18.650: Fundamentals of Statistics November 29, 2023

Lecture 28 — Causal Inference

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

If we see a linear relationship between X and Y in a scatterplot of (X_i, Y_i) pairs, then we know the two are *correlated*:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} \in [-1, 1].$$

Correlation tells us about the presence of a *linear* relationship: the closer the points are to falling on a line of positive slope (negative slope) the closer the correlation coefficient is to one (negative one).

Remark.

> If $Y = aX$ for any $a > 0$, then $\text{Corr}(X, Y) = 1$. The slope of the line does not impact the value of the correlation coefficient.

But the presence of a linear relationship does not mean that X *causes* Y. As you have probably heard many times, correlation/association does not imply causation. Roughly speaking, X *causes* Y if changing X necessarily changes Y (there is typically a time component to this, where X happens before Y). The following are examples in which we are interested in whether X causes Y.

X	Y
Drug dose	response
Financial support	professional success
Tax incentive	jobs

1 Counterfactual model

The counterfactual model is a framework to discuss causation. We let $X \in \{0, 1\}$ denote the random variable which tells us whether or not a "treatment" was applied (0 if no, 1 if yes). Note $X = 0$ is also denoted "control". Here, treatment could be medical but not necessarily; in the above examples, financial support would also be called the treatment. We let $Y \in \mathbb{R}$ be the response — it is sometimes binary but in general can be any real number.

We also introduce two more random variables called the *potential outcomes*:

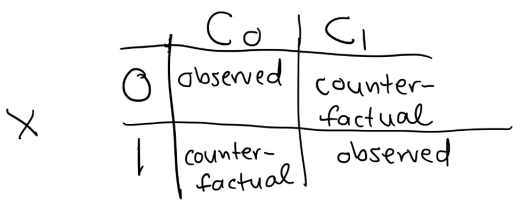

Figure 1: When $X = 0$, then C_0 is observed and C_1 is the counterfactual. When $X = 1$, then C_1 is observed and C_0 is the counterfactual.

C_0, C_1 are random variables such that

$$Y = \begin{cases} C_0 & \text{if } X = 0 \\ C_1 & \text{if } X = 1. \end{cases} \quad \Leftrightarrow Y = C_X$$

For each subject, only one of the two outcomes can be observed. E.g. in a medical trial, if you are assigned to take the drug and go through the treatment you can't then go back in time and see what would have happened had you been assigned to the control group instead. Therefore, when $X = 0$ we call C_0 "observed" and C_1 the *counterfactual*. Conversely, when $X = 1$ we call C_1 "observed" and C_0 the counterfactual.

So as an example, such data could look as follows:

X	Y	C_0	C_1
0	-3	-3	?
0	-2	-2	?
1	3	?	3
1	2.5	?	2.5

The key quantity of interest is the following:

Definition 1.1: Average treatment effect (ATE)

$$\theta = \mathbb{E}[C_1] - \mathbb{E}[C_0]$$

θ tells us the expected difference in the outcome due to the treatment; hence the name "average treatment effect". Another related but different quantity is the *association*:

Definition 1.2: Association

$$\alpha = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0].$$

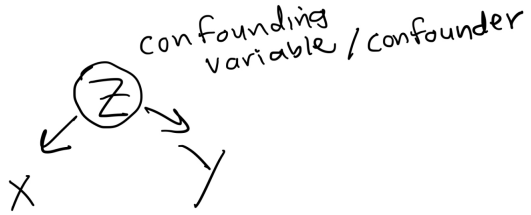

Figure 2: A confounder/confounding variable is a factor which impacts both X (how the decision is made to assign subjects to groups), and Y (the outcome itself).

In general, $\alpha \neq \theta$, as the following example shows.

Example.

> Let $Z \sim \text{Unif}([-1, 1])$, and let X (control vs treatment) be given by
>
> $$X = \mathbb{1}(Z > 0).$$
>
> Let the two outcomes be
>
> $$C_0 = Z, \qquad C_1 = Z.$$
>
> Clearly, the treatment has no effect! In particular, $\theta = \mathbb{E}[C_1] - \mathbb{E}[C_0] = 0$. Next we compute the association α. Note that $Y = Z$ in this set-up, so
>
> $$\mathbb{E}[Y|X = 1] = \mathbb{E}[Z|Z > 0] = 1/2, \quad \mathbb{E}[Y|X = 0] = \mathbb{E}[Z|Z < 0] = -1/2$$
>
> So $\alpha = 1/2 - (-1/2) = 1 \neq 0$, the value of θ.

The reason $\alpha \neq \theta$ is due to the presence of the "*confounding variable*" Z: a factor which impacts both X (how the decision is made to assign subjects to groups), and Y (the outcome itself). This often happens in real life. For example, the decision to give someone a drug could be based on a factor (e.g. current health) which also affects the outcome (e.g. how much their health improves).

2 Computing θ using α

The ideal case is when θ — the quantity we're actually interested — is equal to α. This is because α (unlike θ) can be estimated from the data! Indeed, we can simply use the plug-in estimator

$$\hat{\alpha} = \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i = 1)}{\sum_{i=1}^n \mathbb{1}(X_i = 1)} - \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i = 0)}{\sum_{i=1}^n \mathbb{1}(X_i = 0)}. \tag{1}$$

In some cases, we can design an experiment to guarantee $\theta = \alpha$. This happens in *randomized control trials* (RCTs), also known as "A/B testing" in data science. In RCTs, the decision about who gets the treatment is made by flipping an independent coin. This guarantees there is no confounding factor — you won't have unintentionally impacted the outcome Y.

Theorem 2.1: Randomized control trials

Suppose subjects are assigned independently at random:

$$X_i = \begin{cases} 1 & \text{with prob. } p \in (0,1) \\ 0 & \text{with prob. } 1-p. \end{cases}$$

Then $\theta = \alpha$. Therefore, $\hat{\alpha}$ given in (1) is a consistent estimator of θ.

Proof. Due to the RCT set-up, we know $(C_0, C_1) \perp\!\!\!\perp X$ (this is called "strong unconfoundedness"). We therefore have

$$\alpha = \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] = \mathbb{E}[C_1|X=1] - \mathbb{E}[C_0|X=0] = \mathbb{E}[C_1] - \mathbb{E}[C_0] = \theta.$$

We are able to drop the conditional due to (C_0, C_1) being independent of X. □

Remark.

RCTs can be unethical: if an ailing person could strongly benefit from a drug being tested, denying them this treatment due to a flip of a coin is objectionable.

It can often be challenging to ensure the outcome is independent of X — or unethical, as discussed above. But sometimes we can find a random feature vector Z (e.g. Z = (age,salary, blood pressure, genotype) such that C_0, C_1 are independent of X *conditionally on* Z:

$$C_i \perp\!\!\!\perp X \mid Z \quad \text{for } i = 0, 1.$$

We can then split up the observations based on the value of Z. For example, if Z takes only two values Z_1 and Z_2, then θ will equal α in each of the two cases:

$$\hat{\theta}_1 = \hat{\alpha}(Z_1), \quad \hat{\theta}_2 = \hat{\alpha}(Z_2).$$

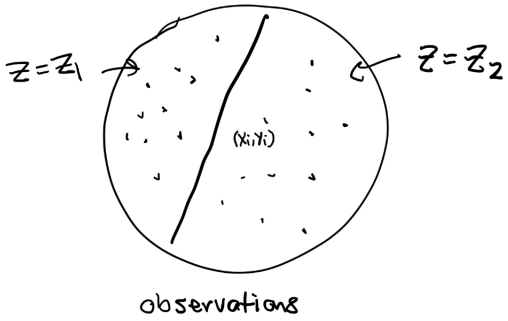

Figure 3: Estimating $\hat{\theta} = \hat{\alpha}$ for each value of Z separately (in the case where X is independent of the outcomes when conditioned on Z.

However, the issue is that often Z takes many more than just 2 values! In fact, Z is usually high-dimensional, so in practice you would have to discretize this very high dimensional space into a large number of bins. Doing so would lead to the issue of there being only few observations in each bin.

This motivates paring Z down to a single number in the unit interval, which is much easier to bin up.

Definition 2.2: Propensity score

$$p(z) = \mathbb{P}(X = 1 | Z = z).$$

The propensity score captures how we decide to assign treatment/control based on the value of z.

The following theorem shows we can condition on $p(z)$ instead of on the possibly very high-dimensional z:

Theorem 2.3: Rosenbaum & Rubin '83

Suppose $C_i \perp\!\!\!\perp X \mid Z$. Then we also have that $C_i \perp\!\!\!\perp X \mid p(Z)$.

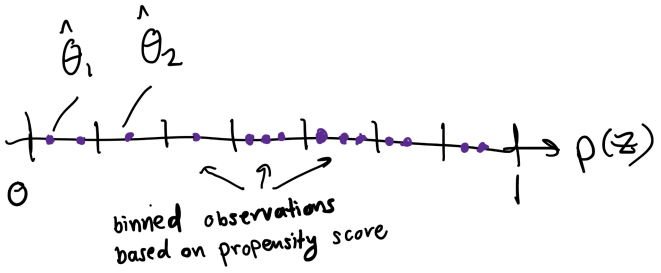

Figure 4: Estimating $\hat{\theta}$ using the propensity score.

We now simply discretize the unit interval and compute $\hat{\theta}_j$ for bin j based on observations whose propensity score falls into that bin.

Remark.

> Typically, $p(z)$ is not known explicitly. However, we can learn $p(z)$ using logistic regression:
>
> $$\text{logit}(p(z)) = z^T \beta^*.$$