

Causal Inference: Some Theory and Practice

Tengyuan Liang¹

DLA Lecture 3: What If

A crash course touching the theoretical and practical side of causal inference from a machine learner's view. Readings: Hardt and Recht² Chapter 9 and 10. Peters, Janzing, and Schölkopf³, Chapter 1 and 6.

Contents

1	<i>In Theory</i>	1
1.1	<i>Observations vs. Actions</i>	1
1.2	<i>Simpson's Paradox</i>	2
1.3	<i>Reichenbach's Common Cause Principle</i>	3
1.4	<i>Causal Models and do-Calculus</i>	4
1.5	<i>Some Graph Structures</i>	7
1.6	<i>Causal Effects and Adjustment Formula</i>	8
1.7	<i>Randomization and the Backdoor Criterion</i>	9
1.8	<i>Potential Outcome Framework vs. Structural Causal Model</i>	10
2	<i>In Practice</i>	11
2.1	<i>Basic Regression Adjustments</i>	12
2.2	<i>Propensity Score Adjustments</i>	12
2.3	<i>Conditional Average Treatment Effects</i>	13
2.4	<i>Doubly Robust Adjustments</i>	13
2.5	<i>Instrumental Variables</i>	14
2.6	<i>Regression Discontinuity</i>	14

1 *In Theory*

1.1 *Observations vs. Actions*

- **Observations** are passive, and reflect the state of the world projected to a set of features we choose to highlight. Data that we collect from passive observation only show a snapshot of the world.
- **Actions** are active, and demonstrate what if we intervene the world in a certain way. We wish to understand how that action will affect the features we choose to observe. Rather than asking for the frequency of an event in the manifested world, this asks for the effect of a hypothetical action.

¹ The University of Chicago
Booth School of Business

² Moritz Hardt and Benjamin Recht. *Patterns, Predictions, and Actions: A Story about Machine Learning*. Princeton University Press, 2022

³ Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017

Figure 1: See Peters, Janzing, and Schölkopf.

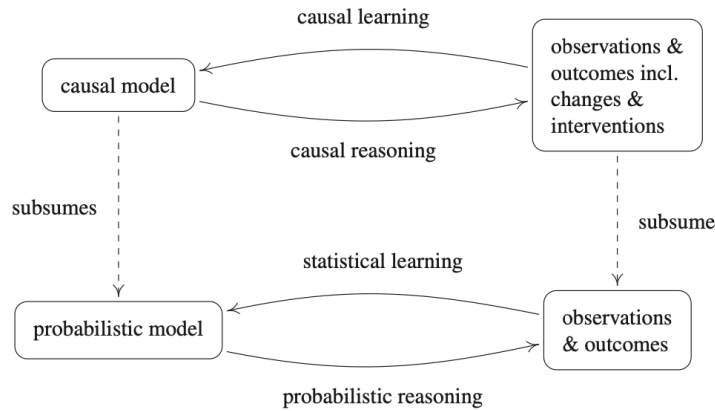


Figure 1.1: Terminology used by the present book for various **probabilistic inference** problems (bottom) and **causal inference** problems (top); see Section 1.3. Note that we use the term “inference” to include both learning and reasoning.

Example. *Correlation question: Do 16-year-old drivers have a higher incidence rate of traffic accidents than 18-year-old drivers?*

Counterfactual question: Would traffic fatalities decrease had one raised the legal driving age by two years?

Causal reasoning is a conceptual and technical framework that addresses the “what if” questions.

1.2 Simpson’s Paradox

- Berkeley’s admission data in 1973 is a venerable example used to demonstrate why observations are limited.
- Historical data show that 12763 applicants were considered for admission to one of 101 departments and inter-departmental majors.
- Of the 4321 women who applied, roughly 35 percent were admitted, while 44 percent of the 8442 men who applied were admitted.
- Standard statistical significance tests suggest that the observed difference would be highly unlikely to be the outcome of sample fluctuation if there were no difference in underlying acceptance rates.
- However, if we condition on the department level, many of the admission rate patterns between men and women are reversed.

Let Y be acceptance, A be female gender, and Z department choice, then Simpson’s paradox can be illustrated as follows

- $\mathbb{P}[Y|A] > \mathbb{P}[Y|\neg A]$
- $\mathbb{P}[Y|A, Z = z] < \mathbb{P}[Y|\neg A, Z = z], \forall z \in \mathcal{Z}$.

Limitation of Observations. Mathematically, Simpson's Paradox has no surprise. However, it poses challenges to observational questions: when can we trust an answer based on empirical observations given that the relationship can be "all reversed" in each subpopulation?

The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

At this point, we have two choices. Causal inference is helpful in either case.

- Experiments: One is to design an experiment (or, better, a randomized study) and collect more data.

On the one hand, causal inference can be used as a guide in the design of new studies. It can help us choose which variables to include, which to exclude, and which to hold constant.

- Assumptions: The other is to resort to our beliefs and plausible assumptions and then argue over which scenario is more likely.

On the other hand, causal models can serve as a mechanism to incorporate scientific domain knowledge and exchange plausible assumptions for plausible conclusions.

1.3 Reichenbach's Common Cause Principle

- "Correlation does not imply causation." Well-known topos.
- To learn causal structures from observational distributions, we need to understand how causal models and observational models relate to each other.
- One may not infer a concrete causal structure, but we may at least infer the existence of causal links from statistical dependences.

Principle (Reichenbach's Common Cause Principle (1956)). *If two random variables X and Y are statistically dependent $X \not\perp Y$, then there exists a third variable Z that causally influences both. Furthermore, this variable screens X and Y from each other in the sense that given Z , they become independent, $X \perp Y|Z$.*

1.4 Causal Models and do-Calculus

- A structural causal model is a generative model where we know exactly the sequence of assignments for generating joint distribution starting from independent noise variables.
- It is an execution sequence of assignments in which we incrementally build a set of jointly distributed random variables.

Definition 1 (Structural Causal Model). *A structural causal model M is given by a set of variables X_1, \dots, X_d and corresponding assignments of the form*

$$X_i := f_i(P_i, U_i), \quad i = 1, \dots, d. \quad (1.1)$$

Here, $P_i \subseteq \{X_1, \dots, X_d\}$ is a subset of variables that we call the parents of X_i . The random variables U_1, \dots, U_d are called exogenous (or noise) variables, which we require to be jointly independent.

The directed graph corresponding to the model has one node for each variable X_i , which has incoming edges from all parents P_i . We call such a graph the causal graph corresponding to the structural causal model.

When M denotes a structural causal model, we write the probability of an event E under the entailed joint distribution as $\mathbb{P}_M(E)$.

Structural causal models are a collection of formal **assumptions** about how certain variables interact. Each assignment specifies a **response function**. We can think of nodes as receiving messages from their parents and acting according to these messages as well as the influence of an exogenous noise variable.

Throughout, we only consider **acyclic assignments**. Many real-world systems are naturally described as stateful dynamical systems with feedback loops. For example, cycles can often be broken up by introducing time-dependent variables, such as investments at time 0 grow the economy at time 1, which in turn grows investments at time 2, continuing so forth until some chosen time horizon t .

Definition 2 (Interventions and the do-Operator). *Given a structural causal model M , we can operate on any assignment of the form*

$$X := f(P, U) \quad (1.2)$$

by replacing it with another assignment. The most common substitution is to assign X a constant value x

$$X := x \quad (1.3)$$

We shall denote the resulting model $M' = M[X := x]$ to indicate the surgery we performed on the original model M . Under this assignment, we

hold X constant by removing the influence of its parent nodes and, thereby, any other variables in the model.

The assignment operator is also called **do-operator** to emphasize the performance of an intervention, denoted as

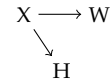
$$\mathbb{P}[E|\text{do}(X := x)], \text{ or, } \mathbb{P}_{M[X:=x]}[E] \tag{1.4}$$

Again, the do-operation (action) is fundamentally different from the conditioning operator (observation).

Example (Exercise, Weight, and Heart Disease). Consider the following structural causal model among variables

- X : regular exercise or not
- W : excessive weight or not
- H : heart disease or not

The first model M is



1. Sample $U_1 \sim \text{Bern}(1/2), U_2 \sim \text{Bern}(1/3), U_3 \sim \text{Bern}(1/3)$

2. $X := U_1$

3. $W := \begin{cases} 0 & \text{if } X = 1 \\ U_2 & \text{o.w.} \end{cases}$

4. $H := \begin{cases} 0 & \text{if } X = 1 \\ U_3 & \text{o.w.} \end{cases}$

This model induces a joint distribution

X	W	H	$\mathbb{P}[X = x, W = w, H = h]$
0	0	0	$1/2 * 2/3 * 2/3 = 2/9$
0	1	0	$1/2 * 1/3 * 2/3 = 1/9$
0	0	1	$1/2 * 2/3 * 1/3 = 1/9$
0	1	1	$1/2 * 1/3 * 1/3 = 1/18$
1	0	0	$1/2$

Consider the intervention (1) $\text{do}(W := 1), M[W := 1]$ and (2) $\text{do}(W := 0), M[W := 0]$. The two resulting causal graphs are much simpler

			$\text{do}(W := 1)$
X	W	H	$\mathbb{P}[X = x, W = w, H = h]$
0	1	0	$1/2 * 2/3 = 1/3$
0	1	1	$1/2 * 1/3 = 1/6$
1	1	0	$1/2$

			do(W := 0)
X	W	H	$\mathbb{P}[X = x, W = w, H = h]$
0	0	0	$1/2 * 2/3 = 1/3$
0	0	1	$1/2 * 1/3 = 1/6$
1	0	0	$1/2$

Intervention differs from conditioning.

- Conditional probability $\mathbb{P}[H = 1|W = 1] = \frac{1/18}{1/18+1/9} = 1/3$ does not equal to substitution probability $\mathbb{P}[H = 1|do(W := 1)] = 1/6$.
- Observing overweight infers higher chance of heart disease as $\mathbb{P}[H = 1|W = 1] = 1/3 > \mathbb{P}[H = 1|W = 0] = 2/15$. However, it does not mean lower body weight could avoid heart disease, because $\mathbb{P}[H = 1|do(W := 1)] = 1/6 = \mathbb{P}[H = 1|do(W := 0)]$
- Based on observational data, one may draw the spurious relationship that lower body weight could result in lower chance of heart disease

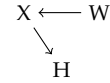
$$\mathbb{E}[H|W = 1] - \mathbb{E}[H|W = 0] > 0 \tag{1.5}$$

whereas the true causal effect should be

$$\mathbb{E}[H|do(W := 1)] - \mathbb{E}[H|do(W := 0)] = 0 \tag{1.6}$$

- This example clearly shows that substitution is fundamentally different from conditioning. However, we can incorporate the causal graph to relate the substitution probability to conditional probability.

Consider now a slight variant of the model M' , which is



1. Sample $U_1 \sim \text{Bern}(1/2), U_2 \sim \text{Bern}(1/3), U_3 \sim \text{Bern}(1/3)$
2. $W := U_2$
3. $X := \begin{cases} 0 & \text{if } W = 0 \\ U_1 & \text{o.w.} \end{cases}$
4. $H := \begin{cases} 0 & \text{if } X = 1 \\ U_3 & \text{o.w.} \end{cases}$

This model induces a joint distribution

X	W	H	$\mathbb{P}[X = x, W = w, H = h]$
0	0	1	$1 * 2/3 * 1/3 = 2/9$
0	0	0	$1 * 2/3 * 2/3 = 4/9$
0	1	1	$1/2 * 1/3 * 1/3 = 1/18$
0	1	0	$1/2 * 1/3 * 2/3 = 1/9$
1	1	0	$1/2 * 1/3 * 1 = 1/6$

In such a case, the do-calculus reads (1) $\text{do}(W := 1)$, $M[W := 1]$ and (2) $\text{do}(W := 0)$, $M[W := 0]$

			$\text{do}(W := 1)$
X	W	H	$\mathbb{P}[X = x, W = w, H = h]$
0	1	0	$1/2 * 2/3 = 1/3$
0	1	1	$1/2 * 1/3 = 1/6$
1	1	0	$1/2$

			$\text{do}(W := 0)$
X	W	H	$\mathbb{P}[X = x, W = w, H = h]$
0	0	0	$2/3$
0	0	1	$1/3$

- Conditional probability $\mathbb{P}[H = 1|W = 1] = \frac{1/18}{1/18+1/9+1/6} = 1/6$,
 $\mathbb{P}[H = 1|W = 0] = \frac{2/9}{2/9+4/9} = 1/3$.
- Substitution probability $\mathbb{P}[H = 1|\text{do}(W := 1)] = 1/6$, $\mathbb{P}[H = 1|\text{do}(W := 0)] = 1/3$
- In such case, the conditional probability equals substitution counterparts, and thus

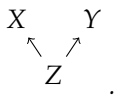
$$\mathbb{E}[H|W = 1] - \mathbb{E}[H|W = 0] = -1/6 \tag{1.7}$$

happens to estimate the true causal effect

$$\mathbb{E}[H|\text{do}(W := 1)] - \mathbb{E}[H|\text{do}(W := 0)] = -1/6 \tag{1.8}$$

1.5 Some Graph Structures

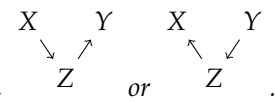
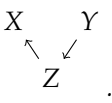
Definition 3 (Forks). A fork is a node Z in a graph that has outgoing edges to two other variables X and Y . Put differently, the node Z is a com-

mon cause of X and Y . In picture,  .

In forks, Z introduces a confounding effect: ignoring Z will introduce a (spurious) correlation between X and Y . The confounding leads to a disagreement between the calculus of conditional probabilities (observation) and do-interventions (action).

In the Example: Exercise, Weight, and Heart Disease M , exercise X influences both the weight W and heart disease H .

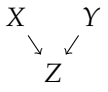
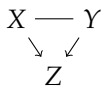
Definition 4 (Mediators). A mediator is a node Z which lies on a directed

path from X to Y . In picture,  or  .

In this case, the path contributes to the total effect of X on Y . It's a causal path and thus one of the ways in which X causally influences Y . That's why Z is not a confounder. We call Z a mediator instead.

In the Berkeley admission example, the department choice is a mediator on the path between Gender and Admission rate.

Definition 5 (Colliders). Z is called a *unshielded collider* if X, Y all

point to Z and X, Y are not adjacent. In picture, . Otherwise, the collider Z is *shielded* and part of a triangle. In picture, .

Collider is also called "inverted forks." A good example for the collider effect is the "**Berkson's paradox**": Two independent diseases can become negatively correlated when analyzing hospitalized patients. The reason is that when either disease (X or Y) is sufficient for admission to the hospital (indicated by variable Z), observing that a patient has one disease makes the other statistically less likely.

TL: Berkson's paradox $\mathbb{P}[X|X \cup Y] \neq \mathbb{P}[X|Y, X \cup Y]$ even if $\mathbb{P}[X|Y] = \mathbb{P}[X]$ (in fact, the $\mathbb{P}[X|X \cup Y] > \mathbb{P}[X|Y, X \cup Y]$).

1.6 Causal Effects and Adjustment Formula

Definition 6 (Causal Effects). Given a structural causal model M , the causal effect of an action $X := x$ on a variable Y refers to the distribution of the variable Y in the model $M[X := x]$.

When X denotes the presence or absence of an intervention or treatment, and thus a binary variable, then the average treatment effect is defined as

$$\tau := \mathbb{E}[Y|\text{do}(X := 1)] - \mathbb{E}[Y|\text{do}(X := 0)] . \tag{1.9}$$

In general, the above causal parameter τ cannot directly be replaced by the conditional probabilities

$$\tau \neq \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0] \tag{1.10}$$

We say that X and Y are **confounded** when the causal effect of action $X := x$ on Y does not coincide with the corresponding conditional probability.

Principle (Adjustment Formula).

$$\mathbb{P}[Y = y|\text{do}(X := x)] = \sum_z \mathbb{P}[Y = y|X = x, PA = z] \mathbb{P}_{M[X:=x]}[PA = z] \tag{1.11}$$

The adjustment formula is one example of what is often called *controlling for a set of variables*: We estimate the effect of X on Y separately in every slice of the population defined by a condition $Z = z$

for every possible value of z . We then average these estimated sub-population effects weighted by the probability of $Z = z$ in the population. To give an example, when we control for age, we mean that we estimate an effect separately in each possible age group and then average out the results so that each age group is weighted by the fraction of the population that falls into the age group.

Example. *Let's revisit the Exercise, Weight, and Heart Disease example and apply the adjustment formula*

$$\begin{aligned} \mathbb{P}[H = 1 | \text{do}(W := 0)] &= \mathbb{P}[H = 1 | W = 0, X = 0] \mathbb{P}_{M[W:=0]}[X = 0] + \mathbb{P}[H = 1 | W = 0, X = 1] \mathbb{P}_{M[W:=0]}[X = 1] \\ &= \mathbb{P}[H = 1 | X = 0] \mathbb{P}_{M[W:=0]}[X = 0] + \mathbb{P}[H = 1 | X = 1] \mathbb{P}_{M[W:=0]}[X = 1] \\ &= 1/3 * \mathbb{P}_{M[W:=0]}[X = 0] + 0 * \mathbb{P}_{M[W:=0]}[X = 1] \\ &= 1/3 * 1/2 = 1/6 \end{aligned}$$

In contrast, the conditional probability calculation reads

$$\begin{aligned} \mathbb{P}[H = 1 | W = 0] &= \mathbb{P}[H = 1 | W = 0, X = 0] \mathbb{P}[X = 0 | W = 0] + \mathbb{P}[H = 1 | W = 0, X = 1] \mathbb{P}[X = 1 | W = 0] \\ &= \mathbb{P}[H = 1 | X = 0] \mathbb{P}[X = 0 | W = 0] + \mathbb{P}[H = 1 | X = 1] \mathbb{P}[X = 1 | W = 0] \\ &= 1/3 * \mathbb{P}[X = 0 | W = 0] + 0 * \mathbb{P}[X = 1 | W = 0] \\ &= 1/3 * 1 = 1/3 \end{aligned}$$

For the three-graph structures

- Fork: controlling for a confounding variable Z in a fork will deconfound the effect of X on Y .
- Mediator: controlling for a mediator Z will eliminate some causal inference of X on Y .
- Collider: controlling for a collider will create a correlation between X and Y . The same is true if we control for a descendant of a collider.

1.7 Randomization and the Backdoor Criterion

Definition 7 (The Backdoor Criterion). *Two variables are cofounded if there is a so-called backdoor path between them. A backdoor path from X to Y is any path starting at X with a backward edge " \leftarrow " into X such as:*

$$X \leftarrow A \rightarrow B \leftarrow C \rightarrow Y$$

- To deconfound a pair of variables we need to select a backdoor set of variables that "blocks" all backdoor paths between the two nodes. A backdoor path involving a chain $A \rightarrow B \rightarrow C$ can be

blocked by controlling for B. Information by default cannot flow through a collider $A \rightarrow B \leftarrow C$. So we only have to be careful not to open information flow through a collider by conditioning on the collider, or descendant of a collider.

- See Pearl's d -separation⁴: valid adjustment set S that d -separates X, Y are those that for any path connect X and Y , either there exists a node Z in S that information flow $\rightarrow Z \rightarrow$ or $\leftarrow Z \leftarrow$ or $\leftarrow Z \rightarrow$, or that neither Z nor any of its descendants are in S and, creates a collider ($\rightarrow Z \leftarrow$). Definition 6.1, Proposition 6.41 in⁵ on backdoor criterion and d -separation.
- The backdoor criterion gives a non-experimental way of eliminating confounding bias given a causal model and a sufficient amount of observational data from the joint distribution of the variables.
- An alternative experimental method of eliminating confounding bias randomization. The idea is simple. If a treatment variable T is an unbiased coin toss, nothing but mere chance influenced its assignment. In particular, there cannot be a confounding variable exercising influence on both the treatment variable and a desired outcome variable.

Thinking in terms of causal models, what this means is that we eliminate all incoming edges into the treatment variable. In particular, this closes all backdoor paths and hence avoids confounding bias.

1.8 Potential Outcome Framework vs. Structural Causal Model

Consider a group of n individuals $i = 1, 2, \dots, n$, and two sequences of outcomes $\{Y_0(i)\}, \{Y_1(i)\}$ ⁶. For an individual, $Y_1(i) \in \mathbb{R}$ denotes the outcome had we applied a treatment, and $Y_0(i) \in \mathbb{R}$ denotes the outcome had we applied a control. We define the actual observed value $Y(i)$ depending on the binary treatment variable $T(i)$

$$\text{consistency : } Y(i) = T(i)Y_1(i) + (1 - T(i))Y_0(i) \quad (1.12)$$

The estimand of interest is

$$\tau_{\text{ATE}} := \mathbb{E}[Y|\text{do}(T := 1)] - \mathbb{E}[Y|\text{do}(T := 0)] = \frac{1}{n} \sum_{i=1}^n Y_1(i) - Y_0(i) \quad (1.13)$$

It is clear that counterfactual reasoning can be achieved as well in this potential outcome framework with experimental data. What happens with observation data? A set of assumptions makes counterfactual reasoning possible with data

⁴ Judea Pearl. *Causality*. Cambridge university press, 2009

⁵ Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017

⁶ Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923

- **Stable Unit Treatment Value Assumption (SUTVA):** The treatment that one unit receives does not change the effect of treatment for any other unit. No spillover.
- **Ignorability:** The potential outcomes are independent of treatment given some deconfounding variables X , namely

$$T \perp\!\!\!\perp (Y_0, Y_1) | X \quad (1.14)$$

In words, the potential outcomes are conditionally independent of treatment given some set of deconfounding variables.

A few remarks follow

- The ignorability assumption on its own cannot be verified or falsified, since we never have access to samples with both potential outcomes manifested. However, we can verify if the assumption is consistent with a given structural causal model, by checking if T, Y is d -separated by the set X ; put differently, if the set X blocks all backdoor paths from treatments T to outcome Y .
- Structural causal model can derive the potential outcome framework. The reverse is not true. A structural equation model encodes more information. However, coming up with a plausible structural causal model is often a daunting task.
- It seems the potential outcome model is much easier to apply for observational data. We will discuss next.

TL: HW: exercise?

2 In Practice

Recall the causal estimand of interest

$$\tau := \mathbb{E}[Y | \text{do}(T := 1)] - \mathbb{E}[Y | \text{do}(T := 0)] \quad (2.1)$$

and define the naive observational estimate

$$\tau_{\text{naive}} := \mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0] \quad (2.2)$$

In the randomized experiment setting, the above approach works beautifully. However, in an observational setup, typically T and Y are confounded, these two estimands are not the same, namely, probabilities do not answer causal questions $\tau_{\text{naive}} \neq \tau$. How to deal with this issue?

In practice, the above issue is addressed by adjustments, by conditioning on a set of covariates to block the confounding. We will go over the few main strategies in this section. The adjustments are done in practice by supervised machine learning methods.

2.1 Basic Regression Adjustments

Lemma 1. Assume $T \perp\!\!\!\perp (Y_0, Y_1) | X$, then

$$\tau = \mathbb{E}_x[\mathbb{E}[Y|T = 1, X = x]] - \mathbb{E}_x[\mathbb{E}[Y|T = 0, X = x]] \quad (2.3)$$

Remark the inside expectation on the RHS is identifiable with observational data.

Proof. We only need to show that, for $i \in \{0, 1\}$

$$\begin{aligned} \mathbb{E}[Y|\text{do}(T := i)] &= \mathbb{E}[Y_i] \\ &= \mathbb{E}_x[\mathbb{E}[Y_i|X = x]] \quad \text{total expectation} \\ &= \mathbb{E}_x[\mathbb{E}[Y_i|X = x, T = i]] \quad \text{ignorability} \\ &= \mathbb{E}_x[\mathbb{E}[Y|X = x, T = i]] \end{aligned}$$

□

2.2 Propensity Score Adjustments

Lemma 2. Assume $T \perp\!\!\!\perp (Y_0, Y_1) | X$ and that

$$e(x) := \mathbb{E}[T|X = x] \in (0, 1) \quad (2.4)$$

then

$$\tau = \mathbb{E} \left[Y \left(\frac{T}{e(X)} - \frac{1-T}{1-e(X)} \right) \right] \quad (2.5)$$

Remark the inside expectation on the RHS is identifiable with observational data.

Proof. We only need to show that

$$\begin{aligned} \mathbb{E}[Y|\text{do}(T := 1)] &= \mathbb{E}[Y_1] \\ &= \mathbb{E}_x[\mathbb{E}[Y_1|X = x]] \\ &= \mathbb{E}_x \left[\frac{1}{e(x)} \mathbb{E}[TY_1|X = x] \right] \quad \text{ignorability} \\ &\text{recall } TY_1 = TY = T(TY_1 + (1-T)Y_0) \\ &= \mathbb{E}_x \left[\frac{1}{e(x)} \mathbb{E}[TY|X = x] \right] \\ &= \mathbb{E} \left[Y \frac{T}{e(X)} \right] \end{aligned}$$

□

2.3 Conditional Average Treatment Effects

Define the conditional average treatment effects (CATE)

$$\tau(x) := \mathbb{E}\left[Y\left(\frac{T}{e(X)} - \frac{1-T}{1-e(X)}\right) \mid X = x\right]$$

by Lemma 2, we know that

$$\mathbb{E}_x[\tau(x)] = \tau = \mathbb{E}[Y \mid \text{do}(T := 1)] - \mathbb{E}[Y \mid \text{do}(T := 0)] \quad (2.6)$$

2.4 Doubly Robust Adjustments

Lemma 3. Assume $T \perp\!\!\!\perp (Y_0, Y_1) \mid X$. Let $g(x)$ and $e(x)$ be two functions, and construct the doubly-robust estimate

$$g(X) + \frac{T}{e(X)}(Y - g(X)) \quad (2.7)$$

then

$$\mathbb{E}[Y \mid \text{do}(T := 1)] = \mathbb{E}\left[g(X) + \frac{T}{e(X)}(Y - g(X))\right] \quad (2.8)$$

if either g or e is well specified in the sense (1) $\mathbb{E}[Y_1 \mid X = x] = g(x)$, or (2) $\mathbb{E}[T \mid X = x] = e(x)$.

Proof. If (1) $\mathbb{E}[Y_1 \mid X = x] = g(x)$ is true, then

$$\begin{aligned} \mathbb{E}[Y \mid \text{do}(T := 1)] &= \mathbb{E}[Y_1] = \mathbb{E}[g(X) + 0] \\ &= \mathbb{E}\left[g(X) + \frac{T}{e(X)}(Y_1 - g(X))\right] \\ &= \mathbb{E}\left[g(X) + \frac{T}{e(X)}(Y - g(X))\right] \end{aligned}$$

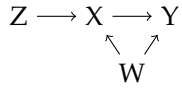
where the last step follows from $TY = TY_1$.

Alternatively, if (2) $\mathbb{E}[T \mid X = x] = e(x)$ holds

$$\begin{aligned} \mathbb{E}\left[g(X) + \frac{T}{e(X)}(Y - g(X))\right] &= \mathbb{E}\left[\mathbb{E}\left[g(X) + \frac{T}{e(X)}(Y - g(X)) \mid X\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[g(X) + \frac{T}{e(X)}(Y_1 - g(X)) \mid X\right]\right] && TY = TY_1 \\ &= \mathbb{E}\left[g(X) + \frac{\mathbb{E}[T \mid X]}{e(X)} \mathbb{E}[Y_1 - g(X) \mid X]\right] && \text{by ignorability} \\ &= \mathbb{E}[\mathbb{E}[Y_1 \mid X]] = \mathbb{E}[Y \mid \text{do}(T := 1)]. \end{aligned}$$

□

2.5 Instrumental Variables



- The instrument variable Z and the outcome Y are unconfounded
- The instrument variable Z has no direct effect on the outcome

$$Y = \alpha + \beta T + \gamma W + N \quad (2.9)$$

where the causal parameter

$$\beta := \mathbb{E}[Y|\text{do}(T := 1)] - \mathbb{E}[Y|\text{do}(T := 0)] \quad (2.10)$$

The moment conditions read

$$\mathbb{E}[Z(Y - \alpha - \beta T)] = 0 \quad (2.11)$$

$$\mathbb{E}[Y - \alpha - \beta T] = 0 \quad (2.12)$$

and thus

$$\beta = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, T)} \quad (2.13)$$

The above is equivalent to the **two-stage least squares**

- Form regression $T \sim Z$ we obtain $\hat{T} = \zeta + \xi Z$ where $\xi = \frac{\text{Cov}(T, Z)}{\xi \text{Var}(Z)}$
- Form regression $Y \sim \hat{T}$

because

$$\beta = \frac{\text{Cov}(Y, \hat{T})}{\text{Var}(\hat{T})} = \frac{\text{Cov}(Y, Z)}{\xi \text{Var}(Z)} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, T)} \quad (2.14)$$

2.6 Regression Discontinuity

References

Moritz Hardt and Benjamin Recht. *Patterns, Predictions, and Actions: A Story about Machine Learning*. Princeton University Press, 2022.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.