DS 102 Data, Inference, and Decisions **Fall 2019**

Lecture 13: Causal Inference

Lecturer: Peng Ding

1 Motivation

Causality is not commonly covered in data science or statistics courses. Instead we often only talk about correlation instead of causation. That is, we often only talk about the association between random variables. But correlation is often not enough. For example, there is a strong correlation between the amount of chocolate consumed and the number of Nobel prize winners in any given year, but this in no way implies that one causes the other. But then if correlation is not enough to imply causation, what is?

1.1 Regression and Causal Inference

Let's begin by considering the simplified setting of linear regression. In linear regression we observe datapoints $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Furthermore we assume that the outputs are related to the inputs according to the following linear equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_d x_{id} + \epsilon,$$

where each β_j is an unknown parameter and ϵ is a zero-mean random variable. From this we can also write that

$$\mathbb{E}[y_i | x_{i1}, \ldots, x_{id}] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_d x_{id}.$$

It is immediately apparent that changing the value of x_{i1} has an effect on the value of y_i. So it seems intuitive to interpret β_1 as telling us what "causal effect" the value of x_{i1} has on y_i.

In the next section we will more precisely define what we mean by a causal effect. There are multiple frameworks used to define what a causal effect means. The one we will use is the potential outcomes framework by Jerzy Neyman.

2 Randomized Experiments

We'll first consider the simpler setting of an experiment where we are interested in determining the causal effect of a treatment on the outcome of the experiment. Furthermore, we'll start by discussing the idealized setting where we get to randomly assign who gets the treatment and who doesn't. This is called a randomized experiment and is used in many different fields.

- When the food and drug administration wishes to determine whether a drug should be approved or not they do so through a randomized controlled trial (RCT). In this setting the FDA randomly assigns patients to take the new drug under trial while also randomly assigning the rest of the patients to take a placebo. After awhile they then compare the outcome between the patients that took the drugs and the ones that took the placebo.
- In social science there is a concept of a field experiment, in which a sub-population will randomly be assigned a specific policy. Recently three economists from MIT and Harvard won the Nobel prize for conducting field experiments in developing countries to determine the effectiveness of various policies in reducing poverty.
- Within technology companies A/B tests, such as tests that determine whether a specific ad placement leads to increased purchases can be seen as a randomized experiment.

In the context of randomized experiments we have units $i = 1, 2, \ldots, n$. Each unit either receives the treatment or the control, where $z_i = 0$ for the control and $z_i = 1$ for the treatment. We then indicate by y_i the outcome observed by unit i. For example i could be the index of patients, y_i could be an indicator on whether the patient gets better or not, and $z_i = 1$ could indicate that the patient was assigned to take the drug, while $z_i = 0$ would indicate that they were assigned the placebo.

So far we have only used notation already available within classical statistics. However we now introduce the concept of potential outcomes $y_i(1)$ and $y_i(0)$ which can be read as the outcome that would have happened if we had given the treatment to unit i and the outcome had we given the control to unit i respectively.

Given this new notation we can now consider the *individual causal effect*

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

In practice we have no hope of finding the individual causal effect as exemplified in Table 13.1 where we see that we only get to observe one outcome, while the other potential outcome, called the *counterfactual*, will forever remain unobserved.

Unit	Z	Y(1)	Y(0)
1	1	✓	?
2	1	✓	?
3	0	?	✓
⋮	⋮	⋮	⋮
n - 1	0	?	✓
n	1	✓	?

Table 13.1: An example of an experiment where a checkmark represents an observed outcome while a question mark represents a counterfactual.

So we have no hope on an individual basis but can we do better on average? That is can we

estimate the average treatment effect (ATE)

$$\begin{aligned}\tau &= \mathbb{E}[Y(1) - Y(0)] \\ &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].\end{aligned}$$

2.1 Inference for τ ?

Without randomization causal inference is hard. For example say we were investigating whether smoking causes lung cancer. In this setting $y_i = 1$ if a person in our dataset contracts lung cancer within their lifetime, while $y_i = 0$ if they do not. Were we to simply take the empirical mean of the outcomes for smokers and non-smokers as our estimates of $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ respectively that is

$$\begin{aligned}\hat{\tau} &= \hat{\mathbb{E}}[Y(1)] - \hat{\mathbb{E}}[Y(0)] \\ &= \frac{1}{n_1} \sum_{i:z_i=1} y_i - \frac{1}{n_0} \sum_{i:z_i=0} y_i.\end{aligned}$$

Where n_1 is the number of smokers in our dataset and n_0 is the number of non-smokers. However, if people who tend to smoke are somehow fundamentally different from those who don't, then the above estimate will be biased. For example, smokers might just naturally be more prone to lung cancer, this corresponds to the case shown in Table 13.2.

Unit	Z	Y(1)	Y(0)
1	1	1	1
2	1	1	1
3	1	1	1
⋮	⋮	⋮	⋮
n - 1	0	0	0
n	0	0	0

Table 13.2: A setting where the treatment Z is not independent from the outcome $Y(1)$ and $Y(0)$.

We will further explore the smoking example later in the lecture, however the key takeaway here is that it is desirable to have the treatment be independent from the outcome when estimating treatment effect. One simple solution is to randomly assign treatments in a randomized experiments, notationally this gives us that

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\}.$$

This is not a trivial fact since in most cases

$$Z \not\perp\!\!\!\perp Y,$$

as in the example shown in Table 13.2. With randomization we are able to further simplify the

average treatment effect to

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

where we have used our independence assumption in the second equality. In this case the naive estimator that we developed at the beginning of this section is unbiased. We can rewrite this estimator as

$$\hat{\tau} = \hat{y}(1) - \hat{y}(0)$$

with

$$\begin{aligned}\hat{y}(1) &= \frac{1}{n_1} \sum_{i:z_i=1} y_i \\ \hat{y}(0) &= \frac{1}{n_0} \sum_{i:z_i=1} y_i,\end{aligned}$$

where n_1 is the number of units assigned the treatment and n_0 is the number of units assigned the control. Since $\hat{y}(1)$ and $\hat{y}(0)$ are independent we have that

$$\mathbb{V}(\hat{\tau}) = \frac{\mathbb{V}(Y(1))}{n_1} + \frac{\mathbb{V}(Y(0))}{n_0}.$$

Hence we can estimate the variance of τ as

$$\hat{v} = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0},$$

where s_1^2, s_0^2 are the corresponding sample variances. Approximating the distribution of $\hat{\tau}$ as a normal distribution[1] gives us a 95% confidence interval of

$$\hat{\tau} \pm 1.96\sqrt{\hat{v}}.$$

2.2 Covariates

Often we have covariates x_i associated with each unit which we would like to use to improve the efficiency of our data. For example in an FDA randomized control trial x_i might include the age of the patient, their medical history, and where they live.

In this setting we can use the dataset of units assigned to the treatment group

$$\{(x_i, y_i)\}_{z_i=1}$$

to train a predictor[2]: $\hat{\mu}_1(x)$ which will output an estimated outcome when the treatment is applied to a unit with covariates x. We can also use the dataset of units assigned to the control group

$$\{(x_i, y_i)\}_{z_i=0}$$

[1] The central limit theorem tells us that this is a valid thing to do asymptotically.

[2] In practice this can be any model, from a linear regression model to a neural network.

to build another predictor: $\hat{\mu}_0(x)$ which will output an estimated outcome when the treatment isn't applied to a unit with covariates x. We then use the two predictors to get the following estimator of the average treatment effect

$$\hat{\tau} = \frac{1}{n}\left[\sum_{z_i=1} y_i + \sum_{z_i=0} \hat{\mu}_1(x_i)\right] - \frac{1}{n}\left[\sum_{z_i=1} \hat{\mu}_0(x_i) + \sum_{z_i=0} y_i\right].$$

Intuitively we are using the two predictors to fill in the missing entries in a table similar to Table 13.1. We are then using these predictions along with our real observed data to estimate the ATE.

In his PhD thesis entitled "Essays on Causal Inference in Randomized Experiments" Winston Lin shows that using linear regressors to estimate the counterfactuals does indeed lead to a better estimator of the ATE. Although the discussion is quite technical so we refer the interested reader to Winston Lin's thesis.

3 Non-randomized studies

Randomized studies aren't always possible. For example we can't force people to smoke or to go to graduate school so we would like to handle non-randomized studies, also known as observational studies, in a principled manner. In non-randomized studies we have

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

This setting is often called the Rubin Causal model or the Neyman-Rubin model. As we saw in the last section we can't use the naive empirical mean estimator in this setting, for example applying this estimator to the data shown in Table 13.2 would lead to a highly biased estimate.

This is a major reason as to why the debate about whether or not smoking causes lung cancer was so hard to resolve. For example, Ronald Fisher, a famous statistician, believed smoking did not cause lung cancer, despite the fact that there was a clear correlation. He instead posited there was a hidden gene that made people more likely to smoke while also increasing the probability they would contract lung cancer. In other words, his theory was that there was a common unobserved cause to both lung cancer and propensity to smoke. We call any such unobserved cause that effects both the variables Z and Y a *confounder*. While we now know that smoking does indeed cause lung cancer, observational studies are still approached with caution. Even today, people will be very suspicious of observational studies since there can always be some unobserved confounder.

One way to get around the issue presented by non-randomized studies is to assume we've collected enough covariates such that

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

this assumption has many names. In epistemology it is called the *unconfoundedness assumption*, in statistics it is called *ignorability* and in economics it is called *selection on observables*.

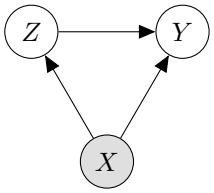

Figure 13.1: A causal graphical model showing the unconfoundedness assumption. Here we see that all common dependencies of Y and Z are captured by the covariates X.

Under ignorability we have

$$\begin{aligned}\tau &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}[\mathbb{E}[Y(1)|X]] - \mathbb{E}[\mathbb{E}[Y(0)|X]] \\ &= \mathbb{E}[\mathbb{E}[Y(1)|X, Z=1]] - \mathbb{E}[\mathbb{E}[Y(0)|X, Z=0]] \\ &= \mathbb{E}[\mathbb{E}[Y|X, Z=1]] - \mathbb{E}[\mathbb{E}[Y|X, Z=0]].\end{aligned}$$

Where we have used the ignorability assumption for the third equality. For a more concrete example let's consider the case where the covariate consists of a single discrete number $X \in \{1, 2, \ldots, K\}$. For example X could indicate whether someone comes from a specific region. Assuming that we have ignorability given X then we have

$$\begin{aligned}\tau &= \mathbb{E}[\mathbb{E}[Y|X, Z=1]] - \mathbb{E}[\mathbb{E}[Y|X, Z=0]] \\ &= \sum_{k=1}^{K} \mathbb{E}[Y|X=k, Z=1]\mathbb{P}(X=k) - \sum_{k=1}^{K} \mathbb{E}[Y|X=k, Z=0]\mathbb{P}(X=k).\end{aligned}$$

We will see how to use this formula in the next lecture.