{P(Y(1)=0)} / \frac{P(Y(0)=1)}{P(Y(0)=0)} = \frac{E[Y(1)]}{E[1-Y(1)]} / \frac{E[Y(0)]}{E[1-Y(0)]}$.

2.2.2.4 Hidden Assumptions

As mentioned, [Rubin(1980)] repopularized this framework by clarifying some important hidden assumptions:

Assumption 2.1 (Consistency)

The treatment levels are well-defined (there are no other versions of the treatment).

Assumption 2.2 (No Interference)

The treatment assigned to other units does not affect the potential outcomes for unit i (no spillover).

These 2 assumptions are called **Stable Unit Treatment Value Assumption (SUTVA).**

For example, our 3rd question/example about cigarette smoking violates both assumptions.

The question isn't well-defined (does smoking entail smoking 1 cigarette a day or smoking a whole pack a day) and thus violates Assumption 1.

It also violates Assumption 2, through non-smokers who might end up passively smoking by being around smokers.

2.2.3 Causal Estimands

Causal Effects are functions of potential outcomes. For example, the causal risk difference is essentially the 'average treatment effect'.

- Unit 'i' has 2 potential outcomes: $Y_i(1)$ and $Y_i(0)$
- **Individualized treatment effect:** $Y_i(1)$ - $Y_i(0)$

Fundamental Problem of Causal Inference (1986 Holland): Never observe both potential outcomes.

2.2.3.1 Add a time-element

We could include a time element - write into our personal journal/smoke a cigarette/receive treatment one day and then go without treatment the next day, to see the 'causal effect' of the treatment.

However, the effect of the treatment might be long-lasting. Thus, one could 'experience'/observe their causal effects even on the days without treatment. Additionally, say for a time-period of 2 days, we actually end up having 4 potential outcomes:

- $Y_{i,day1}(1)$ vs $Y_{i,day1}(0)$
- $Y_{i,day2}(1)$ vs $Y_{i,day2}(0)$

Going back to the potential outcomes framework, we basically end up 'observing' one of the potential outcomes framework.

- The factual/observed outcome is, $Y_i = \begin{cases} Y_i(1) & \text{if } Z_i = 1 \\ Y_i(0) & \text{if } Z_i = 0 \end{cases}$

Equivalently,

$Y_i = Z_i * Y_i(1) + (1 - Z_i) * Y_i(0)$

- The (unobserved) counterfactual, or missing potential outcome is given by:

  $Y_i^{\text{mis}} = Z_i * Y_i(0) + (1 - Z_i) * Y_i(1)$

2.2.4 Simpson's Paradox

Based on the class poll, we saw that the effect of the hint wasn't that strong. There was a confounding variable (row number) that was dampening the effect of the hint (since most of the hints were given to people in the rows

at the back). This was attributed to **Simpson's Paradox**, which might have made it seem like the hints had an **effect reversal** (as if those who got the hints actually ended up doing worse on the poll).

This is mathematically shown as: $\mathbb{P}(Y = 1|Z = 0) > \mathbb{P}(Y = 1|Z = 1)$. However, we can easily counter this by conditioning on the confounding variable (X):

$$\mathbb{P}(Y = 1|Z = 0, X = x) < \mathbb{P}(Y = 1|Z = 1, X = x) \quad \forall x \in X$$

2.2.4.1 Sources of the paradox:

- Confounding variables/factors
- Non-collapsibility

2.3 Lecture Three: Randomized Experiments

Daisy Wang & Mika Lee (Revisions)

2.3.1 Last Lecture: Simpson's Paradox

Simpson's paradox is when the data may originally appear to have one trend, but not when grouped. In stats terms:

$$\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] > 0$$

but is actually

$$\mathbb{E}[Y \mid Z = 1, X = x] - \mathbb{E}[Y \mid Z = 0, X = x] < 0$$

when we condition on the confounding variable X. This is caused by confounding and non-collapsibility.

Note on notation: Potential outcomes $Y(Z = 1) = Y(1)$ and in Hernan, $Y^{Z=1} = Y^1$

2.3.2 Randomized Experiments

Why is randomization so powerful?

- ***Ignorable*** treatment assignment
- Groups are ***exchangeable***, in that groups could swap assignments and still have the same result

2.3.2.1 Exchangeability

$\mathbf{Pr}\, \boldsymbol{Y(1) = 1 \mid Z = 1} = \Pr Y(1) = 1 \mid Z = 0$

$\Pr Y(0) = 1 \mid Z = 1 = \Pr Y(0) = 1 \mid Z = 0 = \mathbf{Pr(Y = 1 \mid Z = 0)}$

where the two bolded equations are identifiable from the data.

Additionally $Y(Z) \perp Z \; \forall z = 0, 1$ which means treatment assignment is independent of potential outcome. In general, treatment assignment is exchangeable which implies ignorable which implied exogenous.

In an **ideal random experiment**, association would be equal to causation:

$$\mathbb{E}[Y \mid Z = 1] = \mathbb{E}[Y(1) \mid Z = 1]$$

Additionally:

$$\text{Risk Difference } \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] = \text{Causal Risk Difference } \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

In tandem with the previous part about the independence of random treatment assignment, it is important to note that $Z \perp Y(Z) \neq Z \perp Y$, in which the relationship on the left side of the inequality contains random treatment assignment and the one on the right side references no treatment assignments.

Finally a **randomized experiment** is where treatments are assigned in a known and probabilistic ("random") manner i.e. Bernoulli random experiments.

2.3.2.2 Completely Randomized Experiment

We denote those who get the treatment as n_1 , and those who get the control as $n_0 = n - n_1$. The treatment assignment is denoted as $\mathbb{Z} = (z_1, ...z_n)$. Then the probability that $\Pr(\mathbb{Z} = z) = \frac{1}{\binom{n}{n_1}}$ where $\mathbb{Z}$ is such that $\sum_{i=1}^{n} z_i = n_1$. Also keep in mind that $\mathbb{Y}(1)$ and $\mathbb{Y}(0)$ are fixed in this situation.

2.3.2.3 Fisher (1935) Fisher's Sharp Null

$H_0 : Y_i(1) = Y_i(0)$ for all $i = 1, ...n$ where $\mathbb{Y} = \mathbb{Y}(1) = \mathbb{Y}(0)$ and the test statistic is $T(Z, \mathbb{Y})$. Here the Z is random and $\mathbb{Y}$ is fixed. This is also known as Neyman's Null Hypothesis (Weak Null).

Under the null, $\{T(z^1, \mathbb{Y})...T(z^{\binom{n}{n_1}}, \mathbb{Y}\}$ is uniform, representing the randomization distribution. Thus we can find the p-value:

$$p = \frac{1}{\binom{n}{n_1}} \sum_{m=1}^{\binom{n}{n_1}} \mathbf{1}\{T(z^m, y) \geq T(Z, Y)\}$$

We can also use the Monte Carlo method to approximate the p-value:

$$\frac{1}{R} \sum_{r=1}^{R} \mathbb{1}\{T(z^r, y) \geq T(Z, Y)\}$$

where $T(z^r, y)$ is a particular fixed value and $T(Z, Y)$ is an observed random variable. All of this is the **Fisher's randomization test**, aka **permutation test**.

2.3.2.4 Choices for the Test-Statistic

2.3.2.4.1 Difference-in-means statistic $\hat{\tau} = \hat{\bar{Y}}(1) - \hat{\bar{Y}}(0)$ where $\hat{\bar{Y}}(1) = \frac{1}{n_1}\sum_{z_i=1} Y_i = \frac{1}{n_1}\sum_{i=1} z_i Y_i$

Note: this statistic is easily ruined by outliers

2.3.2.4.2 Wilcoxon Rank Sum Unlike the difference-in-means statistic, the Wilcoxon rank sum test statistic is robust to outliers. This is because it is defined as follows:

$$R_i : \text{the rank of } Y_i = \#\{j : Y_j \leq Y_i\}, \text{ and so the test statistic itself is } W = \sum_{i=1}^{n} z_i R_i$$

The test statistic is more broad and may miss distributional differences. It can also be viewed as the difference-in-means of the rank under treatment vs control.

2.3.2.4.3 Kolmogorov-Smirnov Statistic The empirical CDF of treated units is $\hat{F}_1(y) = \frac{1}{n_1}\sum_{i=1}^{n} z_1 \mathbf{1}\{Y_i \leq y\}$ and the control is $\hat{F}_0(y) = \frac{1}{n_0}\sum_{i=1}^{n}(1 - z_1)\mathbf{1}\{Y_i \leq y\}$. The test statistic is then

$$D = \max_y |\hat{F}_1(y) - \hat{F}_0(y)|$$

2.4 Lecture Four: Continue on Randomized Experiments

Yulin Zhang & Aadya Agarwal

2.4.1 Last Lecture: Complete Randomized Experiments

- Complete randomized experiment
  - *terms*: let n denotes the total number of test units, n_1 denotes the number of units get treated, $n_0 = 1 - n_1$ denotes the number of units get control.
- exchangability: $y(z) \perp z$, where z denotes treatment, $y(z)$ denotes the outcome of the treatment.
- Fisher's sharp null

2.4.2 Neymanian estimation & inference

2.4.2.1 Finite population counterfactual (potential) quantities

Let $y_i(1)$ and $y_i(0)$ denote the potential outcomes of the ith unit under treatment and control. Define the *individual casual effect* of the ith unit as

$$\tau_i = y_i(1) - y_i(0)$$

Theoretical population level statistics:

$$\bar{y}(1) = \frac{1}{n}\sum_{i=1}^{n} y_i(1)$$

$$\bar{y}(0) = \frac{1}{n}\sum_{i=1}^{n} y_i(0)$$

$$S^2(1) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i(1) - \bar{y}(1))^2$$

$$S^2(0) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i(0) - \bar{y}(0))^2$$

$$S(1,0) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i(1) - \bar{y}(1))(y_i(0) - \bar{y}(0))$$

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \bar{y}(1) - \bar{y}(0)$$

$$S^2(\tau) = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \tau)^2$$

Lemma: $2S(1,0) = S^2(1) + S^2(0) - S^2(\tau)$

In the data:

$$\hat{\bar{y}}(1) = \frac{1}{n_1}\sum_{i=1}^{n} z_i y_i$$

$$\hat{\bar{y}}(0) = \frac{1}{n_0}\sum_{i=1}^{n}(1 - z_i) y_i$$

$$\hat{S}^2(1) = \frac{1}{n_1 - 1}\sum_{i=1}^{n} z_i(y_i - \hat{\bar{y}}(1))^2$$

$$\hat{S}^2(0) = \frac{1}{n_0 - 1}\sum_{i=1}^{n}(1 - z_i)(y_i - \hat{\bar{y}}(0))^2$$

Note, since we could not observe $y_i(1)$ and $y_i(0)$ at the same time, no other theoretical statistics could be calculated at this point.

2.4.2.2 Neyman (1993) theorm

- $\hat{\tau} = \hat{\bar{y}}(1) - \hat{\bar{y}}(0)$ satisfies:
  - unbiasedness: $\mathbb{E}(\hat{\tau}) = \tau$
  - has variance:
$$\mathbb{V}(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n} = \frac{n_0}{n_1 n} S^2(1) + \frac{n_1}{n_0 n} S^2(0) + \frac{2}{n} S(1,0)$$
  - Note in variance terms, $S^2(\tau)$ and $S(1,0)$ are not identifiable from the data, this variance could not be calculated from the data directly.
- Neyman's variance estimator:
$$\hat{\mathbb{V}}(\hat{\tau}) = \frac{\hat{S^2}(1)}{n_1} + \frac{\hat{S^2}(0)}{n_0}$$
  - Note: this is the usual variance estimator under the assumption that the population is infinite and random samples. However, this is not the case we have here: the population is finite and we do not have random samples (treatment and controls are usually mutually exclusive).
  - In the paper, Neyman proved that
$$\mathbb{E}(\hat{V}) - \mathbb{V}(\hat{\tau}) = \frac{S^2(\tau)}{n} \geq 0$$
    where equality holds when $\tau_i = \tau \; \forall i$. This shows that the estimator above is conservative for estimating the variance of $\hat{\tau}$. People call this a "minor miracle that the two mistakes cancel" (Freeman, Pisani & Purues, 2006)
- Regression analysis of the complete randomized experiment

  In OLS settings, we could get $(\hat{\alpha}, \hat{\beta}) = argmin_{a,b} \sum_{i=1}^{n} (y_i - a - bz_i)^2$, where $\hat{\beta} = \hat{\tau}$
  - Usual OLS variance: $\hat{\mathbb{V}}_{OLS}(\hat{\beta}) = \hat{\mathbb{V}}_{OLS}(\hat{\tau}) \approx \frac{\hat{S^2}(1)}{n_0} + \frac{\hat{S^2}(0)}{n_1}$, note this is not equal to the variance we want.
  - Eicker-Huber-White (EHW) Robust Variance Estimator: $\hat{\mathbb{V}}_{EWH}(\hat{\tau}) \approx \frac{\hat{S^2}(1)}{n_1} + \frac{\hat{S^2}(0)}{n_0}$
  - HC2 variant of EHW $= \frac{\hat{S^2}(1)}{n_1} + \frac{\hat{S^2}(0)}{n_0}$

2.4.3 Stratified (Conditional) Random Experiments

2.4.3.1 A potential problem with completely randomized experiment

Bias could be introduced if the different covariate strata have non-random treatments assigned.

- Let $x_i \in \{1..k\}$ denote a covariate that we are interested in.
- Covariate imbalance: bias would be introduced if the proportion of units in stratum k are different across treatment and control groups.

2.4.3.2 Definition

Let $n_{k,1} = \#\{i : x_i = k, z_i = 1\}$, $n_{k,0} = \#\{i : x_i = k, z_i = 0\}$. $n_k = \#\{i : x_i = k\}$

Stratified random experiment (SRE) is to run k independent complete random experiments (CRE) within k strata of discrete covariate x for fixed $n_{k,1}$, $n_{k,0.}$

We could view stratum as block, stratified randomization could also be called block randomization.

$$\tau_k = \frac{1}{n_k} \sum_{x_i=k} \tau_i$$

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \tau_i = \frac{1}{n} \sum_{k=1}^{k} \sum_{x_i=k} \tau_i = \sum_{k=1}^{k} \pi_k \tau_k, \pi_k = \frac{n_k}{n}$$

Comparing SRE and CRE:

- Number of potential assignments: $SRE : \prod_{k=1}^{k} \binom{n_k}{n_{k,1}} < CRE : \binom{n}{n_1}$
- Propensity score (propensity of getting treatment): $e_k = \frac{n_{k,1}}{n_k}$. For SRE, e_k is fixed; for CRE, e_k is random.

2.4.3.3 Fisher Randomization Test

$H_0 : Y_i(0) = Y_i(1)$ for all i

Run Fisher Randomization test by permuting the treatment indicators within each strata X according to the fixed values of treatment assigned for each of those strata

This process is called the **Conditional Randomization Test or Conditional Permutation test**

Test Statistic Choice:

1. Stratified Estimator

$\hat{\tau}_s = \sum_{k=1}^{K} \pi_k \hat{\tau}_k$

2. Combined Wilcoxon-Rank Sum Statistic

W_k : Wilcoxon rank-sum statistic in stratum k

$W_s = \sum_{k=1}^{k} C_k W_k$

$C_k = \frac{1}{n_{k,1} n_{k,0}}$ OR $C_k = \frac{1}{n_k + 1}$

Note: Works well only if k is small

3. Aligned Rank Statistic (Hodges and Lehmann)

$\tilde{Y}_i = Y_i - \bar{Y}_k$ where $Y_k = \frac{1}{n-k} \sum_{X_i = k} Y_k$ (stratum specific mean)

Let $\tilde{Y}_i$ be the rank of $\tilde{Y}_i$

$\tilde{W} = \sum_{i=1}^{n} Z_i \tilde{R}_i$

2.4.3.4 Neymanian Inference for SRE

$\mathbb{E}[\hat{\tau}_k] - \tau_k = 0$

$\hat{\tau}_k$ has variance: Var($\hat{\tau}_k$) = $\frac{S_k^2(1)}{n_{k,1}} + \frac{S_k^2(0)}{n_{k,0}} - \frac{S_k^2(\tau)}{n_k}$

where $S_k^2(1), S_k^2(0), S_k^2(\tau)$ are the stratum specific analogs (stratum specific variance of potential outcomes and causal effects)

We can combine these to get a stratified estimator: $\hat{\tau}_s = \sum_{k=1}^{k} \pi_k \hat{\tau}_k$ where $\pi_k = \frac{n_k}{n}$ and

- $\mathbb{E}[\hat{\tau}_s] - \tau = 0$ (unbiased)
- has variance var($\hat{\tau}_s$) = $\sum_{k=1}^{k} \pi_k^2 var(\hat{\tau}_k)$

If $n_k, 1 >= 2$ and $n_k, 0 >= 2$, we can compute $\hat{S}_k^2(1)$, $\hat{S}_k^2(0)$

$\hat{V}_s = \sum_{k=1}^{k} \pi_k^2 (\frac{\hat{S}_k^2(1)}{n_{k,1}} + \frac{\hat{S}_k^2(0)}{n_{k,0}})$

$\hat{\tau}_s + - Z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_s}$

$H_o : \tau = 0 \quad \epsilon_s = \frac{\hat{\tau}_s}{\sqrt{\hat{V}_s}}$

2.4.3.5 SRE vs CRE

$e_k = e$ for all k, $\hat{\tau} = \hat{\tau}_s$

$var_{CRE}(\hat{\tau}) - var_{SRE}(\hat{\tau}_s) = \sum_{k=1}^{k} \frac{\pi_k}{n} (\sqrt{\frac{n_0}{n_1}} (\bar{Y}_k(1) - \bar{Y}(1)) + \sqrt{\frac{n_1}{n_0}} (\bar{Y}_k(0) - \bar{Y}(0)))^2 >= 0$

$\hat{\tau}_s$ is generally more efficient (has lower variance) than $\hat{\tau}_{CRE}$

2.4.3.6 Post-Stratification

We've already run a CRE

$\mathbf{n} = \{n_{k,1}, n_{k,0}\}_{k=1}^{K}$

$P_{CRE}(\mathbf{Z} = z|\mathbf{n}) = \frac{P(\mathbf{Z}=z,\mathbf{n})}{P(\mathbf{n})}$ where $\mathbf{Z}\epsilon R^n = \{Z_1, ..., Z_n\}$

= $\frac{1}{\pi_{k=1}^{k}\binom{n_k}{n_{k,1}}}$

How to choose k?

$k = 5$ empirically is a good choice OR $k = \frac{n}{2}$ which is the 'matched pairs experiment'

2.5 Lecture Five: Rerandomization and Regression Adjustment

Inigo Artiagoitia & Sai Kolasani

2.5.1 Last Lecture: Discrete Covariate

- Design: Stratification
- Analysis: Post-stratification

2.5.2 Completely randomized experiment

n total experimental units

n_1 units randomly assigned to the treatment group

n_0 units randomly assigned to the control group

number of ways to assign n_1 units to treatment out of n total units in a CRE:

$$\mathbb{Z}\binom{n}{n_1}$$

2.5.2.1 Rerandomization

In rerandomization, we draw Z from the CRE and accept it if and only if the covariates are balanced across treatment and control groups.

Also, we have covariates. For example:

- $x_i \in \mathbb{R}^k$ (e.g., age)
- Outcome y might be fitness, while treatment z could be a health intervention

We want to ensure these covariates are balanced between our treatment and control groups. To check this, we calculate the difference in means of covariates:

$$\hat{\tau}_x = \frac{1}{n_1}\sum_{i=1}^{n} z_i x_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-z_i)x_i$$

Under CRE, $E[\hat{\tau}_x] = 0$, meaning the covariates are perfectly balanced on average. However, in practice, we often see $\hat{\tau}_x \neq 0$.

The covariance of this difference is given by:

$$\text{cov}(\hat{\tau}_x) = \frac{1}{n_1}S_x^2 + \frac{1}{n_0}S_x^2$$

$$= \frac{n}{n_1 n_0}S_x^2$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T$$

2.5.2.2 Mahalanobis Distance

The following Mahalanobis distance measures the difference between the treatment and control groups:

$$M = \hat{\tau}_x^T \,\text{cov}(\hat{\tau}_x)^{-1}\hat{\tau}_x$$

$$= \hat{\tau}_x^T \left(\frac{n}{n_i n_0}S_x^2\right)^{-1}\hat{\tau}_x$$

2.5.2.3 Rerandomization Procedure

For this, we:

1. Draw Z from CRE
2. Accept if and only if $M \leq a$

The value of 'a' determines how strict we are about balance:

- $a = \infty \rightarrow$ We accept any randomization (equivalent to CRE)
- $a = 0 \rightarrow$ We only accept perfect balance (usually impossible)
- $a = 0.001 \rightarrow$ A good threshold used.

This procedure is invariant to linear transformations of covariates.

2.5.2.4 Inference

- We use FRT that simulates Z under the constraint $M \leq a$
- For $a = \infty$: ε be a univariate standard normal RV, and $\hat{\tau} - \tau \overset{D}{\sim} \sqrt{\text{var}(\hat{\tau})}\varepsilon$
- For a close to zero: $\hat{\tau} - \tau \overset{D}{\sim} \sqrt{\text{var}(\hat{\tau})(1 - R^2)}\varepsilon$ where R^2 is the squared correlation between $\hat{\tau}$ and $\hat{\tau}_x$

2.5.3 Regression Adjustment (analysis)

FRT:

1. Regress $Y_i \sim X_i \to \hat{Y}_i$. $\hat{\varepsilon}_i = \hat{Y}_i - Y_i$: pseudo outcome for test statistic.
2. Regress $Y_i \sim (Z_i, X_i) \to \hat{\beta}$ (coefficient of Z_i): Test statistic.
3. Estimation τ:

2.5.3.1 Fisher's ANCOVA (1925)

1. Run an OLS Regression: $Y_i \sim (1, z_i, x_i)$
2. Use β (coefficient of Z) as our estimator for $\tau = \hat{\tau}_F$

However, this method was criticized because:

1. Bias: $E[\hat{\tau}_F] - \tau \neq 0$
2. $\text{Var}(\hat{\tau}_F) > \text{var}(\hat{\tau})$ for $n_1 \neq n_0$
3. Standard errors from OLS are incorrect

2.5.3.2 Lin (2013) Improvements

Lin proposed some improvements to address these issues:

1. The bias $E[\hat{\tau}_F] - \tau$ is small for large n and approaches 0 as $n \to \infty$
2. Run an OLS Regression: $y_i \sim (1, z_i, x_i, x_i * z_i) \to \hat{\tau}_L$ (coefficient of z)
   - $\hat{\tau}_L$ is more efficient than $\hat{\tau}_F$
3. EHW standard error gives conservative estimate for $\hat{\tau}_L$.

2.5.3.3 Difference-in-Differences

- A special case of covariate adjustment, where X is the **lagged outcome before the treatment**.
- The estimator for the average treatment effect is given by:

$$\hat{\tau}_{\text{DiD}} = \frac{1}{n_1}\sum_{i=1}^{n} z_i(Y_i - x_i) - \frac{1}{n_0}\sum_{i=1}^{n}(1 - z_i)(Y_i - x_i)$$

  where n_1 is the number of treated units and n_0 is the number of control units.

This expression represents the **difference in treated and control outcomes after adjusting for pre-treatment covariates (the lagged outcomes x_i).**

- This can be rewritten as:

$$\hat{\tau}_{\text{DiD}} = \left(\hat{Y}(1) - \hat{Y}(0)\right) - \left(\hat{X}(1) - \hat{X}(0)\right)$$

where:

  - $\hat{Y}(1)$ and $\hat{Y}(0)$ are the mean outcomes for the treated and control groups, respectively.
  - $\hat{X}(1)$ and $\hat{X}(0)$ are the mean pre-treatment covariates for the treated and control groups, respectively.

- This estimator is **unbiased**, and the mean of the treatment and control groups are:

$$\hat{g}(C1) = \frac{1}{n_1}\sum_{i=1}^{n} z_i g_i$$

$$\hat{g}(C0) = \frac{1}{n_0}\sum_{i=1}^{n} (1 - z_i) g_i$$

where $g_i = Y_i - X_i$ represents the difference between the outcome and the lagged pre-treatment outcome for unit i.

- The variance of the estimator is:

$$\hat{V} = \frac{1}{n_1(n_1 - 1)}\sum_{i=1}^{n} z_i \left(g_i - \hat{g}(1)\right)^2 + \frac{1}{n_0(n_0 - 1)}\sum_{i=1}^{n} (1 - z_i) \left(g_i - \hat{g}(0)\right)^2$$

2.5.4 Matched Pairs Experiment

2.5.4.1 Notation and Setup

- Each pair contains one treated unit and one control unit. This ensures that comparisons between treatment and control are made within pairs, reducing the influence of extraneous variables.
- There are $2n$ experimental units, where:

$$Z_i = \begin{cases} 1 & \text{if the 1st unit receives treatment} \\ 0 & \text{if the 2nd unit receives treatment} \end{cases}$$

- (i, j) indexes the j-th unit of pair i, with:

$$i = 1, 2, \ldots, n \quad j = 1, 2$$

- **Potential Outcomes Framework**: Each unit (i, j) has potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$, where:
  - $Y_{ij}(1)$ is the outcome if unit (i, j) receives treatment.
  - $Y_{ij}(0)$ is the outcome if unit (i, j) does not receive treatment.

The goal is to estimate the treatment effect by comparing the potential outcomes between treated and

control units.

- The treatment assignment Z_i is drawn *i.i.d.* (independently and identically distributed) from a Bernoulli distribution with probability $\frac{1}{2}$:

$$Z_i \overset{iid}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$$

This means that for each pair, there is an equal chance that either the first unit or the second unit will be assigned to treatment. This randomization ensures that treatment assignment is unbiased.

- The outcomes for an individual pair (i) are defined as:

$$y_{i1} = z_i y_{i1}(1) + (1 - z_i) y_{i1}(0)$$

$$y_{i2} = z_i y_{i2}(0) + (1 - z_i) y_{i2}(1)$$

These equations capture the realized outcomes for the units in the experiment. Depending on the treatment assignment z_i, we observe either the treated or control potential outcome.

2.5.4.2 Fisher's Randomization Test (FRT)

- **Null Hypothesis (H_0)**: The null hypothesis in FRT states that there is no treatment effect, meaning the potential outcomes are the same regardless of whether the unit receives treatment or not. This implies:

$$Y_{ij}(1) = Y_{ij}(0) \quad \text{for all} \quad i = 1, \ldots, n \quad j = 1, 2$$

Under H_0, the assignment of treatment is purely random, and any observed differences between the treated and control units are due to chance.

- **Pairwise Treatment Effect ($\hat{\tau}_i$)**: For each pair i, the treatment effect is estimated as the difference between the outcomes of the two units. The treatment effect for pair i is given by:

$$\hat{\tau}_i = y_{i1} - y_{i2} \quad \text{if} \quad z_i = 1$$

$$\hat{\tau}_i = y_{i2} - y_{i1} \quad \text{if} \quad z_i = 0$$

Here, $\hat{\tau}_i$ represents the observed treatment effect within pair i.

- **Average Treatment Effect ($\hat{\tau}$)**: The overall treatment effect, $\hat{\tau}$, is the average treatment effect across all pairs. This is calculated as:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i$$

The goal is to determine whether this observed treatment effect is significantly different from what would be expected under the null hypothesis.

- **Distribution of $\hat{\tau}$ under H_0**: Under the null hypothesis, the expected value of $\hat{\tau}$ is zero:

$$\mathbb{E}[\hat{\tau}] = 0$$

The variance of $\hat{\tau}$ is calculated as:

$$\text{Var}(\hat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(\hat{\tau}_i)$$

Given that $\hat{\tau}_i = Y_{i1} - Y_{i2}$, the variance can be expanded as:

$$\text{Var}(\hat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(S_i) \left(Y_{i1} - Y_{i2}\right)^2$$

Finally, the variance is simplified to:

$$\text{Var}(\hat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \hat{\tau}_i^2$$

- **Test Statistic**: The test statistic for FRT is constructed by standardizing the observed treatment effect. Under the null hypothesis, the standardized test statistic follows a standard normal distribution:

$$\frac{\hat{\tau}}{\sqrt{\frac{1}{n^2}\sum_{i=1}^{n} \hat{\tau}_i^2}} \sim N(0,1)$$

This allows us to evaluate whether the observed treatment effect $\hat{\tau}$ is significantly different from zero, which would indicate a treatment effect.

2.5.4.3 Kolmogorov-Smirnov Type Statistic

- The observed treatment effects $\hat{\tau}_1, \ldots, \hat{\tau}_n$ are fixed, and the test statistic Δ is defined as:

$$\Delta = \sum_{i=1}^{n} \mathbb{I}\left(\hat{\tau}_i > 0\right)$$

where $\mathbb{I}\left(\hat{\tau}_i > 0\right)$ is an indicator function that takes the value 1 if the treatment effect $\hat{\tau}_i$ is non-negative and 0 otherwise. This statistic counts the number of pairs where the treatment effect is positive.

- **Distribution of Δ under the Null Hypothesis (H_0)**: Under the null hypothesis that there is no treatment effect, the indicator function $\mathbb{I}\left(\hat{\tau}_i > 0\right)$ follows a Bernoulli distribution with probability $\frac{1}{2}$:

$$\mathbb{I}\left(\hat{\tau}_i > 0\right) \overset{iid}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$$

This means that under H_0, there is a 50% chance that the treatment effect is positive in any given pair.

- **Binomial Distribution of Δ**: Since the indicator function follows a Bernoulli distribution, the sum Δ follows a binomial distribution with n trials and probability of success $\frac{1}{2}$:

$$\Delta \sim \text{Binomial}(n, \frac{1}{2})$$

- **Binomial Test and Normal Approximation**: To test whether the observed number of positive treatment effects is significantly different from what we expect under H_0, we can either use the Binomial test with

$p = \frac{1}{2}$ or, for large sample sizes, apply the normal approximation:

$$\frac{\Delta - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \sim N(0,1)$$

2.5.4.4 McNemar's Statistic

Binary Outcome

Pair i	Y_{i1}	Y_{i2}	z_i	Treated Outcome	Control Outcome
1	1	1	1	1	1
2	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
3	1	0	1	1	0
4	1	0	0	0	1

This will be continued in the following lecture...

2.6 Lecture Six: Observational Studies

Yongsi Wu, Hanyang Li, Mika Lee (Revisions)

2.6.1 Review of Last Lecture: Matched Pairs Experiment

One treated unit, one control unit per stratum (pair)

Example: assign 1 student to watch recordings online, 1 to attend in-person where they are comparable in terms of GPA, major, year

- SRE with $\frac{n}{2}$ strata

Unit (i, j) is the jth unit of ith pair for $i = 1, 2, \ldots, n$ and $j = 1, 2$ has potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$

- Define $Z_i = \begin{cases} 1 & \text{if 1st unit is treated} \\ 0 & \text{if 2nd unit is treated} \end{cases}$

FRT

Null: H_0: $Y_{ij}(1) = Y_{ij}(0)\ \forall i, j$

Let $\hat{\tau}_i$ be the outcome under treatment - outcome under control in pair i, i.e.,

$$\hat{\tau}_i = S_i(Y_{i1} - Y_{i2}) \quad \text{with } S_i = 2Z_i - 1$$

We have introduced the following three statistics:

1. **Sign test statistic**

2. **McNemar's statistic:** Binary Outcome Y

Pair i	Y_{i1}	Y_{i2}	z_i	Treated Outcome	Control Outcome
1	1	1	1	1	1
2	0	0	0	0	0
3	1	0	1	1	0
4	1	0	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n					

Denote the number of rows where $Y_{i1} \neq Y_{i2}$ by q. Then the number of times $Z_i = 1$ among q discordant rows is given by

$$m_{10} \sim \text{Binomial}(q, 1/2).$$

This statistic can be computed using the mcnemar.test() function in R.

3. **Kolmogorov-Smirnov-type statistic:**

Under H_0, $|\hat{\tau}_i|$ is fixed and its sign is random with mean 0 and variance 1. Thus, $(\hat{\tau}_1, \cdots, \hat{\tau}_n)$ and -$(\hat{\tau}_1, \cdots, \hat{\tau}_n)$ should have the same distribution.

Usual CDF:

$$\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{\tau}_i \leq t\} \quad \text{for } (\hat{\tau}_1, \cdots, \hat{\tau}_n). \tag{2.1}$$

Recalling the CDF for -X:

$$F_{-X}(t) = \mathbb{P}(-X \leq t) = \mathbb{P}(X > -t) = 1 - F_X(-t),$$

we then get the empirical distribution of $-(\hat{\tau}_1, \cdots, \hat{\tau}_n)$:

$$1 - \widehat{F}(-t-) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{-\hat{\tau}_i \leq t\}, \tag{2.2}$$

where $\widehat{F}(-t-)$ is the left limit of the function $\widehat{F}(\cdot)$ at $-t$. Combining the two pieces (2.1) and (2.2), a Kolmogorov-Smirnov-type statistic is

$$D = \max_t \left| \widehat{F}(t) + \widehat{F}(-t-) - 1 \right|.$$

2.6.2 Neymanian Inference

2.6.2.1 Estimate average casual effect

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \tau_i$$

$$\tau_i = \frac{1}{2}(y_{i1}(1) + y_{i2}(1) - y_{i1}(0) - y_{i2}(0))$$

$$\mathbb{E}[\hat{\tau}_i] = \tau_i$$

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i \text{ and } \mathbb{E}[\hat{\tau}] = \tau$$

$$\hat{V} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\tau}_i - \hat{\tau})^2$$

2.6.2.2 Theorem

$$\mathbb{E}[\hat{V}] - \text{Var}(\hat{\tau}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\tau}_i - \hat{\tau}) \geq 0 \quad \textit{(conservative estimate)}$$

where OLS regress $(\hat{\tau_1}, ..., \hat{\tau_n})^T \sim$ intercept, fitted intercept $= \hat{\tau}$, variance estimate $= \hat{V}$.

2.6.2.3 Covariate adjustment

- Covariate-adjusted FRT
- Regression adjustment

2.6.2.4 General Matched Experiment

- One unit gets treated in each stratum
- m_i units get control in set i
- Define Z_{ij} for unit i, j

$$Y_{ij} = Z_{ij}Y_{ij}(1) + (1 - Z_{ij})Y_{ij}(0)$$

- Within set average casual effect

$$\tau_i = \frac{1}{m_i + 1} \sum_{j=1}^{m_i+1} (Y_{ij}(1) - Y_{ij}(0))$$

$$\hat{\tau_i} = \sum_{j=1}^{m_i+1} Z_{ij}Y_{ij} - \frac{1}{m_i} \sum_{j=1}^{m_i+1} (1 - Z_{ij})Y_{ij}$$

$$\mathbb{E}[\hat{\tau_i}] - \tau_i = 0$$

2.6.2.5 Estimation

$$\tau = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i+1}(Y_{ij}(1) - Y_{ij}(0)) \tag{2.3}$$

$$= \sum_{i=1}^{n} w_i \tau_i \tag{2.4}$$

where N is the number of experiment units, n is the number of sets, and $w_i = \frac{1+m_i}{N}$.

Then the stimator is $\hat{\tau} = \sum_{i=1}^{n} w_i \hat{\tau}_i$, which is unbiased.

$\Rightarrow$ How are m_i's determined?

2.6.3 Observational Studies

Is caffeine bad for babies in uterus?

Women who consume caffeine more likely to have a miscarriage?

2.6.3.1 Observation Study

Why?

1. Caffeine is bad.
2. Selected for women without nausea.
    - Nausea is predictive of good outcomes.

A lot of casual inference problems:

1. Choose target parameter (scientific).
2. Identification
3. Estimation and inference

2.6.3.1.1 Example: In this example, denote $Y =$ miscarriage, $Z =$ drinks more than 2 cups of coffee.

- $\frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]}$
- $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$
- $\mathbb{E}[Y(1) - Y(0) \mid \text{high-risk}]$

Step 1) $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

Step 2) Identification

2.6.3.2 Definition

A parameter τ is identifiable if it can be written as a function of the observed data distribution.

2.6.3.2.1 Continue with example X = confounding factor(nausea) $\rightarrow$ affect both Y and Z.

$\mathbb{E}[Y(1) - Y(0) \mid X = 1]$ vs. $\mathbb{E}[Y \mid X = 1, Z = 1] - \mathbb{E}[Y \mid X = 1, Z = 0]$ holds when $Y(Z) \perp Z \mid X$

Proof:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X]] \tag{2.5}$$

$$= \mathbb{E}[\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]] \tag{2.6}$$

$$= \mathbb{E}[\mathbb{E}[Y(1) \mid Z = 1, X] - \mathbb{E}[Y(0) \mid X, Z = 0]] \tag{2.7}$$

$$= \mathbb{E}[\mathbb{E}[Y \mid Z = 1, X] - \mathbb{E}[Y \mid Z = 0, X]] \tag{2.8}$$

$Y(z) \perp Z \mid X$:

- Exchangeability
- Ignorability
- No unmeasured confounding
- Selection on observable
- Unconfoundedness
- Conditional independence
- Conditional exchangeability

• Untestable:

$$\mathbb{P}(Y(1) = 1 \mid Z = 1, X) = \mathbb{P}(Y(1) = 1 \mid Z = 0, X)$$

We can compute the RHS but not the LHS.

• Have we measured in X all factors that affect both Z and potential outcomes?

$$Y(1) = g_1(x, w)$$

$$Y(0) = g_0(x, u)$$

$$z = g(x, v)$$

1. exchangeability if $(w, u) \perp v$

   eg: V = preference for coffee

   If $Y(z) \not\perp Z \mid X$ but we pretend it is, then we get **omitted variable bias**.

Another assumption:

2. positivity / overlap:

$$\mathbb{P}(0 < \mathbb{P}(Z = 1 \mid X) < 1) = 1$$

* doesn't require every factor to be accounted for

This equation is testable.

3. SUTVA

2.6.4 Target trial emulation

What is the hypothetical ideal randomized experiment that our observational study is trying to emulate?

To be continued...

2.7 Lecture Seven: Observational Studies (cont.)

Youyou Xu & Zhiwei Xiao

2.7.1 Introduction

We continue our discussion to observational studies. Recall that observational studies are used instead of the "ideal" randomized experiments in several situations, primarily due to ethical, practical, or logistical constraints.

2.7.1.1 Example

Denote outcome Y as miscarriage, and treatment assignment Z represent pregnant women that are coffee drinker or not. We are interested in the causal effect, that is the miscarriage rate if all women drink coffee versus if all women do not drink coffee: $\tau = \mathbb{E}[Y(1) - Y(0)]$.

2.7.1.2 Trial Emulation

To target trial emulation, we want to answer what is the hypothetical randomized experiment that our observational study is trying to emulate?

1 Specify target trial

2 Justify how the observational data can be used to emulate the trial

2.7.1.3 Another Example

We want to study the effect of race in criminal justice system. Let outcome Y represent "stopped by the police", and let the Z represent race. Again, we are interested in estimating the causal effect $\tau = \mathbb{E}[Y(1) - Y(0)]$, where we can view $Y(1)$ as the response for one race and $Y(0)$ as the response for another. See that it is very tricky to estimate in this case, as race is a very complex social construct. It's very hard for researchers to intervene on this attribute. And it's difficult to well-define what constitutes as a treatment in this case.

2.7.2 Target Trial Emulation

Target trial emulation helps us make interventions well-defined so that our causal questions are well-defined. Related readings on "no causation without manipulation" : Rubin(1975), Holland (1986).

2.7.2.1 Steps

1. Define causal estimand
2. Identification
3. Estimation.

2.7.2.2 Example of Estimation

One approach is the outcome regression estimator (so called plug-in estimator, outcome modeling), motivated by the identification for τ we've seen before. The target estimand is again $\tau = \mathbb{E}[Y(1) - Y(0)]$. First, see that $\mathbb{E}[\mathbb{E}[Y|X, Z = 1]] = \mathbb{E}[Y|Z = 1]$, which follows from the tower property. Then we have that

$$\tau = \mathbb{E}[\underbrace{\mathbb{E}[Y|X, Z = 1]}_{\mu_1(x)} - \underbrace{\mathbb{E}[Y|X, Z = 0]}_{\mu_0(x)}] \quad \text{where } \mu_z \text{ is the nuisance function.}$$

Hence, $\tau = \mathbb{E}[\mu_1(x) - \mu_0(x)]$, and the estimator for the average treatment effect is $\hat{\tau}_{reg} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu_1}(x) - \hat{\mu_0}(x)]$ where $\hat{\mu_1}(x)$ is a regression model of $Y \sim X|Z = 1$ and $\hat{\mu_0}(x)$ is a regression model of $Y \sim X|Z = 0$. Keep in mind that different regression models all have their own pros and cons.

Example: suppose we have a binary outcome, with logistic regression,

$$\P(Y = 1|Z, X) = \frac{e^{\beta_0 + \beta_z z + \beta_x^T x}}{1 + e^{\beta_0 + \beta_z z + \beta_x^T x}}.$$

Now can we take the β_z as our causal effect estimate? No, there are two problems with this. One unavoidable problem is that

$$\hat{\tau}_{reg} = \frac{1}{n}\sum_{i=1}^{n}\{\frac{e^{\hat{\beta}_0 + \hat{\beta}_z + \hat{\beta_x}^T x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_z + \hat{\beta_x}^T x_i}} - \frac{e^{\hat{\beta}_0 + \hat{\beta_x}^T x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta_x}^T x_i}}\}.$$

If the logistic outcome model is correct, β_z would be equal to the log odds ratio and $\exp(\beta_z) = \frac{\text{odds}(Y(1)=1|X)}{\text{odds}(Y(0)=1|X)}$.

Another problem is that the strong parametric assumption can be misspecified.

Alternatives: random forest, kernel regression, etc.

2.7.3 Alternative Modeling

Alternative to the logistic regression model which introduces parametric assumptions, there are other non-parametric models to estimate β_z. Two examples are random forest model and kernel regression.

Model Type	Pros	Cons
Parametric Models	- Easier to estimate standard errors - Easier to use in high-dimensional settings	- Parametric assumptions may be misspecified - High bias
Non-Parametric Models	- Low bias given no parametric assumptions	- Harder to estimate standard errors - Curse of high dimensionality

Table 2.1: Comparison of Parametric and Non-Parametric Models

Notably, the performance of non-parametric models worsens exponentially as dimensionality increases.

2.7.4 Propensity Scores

Another estimation approach involves propensity scores, which reweighs the sample groups to create a "pseudo-population".

2.7.4.1 Question Setup

We can estimate $\mathbb{E}[Y(1)|Z=1] = \frac{1}{n_1}\sum_{i=1}^{n} z_i y_i$ and $\mathbb{E}[Y(0)|Z=0] = \frac{1}{n_0}\sum_{i=1}^{n}(1-z_i)y_i$. Recall that the estimand is $\tau = \mathbb{E}[Y(1) - Y(0)]$. Then, we reweigh the sample such that it creates a "pseudo-population" where the treatment assignment is as if it were random (unconfounded), given the covariates X. The weights are defined by the inverse of propensity scores of getting treatment or control condition. The propensity score $e(X)$ is:

$$e(X) = \mathbb{P}(Z = 1|X) \tag{2.9}$$

defined by Rosenbaum and Rubin (1983), and it follows that

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\frac{Z \cdot Y}{e(X)} - \frac{(1-Z)\cdot Y}{1-e(X)}] \tag{2.10}$$

2.7.4.2 Proof of Eq.(1.2)

We first prove that $\mathbb{E}[Y(1)] = \mathbb{E}[\frac{Z\cdot Y}{e(X)}]$, and we proceed by leveraging the law of iterated expectations and properties of propensity scores.

By definition, $\mathbb{E}[Y(1)]$ represents the expected outcome if every individual were treated ($Z = 1$). Using **the law of iterated expectations**, we can write:

$$\mathbb{E}\left[\frac{Z \cdot Y}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Z \cdot Y}{e(X)}\middle| X\right]\right].$$

We now evaluate an expectation conditional on X. Given the assumption of **ignorability**, conditioning on X, we can express the expectation as:

$$\mathbb{E}\left[\mathbb{E}\left[\frac{Z \cdot Y}{e(X)}\middle| X\right]\right] = \mathbb{E}\left[\frac{1}{e(X)} \cdot \mathbb{E}[Z \cdot Y|X]\right].$$

Since Z is a binary variable that takes the value 1 when the individual is treated, we know:

$$\mathbb{E}[Z \cdot Y|X] = e(X) \cdot \mathbb{E}[Y(1)|X],$$

because $Z = 1$ with probability $e(X)$ (the propensity score), and when $Z = 1$, the expected outcome is $Y(1)$. Thus, the expectation simplifies to:

$$\mathbb{E}\left[\frac{1}{e(X)} \cdot e(X) \cdot \mathbb{E}[Y(1)|X]\right] = \mathbb{E}\left[\mathbb{E}[Y(1)|X]\right] = \mathbb{E}[Y(1)].$$

Therefore, we have shown that:

$$\mathbb{E}\left[\frac{Z \cdot Y}{e(X)}\right] = \mathbb{E}[Y(1)].$$

This completes the proof for the first part. Similarly, the proof for $\mathbb{E}[Y(0)] = \mathbb{E}\left[\frac{(1-Z)\cdot Y}{1-e(X)}\right]$ follows the same steps. Given these two equations, we are successful in proving **Eq.(1.2)**.

2.7.4.3 An IPW Estimator

To account for the fact that we rarely have access to the true propensity score values in practice, we estimate the propensity scores, $e(X) = \mathbb{P}(Z = 1|X)$, using a model, typically a logistic regression or other classification model. Once we have the estimated propensity scores $\hat{e}(X)$, we can use them to construct an Inverse Probability Weighting (IPW) estimator, also known as the Horvitz-Thompson estimator.

$$\tau_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i \cdot Y_i}{\hat{e}(X_i)} - \frac{(1 - Z_i) \cdot Y_i}{1 - \hat{e}(X_i)}\right) \tag{2.11}$$

The empirical value of the score $\hat{e}(X)$ can be obtained via regressing Z on X.

2.7.4.4 Challenges to Propensity Scores

There are several challenges to the model of propensity scores:

1. Is the propensity score model correct?
2. Propensity scores are not invariant to local transformations. One solution to this issue is normalizing the

weights, thereby proposing an alternative **Hajek estimator**:

$$\tau_{\text{Hajek}} = \frac{\sum_{i=1}^{n}\left(\frac{Z_i \cdot Y_i}{\hat{e}(X_i)}\right)}{\sum_{i=1}^{n}\frac{Z_i}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^{n}\left(\frac{(1-Z_i)\cdot Y_i}{1-\hat{e}(X_i)}\right)}{\sum_{i=1}^{n}\frac{1-Z_i}{1-\hat{e}(X_i)}} \tag{2.12}$$

3. Empirically, $e(X)$ often come close to 0 or Two solutions exist for this issue:

   (a). **Truncate propensity scores.** In truncation, we replace the extreme values of the propensity scores with α_L (lower bound) or α_U (upper bound). The truncated propensity score $e_{\text{trunc}}(X_i)$ is:

   $$e_{\text{trunc}}(X_i) = \max\left(\alpha_L, \min\left(e(X_i), \alpha_U\right)\right)$$

   (b). **Trim observations.** In trimming, we remove observations where the estimated propensity score falls outside the range $[\alpha_L, \alpha_U]$. That is, we trim any observation i for which:

   $$e(X_i) < \alpha_L \quad \text{or} \quad e(X_i) > \alpha_U$$

   For possible upper and lower bound values, Crump et al. (2009) has used $\{\alpha_L = 0.1, \alpha_U = 0.9\}$, and Klumb et al. (2005) has used $\{\alpha_L = 0.05, \alpha_U = 0.95\}$.

2.7.4.5 Alternative Definition of Propensity Scores

An alternative definition to **Eq.(1.1)** that incorporates potential outcomes is

$$e(X, Y(1), Y(0)) = \mathbb{P}(Z = 1|X, Y(1), Y(0)) \tag{2.13}$$

The propensity score defined by this equation is sometime called the "true propensities" while Eq.(1.1) is "nominal propensities".

Note that the two definitions are equal when:

$$\mathbb{P}(Z = 1|X, Y(1), Y(0)) = \mathbb{P}(Z = 1|X) \quad \text{if} \quad Y(1), Y(0) \perp Z|X$$

The next lecture will explore the great features of the propensity score model, including the fact that it offers a great dimensionality reduction tool, which is motivated by:

$$\textbf{Thereom:} \quad \text{if} Z \perp (Y(0), Y(1))|X \quad \text{then} Z \perp (Y(0), Y(1))|e(X)$$

2.8 Lecture Eight: Propensity Scores (cont.)

Dudu Tang & Brian Fernando & Lucas Costa

2.8.1 Final Project & Logistics

Group assignment for STAT 156 final project has been released. Please reach out to your group members and start discussing your project topics. Please review the Group Project Guideline (on course website) for more details.

2.8.2 Questions From Last Lecture

2.8.2.1 Target Trial Emulation

What is the relationship between Target Trial Emulation (TTE) and Stable Unit Treatment Values Assumption (SUTVA)?

TTE encompasses a broader set of assumptions than SUTVA. TTE relies on SUTVA for the causal effects estimated in a target trial emulation to be valid.

Assumptions:

- There is no interference between units
- There is a consistent level of treatment

2.8.2.2 Nominal Propensity Scores vs. True Propensity Scores

Nominal Propensity Score:

$$e_n(x) = P(Z = 1|X = x)$$

True Propensity Score:

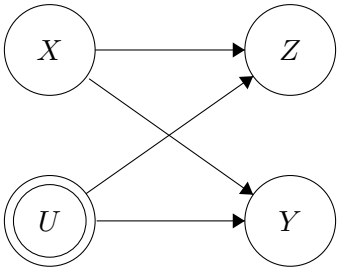
$$e_p(x, Y(1), Y(0)) = P(Z = 1|X = x, Y(1), Y(0))$$

Let X denote the confounding factor: Morning Sickness.

Let Z denote the treatment assignment: Drinking Caffeine

Let Y denote the outcome: Miscarriage

We can see that X can affect both Y and Z. However, they may also be an unobserved confounding factor U that can affect both Y and Z. We can denote U as being cautious.



Example Nominal Propensity Score: $e_n(x) = P(Z = 1|MorningSickness = Yes)$

In this case, even if $X \perp Z|Y(0), Y(1)$, $e_n(x)$ is not sufficient for ignorability due to the unobserved confounding factor U. Therefore, adjusting for propensity scores may not remove all the bias in estimating the treatment effect.

2.8.3 More on Propensity Scores

2.8.3.1 A Dimension Reduction Perspective on Propensity Scores

From the dimension reduction viewpoint, propensity scores help simplify the dimensional problem of covariate adjustment by condensing multiple high-dimensional covariates into a single propensity score — the probability of receiving the treatment given the covariates X. The theorem below reduces the dimension of covariates while maintaining ignorability.

Theorem:

$$\text{If } Z \perp (Y(1), Y(0)) \,|\, X \text{ then } Z \perp (Y(1), Y(0)) \,|\, e(X)$$

2.8.3.2 Propensity Score Stratification

We now introduce a method for estimating causal effect called Propensity Score Stratification.

1. Estimate propensity score by regressing $Z \sim X \to \hat{e}(x)$
2. Discretize $\hat{e}(x)$ into k quantities $\to \hat{e}'(x) = e_k$
3. Analyze as SRE (Stratified Random Experiment)

We see that $Z \perp (Y(1), Y(0)) \,|\, \hat{e}'(x) = e_k$ approximately holds for $k = 1, 2, ..., k$

Empirically, it often appears that getting the correct ordering of propensity scores is more important than obtaining the exact values. The reason lies in how propensity scores are used in practice: their primary function is

to balance covariates between treatment and control groups. This balancing typically relies more on the relative ranking of the propensity scores than their precise numerical values.

Key Questions:

1. How to choose k?
   - If k is too small, then ignorability will be violated
   - If k is too large, then we cannot analyze as SRE
   - In 1985, Rosenbaum and Rubin published a widely regarded paper that recommended to use $k = 5$
   - We can increase k as long as each stratum has enough control and treated units
   - We can decide k after looking at X and Z, however, we should fix k before looking at the outcome Y
2. How to estimate standard errors when propensity score is estimated?
   (a). Use SRE as a conservative estimate; The estimate is conservative because estimated propensity scores decrease asymptotic variance for estimating causal effect.
   (b). Bootstrap (include propensity score estimation); This theory is unclear due to discreteness of the estimator.

2.8.3.3 A Covariate Balancing Perspective on Propensity Scores

The covariate balancing perspective focuses on the role of propensity scores in balancing covariates between treated and untreated groups. The idea is that once individuals are matched or stratified based on their propensity scores, the distribution of covariates should be similar between the two groups, as would be the case in a randomized controlled trial.

$$Z \perp X | e(X)$$

- Within the same level of $e(x)$, covariate distributions are balanced in expectation across treatment and control
- In practice, we can check for covariate balance as a check on how good two propensity score model is

When to use propensity score models vs. outcome models?

1. Depends on how well you think you can estimate one vs. the other
2. Do both and see if their causal estimates align:
   - If their causal estimates differ, this suggests misspecification
   - If estimates align, maybe we got the right causal estimate...
3. Can we combine the propensity score model estimates and outcome regression estimate to get a better estimate of the causal effect?

2.8.4 Doubly-robust Estimators

Doubly-robust estimators are also known as:

1. Bias-corrected plugin estimators (*plugin* is the other word for *outcome regression*)
2. Model-assisted Horvitz-Thompson estimators *or* Model-assisted IPW estimators
3. Semi-parametric estimators *or* Semi-parametric efficient estimators
4. Augmented IPW (AIPW)
5. Double machine learning estimators
6. Debiased machine learning estimators
7. Orthogonal machine learning estimators

2.8.4.1 Introducing $\hat{\tau}_{\text{DR}}$

Recall the formulas for the **IPW estimator** and the **outcome regression estimator** are:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i \cdot Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)\cdot Y_i}{1-\hat{e}(X_i)} \tag{2.14}$$

$$\hat{\tau}_{\text{Reg}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1\left(X_i\right) - \hat{\mu}_0\left(X_i\right)\right) \tag{2.15}$$

One of the reasons why we are using the **outcome regression estimator** is that some of the potential outcomes are missing. Thus we have this formula where each term in the summation is an estimate of the treatment effect for that particular unit.

However, when we have already observed one potential outcome for every unit, why are we still using the predicted outcomes instead of using whichever we've observed? Again, the observed potential outcome is an observation of the conditional mean. We can reduce variance by directly estimating the conditional mean and plugging in that estimate here. However, we cannot estimate the conditional mean perfectly, so we are trading off the lower variance for a bias in estimation.

Here comes another question: is there really a way that we can use the observed outcomes but still get some nice bias-variance trade-offs? To solve this question, we introduce the **Doubly-robust estimator** where we use the errors in the estimates of the outcome regression model as a bias correction.

2.8.4.2 Formula & Interpretation

Formula for the **Doubly-robust estimator** is:

$$\hat{\tau}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}_1\left(X_i\right) + \frac{Z_i\left(Y_i - \mu_i\right)}{\hat{e}\left(X_i\right)} - \hat{\mu}_0\left(X_i\right) - \frac{\left(1-Z_i\right)\left(Y_i - \hat{\mu}_0\left(X_i\right)\right)}{1-\hat{e}\left(X_i\right)}\} \tag{2.16}$$

Here we can see that $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$ are just the outcome regression estimator.

The second and fourth terms, $\frac{Z_i(Y_i-\mu_i)}{\hat{e}(X_i)}$ and $\frac{(1-Z_i)(Y_i-\hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}$, display that Z_i are selecting units that received treatments and they are inversely weighing by their propensity score of treatment and control, respectively.

You can view the two terms as the IPW-weighed, bias-correction terms that are plugged in to reduce the bias of the **outcome regression model**. This explains why the **doubly-robust estimator** is also known as the **bias-corrected plugin estimator**.

We can also rearrange the terms of the equation $\hat{\tau}_{\text{DR}}$:

$$\hat{\tau}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\{\frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{Z_i - \hat{e}(X_i)}{\hat{e}(X_i)}\hat{\mu}_1(X_i) - \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)} - \frac{(\hat{e}(X_i)-Z_i)}{1-\hat{e}(X_i)}\hat{\mu}_0(X_i)\} \tag{2.17}$$

where the first and third terms are usual IPW terms. The second and fourth terms here are the *augmentation terms* that are augmenting the IPW by incorporating the outcome regression model estimates. This explains why the **doubly-robust estimator** is also known as the **augmented IPW (AIPW) estimator**.

Relationship with Machine Learning

It's popular to use Machine Learning to estimate functions such as propensity scores, so people refer to $\hat{\tau}_{\text{DR}}$ also as the **double machine learning** and **debiased machine learning estimators** (bias-corrected as proved above).

2.9 Lecture Nine: Doubly Robust Estimator (cont.)

Taejun Lee & Zach Rewolinski & Reet Mishra

2.9.1 Doubly Robust Estimator

From last lecture:

$$\hat{\tau}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_1(X_i) + \frac{Z_i(Y_i-\hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \hat{\mu}_0(X_i) - \frac{(1-Z_i)(Y_i-\hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}\right\}$$

$\hat{\tau}_{DR}$ is flexible, avoids parametric assumptions, and gets confidence intervals. We call the doubly robust estimator "double machine learning" because we use machine learning to estimate the nuisance functions $e(X)$, $\mu_0(X)$, and $\mu_1(X)$.

2.9.1.1 Advantages of Doubly Robust Estimator (in Parametric Setting)

- Robust to misspecification in either $e(X)$ or $(\mu_1(X), \mu_0(X))$ (hence "doubly robust").
- Can quantify confidence intervals on causal effect.

2.9.1.2 Proof of Robustness to Misspecification

$$\hat{\tau}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \frac{(1-Z_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)} + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right\}$$

$$\mathbb{E}\left[\hat{\tau}_{\text{DR}}\right] = \mathbb{E}\left[\frac{Z(Y - \hat{\mu}_1(X))}{\hat{e}(X)} - \frac{(1-Z)(Y - \hat{\mu}_0(X))}{1-\hat{e}(X)} + \hat{\mu}_1(X) - \hat{\mu}_0(X)\right]$$

Case 1: $\hat{\mu}_1(X) = \mu_1(X)$ **and** $\hat{\mu}_0(X) = \mu_0(X)$

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}_{\text{DR}}\right] &= \mathbb{E}\left[\frac{Z(Y - \mu_1(X))}{\hat{e}(X)} - \frac{(1-Z)(Y - \mu_0(X))}{1-\hat{e}(X)} + \mu_1(X) - \mu_0(X)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{Z(Y - \mu_1(X))}{\hat{e}(X)} - \frac{(1-Z)(Y - \mu_0(X))}{1-\hat{e}(X)} + \mu_1(X) - \mu_0(X)\,\middle|\, X, Z\right]\right] \\
&= \mathbb{E}\left[\frac{Z}{\hat{e}(X)}\mathbb{E}[Y \mid X, Z=1] - \frac{Z}{\hat{e}(X)}\mu_1(X)\right. \\
&\quad \left. - \left(\frac{1-Z}{1-\hat{e}(X)}\mathbb{E}[Y \mid X, Z=0] - \frac{1-Z}{1-\hat{e}(X)}\mu_0(X)\right) + \mu_1(X) - \mu_0(X)\right] \\
&= \mathbb{E}[\mu_1(X) - \mu_0(X)] \\
&= \tau
\end{aligned}$$

Case 2: $\hat{e}(X) = e(X)$

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}_{\text{DR}}\right] &= \mathbb{E}\left[\frac{Z}{\hat{e}(X)}(\mu_1(X) - \hat{\mu}_1(X)) - \frac{(1-Z)}{1-\hat{e}(X)}(\mu_0(X) - \hat{\mu}_0(X)) + \hat{\mu}_1(X) - \hat{\mu}_0(X)\,\middle|\, X\right] \\
&= \mathbb{E}\left[\frac{e(X)}{e(X)}(\mu_1(X) - \hat{\mu}_1(X)) - \frac{1-e(X)}{1-e(X)}(\mu_0(X) - \hat{\mu}_0(X)) + \hat{\mu}_1(X) - \hat{\mu}_0(X)\right] \\
&= \mathbb{E}[\mu_1(X) - \mu_0(X)] \\
&= \tau
\end{aligned}$$

Theorem: $\hat{\tau}_{\text{DR}} - \mathbb{E}[\hat{\tau}_{\text{DR}}] = 0$ **if 1)** $\hat{e}(X) = e(X)$ **OR 2)** $\hat{\mu}_1(X) = \mu_1(X)$ **and** $\hat{\mu}_0(X) = \mu_0(X)$

2.9.1.3 A Few Perspectives on the Doubly Robust Estimator

1. Reduces the bias of the outcome regression estimator.

   Assume estimand is $\mathbb{E}[Y(1)]$. Then we have $\hat{\tau}^1_{reg} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i)$, which has a bias of $\mathbb{E}\left[\hat{\tau}^1_{reg} - Y(1)\right] = \mathbb{E}\left[\hat{\mu}_1(X) - Y(1)\right]$. Notice that we cannot calculate this quantity because we cannot observe the potential outcomes. Thus, we instead do the following estimate of bias:

$$\mathbb{E}\left[\frac{Z(\hat{\mu}_1(X) - Y)}{\hat{e}(X)}\right] = B.$$

The resulting de-biased estimator is

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i) - \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i(\hat{\mu}_1(X_i) - Y_i)}{\hat{e}(X_i)}.$$

2. Reduces the variance of the IPW estimator.

   Recall that

$$\tau_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}$$

   can have high variance because we may be dividing by something which is close to zero.

   Instead, we can divide the residual term by $\hat{e}(X_i)$ so that it blows up less in the event that $\hat{e}(X_i)$ is close to zero or one:

$$\hat{\tau}^1_{DR} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} + \hat{\mu}_1(X).$$

2.9.1.4 Confidence Intervals

1. Nonparametric bootstrap
2. Wald-type normal approximation under certain conditions:
   - Sample-splitting: We use one half of the sample to estimate $e(X), \mu_0(X), \mu_1(X)$ and use the other half to estimate $\hat{\tau}_{DR}$ given $\hat{e}(X), \hat{\mu}_0(X), \hat{\mu}_1(X)$. Repeat the process starting with the second half to estimate the nuisance functions and average the results.
   - Flexible nonparametric methods to estimate nuisance functions $e(X), \mu_0(X), \mu_1(X)$.

$\hat{\tau}_{DR} - \hat{\tau}$ is asymptotically normal with variance $\frac{var(\phi(X,Y,Z))}{n}$ where

$$\phi(X,Y,Z) = \frac{Z}{e(X)}(Y - \mu_1(X)) - \frac{(1-Z)}{1-e(X)}(Y - \mu_0(X)) + \mu_1(X) - \mu_0(X).$$

Caution: doubly fragile in a parametric setting

Bias: $\hat{\tau}^1_{DR} - \mathbb{E}[Y(1)] = \mathbb{E}\left[\frac{e(X)-\hat{e(X)}}{e(X)}(\mu_1(X) - \hat{\mu}_1(X))\right]$

Kang & Schafer (2005) showed this product can make errors large when both $\hat{e}(X)$ and $\hat{\mu}_1(X)$ are misspecified.

2.9.2 Causal Estimands Beyond $E[Y(1) - Y(0)]$

Another estimand of interest is $\tau_T = \mathbb{E}[Y(1) - Y(0)|Z = 1]$

- "Average causal effect on the treated".
- Allows us to learn about effects of removing an exposure.
- Relevant when its infeasible to assign treatment to everyone.

- Example: lottery to attend magnet school.
- Example: invasive surgery for high-risk patients

- Advantage: requires weaker identifying assumptions.

We now need to identify τ_T, specifically the term in red below:

$$\tau_T = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y(0) \mid Z = 1].$$

Assumptions:

- One-sided ignorability: $Z \perp Y(0) \mid X$.
- One-sided positivity: $e(X) < 1$.

2.9.2.1 Theorem

Under one-sided ignorability and positivity,

1. $\tau_T = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}\left[\mathbb{E}[Y \mid Z = 0, X] \mid Z = 1\right]$
2. $\tau_T = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}\left[\frac{e(X)}{P(Z=1)} * \frac{(1-Z)Y}{1-e(X)}\right]$

2.9.2.2 Proof of Theorem 1

Part 1:

We want to show that $\mathbb{E}[Y \mid Z = 1] = \mathbb{E}\left[\mathbb{E}[Y \mid Z = 0, X] \mid Z = 1\right]$.

Proof 2.1

$$\begin{aligned}
\mathbb{E}[Y \mid Z = 1] &= \mathbb{E}\left[\mathbb{E}[Y(0) \mid Z = 1, X] \mid Z = 1\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y(0) \mid Z = 0, X] \mid Z = 1\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y \mid Z = 0, X] \mid Z = 1\right]
\end{aligned}$$

END OF LECTURE.

2.10 Lecture Ten: Matching in Observational Studies

Nikhil Shanbhag & Boyu Fan & Yichen Xu

2.10.1 Estimator for Average Causal Effect on the Treated

2.10.1.1 Assumptions

- One-sided ignorability: $Z \perp Y(0)|X$
- One-sided positivity: $P(e(X) > 0) = 1$

2.10.1.2 Definition

Define estimator for the average causal effect on the treated group to be

$$\tau_T = \mathbb{E}[Y(1) - Y(0) \mid Z = 1] = E[Y \mid Z = 1] - E[Y \mid Z = 0].$$

2.10.1.3 Theorem

$$\tau_T = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[\mathbb{E}[Y \mid Z = 0, X] \mid Z = 1] = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}\left[\frac{e(X)}{P(Z = 1)} \cdot \frac{(1 - Z)}{(1 - e(X))} Y \mid Z = 1\right].$$

2.10.1.4 Proof

One-sided ignorability implies $Y(0) \perp Z \mid X$ and one-sided positivity implies $P(e(X) > 1) = 1$.

To show the first line of the theorem,

$$\mathbb{E}[Y(0) \mid Z = 1] = \mathbb{E}[\mathbb{E}[Y(0) \mid Z = 1, X] \mid Z = 1] =$$

$$\mathbb{E}[\mathbb{E}[Y(0) \mid Z = 0, X] \mid Z = 1] = \mathbb{E}[\mathbb{E}[Y \mid Z = 0, X] \mid Z = 1].$$

Notice we can condition on X to rewrite:

$$\mathbb{E}\left[\frac{ZY(0)}{e}\right] = \frac{\mathbb{E}\left[\mathbb{E}[ZY(0) \mid X]\right]}{e} = \frac{\mathbb{E}\left[\mathbb{E}[Z \mid X]\mathbb{E}[Y(0) \mid X]\right]}{e} = \frac{\mathbb{E}\left[e(X)\mathbb{E}[Y(0) \mid X]\right]}{e}.$$

The RHS can be written as

$$\mathbb{E}\left[\frac{e(X)}{e} \cdot \frac{(1 - Z)Y}{1 - e(X)} \mid X\right] = \mathbb{E}\left[\frac{e(X)}{e} \cdot \frac{\mathbb{E}[(1 - Z)Y \mid X]}{1 - e(X)}\right] = \mathbb{E}\left[\frac{e(X)}{e}\mathbb{E}[Y \mid Z = 0, X]\right] =$$

$$\mathbb{E}\left[\frac{e(X)}{e}\mathbb{E}[Y(0) \mid X]\right]$$

by using the Tower property. This completes the proof.

2.10.2 Regression estimator

2.10.2.1 Definition

Consider

$$\tau_T = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[\mathbb{E}[Y \mid Z = 0, X] \mid Z = 1].$$

Define

$$\hat{\tau}_T^{\text{reg}} = \frac{1}{n_1}\sum_{i=1}^{n} Y_i Z_i - \frac{1}{n_1}\sum_{i=1}^{n} Z_i \hat{\mu}_0(X_i),$$

where $\hat{\mu}_0(X)$ is the outcome model for $E[Y|Z = 0, X]$ learned by regressing $Y \sim X \mid Z = 0$.

In order to compute

$$\frac{1}{n_1}\sum_{i=1}^{n} Z_i \left(Y_i - \hat{\mu}_0(X_i)\right),$$

recall the IPW estimators and specifically the 2nd identification, which is

$$\tau_T = E[Y \mid Z = 1] = \mathbb{E}\left[\frac{e(X)}{e} \cdot \frac{(1 - Z)Y}{1 - e(X)}\right].$$

2.10.2.2 Derivation of Other Estimators

This allows us to derive three estimators: (1) Horvitz-Thompson, (2) Hajek, and (3) Odds Ratio. They can written as follows:

$$\hat{\tau}_T^{\text{ht}} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_1}\sum_{i=1}^{n} \frac{\hat{e}(X_i)(1 - Z_i)Y_i}{1 - \hat{e}(X_i)}$$

$$\hat{\tau}^{\text{Hajek}} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{\sum_{i=1}^{n} \frac{\hat{e}(X_i)(1-Z_i)Y_i}{1-\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{\hat{e}(X_i)(1-Z_i)}{1-\hat{e}(X_i)}}$$

$$\tau^{\text{OR}} = \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i + \text{DR}(\mathbb{E}[Y(0) \mid Z = 1]) = \frac{1}{n_1}\sum_{i=1}^{n} \left\{\frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)}(1 - Z_i)(Y_i - \hat{\mu}_0(X_i)) + Z_i \hat{\mu}_0(X_i)\right\},$$

where DR represents the doubly robust estimator.

2.10.2.3 Theorem

Under one-sided ignorability and overlap, if either $\hat{e}(X) = e(X)$ or $\hat{\mu_0}(X) = \mu_0(X)$, then this estimator is unbiased.

2.10.2.4 Proof

The following is an alternative way to rewrite the doubly robust estimator:

$$\hat{\tau}_T^{\text{DR}} = \hat{\tau}_T^{\text{Reg}} - \frac{1}{n_1}\sum_{i=1}^{n}\left[\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(1-Z_i)(Y_i - \hat{\mu}_0(X_i))\right] = \hat{\tau}_T^{\text{ht}} - \frac{1}{n_1}\sum_{i=1}^{n}\left[\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(1-Z_i)Z_i\right]\hat{\mu}_0(X_i).$$

The causal effect on the overlap population is then determined by the estimand

$$\tau_0 = \frac{\mathbb{E}[e(X)(1-e(X))\tau(X)]}{\mathbb{E}[e(X)(1-e(X))]},$$

where $\tau(X)$ represents largest weights for units with $e(X) = \frac{1}{2}$.

Recall:

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$

$$\tau_T = \mathbb{E}[Y(1) - Y(0) \mid Z = 1]$$

$$\tau_0 = \frac{\mathbb{E}[e(X)(1-e(X))\tau(X)]}{\mathbb{E}[e(X)(1-e(X))]}.$$

2.10.3 Heterogeneous Causal Effects and Effect Modification

2.10.3.1 Definitions

Let V be a covariate, e.g., male/female. The heterogeneous causal effect is defined as:

$$\tau_{HCE} = \mathbb{E}[Y(1) - Y(0)|V = v]$$

A covariate V is called an *effect modifier* if:

$$\mathbb{E}[Y(1) - Y(0)|V = v] \neq \mathbb{E}[Y(1) - Y(0)]$$

and V is not affected by the treatment Z. If V is affected by treatment, this would instead be called a mediator.

The value of τ_{HCE} depends on the causal estimand.

2.10.3.2 Example

$$P(Y(0) = 1|V = 1) = 0.8$$

$$P(Y(1) = 1|V = 1) = 0.9$$

$$P(Y(0) = 1|V = 0) = 0.1$$

$$P(Y(1) = 1|V = 0) = 0.2$$

The heterogeneous causal effect is:

$$\tau_{HCE} = 0.1$$

Now consider:

$$\frac{P(Y(1) = 1|V = 1)}{P(Y(0) = 1|V = 1)} = \frac{0.9}{0.8} = \frac{9}{8}$$

$$\frac{P(Y(1) = 1|V = 0)}{P(Y(0) = 1|V = 0)} = 2$$

We have "effect measure modifier" and "effect heterogeneity"

2.10.3.3 Why Do We Care?

1. τ could be zero, but $\tau_{HCE} \neq 0$, meaning the effect differs for different subpopulations.
2. We want to identify who benefits from the treatment.

2.10.3.4 Examples

2.10.3.4.1 Moving to Opportunity (MTO) Experiment: A randomized experiment gave vouchers to families in public housing to move to richer neighborhoods. One outcome was mental health. The results showed that mental health improved for females but not for males, indicating an effect modification based on gender.

2.10.3.4.2 Direct Cash Transfers in Kenya: The effect of cash transfers on reducing childhood mortality was particularly strong for families who gave birth during the program. This suggests targeting families accordingly in future interventions. Additionally, the overall effect could be muted in geographies with fewer people of child-bearing age.

Effect modifiers are also referred to as *effect-measure modifiers* or *effect heterogeneity*.

2.10.3.5 Definition

Transportability refers to the extrapolation of causal effects computed in one population to a second. Lack of transportability corresponds to a lack of external validity.

2.10.3.6 Example: Transportability Issue

An example where transportability was questioned is the study by Smith and Pell (2003), which found no effect modifiers for the effect of parachutes on high-altitude jumping, showing that the reduction in death after a jump may not be generalizable across populations or situations.

2.10.4 RCT vs. Observational Studies

2.10.4.1 Randomized Controlled Trials (RCT)

Advantages:

- High confidence in causal effect for a particular context.

Disadvantages:

- Expensive to scale.

2.10.4.2 Observational Studies

Advantages:

- Easy to scale.
- Allows investigation across different contexts.

Disadvantages:

- Make strong assumptions.

END OF LECTURE.

2.11 Lecture Eleven: Matching in Observational Studies & Causal DAGs

Grace Yin & Haodong Ling & Zhengxing Cheng

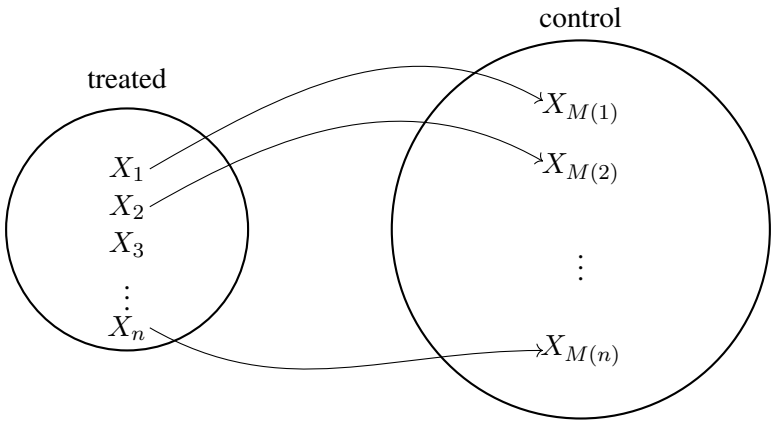
2.11.1 Last Time:

- RCT shows conflicting results. So the solution is using a meta-analysis of RCTs and observational studies clarified that certain outcomes, like pulmonary embolism, increased under HRT, while others, such as myocardial infarction, decreased.

2.11.2 Matching in Observational Studies

- **Goal**: Construct a subset of the population in which covariates have the same distribution in the treated and control groups.
- Intuitive and interpretable.

2.11.2.1 Matching with Many More Control Units



where X_i is matched to $X_{M(i)}$.

- **Ideal Settings:** we would find exact matches: $X_i = X_{M(i)} \Rightarrow e(X_i) = e(X_{M(i)})$.
  - Matches pair experiment assuming $Y(0), Y(1) \perp Z|X$.
  - Positivity holds in matching by construction.
- **More Realistic Settings: Imperfect Matching**
  - Almost always the case in observational studies because the covariate space can be large (e.g., continuous features).

$$m(i) = \arg \min_{k:Z_k=0} d(X_i, X_k)$$

where $d(X_i, X_k)$ is a distance metric.

2.11.2.2 Distance Metrics for Matching

Two Distance Metrics examples are:

- **Euclidean**:

$$d(X_i, X_k) = (X_i - X_k)^\top (X_i - X_k)$$

- **Mahalanobis**:

$$d(X_i, X_k) = (X_i - X_k)^\top \Sigma^{-1} (X_i - X_k)$$

where Σ is the sample covariance matrix of X's in the population.

  - Accounts for differences in scale across covariates and correlations between them.

2.11.2.3 Covariate Adjustment

Additionally, use covariate adjustment in analysis to correct for the residual covariate imbalance.

2.11.2.4 Handling High-Dimensional Covariates

Problem: When X is high-dimensional, for some X_i, $\min_{k:Z_k=0} d(X_i, X_k)$ is too large.

Solutions:

1. Discard these X_i's:
   - Effectively changes the study population.
2. Dimension reduction before matching:
   - e.g., Propensity score matching.
   - $m(i) = \arg\min_{k:Z_k=0} |\hat{e}(X_i) - \hat{e}(X_{m(i)})|$

2.11.2.5 Algorithm for Matching Multiple Control Units for Treated Unit

Allow for M_i Control Units for Treated Unit X_i:

- Input to the matching algorithm: Specify the number of control units, as well as lower or/and upper bounds on M_i.
- Matching algorithm outputs matched sets, M_i chosen to minimize:

$$\frac{1}{M_i} \sum_{j \in J_i} d(X_i, X_j)$$

where $d(X_i, X_j)$ is a distance function and J_i is the set of indices for control units matched to treated unit X_i.

2.11.2.6 Assessing the Distance Metric

- **Recap of Ignorability**: $Y(0), Y(1) \perp Z \mid X$ is impossible to empirically verify.
- **How to assess whether the distance metric is good?**
  - Analyze covariate distributions in treated vs. matched controls using visualizations like boxplots, or by computing moments of empirical distributions.

2.11.2.7 Matching Paradigm and Estimand

- Matching paradigm connects to our earlier discussion of the estimand: τ_T (average causal effect among treated).

$$\tau_T = \mathbb{E}[Y(1) - Y(0) \mid Z = 1] = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[\mathbb{E}[Y \mid Z = 0, X] \mid Z = 1]$$

- **Matching Estimator**:

$$\hat{\tau}_T = \frac{1}{n} \sum_{i=1}^{n} Z_i Y_i - Z_i \hat{\mu}(X_i)$$

where $\hat{\mu}(X_i) = \frac{1}{m_i} \sum_{j \in J_i} Y_j$.

This concludes our discussion on matching...

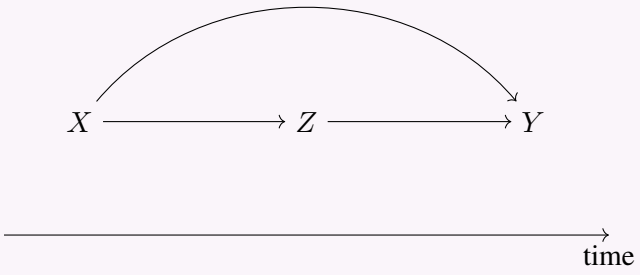
2.11.3 Causal DAGs

From now on, we begin our discussion on Causal Directed Acyclic Graphs (DAGs), which are powerful tools in causal inference. A famous quote, "*Draw your assumptions before your conclusions,*" emphasizes the importance of understanding underlying assumptions in any causal analysis.

In the computer science (CS) and artificial intelligence (AI) literatures, causal DAGs are extensively used to model relationships between variables. Pioneering work by researchers like Judea Pearl, Peter Spirtes, and Clark Glymour.

Example 2.1

Consider the following scenario: X, a confounder (e.g., nausea); Z, a treatment (e.g., drinking coffee); and Y, an outcome (e.g., miscarriage). In the following diagram, the DAG illustrate how the confounder X might influence both the treatment Z and the outcome Y.



Here are the meanings of some common notations in causal DAGs:

- $V \to W$ means V has a direct causal effect on W (an effect not mediated by another variable in the graph) for at least one individual.
- Lack of an arrow encodes the assumption that there is no direct causal effect of V on W for anyone.

Now, let's formally introduce Causal Directed Acyclic Graphs (Causal DAGs), which are essential tools for representing and analyzing causal relationships between variables in a structured, visual manner. Let's break down the key components:

- **Directed**: Causal DAGs include directed arrows between variables to indicate causal effects. For example, $X \to Z$ means X has a direct causal effect on Z.
- **Acyclic**: The graph is acyclic, meaning it cannot contain cycles like $X \rightleftarrows Z$ (where X causes Z and Z causes X), as such relationships are not allowed.

Some key definitions are listed below:

Definition 2.8 (Causal DAG)

A DAG that satisfies the Causal Markov Assumption.

Definition 2.9 (Causal Markov Assumption)

Conditional on its direct causes, any variable in a causal DAG is independent of any other variable that it does not cause.

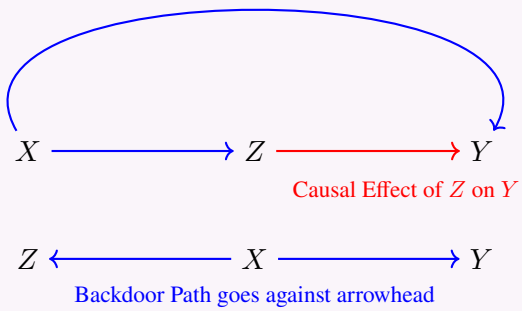
Causal DAGs represent causal relationships through properly ordered arrows (e.g., $X \to Y \to Z$) and associations through improperly ordered arrows (e.g., $X \to Z \leftarrow Y$).

2.11.3.1 Backdoor Path

A backdoor path is a non-causal path between two variables that can create spurious correlations. Backdoor paths are important because they can make it difficult to determine if an association between two variables is a result of a causal effect or a backdoor path.

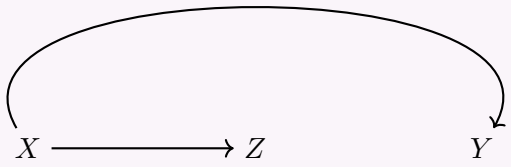
Example 2.2

Z and Y are linked by a backdoor path, so that are associated. This type of path creates confounding and needs to be accounted for when estimating causal effects.



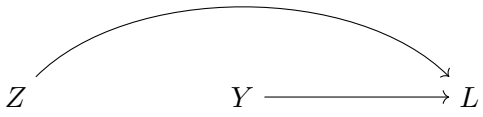
Example 2.3

Suppose variable Z is Carries lighter, Y means Lung cancer, and X means cigarette smoker. Z and Y have an association through X by a backdoor path, but no direct causal effect.



2.11.3.2 Colliding

In causal DAGs, a variable is a collider when it is causally influenced by two or more variables. For example, in the following diagram, L is a collider, as it is causally influenced by Z and Y.

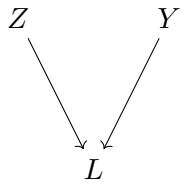


Definition 2.10 (collider)

Common effect is a **collider**.

Definition 2.11

Graphical structure of a collider is a **V-structure**, shown in the diagram below.



In the V-structure, we have:

- No association unconditionally between Z and Y: $Z \perp\!\!\!\perp Y$
- Z and Y are not independent when conditioning on L: $Z \not\perp\!\!\!\perp Y \mid L$

Example 2.4

Consider Z means poor time management, Y means family emergency and L means missed class. The relationship of Z, Y and L are shown as the V-structure, then $Z \perp\!\!\!\perp Y$. However, if I know that a student missed class, then if you tell me whether the student has poor time management skills, it is relevant for my guess about a family emergency occurring. This phenomenon is what we called *"explaining away"*.

2.11.3.3 Dependency in Causal DAGs

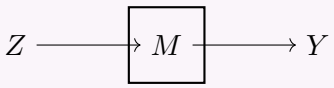
Now we introduce how we represent conditioning in causal DAGs.

Example 2.5

Consider the following diagram. Suppose Z means taking aspirin, M means platelet aggregation and Y means heart disease,

$$Z \longrightarrow M \longrightarrow Y$$

Is there an association between Z and Y within levels of M (e.g., conditional on M)? To study this question, we use the following diagram, where the box around M indicates conditioning on M.



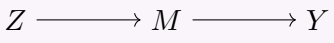
2.12 Lecture 12: Paths of association and d-separation

James Bowden

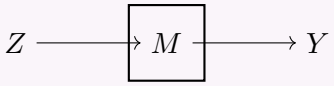
2.12.1 Last Time: Causal DAG definitions

Example 2.6

Consider the following diagram. Suppose Z: taking aspirin, M: platelet aggregation, and Y: heart disease.



Is there an association between Z and Y when conditioned on M? To study this question, we use the following diagram, where the box around M indicates conditioning on M.

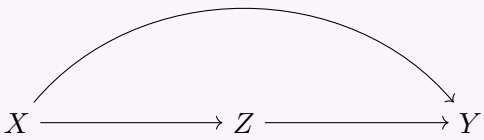


New objects introduced in the context of causal graphs include confounders, backdoor paths (non-causal), colliders (variables causally influenced by two or more variables), and mediators (which yield causal independence when conditioned upon).

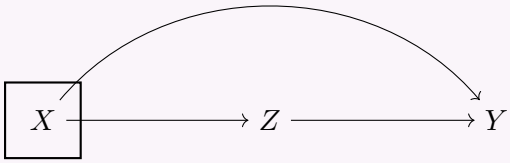
2.12.2 Revisiting common cause structure

Example 2.7

In this diagram, there's a direct causal path, $Z \to Y$, and an associative path, $Z \leftarrow X \to Y$. X is called a **confounder**.



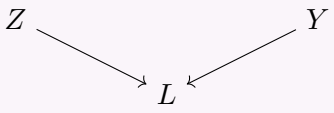
We can also denote blocking the associative path with a box, meaning $Z \perp Y(z)|X$. In this case, X no longer confounds the relationship between Z and Y; that is, we will only observe how changes in Z not induced by X affect Y.



2.12.3 Common effect structure

Example 2.8

Consider the following causal DAG: L is wet grass, Z is sprinklers ran overnight, Y is that it rained.



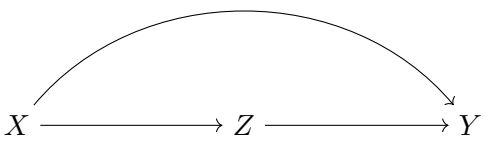
Here, $Z \perp Y$ but $Z \not\perp Y|L$. L is called a **collider**. Paths of association do not flow through colliders (i.e., $Z \perp Y$). However, if we were to condition on L (or some descendent thereof), denoted in a causal DAG as a box around L, then association would flow through L (i.e., $Z \not\perp Y|L$). This is because when L has been fixed as a known quantity, knowing something about Z will provide information about Y (and vice versa). One can see this through the simple relationship $Z + Y = L$ where $Z \perp Y$ but if L is known, then knowing Z yields Y as $Y = L - Z$.

2.12.4 Recap

2.12.4.1 Why might 2 variables be associated? Structural reasons:

1. One causes the other
2. The two share a common cause (a confounder)
3. The two share a common effect, and our analysis looks at a certain level that effect (i.e., conditional on a value of $L = l$ in the diagram immediately above).

2.12.4.2 A causal DAG implies a set of structural equations



What equations does this graph imply?

- $X \sim F_X(X)$

- $Z \sim g_Z(X, \epsilon_Z)$
- $Y \sim g_Y(X, Z, \epsilon_Y)$

2.12.5 Exchangeability

2.12.5.1 Causal graphs + exchangeability

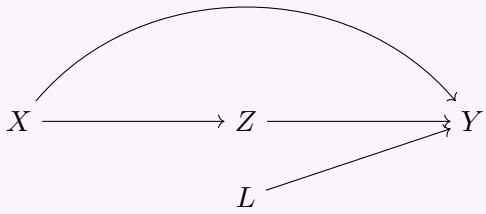
If we know the true causal DAG, then we can determine whether there exists a set of variables X s.t. $Z \perp Y(1), Y(0)|X$. This leads us to the backdoor criterion.

Definition 2.12 (Backdoor criterion)

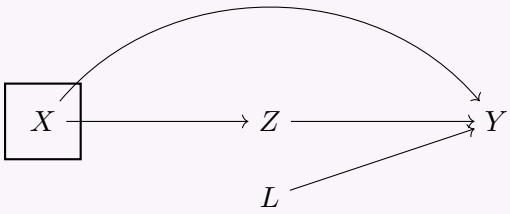
The backdoor criterion holds for a set of covariates X when all backdoor paths (non-causal associative paths) between Z and Y are blocked by conditioning on X, and X contains no variables that are descendents of Z. The backdoor criterion implies exchangeability conditional on X.

Example 2.9

In the below DAG, there is a backdoor path through X.



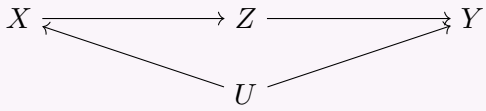
In order to make the backdoor criterion hold, we must condition on X, denoted by a box:



Now there are no backdoor paths, so the backdoor criterion is fulfilled.

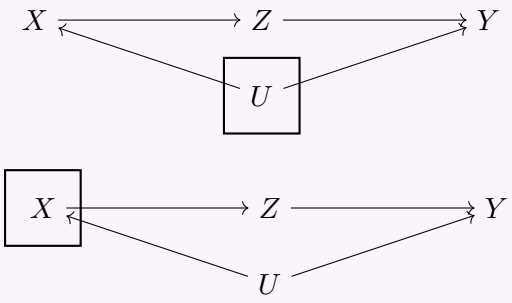
Example 2.10

Let Z be aspirin, Y is stroke, and X is doctor diagnosis of previous underlying health conditions.



What paths go from Z to Y? There's the direct causal path $Z \rightarrow Y$ as well as a backdoor path, $Z \leftarrow X \leftarrow U \rightarrow Y$. As such, U is a **common cause**.

To fulfill the backdoor criterion, we can block the backdoor path in two different ways: via conditioning on either U or X.



2.12.5.2 Recap

Causal graphs are useful for thinking about confounding because:

- We make assumptions about the data generating process explicit.
- They enable us to find the set of confounders (if possible) to achieve exchangeability under those assumptions.

2.12.5.3 Other methods for assessing exchangeability

Recall that exchangeability, $Z \perp Y(1), Y(0)|X$ is untestable. This is because it implies

$$P(Y(1) = 1|Z = 1, X) = P(Y(1) = 1|Z = 0, X)$$

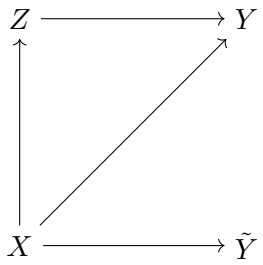
and while the former term is observable, the latter is the counterfactual and is not ever known. *Note that we can write the analogous equation for $Y(0)$ and observe the same issue.*

As such, we cannot ever test this assumption formally. We can, however, assess the strength of evidence in support of exchangeability to validate (or call into question) our assumption.

1. Use data on other outcomes: "negative outcomes".

   Suppose we observe $\tilde{Y}$. Assume:

   - The confounding of $\tilde{Y}$ is the same as the confounding for Y w.r.t. Z
   - Z has no effect on $\tilde{Y}$



In terms of potential outcomes, we would write:

- $Z \perp Y(Z)|X$ and $Z \perp Y(\tilde{Z})|X$
- $\mathbb{E}[\tilde{Y}|Z=1] - \mathbb{E}[\tilde{Y}|Z=0] \neq 0$ if we have confounding

Example 2.11

From Cornfield (1959): Y is lung cancer, Z is cigarette smoking, $\tilde{Y}$ is car accidents.

He found evidence that $\mathbb{E}[\tilde{Y}|Z=1] = \mathbb{E}[\tilde{Y}|Z=0]$.

To the extent that there are common confounders for car accidents and lung cancer, then this supports the claim that the association between lung cancer and smoking is causal. It is noted that there's no real intuition for why car accidents and lung cancer are related, but this example may be illustrative anyway, as an extreme case.

Example 2.12

From Jackson et al. (year?): Let Y be hospitalization during flu season, $\tilde{Y}$ hospitalization before flu season, Z getting the flu vaccine.

They found evidence that $\mathbb{E}[\tilde{Y}|Z=1] - [\tilde{Y}|Z=0]$ was very large. From this, we can conclude that there were unmeasured confounders, and that the assumption of exchangeability was violated / not reasonable.

2. To be continued...

2.13 Lecture 13: Negative Outcomes and Instrumental Variables

Jean Lee

2.13.1 Last Time

- Negative outcomes were used to assess the exchangeability assumption.
- Negative outcomes are also known as negative outcome controls or placebo tests.

Recall example:

- y: hospitalization during flu season
- $\tilde{y}$: hospitalization before flu season
- z: vaccine

$\mathbb{E}[\tilde{Y}|Z=1] - \mathbb{E}[\tilde{Y}|Z=0]$ large: suggests unmeasured confounding

2.13.1.1 Another Example: lagged outcomes (Imbens and Rubin, 2015)

- Use outcome right before treatment as a negative outcome
- Treatment can't affect something that happened before the treatment
- The confounding structure may be similar for lagged outcomes

Assume the confounding follows:

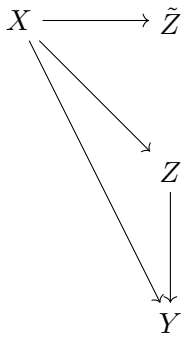
$$\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0] = \mathbb{E}[\tilde{Y}|Z=1] - \mathbb{E}[\tilde{Y}|Z=0]$$

This leads to a difference-in-differences model.

2.13.2 Assessing Confounding Using Negative Exposures

Another method to assess confounding is by using data on other treatments/exposures: "Negative Exposures"

- $\tilde{Z}$: treatment/exposure variable that shares the same confounding as Z w.r.t Y



Assume: $Z \perp\!\!\!\perp Y(Z) \mid X$ and $\tilde{Z} \perp\!\!\!\perp Y(Z) \mid X$

$$\mathbb{E}[Y(\tilde{Z}=1) - Y(\tilde{Z}=0)] = 0$$

2.13.2.1 Example: Sanderson et al 2017

- Let Z be maternal exposure during pregnancy
- Let $\tilde{Z}$ be paternal exposure during pregnancy
- Let Y be the child's BMI or autism spectrum disorder

2.13.3 Proximal Causal Inference

Proximal causal inference: when we have both negative outcomes and negative exposures

- Conditions under which we can nonparametrically identify the casual effect in the presence of unmeasured confounding
  - Example:

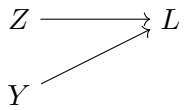
    u: unobserved confounder (discrete)

    $\tilde{y}$: negative outcome (discrete)

    $\tilde{z}$: negative exposure (discrete)
- when $\tilde{y}$ and $\tilde{z}$ have as many levels as u, then we can identify the causal effect

2.13.4 Colliders and Over-Adjustment Problems

Recall collider L:



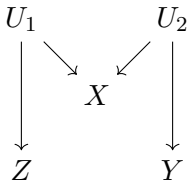
Rule of d-separation:

$\rightarrow\leftarrow$ closed

$\rightarrow \boxed{L} \leftarrow$ opens $Z \perp\!\!\!\perp Y \mid L$

2.13.4.1 Problem 1: M-Bias

M-bias:

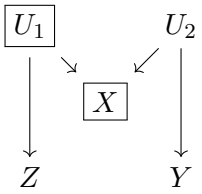


$$\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] = 0$$

- valid unbiased estimator for causal effect because there's no confounding

If "over-adjust" by condition on X: $Z \perp\!\!\!\perp Y(Z) \mid X$

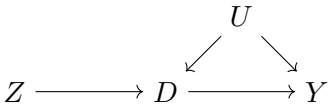
$$\mathbb{E}[\mathbb{E}[Y \mid Z = 1, X] - \mathbb{E}[Y \mid Z = 0, X]] \neq 0$$



blocks path of association

2.13.4.2 Problem 2: Z-bias

Z-bias:



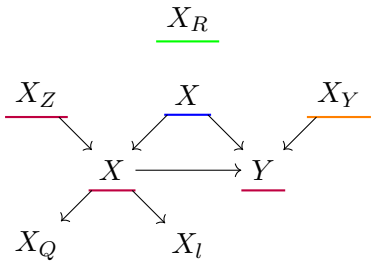
Key Assumptions:

- Instrument Independence: $Z \perp\!\!\!\perp Y$
- Instrument Relevance: $Z \not\perp\!\!\!\perp D$
- Z only affects y through D
- Generally conditioning on Z will make our estimates more biased
  - D: treatment received
  - Y: outcome
  - Ch. 16 of A first course example with linear models
- Z is not a confounder, but U is and we don't observe U
- Conditioning on Z makes D less random and amplifies the role of U in the remaining randomness of D
- Z: instrumental variable
- Z-bias = instrumental variable bias

Example:

- D: prison sentence length
- Y : recommit an offense (re-arrest) after release
- Z: random assignment of cases to judges
- U: personal characteristics, family support

What covariates should we adjust for in observational studies?



X_l: post-treatment variable

Necessary for Identification	Helpful to Reduce Variance	Harmful
X	X_Y	X_R
"confounding"	"effect modifier"	X_Z
		X_l

What to do when it's unreasonable to assume exchangeability

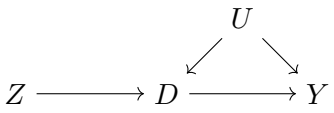
Today: instrumental variables

Next week: sensitivity analysis + bounds + partial identification

2.13.5 Instrumental Variables

In experiments, often participants may not adhere/comply to treatment assignment

- Z: treatment assignment
- D: the treatment taken, "adherence"
- U : confounders that affect Y and D
- Y: outcome



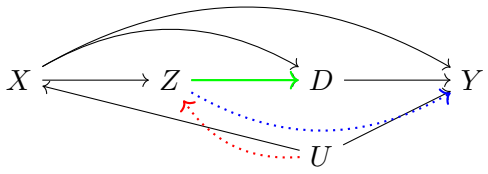
Example

- D: aspirin taken
- Y: stroke
- Z: whether patient was assigned aspirin
- U: behavioral factors

- Challenge: can't measure everything in U
- Solution: leverage randomness in instrument Z

Definition: A **instrumental variable** is a random variable that meets 3 condition:

1. **Relevance**: $Z \not\perp\!\!\!\perp D$
   - Z is associated with the treatment D
2. **Exclusion Restriction**: Z only affects Y only through D
   - No direct effect of Z on Y
   - e.g., holds double-blind experiment
3. **Exchangeable / Unconfounded IV** Z and Y don't share unmeasured confounders



* dotted arrow: not an edge

Example: Minneapolis Domestic Violence Experiment (1980s)

- randomly assigned penalties:
  1. arrest (1/3)
  2. counseling (1/3)
  3. separation (1/3)

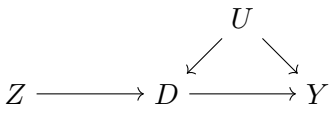
  (arrest, counseling, separation are all D)
- Y : re-offense
- Officers did not always comply

2.14 Lecture 14: Instrumental Variables Continued

Jaeeun Park & Zekai Wang & Claire Hsu & Jiarong Zhou(revisions)

2.14.1 Review from Last Lecture

Last lecture we introduced the notion of instrumental variable and gave a high level definition for it. Assume Z is the binary instrumental variable (e.g., treatment assigned), D is the binary treatment a unit actually received, Y is the outcome, and U are (unmeasured) confounders that affect both the treatment received and the outcome.



Definition 2.13

Informally, the Instrumental Variable (IV) Z is valid if it meets the following three conditions:

1. Relevance: Z has a direct effect on D. This can be seen from the $Z \to D$ path in the causal DAG.
2. Exclusion Restriction: Z only affects Y through D. This can be seen from the lack of a $Z \to Y$ path in the causal DAG.
3. Exchangeability (Unconfoundness): Z and Y don't share unmeasured confounders. This can be seen from the lack of a $U \to Z$ path in the causal DAG.

As a concrete example, in the Minneapolis Domestic Violence Experiment, police officers are randomly assigned to one of three penalties (arrest, counseling, and separation from partner) when dealing with domestic violence perpetrators. However, officers are not required to strictly follow the assignment. In this example, Z is the random assignment of the three penalties, D is the actual penalty given, Y is the indicator for re-offense, and U is some latent cause (e.g., propensity to commit violence, severity of the offense). *Relevance* holds since the officers are encouraged to follow the penalty assignment, *Exclusion Restriction* holds since it is reasonable to assume the assignment does not directly affect the chances of re-offense, and *Exchangeability* holds by design since Z is randomly assigned in the experiment. However, as we will see later, it is usually hard to identify a valid IV in observational studies since doing so involves making untestable assumptions.

2.14.2 Instrumental Variables

2.14.2.1 Formal Definition of IV

To rigorously analyze IVs, we first need to introduce some additional notations.

- $D(Z = z)$ is the potential outcomes for D, i.e., what treatment a unit would actually receive if the IV takes value z. In the Minneapolis Domestic Violence Experiment, this is what penalty would be chosen if the officer is suggested to give penalty z. In the case of binary outcome, we often abbreviate $D(Z = 1) = D(1), D(Z = 0) = D(0)$.
- $Y(D = d)$ is the potential outcomes for Y given the unit actually receives treatment d. In the Minneapolis Domestic Violence Experiment, this is the indicator for re-offense if the officer gives penalty d to the perpetrator.
- $Y(Z = z, D = d)$ is the potential outcomes for Y given the unit has IV value z and actually receives treatment d. In the Minneapolis Domestic Violence Experiment, this is the indicator for re-offense if the officer is suggested to give penalty z and actually gives penalty d.
- $Y(Z = z) = Y(Z = z, D = D(Z = z))$ is the potential outcomes we would observe if a unit has IV value z. In the Minneapolis Domestic Violence Experiment, this is the indicator for re-offense if the officer is suggested to give penalty z but we don't have information about what penalty the officer actually gives.

With these, we are ready to formally define IVs.

Definition 2.14

A variable Z is a valid Instrumental Variable (IV) if it satisfies the following three conditions:

1. **Relevance**: $Z \not\perp\!\!\!\perp D$, meaning Z is correlated with D and thus has a measurable effect on the treatment variable D.
2. **Exclusion Restriction**: The potential outcome Y depends on the treatment D but not directly on Z. Formally, $Y(Z = z, D = d) = Y(D = d)$ for all possible values of z and d.
3. **Exchangeability (Unconfoundedness)**: The IV Z is independent of the potential outcomes of D and Y. Specifically, $Z \perp (D(Z = z), Y(Z = z))$. In the case of a binary IV, this implies $Z \perp (D(Z = 0), D(Z = 1), Y(Z = 0), Y(Z = 1))$. If covariates X are included, this condition becomes $Z \perp (D(Z = z), Y(Z = z)) \mid X$.

The *Relevance* assumption has a testable implication: $\Pr D = 1 \mid Z = 1 - \Pr D = 1 \mid Z = 0 \neq 0$. Note when its value is small but nonzero, Z is considered to be a weak instrument. However, the other two assumptions are generally untestable.

2.14.2.2 Some Estimands

Sadly, if we only assume Z is a valid IV (i.e., we only assume *Relevance, Exclusion Restriction, and Exchangeability*), we are not able to identify the Average Treatment Effect (ATE) $\mathbb{E}[Y(D=1) - Y(D=0)]$. Nevertheless, below are some estimands.

- Intention-to-treat effect: $\tau_{ITT} = \mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]$
- As-treated estimand: $\tau_{AT} = \mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0]$
- Per protocol estimand: $\tau_{PP} = \mathbb{E}[Y \mid D=Z=1] - \mathbb{E}[Y \mid D=Z=0]$

Unfortunately, the As-treated estimand and the Per protocol estimand do not have a clear causal meaning. For the Intention-to-treat effect, under *Relevance* and *Exchangeability*, it equals the causal effect of instrument Z on Y:

$$\begin{aligned}\tau_{ITT} &= \mathbb{E}[Y(Z=1) \mid Z=1] - \mathbb{E}[Y(Z=0) \mid Z=0] \\ &= \mathbb{E}[Y(Z=1) - Y(Z=0)]\end{aligned}$$

If we additionally assume *Exclusion Restriction*,

$$\begin{aligned}\tau_{ITT} &= \mathbb{E}[Y(Z=1) - Y(Z=0)] \\ &= \mathbb{E}[Y(Z=1, D=D(1)) - Y(Z=0, D=D(0))] \\ &= \mathbb{E}[Y(D=D(1)) - Y(D=D(0))]\end{aligned}$$

Where $Y(D=D(0)), Y(D=D(1))$ are called stochastic intervention since the actual treatment received is a random variable (here they are $D(0)$ and $D(1)$ respectively).

Recall Fisher's Sharp Null is $H_0 : Y(D=1) = Y(D=0)$ (for all units). It is stronger than the "Neyman's weak null," which only asserts $\mathbb{E}[Y(D=1) - Y(D=0)] = 0$ in expectation. Under Fisher's Sharp Null and assuming binary instrument and treatment,

$$\begin{aligned}\tau_{ITT} &= \mathbb{E}[Y(D=D(1)) - Y(D=D(0))] \\ &= \mathbb{E}[(Y(D=1) - Y(D=0)\mathbf{1}\{D(1) > D(0)\}) + \\ &\quad (Y(D=0) - Y(D=1)\mathbf{1}\{D(1) < D(0)\})] \\ &= \mathbb{E}[(Y(D=1) - Y(D=0)) \times \\ &\quad (\mathbf{1}\{D(1) > D(0)\} - \mathbf{1}\{D(1) < D(0)\})] \\ &= 0\end{aligned}$$

One limitation of the Intention-to-treat effect is that it might not answer scientific questions about the effect of taking treatment (which is measured by the ATE).

2.14.2.3 Dealing with Noncompliance

In the case of binary instrument and treatment, we can view the instrument as the treatment assigned. At a high level, we aim to stratify the population based on how each unit complies with the assigned treatment. We begin by defining a latent (unobservable) variable U_i (this is not the confounder U) for each unit i according to the table below.

$D_i(1)$	$D_i(0)$	Label	U_i
1	1	Always Taker	a
1	0	Complier	c
0	1	Defier	d
0	0	Never Taker	n

Usually we have to make the additional assumption of *Monotonicity*.

Assumption 2.3

Monotonicity means there is no defier in the population: for all units i,

$$D_i(1) \geq D_i(0)$$

Monotonicity also has a testable implication that $\Pr D = 1 \mid Z = 1 \geq \Pr D = 1 \mid Z = 0$. When control units don't have access to treatment (e.g., when treatment is a new drug unavailable on the market), $Z = 0$ would imply $D = 0$, i.e., we have one-sided noncompliance (there can be compliers and never takers, but there cannot be any defier or always taker). One-sided noncompliance further implies *Monotonicity*.

Under *Relevance, Exclusion Restriction, Exchangeability* and if we assume one-sided noncompliance (meaning $Z = 0$ would imply $D = 0$), we can identify the effect of removing treatment on the whole population or among the treated.

$$\begin{aligned}\mathbb{E}[Y - Y(D=0)] &= \mathbb{E}[Y] - \mathbb{E}[Y(D=0) \mid Z=0] \\ &= \mathbb{E}[Y] - \mathbb{E}[Y \mid Z=0] \\ \mathbb{E}[Y - Y(D=0) \mid D=1] &= \frac{\mathbb{E}[Y] - \mathbb{E}[Y \mid Z=0]}{\Pr D=1}\end{aligned}$$

Where the last equation relies on the fact that $\mathbb{E}[Y - Y(D = 0) \mid D = 0] = 0$. We can estimate these values using various techniques, for example the Regression, Inverse Probability Weighted, and Double Robust estimators.

2.15 Lecture 15: Compiler/Local Average Treatment Effect (CATE/LATE)

Tobias Kreiman & Zixun Tan Jiarong Zhou(revisions)

2.15.1 Last time:

Monotonicity: $D_i(1) \geq D_i(0) \quad$ for all i

- Holds when there is one-sided non-compliance:

$$\Pr(D = 1 \mid Z = 1) \geq \Pr(D = 1 \mid Z = 0)$$

- Why $\geq$ not $>$?

  If $\quad \Pr(D = 1 \mid Z = 1) = \Pr(D = 1 \mid Z = 0), \quad$ then the relevance assumption is violated

2.15.2 CATE/LATE

$$\tau_c = \mathbb{E}\left[Y(D = 1) - Y(D = 0) \mid U = c\right]$$

- U is the unobservable adherence random variable, c is the complier.

$$= \mathbb{E}\left[Y(D = 1) - Y(D = 0) \mid D(1) \geq D(0)\right]$$

$$= \mathbb{E}\left[Y(D = 1) - Y(D = 0) \mid D(1) = 1, D(0) = 0\right]$$

- Identifiable under monotonicity + 3 standard IV assumptions (Imbens and Angrist, 1994).

$$\tau_c = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]}$$

Proof of identification:

Numerator:

$$\begin{aligned}
\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] &= \mathbb{E}[Y(Z = 1) \mid Z = 1] - \mathbb{E}[Y(Z = 0) \mid Z = 0] \quad \text{(consistency)} \\
&= \mathbb{E}[Y(Z = 1)] - \mathbb{E}[Y(Z = 0)] \quad \text{(exchangeability)} \\
&= \mathbb{E}[Y(Z = 1, D = D(1)) - Y(Z = 0, D = D(0))] \\
&= \mathbb{E}[Y(D = D(1)) - Y(D = D(0))]
\end{aligned}$$

$$\mathbb{E}\left[Y(D = 1) - Y(D = 0) \mid D(1), D(0)\right] = \begin{cases} \mathbb{E}\left[Y(D = 1) - Y(D = 0)\right] & \text{if } D(1) = 1, D(0) = 0 \\ \mathbb{E}\left[Y(D = 0) - Y(D = 1)\right] & \text{if } D(1) = 0, D(0) = 1 \end{cases}$$

$$\begin{aligned}\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0] &= \mathbb{E}[D(Z=1) \mid Z=1] - \mathbb{E}[D(Z=0) \mid Z=0] \\ &= \mathbb{E}[D(Z=1) - D(Z=0)] \\ &= P(D(Z=1)=1) - P(D(Z=0)=1) \\ &= P(D(1) > D(0)) + P(D(1) = D(0) = 1) \\ &\quad - P(D(1) < D(0)) - P(D(1) = D(0) = 1) \\ &= P(D(1) > D(0)) - P(D(1) < D(0))\end{aligned}$$

$$\frac{\text{Numerator}}{\text{Denominator}} = \frac{\mathbb{E}\left[(Y(D=1) - Y(D=0))\left(\mathbf{1}\{D(1) > D(0)\} - \mathbf{1}\{D(1) < D(0)\}\right)\right]}{P(D(1) > D(0)) - P(D(1) < D(0))}$$

$$\text{Apply monotonicity:} \quad \mathbf{1}\{D(1) < D(0)\} = 0 \quad \Rightarrow \quad P(D(1) < D(0)) = 0$$

$$= \frac{\mathbb{E}\left[(Y(D=1) - Y(D=0))\,\mathbf{1}\{D(1) > D(0)\}\right]}{P(D(1) > D(0))}$$

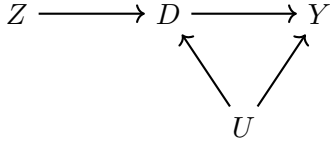
$$= \mathbb{E}\left[(Y(D=1) - Y(D=0)) \mid D(1) > D(0)\right]$$

where $D(1) > D(0)$ is an unidentifiable subgroup.

In the context of the Minneapolis Domestic Violence Experiment:

- $Y =$ re-offense
- $Z =$ random assignment to arrest or counseling
- $D = 1$: Arrest, occurs with probability $1/3$
- $D = 0$: Counseling/Separation, occurs with probability $2/3$

$$\tau_c = \frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0]}$$



Among compliers, $Z \to Y$ and $Z \to D \to Y$

2.15.3 Classical IV Models

Pre-1990 Parametric Two-Stage Model

$$D = \alpha_0 + \alpha_1 Z + \gamma \tag{A}$$

$$Y = \beta_0 + \beta_1 D + \epsilon \tag{A}$$

where ϵ and γ are random error terms.

Z is the IV that follows:

1. Relevance:

$$\text{Cov}(D, Z) \neq 0 \Rightarrow \alpha_1 \neq 0$$

2. Exclusion Restriction: Z does not appear in the second-stage OLS regression, meaning Z affects Y only through D.
3. Unconfoundedness:

$$\text{Cov}(Z, Y) = 0 \quad \text{and} \quad \text{Cov}(Z, \epsilon) = 0$$

With covariates, the models become:

$$D = \alpha_0 + \alpha_1 Z + \alpha_2 X^\top + \gamma$$

$$Y = \beta_0 + \beta_1 D + \beta_2 X^\top + \epsilon$$

with the assumptions:

$$\text{Cov}(Z, \epsilon \mid X) = 0 \quad \text{and} \quad \text{Cov}(Z, \gamma \mid X) = 0$$

These models are causal if the structural assumptions hold.

Structural Equation Model

$$D(Z) = \alpha_0 + \alpha_1 Z + \gamma \tag{B}$$

$$Y(Z, D) = \beta_0 + \beta_1 D + \epsilon \tag{B}$$

Theorem: Under the assumptions of SUTVA (Stable Unit Treatment Value Assumption) and the standard IV conditions:

1. Relevance: $Z \not\perp\!\!\!\perp D$
2. Exclusion Restriction: $Y(Z = z, D = d) = Y(D = d)$
3. Exchangeable IV: $Z \perp\!\!\!\perp (D(Z), Y(Z))$.

The structural equation model (B) implies the observed model (A) under the following assumptions:

1. Relevance: $\text{Cov}(D, Z) \neq 0$.
2. Exclusion Restriction: Z does not appear in the second-stage OLS regression.
3. Unconfounded IV: $\text{Cov}(Z, Y) = 0$ and $\text{Cov}(Z, \epsilon) = 0$.

Under the structural model, $\beta_1 = \mathbb{E}[Y(D = 1) - Y(D = 0)]$. Additionally, under IV, $\beta_1 = \frac{\text{Cov}(Y,Z)}{\text{Cov}(D,Z)}$ (see the "Wald Estimand" from Wald (1940)). The proof of this is as follows:

Proof 2.2

$$D = \alpha_0 + \alpha_1 Z + \nu$$

$$Y = \beta_0 + \beta_1 D + \varepsilon$$

$$\text{Cov}(Y, Z) = \text{Cov}(\beta_0 + \beta_1 D + \varepsilon, Z) = \beta_1 \text{Cov}(D, Z) + \text{Cov}(\varepsilon, Z) = \beta_1 \text{Cov}(D, Z),$$

where in the last step we used the unfounded IV assumption that $\text{Cov}(\varepsilon, Z) = 0$. This gives us the result that (by dividing by $\text{Cov}(D, Z)$):

$$\beta_1 = \frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)},$$

which is valid due to relevance since $\text{Cov}(D, Z) \neq 0$.

2.15.3.1 2 Stage Estimate of β_1

We now describe the 2-stage least squares estimate of β_1.

1. OLS regress $D \sim Z$. This gives predictions for $\hat{D} = \hat{\alpha}_1 Z$.
2. OLS regress $Y \sim \hat{D}$, where $\hat{D} = \hat{\alpha}_1 Z$. This gives a predicted $\hat{\beta}_1$.

This procedure leads to the following theorem: Under the 2 stage model with $\text{Cov}(D, Z) \neq 0$ and $\text{Cov}(Z, \nu) = \text{Cov}(Z, \varepsilon) = 0$, β_1 is the coefficient in a population regression of Y on the predicted value from a population regression of D on Z.

As a side note, we can also control for covariates by conditioning on them:

$$\beta_1 = \frac{\text{Cov}(Y, Z|X)}{\text{Cov}(D, Z|X)}$$

The two stage approach with covariates becomes:

1. OLS regress $D \sim Z, X$, giving predicted $\hat{D} = \hat{\alpha}_1 Z + \hat{\alpha_2}^T X$
2. OLS regress $Y \sim \hat{D}$ yielding $\hat{\beta}_1$.

This 2 stage model is easy to implement, but it has strong parametric assumptions, such as linearity, additivity, and constant effects.

Putting all together through the lens of CACE, the average causal effect under compliers is:

$$\beta_1 = \frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)} = \frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]} =$$

And the right hand side can be written as (focusing on compliers):

$$= \mathbb{E}[Y(D = 1) - Y(D = 0)|D(1) > D(0)]$$

. We reiterate that this assumes monotonicity. Additionally, remember that this is untestable, since we cannot observe whether $D_i(1) \geq D_i(0)$.

2.15.3.2 Some Practical Considerations

In practice, several challenges arise when using preference-based instruments. For example, consider the random assignment of cases to judges (Z) as an instrument. Here, D represents whether there is a pre-trial detention, and Y indicates whether there is a future offense. If $Z = 1$, it implies that the case was randomly assigned to a harsh judge. However, it is not always possible to assume that $D_i(1) \geq D_i(0)\ \forall i$, as there may be exceptions. For instance, a harsh judge might show leniency in specific cases due to personal biases or "soft spots" for particular defendants.

Another practical issue is the limited policy relevance of analyzing effects for unidentified subgroups. Eaton (2010) critiques this approach, stating:

> "This goes beyond the story of looking for an object where the light is strong enough to see. Rather, we have control over the light, but we choose to let it fall where it may and then proclaim that whatever the light illuminates is what we were looking for all along."

This highlights the importance of a structured approach: (1) defining the estimand, (2) ensuring identification, and (3) performing estimation.

In contrast, a policy-relevant analysis focuses on identifiable subgroups. For example, Kennedy et al. (2019) suggest that predicting which individuals are likely to comply has practical significance. Formally, this can be expressed as:

$$\Pr(D(1) > D(0)) = \mathbb{E}[D|X, Z = 1] - \mathbb{E}[D|X, Z = 0].$$

Such an approach ensures that findings align with actionable insights, making them more useful for policymaking.

2.16 Lecture 16: CACE, IV Methods, and Overlap Violations in RDD

Minji, Hyemin Park, Jiarong Zhou(revisions)

2.16.0.1 Key Definitions:

- Z_i: Treatment assigned to unit i (1 for treatment, 0 for control).
- D_i: Treatment received by unit i (1 for treatment, 0 for control).
- Y_i: Outcome of interest for unit i.

2.16.1 CACE (Complier Average Causal Effect)

The average causal effect of the treatment for individuals who comply with the assigned treatment.

2.16.1.1 Estimation of CACE

To estimate the CACE under IV assumptions and monotonicity, the formula is given as:

$$\mathbb{E}\left[y(D=1) - y(D=0) \mid D(1) > D(0)\right] =$$

This can be broken down as follows:

$$= \frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0]} = \frac{\mathbb{E}\left[\mathbb{E}[Y \mid X, Z=1] - \mathbb{E}[Y \mid X, Z=0]\right]}{\mathbb{E}\left[\mathbb{E}[D \mid X, Z=1] - \mathbb{E}[D \mid X, Z=0]\right]}$$

2.16.1.2 How to estimate?

- regression: $\mu_Z(x) = \mathbb{E}[y \mid X, Z=z]$
- weighting: $\lambda_Z(x) = \mathbb{E}[D \mid X, Z=z]$
- doubly-robust: $e(X) = P(Z=1 \mid X)$

$$\hat{\mu}_2(x) \text{ by regress } y \sim X \mid Z=z$$

$$\hat{\lambda}_z(x) \text{ by regress } D \sim X \mid Z=z$$

2.16.2 Estimators

- **Regression Estimator**:

$$\hat{\tau}_{i,reg} = \frac{\sum_{i=1}^{n} \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}}{\sum_{i=1}^{n} \left\{\hat{\lambda}_1(X_i) - \hat{\lambda}_0(X_i)\right\}}$$

  This estimator uses the regression estimates for the potential outcomes and the treatment assignment, providing a way to compute the average causal effect.

- **Inverse Probability Weighting (IPW) Estimator**:

$$\hat{\tau}_{i,IPW} = \frac{\sum_{i=1}^{n} \left\{ \frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)} \right\}}{\sum_{i=1}^{n} \left\{ \frac{Z_i D_i}{\hat{e}(X_i)} - \frac{(1-Z_i)D_i}{1-\hat{e}(X_i)} \right\}}$$

  This estimator accounts for bias by weighting the observed outcomes according to the propensity scores.

- **Doubly Robust Estimator**:

$$\hat{\tau}_{i,DR} = \frac{\sum_{i=1}^{n} \left\{ \left(\frac{Z_i}{\hat{e}(X_i)} - \frac{(1-Z_i)}{1-\hat{e}(X_i)} \right) (Y_i - \hat{\mu}_z(X_i)) + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right\}}{\sum_{i=1}^{n} \left\{ \left(\frac{Z_i}{\hat{e}(X_i)} - \frac{(1-Z_i)}{1-\hat{e}(X_i)} \right) \left(D_i - \hat{\lambda}_z(X_i) \right) + \hat{\lambda}(X_i) - \hat{\lambda}_0(X_i) \right\}}$$

  This estimator combines both regression and weighting approaches, ensuring validity even if one model is misspecified.

2.16.2.1 Question to Consider

Should you always use the same covariate to condition on D and Y?

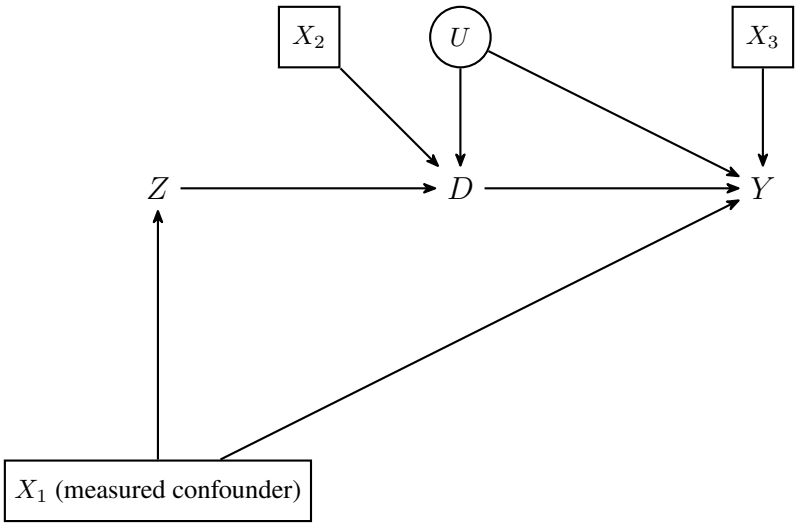
2.16.2.2 Graphical Interpretation

The graph outlines the relationships where:

- X_1: Represents measured confounders that affect both Z and Y.

- X_2: Influences treatment D but not directly associated with the outcome.

- X_3: Affects outcome Y without influencing treatment D.



- We should always include X_1 in λ and μ regressions to achieve exchangeability.
- You can additionally condition on X_2 in λ regress for efficiency.
- You can condition on X_3 in μ regression for efficiency.
- If we condition on X_1, X_2, and X_3 in both λ and μ, that is fine.

2.16.2.3 Question

Which among the regression, IPW, and DR estimators should we use?

- If $e(X)$ is known: Use IPW or DR.
- If $e(X)$ is unknown but X is discrete and low-dimensional, any estimator is good. They are numerically equivalent and optimally efficient.
- If $e(X)$ is unknown and X includes continuous components or is high-dimensional, choose the doubly robust (DR) estimator.

Additional Explanation:

In experimental settings where the instrument propensity $e(X)$ is known (as determined by the experimental design), using the IPW estimator is generally robust. The robust estimator is also suitable because it utilizes the true $e(X)$ while also incorporating estimated models.

In cases where $e(X)$ is unknown but X is discrete and low-dimensional, we can estimate nuisance functions using empirical distributions, making any estimator effective.

However, when $e(X)$ is unknown and involves continuous components or is high-dimensional, the doubly robust estimator is preferable. It offers better theoretical properties under milder assumptions, allowing it to perform well even if only some of the nuisance functions are estimated accurately.

2.16.3 Beyond the CACE : Partial Identification of Average Causal Effect

The average causal effect τ is defined as:

$$\tau = E[Y(D=1)] - E[Y(D=0)]$$

This represents the expectation if we treated everyone versus treating no one.

Under the monotonicity and exclusion restriction, we can establish bounds:

$$\tau_L \leq \tau \leq \tau_U$$

Where:

$$\tau_L = E[Y|Z=1] - E[Y|(1-D)+D|Z=0]$$

$$\tau_U = \tau_L + 1 - E[D|Z=1] + E[D|Z=0]$$

The length of the bound is determined by the proportion of non-compliers:

$$1 - (E[D|Z=1] - E[D|Z=0]) = P(D(1) \leq D(0))$$

This formulation shows that we can partially identify the average causal effect, giving us lower and upper bounds

based on observable quantities that can be estimated from the data.

2.16.4 Examples of IVs (Instrumental Variables)

2.16.4.1 Challenge: Identifying and Justifying the IV

One of the challenges with using instrumental variables (IVs) is identifying a valid instrument and justifying its use. The instrument must satisfy key assumptions, including the exclusion restriction, which states that it affects the outcome only through the treatment.

2.16.4.2 Experiments with Non-compliance

Example: Minneapolis Domestic Violence Experiment - Z (Instrument): The penalty recommendation by officers responding to a domestic violence incident. This recommendation was randomly assigned. - D (Treatment): Whether the officer adhered to the recommended penalty. - Y (Outcome): Whether the individual re-offended.

This experimental setup involves non-compliance, as officers had the discretion to deviate from the randomly assigned recommendation. In such cases, Instrumental Variable (IV) methods are particularly useful to estimate the causal effect of the recommendation on the likelihood of re-offending.

2.16.4.3 Distance-based Measures

Example: Family Visitation and Re-offense Rates (Mauro et al., 2008) - Z (Instrument): The proximity of an inmate's family to the jail. - D (Treatment): Whether the family visited the inmate. - Y (Outcome): The re-offense rate after release from jail.

This approach assumes that the distance between the family's residence and the jail affects the likelihood of visitation but does not directly influence the re-offense rate except through its impact on visitation. By using distance as an instrument, it becomes possible to estimate the causal effect of family visits on post-release behavior.

2.16.4.4 Preference-based Measures

Example :

Z (Instrument): Doctor's preference for prescribing Paxlovid (an antiviral drug).

D (Treatment): Whether the patient was prescribed Paxlovid.

Y (Outcome): Potential liver damage as a side effect.

The instrument is the variation in doctors' prescribing behavior in this case. Doctors with different preferences might prescribe Paxlovid differently, and this variation can be exploited as an IV to study the effect of Paxlovid on liver damage.

2.16.4.5 Time-based Measures (Change in Policy)

Example :

Z (Instrument): Expansion in the mandated reporter law in 2014, which increased the number of people required to report cases of child abuse or neglect.

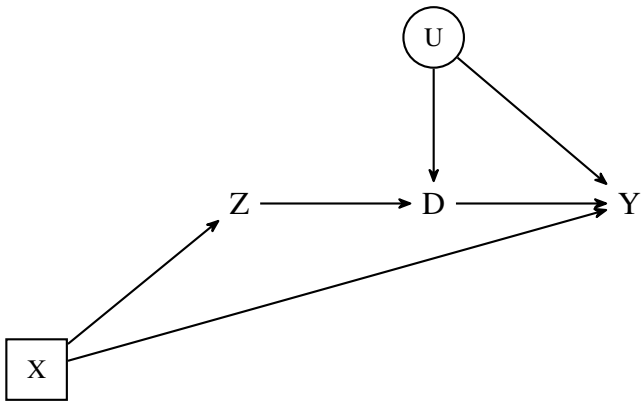
D (Treatment): Whether a child welfare investigation was opened.

Y (Outcome): Educational outcomes of children in the family.

Here, a policy change creates variation in reporting, which serves as an instrument for the likelihood of child welfare investigations. This variation allows researchers to study the causal effect of an investigation on children's education.

$$X \to Z \to D \to Y$$

Where X are covariates that may affect both the treatment (D) and the outcome (Y), while Z (the instrument) affects D but only affects Y indirectly through D.



2.16.4.6 Genes: Mendelian Randomization

- **Mendel's 2nd law:** The law of random assortment states that the inheritance of one trait is independent of other traits.
- Mendelian randomization uses genetic variation as an IV to study the causal effects of exposure (e.g., cholesterol levels) on an outcome (e.g., cancer).

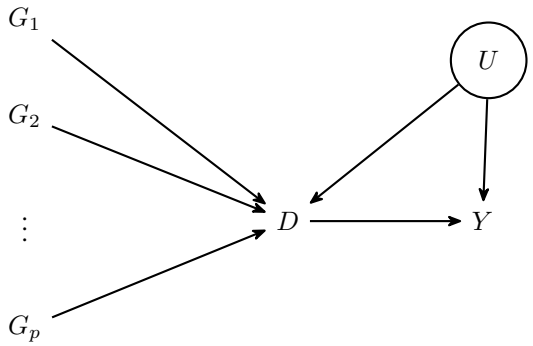
Example :

Z (Instrument): Genes

D (Treatment): Cholesterol levels

Y (Outcome): Likelihood of developing cancer

In Katan (1986), apolipoprotein E genes were used as an instrument to study the effect of cholesterol levels on cancer risk. The assumption here is that genes affect cholesterol levels but do not directly affect cancer, satisfying the exclusion restriction.



$$G_1, G_2, \ldots, G_p \to D \to Y$$

Where $G_1, G_2, \ldots, G_p$ represent different genetic variants (instruments), D represents cholesterol levels, and Y represents the outcome (cancer).

$$G_1, G_2, \ldots, G_p \to U$$

Where U represents unobserved confounders.

Standard Linear IV Model:

$$Y = \beta_0 + \beta D + \beta_\mu U + \epsilon_Y$$

$$D = \gamma_0 + \gamma_1 G_1 + \cdots + \gamma_p G_p + \gamma_U U + \epsilon_D$$

Here, Y is the outcome, D is the treatment, $G_1, G_2, \ldots, G_p$ are the genetic instruments, U is an unobserved confounder, and ϵ_Y and ϵ_D are error terms.

Reduced Form:

$$Y = \beta_0 + \beta_D\gamma_0 + \beta_D\gamma_1 G_1 + \cdots + \beta_D\gamma_p G_p + (\beta_U + \beta_D\lambda_U)U + \epsilon_Y$$

Apply Two-Stage Least Squares Estimators to Estimate β: See Chapter 25 of *Ding* for further details.

Critiques of Mendelian Randomization Analyses

Mendelian Randomization is a powerful tool, but there are several potential issues to consider:

- **SUTVA (Stable Unit Treatment Value Assumption)** or **target trial emulation** may not be satisfied
  - For example, treatment variables like BMI or cholesterol levels may not satisfy SUTVA.
  - There may be situations where the treatment status of one individual affects the outcome of another, or where the treatment is not well-defined.
- **Exclusion restriction** may be violated
  - There may be alternative pathways through which the genes affect the outcome. For example, genes may directly affect cancer development, in addition to affecting cholesterol levels, which would violate the exclusion restriction.
- **Linearity assumption in the IV model** may be violated
  - Mendelian Randomization typically assumes a linear relationship between the treatment and the outcome, but in reality, the relationship may be nonlinear.
- **Measurement error** may exist
  - The treatment variable (e.g., cholesterol levels) or the outcome variable (e.g., cancer diagnosis) may not be measured accurately, which could lead to biased results.

2.16.5 Violations of Overlap / Positivity

2.16.5.1 Overlap Condition:

$$P(0 < e(X) < 1) = 1$$

The overlap or positivity assumption requires that all units have some probability of receiving either treatment or control (i.e., no deterministic treatment assignment). Here. $e(X)$ represents the propensity score.

2.16.5.2 Tension between Overlap and Exchangeability:

There is a tension between the overlap assumption and the exchangeability assumption. As we condition on more covariates to satisfy exchangeability, we may reduce the randomness in the treatment assignment, which can lead to a violation of overlap.

Example: Child Welfare Investigation

X: Administrative records, allegations.

D: Child welfare investigation

In this context, administrative records and the natural language of allegations are used to determine whether a child welfare investigation is opened. Conditioning on these covariates may improve exchangeability but can reduce overlap if these features deterministically affect the treatment decision.

2.16.6 Conceptual Challenge: Deterministic Treatment Assignment and Counterfactuals

In settings where overlap is violated (i.e., units deterministically receive treatment or control), conceptual challenges arise. For instance, what does it mean to consider a counterfactual outcome for a unit that always receives treatment? This raises questions about the definition of potential outcomes for such deterministic units.
To address this, overlap requires that propensity scores be bounded away from 0 and 1:

$$\epsilon < e(X) < 1 - \epsilon$$

where ϵ is a small positive constant. This boundedness assumption is necessary for many causal inference techniques, particularly in high-dimensional settings.

2.16.6.1 Handling Overlap Violations: Trimming

One common approach to handle overlap violations is trimming, where observations with extreme propensity scores are removed from the analysis. The downside is that this changes the estimand, meaning we are no longer estimating the causal effect for the entire population but rather for a subset where overlap holds.

2.16.7 Regression Discontinuity Design (RDD)

When overlap is violated, an alternative approach is to use a **regression discontinuity design (RDD)**. RDD is useful in settings with a threshold-based decision rule, where the decision threshold can be considered somewhat arbitrary. By examining outcomes near the threshold, we can assume that the groups on either side are exchangeable, and we can estimate causal effects by comparing outcomes just above and below the threshold.

2.16.7.1 Example: COVID-19 Treatment and Oxygen Levels

Consider a new antiviral COVID-19 treatment. We want to measure its impact on oxygen levels (a continuous outcome) one week after treatment. During the pandemic, this treatment was restricted to high-risk patients, specifically those aged 65 and older.

- **Treatment Group**: Patients aged 65 and older.
- **Control Group**: Patients under 65.

Since there is no randomness in who receives the treatment (i.e., there is no overlap), we cannot apply typical causal inference methods. However, we can use RDD by assuming continuity of the outcome regression functions around the threshold (age 65).

2.16.7.2 Assumptions for RDD:

The key assumption in RDD is that the outcome regression functions are continuous at the threshold:

$$\lim_{X \to 65^-} E[Y|X] = \lim_{X \to 65^+} E[Y|X]$$

where X is the age variable. Under this assumption, the difference in outcomes around age 65 can be attributed to the causal effect of the treatment.

Interpretation The causal effect of the treatment is estimated by the difference in the observed means just above and below the threshold:

$$\text{Causal Effect} = E[Y|X = 65^+] - E[Y|X = 65^-]$$

In this example, we would observe oxygen levels (Y) for patients just below 65 (control group) and just above 65 (treatment group) and interpret the difference in means as the causal effect of the COVID-19 treatment.

Next Class

How to implement a regression discontinuity analysis in practice.

Reminder: there is a midterm exam scheduled for next Tuesday.

2.17 Lecture 17: Sharp Regression Discontinuity Design (RDD)

Carlos Guirado, Stella Jia, Kayla Sim(revisions)

2.17.1 Regression Discontinuity Design (RDD)

2.17.1.1 Introduction and Motivating Example

In the last lecture, Regression Discontinuity (RD) was introduced through a COVID treatment example:

- Treatment (D): New antiviral COVID treatment
- Outcome (Y): Patient oxygen levels
- Running variable (X): Age

- Policy: Treatment restricted to those 65 and older

This creates a sharp discontinuity in treatment probability:

$$P(D = 1|X \geq 65) = 1$$

$$P(D = 1|X < 65) = 0$$

2.17.1.2 Exchangeability in RD

A key feature of RD is that exchangeability holds by design. This can be understood by examining the causal structure:

1. The running variable X (age) determines treatment D:

$$D = \mathbf{1}[X \geq x_0]$$

2. Traditional settings require:
   - Both exchangeability and overlap
   - D as a function of X and other variables V:

$$D = f(X, V)$$

   - V affects D but not Y for overlap to hold
3. In sharp RD:
   - Overlap does not hold
   - D is determined solely by X (i.e. deterministic)
   - Conditioning on X eliminates all unobserved confounding
   - Exchangeability holds automatically due to treatment assignment mechanism

2.17.1.3 Examples of RD in Practice

1. Thistlethwaite & Campbell (1960) - First RD study:
   - D: Certificate of merit
   - Treatment rule:

$$D = \mathbf{1}[\text{test score} \geq 10.5]$$

   - Y: Plans for graduate study and scientific research
   - Demonstrated how seemingly arbitrary cutoffs can be leveraged for causal inference
2. Carpenter & Dobkin (2009):
   - D: Legal drinking status

- Treatment rule:

$$D = \mathbf{1}[\text{age} \geq 21]$$

- Y: Death outcomes
- Examined different types of mortality:
  - Overall mortality
  - Alcohol-related deaths
  - External vs. internal causes of death

2.17.1.4 Formal Treatment and Identification

RD identifies the local average treatment effect at the threshold:

$$\tau(x_0) = \mathbb{E}[Y(1) - Y(0)|X = x_0]$$

The identification strategy differs from traditional approaches:

1. For treated potential outcome:

$$\begin{aligned}\mathbb{E}[Y(1)|X = x_0] &= \lim_{\epsilon \to 0^+} \mathbb{E}[Y(1)|X = x_0 + \epsilon] \\ &= \lim_{\epsilon \to 0^+} \mathbb{E}[Y(1)|D = 1, X = x_0 + \epsilon] \\ &= \lim_{\epsilon \to 0^+} \mathbb{E}[Y|D = 1, X = x_0 + \epsilon]\end{aligned}$$

2. For control potential outcome:

$$\mathbb{E}[Y(0)|X = x_0] = \lim_{\epsilon \to 0^+} \mathbb{E}[Y|D = 0, X = x_0 - \epsilon]$$

3. The treatment effect:

$$\tau(x_0) = \lim_{\epsilon \to 0^+} \mathbb{E}[Y|D = 1, X = x_0 + \epsilon] - \lim_{\epsilon \to 0^+} \mathbb{E}[Y|D = 0, X = x_0 - \epsilon]$$

2.17.1.5 Local Linear Regression Approach

Implementation typically uses local linear regression:

1. Basic procedure:
   - Select observations near threshold $x = x_0$
   - Fit separate linear models on each side of the threshold
   - Compare fitted values at threshold
2. Critical bandwidth choice considerations:
   - Too small (e.g., 1 day from threshold):

        - Few data points
        - High variance in estimates
    - Too large (e.g., 2 years from threshold):
        - Bias from comparing non-exchangeable individuals
        - Other factors may vary over wider window
3. Practical recommendations:
    - Report estimates for multiple bandwidths
    - Include confidence intervals
    - Use `rdrobust` package in R
    - Consider covariate adjustment for wider bandwidths

2.17.1.6 Limitations and Potential Problems

1. Data sensitivity:
    - Results vulnerable to leverage points near cutoff
    - Individual observations can heavily influence slope estimates
2. Continuity assumption:
    - Requires continuity in potential outcome regression functions
    - Not directly testable as counterfactual outcomes unobservable
    - $\mu_1(x)$ and $\mu_0(x)$ not fully observable
3. Multiple threshold effects:
    - Other treatments/policies may change at threshold
    - Example: Medicare eligibility at age 65
    - Complicates isolation of specific treatment effect
    - May affect interpretation of results
4. Bandwidth selection challenges:
    - Tradeoff between bias and variance
    - Results may be sensitive to choice
    - Need to demonstrate robustness across choices

2.17.1.7 Comparison to Traditional Experiments

When considering RD designs, it's useful to think about the ideal experiment:

- Traditional experiment might randomize treatment directly

- RD approximates experiment around threshold
- Treatment effect identified only for subpopulation near threshold
- Requires additional assumptions for broader generalization

Class Survey

This lecture was shorter than usual as students were given time to complete a course evaluation survey.

Next Class

Fuzzy regression discontinuity design.

2.18 Lecture 18: Fuzzy Regression Discontinuity Design (RDD)

Harish Srinivasan, Aditya Vunnum, Simon Cha, Kayla Sim (revisions)

2.18.1 Wrapping Up Regression Discontinuity

2.18.1.1 Sharp RDD Recap

In the previous lecture, we explored the concept of Sharp Regression Discontinuity Design (RDD), which uses a strict cutoff point in the running variable to determine treatment assignment.

$$D = \mathbf{1}[X \geq x_0]$$

A key example was a COVID treatment policy:

- **Treatment (D)**: Administration of a new antiviral COVID treatment.
- **Outcome (Y)**: Patient oxygen levels.
- **Running Variable (X)**: Age.
- **Policy Rule**: Treatment was provided only to patients aged 65 and older.

This setup created a sharp discontinuity in treatment probability:

$$P(D = 1|X \geq 65) = 1, \quad P(D = 1|X < 65) = 0$$

The sharp cutoff at age 65 made treatment assignment deterministic based on age alone, allowing for causal inference by comparing outcomes immediately on either side of the threshold.

Exchangeability holds by design in Sharp RDD. Since treatment is fully determined by the running variable, conditioning on the running variable (age in this case) controls for unobserved confounding. This setup does

not require overlap between treatment and control groups, as assignment is non-random and based purely on the cutoff.

2.18.1.2 Transition to Fuzzy RDD

In practice, strict cutoff rules are not always feasible, leading to the development of Fuzzy RDD. In Fuzzy RDD, the probability of treatment changes at the threshold, but treatment is not strictly assigned based on the cutoff. Instead, the threshold increases the likelihood of treatment, introducing "fuzziness."

2.18.2 Fuzzy Regression Discontinuity Design (RDD)

2.18.2.1 Introduction to Fuzzy RDD

Fuzzy RDD relaxes the strict assignment of treatment, allowing for a probabilistic jump in treatment assignment at the threshold. This is useful in settings where the treatment may be recommended but not enforced strictly based on the running variable.

Today's example uses a treatment indicator Z defined as:

$$Z = \mathbf{1}[X \geq x_0]$$

where X is the running variable and x_0 is the cutoff. With Fuzzy RDD, the probability of treatment $P(D = 1|X = x)$ has a jump at $X = x_0$, but treatment is not assigned deterministically.

2.18.2.2 Examples of Fuzzy RDD in Practice

2.18.2.2.1 Example 1: College Admissions

- A fuzzy cutoff is applied with SAT scores around 1300.
- Colleges may admit students who score below the threshold if they excel in other areas, and may reject students above the threshold if their grades are poor.

2.18.2.2.2 Example 2: Eligibility for Social Assistance

- A social assistance program uses a predicted income threshold $X \leq x_0$ to determine eligibility.
- Not all eligible families enroll, and other eligibility criteria may apply.

The estimand for the local complier average causal effect in this setup is:

$$\tau_c(x_0) = \mathbb{E}\left[Y(D = 1) - Y(D = 0) \mid D(1) > D(0), X = x_0\right]$$

2.18.3 Theorem for Fuzzy RDD

Under the following conditions:

1. Monotonicity: $D(1) \geq D(0)$
2. Exclusion restriction: $(Y(Z = z, D = d) = Y(D = d))$

the local complier average causal effect is given by:

$$\tau_c(x_0) = \frac{\mathbb{E}\left[Y(D=1) - Y(D=0) \mid X = x_0\right]}{\mathbb{E}\left[D(1) - D(0) \mid X = x_0\right]}$$

Under continuity conditions, and if $P(D = 1|X = x)$ has a jump at x_0, then $\tau_c(x_0)$ is identified as:

$$\tau_c(x_0) = (\lim_{\epsilon \to 0^+} \mathbb{E}[Y|Z = 1, X = x_0 + \epsilon] - \lim_{\epsilon \to 0^+} \mathbb{E}[Y|Z = 0, X = x_0 - \epsilon])$$
$$\div(\lim_{\epsilon \to 0^+} \mathbb{E}[D|Z = 1, X = x_0 + \epsilon] - \lim_{\epsilon \to 0^+} \mathbb{E}[D|Z = 0, X = x_0 - \epsilon])$$

For more details, see Chapter 24 of *Imbens & Lemieux (2008)*.

2.18.4 Causal Inference under Unobserved Confounding

2.18.4.1 Sensitivity to Ignorability Assumption

It's essential to understand how sensitive results are to the assumption of ignorability.

2.18.4.1.1 Example: Smoking and Lung Cancer

- Doll and Hill (1950) found that the risk ratio for smoking on lung cancer was 9.
- However, Fisher (1957) proposed that a common genetic cause might be responsible for both smoking and lung cancer.

The sensitivity parameter describes the strength of this unobserved confounding.

2.18.4.2 Sensitivity Parameters

For binary Y and binary U:

$$Z \not\perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$$

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X, U$$

Define two sensitivity parameters:

1. $\text{RR}_{ZU|X} = \frac{P(U=1|Z=1,X=x)}{P(U=1|Z=0,X=x)}$
2. $\text{RR}_{YU|X} = \frac{P(Y=1|U=1,X=x)}{P(Y=1|U=0,X=x)}$

2.18.4.3 Observed Risk Ratio

The observed risk ratio is given by:

$$\text{RR}_{ZY|X}^{obs} = \frac{P(Y=1|Z=1,X=x)}{P(Y=1|Z=0,X=x)}$$

Under unmeasured confounding, the true risk ratio differs:

$$\text{RR}_{ZY|X}^{true} = \frac{P(Y(1)=1|X=x)}{P(Y(0)=1|X=x)}$$

2.18.4.4 Theorem for Sensitivity Analysis

The observed risk ratio under unobserved confounding is bounded by:

$$\text{RR}_{ZY|X}^{\text{obs}} \leq \frac{\text{RR}_{ZU|X} \cdot \text{RR}_{YU|X}}{\text{RR}_{ZU|X} + \text{RR}_{YU|X} - 1}$$

Assuming $Z \perp Y \mid (X, U)$ and, without loss of generality, that:

$$\text{RR}_{ZY|X}^{\text{obs}} > 1, \quad \text{RR}_{ZU|X} > 1, \quad \text{RR}_{YU|X} > 1.$$

2.18.4.5 Implications of the Theorem

The theorem implies that:

$$\text{RR}_{ZU|X} \geq \text{RR}_{ZY|X}^{\text{obs}}$$

$$\text{RR}_{YU|X} \geq \text{RR}_{ZY|X}^{\text{obs}}$$

This is known as the **Cornfield inequality**.

To explain away the observed relative risk, both confounding measures $\text{RR}_{ZU|X}$ and $\text{RR}_{YU|X}$ must be at least as large as $\text{RR}_{ZY|X}^{\text{obs}}$.

2.18.4.6 Bounding the Average Causal Effect without Sensitivity Parameters

Assume $\underline{Y} \leq Y \leq \overline{Y}$, meaning the outcome is bounded. The average causal effect can be bounded as:

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$

$$\tau \leq \overline{Y} - \underline{Y}$$

If τ is partially identified, multiple values of τ are compatible with the observed data distribution, as discussed in *Manski (1990, 2003)*. However, bounds typically cover zero.

More informative bounds can be obtained when combined with other assumptions, such as monotonicity:

$$Y(1) \geq Y(0) \Rightarrow \epsilon_0 \leq 0$$

Define the sensitivity parameters:

$$\epsilon_1(X) = \frac{\mathbb{E}[Y(1) \mid Z = 1, X]}{\mathbb{E}[Y(1) \mid Z = 0, X]}$$

$$\epsilon_0(X) = \frac{\mathbb{E}[Y(0) \mid Z = 1, X]}{\mathbb{E}[Y(0) \mid Z = 0, X]}$$

2.18.4.7 Theorem

$$\mathbb{E}[Y(1) \mid Z = 0] = \mathbb{E}\left[\frac{\mathbb{E}[Y \mid Z = 1, X]}{\epsilon_1(X)} \mid Z = 0\right]$$

$$\mathbb{E}[Y(0) \mid Z = 1] = \mathbb{E}\left[\mathbb{E}[Y \mid Z = 0, X]\epsilon_0(X) \mid Z = 1\right]$$

where $\epsilon_1(X)$ and $\epsilon_0(X)$ are sensitivity parameters.

2.18.5 Estimators

2.18.5.1 Estimators

Define the following estimators:

1. $\hat{\tau}_{\text{ht}}$:

$$\hat{\tau}_{\text{ht}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i Y_i\left(\hat{e}(X_i) + (1 - \hat{e}(X_i))/\epsilon_1(X_i)\right)}{\hat{e}(X_i)} - \frac{\hat{e}(X_i)\epsilon_0(X_i) + (1 - \hat{e}(X_i))(1 - Z_i)Y_i}{1 - \hat{e}(X_i)}\right)$$

2. $\hat{\tau}_{\text{haj}}$:

$$\hat{\tau}_{\text{haj}} = \frac{\sum_{i=1}^{n} \frac{\hat{e}(X_i)\epsilon_0(X_i) + (1-\hat{e}(X_i))(1-Z_i)Y_i}{1-\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{1-Z_i}{1-\hat{e}(X_i)}}$$

3. $\hat{\tau}_{\text{OR}}$:

$$\hat{\tau}_{\text{OR}} = \hat{\tau}_{\text{HT}} - \frac{1}{n}\sum_{i=1}^{n}(Z_i - \hat{e}(X_i))\left(\frac{\hat{\mu}_1(X_i)}{\hat{e}(X_i)\epsilon_1(X_i)} + \frac{\hat{\mu}_0(X_i)\epsilon_0(X_i)}{1 - \hat{e}(X_i)}\right)$$

2.18.6 Calibrating Sensitivity Parameters Using Observed Covariates

Using a leave-one-out approach:

$$\epsilon_z(X_{-j}) = \frac{\mathbb{E}[Y(z)|Z = 1, X_{-j}]}{\mathbb{E}[Y(z)|Z = 0, X_{-j}]}$$

Under $Z \perp Y(z) \mid X$, we have:

$$\epsilon_z(X) = \exp(\alpha z + \beta^T z)$$

2.19 Lecture 19: Sensitivity Analysis

Isabel Moreno, Jane Chen, Xuanlin Mao,

Kayla Sim(revisions)

2.19.1 Strategies for causal inference under unobserved confounding

2.19.1.1 Cornfield-style inequalities and Manski-style bounds: bound estimand using outcome bounds - recap

In the previous lecture, the first two strategies were introduced:

- 1. **Cornfield-style** inequalities provide a way to bound the *relative risk* (RR) in the presence of unobserved confounding variables. These inequalities express a relationship between observed treatment effects and the potential impact of confounders.

$$RR^{obs}_{zy|x} \leq min(RR_{zu|x}, RR_{uy|x})$$

- 2. **Manski-style bounds** are used to bound the estimand by placing bounds on the outcomes. These can be represented as:

$$\underline{y} \leq y \leq \overline{y}$$

For treatment outcomes:

$$\underline{y} \leq E[y(0)|z = 1] \leq \overline{y}$$

$$\underline{y} \leq E[y(1)|z = 0] \leq \overline{y}$$

2.19.2 Rosenbaum-Style Sensitivity Analysis

Rosenbaum (1987) introduced a sensitivity analysis for one-to-one matched observational studies to test the sharp null hypothesis of no individual treatment effect.

Example: Consider the following variables:

- Z = in-class lecture attendance (treatment),
- Y = grade (outcome),
- X = graduate student, statistics concentration (covariate),
- U = studiousness, passion for the subject, etc. (unobserved confounder).

The sharp null hypothesis asserts that attending the lecture in person has no effect on anyone's grade.

2.19.2.1 Notation and Setup

Let (i, j) index pair j in unit i, where $i = 1, 2, \ldots, n$ and $j = 1, 2$.

$$H_0^f : Y_{ij}(1) = Y_{ij}(0) \quad \text{for} \quad i = 1, 2, \ldots, n, \quad j = 1, 2$$

Let:

$$z_{ij} = \text{treatment assigned to unit } (i, j),$$

$$e_{ij} = P(Z_{ij} = 1 | X_i, Y_{ij}(1), Y_{ij}(0)) \quad \text{(true propensity score)},$$

$$X_i = \text{covariates of the } i\text{th pair (assume perfect matching)}.$$

$$O_{ij} = \frac{e_{ij}}{1 - e_{ij}} \quad \text{(odds of treatment for unit } (i, j)).$$

2.19.2.2 Rosenbaum Sensitivity Analysis Model

The sensitivity model is defined as:

$$\frac{\sigma_{i1}}{\sigma_{i2}} \leq \Gamma, \quad \frac{\sigma_{i2}}{\sigma_{i1}} \leq \Gamma \quad \text{for} \quad i = 1, 2, \ldots, n,$$

which is equivalent to:

$$\frac{1}{1 + \Gamma} \leq \pi_{i1} \leq \frac{\Gamma}{1 + \Gamma} \quad \text{for} \quad i = 1, 2, \ldots, n,$$

Where:

$$\pi_{i1} = P(Z_{i1} = 1 | X_i, Z_{i1} + Z_{i2} = 1, Y_{i1}(0), Y_{i2}(0), Y_{i1}(1), Y_{i2}(1))$$

$$\pi_{i1} = \frac{e_{i1}(1 - e_{i2})}{e_{i1}(1 - e_{i2}) + e_{i2}(1 - e_{i1})}$$

where π_{i1} is the probability of receiving treatment $Z_{i1} = 1$, conditioned on covariates and potential outcomes.

If there is no unobserved confounding, we have:

$$e_{ij} = P(Z_{ij} = 1 | X_i),$$

which leads to:

$$e_{i1} = e_{i2}, \quad \pi_{i1} = \frac{1}{2}, \quad \Gamma = 1 = \frac{\sigma_{i1}}{\sigma_{i2}}.$$

2.19.2.3 Testing the Sharp Null Hypothesis

To test the sharp null hypothesis, we examine the within-pair sign. Under the null hypothesis (H_0^f), the distribution of the signS_i follows a Bernoulli(1/2):

Let

$$\hat{\tau}_i = (2Z_{ij} - 1)(Y_{i1} - Y_{i2})$$

$$S_i \sim \text{Bernoulli}(\pi_{i1}),$$

where the worst-case scenario under the Rosenbaum model is:

$$S_i \sim \text{Bernoulli}\left(\frac{\Gamma}{1+\Gamma}\right).$$

2.19.2.3.1 Test Statistics Two common test statistics are used in matching:

- 1. **Sign Statistic:**

$$T = \sum_{i=1}^{n} S_i.$$

- 2. **Wilcoxon Signed-Rank Statistic:**

$$T = \sum_{i=1}^{n} S_i R_i,$$

  where R_i is the rank of the absolute value of T_i.

Generalization:

$$T = \sum S_i q_i$$

The distribution of T has the following properties:

$$E_\Gamma(T) = \frac{\Gamma}{1+\Gamma} \sum_{i=1}^{n} q_i,$$

$$\text{Var}_\Gamma(T) = \frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{n} q_i^2.$$

The normal approximation of T is given by:

$$\frac{T - \frac{\Gamma}{1+\Gamma} \sum_{i=1}^{n} q_i}{\sqrt{\frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{n} q_i^2}} \to N(0,1).$$

2.19.2.4 Generalizing the Rosenbaum Model Beyond 1-1 Matching

In the general model, the assumption is:

$$Z \not\perp\!\!\!\perp (Y(1), Y(0))|X, \quad Z \perp\!\!\!\perp (Y(1), Y(0))|X, U.$$

The odds ratio model becomes:

$$\frac{1}{\Gamma} \leq \frac{\text{odds}(Z = 1|X, U = u)}{\text{odds}(Z = 1|X, U = u')} \leq \Gamma \quad \text{for any} \quad u, u'.$$

Where Γ is a measure of how far we are from the perfect unconfounded experiment

The marginal sensitivity model is:

$$\frac{1}{\Gamma} \leq \frac{\text{odds}(Z = 1|X, U)}{\text{odds}(Z = 1|X)} \leq \Gamma,$$

which is equivalent to:

$$\frac{1}{\Gamma} \leq \frac{\frac{P(Z=1|X,U=u)}{P(Z=0|X,U=u)}}{\frac{P(Z=1|X)}{P(Z=0|X)}} \leq \Gamma.$$

Once we specify Γ, we can partially identify our estimand under the marginal sensitivity model (MSM) and proceed with the estimation of the partially identified target.

2.20 Lecture 20: Principal Stratification and Mediation

Yixin Feng, Keira Chiu, Allie Huang, Fangyuan Li(revisions)

Announcements:

- No lecture Thursday - BSTARS 1-5
- 256: submit video on Gradescope
- Project signups
- Midterm grades released

Challenge example:

- Officers are as likely to shoot a white person they stop as a black person (Fryer 2019)
- Officers are more likely to use unjustified force on black people
- Example of post-treatment bias:
  - Treatment: race
  - Post-treatment variable: stop
  - Outcome: shoot

Post-treatment Variables

- **Police bias:**
  - Y: use of force
  - Z: race of driver/pedestrian
  - M: stopped driver/pedestrian
- **Job training:**
  - Z: job training
  - Y: wage
  - M: employment status
- **Clinical trial:**
  - Z: HIV treatment
  - Y: 30-year survival
  - M: surrogate endpoints
  - e.g., CD4 cell counts

Completely Randomized Experiment

$$Z = \{Y(0), Y(1), M(0), M(1)\}$$

Goal: Compare $P(Y(1)|M = m)$ **vs** $P(Y(0)|M = m)$

Can we use $P(Y|Z = 1, M = m)$ vs $P(Y|Z = 0, M = m)$?

By CRE:

$$P(Y|Z = 1, M = m) = P(Y(1)|Z = 1, M(1) = m) = P(Y(1)|M(1) = m)$$

$$P(Y|Z = 0, M = m) = P(Y(0)|Z = 0, M(0) = m) = P(Y(0)|M(0) = m)$$

In the challenge example, comparing $P(Y \mid Z = 1, M = m)$ and $P(Y \mid Z = 0, M = m)$ involves two different populations. Specifically, $Y(1)$ represents the outcome for all black drivers, while $M(1) = 1$ corresponds to black drivers who are stopped. Similarly, $Y(0)$ represents the outcome for all white drivers, and $M(0) = 0$ corresponds to white drivers who are stopped. Since these comparisons condition on two completely different populations, it is not possible to determine whether the causal effect is due to the causal relationship between Y and Z, or if it arises because we are not comparing equivalent groups. Comparing causal effects using different populations can introduce bias, as the comparisons may not accurately reflect the true causal relationships.

Another example: truncation by death

Z: severe disease treatment

Y: quality of life

M: survival status

Suppose treatment affects survival status, so treatment can save more weak patients than control.

$$P(Y(1)|M(1)=1) \text{ vs } P(Y(0)|M(0)=1)$$

where the first includes weaker patients, and the second does not.

2.20.1 Principal Stratification: conditioning on potential

2.20.1.1 Values of post-treatment variables (Frangakis and Rubin 2002):

Goal: compare $P(Y(1)|M(1)=m_1, M(0)=m_0)$ vs $P(Y(0)|M(1)=m_1, M(0)=m_0)$ for some m_1, m_0.

- $M(1)$, $M(0)$ are pre-treatment covariates:
- Principal strata defined by potential values of post-treatment variable

Ex: Binary M

Principal strata, severe disease example:

$M(1)=1$, $M(0)=1$ survive always

$M(1)=1$, $M(0)=0$ need treatment to survive

$M(1)=0$, $M(0)=1$ harmed by treatment

$M(1)=0$, $M(0)=0$ doomed

Define:

$$\tau(m_1, m_0) = E[Y(1) - Y(0) \mid M(1)=m_1, M(0)=m_0]$$

as the Principal Stratification Average Causal Effect for subgroup with $M(1)=m_1$ and $M(0)=m_0$.

For binary M, we have four effects:

$$\tau(1,1) = E[Y(1) - Y(0) \mid M(1)=1, M(0)=1]$$

$$\tau(1,0) = E[Y(1) - Y(0) \mid M(1)=1, M(0)=0]$$

$$\tau(0,1) = E[Y(1) - Y(0) \mid M(1)=0, M(0)=1]$$

$$\tau(0,0) = E[Y(1) - Y(0) \mid M(1)=0, M(0)=0]$$

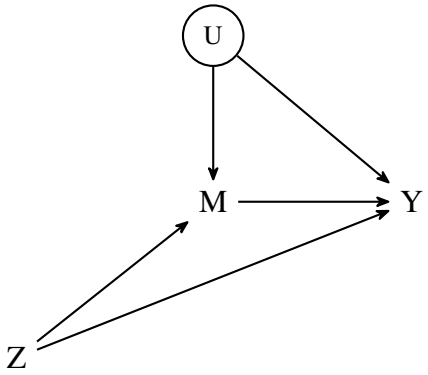
where $\tau(1,1)$ and $\tau(0,0)$ are measures of dissociative effects, and $\tau(1,0)$ and $\tau(0,1)$ are measures of associative effects.

Example: experiment with noncompliance

Z: treatment assigned

Y: treatment received (we used D for this previously)

M: outcome



$\tau(1,1)$ always takers effect

$\tau(1,1)$ compliers effect

$\tau(1,1)$ defiers effect

$\tau(1,1)$ never takers effect

Example: severe disease

- $Y(M=0)$ is not well defined (Y is quality of life).
- Only $\tau(1,1)$ is well defined.
  - Survivor average causal effect (Rubin 2006)

Example: Unemployment

- Y: wage
- Z: job training
- M: employment status

Note: Only $\tau(1,1)$ is well-defined.

Example: Clinical Trial

- A good surrogate meets two criteria:
  1. If treatment doesn't affect the surrogate, then it doesn't affect the outcome (**causal necessity**).
  2. If treatment affects the surrogate, then it affects the outcome too (**causal sufficiency**).

- **Causal necessity** requires:

$$\tau(1,1) \text{ and } \tau(0,0) = 0.$$

- **Causal sufficiency** requires:

$$\tau(1,0) \text{ and } \tau(0,1) \neq 0.$$

2.20.2 Identification of $\tau(m_1, m_0)$

To identify $\tau(m_1, m_0) = E[Y(1) - Y(0) \mid M(1) = m_1, M(0) = m_0]$, we rely on the following assumptions:

2.20.2.1 Assumptions

1. **Completely Randomized Experiment (CRE)**:

$$Z \perp \{Y(1), Y(0), M(1), M(0)\}$$

2. **Monotonicity**: $M(1) \geq M(0)$.

   Example: no defiers in compliance settings.

2.20.2.2 Latent Strata Proportions

Under these assumptions, we can identify the proportions of three latent strata:

$$\pi(1,1) = P(M(1) = 1, M(0) = 1)$$

$$\pi(0,0) = P(M(1) = 0, M(0) = 0)$$

$$\pi(1,0) = P(M(1) = 1, M(0) = 0).$$

- By monotonicity, $P(M(1) < M(0)) = 0$.

Example: Severe disease and truncation by death.

Only $\tau(1,1)$ is well-defined because $Y(M = 0)$ (e.g., quality of life for non-survivors) is not observable.

2.20.2.3 Key Derivations

To identify $E[Y(1) \mid M(1) = 1, M(0) = 1]$:

$$E[Y(1) \mid M(1) = 1, M(0) = 1] = E[Y \mid Z = 1, M = 1]$$

However, due to CRE, this cannot directly identify $E[Y(1) \mid M(1) = 1, M(0) = 1]$, as:

$$E[Y \mid Z = 1, M = 1] \neq E[Y(1) \mid M(1) = 1, M(0) = 1].$$

Instead:

$$E[Y(1) \mid M(1)=1, M(0)=1] = E[Y(1) \mid M(1)=1, M(0)=1] \cdot P(M(0)=1 \mid M(1)=1) + E[Y(1) \mid M(1)=1, M(0)=0] \cdot P(M(0)=0 \mid M(1)=1).$$

2.20.2.4 Partial Identification

As there is one equation with two unknowns, we cannot point-identify $E[Y(1) \mid M(1)=1, M(0)=0]$.

Solution: Partial identification techniques can provide bounds for the unobservable quantities.

2.21 Lecture 21: Guest Lecture about Panel Causal Model

Lingxi Zhong, Dominic Fannjiang, Fangyuan Li(revisions)

Note:

- This lecture is a guest lecture offer by professor Avi Feller

Panel Data Overview

- **Cross-sectional data:** Observe a single snapshot in time
  - Standard observational data - **assume away unmeasured confounding.**
  - Most useful when we have lots of Xs and large N
- **Panel Data:** Observe same units over time. Usually in settings with a small number of units.
  - Almost all analysis is model-based, not design-based. Therefore, in order to believe the results, one needs to be convinced about the underlying model.
  - Can we do better than assuming away unmeasured confounding?
- Assume **unmeasured confounding is "stable" over time**

Longitudinal Designs

- A) Pre/post
- B) Interrupted time-series
- C) Difference-in-difference
- D) Comparative interrupted time-series
- **WARNING**
  - Some people call every panel data setting **diff-in-diff**

- Elsewhere, common to reserve **diff-in-diff** for **2 period case only**. More general setting is Comparative Interrupted Time-Series

Panel Data By Any Other Name

- **Treated unit only:** Pre-post, Interrupted Time-Series(ITS)
- **+Comparison unit:** Difference-in-difference(DiD), Comparative Interrupted Time-Series(CITS), Two-Way Fixed Effect(TWFE)
- **Variation:** Event Study, Difference-in-Difference-in-Difference(DDD), Synthetic Control Method.

Part 1: BEFORE-AFTER(Pre-Post)

- Pre-Post: Basic Setup
  - We observe many schools: Albany HS, Berkeley HS, Cerrito HS, ...
  - **All** schools adopt a new tutoring program
  - Observe test scores before/after
- Pre/Post
  - Outcome levels at two time points
  - **Treated unit/group only**
  - Attribute all changes to intervention

$$\widehat{\text{Pre-Post}} = \bar{Y}_{T,\text{Post}} - \bar{Y}_{T,\text{Pre}}$$

Part 2: Canonical 2x2 DiD, the basics

- 2x2 DiD: The Setup
  - We observe many schools: Albany HS, Berkeley HS, Cerrito HS, ...
  - **Some** schools adopt a new tutoring program
    - Treatment = Z
  - School-average assessments before and after
    - Scores = Pre and Post
  - 2x2 Differences-in-Differences(DiD) setting
    - Two groups and two time periods
- Differences-in-Differences
  - Other units comparison

- Relative change in levels
- **Assumed counterfactual:** outcome would have changed as much as the non-treatment group if not for the intervention. This usually involves an assumption of "parallel trends", which is that if the treated unit were not treated, its different in potential outcomes would be the same as that of the control unit.

$$\widehat{\text{DiD}} = \underbrace{\left(\bar{Y}_{T,\text{ Post}} - \bar{Y}_{T,\text{ Pre}}\right)}_{\text{Gain }_T} - \underbrace{\left(\bar{Y}_{C,\text{ Post}} - \bar{Y}_{C,\text{ Pre}}\right)}_{\text{Gain }_c}$$

$$\widehat{\text{DiD}} = (\bar{Y}_{T,\text{Post}} - \bar{Y}_{C,\text{Post}}) - (\bar{Y}_{T,\text{Pre}} - \bar{Y}_{C,\text{Pre}})$$

Part 3: Interrupted time-series

One timeseries, but you compare multiple data points of before and after the time of treatment.
Emphasis on the fact that multiple points in time can be used to define the estimand of interest. Note that this is different from regression discontinuity precisely because the estimand of interest is not merely what happens around the threshold (which, in this case, is the time point where the treatment occured).

Part 4: Comparative interrupted time-series

The same as part 3, but with multiple time series, where some are considered control and other are considered treated. ex) You have a time series of gun-related deaths in each state where some states implemented some law related to gun control, and other have not. Often times, the goal is to use the "control" timeseries to be an example of what would have happened to the treated state had it not been treated.

2.22 Lecture 23: Post-Treatment Variables

Max Medina, Shana Kim, Jake Derr,

Fangyuan Li(revisions)

Recap

Recall that **Principal Stratification** is conditioning on potential outcomes of post treatment variables. For example: with Z a treatment for a disease, Y a health-related outcome and M a post-treatment variable, like the quality of life, we define the ATE in each strata as $\tau(m_1, m_0) = E(Y(1) - Y(0)|M(1) = m_1, M(0) = m_0)$. Is $\tau(1, 1)$ well-defined? Partially:

- Under CRE + monotonicity: $E(Y(0)|M(1) = 1, M(0) = 1) = E(Y|Z = 0, M = 1)$, all observables

- However $\mathbb{E}[Y(1) \mid M(1) = 1, M(0) = 1]$ cannot be identified. Specifically:

$$\mathbb{E}[Y \mid Z = 1, M = 1] = \Psi \cdot \mathbb{P}(M(1) = 1 \mid M(0) = 1) + E(Y(1) \mid M(1) = 1, M(0) = 0]P(M(1) = 1 \mid M(0) = 0),$$

  where the last term is unobservable.

Mediation Analysis

Why Mediation Analysis?

- Reveals mechanisms
- Easier to intervene on M than Z (e.g., Z: smoking, M: blood lipid levels, Y: cardiovascular disease).
- Y may be expensive or time-consuming to observe; M serves as a surrogate (proxy) outcome.

Notation:

- $M(Z)$: Potential outcome of M given Z.
- $Y(Z)$: Potential outcome of Y given Z.
- $Y(Z, M)$: Potential outcome of Y given Z and M.
- $Y(Z, M(Z'))$: Counterfactual outcome of Y under Z and M had Z been Z'.

Example (binary Z, M): $Y(1, M_0)$ and $Y(0, M_1)$ are unobservable *cross-world counterfactuals*.

Total Effect Decomposition

Why care about cross-world counterfactuals?

$$\text{Total Effect (TE)} = \mathbb{E}[Y(1) - Y(0)].$$

$$\text{Natural Direct Effect (NDE)} = \mathbb{E}[Y(1, M_0) - Y(0, M_0)].$$

$$\text{Natural Indirect Effect (NIE)} = \mathbb{E}[Y(0, M_1) - Y(0, M_0)].$$

Identification (Pearl, 2001)

Mediation Formula: let X denote any observed covariate.

1. $Z \perp Y(z, m) \mid X$
2. $M \perp Y(z, m) \mid X, Z$
3. $Z \perp M(z) \mid X$
4. $Y(Z, M) \perp M(Z') \mid X$ for all z, z' (Cross-world assumption, untestable)

Note that 1 and 2 together are called *sequential ignorability* and is equivalent to joint randomization $(Z, M) \perp Y(z, m)|X$.

Under 1–4:

$$\text{NDE}(x) = \mathbb{E}[Y(1, m_0) - Y(0, m_0) \mid X = x]$$

$$= \sum_m \Big\{ (\mathbb{E}[Y \mid Z = 1, M = m, X = x] - \mathbb{E}[Y \mid Z = 0, M = m, X = x] \Big\} \mathbb{P}(M = m \mid Z = 0, X = x),$$

$$\text{NIE}(x) = \sum_m \mathbb{E}[Y \mid Z = 1, M = m, X = x] \Big\{ \mathbb{P}(M = m \mid Z = 1, X = x) - \mathbb{P}(M = m \mid Z = 0, X = x) \Big\}.$$

Finally:

$$\text{NDE} = \int \text{NDE}(x) p(x) dx, \quad \text{NIE} = \int \text{NIE}(x) p(x) dx.$$

Baron-Kenny Method

Assume linear relationships and a binary M. We adopt a parametric model and use the regressions $M \sim Z, M$ and $Y \sim Z, M, X$ to estimate the direct and indirect effects. Specifically:

$$\mathbb{E}[M \mid Z, X] = \beta_0 + \beta_1 Z + \beta_2^T X, \tag{2.18}$$

and:

$$\mathbb{E}[Y \mid M, Z, X] = \theta_0 + \theta_1 Z + \theta_2 M + \theta_3^T X. \tag{2.19}$$

The coefficients in this regressions can be identified in the edges of the following causal diagram.

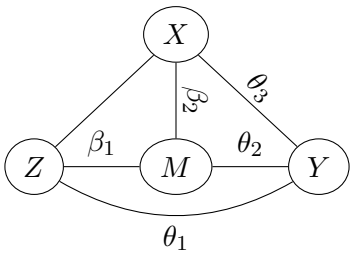


Figure 2.1: Causal relationships between Z, X, M, Y

Effects:

- NDE $= \theta_1$
- NIE $= \theta_2 \beta_1$

These can be estimated using normal OLS regression, with mean and variance. **Pros:** Running regressions is straightforward. **Cons:** Relies on strong parametric assumptions, which may be unstable.

Controlled Effects

Controlled Direct Effect (CDE):

$$\text{CDE}(m) = \mathbb{E}[Y(1, m) - Y(0, m)].$$

Controlled Indirect Effect (CIE):

$$\text{CIE}(m, m') = \mathbb{E}[Y(z, m) - Y(z, m')].$$

Identification of CDE

Assumptions

1. $Z \perp\!\!\!\perp Y(Z, M) \mid X$
2. $M \perp\!\!\!\perp Y(Z, M) \mid Z, X$

For Binary Z and M

- Define $\mu_{zm} = \mathbb{E}[Y(Z, M)]$, so $\text{CDE}(m) = \mu_{1m} - \mu_{0m}$
- $\mu_{zm}(X) = \mathbb{E}[Y \mid Z = z, M = m, X = x]$
- $e_{zm}(X) = P(Z = z, M = m \mid X = x) = P(Z = z \mid X = x) \cdot P(M = m \mid Z = z, X = x)$

Under $(Z, M) \perp\!\!\!\perp Y(Z, M) \mid X$

$$\mu_{zm} = \mathbb{E}[\mu_{zm}(X)]$$
$$\mu_{zm} = \mathbb{E}\left[\frac{\mathbb{I}\{Z = z, M = m\}Y}{e_{zm}(X)}\right]$$

Controlled Direct Effect (CDE)

$$\text{CDE}(m) = \mu_{1m} - \mu_{0m}$$

Doubly Robust Formula

Given the following working models

- $e_{zm}(X, \alpha)$ for $e_{zm}(X)$,
- $\mu_{zm}(X, \beta)$ for $\mu_{zm}(X)$,

we can employ a DR approach as follows:

$$\mu_{zm} = \mathbb{E}\left[\mu_{zm}(X,\beta)\right] + \mathbb{E}\left[\frac{\mathbb{I}\{Z=z, M=m\}(Y-\mu_{zm}(X,\beta))}{e_{zm}(X,\alpha)}\right]$$
$$= \mu_{zm} \text{ if either } e_{zm}(X,\alpha) = e_{zm}(X) \text{ or } \mu_{zm}(X,\beta) = \mu_{zm}(X).$$

Recap on Post-Treatment Variables

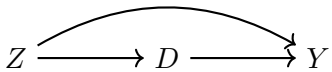
Framework	Direct Effect	Indirect Effect
Principal Stratification	$Y(1,1), Y(0,0)$	-
Mediation Analysis	NDE	NIE
Controlled Effects	CDE(m)	CIE$(m, m') \mid Z$

2.23 Lecture 24: Time-Varying Treatment and Confounders

Will Rathgeb, Qinhan Zhou, Ziyuan Jin

Recap

Mediation Analysis via Natural and Controlled Effects

- M: Mediator 

- Direct Effect: $\mathbb{E}[(Y(1, M=m) - Y(0, M=m)]$

  Natural effects: $m = M(1)$ or $m = M(0)$

  Controlled effects: $m \in \mathcal{M}$ (set of all possible values of M)
- Indirect Effect: $\mathbb{E}[(Y(Z, m_1) - Y(Z, m_0)]$

  Natural effects: $m_1 = M(1),\ m_0 = M(0)$

  Controlled effects: $m_0, m_1 \in \mathcal{M}$

Time-Varying Treatment + Confounding

ex: HIV patients take antiretroviral medicines on + off over time ex: candidates adjust their campaign strategy based on polls + opponent's behavior

Suppose we have two time points.

Temporal order:

$$X_0 \to Z_1 \overset{a}{\to} X_1 \to Z_2 \to Y$$

- X_0: baseline covariates
- Z_1: treatment at time point 1 (binary): with 1-4 hours
- X_1: time-varying covariates observed between treatments
- Z_2: treatment at time point 2 (binary): with 4-8 hours
- Y: Outcome

4 potential outcomes: Y(0,0), Y(0,1), Y(1,0), Y(1,1) Observed outcome:

Estimates:

E[Y(Z_1, Z_2)]

- E[Y(1,0) - Y(0,0)]
- E[Y(0,1) - Y(0,0)]
- E[Y(1,1) - Y(0,0)]

Our choice of estimand depends on policy/ science question

Suppose our estimate is E[Y(Z_1, Z_2)]

Identifications: Assumption: Sequential ignorability: treatments are sequentially randomized given observed history 1. Z_1 is randomized given $C(X_0)$.

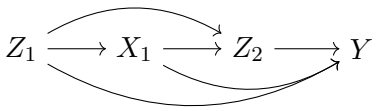
$$Z_1 \perp\!\!\!\perp Y(Z_1, Z_0) \mid Z_0,$$

for $z_1, z_2 \in \{0, 1\}$. 2. Z_2 is randomized given $C(Z_1, X_1, X_0)$.

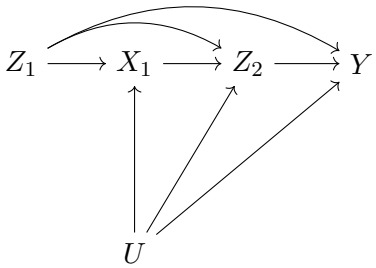
$$Z_2 \perp\!\!\!\perp Y(Z_1, Z_0) \mid Z_1, X_1, X_0$$

for $z_1, z_2 \in \{0, 1\}$

Satisfied under a DAG



Also satisfied under OAG



Recall identification under a single time-point setting: $\mathbb{E}[Y(Z)] = \mathbb{E}[\mathbb{E}[Y|Z = z, X]]$

- discrete X: $\sum_X \mathbb{E}[Y|Z=z, X=x]\mathbb{P}(X=x)$
- continuous X: $\int_X \mathbb{E}[Y|Z=z, X=x]p(x)dx$
- For discrete X:

$$\sum_x \mathbb{E}[Y \mid Z=z, X=x]P(X=x)$$

- For continuous X:

$$\int_x \mathbb{E}[Y \mid Z=z, X=x]p(X)\,dx$$

Theorem

Under sequential ignorability: $\mathbf{E}[Y(Z_1, Z_2)] = \mathbf{E}\left[\mathbf{E}\left[\mathbf{E}[Y \mid Z_2 = Z_2, Z_1 = Z_1, X_1, X_0] \mid Z_1 = Z_1, X_0\right]\right]$

"g-formula" - Jamie Robins:

- discrete X_0, X_1

  $= \sum_{X_0}\sum_{X_1} \mathbb{E}[Y|z_2, z_1, X_1, X_0]\mathbb{P}(X_1|z_1, X_0)\mathbb{P}(X_0)$
- continuous X_0, X_1

  $= \int\int \mathbb{E}[Y|Z_2, Z_1, X_1, X_0]p(X_1|Z_1, X_0)p(X_0)dx_1dx_0$

Proof

-

$$\begin{aligned}
\mathbb{E}[Y(Z_1,Z_2)] &= \mathbb{E}[\mathbb{E}[Y(Z_1,Z_2) \mid X_0]] \\
&= \mathbb{E}[\mathbb{E}[Y(Z_1,Z_2) \mid Z_1, X_0]] \quad \text{by sequential ignorability} \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y(Z_1,Z_2) \mid Z_1, X_1, X_0] \mid Z_1, X_0]] \quad \text{by Tower Property} \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y(Z_1,Z_2) \mid Z_2, Z_1, X_1, X_0] \mid Z_1, X_0]] \quad \text{by sequential ignorability} \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y \mid Z_2, Z_1, X_1, X_0] \mid Z_1, X_0]]
\end{aligned}$$

Estimators

1. Outcome Modeling

- Recall outcome modeling in a single time-point setting:

$$\mathbb{E}[Y(Z)] = \mathbb{E}[\mathbb{E}[Y \mid Z=z, X]]$$

- Steps:
  1. Regress $Y \sim X \mid Z=z$ to obtain $\hat{Y}_i(z)$ for all units.
  2. Compute:

$$\widehat{\mathbb{E}}[Y(Z)] = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i(z)$$

- Now, in a time-varying setting:
  1. Regress $Y \sim X_1, X_0 \mid Z_1 = z_1, Z_2 = z_2$ to get $\hat{Y}_{i2}(z_1, z_2)$.
  2. Regress $\hat{Y}_{i2}(z_1, z_2) \sim X_0 \mid Z_1 = z_1$ to get $\hat{Y}_i(z_1, z_2)$.
  3. Compute:
$$\widehat{\mathbb{E}}[Y(Z_1, Z_2)] = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i(z_1, z_2)$$

2. IPW Estimator

- Recall for a single time-point, the identification:
$$\mathbb{E}[Y(Z)] = \mathbb{E}\left[\frac{\mathbb{I}\{Z = z\}Y}{P(Z = z \mid X)}\right]$$
- For a two-time-point setting:
  - Define:
$$e(z_1, X_0) = P(Z_1 = z_1 \mid X_0)$$
$$e(z_2, z_1, X_1, X_0) = P(Z_2 = z_2 \mid Z_1 = z_1, X_1, X_0)$$
- Theorem (under sequential ignorability):
$$\mathbb{E}[Y(Z_1, Z_2)] = \mathbb{E}\left[\frac{\mathbb{I}\{Z_1 = z_1\}\mathbb{I}\{Z_2 = z_2\}Y}{e(z_1, X_0)e(z_2, z_1, X_1, X_0)}\right]$$
- Highlights the overlap assumption:
$$0 < e(z_1, X_0) < 1, \quad 0 < e(z_2, z_1, X_1, X_0) < 1, \quad \forall z_1, z_2$$
- IPW Estimator:
  1. Regress $Z_1 \sim X_0$ to estimate $\hat{e}(z_1, X_0)$.
  2. Regress $Z_2 \sim Z_1, X_1, X_0$ to estimate $\hat{e}(z_2, z_1, X_1, X_0)$.
  3. Compute:
$$\widehat{\mathbb{E}}[Y(Z_1, Z_2)] = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}\{Z_{1i} = z_1\}\mathbb{I}\{Z_{2i} = z_2\}Y_i}{\hat{e}(z_1, X_{0i})\hat{e}(z_2, z_1, X_{1i}, X_{0i})}$$

Limitations and Solutions

Limitations

For K time points, we have 2^K treatment combinations. This approach may not work well for settings with more than a few time points.

Solution: Marginal Structural Model (MSM)

References:

- Robins 2000
- Hernán 2000

Detour:

Point treatment, continuous, or multivariate/multi-dimensional treatments.

Examples:

1. **Dose:** Drug dose, e.g., mg of aspirin over time.
2. **Duration:** e.g., effect of education duration on wages.
   - Multidimensional: also degree level (undergrad, master's, etc.).
3. **Frequency:** e.g., estimate effects of hospital nurse staffing on readmission outcomes.
   - Nurse staffing = nurse hours/day.

2.23.0.1 Identification:

$\mathbb{E}[Y(Z=z)] = \mathbb{E}[\mathbb{E}[Y|X, Z=z]]$

2.23.0.2 MSM Estimation:

- Assume a model:

$$\mathbb{E}[Y(Z)] = g(z, \beta)$$

  where $g(z, \beta)$ is a known function up to a finite-dimensional parameter β.

- Example (Semiparametric model assumption):

$$\mathbb{E}[Y(Z)] = \beta_0 + \beta_1 z + \beta_2 z^2$$

Semiparametric Approach and IPW Estimator under MSM

Definition

$$\beta = \arg\min_{\beta} \int \left(\mathbb{E}[Y(Z)] - g(Z;\beta)\right)^2 P(Z) dZ$$

Back to the Time-Varying Setting (2 Time Points)

Semiparametric Approach Assume:

$$\mathbb{E}[Y(Z_1, Z_2)] = g(Z_1, Z_2; \beta)$$

Best Approximation Approach

$$\beta = \arg\min_{\beta} \sum_{z_1} \sum_{z_2} \left(\mathbb{E}[Y(Z_1, Z_2)] - g(Z_1, Z_2, X_0; \beta)\right)^2$$

IPW under MSM with Sequential Ignorability

$$\beta = \arg\min_{\beta} \sum_{z_1} \sum_{z_2} \mathbb{E}\left[\frac{\mathbb{I}\{Z_1 = z_1\}\mathbb{I}\{Z_2 = z_2\}}{e(z_1, X_0)e(z_2, z_1, X_1, X_0)} \left(Y - g(Z_1, Z_2, X_0; \beta)\right)^2\right]$$

IPW Estimator under MSM

1. Estimate two propensity scores (same as IPW steps described above):
   - $e(z_1, X_0) = P(Z_1 = z_1 \mid X_0)$
   - $e(z_2, z_1, X_1, X_0) = P(Z_2 = z_2 \mid Z_1 = z_1, X_1, X_0)$
2. $\hat{\beta}$ is the coefficient of a Weighted Least Squares (WLS) fit of:

   $$Y_i \sim (1, Z_{1i}, Z_{2i}, X_{0i})$$

   with weights:

   $$w_i = \frac{1}{\hat{e}(Z_{1i}, X_{0i})\hat{e}(Z_{2i}, Z_{1i}, X_{1i}, X_{0i})}.$$

Chapter 3 Predicting Long-term Outcomes

3.1 Surrogate Index: [Athey et al.(2019)]

Define two samples N_E (Experimental) and N_O (Observational), $N = N_E + N_O$, where we use $P_i \in \{O, E\}$ as a binary indicator of whether an individual i is in N_E or N_O.

For each individual, we have

X_i	vector $\in \mathbb{X}$	pre-treatment covariates for each unit
W_i	$\in \mathbb{W} := \{0, 1\}$	binary treatment for each unit (unobserved if $P_i = O$)
Y_i	scalar $\in \mathbb{Y}$	primary outcome (unobserved if $P_i = E$)
S_i	vector $\in \mathbb{S}$	surrogates (intermediate outcomes)

The primary outcome Y_i is unobservable for individuals in N_E (i.e., in experimental sample, $P_i = E$).

Following the potential outcome framework or Rubin Causal Model, each individual has two pairs of potential outcomes:

$$Y_i \equiv Y_i(W_i) = \begin{cases} Y_i(0), & W_i = 0 \\ Y_i(1), & W_i = 1 \end{cases}, \; S_i \equiv S_i(W_i) = \begin{cases} S_i(0), & W_i = 0 \\ S_i(1), & W_i = 1 \end{cases}$$

Overall, the units are characterized by $(Y_i(0), Y_i(1), S_i(0), S_i(1), X_i, W_i, P_i)$.

- In the experimental sample, $P_i = E$, we observe

$$(X_i, W_i, S_i) \in (\mathbb{X}, \mathbb{W}, \mathbb{S})$$

- In the observational sample, $P_i = O$, we observe

$$(X_i, S_i, Y_i) \in (\mathbb{X}, \mathbb{S}, \mathbb{Y})$$

For simplicity, we analyze the data as we observe

$$(P_i, X_i, S_i, \mathbf{1}_{P_i=E} W, \mathbf{1}_{P_i=O} Y_i)$$

The estimand we are interested in is the Average Treatment Effect (ATE) on the primary outcome in the population from which the experimental sample is drawn:

$$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid P_i = E].$$

3.1.1 Critical Assumptions

Unconfoundedness

For the individuals in the experimental group, the propensity score is the conditional probability of receiving the treatment: $\rho(x) \equiv \Pr(W_i = 1 \mid X_i = x, P_i = E)$.

Assumption 3.1 (Unconfounded Treatment Assignment / Strong Ignorability)

1. For individuals in the experimental group, treatment assignment is unconfounded:

$$W_i \perp (Y_i(0), Y_i(1), S_i(0), S_i(1)) \mid X_i, P_i = E$$

2. We have overlap in the distribution of pre-treatment variables between the treatment and control groups:

$$\rho(x) \in (0, 1), \ \forall x \in \mathbb{X}$$

Surrogacy

Surrogate is a post-treatment variable where conditioning on it makes the outcome and the treatment independent. We can define two scalar functions of surrogates:

Definition 3.1 (Surrogate Index and Surrogate Score)

1. The *surrogate index* is the conditional expectation of the primary outcome given the surrogate outcomes and the pre-treatment variables, conditional on the sample:

$$\mu(s, x, p) \equiv \mathbb{E}[Y_i \mid S_i = s, X_i = x, P_i = p]$$

2. The *surrogate score* is the conditional probability of having received the treatment given the value for the surrogate outcomes and the covariates in the experimental sample:

$$\rho(s, x) \equiv \Pr(W_i = 1 \mid S_i = s, X_i = x, P_i = E)$$

Assumption 3.2 (Surrogacy)

The treatment is independent of the primary outcomes conditional on the surrogates on the experimental group:

$$W_i \perp Y_i \mid S_i, X_i, P_i = E$$

and $\rho(s, x) \in (0, 1), \forall s \in \mathbb{S}, x \in \mathbb{X}$, and $\Pr(P_i = E) \in (0, 1)$.

Proposition 3.1 (Surrogate Score)

Suppose Surrogacy (Assumption 3.2) holds, then

$$W_i \perp Y_i \mid \rho(S_i, X_i), P_i = E.$$

Note *[(The More The Better)] Having multiple short-term variables can make a surrogacy approach more plausible, the same way multiple pre-treatment variables can make the unconfoundedness assumption more plausible.*

Comparability

Surrogacy and Unconfoundedness are not sufficient for consistent estimation of τ because they do not place restrictions on how the relationship between Y_i and S_i in the observational sample compares to that in the experimental sample.

Definition 3.2 (Sampling Score)

The *sampling score* is $\varphi(s,x) \equiv \Pr(P_i = E \mid S_i = s, X_i = x)$.

Assumption 3.3 (Comparability of Samples)

The conditional distribution of Y_i given (S_i, X_i) in the observational sample is the same as it in the experimental sample:

$$P_i \perp Y_i \mid S_i, X_i$$

and $\varphi(s,x) \in (0,1), \forall s \in \mathbb{S}, x \in \mathbb{X}$.

Proposition 3.2 (Surrogate Index)

1. Suppose Surrogacy (Assumption 3.2) holds, then
$$\mu(s,w,x,E) = \mu(s,x,E), \forall s \in \mathbb{S}, x \in \mathbb{X}, w \in \mathbb{W}$$
2. Suppose Comparability (Assumption 3.3) holds, then
$$\mu(s,x,E) = \mu(s,x,O), \forall s \in \mathbb{S}, x \in \mathbb{X}$$
3. Suppose Surrogacy (Assumption 3.2) and Comparability (Assumption 3.3) hold, then
$$\mu(s,w,x,E) = \mu(s,x,O), \forall s \in \mathbb{S}, x \in \mathbb{X}, w \in \mathbb{W}$$

3.2 Using Survival Models: [Chandar et al.(2022)]

We want to estimate the ATE in experimental dataset (D_E),

$$\delta_E = \mathbb{E}_{D_E}[Y_i(1) - Y_i(0)] = \frac{1}{N_{W^1}} \sum_{i \in W^1} Y_i(1) - \frac{1}{N_{W^0}} \sum_{i \in W^0} Y_i(0)$$

where $W^c = \{i \in D_E \mid W_i = c\}$ and $N_{W^c} = |W^c|$.

Bibliography

[Athey et al.(2019)] Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research.

[Chandar et al.(2022)] Chandar, P., St. Thomas, B., Maystre, L., Pappu, V., Sanchis-Ojeda, R., Wu, T., Carterette, B., Lalmas, M., and Jebara, T. (2022). Using survival models to estimate user engagement in online experiments. In *Proceedings of the ACM Web Conference 2022*, pages 3186–3195.

[Fillmore et al.(2007)] Fillmore, K. M., Stockwell, T., Chikritzhs, T., Bostrom, A., and Kerr, W. (2007). Moderate alcohol use and reduced mortality risk: systematic error in prospective studies and new hypotheses. *Annals of epidemiology*, 17(5):S16–S23.

[Neyman(1923)] Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science*, (9):S465—S472.

[Rubin(1980)] Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.