LECTURE NOTE 9: CAUSALITY

1 Introduction

A great deal of social science and policy research takes interest in the "causal effect" of some policy, program, treatment, or experience. However, in this course, we have in large part maintained a conservative view of regression as linear projection (i.e., correlation or association). We have at times discussed whether our coefficients reflect "causal effects," but we have used this term rather loosely. This lecture note aims to clarify the meaning of causality.

2 Potential Outcomes

The most common approach to defining causality involves the concept of potential outcomes. This approach is called the *potential outcomes framework*, or the *Rubin Causal Model*, after its creator Don Rubin, currently a professor of statistics at Harvard.[1] The key idea is that each individual has her own set of potential outcomes, $Y_i(t)$, where t is a treatment level. In general, t can take on many values, but we will primarily focus on a binary treatment, so that each individual has two potential outcomes, $Y_i(1)$ and $Y_i(0)$. These potential outcomes reflect the outcome the individual would experience with and without the treatment, respectively. The *causal effect* of t for individual is $\alpha_i = Y_i(1) - Y_i(0)$. Note that causal effects may vary across individuals; i.e., they may be heterogeneous.

Unfortunately, we do not observe both $Y_i(1)$ and $Y_i(0)$ for each individual, so we cannot calculate the causal effect for each individual. To see this, let T_i denote the treatment level assigned to i. (In the binary case, $T_i = 0$ or 1.) For each individual, we observe only the realized outcome:

$$Y_i = Y_i(T_i) = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

We wish to make inferences about $\alpha_i = Y_i(1) - Y_i(0)$, but we only have data on (Y_i, T_i).

[1] The attribution of the potential outcomes framework to Don Rubin is a little unfair to Jerzy Neyman, who conceived of a very similar model in the 1920s.

3 Common Estimands of Interest

In making inferences about the distribution of α_i, we will focus on two averages:

- The *Average Treatment Effect (ATE)* for the whole population:

$$ATE = E[Y_i(1) - Y_i(0)] = E[\alpha_i]$$

- The average treatment effect on the individuals who were treated. This is known as the *mean effect of Treatment on the Treated (TOT)*:

$$TOT = E[Y_i(1) - Y_i(0)|T_i = 1] = E[\alpha_i|T_i = 1]$$

When the treatment effect is homogeneous in the population (so that $\alpha_i = \alpha$ for all i), the ATE and the TOT are the same. But when treatment effects are heterogeneous, the two estimands are different. For example, suppose we are estimating the effect of college education on earnings. If high-ability children are more likely to attend college but also gain more from attending, then the TOT will be larger than the ATE. If poor children have high returns from attending college but are unable to attend because of credit constraints, then the TOT will be smaller than the ATE. Note that these differences are not due to bias; they simply reflect averages of different distributions of treatment effects.

4 Selection Bias

We are now equipped to characterize selection bias. Suppose we observe treatment status and an outcome for a sample of individuals. Intuition might lead us to try estimating the effect of the treatment on the outcome by taking the difference in mean outcomes between the treated and untreated groups:

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$$

If we rewrite this expression using the underlying potential outcomes, we will notice a potential problem:

$$\begin{aligned}
E[Y_i|T_i = 1] - E[Y_i|T_i = 0] &= E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] \\
&= (E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0]) + (E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 1]) \\
&= \underbrace{(E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1])}_{TOT} + \underbrace{(E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0])}_{selection\ bias} \\
&= \underbrace{E[\alpha_i|T_i = 1]}_{TOT} + \underbrace{(E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0])}_{selection\ bias}
\end{aligned}$$

So the difference in the group-level means is equal to the TOT plus a selection bias term. The selection bias term reflects differences in the distributions of baseline outcomes ($Y_i(0)$) between individuals with $T_i = 0$ and $T_i = 1$. For the difference in means to have a causal interpretation, we must assume:

$$E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$$

which implies that baseline outcomes for the treatment and control groups have the same means. Under this condition, the difference in means is an unbiased estimator of the mean effect of treatment on the treated. Note that $E[Y_i(0)|T_i = 1]$ is unobservable, so the assumption above is fundamentally untestable.[2]

5 Unconfoundedness

The condition that $E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$ is somewhat abstract, so researchers usually make a more general assumption:

$$(Y_i(0), Y_i(1)) \perp T_i$$

where $\perp$ denotes independence. This assumption is called the *unconfoundedness assumption*. It states that potential outcomes are independent of treatment, and it implies that $E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$. A properly-implemented randomized trial (in which T_i is randomly assigned to members of the population) guarantees unconfoundedness.

Sometimes we wish to condition on covariates, as in a regression setting. The unconfoundedness assumption can easily accommodate covariates:

$$(Y_i(0), Y_i(1)) \perp T_i \big| X_i$$

This slightly-expanded condition states that, conditional on covariates, potential outcomes are independent of treatment. The expanded unconfoundedness assumption has several alternative names: *ignorable treatment assignment*, the *conditional independence assumption*, and *selection on observables.* As with the conditional mean assumption in Section 4, these assumptions are untestable.

6 Randomized Experiments

The method most closely identified with unconfoundedness is the randomized experiment. This section describes two approaches to randomization in policy research and discusses the estimation issues that arise

[2] Although the assumption is untestable, we can still shed light on it with "balance tests" for mean differences between the treatment and control groups in variables that should not have been affected by the treatment.

in each. Importantly, in both instances, we will assume that randomization does not change the pool of applicants or their behavior.

In one common type of policy experiment, the experimenter controls T_i directly. For example, if a social program is over-subscribed (so the number of applicants exceeds the capacity of the program), the program administrator might randomly choose which applicants to accept. Applicants have already expressed their interest in program participation, so if they are accepted, they will participate.[3] As a result, the difference in means between the treatment (accepted) and control (unaccepted) groups provides an unbiased estimator of the TOT.

Another common type of policy experiment is eligibility randomization. Here, the experimenter randomly chooses individuals in the population to be given eligibility to participate in the program. Some treatment group individuals will decide to participate; some will not. In consequence, the difference in means between the treatment and control groups measures the effect of *eligibility*, not *treatment*. The average effect of eligibility is known as the *intent-to-treat effec*t:

$$ITT = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$

where Z_i denotes program eligibility. Under eligibility randomization, we can still retrieve the TOT. We need three assumptions:

1. *Eligibility is randomly assigned.* This assumption guarantees identification of the ITT.
2. *The effect of group assignment on outcomes only operates through treatment.* This assumption (known as an *exclusion restriction*) is important because it allows us to exclude Z_i from the conditional expectation of Y_i:

   $$E[Y_i|Z_i, T_i] = E[Y_i|T_i]$$

   Conditional on treatment status, eligibility does not affect the conditional expectation of Y_i. More broadly, we might define a potential outcome $Y_i(z, t)$ that depends on both the eligibility level z and the treatment level t. The assumption says that we can exclude e from the potential outcome function without losing any information:

   $$Y_i(z, t) = Y_i(t)$$

   The potential outcomes assumption guarantees the conditional expectation assumption, but not vice versa.
3. *Ineligibles cannot participate in the program.* In other words, $Pr[T_i = 1|Z_i = 0] = 0$.

[3] If many accepted applicants end up not participating in the program, then the experiment will be more similar to eligibility randomization, described below.

Under assumption (3), we can separate individuals into two groups: compliers and never-takers. The compliers have $T_i = 1$ if $Z_i = 1$ and $T_i = 0$ if $Z_i = 0$. The never-takers have $T_i = 0$ regardless of the value of Z_i. In the eligible group, we can directly identify compliers and never-takers based on who opts into treatment. In the ineligible group, we cannot directly identify compliers and never-takers, but we know they exist. By asumption (1), the fraction of compliers (called the "compliance rate") is the same in the eligible and ineligible groups. Then we can write:

$$\begin{aligned} ITT \quad = \quad & (\text{mean effect of eligibility on outcomes of compliers})\,(\text{compliance rate}) \\ & + (\text{mean effect of eligibility on outcomes of never-takers})\,(1 - \text{compliance rate}) \end{aligned}$$

Note that the mean effect of eligibility on the on the outcomes of compliers (the first term in the first line) is the TOT. Furthermore, by assumption (2), the mean effect of eligibility on the on the outcomes of never-takers (the first term in the second line) is zero. Thus:

$$ITT = TOT \cdot Pr[T_i = 1|Z_i = 1] + 0 \cdot Pr[T_i = 0|Z_i = 1]$$

Rearrange to obtain:

$$TOT = \frac{ITT}{Pr[T_i = 1|Z_i = 1]}$$

So the TOT equals the ITT divided by the compliance rate. This result is intuitive. A nonzero ITT reflects a 'watered down' average effect of eligibility among compliers, where the 'watering down' is due to the presence of never-takers, who by assumption (2) could not be affected by eligibility. So to obtain the TOT, we merely need to rescale the ITT by the share of compliers.