36-705: Intermediate Statistics **Fall 2019**

Lecture 29: November 11

Lecturer: Siva Balakrishnan

We will continue our discussion of causal inference, and if time permits then we will turn our attention to the problem of regression.

29.1 Things worth noting

In the last lecture we talked about estimating the average treatment effect in a randomized trial, i.e. when

$$W \perp\!\!\!\perp (Y(0), Y(1)).$$

When this assumption holds we noted that we could estimate the ATE τ using the difference of the average outcome in the treatment group (individuals for whom $W = 1$) and the average outcome in the control group (individuals for whom $W = 0$).

Randomly assigning treatment allows us to estimate causal quantities (which depend on counterfactuals) using only observed outcomes.

When we discussed point estimation in this class, we started with an "identified parameter" (i.e. something that could be estimated from the observed data) and then designed an estimate for it. When we moved to causal inference, the parameter of interest:

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

is not expressed in terms of observed quantities, and we need an assumption (for instance, randomization of W) to make progress. Causal inference is most clearly thought about in two steps:

1. **Identification:** Leveraging some set of "causal assumptions" in order to link the parameter of interest to something that can be derived from the observed data distribution. In a simple randomized trial, we used the assumption $W \perp\!\!\!\perp (Y(0), Y(1))$ to say that,

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

2. **Estimation:** Once we've "identified" the parameter (i.e. written it in terms of observed quantities) we can design an estimator for it, using for instance techniques we introduced for point estimation.

The final point that is worth noting, is that the potential outcomes notation already hides some assumption that are worth thinking about. In particular, if we say that we observe for each individual i:

$$Y_i^{\text{obs}} = Y_i(1)W_i + (1 - W_i)Y_i(0),$$

then we are implicitly assuming:

1. The treatment is binary and well-defined, i.e. it cannot be the case that some people take 10 aspirins, and some people take 2 when assigned treatment (or we would need to define more potential outcomes).

2. Similarly, and more substantially we are implicitly assuming that the observed potential outcome is not dependent in any way on the treatment status of other individuals, i.e. Y_i^{obs} only depends on W_i and not on the other Ws. This might seem reasonable, but is for example not generally true in a vaccine trial where if all my friends are treated (take some vaccine) this likely reduces my own chance of getting the disease even if I don't get treated (so my potential outcomes depend on the treatment status of all my friends).

These assumptions are sometimes called the Stable Unit Treatment Value Assumption (SUTVA).

Finally, we won't belabor this point, but in general causal inference in a randomized trial doesn't rely on anything other than the treatment being random, i.e. the units can be fixed/randomly drawn from some population, the potential outcomes can be fixed/random, and everything we've said so far will make sense (with just a little bit of care). However, this won't necessarily be the case when we start discussing observational studies. For observational studies we will assume that: e observe triplets from some distribution $(X, W, Y^{\text{obs}}) \sim F$, where as usual

$$Y^{\text{obs}} = WY(1) + (1 - W)Y(0).$$

We will assume all of these are stochastic. Sometimes we will use the subscript $\mathbb{E}_X$ to denote taking an expectation with respect to the randomness of X.

29.2 Confounding

For most interesting policy questions, we cannot actually do a randomized trial. For instance, if I wanted to know if smoking caused lung cancer, there are ethical issues with trying to run a randomized trial. In this case, we have to use what is called *observational data*, i.e. we have information about many people who are smokers and not, and whether they have lung cancer or not.

It is clear that we can measure the correlation between smoking and lung cancer: the main question is when, if ever, can we claim a causal relationship?

The main problem is again selection bias.

Here is a motivating example: Suppose that our population has two kinds of people, those who are always healthy ($Y_i(1) = Y_i(0) = 1$) irrespective of whether they take the treatment or not, and those who are always unhealthy ($Y_i(1) = Y_i(0) = 0$) irrespective of whether they take the treatment or not. Suppose further that mostly healthy people take the treatment, while the unhealthy ones do not take the treatment.

The causal effect $\tau = 0$, but the estimator above would yield, $\widehat{\tau} \approx 1$, and we might incorrectly conclude that the treatment is beneficial.

Suppose however, that we knew who the healthy people were and who the unhealthy people were (we could gather such information by asking people questions about their lifestyle and other things). Then we could try to compare healthy people who took the treatment with healthy people who did not and similarly compare unhealthy people who took the treatment with unhealthy people who did not (and then try to combine these two estimates in some way). In this case, when we compared two healthy people who took the treatment and who did not we would see the treatment had no effect, and similarly for the unhealthy ones. We would likely correctly conclude that the treatment has no effect.

The key assumption that makes causal inference from observational data possible is the assumption of *no unmeasured confounding* or *selection on observables* or *ignorability*. Formally, we suppose that we have access to covariates X (think demographic information) such that,

$$W \perp\!\!\!\perp (Y(1), Y(0))|X.$$

This is an assumption. Roughly the assumption is is plausible in settings where we believe we can measure all of the covariates that explain the decision to take the treatment. We also need some other assumptions (SUTVA from the previous lecture, and the assumption that $\mathbb{P}(W = 1|X = x)$ is bounded away from 0 and 1, i.e. every individual has some non-zero chance of being either treated or in the control group) but we will ignore these and focus on no unmeasured confounding.

One way to think about this assumption, is that conditional on X we have a randomized trial, i.e. the treatment is independent of the potential outcomes. So if we condition on X (they are the confounders) we no longer have any selection bias. Alternatively, within levels of the covariate treatment is decided by (a biased) coin flip.

The other way to think about it is to think about a procedure called *matching*, i.e. suppose temporarily that X is discrete as it was in the healthy/unhealthy example. Then we could *match* people who had the same covariates but different treatments (i.e. some of them received treatment and some of them received control), and take the difference in their outcomes. For these matched people, under the assumption of no unmeasured confounding we are effectively observing *both potential outcomes*.

29.3 Identification under no unmeasured confounding

We want to estimate:

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

and to do this (as we did previously) we need to be able to write it in terms of observable quantities. Notice that by the law of total expectation,

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

Furthermore,

$$\tau = \mathbb{E}_X[\mathbb{E}[Y(1)|X, W = 1]] - \mathbb{E}_X[\mathbb{E}[Y(0)|X, W = 0]],$$

since we have that $W \perp\!\!\!\perp (Y(1), Y(0))|X$. This can we written as:

$$\tau = \mathbb{E}_X[\mathbb{E}[Y^{\text{obs}}|X, W = 1]] - \mathbb{E}_X[\mathbb{E}[Y^{\text{obs}}|X, W = 0]],$$

which is just a function of the observed data, and so we can turn our attention to estimating this quantity.

29.4 Estimation under no unmeasured confounding

The most direct way to estimate τ is to estimate:

$$\mu_0(x) = \mathbb{E}[Y^{\text{obs}}|X = x, W = 0]$$
$$\mu_1(x) = \mathbb{E}[Y^{\text{obs}}|X = x, W = 1].$$

These are two functions of the covariates X, one of them is the average outcome of the treatment group as a function of the covariates, and the other is the average outcome of the control group as a function of the covariates.

Estimating a conditional expectation is a problem is probably the most common problem in statistics – it is known as *regression*. We will delve into this formally in the next few lectures but for now let us suppose that someone hands us estimators $\widehat{\mu}_0$ and $\widehat{\mu}_1$ of these two functions.

Then we can compute the plug-in estimator:

$$\begin{aligned}\widehat{\tau} &= \widehat{\mathbb{E}}_X\left[\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right], \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\right]\end{aligned}$$

which is just the average of the difference between two regression functions. One approximately correct way to think about this is that we are using regression to impute the missing potential outcomes for each individual.

There are other ways to try to estimate τ. The other popular estimator is called the Horvitz-Thompson estimator or the inverse propensity score estimator. I will explain the basic idea here.

First let us define a quantity, known as the *propensity score*:

$$\pi(x) = \mathbb{P}(W = 1|X = x),$$

which represents the probability that a unit with covariates x receives treatment. Note that,

$$\begin{aligned}\mathbb{E}[W|X = x] &= \pi(x)\\ \mathbb{E}[1 - W|X = x] &= 1 - \pi(x).\end{aligned}$$

Returning to the average treatment effect,

$$\begin{aligned}\tau &= \mathbb{E}[Y(1) - Y(0)]\\ &= \mathbb{E}_X[\mathbb{E}[Y(1) - Y(0)|X = x]]\\ &= \mathbb{E}_X\left(\mathbb{E}_W\left[\mathbb{E}\left[\frac{Y(1)W}{\pi(x)}|X = x\right]\right] - \mathbb{E}_W\left[\mathbb{E}\left[\frac{Y(0)(1-W)}{1-\pi(x)}|X = x\right]\right]\right).\end{aligned}$$

This in turn can be written in terms of only observed quantities:

$$\begin{aligned}\tau &= \mathbb{E}_X\left(\mathbb{E}_W\left[\mathbb{E}\left[\frac{Y^{\text{obs}}W}{\pi(x)}|X = x\right]\right] - \mathbb{E}_W\left[\mathbb{E}\left[\frac{Y^{\text{obs}}(1-W)}{1-\pi(x)}|X = x\right]\right]\right)\\ &= \mathbb{E}\left[\frac{Y^{\text{obs}}W}{\pi(x)}\right] - \mathbb{E}\left[\frac{Y^{\text{obs}}(1-W)}{(1-\pi(x))}\right]\end{aligned}$$

which we can estimate as:

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i^{\text{obs}}W_i}{\pi(X_i)} - \frac{Y_i^{\text{obs}}(1-W_i)}{1-\pi(X_i)}\right].$$

This is called the Horvitz-Thompson estimator or the inverse propensity score estimator. In practice, we do not know the propensity score so we would estimate it from the data: this is again a problem of regression except the outcome is binary (i.e. since $\pi(X)$ is simply a conditional expectation we can estimate it by regressing W on X).

29.5 Advanced topics

This is just the tip of the iceberg. If you take a course in Causal Inference (next semester!) you will see many other neat things (here is just a small subset):

1. No unmeasured confounding is just one assumption that leads to identification of a causal effect. More broadly, in economics, political science and other fields people look for what are called natural experiments, i.e. roughly some subset of the population for which the assignment to treatment/control is nearly random.

2. Even in a randomized trial you might have something called non-compliance, i.e. some people don't do what they are told. In this case, you need to adjust your estimates. This is a canonical example of something called an instrumental variable problem.

3. There are many things beyond the average treatment effect that you might want to estimate. They all have different assumptions under which they are identified (i.e. can be written in terms of observable quantities) and there are different strategies to estimate them.

4. There is a very nice/simple way to combine the regression-based and propensity-score based estimators from above to construct what are called *doubly robust* estimators. These have the property that they are consistent if you can estimate either the regression function or the propensity score well (i.e. you do not need to estimate both well).

5. From a purely statistical standpoint, the average treatment effect under no unmeasured confounding:

$$\tau = \mathbb{E}_X[\mathbb{E}[Y^{\text{obs}}|X, W = 1]] - \mathbb{E}_X[\mathbb{E}[Y^{\text{obs}}|X, W = 0]],$$

   is an example of a functional. The regression method we used is basically the plug-in estimator of the functional. Plug-in estimators for functionals are usually not optimal. Optimally estimating functionals is a challenging statistical problem and there is an analog of the Cramer-Rao, efficiency, score functions, Fisher information etc. for functional estimation problems. This is known as semi-parametric efficiency theory.

6. There are many different languages for talking about causality and causal inference. We used potential outcomes. Many people (including many people in Philosophy here) use a language pioneered by Judea Pearl based on (causal) directed graphs. Most things can be translated from one language to another. People who work on causal graphs often focus effort on a much harder problem of *causal discovery*. We considered the case where there was a treatment and an outcome and we wanted to know if the treatment had a causal effect on the outcome. In causal discovery we have a collection of variables and want to discover causal relationships between them. This requires even stronger assumptions.

29.6 Non-parametric Regression

Broadly in regression we observe datapoints $\{(X_1, y_1), \ldots, (X_n, y_n)\}$ and our goal is to estimate the regression function

$$r(x) = \mathbb{E}[Y|X = x].$$

Unlike the CDF which we could estimate with no assumptions about the distribution, here we will need *smoothness* assumptions, i.e. we will need to assume that $r(x)$ is a smooth function of x. This allows us to gain statistical strength by averaging near by points.

Suppose we construct an estimate $\widehat{r}(x)$. Then a natural measure of how well we do is the squared loss, except since these are functions this is called the *integrated* squared loss, i.e.:

$$L(\widehat{r}, r) = \int (\widehat{r}(x) - r(x))^2 dx.$$

The risk is then just the expected loss, i.e.:

$$R(\widehat{r}, r) = \mathbb{E}\left(\int (\widehat{r}(x) - r(x))^2 dx\right).$$

As in the case of point estimation we have a bias variance decomposition. First we define the point-wise bias:

$$b(x) = \mathbb{E}(\widehat{r}(x)) - r(x),$$

and the point-wise variance:

$$v(x) = \mathbb{E}(\widehat{r}(x) - \mathbb{E}(\widehat{r}(x)))^2.$$

Now, as before we can verify that:

$$R(\widehat{r}, r) = \int b^2(x) dx + \int v(x) dx.$$

A natural strategy in non-parametric regression is to locally average the data, i.e. our estimate of the regression function at any point will be the average of the Y values in a small neighborhood of the point.

The width of this neighborhood will determine the bias and variance. Too large a neighborhood will result in high bias and low variance (this is called oversmoothing) and too small a neighborhood will result in low bias but large variance (this is known as undersmoothing).

29.7 Optimal Regression Function

Suppose we knew the joint distribution over (X, Y). One could alternatively begin by defining the risk of an estimate $\widehat{r}$ as

$$R(\widehat{r}) = \mathbb{E}(Y - \widehat{r}(X))^2.$$

This risk simply measures the prediction error, i.e. the expected error we make in predicting Y when we use the function $\widehat{r}(X)$. This risk is minimized by the conditional expectation, i.e. we have the following theorem.

Theorem 29.1 *The risk R is minimized by*

$$r(x) = \mathbb{E}(Y|X = x).$$

Proof: Let $g(x)$ be any function of x. Then

$$\begin{aligned}
R(g) &= \mathbb{E}(Y - g(X))^2 = \mathbb{E}(Y - r(X) + r(X) - g(X))^2 \\
&= \mathbb{E}(Y - r(X))^2 + \mathbb{E}(r(X) - g(X))^2 + 2\mathbb{E}((Y - r(X))(r(X) - g(X))) \\
&\geq \mathbb{E}(Y - r(X))^2 + 2\mathbb{E}((Y - r(X))(r(X) - g(X))) \\
&= \mathbb{E}(Y - r(X))^2 + 2\mathbb{E}\mathbb{E}\left((Y - r(X))(r(X) - g(X)) \,\middle|\, X \right) \\
&= \mathbb{E}(Y - r(X))^2 + 2\mathbb{E}\left((\mathbb{E}(Y|X) - r(X))(r(X) - g(X)) \right) \\
&= \mathbb{E}(Y - r(X))^2 + 2\mathbb{E}\left((r(X) - r(X))(r(X) - g(X)) \right) \\
&= \mathbb{E}(Y - r(X))^2 = R(r).
\end{aligned}$$

■

For what we will do in class it will not matter which definition of risk we use so we will use the one from the previous section in the next lecture.