Causal Inference via Conditional Ignorability

Gauranga Kr. Baishya, Chennai Mathematical Institute (CMI)

May 14, 2025

Introduction

Here we discuss how average causal effects may be identified using regression when treatment is not randomly assigned but instead depends on observed covariates.

We discuss the conditional or adjustment method, which relies on comparing the average difference between expected outcomes for treated and untreated units that are comparable (formally, identical) in terms of their characteristics X.

If treatment is as good as randomly assigned conditional on X, then this approach recovers average causal or treatment effects. This key condition is commonly referred to as conditional ignorability, conditional exogeneity, or unconfoundedness.

Key Variables

1. D is the treatment variable, which indicates whether an individual or unit received the treatment. For example, $D = 1$ might mean a person got a new drug, and $D = 0$ means they did not.
2. Y is the observed outcome, the result you measure after the treatment is applied.
3. $Y(d)$ represents the potential outcome if the treatment were set to d. For instance, $Y(1)$ is what the outcome would be if the individual were treated, and $Y(0)$ is what it would be if they were not.
4. X is a set of observed covariates or characteristics (such as age, education, income, etc.) that might influence both the treatment assignment D and the outcome Y.

Assumption 1: Conditional Ignorability and Consistency

Ignorability: Treatment status D is independent of potential outcomes $Y(d)$ conditional on a set of covariates X. In other words, for each d,

$$D \perp\!\!\!\perp Y(d) \mid X.$$

We can figure out how a treatment D affects an outcome Y by looking only at people who share the same characteristics X. Within each group of people who have the same X, any differences in outcomes between those who got the treatment and those who did not can be attributed to the treatment itself, assuming *conditional ignorability* holds (meaning D is effectively random once we account for X).

Consistency: The observed outcome Y is generated by the potential outcome corresponding to the treatment received, that is,

$$Y := Y(D).$$

Selection Bias

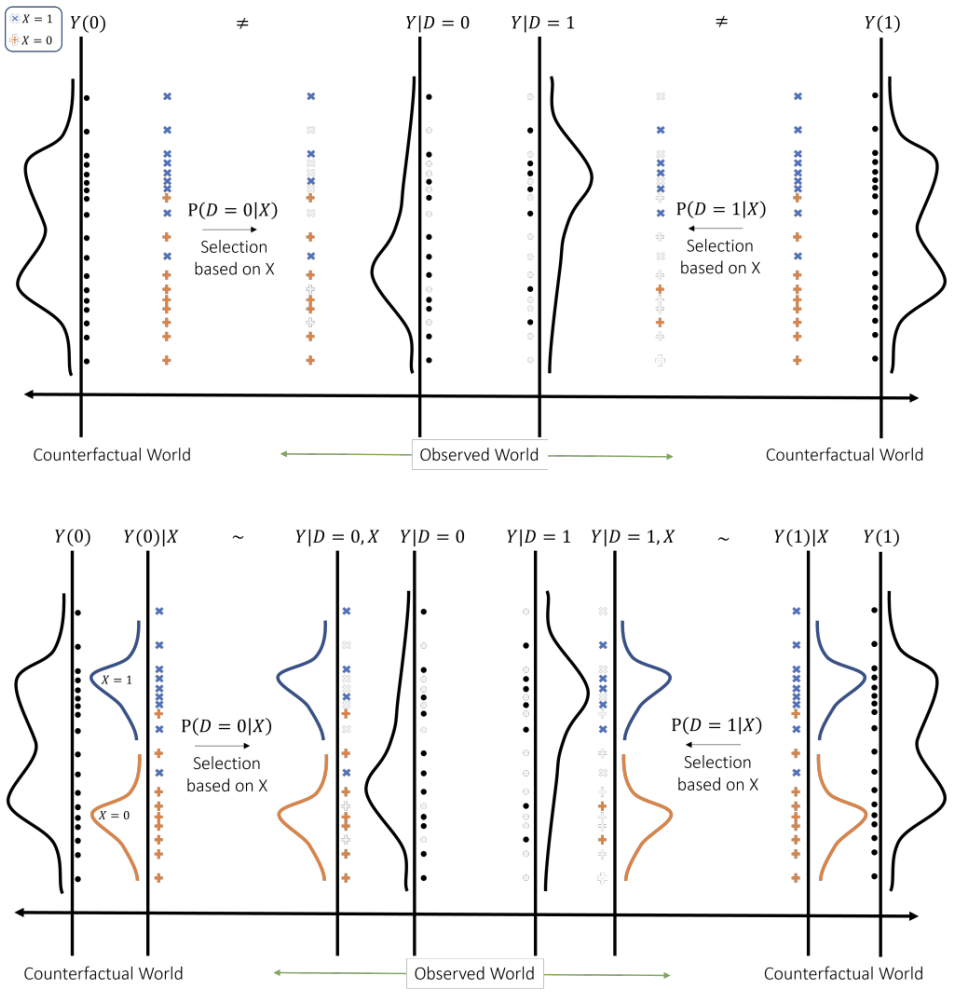


Figure 5.4: Pictorial representation of how selection on X can lead to biased observed outcomes between treated and control populations, while conditioning on X removes the selection bias. In this example, the potential outcomes $Y(0)$ and $Y(1)$ have identical distributions shown in the far left and right of the figure. We also have a binary covariate X that is related to treatment probability in the sense that $\mathrm{P}(D = 1|X = 1) > \mathrm{P}(D = 1|X = 0)$ and $\mathrm{P}(D = 0|X = 1) < \mathrm{P}(D = 0|X = 0)$ which leads to selection bias when we do not condition on X. This bias is illustrated by the difference in the distribution of (observed) Y given $D = 0$ and $D = 1$ shown in the black curves in the middle of the figure. The bottom panel then shows that selection bias is removed by conditioning on X as the distribution of potential outcomes given X (blue and orange curves under $Y(0)|X$ and $Y(1)|X$) equals the distribution of observed outcomes given D and X (blue and orange curves under $Y|D = 0, X$ and $Y|D = 1, X$).

Assumption 2: Overlap/Full Support

The probability of receiving treatment given X, called the *propensity score* and defined as

$$p(X) := P(D = 1 \mid X),$$

is non-degenerate. This means that for every value of X, the probability of treatment lies strictly between 0 and 1:

$$P\big(0 < p(X) < 1\big) = 1.$$

i.e. for every combination of characteristics X, there is always a positive chance of receiving the treatment/ not receiving it. & no matter what X is, the probability of treatment is neither 0 nor 1. This is important because it ensures that for every type of individual (defined by X), we have both treated and untreated cases to compare, which is essential for estimating causal effects.

Conditioning on X Removes Selection Bias

Under Conditional Ignorability and Overlap, the conditional expectation of the observed outcome Y given $D = d$ and X recovers the conditional expectation of the potential outcome $Y(d)$ given X:

$$E[Y \mid D = d, X] = E[Y(d) \mid D = d, X] = E[Y(d) \mid X].$$

The overlap assumption ensures that we can condition on the events $\{D = 0, X\}$ and $\{D = 1, X\}$ at any value in the support of X, and the second equality holds by ignorability.
Hence, the Conditional Average Predictive Effect (CAPE),

$$\pi(X) = E[Y \mid D = 1, X] - E[Y \mid D = 0, X],$$

is equal to the Conditional Average Treatment Effect (CATE),

$$\delta(X) = E[Y(1) \mid X] - E[Y(0) \mid X].$$

Thus,

$$APE = \delta = E[\delta(X)] = E[\pi(X)] = \pi = ATE.$$

A Directed Acyclic Graph (DAG) Illustrating Ignorability

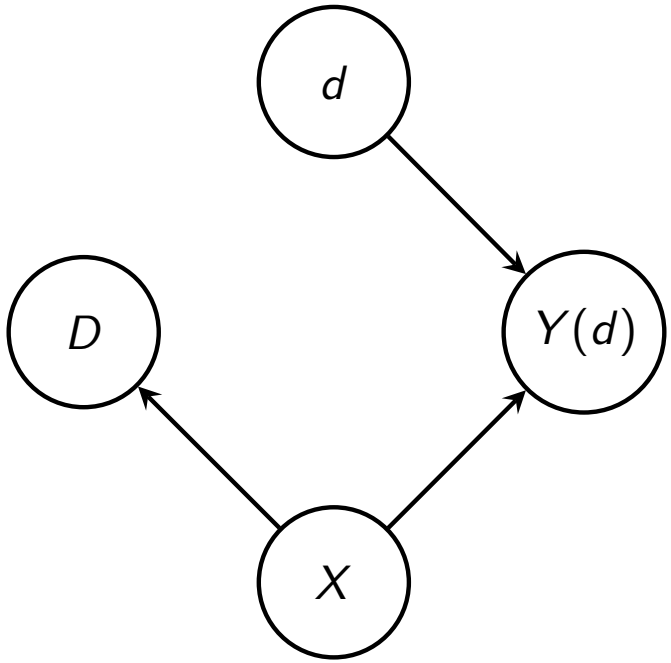
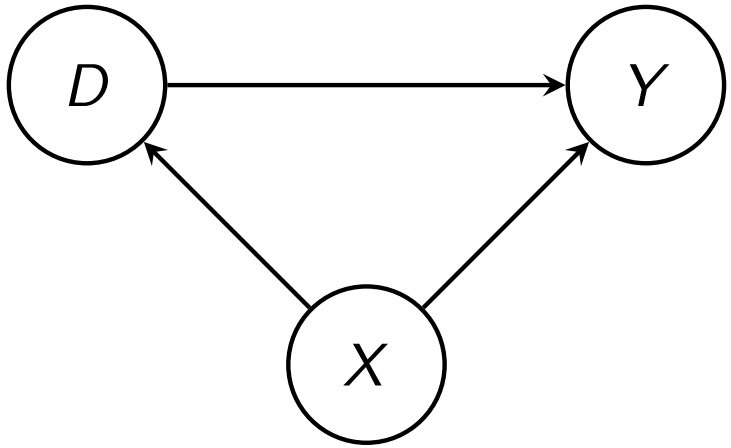


fig: Potential outcome $Y(d)$ has a node and the potential treatment status d has another node. The node d is deterministic, with an arrow from d to $Y(d)$ indicating how the potential outcome depends on the treatment status. The pre-treatment covariates X affect both the realized treatment variable D and the potential outcomes $Y(d)$, as shown by arrows from X to D and X to $Y(d)$. Under Assumption of Ignorability, once we condition on X, the assigned treatment D is independent of $Y(d)$.

Example: Drug Treatment Scenario

1. $\mathbf{X}$ represents the patient's pre-treatment characteristics (e.g., age, weight, pre-existing conditions).
2. $\mathbf{d}$ denotes the (hypothetical) potential treatment status, where $d = 0$ means no drug and $d = 1$ means drug.
3. $\mathbf{Y}(\mathbf{d})$ is the potential outcome - the blood pressure the patient would have if the treatment were set to d.
4. $\mathbf{D}$ is the actual, observed treatment assignment.
5. $\mathbf{X} \rightarrow \mathbf{D}$ and $\mathbf{X} \rightarrow \mathbf{Y}(\mathbf{d})$ show that the patient's characteristics can affect both the chance of receiving the drug (D) and how the patient's blood pressure responds to it (Y(d)).
6. $\mathbf{d} \rightarrow \mathbf{Y}(\mathbf{d})$ shows that the potential outcome depends directly on the (potential) treatment.
7. Ignorability assumption: once we condition on X, treatment assignment, D is random. Differences in outcomes among patients with the same X can be attributed solely to the drug.

Causal Diagram: $X \to D$, $X \to Y$, $D \to Y$



1. Imagine a process where we first observe or "generate" a set of background characteristics X for each individual. X might include things like age, education, or income.
2. Next, based on these characteristics, each individual is assigned a treatment D (for example, whether they receive a new drug or not).
3. Finally, the outcome Y is determined by the potential outcome function ($Y(d)$); however, in reality, we only observe the outcome corresponding to the treatment the individual actually received (that is, $Y = Y(D)$). Hence $X \to D \to Y$.

Linear Regression to estimate $ATE_{Conditional-Ignorability}$

Imagine you want to find out how a treatment D affects an outcome Y, while also taking into account other variables X that might influence both D and Y. Under the conditional ignorability assumption (meaning that once you account for X, the treatment is effectively random), one way to do this is by using a linear regression model of the form

$$E[Y \mid D, X] = \alpha D + \beta' W,$$

In practice, this corresponds to fitting the regression

$$Y = \alpha D + \beta' W + \epsilon, \quad E[\epsilon \mid D, X] = 0$$

This means that once you've accounted for D and X, $E[\epsilon] = 0$. The coefficient α tells you how much Y changes on average when D switches from 0 to 1, holding the other variables W fixed. In other words, $ATE = \delta = \alpha$.

Linear Regression to estimate $ATE_{Conditional-Ignorability}$

The assumption of linearity and homogeneous treatment effects is restrictive. A simple way to relax this is to consider interactions between W and D :

$$E[Y \mid D, X] = \alpha_1 D + \alpha_2'(W \cdot D) + \beta_1 + \beta_2' W; E[W] = 0$$

1. **ATE:** Averaging over X (and noting that $E[W] = 0$) yields the Average Treatment Effect (ATE):

$$\delta = E[\delta(X)] = \alpha_1.$$

2. **CATE:** The Conditional Average Treatment Effect (CATE),

$$\begin{aligned}\delta(X) &= E[Y \mid D = 1, X] - E[Y \mid D = 0, X] \\ &= (\alpha_1 + \alpha_2' W + \beta_1 + \beta_2' W) - (\beta_1 + \beta_2' W) \\ &= \alpha_1 + \alpha_2' W\end{aligned}$$

Identification Using Propensity Scores: Horvitz-Thompson method

The identification by conditioning approach requires accurately modeling the "outcome process," i.e. the conditional expectation function

$$E[Y \mid D, X].$$

In many real-world cases, this function can be very complex and hard to approximate. However, we might have a better understanding of the "treatment selection process," represented by the propensity score,

$$p(X) = P(D = 1 \mid X).$$

An alternative approach, known as the Horvitz-Thompson method , uses propensity score reweighting to recover averages of potential outcomes. This strategy is particularly useful when X is high-dimensional and $p(X)$ is either known or can be accurately approximated.

Identification Using Propensity Scores: Horvitz-Thompson method

Theorem: Horvitz-Thompson Propensity Score Reweighting

Under Conditional Ignorability and Overlap, the conditional expectation of an appropriately reweighted observed outcome Y, given X, identifies the conditional average of the potential outcome $Y(d)$ given X:

$$E\left[\frac{Y \cdot 1(D = d)}{P(D = d \mid X)} \,\middle|\, X\right] = E\big[Y(d) \mid X\big].$$

Then, averaging over X identifies the average potential outcome:

$$E\left[\frac{Y \cdot 1(D = d)}{P(D = d \mid X)}\right] = E\big[Y(d)\big].$$

Identification Using Propensity Scores: Horvitz-Thompson method

Proof.

$$
\begin{aligned}
E\left[\frac{Y \cdot 1(D=d)}{P(D=d \mid X)} \,\middle|\, X\right] &= E\left[\frac{Y(d) \cdot 1(D=d)}{P(D=d \mid X)} \,\middle|\, X\right] \quad (\because\ Y = Y(d) \text{ if } D=d) \\
&= \frac{1}{P(D=d \mid X)} E\Big[Y(d) \cdot 1(D=d) \mid X\Big] \\
&= \frac{1}{P(D=d \mid X)} \Big(E\big[Y(d) \mid X, D=d\big]\, P(D=d \mid X)\Big) \\
&= \frac{1}{P(D=d \mid X)} \Big(E\big[Y(d) \mid X\big] \cdot P(D=d \mid X)\Big) \\
&= E\big[Y(d) \mid X\big].
\end{aligned}
$$

The last step is due to Conditional Ignorability: $Y(d) \perp D|X$. $\square$

Horvitz-Thompson Transform

As a consequence, we can identify average treatment effects by simple averaging of transformed outcomes:

$$\delta = E[Y \cdot H],$$

where the Horvitz-Thompson transform is defined as

$$H = \frac{1\{D = 1\}}{P(D = 1 \mid X)} - \frac{1\{D = 0\}}{P(D = 0 \mid X)}.$$

Similarly, the conditional average treatment effect is identified by the conditional expectation

$$\delta(X) = E[Y \cdot H \mid X].$$

A remark

Suppose we are studying the effect of a job training program on wages. The propensity score might be calculated from various characteristics such as age, education, and work experience. Two individuals could end up with the same propensity score, meaning they have a similar overall probability of being selected for training. However, one might have a lot of work experience while the other has very little.

Using both reweighting and a regression that incorporates all these extra details (as in double machine learning) can "de-noise" the outcome-meaning it reduces unexplained variation-and provide a more precise, efficient estimate of the true treatment effect.

Covariate Balance Checks

To verify that randomization was successful, we perform a covariate balance check. Under conditional ignorability, we have

$$E[H \mid X] = 0.$$

Thus, if covariates predict H, then conditional ignorability does not hold. Heuristically, if covariates are able to predict H, it means that even after reweighting there remain systematic differences in X across treatment and control groups.

In a low-dimensional linear model framework, a covariate balance check can be performed by regressing H on W, a dictionary of transformations of X, and testing whether W significantly predicts H. If W predicts H, this indicates that the RCT's randomization protocol did not work as planned.