Potential Outcomes

Francis J. DiTraglia

Econ 224

Introduction

People who have been to a hospital in the past month are more likely to die than people who haven't. But this doesn't mean that going to the hospital *causes* them to die. Instead it reflects a phenomenon that economists call *selection bias*. People who have been to the hospital are not representative of the population as a whole; they are *selected* for being sicker than average. The real causal story is that sicker people are more likely to go the hospital *and* more likely to die. Correlation is not causation, and it is very difficult to learn causal effects using observational rather than experimental data.

But why exactly is experimental data better? How do we know that it won't be subject to selection bias? To answer this question, we will introduce an idea called *potential outcomes*. Potential outcomes will be our tool for thinking about causality throughout the course, both in experimental and observational settings. This material is very important, but can be confusing when you first encounter it, so make sure to take the time to think carefully about this note and the associated reading assignment.

Basic Notation

When we talk about causes and effects in this course, we'll always use the same notation: Y will be an outcome of interest, and D will be a binary treatment. So D is the (potential) cause, and Y is the effect. When $D = 1$, this indicates that a person was treated; when $D = 0$, this indicates that a person was not treated. We'll use the subscript i to refer to an individual. So Y_i is the observed outcome for person i, and D_i indicates whether or not person i was treated.

Observed Outcomes versus Potential Outcomes

To motivate potential outcomes, let's think about a simple example involving my dog Luna. Every month I apply a special chemical to Luna's fur to prevent her from getting fleas or ticks. How do I know that this treatment is effective? In the past I've always given Luna the treatment and she's never gotten fleas or ticks, but does that mean that the treatment *caused* her not to get fleas or ticks? Maybe she wouldn't have gotten any even if I *hadn't* given her the treatment. Fundamentally, potential outcomes are about asking a what-if question just like this one: what would have happened if the treatment had been different?

Let D be the indicator for whether or not Luna was given the treatment: $D = 1$ means she got the treatment, and $D = 0$ means she didn't. The outcome in this case is also binary: let $Y = 0$ indicate that Luna didn't get any fleas or ticks and $Y = 1$ indicate that she did. On July 1st I gave Luna the treatment, $D = 1$, and she didn't get fleas or ticks that month, $Y = 0$. Notice that Y is the *actual outcome that I observed given the treatment that Luna actually received*. In other words, I *know* the value of Y, just like I know the value of D. What I don't know is what Y *would have been* if D had been zero. In other words I don't know Luna's *counterfactual outcome*. But just because I don't, and indeed *can't*, know Luna's counterfactual outcome, there's nothing to stop me from defining some notation to refer to it.

Luna's *potential* outcomes are defined by the pair (Y_0, Y_1). The first element Y_0 indicates whether Luna *would* have gotten fleas or ticks in July if she *hadn't* been treated. The second element Y_1 indicates whether Luna *would* have gotten fleas or ticks in July if she *had* been treated. Notice that the potential outcomes are defined without reference to Luna's actual treatment. As I will make clear in a moment, we can never

observe both Y_0 and Y_1 for the same individual, so you should consider them as a *thought experiment.* If, for example, $(Y_0, Y_1) = (1, 0)$, this means that Luna will get fleas or ticks unless she is treated. If on the other hand, $(Y_0 = 0, Y_1 = 0)$, this means that Luna will not get fleas or ticks, regardless of whether she is treated. The *difference* $Y_1 - Y_0$ of potential outcomes is the causal effect of treatment for Luna.

The potential outcomes (Y_0, Y_1) are linked to the observed outcome Y via the treatment assignment D. In particular, Luna's observed outcome Y equals Y_0 if she is not treated, and Y_1 if she is treated, or in mathematical notation:

$$Y = Y_0(1 - D) + Y_1 D.$$

There two mutually exclusive possibilities: either I treated Luna in July or I did not. This means that I can *never* observe both potential outcomes; I can only observe the one that corresponds to the *actual* treatment that Luna received. This also means that I can never observe Luna's causal effect $Y_1 - Y_0$.

Selection Bias and Random Assignment

So how can we do causal inference? One idea would be to treat Luna in July but not in August and then compare her outcomes in each month. But maybe the flea and tick treatment takes more than a month to wear off, in which case Luna is really "treated" in August as well in July. Even if this were not the case, there are other problems. For one, fleas and ticks respond to weather conditions, so the chance that Luna will be affected by them could change from month to month. If the only dog we study is Luna, there is no way to make an apples-to-apples comparison.

But suppose we had a whole *bunch* of Belgian Malinois – the breed of dog that Luna is. In particular suppose we observe (Y_i, D_i) for a random sample of n Belgian Malinois: for each dog, i, we know whether this dog was treated, D_i, and whether she got fleas or ticks, Y_i. If we took the sample average of Y_i among all Malinois with $D_i = 1$ this would give us an unbiased estimator of $E[Y_i|D_i = 1]$. If we took the sample average of Y_i among all dogs with $D_i = 0$ this would give us an unbiased estimator of $E[Y_i|D_i = 0]$. Remember that sampling variability decreases with sample size, so if n is very large, our estimates will be very precise; by gathering data on enough Malinois we can learn the values of $E[Y_i|D_i = 0]$ and $E[Y_i|D_i = 1]$ as precisely as we like. But what good is it to know these expectations? Do they provide any information about the effectiveness of flea and tick treatment for Belgian Malinois? The answer depends on how D_i was assigned.

Luna is a very pampered Malinois. She spends most of her time indoors, eats expensive dog food, gets regular visits to the vet, and is bathed monthly. Some of these factors make her much less likely to get fleas and ticks *regardless* of whether she gets the treatment. If owners who take very good care of their dogs in these ways are also more likely to give their dogs flea and tick treatment, a naive comparison of $E[Y_i|D_i = 1]$ to $E[Y_i|D_i = 0]$ will not give us the causal effect of treatment. If, on the other hand, D_i were *randomly assigned*, there would be no relationship between these other factors and treatment assignment. In this case, comparing $E[Y_i|D_i = 1]$ to $E[Y_i|D_i = 0]$ would allow us to draw a causal conclusion. To be more precise about this, we'll use the language of potential outcomes. But first we need to define the causal quantity that we hope to learn: the *average treatment effect.*

Define $\Delta_i = Y_{1i} - Y_{0i}$. This is the causal effect of treatment for individual i. The *average treatment effect* or ATE for short is the mean causal effect over all individuals in the population, namely $E[\Delta_i]$. There is no reason why the treatment effect Δ_i should be the same for everyone. If you are allergic to penicillin, the same drug that would save my life could threaten yours. When treatment effects can vary across individuals, we say that there are *heterogeneous treatment effects.* The ATE summarizes heterogeneous treatment effects by taking an average over everyone in the population. This tells us whether the treatment is beneficial *on average.* Returning to the example from above, a Malinois who lives in an area where there are very few ticks and fleas could have $(Y_{0i}, Y_{1i}) = (0, 0)$ so that $\Delta_i = 0$, while a Malinois who lives in a flea-and-tick-infested area could have $(Y_{1i}, Y_{0i}) = (1, 0)$ so that $\Delta_i = -1$. The treatment does nothing for the first Malinois, and prevents the second from getting fleas and ticks. In this example, the ATE give the *decrease* in the probability of getting fleas and ticks caused by the treatment. For this particular example, a negative ATE is good since we'd rather have Y equal to zero (no fleas or ticks) rather than one. Since we can never observe both

potential outcomes for the same individual, we can never observe Δ_i. This means that we cannot take the sample average of Δ_i to learn the ATE. Fortunately we don't have to observe Δ_i to calculate the ATE. Using the linearity of expectation,

$$\text{ATE} = E[\Delta_i] = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}].$$

If we observe Y_{1i} for a representative sample of individuals, we can estimate $E[Y_{1i}]$; if we observe Y_{0i} for a representative samples we can estimate $E[Y_{0i}]$. What the preceding expression tells us is that the difference of these two quantities is the *same thing* as the ATE. Go back and re-read this paragraph again to make sure you realize how amazing this is.

So what is the relationship between $E[Y_i|D_i = 0]$ and $E[Y_i|D = 1]$ on the one hand, and $E[Y_{0i}]$ and $E[Y_{1i}]$ on the other? In words, what is the relationship between the conditional means of the *observed* outcome Y_i and the mean of the *potential* outcomes Y_{0i} and Y_{1i}? Recall from above that the observed outcome Y_i for individual i is related to her potential outcomes (Y_{0i}, Y_{1i}) according to

$$Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i.$$

Taking conditional expectations, we obtain $E[Y_i|D_i = 0] = E[Y_{0i}|D_i = 0]$ and $E[Y_i|D_i = 1] = E[Y_{1i}|D_i = 1]$. Combining these,

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

If we add and subtract $E[Y_{0i}|D_i = 1]$ from the preceding expression, we obtain a very useful expression, namely

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{Observed Difference}} = \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Treatment on the Treated}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Selection Bias}}$$

If treatment assignment D_i is *independent* of the potential outcomes (Y_{0i}, Y_{1i}), then $E[Y_{1i}|D_i = 1] = E[Y_{1i}]$ and $E[Y_i|D_i = 0] = E[Y_{0i}]$ so $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$ equals the ATE. This is the case in an experimental dataset: since D_i is randomly assigned, it is definitely independent of the potential outcomes.

But in an observational dataset, potential outcomes and treatment assignment are often correlated: sicker people are more likely to go to the hospital, and dogs who are unlikely to get fleas or ticks because they have attentive owners are also more likely to be given flea-and-tick treatment.