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

Lecture 14: Causal Inference II

Lecturer: Peng Ding

1 Review

In the last lecture we introduced the concept of causal inference. One of the simpler settings in causal inference is the case where we want to determine whether a treatment has an effect on a given outcome. In this setting we have units $i \in \{1, 2, \ldots, n\}$ with potential outcomes $Y_i(0)$ and $Y_i(1)$, which can be read as the outcome that would have occurred had the unit been assigned the control or the treatment respectively. The main quantity of interest is the individual causal effect defined as

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

In nearly all situations we have no hope of looking at the individual causal effect but we have some hope of estimating the average causal effect, also known as the average treatment effect.

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

In the last lecture we looked at randomized experiments where we randomly assign whether each unit gets a treatment or not (Z_i) at random, in that case we have

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

which gives us

$$\tau = \mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0].$$

2 Non-randomized studies continued

In the last lecture we also started looking at observational studies in which we do not control the assignment of Z_i, that is we might have

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

Our only hope then is that we have collected enough covariates X such that

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

this is called ignorability or unconfoundedness and it implies that

$$\tau = \mathbb{E}\left[\mathbb{E}[Y|Z = 1, X = x] - \mathbb{E}[Y|Z = 0, X = x]\right].$$

This is an incredibly strong assumption that is also untestable! To see this note that

$$Z \perp\!\!\!\perp Y(1)|X \iff \mathbb{P}(Y(1)|Z=1,X) = \mathbb{P}(Y(1)|Z=0,X)$$

and

$$Z \perp\!\!\!\perp Y(0)|X \iff \mathbb{P}(Y(0)|Z=1,X) = \mathbb{P}(Y(0)|Z=0,X)$$

but the second probability in the first equation and the first probability in the second equation are counterfactuals!

2.1 Estimators

Assuming that we do indeed have ignorability given X we can then form an unbiased estimator of τ. In this section we explore a few different variations on estimators.

Discrete X

Recall the simplified case from last lecture where $X \in \{1, 2, \ldots, K\}$ is a discrete random variable. We showed that ignorability gives us that

$$\tau = \sum_{i=1}^{K} (\mathbb{E}[Y|X=k, Z=1] - \mathbb{E}[Y|X=k, Z=1])\mathbb{P}(X=k)$$

For example, we might be interested in whether being tall (Z) has a causal effect on longevity (Y). While being tall might result in a shorter expected lifespan, there might be some regions (X) where people live longer and are also taller. The above equation tells us that if we have ignorability conditioned on X, we can simply consider every region separately when computing τ. Splitting along the regions allows us to treat each sub-region as if they were randomized experiments[1]. Given the above formula for τ, ignorability gives us the following natural estimator when X is discrete

$$\hat{\tau} = \sum_{k=1}^{K} (\hat{Y}_{k1} - \hat{Y}_{k0}) \frac{n_k}{n},$$

with

$$\hat{Y}_{k1} = \frac{1}{n_{k1}} \sum_{\substack{i:z_i=1,\\ x_i=k}} y_i$$

$$\hat{Y}_{k0} = \frac{1}{n_{k0}} \sum_{\substack{i:z_i=0,\\ x_i=k}} y_i.$$

Where n_k is the number of units with covariate equal to k, n_{k0} is the number of units that take the control and have covariate equal to k, and n_{k1} is the number of units that take the treatment and

[1] Of course this is making the strong assumption that the only confounder is the region, which is not true in practice.

have covariate equal to k. We are essentially estimating the average treatment effect per sub-group and then computing the overall ATE by reweighting the subgroup ATEs based on the size of each subgroup's population.

Continuous X

In the case where X is not a discrete random variable but is instead continuous we have that:

$$\begin{aligned}\tau &= \mathbb{E}\left[\mathbb{E}[Y|Z=1,X=x] - \mathbb{E}[Y|Z=0,X=x]\right] \\ &= \int \left(\mathbb{E}[Y|Z=1,X=x] - \mathbb{E}[Y|Z=0,X=x]\right)\mathbb{P}(x)dx,\end{aligned}$$

estimating the integral (i.e. the outer expectation) can just be achieved by sampling x from its distribution, just as you saw in Lecture 10. Hence we can just take the empirical mean over the covariates in our dataset to estimate the outer expectation. Unfortunately we can't repeat the same procedure for the inner expectation since we won't have enough samples to estimate both inner expectations for all values of x we encounter. We instead fit two models[2] to estimate the two inner expectations:

$$\begin{aligned}\hat{\mathbb{E}}[Y|Z=1,X] &= \hat{\mu}_1(X) \\ \hat{\mathbb{E}}[Y|Z=0,X] &= \hat{\mu}_0(X),\end{aligned}$$

where $\hat{\mu}_1$ and $\hat{\mu}_0$ are trained on all data from units that have taken the treatment and control respectively. We then have

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(x_i) - \hat{\mu}_0(x_i).$$

However these types of estimators have come under criticism by the statistical community since researchers can easily search over the space of possible models to represent $\hat{\mu}_1, \hat{\mu}_0$ until they reach a desired value of $\hat{\tau}$ that matches their desired narrative.

Stratified propensity score

Given the problems with the previous model we consider two alternate methods of estimating τ under ignorability that can't be gamed quite as easily. Both these methods are based on the concept of a propensity score defined as

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

that is the probability that a unit with covariate X takes the treatment. Unfortunately this quantity is hidden to us so we instead estimate it as $\hat{e}(X)$ using a logistic regression model.

Rosenbaum & Rubin proposed a method of estimating τ by stratifying over our estimate of $e(X)$. The fundamental assumption is here is that $Z \perp\!\!\!\perp (Y(1), Y(0))|e(X)$. The procedure works as follows

1. Fit a model $\hat{e}(X)$

[2] Just as in Section 2.2 of Lecture 13 these models can be anything, from a linear regression model to a neural network. Although the canonical choice is usually a linear model.

2. Discretize $\hat{e}(X)$ into K strata: $\left\{(0, \frac{1}{K}), (\frac{2}{K}, \frac{3}{K}) \ldots, (\frac{K-1}{K}, 1)\right\}$.

3. Assuming that in each stratum k assume we have a randomized control trial then we may re-use the estimator for discrete covariates by treating the stratified $\hat{e}(X)$ as our covariate:

$$\hat{\tau} = \sum_{k=1}^{K} (\hat{Y}_{k1} - \hat{Y}_{k0}) \frac{n_k}{n},$$

where

$$\hat{Y}_{k1} = \frac{1}{n_{k1}} \sum_{\substack{i: z_i = 1 \\ x_i \in S_k}} y_i$$
$$\hat{Y}_{k0} = \frac{1}{n_{k0}} \sum_{\substack{i: z_i = 0 \\ x_i \in S_k}} y_i.$$

With $S_k = [\frac{k-1}{K}, \frac{k}{K})$, n_k as the number of units with $\hat{e}(X)$ in stratum k, n_{k0} as the number of units in stratum k who have taken the control, and n_{k1} being the corresponding number of units within that same stratum that have taken the treatment.

Inverse Propensity Weighting

A more natural estimator that also uses the propensity estimate $\hat{e}(X)$ is called inverse propensity weighting. In particular it makes us of the fact that assuming ignorability we have

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

To see this note that

$$\begin{aligned}
\mathbb{E}\left[\frac{ZY}{e(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{ZY}{e(X)}\middle| X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{ZY(1)}{e(X)}\middle| X\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[Y(1)|X\right]\mathbb{E}[Z|X]}{e(X)}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y(1)|X\right]\right] \\
&= \mathbb{E}[Y(1)].
\end{aligned}$$

Where we have used ignorability in the third equality and the fact that the mean of a Bernoulli random variable is equal to the probability that it will be 1 in the fourth equality. The proof of the

second equality is similar. Hence we can directly estimate both $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ by using the above equalities and taking empirical means to get an estimate of τ:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \frac{z_i y_i}{\hat{e}(x_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-z_i)y_i}{1-\hat{e}(x_i)}.$$

3 Instrumental Variables

We've looked at two extremes: the randomized control trial where we have complete control of the treatments, and the observational study where we have no control of the treatments. Instrumental variables look at a middle.

In the case of observational studies, while we can collect many covariates as possible, there always exists the possibility that there is some unobserved confounding variable. Going back to the example in Lecture 13, this was exactly Fisher's argument as to why smoking might not cause cancer. In trying to determine whether smoking causes cancer, we can't conduct a randomized control trial either due to ethical issues.

To remedy this issue we might employ *encouragement design*. In this setting we randomly select people and encourage them not to smoke by giving them \$500. The encouragement should have no effect on the confounders or the probability of cancer directly but does have an effect on whether or not someone will smoke.

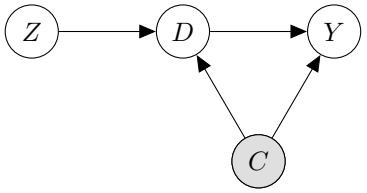

Figure 14.1: A causal graphical model showing the instrumental variable setup. The instrumental variable Z does not affect any confounders C, nor does it affect the outcome Y, except through the treatment D.

This is exactly the situation shown in Figure 14.1, where Y is whether someone will contract cancer, D is whether someone smokes, and Z is whether someone got encouraged not to smoke. The encouragement Z is called the *instrumental variable*. Instrumental variables must satisfy two conditions: (1) they can not have any effect on confounding variables, and (2) they can not affect the outcome Y directly, instead they must affect the treatment variable D.

In this setting we have four potential outcomes to consider with respect to Z: $D(1)$ which corresponds to whether someone smoked when getting encouraged not to, $D(0)$ which corresponds to whether someone smoked without receiving any encouragement, $Y(1)$ which corresponds to whether someone contracted lung cancer when getting encouraged not to smoke, and $Y(0)$ which corresponds to whether someone contracted lung cancer when no receiving any encouragement.

Now using the law of total probability gives us

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

Since the instrumental variable can only affect the outcome Y through D it follows that $D(1) = D(0) \implies Y(1) = Y(0)$, hence the first and fourth terms in the sum are equal to 0. Furthermore we assume that our encouragement does not have a counter-productive effect. For example, paying people to stop smoking will not incentivize them to start smoking instead. This is also called the no defiers assumption, expressed mathematically it states that $D(1) \geq D(0)$ which in turn implies that $\mathbb{P}[D(1) = 0, D(0) = 1] = 0$. Hence the third term is also equal to 0.

Hence all terms except the second line are zero and we have

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

Repeating the above argument with Y replaced with D we can show that

$$\mathbb{E}[D(1)-D(0)] = \mathbb{P}(D(1)=1,D(0)=0).$$

Combining the above results and using the fact that $Z \perp\!\!\!\perp \{Y(1), Y(0), D(1), D(0)\}$ gives us

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

this quantity is called the complier average causal effect or the local average treatment effect (LATE) and can easily be estimated by computing empirical means of the four conditional expectations. However, we note that this is only an estimator of the causal effect for compliers, that is people for which the encouragement actually works to change their behaviour.

Our presentation of instrumental variables is fairly recent. An alternative viewpoint is the two-stage least squares problem which works as follows:

1. First fit a linear model with Z as inputs and predict on D as $\hat{D}$.
2. Then fit a linear model to predict Y on the fitted value, $\hat{D}$ as $\hat{Y}$.

We can show that

$$\hat{\tau} = \frac{\hat{Y}_1 - \hat{Y}_0}{\hat{D}_1 - \hat{D}_0}.$$

We delve into this algorithm further in Discussion 7.