Lecture Notes on Treatment Effects

(or *Completely Innocuous Econometrics*)

Francis J. DiTraglia

University of Oxford

This Version: 2022-09-06 22:16

Latest Version: `https://treatment-effects.com`

Abstract

These are lecture notes that I use to accompany my graduate teaching in Econometrics at Oxford. For more details, along with some short videos to accompany these notes, see the course website: treatment-effects.com. If you spot any typos, send me a Github pull request or an email: francis.ditraglia@economics.ox.ac.uk.

Contents

Chapter 1

Introduction

In this chapter we set the stage for the material to come, introducing the fundamental problem of causal inference, developing notation for later use, and reviewing some important facts concerning random variables.

1.1 What are these notes about?

Will earning an MPhil in Economics from Oxford increase your lifetime earnings? Does eating bacon sandwiches cause cancer? Does watching Fox News cause people to vote Republican? Will owning a dog increase your lifespan? Each of these questions concerns the **causal effect** of a **treatment** D on an **outcome** Y. The terminology "treatment" evokes a medical trial, but we will use the term much more broadly to refer to any variable D whose causal effect we hope to learn. For us, a treatment could be earning an MPhil, eating bacon sandwiches, watching Fox news, or owning a dog. These notes will focus on the case in which D is *binary*: either zero or one. If you have $D = 1$ we say that you are **treated**; if $D = 0$ we say that you are **untreated**. We will be particularly interested in methods for learning causal effects when the treatment variable is not randomly assigned, as would be the case in an **observational** rather than experimental study. So far as I know, no experiment has yet been carried out in which subjects are randomly compelled to be dog owners or forced to watch Fox News. Nonetheless papers have been written and published that attempt to estimate the causal effects of both of these treatments. We will study methods and assumptions under which observational data can be used to recover causal effects. We will also consider experiments in which subjects may fail to *comply* with their assigned treatments. In this case, the treatments that subjects actually receive are no longer randomly assigned, even if the treatments that they have been offered actually were.

1.2 The Fundamental Problem of Causal Inference

The fundamental problem of causal inference is that we can never observe a person's **counterfactual** outcome. In other words, we can never know what her outcome *would have been* if her treatment *had been different*. After finishing her undergraduate degree, Alice earned an MPhil in Economics at Oxford. She now makes £75,000 per year. Would she still have earned as much if she had gone straight to work after finishing her undergraduate degree? Barry was a vegetarian so he never ate bacon sandwiches. He lived to the ripe old age of 90 and died in a hang-gliding accident, never having developed cancer. If he had eaten bacon sandwiches every day, would he have died of cancer at the age of 60 instead? Donald watches Fox News 10 hours a day and always votes for the Republican candidate. If he hadn't watched Fox News, would he instead vote for the Democrats?

A counterfactual is a **within person** comparison: it asks how a given person's outcome would have been different if her treatment had been different. Because we can never observe the same person in two different treatment states, we can never actually make this comparison. You may be wondering about a before-and-after comparison. For example, what if we looked at Alice's wage immediately before she earned the MPhil and then immediately afterwards. Tracking the same person over time can be an extremely helpful way to untangle cause-and-effect, as we'll explore in a later chapter. It cannot, however, solve the fundamental problem of causal inference: comparing Alice's wages at two *different* points in time is not the same as comparing her wage at the *same* point in time across two "parallel universes," one in which she went straight to work and another in which she went to Oxford. Most people's income increases as they gain additional experience, for example. Comparing Alice's income before and after might confuse the effect of more experience in the labor force with the effect of earning an MPhil in Economics. Or perhaps Alice started the MPhil during an economic boom and finished during a severe downturn. If so, the fact that her income fell after the MPhil would tell us little of value: perhaps it would have fallen by *more* without the degree. Because the idealized within person comparison is impossible, we will need to develop methods and assumptions that allow us to substitute a **between-person** comparison.

1.3 The Potential Outcomes Framework

In order to study causal effects we need a framework that allows us to formally define them and manipulate them mathematically. Following the bulk of the treatment effects literature, we will adopt the **potential outcomes framework**, also know as the Neyman-Rubin Causal Model. With each person i we associate a pair of **potential outcomes** (y_{i0}, y_{i1}). These are precisely the counterfactual outcomes that I discussed in the preceding section. Suppose, for example, that Alice is person i. Then y_{i0} is her wage if

she doesn't earn the MPhil and y_{i1} is her wage if she does. Even though we can never observe both y_{i0} and y_{i1} for the same person, we can still *imagine* that there is a fact of the matter regarding what Alice's wage would have been in a parallel universe where her treatment had been different. Using this notation, $(y_{i1} - y_{i0})$ is the causal effect *for Alice* of earning the Oxford MPhil. This need not be the same as the causal effect for Bob of earning an Oxford MPhil, or indeed the same as the causal effect of anyone else. In other words, we will allow for the possibility that treatment effects are **heterogeneous**.

While we never observe both y_{i0} and y_{i1}, we always observe one of them. If Alice is treated then we observe y_{i1}; otherwise we observe y_{i0}. We can express this as follows

$$y_i = (1 - d_i)y_{i0} + d_i y_{i1} = y_{i0} + d_i(y_{i1} - y_{i0}) \tag{1.1}$$

where y_i is person i's **observed outcome** and d_i is an indicator that equals one if she was treated and zero otherwise. Implicit in this equation and the potential outcomes notation that we have adopted is a very important assumption that we will maintain throughout these notes: the **stable unit treatment value assumption** (SUTVA). This requires that Alice's outcome depends only on her own treatment and not the treatments of anyone else. SUTVA is a strong assumption and it is easy to think of settings where it doesn't hold. For example, if Alice gets a flu vaccine this makes Bob less likely to get the flu regardless of whether he was vaccinated. Finding ways to relax the SUTVA assumption is a challenging and active area of research that we'll explore in a later chapter.

1.4 Populations, Observables, and Random Variables

The first step of any causal analysis is to specify the **population of interest**. Suppose that we hope to learn the causal effect of watching Fox News on voting behavior. Whose voting behavior are we interested in? All US voters? Swing voters? Often the choice of population is dictated by circumstance. Perhaps we have access to a fantastic dataset on Pennsylvania voters but no information about voters from other states. If so, the causal claims we can make will necessarily be limited to Pennsylvania: the effect of Fox News could be markedly different, say, in Florida.

For the most part, these notes will assume that we have already specified a population of interest and observed a random sample from it. If our population is Pennsylvania voters, this assumes that we have observed a representative sample of n voters from the state. But what, precisely, do we observe? As discussed in the previous section, we can only observe *one* of a person's potential outcomes (y_{i0}, y_{i1}), namely the one that corresponds to her treatment d_i, as shown in (1.1). At a bare minimum, we will always assume that both y_i and d_i are observed for each person i in our sample. Most of the methods we describe below will in fact rely on observing some *additional* information $\boldsymbol{w}_i$.

For this reason, I will refer to $(y_i, d_i, \boldsymbol{w}_i)$ as the **observables** for person i.

Throughout this section and the preceding one I have used lowercase letters: y_i rather than Y_i and d_i rather than D_i, for example. I did this to emphasize that we are talking about specific values for a particular person. There is, in principle, nothing random about Alice's treatment, her observed outcome, or her potential outcomes. Randomness enters only when we view her as merely one member of a *population* from which we will draw a random sample. From this point onwards, we will stop thinking about the values for a particular person and instead think about random variables that represent the notion of *randomly drawing someone* from the population of interest.

The idea is as follows. Suppose that 35% of voters in Pennsylvania watch Fox News ($d_i = 1$). Then if I randomly sample a single voter, there is a 35% chance that she watches Fox News. We can represent this as a *random variable* D with a Bernoulli(0.35) distribution. Similarly, if we knew the values of y_i and $\boldsymbol{w}_i$ for every voter in Pennsylvania, we could construct random variables Y and $\boldsymbol{W}$ that represent the idea of randomly selecting a voter and observing her values of y_i and $\boldsymbol{w}_i$. Using this abstraction, we will view the observables $(y_i, d_i, \boldsymbol{w}_i)$ for any given person a *realization* from the joint distribution of a collection of random variables $(Y, D, \boldsymbol{W})$. The thought experiment is that we reach into the state of Pennsylvania, pull out a voter at random, and observe $(y_i, d_i, \boldsymbol{w}_i)$. Viewed in this way, knowing the values of $(y_i, d_i, \boldsymbol{w}_i)$ for everyone in the population is the same thing as knowing the joint distribution of $(Y, D, \boldsymbol{W})$.

Although we can never actually observe the pair (y_{i0}, y_{i1}) for the same person, we can still *imagine* reaching into the state of Pennsylvania and learning $(y_{i0}, y_{i1}, d_i, \boldsymbol{w}_i)$ for a particular person. As above, we can represent this idea using a collection of random variables: $(Y_0, Y_1, D, \boldsymbol{W})$. Knowing $(y_{i0}, y_{i1}, d_i, \boldsymbol{w}_i)$ for everyone in the population would be equivalent to knowing the joint distribution of $(Y_0, Y_1, D, \boldsymbol{W})$. Because these random variables are constructed from the values for each individual in the population, the relationship from (1.1) continues to apply, that is

$$Y = (1 - D)Y_0 + DY_1 = Y_0 + D(Y_1 - Y_0). \tag{1.2}$$

Equation 1.2 shows that knowledge of the joint distribution of $(Y_0, Y_1, D, \boldsymbol{W})$ implies knowledge of the joint distribution of $(Y, D, \boldsymbol{W})$, because Y is a function of (Y_0, Y_1, D). The converse, however, is false: knowledge of a person's observed outcome and her treatment does not allow us to reconstruct both of her potential outcomes.

1.5 Identification Versus Estimation

It's important to distinguish the problem of **identifying** causal effects from the problem of **estimating** them. Roughly speaking, identification is about finding the limits of what

we could ever possibly hope to learn from observable data, while estimation is about doing our best with the limited data we *actually* observe.

When studying identification we pretend that we could observe every individual in the population. We then ask whether this knowledge would provide be *enough* to answer a particular question. A bit more formally, suppose that we know the joint distribution of $(Y,D,\boldsymbol{W})$ and hope to learn the value of some quantity θ in our population of interest. As explained in the preceding section, knowing the distribution of $(Y,D,\boldsymbol{W})$ is the same as knowing the values of $(y_i,d_i,\boldsymbol{w}_i)$ for everyone in the population. If this knowledge is sufficient to uniquely pins down θ, then we say that θ is **identified**; otherwise we say that it is **unidentified**.[1] The challenge of identifying causal effects is that we observe not the joint distribution of potential outcomes (Y_0,Y_1) but only that of $(Y,D,\boldsymbol{W})$. Our identification question is whether this observed information, combined with appropriate assumptions, will allow us determine whether D causes Y.

Identification is about populations rather than samples. Estimation, on the other hand, asks how we can use a sample of observed data to produce a "best guess" of some quantity of interest θ. In the simplest case, we assume that the researcher observes a collection of n iid draws $(Y_i,D_i,\boldsymbol{W}_i)$ from the population and ask how this information can be used to construct an estimator $\widehat{\theta}$ of θ with desirable properties. These notes mainly focus on identification because estimation is meaningless without it: if there is no way to learn the causal effect of D on Y from knowledge of $(y_i,d_i,\boldsymbol{w}_i)$ for *everyone* in the population, there is no way to estimate it using a random sample from this population.

1.6 Why Study Average Treatment Effects?

When treatment effects are heterogeneous, every person in the population could have her own, unique causal effect: $(y_{i1}-y_{i0})$. Collecting the individual treatment effects for each person in our population of interest gives rise to a *distribution* of causal effects. Using the random variables defined above, we can represent this distribution using the random variable (Y_1-Y_0). If (Y_1-Y_0) were simply a constant, i.e. if treatment effects were **homogeneous**, asking whether D causes Y would be the same thing as asking if $(Y_1-Y_0)=0$. The sign and magnitude of (Y_1-Y_0) would then tell us the direction and importance of the effect. When treatment effects are heterogeneous, however, the yes-or-no question "does D cause Y?" no longer makes sense. Watching Fox News will not make Bernie Sanders vote Republican, but it might still affect the average swing voter in western Pennsylvania, for example. Faced with effects that vary across people, the natural

[1] Notice the use of the word *sufficient* in the definition of identification. Saying that θ is identified doesn't mean that knowing the joint distribution of $(Y,D,\boldsymbol{W})$ is *necessary* to uniquely pin down θ. For example, uniquely determining the vector of slope coefficients from a regression of Y on $(D,\boldsymbol{W})$ would only require us to know the means, covariances, and variances of these random variables.

question is "how do they vary?" In other words, what can we say about the *distribution* of $(Y_1 - Y_0)$? If we could learn the distribution of $(Y_1 - Y_0)$ across the population, we could answer a variety of interesting questions. For example: "what fraction of people benefit from this treatment?" or "what is the variance of treatment effects?"

Unfortunately it is impossible to learn the distribution of treatment effects. As we discussed above, the fundamental problem of causal inference is that we can never observe both y_{i1} and y_{i0} for the same person. For this reason, there is no way to identify the joint distribution of (Y_0, Y_1). If we want to determine the correlation between height and weight, we need observations of both variables for *the same people*. So too, identifying the joint distribution between (Y_0, Y_1) would require observations of both potential outcomes for the same people. Because we observe Y_0 for a subset of the population and Y_1 for another subset, we can learn the *marginal* distributions of Y_0 and Y_1. What we can never learn is the *dependence* between them.

This problem severely limits our ability to characterize the distribution of $(Y_1 - Y_0)$. Suppose, for example, that we wanted to determine $\text{Var}(Y_1 - Y_0)$. By the formula for the variance of a difference,

$$\text{Var}(Y_1 - Y_0) = \text{Var}(Y_0) + \text{Var}(Y_1) - 2\text{Cov}(Y_0, Y_1).$$

Because the variance of the treatment effects depends on a feature of the joint distribution of (Y_0, Y_1)—namely the covariance—this quantity cannot be identified. If we were willing to assume that Y_0 and Y_1 are uncorrelated, then we could indeed identify $\text{Var}(Y_1 - Y_0)$ based on knowledge of $\text{Var}(Y_0)$ and $\text{Var}(Y_1)$. In most examples, however, this assumption is untenable. Consider the problem of identifying the returns to an Oxford MPhil in Economics. More than likely, people who would earn a higher than average wage without the MPhil (high Y_0) would *also* earn a higher than average wage *with* an MPhil (high Y_1), implying a positive correlation between between Y_0 and Y_1.

It seems as though we have reached an impasse. How can we say anything useful about $(Y_1 - Y_0)$ without knowledge of the joint distribution of (Y_0, Y_1)? Recall a fundamental property of expectation: *linearity*. The expectation of a sum equals the sum of the expectations, and the expectation of a difference equals the difference of expectations. Thus, taking expectations of both sides

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0].$$

We call $\mathbb{E}[Y_1 - Y_0]$ the **average treatment effect** and abbreviate it ATE. The ATE measures how large the individual treatment effects $(y_{i1} - y_{i0})$ are *on average* across everyone in the population. If the ATE is positive, then the treatment is beneficial on average; if it is negative, then the treatment is harmful on average. If the ATE is zero,

then the treatment has no effect on average. The primary goal of the treatment effects literature is to identify the ATE or, failing that, at least an average treatment for some *subset* of the population. Undeniably the ATE is a valuable summary of $(Y_1 - Y_0)$, but it sweeps many important questions under the rug. What fraction of people would be *harmed* by the treatment? Is the treatment effect highly variable, or very similar for nearly everyone? These are important questions, but they are extremely difficult to answer. We study average treatment effects not because they are an ideal measure of the causal effect of D on Y, but because they give us a way around the fundamental problem of causal inference.

1.7 What about Quantile Treatment Effects?

If you have studied quantile regression, you may have encountered the term *quantile treatment effect*. How does this concept relate to our discussion of average treatment effects from above? Let Q_0 be the quantile function of Y_0 and Q_1 be the quantile function of Y_1. Then $Q_0(0.5)$ is the median of Y_0 while $Q_1(0.5)$ is the median of Y_1. Both of these quantities are identified from the marginal distributions of the potential outcomes. Indeed, for any quantile τ, both $Q_0(\tau)$ and $Q_1(\tau)$ are identified from these marginal distributions. The difference $\delta(\tau) \equiv Q_1(\tau) - Q_0(\tau)$ is called the **quantile treatment effect** of D on Y. Suppose that Alice's potential outcome without treatment y_{i0} falls at the τth quantile of the distribution of Y_0. In other words suppose that $\tau \times 100\%$ of people have a lower value of Y_0 than Alice, and $(1-\tau) \times 100\%$ have a higher value of Y_0. Then $\delta(\tau)$ tells us how much higher Alice's value of y_{i1} would *need to be* in order for her to fall at the τth quantile of the distribution of Y_1 as well.

It might be tempting to suppose that $\delta(0.5)$ represents the *median* of the distribution of treatment effects $(Y_1 - Y_0)$. For this to hold, however, requires an assumption called **rank invariance**. This stipulates that if Alice occupies the τth quantile of the Y_0 distribution, then she *also* occupies the τth quantile of the Y_1 distribution. In other words, if you lined everyone up based on their values of Y_0 they would be in the same order as if you had lined them up based on their values of Y_1: the treatment effect can have different effects on different people but doesn't change anyone's "position in line." Under rank invariance, $\delta(\tau)$ is the causal effect of D on Y for a person who *would have fallen* at the τth quantile of Y_0 had she not been treated. Rank invariance is a very strong assumption, and without it quantile treatment effects generally lack a clear causal interpretation.

1.8 The problem to overcome: selection bias

We know from above that $Y = (1-D)Y_0 + DY_1$. For a person who is treated we observe Y_1 and for a person who is not we observe Y_0. So to estimate ATE $\equiv \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$, why not simply compare the average value of Y among those with $D = 1$ to the average value of Y among those with $D = 0$? Because D is binary, this idea is *precisely* equivalent to running a regression of Y on D. To see this we use the following lemma.[2]

Lemma 1.1. *Let W be a binary random variable with $\mathbb{P}(W = 1) = p$. Then for any random variable X, we have $Cov(X, W) = p(1-p)\left[\mathbb{E}(X|W = 1) - \mathbb{E}(X|W = 0)\right]$ provided that the requisite expectations exist.*

Since D is binary, $\text{Var}(D) = \mathbb{P}(D = 1)\left[1 - \mathbb{P}(D = 1)\right]$. Thus, applying Lemma 1.1,

$$\beta_{OLS} \equiv \frac{\text{Cov}(D, Y)}{\text{Var}(D)} = \mathbb{E}(Y|D = 1) - \mathbb{E}(Y|D = 0). \tag{1.3}$$

Does β_{OLS} equal the ATE? To find out, we substitute (1.2) into (1.3) yielding

$$\begin{aligned}
\beta_{OLS} &= \mathbb{E}(Y|D = 1) - \mathbb{E}(Y|D = 0) \\
&= \mathbb{E}\left[(1-D)Y_0 + DY_1|D = 1\right] - \mathbb{E}\left[(1-D)Y_0 + DY_1|D = 0\right] \\
&= \mathbb{E}\left[Y_1|D = 1\right] - \mathbb{E}\left[Y_0|D = 0\right].
\end{aligned}$$

These manipulations show that β_{OLS} *may not* equal the ATE. The unconditional mean $\mathbb{E}(Y_1)$ need not equal the conditional mean $\mathbb{E}(Y_1|D = 1)$, and similarly $\mathbb{E}(Y_0)$ need not equal $\mathbb{E}(Y_0|D = 0)$, because D may be *related* to the potential outcomes. This problem is called called **selection bias**. To better understand it, consider the following example: let $D = 1$ if you graduated from university and let Y be your income at age 30. Adding and subtracting $\mathbb{E}(Y_0|D = 1)$ from the expression for β_{OLS}, we have

$$\beta_{OLS} = \underbrace{\mathbb{E}(Y_1 - Y_0|D = 1)}_{\text{TOT}} + \underbrace{\left[\mathbb{E}(Y_0|D = 1) - \mathbb{E}(Y_0|D = 0)\right]}_{\text{Selection Bias}}. \tag{1.4}$$

The first term in (1.4) is the average causal effect of the **treatment on the treated** abbreviated TOT. This measures causal effect of graduating from university on income averaged over all the people in the population who *chose* to graduate from university. When treatment effects are heterogeneous the TOT need not equal the ATE. Mark Zuckerberg famously dropped out of Harvard University in his sophomore year ($D = 0$) but is currently one of the highest earning people on the planet. Presumably his decision to leave university was motivated by a belief that his personal treatment effect $y_{i1} - y_{i0}$ was *negative*: the time he would have spent studying could be put to more lucrative use developing

[2] For a proof, see the appendix to this chapter.

Facebook. If people have some knowledge of their personal treatment effects and are to some extent free to choose their treatment, then we would expect $\mathbb{E}(Y_1 - Y_0|D = 1)$ to be *higher* than the ATE and $\mathbb{E}(Y_1 - Y_0|D = 0)$ to be *lower*.[3]

While the TOT does not in general equal the ATE, it is a meaningful and interesting causal quantity, answering the question "does this treatment help the people who choose to take it?" The second term in (1.4), on the other hand, totally destroys any hope of untangling cause-and-effect. This term, called **selection bias** by most authors, measures the difference in average values of Y_0 between the treated and the untreated. In the university and income example, it measures something like "difference of outside options" between those who ultimately chose to attend university and those who did not. If higher ability people are more likely to graduate from university ($D = 1$) and also have a higher-paying outside option Y_0, say because ability has a direct effect on income, the second term in (1.4) will be *positive*. Thus, even if the TOT *were* equal to the ATE, β_{OLS} will not in general identify the average causal effect of D on Y when individuals can choose their treatment status.

Once you start looking for it, you will find examples of selection bias *everywhere*. People who are admitted to hospitals are more likely to die in the next year than people who are not. This isn't because hospitals kill people: it's because sick people are more likely to go to hospitals. Dog owners are less likely to die over a five year horizon, but this may simply reflect the fact that healthy people are more likely to get a dog than sick people: taking care of an animal is a lot of work! Watching Fox News may cause you to vote Republican, or perhaps voting Republican causes you to watch Fox News.

1.9 Appendix: Proofs and Probability Review

The mathematical level of these notes is fairly modest. I assume throughout, however, that you are familiar with basic properties of random variables, expectation, variance, and covariance. In case you need to refresh your memory, this section lists some important properties that are used throughout the document.

Proof of Lemma 1.1. Let $p = \mathbb{P}(W = 1) = \mathbb{E}(W)$ and define $m_0 = \mathbb{E}(X|W = 0)$ and $m_1 = \mathbb{E}(X|W = 1)$ By the shortcut formula and iterated expectations,

$$\begin{aligned}\text{Cov}(X, W) &= \mathbb{E}(XW) - \mathbb{E}(X)\mathbb{E}(W) = \mathbb{E}\left[W\mathbb{E}(X|W)\right] - \mathbb{E}(X)p \\ &= \mathbb{E}(X|W = 1)p - \mathbb{E}(X)p = pm_1 - p\mathbb{E}(X)\end{aligned}$$

[3] By the Law of Iterated Expectations (Lemma 1.2), the ATE $\mathbb{E}(Y_1 - Y_0)$ is a convex combination of $\mathbb{E}(Y_1 - Y_0|D = 1)$ and $\mathbb{E}(Y_1 - Y_0|D = 0)$, so it necessarily lies *between* them.

Applying iterated expectations a second time,

$$\mathbb{E}(X) = \mathbb{E}\left[E(X|W)\right] = m_0(1-p) + pm_1$$

and substituting this equation into the expression for $\mathrm{Cov}(X, W)$,

$$\begin{aligned}\mathrm{Cov}(X, W) &= pm_1 - p\left[m_0(1-p) + pm_1\right] = (p - p^2)m_1 - p(1-p)m_0 \\ &= p(1-p)(m_1 - m_0) = p(1-p)\left[\mathbb{E}(X|W=1) - \mathbb{E}(X|W=0)\right]\end{aligned}$$

□

Lemma 1.2 (The Law of Iterated Expectations)**.**

$$\mathbb{E}[Y] = \mathbb{E}_X\left[\mathbb{E}(Y|X)\right], \quad \mathbb{E}[Y|Z] = \mathbb{E}_{X|Z}\left[\mathbb{E}(Y|X, Z)\right]$$

Lemma 1.3 (Taking out what is known)**.** *If f is a measurable function, then*

$$\mathbb{E}[f(X)Y|X] = f(X)\mathbb{E}[Y|X]$$

Lemma 1.4 (The Law of Total Probability)**.** *For discrete random variables X and Y*

$$\mathbb{P}(Y = y) = \sum_{all\ x} \mathbb{P}(Y = y|X = x)\mathbb{P}(X = x)$$

Lemma 1.5 (Linearity of Expectation)**.** *For RVs X, Y, Z and constants a, b, c*

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c, \quad \mathbb{E}[aX + bY + c|Z] = a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z] + c$$

Lemma 1.6 (Bayes' Theorem)**.**

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}, \quad \mathbb{P}(A|B, C) = \frac{\mathbb{P}(B|A, C)\mathbb{P}(A|C)}{\mathbb{P}(B|C)}$$

Definition 1.1 (Variance and Conditional Variance)**.**

$$\mathrm{Var}(X) \equiv \mathbb{E}\left[(X - \mathbb{E}\left\{X\right\})^2\right], \quad \mathrm{Var}(X|Z) \equiv \mathbb{E}\left[(X - \mathbb{E}\left\{X|\,Z\right\})^2\middle|\,Z\right]$$

Definition 1.2 (Covariance and Conditional Covariance)**.**

$$\begin{aligned}\mathrm{Cov}(X, Y) &\equiv \mathbb{E}\left[\left(X - \mathbb{E}\left\{X\right\}\right)\left(Y - \mathbb{E}\left\{Y\right\}\right)\right] \\ \mathrm{Cov}(X, Y|Z) &\equiv \mathbb{E}\left[\left(X - \mathbb{E}\left\{X|Z\right\}\right)\left(Y - \mathbb{E}\left\{Y|Z\right\}\right)|\,Z\right]\end{aligned}$$

Lemma 1.7 (Shortcut Rule for Variance and Covariance)**.**

$$\begin{aligned} Var(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ Var(X|Z) &= \mathbb{E}[X^2|Z] - \mathbb{E}[X|Z]^2 \\ Cov(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ Cov(X,Y|Z) &= \mathbb{E}[XY|Z] - \mathbb{E}[X|Z]\mathbb{E}[Y|Z] \end{aligned}$$

Lemma 1.8 (Properties of Variance and Covariance)**.**

(i) $Cov(X,X) = Var(X)$

(ii) $Var(aX+c) = a^2\,Var(X)$

(iii) $Var(aX+bY+c) = a^2\,Var(X) + b^2\,Var(Y) + 2ab\,Cov(X,Y)$

(iv) $Cov(aX+bY+c,Z) = a\,Cov(X,Z) + b\,Cov(Y,Z)$

Lemma 1.9 (Properties of Conditional Variance and Covariance)**.**

(i) $Var(X|X) = 0$

(ii) $Cov(X,Y|X) = 0$

(iii) $Cov(X,X|Z) = Var(X|Z)$

(iv) $Var(aX+c|Z) = a^2\,Var(X|Z)$

(v) $Var(aX+bY+c) = a^2\,Var(X) + b^2\,Var(Y) + 2ab\,Cov(X,Y)$

(vi) $Cov(aX+bY+c,Z|W) = a\,Cov(X,Z|W) + b\,Cov(Y,Z|W)$

Lemma 1.10 (The Law of Total Variance)**.**

$$Var(Y) = \mathbb{E}\left[\,Var(Y|X)\right] + \,Var\left(\mathbb{E}[Y|X]\right)$$

Lemma 1.11 (The Law of Total Covariance)**.**

$$Cov(X,Y) = \mathbb{E}\left[Cov(X,Y|Z)\right] + Cov\left[\mathbb{E}(X|Z),\mathbb{E}(Y|Z)\right]$$

Chapter 2

Conditional Independence

To understand the literature on treatment effects, you will need to develop some familiarity with the notion of **conditional independence** and its properties. This chapter provides an overview. We begin by defining independence and the closely related idea of conditional independence, and go on to explain the consequences that these notions have for *expectations*. This allows us to propose our first solution to the problem of selection bias: randomly assigning individuals to treatment.

The remainder of the chapter discusses a set of *axioms* that allow us to manipulate conditional independence relationships. Defining conditional independence and deriving its axioms for *all possible* kinds of random variables requires some measure theory. If have the appropriate background, I recommend reading the technical appendix, section 2.6, alongside the rest of the chapter. If you are not familiar with measure theory, don't worry: you will be able to understand everything except the technical appendix. There are only two terms from measure theory that I use in the body of the chapter. The first is that of a **measurable function**. If you haven't encountered this term before, it is just a particular way of saying that a function is "well-behaved." Any continuous function is measurable, as is any discontinuous function with a finite or countable number of discontinuities. The second is the terminology "W is Y-measurable." In words, this simply means that if we know the realization of the random variable Y then we also know the realization of the random variable W.

2.1 Intuition and Notation

Two continuous random variables X and Y are **independent** if and only if their joint density equals the product of their marginal densities: $f(x, y) = f(x)f(y)$ for all x, y in the support sets of X and Y.[1] By the definition of a conditional density, $f(y|x) = f(x, y)/f(x)$ so an *equivalent* definition of statistical independence is $f(y|x) = f(y)$ for

[1] For discrete RVs, replace densities with mass functions throughout, e.g. $p(x, y) = p(x)p(y)$.

all x, y in the support sets of X and Y. In other words, X and Y are independent if and only if knowing X provides *no additional information* about Y: the conditional density of Y given X is the same as the marginal density of Y. Of course we could just have easily reversed the roles of X and Y: an additional equivalent definition of conditional independence is $f(x|y) = f(x)$.

A closely related property is **conditional independence**. Two continuous random variables X and Y are conditionally independent given a third random variable Z if and only if $f(x, y|z) = f(x|z)f(y|z)$ for all x, y, z in the support sets of X, Y, Z. Using the definition of a conditional density, $f(y|x, z) = f(x, y|z)/f(x|z)$, this is equivalent to $f(y|x, z) = f(y|z)$. Reversing the roles of y and x, it is *also* equivalent to $f(x|y, z) = f(x|z)$.[2] If X and Y are conditionally independent given Z, this means that any dependence between X and Y comes solely from the fact that both are dependent on Z. In words: if we already know Z, then knowing X tells us nothing additional about Y, and vice-versa. We define conditional independence for continuous random *vectors* analogously: $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent given $\boldsymbol{Z}$ if $f(\mathbf{x}, \mathbf{y}|\mathbf{z}) = f(\mathbf{x}|\mathbf{z})f(\mathbf{y}|\mathbf{z})$, or equivalently if $f(\mathbf{y}|\mathbf{x}, \mathbf{z}) = f(\mathbf{y}|\mathbf{z})$ or $f(\mathbf{x}|\mathbf{y}, \mathbf{z}) = f(\mathbf{x}|\mathbf{z})$. For discrete random vectors, replace densities with mass functions.[3]

Independence, conditional and unconditional, is such an important concept in statistics and econometrics that it has its own symbol: "$\perp\!\!\!\perp$." If we write $X \perp\!\!\!\perp Y$ this means that X is independent of Y; if we write $X \perp\!\!\!\perp Y|Z$, this means that X is independent of Y, given Z. The same notation is used for random variables and random vectors.

2.2 Independence versus Mean Independence

Because our goal is to identify average treatment effects, we will be particularly interested in the consequences that conditional independence has for *means*.

Lemma 2.1. *Let X, Y, Z be random variables. If $X \perp\!\!\!\perp Y|Z$ then*

(i) $\mathbb{E}[XY|Z] = \mathbb{E}[X|Z]\mathbb{E}[Y|Z]$

(ii) $\mathbb{E}[Y|X, Z] = \mathbb{E}[Y|Z]$

(iii) $\mathbb{E}[X|Y, Z] = \mathbb{E}[X|Z]$.

Proof. The general case follows as a corollary of Proposition 2.1. Here we will assume that that X, Y, Z are continuous random variables. Results for discrete RVs follow by

[2] There are in fact many equivalent definitions of conditional independence. For full details see the Technical Appendix (section 2.6).

[3] For a fully general definition of conditional independence, see the Technical Appendix (section 2.6).

replacing integrals with sums. For (i), use $f(x, y|z) = f(x|z)f(y|z)$ and the definition of conditional expectation to write

$$\begin{aligned} E[XY|Z=z] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y|z)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x|z)f(y|z)\,dx\,dy \\ &= \int_{-\infty}^{\infty} yf(y|z)\left(\int_{-\infty}^{\infty} xf(x|z)dx\right)dy = \mathbb{E}[X|Z=z]\int_{-\infty}^{\infty} yf(y|z)dy \\ &= \mathbb{E}[X|Z=z]\mathbb{E}[Y|Z=z]. \end{aligned}$$

For (ii), use $f(y|x, z) = f(y|z)$ and the definition of conditional expectation to write

$$\mathbb{E}[Y|X=x, Z=z] = \int_{-\infty}^{\infty} yf(y|x,z)\,dy = \int_{-\infty}^{\infty} yf(y|z)\,dy = \mathbb{E}[Y|Z=z].$$

The argument for (iii) is nearly identical, combining $f(x|y, z) = f(x|z)$ with the definition of conditional expectation. □

Properties (ii) and (iii) of the lemma are often called **mean independence**. It is important to remember that conditional independence implies mean independence but *not the other way around*. Conditional independence is the stronger assumption. There is also a version of Lemma 2.1 that holds without conditioning on Z: $X \perp\!\!\!\perp Y$ implies that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, $\mathbb{E}[Y|X] = \mathbb{E}[Y]$, and $\mathbb{E}[X|Y] = \mathbb{E}[X]$. A good exercise would be to prove these implications for yourself if X and Y are continuous. Similar results also hold for random *vectors*: if $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$ then $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}] = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}] = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$. Moreover, if $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ then $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \mathbb{E}[\boldsymbol{Y}]$ and $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}]$.

2.3 Randomize treatments to eliminate selection bias.

Now that we know something about mean independence, we can propose our first solution to the problem of selection bias, as described in section 1.8 above. Suppose that, instead of arising naturally from the decisions people make, treatments were *randomly assigned* to people, independently of any of their characteristics. In this case, D would be *independent* of (Y_0, Y_1). By an argument nearly identical to that in Lemma 2.1 only without the "Z" this would imply that $\mathbb{E}(Y_0|D) = \mathbb{E}(Y_0)$ and $\mathbb{E}(Y_1|D) = \mathbb{E}(Y_1)$. Thus,

$$\begin{aligned} \beta_{OLS} &= \mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) \\ &= \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=0) \\ &= \mathbb{E}(Y_1 - Y_0) = \text{ATE} \end{aligned}$$

if $D \perp\!\!\!\perp (Y_0, Y_1)$. In words: there is no selection bias in a randomized experiment in which subjects are not free to choose their treatment.[4] Because randomized experiments are immune to selection bias, experimental studies are considered by many to be a "gold standard" against which other kinds of studies, such as those based on observational data, are to be judged. Valuable though they can be when applied carefully and interpreted correctly, however, randomized controlled trials are no panacea. For a recent critique, see Deaton and Cartwright (2018).

2.4 The Axioms of Conditional Independence

Now that we understand what conditional independence means, we have to learn how to work with it mathematically. Our approach will be *axiomatic*: we will state a number of abstract properties that the independence operator $\perp\!\!\!\perp$ satisfies and see how to use these to derive new properties. The result will be a kind of "algebra" of conditional independence: we will learn a number of rules with which we can manipulate a given conditional independence assumption to transform it into new conditional independence assumptions. All of the axioms of conditional independence can be rigorously proved from first principles: see the Technical Appendix for details (section 2.6). The names attached to axioms (i) and (iii)–(v) are taken from Pearl (1988). Axiom (ii) has not been given a name in the literature, so I have christened it the "redundancy" property. Note that when we write $W = h(Y)$ where h is a measurable function, this is equivalent to saying that W is Y-measurable: in other words, knowing the realization of Y tells us with certainty the realization of W.

Theorem 2.1 (Axioms of Conditional Independence)**.** *Let X, Y, Z, W be random variables defined on a common probability space, and let h be a measurable function. Then:*

(i) (Symmetry): $X \perp\!\!\!\perp Y|Z \implies Y \perp\!\!\!\perp X|Z$.

(ii) (Redundancy): $X \perp\!\!\!\perp Y|Y$.

(iii) (Decomposition): $X \perp\!\!\!\perp Y|Z$ *and* $W = h(Y) \implies X \perp\!\!\!\perp W|Z$.

(iv) (Weak Union): $X \perp\!\!\!\perp Y|Z$ *and* $W = h(Y) \implies X \perp\!\!\!\perp Y|(W, Z)$.

(v) (Contraction): $X \perp\!\!\!\perp Y|Z$ *and* $X \perp\!\!\!\perp W|(Y, Z) \implies X \perp\!\!\!\perp (Y, W)|Z$.

We begin with some important discussion of what these properties mean, how they can be used, and how they relate to properties used by other authors.

[4] This rules out settings in which some experimental subjects refuse to comply with the treatment they have been randomly assigned. We take up this more challenging case in a later chapter.

Random Variables vs. Vectors All of the results from above and the Technical Appendix, including Proposition 2.1 and Theorem 2.1, hold regardless of whether X, Y, Z, W are real-valued random variables, random vectors, or arbitrary collections of random variables and vectors. This is important, as it is typically necessary to find "clever" choices of X, Y, Z, W when applying the axioms of conditional independence. Often this requires defining one or more of these to be a *collection* of random variables, as we will see in many of the examples below.

Conditional vs. Unconditional Axioms Axioms (i) and (iii)–(v) are stated conditional on Z, but these same statements also hold *unconditionally* by dropping Z.[5] Because it is easier to put these unconditional versions of the axioms into words, I omit explicit conditioning on Z in some of the verbal explanations below.

Symmetry The symmetry property says that if learning Y does not give us any information about X, then learning X does not give us any information about Y. This is actually somewhat surprising, as the equality $\mathbb{E}\left(\mathbb{1}\left\{A_X\right\}|Y, Z\right) = \mathbb{E}\left(\mathbb{1}\left\{A_X\right\}|Z\right)$ does *not* treat X and Y symmetrically. Symmetry only becomes intuitively clear after establishing Proposition 2.1.

Redundancy The redundancy property says that if I already know Y, then learning Y *a second time* provides no additional information about X. Since $X \perp\!\!\!\perp Y|Y$ implies $Y \perp\!\!\!\perp X|Y$ by symmetry, another way of interpreting this condition is that, conditional on itself, a random variable Y is independent of *any other random variable*. In fact we can establish a more general result using similar reasoning, namely $X \perp\!\!\!\perp W|Y$ if W is Y-measurable. A proof of this fact using the axioms of conditional independence appears in the following section.

Decomposition The decomposition property says that if learning Y provides no information about X, then learning a *function* of Y likewise provides no information about X. If W is a measurable function of Y than it contains *at most* the same information content as Y. A common use of decomposition is to *drop* a random variable from a conditional independence statement. For example, suppose that $X_1 \perp\!\!\!\perp (X_2, X_3)|Z$. Since X_2 is (X_2, X_3)-measurable, it follows that $X_1 \perp\!\!\!\perp X_2|Z$. Analogously, $X_1 \perp\!\!\!\perp X_3|Z$. This *consequence* of the decomposition axiom is what some authors call "the decomposition property."

Weak Union The weak union property says that if learning Y provides no information about X, then learning Y after having *already learned* a function of Y likewise provides no

[5] Formally, this is equivalent to taking $\sigma(Z) = \emptyset$.

information about X. In effect, weak union allows us to *add* something to our conditioning set. A common application of this property is to move a random variable from the "left" of the conditioning bar to the "right." For example, suppose that $X_1 \perp\!\!\!\perp (X_2, X_3)|Z$. Since X_2 is (X_2, X_3)-measurable, weak union gives $X_1 \perp\!\!\!\perp (X_2, X_3)|(X_3, Z)$. It follows by decomposition that $X_1 \perp\!\!\!\perp X_2|(X_3, Z)$. Naturally, the same logic shows that $X_1 \perp\!\!\!\perp X_3|(X_2, Z)$. This *consequence* of the weak union and decomposition axioms is what some authors call the "weak union property."

Contraction The contraction property is a bit complicated to put into words. In effect, it allows us to move a random variable from the "right" of the conditioning bar to the "left". For example, suppose that $X_1 \perp\!\!\!\perp X_2|(X_3, X_4)$ and we want to show that $X_1 \perp\!\!\!\perp (X_2, X_3)|X_4$. If $X_1 \perp\!\!\!\perp X_3|X_4$, then contraction will give us our desired result.

2.5 More Properties of Conditional Independence

The axioms of conditional independence from Theorem 2.1 provide a simple but powerful way to deduce new conditional independence relationships from old ones.

Corollary 2.1. *$X \perp\!\!\!\perp Y|Z$ implies $(X, Z) \perp\!\!\!\perp Y|Z$.*

Proof of Corollary 2.1. By symmetry,

$$Y \perp\!\!\!\perp X|Z \tag{2.1}$$

and by redundancy,

$$Y \perp\!\!\!\perp (X, Z)|(X, Z). \tag{2.2}$$

Now, applying the decomposition property to (2.2)

$$Y \perp\!\!\!\perp Z|(X, Z) \tag{2.3}$$

and hence, applying the contraction property to (2.1) and (2.3), we obtain $Y \perp\!\!\!\perp (X, Z)|Z$. The result follows by symmetry. □

Another simple result that can be derived from the axioms of conditional probability is the following extension of the redundancy property. This does not appear in any references that I have seen, but it is easy to establish using the axioms of conditional independence.

Corollary 2.2. *Let $W = h(Y)$ where h is a measurable function. Then $X \perp\!\!\!\perp W|Y$.*

Proof of Corollary 2.2. By redundancy $X \perp\!\!\!\perp Y|Y$. By decomposition, taking Y to be "Z," this yields $X \perp\!\!\!\perp W|Y$. □

The well known-result that $X \perp\!\!\!\perp Y|Z$ implies $f(X) \perp\!\!\!\perp g(Y)|Z$ also follows directly from the axioms of conditional independence.

Corollary 2.3. *Let f and g be measurable functions. Then $X \perp\!\!\!\perp Y|Z \implies f(X) \perp\!\!\!\perp g(Y)|Z$.*

Proof of Corollary 2.3. By decomposition, $X \perp\!\!\!\perp g(Y)|Z$. Hence, by symmetry $g(Y) \perp\!\!\!\perp X|Z$. Applying decomposition a second time, $g(Y) \perp\!\!\!\perp f(X)|Z$. The result follows by a final application of symmetry. □

2.6 Appendix: Technical Details

Definition 2.1 (Conditional Independence)**.** Let X, Y, Z be random variables defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We say that X is conditionally independent of Y given Z (with respect to $\mathbb{P}$), written $X \perp\!\!\!\perp Y|Z$ if for all events $A_X \in \sigma(X)$ we have $\mathbb{E}\left(\mathbb{1}\left\{A_X\right\}|Y, Z\right) = \mathbb{E}\left(\mathbb{1}\left\{A_X\right\}|Z\right)$, $\mathbb{P}$-almost surely.

Proposition 2.1 (Equivalent Definitions of Conditional Independence)**.** *Let X, Y, Z be random variables defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then the following statements are equivalent:*

(i) $X \perp\!\!\!\perp Y|Z$

(ii) For all real, bounded, measurable functions f, $\mathbb{E}\left[f(X)|Y, Z\right] = \mathbb{E}\left[f(X)|Z\right]$

(iii) For all, real, bounded, measurable functions f, g, $\mathbb{E}\left[f(X)g(Y)|Z\right] = \mathbb{E}\left[f(X)|Z\right]\mathbb{E}\left[g(Y)|Z\right]$

(iv) For all $A_X \in \sigma(X)$ and all $A_Y \in \sigma(Y)$, $\mathbb{E}\left[\mathbb{1}\left\{A_X \cap A_Y\right\}|Z\right] = \mathbb{E}\left[\mathbb{1}\left\{A_X\right\}|Z\right]\mathbb{E}\left[\mathbb{1}\left\{A_Y\right\}|Z\right]$

where all equalities of conditional expectations are understood to hold $\mathbb{P}$-almost surely.

Proof of the Symmetry Property. The symmetry property follows immediately from the alternative definition of conditional independence given in Proposition 2.1 (iii). □

Proof of the Redundancy Property. Let f and g be real-valued, bounded, measurable functions. Since $g(Y)$ is Y-measurable,

$$\mathbb{E}\left[f(X)g(Y)|Y\right] = \mathbb{E}[f(X)|Y]g(Y) = \mathbb{E}\left[f(X)|Y\right]\mathbb{E}[g(Y)|Y]$$

so the result follows by Proposition 2.1 (iii). □

Proof of the Decomposition Property. Let f be a real-valued, bounded, measurable function. Since W is a measurable function of Y, we have $\sigma(W) \subseteq \sigma(Y)$ and consequently $\sigma(W, Z) \subseteq \sigma(Y, Z)$. Hence, by the *tower property* of conditional expectation,

$$\mathbb{E}\left[f(X)|W, Z\right] = \mathbb{E}\left\{\mathbb{E}\left[f(X)|Y, Z\right]|W, Z\right\}.$$

But since $X \perp\!\!\!\perp Y|Z$, Proposition 2.1 (ii) gives $\mathbb{E}\left[f(X)|Y,Z\right] = \mathbb{E}\left[f(X)|Z\right]$. And because $\mathbb{E}\left[f(X)|Z\right]$ is (W,Z)-measurable,

$$\mathbb{E}\left\{\mathbb{E}\left[f(X)|Z\right]|W,Z\right\} = \mathbb{E}\left[f(X)|Z\right]\mathbb{E}\left[1|W,Z\right] = \mathbb{E}\left[f(X)|Z\right].$$

Thus, $\mathbb{E}\left[f(X)|W,Z\right] = \mathbb{E}\left[f(X)|Z\right]$ so the result follows by Proposition 2.1 (ii). □

Proof of the Weak Union Property. Let f be a real-valued, bounded, measurable function. Since W is a measurable function of Y, we have $\sigma(W) \subseteq \sigma(Y)$. As a result, it follows that $\sigma(Y,W,Z) = \sigma(Y,Z)$ and hence $\mathbb{E}[f(X)|Y,W,Z] = \mathbb{E}[f(X)|Y,Z]$. Now, since $X \perp\!\!\!\perp Y|Z$, Proposition 2.1 (ii) gives $\mathbb{E}[f(X)|Y,Z] = \mathbb{E}[f(X)|Z]$. Finally, since $X \perp\!\!\!\perp Y|Z$ and W is Y-measurable, the decomposition property, Theorem 2.1 (iii), gives $X \perp\!\!\!\perp W|Z$ and hence $\mathbb{E}\left[f(X)|Z\right] = \mathbb{E}\left[f(X)|Z,W\right]$. Hence, the result follows by Proposition 2.1 (ii). □

Proof of the Contraction Property. Let f be a real, bounded, measurable function. Now, since $X \perp\!\!\!\perp W|(Y,Z)$ we have $\mathbb{E}[f(X)|Y,W,Z] = \mathbb{E}[f(X)|Y,Z]$ by Proposition 2.1 (ii). Similarly, since $X \perp\!\!\!\perp Y|Z$ we have $\mathbb{E}[f(X)|Y,Z] = \mathbb{E}[f(X)|Z]$. Combining these equalities gives $\mathbb{E}[f(X)|Y,W,Z] = \mathbb{E}[f(X)|Z]$ so the result follows by Proposition 2.1 (i). □

Chapter 3

Partial Identification

Sections 1.5, 1.6, and 1.7 of chapter 1 touched on identification and the limits of causal inference. Because we can never observe Y_0 and Y_1 for the same person, many interesting causal quantities are *fundamentally* unidentifiable. It's not a matter of getting more or better data, or thinking up new and more powerful statistical techniques; we're simply out of luck. But, at least as defined in section 1.5, identification is an extremely demanding concept: it requires us to be able to *uniquely* determine θ given knowledge of the distribution of $(Y, D, \boldsymbol{W})$. In many problems, however, there is an interesting middle ground. This chapter adds some nuance to our earlier discussion of identification and the limits of causal inference by introducing the concept of **partial identification**. The rough idea is as follows. Even if we can't pin it down *exactly*, we may still be able to rule out a wide range of values for θ. By ruling out enough values, we may even be able to answer our research question *without* determining the precise value of θ.

3.1 What is Partial Identification?

Here's a simple example that you may have seen before: linear regression with classical measurement error. Let α and β be the intercept and slope coefficients from a population linear regression of Y on X, so that $Y = \alpha + \beta X + U$ where

$$\beta \equiv \frac{\text{Cov}(X,Y)}{\text{Var}(X)}, \quad \alpha \equiv \mathbb{E}[Y] - \beta\mathbb{E}[X], \quad U \equiv Y - \alpha - \beta X, \quad \mathbb{E}(XU) = \mathbb{E}(U) = 0.$$

Asking whether a parameter is identified amounts to asking whether we could determine its exact value, given unlimited data. To be more precise, replace "given unlimited data" with "assuming that we know the joint distribution of any observed variables." This is worth emphasizing: *whether a parameter is identified depends on what we can observe.* For example, suppose that we could observe both X and Y. Then our identification exercise would begin by assuming that we know the joint distribution of (X, Y), call it

$f(x, y)$. This is enough information to uniquely determine α and β, since

$$\begin{aligned}
\mu_Y \equiv \mathbb{E}[Y] &= \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f(x, y)\, dx \right] dy \\
\mu_X \equiv \mathbb{E}[X] &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f(x, y)\, dy \right] dx \\
\text{Cov}(X, Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y)\, dx\, dy \\
\text{Var}(X) &= \int_{-\infty}^{\infty} (x - \mu_X)^2 \left[\int_{-\infty}^{\infty} f(x, y)\, dy \right] dx.
\end{aligned}$$

In the parlance of chapter 1, we would say that α and β are identified. In this chapter, however, we need to be a bit more precise. Because we can rule out all possible values for α and β except for a *single point*, we say that these parameters are **point identified**.

Now let $\widetilde{X} \equiv X + W$ where W is **classical measurement error**: $\text{Cov}(W, X) = 0$, $\text{Cov}(W, U) = 0$, and $\mathbb{E}(W) = 0$. What happens if we observe the noisy "proxy" variable $\widetilde{X}$ instead of the true regressor X? Can we uniquely determine α and β from the joint distribution of $(\widetilde{X}, Y)$? Call the associated density $f(\tilde{x}, y)$. As above, we can "integrate out" $\widetilde{X}$ to learn the marginal density $f(y)$ of Y, from which we can calculate $\mathbb{E}(Y)$. What about the remaining quantities? First, $\mathbb{E}[X] = \mathbb{E}[\widetilde{X} - W] = \mathbb{E}[\widetilde{X}]$ since $\mathbb{E}[W] = 0$. Similarly, since $\text{Cov}(W, U) = \text{Cov}(W, X) = 0$,

$$\begin{aligned}
\text{Cov}(\widetilde{X}, Y) &= \text{Cov}(X + W, Y) = \text{Cov}(X, Y) + \text{Cov}(W, Y) \\
&= \text{Cov}(X, Y) + \text{Cov}(W, \alpha + \beta X + U) \\
&= \text{Cov}(X, Y) + \text{Cov}(W, U) + \beta \text{Cov}(W, X) \\
&= \text{Cov}(X, Y)
\end{aligned}$$

so $\text{Cov}(X, Y)$, which we cannot directly compute, equals $\text{Cov}(\widetilde{X}, Y)$, which we can. The problem is $\text{Var}(X)$. Since X and W are uncorrelated,

$$\text{Var}(\widetilde{X}) = \text{Var}(X + W) = \text{Var}(X) + \text{Var}(W)$$

but because we do not know $\text{Var}(W)$, we cannot compute $\text{Var}(X)$. We know that it cannot be *greater* than $\text{Var}(\widetilde{X})$, but this still leaves us unable to calculate α and β since

$$\beta \equiv \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X}) - \text{Var}(W)}, \quad \alpha \equiv \mathbb{E}[Y] - \beta \mathbb{E}[X] = \mathbb{E}[Y] - \beta \mathbb{E}[\widetilde{X}].$$

Because we cannot pin down $\text{Var}(W)$, neither α nor β are point identified.[1] Nevertheless,

[1] If you're reading carefully, you may object that I haven't *proven* that we can't pin down $\text{Var}(W)$. I'll address this point in my discussion of "sharp" versus "tight" bounds below.

we can still say *something* about $\text{Var}(W)$ given what we can observe, and this will allow us to construct bounds for α and β. Since $\text{Cov}(X, Y) = \text{Cov}(\widetilde{X}, Y)$, we have

$$\frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})} = \frac{\text{Cov}(X, Y)}{\text{Var}(X) + \text{Var}(W)} = \frac{\text{Cov}(X, Y)/\text{Var}(X)}{1 + \text{Var}(W)/\text{Var}(X)} = \frac{\beta}{1 + \text{Var}(W)/\text{Var}(X)}.$$

But since $1 + \text{Var}(W)/\text{Var}(X) \geq 1$, it follows that $\text{Cov}(\widetilde{X}, Y)/\text{Var}(\widetilde{X})$ has the same *sign* as β and moreover

$$\left| \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})} \right| \leq |\beta|.$$

In other words, a regression of Y on the observed proxy $\widetilde{X}$ tells us the sign of β and provides a *lower bound* on its magnitude. To obtain an upper bound, we carry out the *reverse regression*: $\widetilde{X}$ on Y. This gives

$$\frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(Y)} = \frac{\text{Cov}(X, Y)}{\beta^2 \text{Var}(X) + \text{Var}(U)} = \frac{\beta \text{Var}(X)}{\beta^2 \text{Var}(X) + \text{Var}(U)}.$$

Taking the reciprocal of this expression,

$$\frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} = \beta + \frac{\text{Var}(U)}{\beta \text{Var}(X)} = \beta \left[1 + \frac{\text{Var}(U)}{\beta^2 \text{Var}(X)} \right].$$

Since the term in the square brackets is strictly *larger* than one, $\text{Var}(Y)/\text{Cov}(\widetilde{X}, Y)$ has the same sign as β and

$$\left| \frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} \right| \geq |\beta|.$$

Combining the two inequalities, we have shown that β lies *between* $\text{Cov}(\widetilde{X}, Y)/\text{Var}(\widetilde{X})$ and $\text{Var}(Y)/\text{Cov}(\widetilde{X}, Y)$. These are often called the **reverse regression bounds**. Since β is bounded and α depends only on β and the observed means $\mathbb{E}[Y]$ and $\mathbb{E}[\widetilde{X}]$, the regression intercept is likewise bounded.

The preceding is perhaps the simplest interesting example of **partial identification**: although we cannot pin down α and β exactly–neither is point identified–we can construct meaningful bounds for both. A parameter that is not point identified but can be bounded using what we observe is said to be **partially identified**.[2] We call the range of values that lie inside the bounds, i.e. the possible values for the parameter, the **identified set**. All else equal, we'd prefer our parameters to be point identified. So when presented with the consolation prize of partial identification, it's natural to wonder if our identified set is of any use. There are two related but distinct concepts lurking here: *sharpness* and *tightness*. We say that an identified set is **sharp** if it represents the *best possible* bounds

[2]Some authors prefer the term "set identified."

for our parameter of interest given what we can observe and what we have assumed. A sharp identified set exploits all available information. In contrast, an identified set that is *not sharp* leaves money on the table: we could improve our bounds by doing some extra work, without the need to observe anything extra or make any additional assumptions. Proving that a particular identified set is sharp can sometimes be very challenging. Even when a proof exists, actually *computing* the sharp identified set from a given dataset may be a Herculean task. For this reason, non-sharp bounds are fairly common. It turns out that our bounds for β in the classical measurement error problem *are in fact sharp*. They cannot be improved unless we observe something additional, like a second measure of X, or make an additional assumption, such as independence between X and (U, W) rather than uncorrelatedness. As an example of a non-sharp bound, consider the statement "β lies between zero and $\text{Var}(Y)/\text{Cov}(\widetilde{X}, Y)$." This is perfectly true, but we know that a better bound exists because we've calculated it!

In contrast to sharpness, which has a precise definition, *tightness* is to a certain extent in the eye of the beholder. We say that an identified set is **tight** when it is *small enough to be useful in practice*. An identified set could be sharp but not tight, tight but not sharp, both sharp and tight, or neither sharp nor tight. To get a better sense of what tightness could mean in real life, suppose we were interested in learning the sign of a parameter θ. An identified set that excludes zero could then be considered tight, as it allows us to answer our research question. Similarly, there might be an economic theory that implies $\phi = 1$. If we could construct an identified set for ϕ that excludes one, this would be enough to cast doubt on the theory. Although tightness is a desirable property, it is possible to have too much of a good thing. It is possible for the identified set to be *so tight* that it is *empty*! An **empty identified set** indicates that there are *no values* of the parameter θ that are compatible with both our assumptions and the joint distribution of what we can observe. For this reason, checking whether the identified set is empty can be used as a kind of *model specification test*.

Despite the inherent vagueness of the concept, we can say something more concrete about tightness in our measurement error example from above. Let r be the correlation between Y and $\widetilde{X}$. Then, we have

$$r^2 \equiv \frac{\text{Cov}(\widetilde{X}, Y)^2}{\text{Var}(\widetilde{X})\text{Var}(Y)} = \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})} \cdot \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(Y)}.$$

Re-arranging, it follows that

$$r^2 \cdot \frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} = \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})}.$$

This allows us to compute a simple expression for the width of the reverse regression

bound for β as follows:

$$\text{Width} = \left| \frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} - \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})} \right| = (1 - r^2) \left| \frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} \right|.$$

All else equal, the bound for β is *narrower*, and hence tighter, when the correlation between $\widetilde{X}$ and Y is large in absolute value. This makes intuitive sense. Since W is uncorrelated with U and X by assumption, any correlation between $\widetilde{X} = X + W$ and $Y = \alpha + \beta X + U$ must be driven by correlation between X and Y. If $\widetilde{X}$ and Y were perfectly correlated, this would imply that $\text{Var}(U) = \text{Var}(W) = 0$. The stronger the observed correlation between $\widetilde{X}$ and Y the less "room" is left for measurement error in either variable.

3.2 Bounding the ATE

Now that we've covered the basics of partial identification, it's time to apply what we've learned to the ostensible topic of these notes, *treatment effects*! This section considers a number of simple methods for bounding the average treatment effect (ATE) *without* assuming that the treatment is randomly assigned, following Manski (2003). In broad strokes, the idea is to ask what can be learned by making assumptions that are *weaker* than those typically used to obtain point identification.

We start with the simple observation that in many real-world applications the potential outcomes (Y_0, Y_1) are *themselves* bounded. To keep things simple, in this section we'll assume that both the treatment and the potential outcomes are *binary*. As a concrete example: does attending an Ivy League university ($D = 1$) *cause* you to continue your studies and earn a PhD ($Y = 1$)?[3] Or does it merely reflect the fact that academically-inclined students are more likely to apply to and be accepted at elite undergraduate institutions? As usual, the observed outcome Y equals Y_0 if $D = 0$ and Y_1 if $D = 1$, i.e.

$$Y = (1 - D)Y_0 + DY_1 = Y_0 + D(Y_1 - Y_0).$$

If (Y_0, Y_1) were *conditionally mean independent* of D–$\mathbb{E}[Y_0|D] = \mathbb{E}[Y_0]$ and $\mathbb{E}[Y_1|D] = \mathbb{E}[Y_1]$–the ATE would be point identified:

$$\mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0] = \mathbb{E}[Y_1|D = 1] - \mathbb{E}[Y_0|D = 0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \text{ATE}.$$

Notice that $\mathbb{E}[Y_1|D] = \mathbb{E}[Y_1]$ is equivalent to $\mathbb{E}[Y_1|D = 1] = \mathbb{E}[Y_1|D = 0]$. In the (Ivy League $\to$ PhD) example, this says that people who *do not* attend an Ivy League

[3] The "Ivy League" is a group of eight elite, east coast private Universities in the US: Brown, Columbia, Cornell, Dartmouth, Harvard, UPenn, Princeton, and Yale.

institution would go on to receive PhD degrees at the same rate as people who *do* attend an Ivy League institution, provided that we forcibly enrolled them at Harvard or Yale. Similarly, $\mathbb{E}[Y_0|D] = \mathbb{E}[Y_0]$ is equivalent to $\mathbb{E}[Y_0|D=0] = \mathbb{E}[Y_0|D=1]$. This says that people who *do* attend an Ivy League institution would go on to receive PhD degrees at the same rate as people who *do not* attend an Ivy League institution, provided that we forcibly *prevented* them from going to Harvard. Together, these conditions completely rule out selection into treatment. Given what we know about Ivy League admissions and what it takes to earn a PhD, this is wildly implausible. Our goal in this section is to say something useful about the average treatment effect $\mathbb{E}[Y_1 - Y_0]$ while *allowing* for self-selection into treatment.

Let's start off by asking what we know about the ATE *before* observing any data. Because expectation is linear, $\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$. And since the the outcome is binary, $0 \leq Y_0 \leq 1$ and $0 \leq Y_1 \leq 1$. Expectation preserves inequalities, so it follows that $0 \leq \mathbb{E}[Y_1] \leq 1$ and $0 \leq \mathbb{E}[Y_0] \leq 1$. Therefore $-1 \leq \text{ATE} \leq 1$. This isn't a terribly exciting bound, but you have to start somewhere! In the remainder of this section, we'll explore different ways of improving it.

First it will be helpful to introduce a bit of shorthand. Define

$$\begin{aligned} P_{11} &\equiv \mathbb{P}(Y=1|D=1) = \mathbb{E}[Y|D=1] = \mathbb{E}[Y_1|D=1] \\ P_{10} &\equiv \mathbb{P}(Y=1|D=0) = \mathbb{E}[Y|D=0] = \mathbb{E}[Y_0|D=0] \\ p &\equiv \mathbb{P}(D=1) = \mathbb{E}(D). \end{aligned}$$

We will assume throughout that the joint distribution of (Y, D) is observed. Since P_{11}, P_{10}, and p are all quantities that we can compute from this joint distribution, we can consider them "observed" as well.

3.2.1 Assumption-Free Bounds

Before observing the data, we know that $-1 \leq \text{ATE} \leq 1$. Now suppose that we observe the joint distribution of Y and D. If we are unwilling to make *any assumptions whatsoever* about the relationship between D and the underlying potential outcomes (Y_0, Y_1), is there any way to improve the bound? By iterated expectations,

$$\begin{aligned} \mathbb{E}[Y_1] &= \mathbb{E}_D\left[\mathbb{E}\left(Y_1|D\right)\right] = \mathbb{E}[Y_1|D=1]\mathbb{P}(D=1) + \mathbb{E}[Y_1|D=0]\mathbb{P}(D=0) \\ &= P_{11}p + \mathbb{E}[Y_1|D=0](1-p) \end{aligned}$$

Since p and P_{11} are observed, $\mathbb{E}[Y_1|D=0]$ is the only unknown in the equation for $\mathbb{E}[Y_1]$. But since Y_0 is binary, $0 \leq \mathbb{E}[Y_1|D=0] \leq 1$. Substituting gives a bound for $\mathbb{E}[Y_1]$,

$$pP_{11} \leq \mathbb{E}[Y_1] \leq pP_{11} + (1-p).$$

We can use the same idea to bound $\mathbb{E}[Y_0]$. First,

$$\begin{aligned}\mathbb{E}[Y_0] &= \mathbb{E}_D\left[\mathbb{E}(Y_0|D)\right] = \mathbb{E}[Y_0|D=1]\mathbb{P}(D=1) + \mathbb{E}[Y_0|D=0]\mathbb{P}(D=0) \\ &= \mathbb{E}[Y_0|D=1]p + P_{10}(1-p).\end{aligned}$$

And since $\mathbb{E}[Y_0|D=1]$ is bounded between 0 and 1,

$$(1-p)P_{10} \leq \mathbb{E}[Y_0] \leq p + (1-p)P_{10}$$

Combining the bounds for $\mathbb{E}[Y_1]$ and $\mathbb{E}[Y_0]$ gives

$$pP_{11} - (1-p)P_{10} - p \leq \text{ATE} \leq pP_{11} - (1-p)P_{10} + (1-p).$$

Defining the shorthand $q \equiv [pP_{11} - (1-p)P_{10} - p]$, we have shown that

$$q \leq \text{ATE} \leq (q+1), \quad q \equiv [pP_{11} - (1-p)P_{10} - p] \tag{3.1}$$

Notice that the bound from (3.1) always has width equal to one. This means that it is always *half as wide* as our earlier bound $-1 \leq \text{ATE} \leq 1$. The assumption-free bounds are big improvement, but since $q \leq 0$ and $q+1 \geq 0$, they always include zero.[4] If we hope to determine the *sign* of the ATE, we'll need to impose stronger restrictions.

3.2.2 Monotone Treatment Selection (MTS)

The bound from (3.1) was completely assumption-free. As such, we would expect that adding additional information in the form of extra assumptions should tighten our bounds for the ATE. We now consider one such assumption: **monotone treatment selection** or MTS for short. In the previous section we made no assumption whatsoever about the relationship between D and (Y_0, Y_1). As such, we allowed *arbitrary* self-selection into treatment. MTS still allows self-selection into treatment, but it assumes we know the *direction* of any selection effect that may be present. For *positive* selection into treatment, MTS posits that

$$\mathbb{E}[Y_1|D=0] \leq \mathbb{E}[Y_1|D=1], \quad \text{and} \quad \mathbb{E}[Y_0|D=0] \leq \mathbb{E}[Y_0|D=1].$$

For *negative* selection into treatment, we simply reverse the preceding inequalities. For concreteness, we will assume positive selection for the rest of the section. Recall from above that,

$$\mathbb{E}[Y_1] = pP_{11} + (1-p)\mathbb{E}[Y_1|D=0]$$

[4] Notice that $q = -p(1-P_{11}) - (1-p)P_{10} \leq 0$ while $q+1 = (1-p)(1-P_{10}) + pP_{11} \geq 0$.

by iterated expectations. Under positive MTS, $\mathbb{E}[Y_1|D=0] \leq \mathbb{E}[Y_1|D=1]$ and thus

$$\begin{aligned}\mathbb{E}[Y_1] &\leq pP_{11} + (1-p)\mathbb{E}[Y_1|D=1] \\ &= pP_{11} + (1-p)P_{11} = P_{11}.\end{aligned}$$

Similarly, recall from above that

$$\mathbb{E}[Y_0] = p\mathbb{E}[Y_0|D=1] + (1-p)P_{10}.$$

Since $\mathbb{E}[Y_0|D=0] \leq \mathbb{E}[Y_0|D=1]$ under positive MTS, it follows that

$$\begin{aligned}\mathbb{E}[Y_0] &\geq p\mathbb{E}[Y_0|D=0] + (1-p)P_{10} \\ &= pP_{10} + (1-p)P_{10} = P_{10}.\end{aligned}$$

We have thus shown that positive MTS implies $\mathbb{E}[Y_1] \leq P_{11}$ and $P_{10} \leq \mathbb{E}[Y_0]$. Combining the largest possible value of $\mathbb{E}[Y_1]$ with the smallest possible value of $\mathbb{E}[Y_0]$ gives us an new *upper bound* for the ATE, namely ATE $\leq P_{11} - P_{10}$. But is this actually a better bound that ATE $\leq (q+1)$ from (3.1)? The answer is yes. Rearranging, we see that

$$\begin{aligned}q+1 &= pP_{11} - (1-p)P_{10} - p + 1 \\ &= (1-p) + pP_{11} - P_{10} + pP_{10} \\ &= (1-p) + pP_{11} - P_{10} + pP_{10} + (P_{11} - P_{11}) \\ &= (P_{11} - P_{10}) + (1-p) - (1-p)P_{11} + pP_{10} \\ &= (P_{11} - P_{10}) + [(1-p)(1-P_{11}) + pP_{10}] \\ &\geq P_{11} - P_{10}\end{aligned}$$

since $(1-p)$, $(1-P_{11})$, p, and P_{10} are all probabilities, implying that $[(1-p)(1-P_{11}) + pP_{10}]$ cannot be negative. Therefore the upper bound ATE $\leq P_{11} - P_{10}$ obtained under positive MTS *cannot be less informative* than the no-assumptions upper bound ATE $\leq (q+1)$. Positive MTS does not affect the no-assumptions *lower bound* for the ATE. Therefore our final result is

$$q \leq \text{ATE} \leq P_{11} - P_{10}, \quad q \equiv [pP_{11} - (1-p)P_{10} - p]. \tag{3.2}$$

Since $P_{11} \equiv \mathbb{E}[Y_1|D=1]$ while $P_{10} \equiv \mathbb{E}[Y_0|D=0]$, notice that

$$P_{11} - P_{10} = \mathbb{E}[Y_1|D=1] - \mathbb{E}[Y_0|D=0] = \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0].$$

Under no selection into treatment, $\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]$ *equals* the ATE; under positive MTS it is the *upper bound* for the ATE in our binary outcome example.

3.2.3 Monotone Treatment Response (MTR)

Whereas MTS places an assumption of the direction of *selection*, **monotone treatment response**, or MTR for short, places an assumption on the direction of *causal effects*. Positive MTR imposes $Y_1 - Y_0 \geq 0$; for negative MTR, simply reverse the inequality.[5] Positive MTR says that *no one experiences a negative treatment effect*. Depending on the application, this assumption may be reasonable or it may be absurd. In a medical trial, it's likely that at least some people are *allergic* to a given medication. Even if it helps the vast majority of people, a treatment that harms *anyone* does not satisfy positive MTR. In the (Ivy League → PhD) example, the question becomes: are there any people who would *not* earn a PhD if they attended an Ivy League undergraduate institution but *would* earn a PhD if they did not? It is unclear whether positive MTR holds in this example. On the one hand you might argue that access to elite educational opportunities shouldn't hurt one's chance of continuing to doctoral studies. On the other hand, there may be students who would flourish in a small liberal arts college but feel intimidated and "lost in the crowd" at Harvard.

For the sake of argument, let's suppose that positive MTR *does hold*. We will show that this implies ATE ≥ 0. Since $Y_1 > Y_0$ for *everyone* under positive MTR, it is also true that $Y_1 > Y_0$ when we restrict attention to people with $D = 0$ or people with $D = 1$. Since expectation preserves inequalities, it follows that

$$\begin{aligned} 0 \leq \mathbb{E}[Y_0|D=1] \leq \mathbb{E}[Y_1|D=1] = \mathbb{E}[Y|D=1] = P_{11} \\ 1 \geq \mathbb{E}[Y_1|D=0] \geq \mathbb{E}[Y_0|D=0] = \mathbb{E}[Y|D=0] = P_{10}. \end{aligned}$$

Now, recall from above that

$$\begin{aligned} \mathbb{E}[Y_1] &= pP_{11} + (1-p)\mathbb{E}[Y_1|D=0] \\ \mathbb{E}[Y_0] &= p\mathbb{E}[Y_0|D=1] + (1-p)P_{10}. \end{aligned}$$

When using these equalities to construct the assumption-free ATE upper bound in (3.1), we replaced $\mathbb{E}[Y_1|D=0]$ with its largest possible value, one, and $\mathbb{E}[Y_0|D=1]$ with its smallest possible value, zero. Positive MTR doesn't provide any additional information that would help us tighten this part of the bound. When constructing our assumption-free ATE *lower bound*, on the other hand, we replaced $\mathbb{E}[Y_1|D=0]$ with its *smallest possible value*, zero, and we replaced $\mathbb{E}[Y_0|D=1]$ with its *largest possible value*, one. But under positive MTR,

$$\mathbb{E}[Y_1|D=0] \geq \mathbb{E}[Y_0|D=0] = P_{10} \geq 0$$

[5] If you're feeling pedantic, strictly speaking we only need $\mathbb{P}(Y_1 - Y_0 \leq 0) = 0$.

and similarly,

$$\mathbb{E}[Y_0|D=1] \leq \mathbb{E}[Y_1|D=1] = P_{11} \leq 1.$$

Substituting P_{10} for $\mathbb{E}[Y_1|D=0]$ in the expression for $\mathbb{E}[Y_1]$ and P_{11} for $\mathbb{E}[Y_0|D=1]$ in the expression for $\mathbb{E}[Y_0]$, we obtain

$$\text{ATE} \geq [pP_{11} + (1-p)P_{10}] - [pP_{11} + (1-p)P_{10}] = 0.$$

Therefore positive MTR ($Y_0 \leq Y_1$) implies $0 \leq$ ATE. Since positive MTR does not affect the no-assumptions *upper bound* for the ATE, our final result becomes

$$0 \leq \text{ATE} \leq q+1, \quad q \equiv [pP_{11} - (1-p)P_{10} - p]. \tag{3.3}$$

3.2.4 A Comparison of Bounds

Thus far we have derived three bounds: the "no-assumptions" bound from (3.1), the positive MTS bound from (3.2), and the positive MTR bound from (3.3). By imposing both positive MTS and positive MTR *simultaneously*, we obtain a fourth bound, namely

$$0 \leq \text{ATE} \leq (P_{11} - P_{10}), \quad q \equiv [pP_{11} - (1-p)P_{10} - p]. \tag{3.4}$$

While we haven't shown this explicitly, each of these bounds is *sharp* given the assumptions that it makes. Of course different assumptions lead to different bounds. In this section we'll compare the *tightness* of the various bounds in a simple numerical example. Suppose that 8% of Ivy League graduates go on to earn a PhD ($P_{11} = 0.08$) compared to 1.5% of the general public ($P_{10} = 0.015$). Suppose further that 0.2% of people attend an Ivy League institution ($p = 0.002$). Then,

$$\begin{aligned} q &\equiv pP_{11} - (1-p)P_{10} - p \\ &= 0.002 \times 0.08 - (1-0.002) \times 0.015 - 0.002 \approx -0.017 \end{aligned}$$

and $(q+1) \approx 0.983$. This gives the following results for our bounds from above:

$$\begin{aligned} \text{No Asumptions:} &\quad [q, q+1] \approx [-0.017, 0.983] \\ \text{Positive MTS:} &\quad [q, P_{11} - P_{10}] \approx [-0.017, 0.065] \\ \text{Positive MTR:} &\quad [0, q+1] \approx [0, 0.983] \\ \text{Positive MTS + MTR:} &\quad [0, P_{11} - P_{10}] = [0, 0.065]. \end{aligned}$$

If $P_{10} = 0.015$, $P_{11} = 0.08$ and $p = 0.002$, then without making any assumptions whatsoever we can be sure that the ATE, if negative, is not especially large in absolute value. In contrast, we can't say much about the upper bound for the ATE unless we are willing

to impose further restrictions. Since positive MTR simply replaces the lower bound of q with a lower bound of zero, it does little to help us in this example. In contrast, positive MTS drastically reduces our upper bound.

So are these bounds useful, or have I merely wasted your time for the past five pages? In particular, given that these bounds can never exclude zero, you might ask "what's the point?" In some applications, an upper bound of 0.065 for an ATE would be enough to tell us that the treatment isn't worth bothering about. In this case, the partial identification exercise would tell us all we need to know. More broadly, this exercise shows us just *how much* extra information random assignment of the treatment, selection-on-observables assumptions, or instrumental variables bring to the table. Without them, the causal information that we can extract from observational data is much more limited. It is also worth pointing out that the preceding bounds are far from the end of the story: they are really the simplest non-trivial example of applying partial identification to the study of causal inference. There are *many* other ways in which partial identification can shed light on causal questions. In the next section we'll explore a more interesting example.

3.3 Bounding the Distribution of Treatment Effects

A deadly and contagious disease is sweeping through the land of Erewhon. Fortunately a promising new treatment has just been subjected to a randomized, double-blind, placebo-controlled trial. But the treatment comes with its own risks. Before deciding whether to roll it out on a large scale, Erewhonian physicians want to know not only the ATE, but also the fraction of people it would *help* and the fraction it would *harm*. This is a tall order. As discussed in chapter 1, the fundamental problem of causal inference is that we can never observe both Y_0 and Y_1 for the same person. For this reason, a randomized controlled trial allows us to learn the *marginal* distributions of Y_0 and Y_1, but not their *joint* distribution. Features of the distribution of treatment effects that depend only on the marginals–notably the ATE–are point identified, while features that depend on the joint distribution–for example the variance of treatment effects–are not. But point identification isn't everything. In this section we'll ask what can be learned about the distribution of treatment effects from a partial identification perspective.

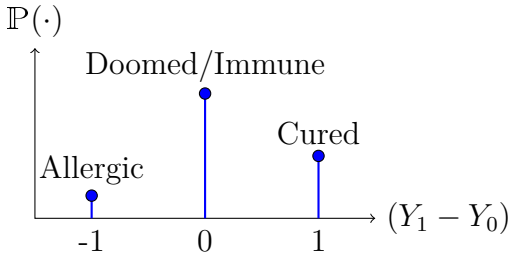
3.3.1 A Simple Example with Binary Outcomes

To keep things simple, we'll begin by considering a simple setting with a binary treatment and binary outcome. In the next section, we'll consider the general case. Continuing the disease example introduced above, let $Y = 1$ be *survive* and $Y = 0$ be *perish*. As usual $D = 1$ means treated and $D = 0$ means control. Since Y is binary, so are the potential outcomes. This means that there are four possible pairs (Y_0, Y_1). For later

reference, we will give each pair an evocative name. There are two groups of people who are unaffected by the treatment; their treatment effect is exactly zero. Those with $(Y_0 = 1, Y_1 = 1)$ will survive with or without the treatment: they are **immune**. Those with $(Y_0 = 0, Y_1 = 0)$, on the other hand, will perish regardless: they are **doomed**. The remaining two groups are affected by treatment. Those with $(Y_0 = 0, Y_1 = 1)$ would die without treatment, but survive with treatment: they are **cured** by the treatment. Finally, those with $(Y_0 = 1, Y_1 = 0)$ would survive without the treatment, but die if given the treatment: they are **allergic** to the treatment.

		Y_1	
		0	1
Y_0	0	$\mathbb{P}$(Doomed)	$\mathbb{P}$(Cured)
	1	$\mathbb{P}$(Allergic)	$\mathbb{P}$(Immune)

(a) Joint Distribution of Potential Outcomes

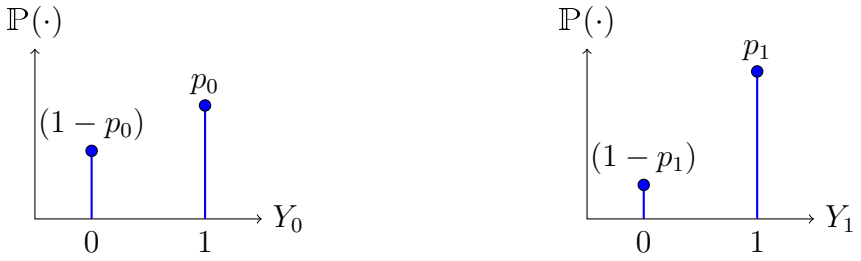


(b) Distribution of Treatment Effects

Figure 3.1: Unobserved Quantities. Neither the joint distribution of potential outcomes (left) nor the distribution of treatment effects (right) is point identified.

Figure 3.1 depicts the joint distribution of potential outcomes, (3.1a) along with the corresponding distribution of treatment effects (3.1b) in this example. The fraction of people with a treatment effect of -1 is $\mathbb{P}$(Allergic), the fraction with a treatment effect of 1 is $\mathbb{P}$(Cured), and the fraction with a treatment effect of zero is $\mathbb{P}$(Doomed) plus $\mathbb{P}$(Immune). The Erewhonian physicians, introduced above, want to compare $\mathbb{P}$(Cured) to $\mathbb{P}$(Allergic) to decided whether the cure is worse than the disease. Unfortunately, neither of these quantities is observed, since they both depend on the joint distribution of potential outcomes. Figure 3.2 depicts the quantities that *are* observed: the marginal distribution of Y_0 (3.2a) and the marginal distribution of Y_1 (3.2b). We know these marginal distributions because the treatment was assigned via a randomized, double-blind, placebo-controlled experiment: there is no self-selection into treatment, and there are no placebo effects. And because the outcome is binary, survive or perish, each of the marginals boils down to a single number: $p_0 \equiv \mathbb{P}(Y_0 = 1)$ for Y_0 and $p_1 \equiv \mathbb{P}(Y_1 = 1)$.

In short, then, our partial identification exercise amounts to asking what we can say about $\mathbb{P}$(Cured) versus $\mathbb{P}$(Allergic) having observed p_0 and p_1. The trick is to relate the joint distribution, which we can't observe, to the marginals, which we can. From



(a) Marginal Distribution of Y_0 **(b)** Marginal Distribution of Y_1

Figure 3.2: Observed Quantities. In a randomized controlled trial, the marginal distributions of the potential outcomes are point identified.

Figure 3.1a, we see that

$$\begin{aligned} 1 - p_0 &= \mathbb{P}(Y_0 = 0) = \mathbb{P}(\text{Doomed}) + \mathbb{P}(\text{Cured}) \\ p_0 &= \mathbb{P}(Y_0 = 1) = \mathbb{P}(\text{Allergic}) + \mathbb{P}(\text{Immune}) \\ 1 - p_1 &= \mathbb{P}(Y_1 = 0) = \mathbb{P}(\text{Doomed}) + \mathbb{P}(\text{Allergic}) \\ p_1 &= \mathbb{P}(Y_1 = 1) = \mathbb{P}(\text{Cured}) + \mathbb{P}(\text{Immune}) \end{aligned}$$

Now, introduce the shorthand $\alpha \equiv \mathbb{P}(\text{Allergic})$. Recall that this equals the fraction of people who would be *harmed* by the treatment. Re-arranging the expressions for p_0 and $1 - p_1$, it follows that

$$\mathbb{P}(\text{Immune}) = p_0 - \alpha, \quad \mathbb{P}(\text{Doomed}) = 1 - p_1 - \alpha.$$

And since the fraction of people who are Cured necessarily equals the fraction of people who are *not* Immune, Doomed, or Allergic,

$$\mathbb{P}(\text{Cured}) = 1 - (p_0 - \alpha) - (1 - p_1 - \alpha) - \alpha = (p_1 - p_0) + \alpha.$$

We see that the fraction of people who are *helped* (Cured) by the treatment equals the ATE plus the fraction who are *harmed* (Allergic). This may seem strange, so here's another way to look at it. Because the treatment was randomly assigned, the fraction of each "type" of person is the same on average in both the treatment groups. In the control group, p_0 survive: the Allergic and the Immune. In the treatment group, p_1 survive: the Cured and the Immune. Taking the difference $p_1 - p_0$, cancels out the contribution of the Immune to p_1 and p_0. What remains is a positive contribution from $\mathbb{P}(\text{Cured})$ and a negative contribution from $\mathbb{P}(\text{Allergic})$. Thus $\text{ATE} = (p_1 - p_0) = \mathbb{P}(\text{Cured}) - \mathbb{P}(\text{Allergic})$. Because Y is binary, in this example a positive average effect necessarily implies that the

treatment **helps more people than it harms**. Collecting what we've learned so far:

$$\mathbb{P}(\text{Allergic}) \equiv \alpha \qquad \mathbb{P}(\text{Cured}) = (p_1 - p_0) + \alpha$$
$$\mathbb{P}(\text{Doomed}) = 1 - p_1 - \alpha \qquad \mathbb{P}(\text{Immune}) = p_0 - \alpha.$$

Since we learn p_0 and p_1 from carrying out the randomized controlled trial, this shows that we would know the fraction of all four "types," and hence the full distribution of treatment effects, if only we knew α. While we cannot point identify α, however, we can *bound* it. The idea is as follows. If you tell me a potential value $\alpha^* \in [0, 1]$ for the share of people harmed by the treatment, I can immediately work out the implied joint probabilities from Figure 3.1a using the preceding four equalities. These equalities automatically ensure that the joint distribution implied by your choice α^* sums to one and is compatible with the observed marginal distributions, p_0 and p_1. They *do not*, however, ensure that each of the joint probabilities lies between zero and one. Any proposed value of α^* that yields a value outside of $[0, 1]$ for $\mathbb{P}(\text{Cured})$, $\mathbb{P}(\text{Doomed})$ or $\mathbb{P}(\text{Immune})$ can be *ruled out* as incompatible with what we observe. By collecting all values $\alpha^* \in [0, 1]$ that are *not ruled out*, we end up with the *sharp identified set* for α. By substituting the endpoints of this set into our expressions for $\mathbb{P}(\text{Cured})$ etc., we obtain the sharp identified set for these quantities as well.

We now construct the identified set for α. Imposing the constraints $0 \leq \mathbb{P}(\cdot) \leq 1$ for the Cured, Doomed, and Immune yields the following set of inequalities:

$$0 \leq (p_1 - p_0) + \alpha \leq 1, \quad 0 \leq 1 - p_1 - \alpha \leq 1, \quad 0 \leq p_0 - \alpha \leq 1.$$

Re-arranging, and using the fact that $\text{ATE} = p_1 - p_0$ in this example, we obtain

$$-\text{ATE} \leq \alpha \leq 1 - \text{ATE}, \quad -p_1 \leq \alpha \leq 1 - p_1, \quad p_0 - 1 \leq \alpha \leq p_0.$$

But since α is a probability, and both $-p_1$ and $p_0 - 1$ are always negative, we can replace these lower bounds with zero, yielding

$$-\text{ATE} \leq \alpha \leq 1 - \text{ATE}, \quad 0 \leq \alpha \leq 1 - p_1, \quad 0 \leq \alpha \leq p_0.$$

But since we have

$$1 - \text{ATE} = 1 - (p_1 - p_0) = (1 - p_1) + p_0 \geq (1 - p_1)$$

and since the upper bounds $\alpha \leq (1 - \text{ATE})$, $\alpha \leq 1 - p_1$ and $\alpha \leq p_0$, must hold *simulta-*

neously, we can simplify these inequalities to

$$\max\{-\text{ATE}, 0\} \leq \alpha \leq \min\{p_0, 1 - p_1\}, \quad \text{ATE} = (p_1 - p_0). \tag{3.5}$$

From (3.5), we see that there are two cases. When the average treatment effect is negative, we can be *sure* that some people are harmed by the treatment: the sharp identified set for α excludes zero. In contrast, when the average treatment effect is positive we cannot exclude $\alpha = 0$. In either case, the upper bound for α is $\min\{p_0, 1 - p_1\}$.

Now we have all the information we need to bound the distribution of treatment effects from 3.1b. Define the shorthand $\underline{\alpha} \equiv \max\{-\text{ATE}, 0\}$ and $\overline{\alpha} \equiv \min\{p_0, 1 - p_1\}$ so the bound for α becomes $\underline{\alpha} \leq \alpha \leq \overline{\alpha}$. By definition $\mathbb{P}(Y_1 - Y_0 = -1) \equiv \alpha$ and is hence bounded between $\underline{\alpha}$ and $\overline{\alpha}$. Since the fraction of people with a treatment effect of zero equals the fraction of Doomed plus the fraction of Immune,

$$\begin{aligned}\mathbb{P}(Y_1 - Y_0 = 0) &= \mathbb{P}(\text{Doomed}) + \mathbb{P}(\text{Immune}) = (1 - p_1 - \alpha) + (p_0 - \alpha) \\ &= (1 - p_1) + p_0 - 2\alpha.\end{aligned}$$

Since α enters this expression with a minus sign in front of it, we obtain

$$(1 - p_1) + p_0 - 2\overline{\alpha} \leq \mathbb{P}(Y_1 - Y_0 = 0) \leq (1 - p_1) + p_0 - 2\underline{\alpha}.$$

Finally, since the fraction of people with a treatment effect of one equals the fraction of Cured, we have

$$\mathbb{P}(Y_1 - Y_0 = 1) = \mathbb{P}(\text{Cured}) = \text{ATE} + \alpha$$

and thus, we obtain the bounds

$$\text{ATE} + \underline{\alpha} \leq \mathbb{P}(Y_1 - Y_0 = 1) \leq \text{ATE} + \overline{\alpha}.$$

An interactive webapp posted at the following url illustrates how the bounds we have derived for the distribution of $(Y_1 - Y_0)$ vary with p_0 and p_1:

https://fditraglia.shinyapps.io/binary-treatment-effect-bounds/.

It is important to stress that these bounds are only sharp **pointwise**. In other words, for any possible value k of the treatment effect $(Y_1 - Y_0)$, we have constructed the best possible bound for $\mathbb{P}(Y_1 - Y_0 = k)$ *in isolation*. If you are interested in bounding a quantity that depends on *more than one* of the probabilities in Figure 3.1b, it is possible to do better than simply combining the pointwise bounds. For example, suppose you wanted to bound $\mathbb{P}(\text{Cured}) - \mathbb{P}(\text{Allergic})$. If $p_1 = 0.7$ and $p_0 = 0.5$ then the ATE is $0.7 - 0.5 = 0.2$ and we know from above that this equals $\mathbb{P}(\text{Cured}) - \mathbb{P}(\text{Allergic})$. The

difference in proportions of Cured versus Allergic is *point identified* in this example. Since

$$\underline{\alpha} = \{-\text{ATE}, 0\} = \{-0.2, 0\} = 0, \quad \overline{\alpha} = \min\{p_0, 1 - p_1\} = \min\{0.5, 0.3\} = 0.3,$$

it follows that $\mathbb{P}(\text{Allergic}) \in [0, 0.3]$ while $\mathbb{P}(\text{Cured}) \in [0.2, 0.3]$. We see that the lower bound for $\mathbb{P}(\text{Cured})$ lies *below* the upper bound for $\mathbb{P}(\text{Allergic})$. But didn't we just say that $\mathbb{P}(\text{Cured}) - \mathbb{P}(\text{Allergic})$ is point identified and in fact positive? While this may look like a contradiction, it isn't: the lower bound for $\mathbb{P}(\text{Cured})$ corresponds to $\alpha = \underline{\alpha}$ while the upper bound for $\mathbb{P}(\text{Allergic})$ corresponds to $\alpha = \overline{\alpha}$. In other words, the seeming contradiction comes from evaluating the bounds at two *different* values of α. To form a **joint** rather than pointwise bound, we would need to use the same value of α for both. Doing do would give us precisely the ATE.

3.3.2 The General Case: Fan & Park (2010)

In the previous section we constructed (pointwise) sharp bounds for the distribution of treatment effects when the outcome was binary. To accomplish this, we asked which joint distributions for (Y_0, Y_1) were *ruled out* by marginal distributions of Y_0 and Y_1 that we observed. We then examined all of the joint distributions that were *not ruled out* to find the largest and smallest possible values of $\mathbb{P}(Y_1 - Y_0 = -1)$, $\mathbb{P}(Y_1 - Y_0 = 0)$, and $\mathbb{P}(Y_1 - Y_0 = 1)$. In this section we will discuss a result from Fan and Park (2010) that applies regardless of whether the outcome is binary, discrete, continuous, or a mixture of these. The logic is nearly identical to the binary outcome case considered above, but as the mathematical details are more involved, we will limit ourselves to stating the key results, and interpreting them in particular examples.[6]

As above, suppose that we observe the marginal distributions of Y_1 and Y_0. Let $F_0(y)$ be the CDF of Y_0 and $F_1(y)$ be the CDF of Y_1. Our goal is to bound the CDF $F(\delta)$ of $\Delta \equiv Y_1 - Y_0$, i.e. the distribution of treatment effects. To state the main result of Fan and Park (2010) we need to define some additional notation. First, let

$$\underline{F}(\delta) \equiv \sup_y F_1(y) - F_0(y - \delta). \tag{3.6}$$

(If you're unfamiliar with the notation "sup," cross it out and write max and you'll still get the basic idea.) This is a somewhat complicated expression, so let's disassemble it before proceeding. We compute $\underline{F}(\delta)$ by choosing a value of δ, then taking the difference between $F_1(y)$ and $F_0(y-\delta)$ at all values of y, and then finding the value of y at which this difference is maximal. For example, to compute $\underline{F}(0.2)$ we would maximize $F_1(y) - F_0(y - 0.2)$ over

[6] There is one important difference between the binary outcome case and the more general setting considered in this section: a positive ATE does **not in general** imply that more people are helped by the treatment than harmed by it. A treatment that helps a tiny fraction of people by a large amount and harms everyone else by a small amount can still have a positive ATE.

y. We defined $F_1(y)$ to be the CDF of Y_1, but what on earth is $F_0(y-\delta)$? When $\delta = 0$ this becomes $F_0(y)$, the CDF of Y_0. More generally,

$$F_0(y-\delta) = \mathbb{P}(Y_0 \leq y - \delta) = \mathbb{P}(Y_0 + \delta \leq y).$$

In other words, $F_0(y-\delta)$ is the CDF of a random variable constructed by *shifting* Y_0 a total of δ units to the right. The amount that we shift is determined by the location at which we want to evaluate $\underline{F}$. Figure 3.3 shows an example in which we shift F_0 two units to the right. Now define

$$\overline{F}(\delta) \equiv 1 + \left[\inf_y F_1(y) - F_0(y-\delta)\right]. \tag{3.7}$$

(If you're unfamiliar with the notation "inf," cross it out and write "min." and you'll still get the basic idea.) The definition of $\overline{F}(\delta)$ is very similar to that of $\underline{F}(\delta)$ except that we compute the *minimum* difference between $F_1(y)$ and $F_0(y-\delta)$. No we're ready to construct the sharp bounds for $F(\delta)$.

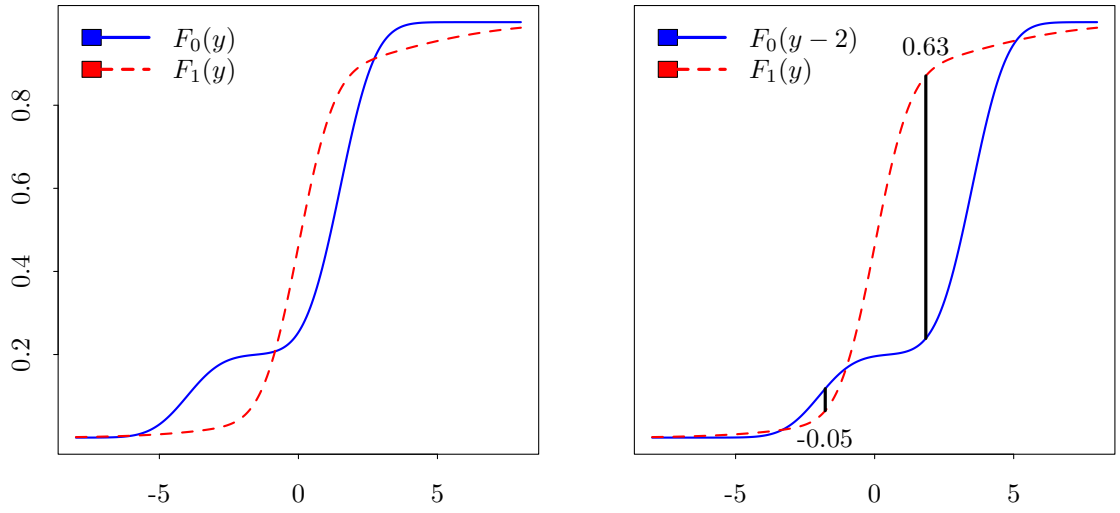


Figure 3.3: An illustration of (3.6) and (3.7) in a hypothetical example with continuous potential outcomes. The left panel depicts the marginal CDFs of Y_0 and Y_1. The right panel depicts the Fan and Park (2010) procedure for bounding $F(2) \equiv \mathbb{P}(Y_1 - Y_0 \leq 2)$. Shift the CDF of F_0 two units to the right and then compute the largest and smallest distances between the curves, yielding $F(2) \in [0.63, 1 - 0.05]$.

Theorem 3.1. *Let $F_0(\cdot)$ be the CDF of Y_0, let $F_1(\cdot)$ be the CDF of Y_1, and let $F(\cdot)$ be the CDF of $\Delta \equiv Y_1 - Y_0$. For any δ,*

$$0 \leq \underline{F}(\delta) \leq F(\delta) \leq \overline{F}(\delta) \leq 1.$$

where $\underline{F}(\cdot)$ and $\overline{F}(\cdot)$ are defined as in (3.6) and (3.7). These bounds bound are sharp.

Theorem 3.1 gives the sharp bounds for the distribution of treatment effects, $F(\cdot)$, in terms of the functions $\underline{F}(\cdot)$ and $\overline{F}(\cdot)$. Our bounds for the binary outcome example from above constitute a special case of this result. To see why, recall that in the binary outcome case,

$$F_0(y) = \begin{cases} 0, & y < 0 \\ 1 - p_0, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}, \quad F_1(y) = \begin{cases} 0, & y < 0 \\ 1 - p_1, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}$$

When Y is binary, $F(-0.5)$ equals the fraction of people harmed by the treatment.[7] Substituting $\delta = -0.5$ gives

$$F_0(y - \delta) = F_0(y + 0.5) = \begin{cases} 0, & y < -0.5 \\ 1 - p_0, & -0.5 \leq y < 0.5 \\ 1, & y \geq 0.5 \end{cases}$$

and subtracting this from $F_1(y)$ we obtain,

$$F_1(y) - F_0(y - \delta) = F_1(y) - F_0(y + 0.5) = \begin{cases} 0, & y < -0.5 \\ -(1 - p_0), & -0.5 \leq y < 0 \\ (1 - p_1) - (1 - p_0), & 0 \leq y < 0.5 \\ (1 - p_1) - 1, & 0.5 \leq y < 1 \\ 0, & y \geq 1 \end{cases}$$

Notice that $F_1(y) - F_0(y + 0.5)$ only takes on four distinct values. It follows that

$$\sup_y F_1(y) - F_0(y + 0.5) = \max\{0,\ (p_0 - 1),\ (p_0 - p_1),\ -p_1\} = \max\{0,\ (p_0 - p_1)\}$$
$$\inf_y F_1(y) - F_0(y + 0.5) = \min\{0,\ (p_0 - 1),\ (p_0 - p_1),\ -p_1\} = \min\{(p_0 - 1),\ -p_1\}$$

since $(p_0 - 1) \leq (p_0 - p_1) \leq -p_1$. This gives us precisely the same bounds for $\mathbb{P}(\text{Allergic})$ as we derived by hand in the preceding section, namely

$$\underline{F}(0.5) = \max\{0, -(p_1 - p_0)\}, \quad \overline{F}(0.5) = 1 + \min\{(p_0 - 1), -p_1\} = \min\{p_0, 1 - p_1\}.$$

This should give us at least some confidence in Theorem 3.1 in spite of our not having proved it!

Figure 3.3 depicts a more interesting example in which Y_0 and Y_1 are continuous random variables. To bound $F(2)$, i.e. $\mathbb{P}(Y_1 - Y_0 \leq 2)$, we shift the CDF of Y_0 two units to the right. The largest value of $F_1(y) - F_0(y - 2)$ is approximately 0.63, while the

[7] Since $(Y_1 - Y_0) \in \{-1, 0, 1\}$, any value of δ between -1 and 0 also equals the fraction harmed.

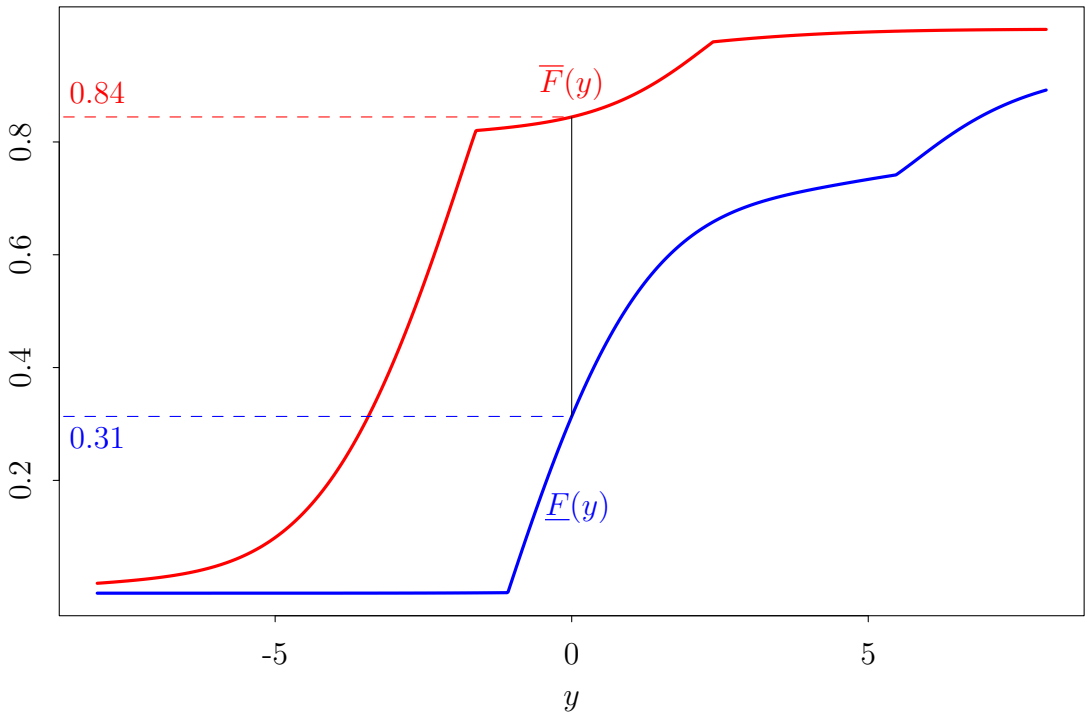


Figure 3.4: The full set of Fan and Park (2010) bounds for the marginal distributions given in Figure 3.3.

smallest is approximately -0.05. Therefore $0.63 \leq F(2) \leq 0.95$. Repeating this process for a wide range of values of δ gives the full set of (pointwise) bounds for $F(\cdot)$, as shown in Figure 3.4. This allows us to read of the bound for $F(y)$ for any desired value of y. For example, from the figure we see that $0.31 \leq F(0) \leq 0.84$. This means that at least 31% and no more than 84% of people are harmed by this particular treatment.

Chapter 4

Selection on Observables

As we saw in chapter 2, there is no selection bias when D is randomly assigned: a simple comparison of mean outcomes between treated and untreated individuals identifies the ATE. In many examples, however, carrying out a randomized controlled trial may be infeasible, unethical, or even impossible. In this chapter we will consider an assumption called *selection on observables* that allows us to identify the ATE from observational data by conditioning on observed characteristics $\boldsymbol{X}$. We'll consider two different approaches to identification that both rely on the selection on observables assumption: one based on regression adjustment and another based on propensity score weighting.

4.1 Does education cause political participation?

This example is based on Kam and Palmer (2008). University graduates are more likely to vote, volunteer for political campaigns, contact their elected representatives, and participate in demonstrations. Does this show that education *causes* political participation? Let $D = 1$ if you attended university and $D = 0$ otherwise. Further let Y be an index of political participation, where high values indicate greater participation and lower values indicate less. It seems hard to believe that D could be independent of the potential outcomes (Y_0, Y_1) in this example. University graduates differ from non-graduates in myriad ways that could also influence political participation. People from wealthy backgrounds are more likely to graduate from college. They are also more likely to have the leisure time required for political participation; if you are struggling to make ends meet it will be hard to find time to attend a political rally. Because it seems far-fetched to imagine anyone carrying out an experiment that forced some people to attend college and others not to, observational data is the best we can hope for if our goal is to identify the causal effect of education on political participation.

The assumption that $(Y_0, Y_1) \perp\!\!\!\perp D$ is clearly untenable, so what could we use instead? Our main reason for doubting that a simple comparison of mean political participation

across groups could be given a causal interpretation was that university graduates are different from non-graduates in more ways than their education level. But perhaps if we were to *condition* on these differences, effectively holding them fixed, we could find a way to make progress. In other words, even if (Y_0, Y_1) are not independent of D, perhaps there is a collection of observable individual characteristics $\boldsymbol{X}$ such that $(Y_0, Y_1) \perp\!\!\!\perp D|\boldsymbol{X}$. For example, perhaps by conditioning on sex, race, family background and so on we could break the dependence between college graduation and the potential outcomes. This idea is called *selection on observables* because it assumes that selection bias operates solely through characteristics that we can observe.

4.2 Selection on Observables and Overlap

The methods explored in this chapter rely on two assumptions. First is *selection on observables*, as outlined in the previous section. The precise version of this condition that we will rely on below is as follows.

Assumption 4.1 (Selection on Observables)**.**

$$\mathbb{E}(Y_0|\boldsymbol{X}, D) = \mathbb{E}(Y_0|\boldsymbol{X}), \quad \textit{and} \quad \mathbb{E}(Y_1|\boldsymbol{X}, D) = \mathbb{E}(Y_1|\boldsymbol{X}).$$

Assumption 4.1 says that the potential outcomes (Y_0, Y_1) are *mean independent* of the treatment D conditional on $\boldsymbol{X}$. This is weaker than but implied by the conditional independence assumption, namely $(Y_0, Y_1) \perp\!\!\!\perp D|\boldsymbol{X}$, described in the previous section.[1] Because our goal is to identify a mean, the ATE, we only require a mean independence assumption. To introduce our second assumption we require the following definition.

Definition 4.1 (Propensity Score)**.** The probability $p(\boldsymbol{X}) \equiv \mathbb{P}(D = 1|\boldsymbol{X})$ of treatment conditional on an observed random vector $\boldsymbol{X}$ is called the propensity score.

Assumption 4.2 (Overlap)**.** $0 < p(\boldsymbol{x}) < 1$ *for all* $\boldsymbol{x}$ *in the support of* $\boldsymbol{X}$*.*

Assumption 4.2 states that the propensity score is *strictly* between zero and one for any value that the covariates $\boldsymbol{X}$ could take on. Since $p(\boldsymbol{x}) \equiv \mathbb{P}(D = 1|\boldsymbol{X} = \boldsymbol{x})$, this requires that, among people with any fixed value $\boldsymbol{x}$ of the covariates $\boldsymbol{X}$, some are treated $(D = 1)$ and some are untreated $(D = 0)$.

Both Assumption 4.1 and Assumption 4.2 are crucial for the methods described below. Unfortunately the two are somewhat at odds with each other. The more observed controls $\boldsymbol{X}$ that we condition on, the more plausible the selection on observables assumption (Assumption 4.1) becomes.[2] At the same time, conditioning on a richer set of

[1] See part (ii) of Lemma 2.1.

[2] But beware of *bad controls*! See section 4.8 for details.

controls makes it *harder* to satisfy the overlap condition. Suppose that $\boldsymbol{X}$ includes race, sex, whether or not you attended an independent secondary school, year of birth, and post code. It is distinctly possible that *every* white male who attended an independent secondary school and was born in 1995 to a wealthy North Oxford family in fact graduated from university. If so, the overlap assumption fails for this particular value of $\boldsymbol{x}$. A common although not entirely satisfactory solution to the failure of Assumption 4.2 is to redefine the population of interest by restricting attention to only those values $\boldsymbol{x}$ for which overlap holds. For example, we might be forced to exclude people born to wealthy North Oxford families from our population of interest. Note that if we take this route, we will identify a *different ATE* than the one we initially set out to recover: one that corresponds to the restricted population.

4.3 Identification by Regression Adjustment

Our first approach to identifying the ATE using Assumption 4.1 and Assumption 4.2 is called *regression adjustment*. The idea is to compare mean values of Y between treated and untreated individuals within *strata* defined by a common value $\boldsymbol{x}$ of the covariates. This yields a conditional ATE given that $\boldsymbol{X} = \boldsymbol{x}$. This quantity, which we denote $\text{ATE}(\boldsymbol{x})$, is the average treatment effect for a certain kind of person, namely someone with covariates equal to $\boldsymbol{x}$, e.g. a white male born to a wealthy North Oxford family in 1995. To convert this into an unconditional ATE we average $\text{ATE}(\boldsymbol{x})$ over the distribution of $\boldsymbol{X}$ in the population using the law of iterated expectations (Lemma 1.2).

Theorem 4.1. *Under Assumption 4.1 and Assumption 4.2,*

$$ATE \equiv \mathbb{E}(Y_1 - Y_0) = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}(Y|\boldsymbol{X}, D=1)\right] - \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}(Y|\boldsymbol{X}, D=0)\right].$$

Proof. Since $Y = Y_0 + D(Y_1 - Y_0)$, under selection on observables

$$\begin{aligned}\mathbb{E}(Y|\boldsymbol{X}, D) &= \mathbb{E}(Y_0|\boldsymbol{X}, D) + D\left[\mathbb{E}(Y_1|\boldsymbol{X}, D) - \mathbb{E}(Y_0|\boldsymbol{X}, D)\right] \\ &= \mathbb{E}(Y_0|\boldsymbol{X}) + D\left[\mathbb{E}(Y_1|\boldsymbol{X}) - \mathbb{E}(Y_0|\boldsymbol{X})\right]\end{aligned}$$

where the first equality follows by the properties of conditional expectation, and the second from Assumption 4.1. Substituting $D = 0$ and $D = 1$ into the preceding expression,

$$\mathbb{E}(Y|\boldsymbol{X}, D=0) = \mathbb{E}(Y_0|\boldsymbol{X}), \quad \mathbb{E}(Y|\boldsymbol{X}, D=1) = \mathbb{E}(Y_1|\boldsymbol{X})$$

which in turn implies that

$$\text{ATE}(\boldsymbol{X}) = \mathbb{E}(Y_1 - Y_0|\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X}, D=1) - \mathbb{E}(Y|\boldsymbol{X}, D=0).$$

The overlap assumption (Assumption 4.2) implies that $\text{ATE}(\boldsymbol{X})$ is well-defined for all points in the support of $\boldsymbol{X}$, since it ensures that there are individuals with $D = 1$ and $D = 0$ for any value of the covariates. Hence, taking the expectation of both sides,

$$\text{ATE} = \mathbb{E}_{\boldsymbol{X}}\left[\text{ATE}(\boldsymbol{X})\right] = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}(Y|\boldsymbol{X}, D = 1)\right] - \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}(Y|\boldsymbol{X}, D = 0)\right]$$

by the law of iterated expectations. □

4.4 Estimation by Regression Adjustment

Let $\widehat{\mu}_0(\boldsymbol{X})$ be a consistent estimator of $\mathbb{E}(Y|\boldsymbol{X}, D = 0)$ and $\widehat{\mu}_1(\boldsymbol{X})$ be a consistent estimator of $\mathbb{E}(Y|\boldsymbol{X}, D = 1)$. Then, under general conditions,

$$\widehat{\text{ATE}}_{RA} \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mu}_1(\boldsymbol{X}_i) - \widehat{\mu}_0(\boldsymbol{X}_i)\right]$$

is a consistent estimator of the ATE, where RA stands for *regression adjustment*. The question remains: how do we obtain $\widehat{\mu}_0(\cdot)$ and $\widehat{\mu}_1(\cdot)$? If $\boldsymbol{X}$ is discrete and takes on a small number of values, we can simply calculate the sample mean of Y at each combination of $(D = 0, \boldsymbol{X} = \boldsymbol{x})$ for $\widehat{\mu}_0(\boldsymbol{x})$ and at each combination of $(D = 1, \boldsymbol{X} = \boldsymbol{x})$ for $\widehat{\mu}_1(\boldsymbol{x})$. If $\boldsymbol{X}$ contains any continuous variables, or is discrete but takes on a large number of values, however, this approach fails. Non-parametric methods, either series or kernel-based, provide an alternative but perform poorly when the dimension of $\boldsymbol{X}$ is large. Model-based approaches are also possible, e.g. assuming that $\mathbb{E}[Y|D = d, \boldsymbol{X}]$ is linear in $\boldsymbol{X}$ for a given value of d. If the model is a poor description of the true conditional mean function, however, this can produce misleading results. Model-based approaches can also mask failures of the overlap assumption: they will always generate a value for $\mathbb{E}[Y|D = d, \boldsymbol{X} = \boldsymbol{x}]$ even if there are no individuals in the dataset with $(D = d, \boldsymbol{X} = \boldsymbol{x})$. The model extrapolates from values that are actually contained in the dataset. Whichever method is used to construct estimates $\widehat{\mu}_0(\cdot)$ and $\widehat{\mu}_0(\cdot)$, a simple way to carry out inference that correctly accounts for this first-stage estimation step is to bootstrap pairs $(Y_i, \boldsymbol{X}_i)$.

4.5 Identification by Propensity Score Weighting

Our second approach to identifying the ATE using Assumption 4.1 and Assumption 4.2 is called *propensity score weighting*. Whereas regression adjustment compares average values of Y between treated and untreated individuals with the same value of $\boldsymbol{X}$, propensity score weighting calculates the average value of Y across *everyone in the population* with weights that depend on each person's actual treatment D and her predicted probability of treatment: the propensity score.

Theorem 4.2. *Under Assumption 4.1 and Assumption 4.2,*

$$ATE \equiv \mathbb{E}(Y_1 - Y_0) = \mathbb{E}\left[\frac{\{D - p(\boldsymbol{X})\}Y}{p(\boldsymbol{X})\{1 - p(\boldsymbol{X})\}}\right].$$

Proof. Since D is binary, $D^2 = D$, $(1-D)^2 = (1-D)$, and $D(1-D) = 0$. Hence,

$$DY = D^2Y_1 + D(1-D)Y_0 = DY_1$$
$$(1-D)Y = (1-D)DY_1 + (1-D)^2Y_0 = (1-D)Y_0$$

since $Y = DY_1 + (1-D)Y_0$. Thus,

$$\mathbb{E}\left[\frac{DY}{p(\boldsymbol{X})}\middle|\boldsymbol{X}\right] = \frac{1}{p(\boldsymbol{X})}\mathbb{E}\left[DY_1|\boldsymbol{X}\right] \tag{4.1}$$
$$\mathbb{E}\left[\frac{(1-D)Y}{1-p(\boldsymbol{X})}\middle|\boldsymbol{X}\right] = \frac{1}{1-p(\boldsymbol{X})}\mathbb{E}\left[(1-D)Y_0|\boldsymbol{X}\right]. \tag{4.2}$$

Now, by iterated expectations and Assumption 4.1,

$$\begin{aligned}\mathbb{E}[DY_1|\boldsymbol{X}] &= \mathbb{E}_{D|\boldsymbol{X}}\left[\mathbb{E}\left(DY_1|D,\boldsymbol{X}\right)\right] = \mathbb{E}_{D|\boldsymbol{X}}\left[D\mathbb{E}\left(Y_1|D,\boldsymbol{X}\right)\right] = \mathbb{E}_{D|\boldsymbol{X}}\left[D\mathbb{E}\left(Y_1|\boldsymbol{X}\right)\right] \\ &= \mathbb{E}[D|\boldsymbol{X}]\mathbb{E}[Y_1|\boldsymbol{X}] = p(\boldsymbol{X})\mathbb{E}[Y_1|\boldsymbol{X}]\end{aligned}$$

where the final equality uses $\mathbb{E}[D|\boldsymbol{X}] = \mathbb{P}(D=1|\boldsymbol{X})$. Similarly,

$$\begin{aligned}\mathbb{E}[(1-D)Y_0|\boldsymbol{X}] &= \mathbb{E}_{D|\boldsymbol{X}}\left[\mathbb{E}\left\{(1-D)Y_0|D,\boldsymbol{X}\right\}\right] = \mathbb{E}_{D|\boldsymbol{X}}\left[(1-D)\mathbb{E}\left(Y_0|D,\boldsymbol{X}\right)\right] \\ &= \mathbb{E}_{D|\boldsymbol{X}}\left[(1-D)\mathbb{E}\left(Y_0|\boldsymbol{X}\right)\right] = \mathbb{E}[1-D|\boldsymbol{X}]\mathbb{E}[Y_0|\boldsymbol{X}] \\ &= [1-p(\boldsymbol{X})]\,\mathbb{E}[Y_0|\boldsymbol{X}]\end{aligned}$$

Substituting these expressions for $\mathbb{E}[DY_1|\boldsymbol{X}]$ and $\mathbb{E}[(1-D)Y_0|\boldsymbol{X}]$ into (4.1) and (4.2)

$$\mathbb{E}\left[\frac{DY}{p(\boldsymbol{X})}\middle|\boldsymbol{X}\right] = \mathbb{E}(Y_1|\boldsymbol{X}), \quad \mathbb{E}\left[\frac{(1-D)Y}{1-p(\boldsymbol{X})}\middle|\boldsymbol{X}\right] = \mathbb{E}(Y_0|\boldsymbol{X})$$

so we see that

$$\begin{aligned}\mathrm{ATE}(\boldsymbol{X}) \equiv \mathbb{E}(Y_1 - Y_0|\boldsymbol{X}) &= \mathbb{E}\left[\frac{DY}{p(\boldsymbol{X})} - \frac{(1-D)Y}{1-p(\boldsymbol{X})}\middle|\boldsymbol{X}\right] \\ &= \mathbb{E}\left[\frac{DY\{1-p(\boldsymbol{X})\} - (1-D)Yp(\boldsymbol{X})}{p(\boldsymbol{X})\{1-p(\boldsymbol{X})\}}\middle|\boldsymbol{X}\right] \\ &= \mathbb{E}\left[\frac{DY - DYp(\boldsymbol{X}) - Yp(\boldsymbol{X}) + DYp(\boldsymbol{X})}{p(\boldsymbol{X})\{1-p(\boldsymbol{X})\}}\middle|\boldsymbol{X}\right] \\ &= \mathbb{E}\left[\frac{\{D-p(\boldsymbol{X})\}Y}{p(\boldsymbol{X})\{1-p(\boldsymbol{X})\}}\middle|\boldsymbol{X}\right].\end{aligned}$$

Therefore, taking iterated expectations,

$$\text{ATE} = \mathbb{E}_{\boldsymbol{X}}\left[\text{ATE}(\boldsymbol{X})\right] = \mathbb{E}_{\boldsymbol{X}}\left(\mathbb{E}\left[\frac{\{D - p(\boldsymbol{X})\}\, Y}{p(\boldsymbol{X})\,\{1 - p(\boldsymbol{X})\}}\middle|\, \boldsymbol{X}\right]\right) = \mathbb{E}\left[\frac{\{D - p(\boldsymbol{X})\}\, Y}{p(\boldsymbol{X})\,\{1 - p(\boldsymbol{X})\}}\right].$$

□

4.6 Estimation by Propensity Score Weighting

Suppose we already have a consistent estimator $\widehat{p}(\cdot)$ of the propensity score. Then,

$$\widehat{\text{ATE}}_{PSW} \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{[D_i - \widehat{p}(\boldsymbol{X}_i)]\, Y_i}{\widehat{p}(\boldsymbol{X}_i)\,[1 - \widehat{p}(\boldsymbol{X}_i)]}$$

where PSW stands for *propensity score weighting* is a consistent estimator of the ATE under Assumption 4.1, Assumption 4.2, and appropriate regularity conditions. But how can we estimate the propensity score? If $\boldsymbol{X}$ is discrete and only takes on a small number of values, we can estimate the propensity score directly using the sample fraction of observations with $\boldsymbol{X} = \boldsymbol{x}$. This approach is no longer possible when any of the elements of $\boldsymbol{X}$ is continuous and can perform poorly even for discrete $\boldsymbol{X}$ if some values $\boldsymbol{x}$ are shared by only a small number of people in the sample. A common model-based approach is to fit a "flexible" logit model, including levels, squares, and interactions of $\boldsymbol{X}$. Although fairly widespread and convenient, this approach has the potential to mask failures of overlap: the logit model will never give $p(\boldsymbol{X}) = 0$ or 1 regardless of whether there are values of $\boldsymbol{x}$ for which everyone in the sample is either treated or untreated. Moreover, the particular logit model that we specify could be a poor reflection of the true propensity score. Another approach uses non-parametric methods, either series or kernel based, to estimate the propensity score. While less prone to mis-specification than model-based approaches, non-parametric methods perform poorly when $\boldsymbol{X}$ is high-dimensional. Regardless of the particular method used, inference for propensity score weighting is somewhat complicated by the first-stage estimation of $\widehat{p}(\boldsymbol{X})$. An easy solution is to bootstrap pairs $(\boldsymbol{X}_i, Y_i)$.

4.7 Regression Adjustment versus Propensity Score Weighting

In theory, both Theorem 4.1 and Theorem 4.2 identify the *same* quantity, namely the ATE.[3] In practice, however, because they require us to use the data in different ways, estimators based on regression adjustment and propensity score weighting will differ,

[3] If Assumption 4.2 fails and we are forced to restrict attention individuals with values of $\boldsymbol{X}$ for which overlap holds, then both theorems identify the ATE for this *restricted* population.

sometimes substantially. Recall that regression adjustment requires us to model and estimate the conditional mean of Y given $(D = 0, \boldsymbol{X})$ and $(D = 1, \boldsymbol{X})$ whereas propensity score weighting requires us to model and estimate the conditional probability that $D = 1$ given $\boldsymbol{X}$. A particular challenge for propensity score weighting is values of $\widehat{p}(\boldsymbol{X}_i)$ that are close to zero or one, as this causes the fraction in $\widehat{\text{ATE}}_{PSW}$ to become unstable.

4.8 Don't condition on an intermediate outcome!

The key message of this chapter is that conditioning on the right information can allow us to identify causal effects even when treatment is not randomly assigned. The key message of this section is that conditioning on the *wrong* information can lead us to draw erroneous causal conclusions even when treatment *is indeed* randomly assigned. This problem is commonly known as **bad control** or **conditioning on an intermediate outcome**. We'll use a simple example to explain the problem and how to avoid it. For simplicity our discussion will be limited to a binary covariate X that is potentially a "bad control." Very similar reasoning applies to any covariate, binary or not.

Gwynaeth attended a bilingual French and English high school in Canada. She is now a university senior lecturer and earns a good living. Did attending a bilingual high school *cause* her earnings to be higher than they otherwise would have been? Let Y be a person's wage, and define $D = 1$ if she attends a bilingual high school and zero otherwise. Gwynaeth chose to attend a bilingual high school: her D was not randomly assigned. But imagine that we were to carry out an experiment in which we *did* randomly assign D, sending half of a group of students to a bilingual high school and the rest to a regular high school. Since $D \perp\!\!\!\perp (Y_0, Y_1)$, we have $\mathbb{E}(Y_0|D) = \mathbb{E}(Y_0)$ and $\mathbb{E}(Y_1|D) = \mathbb{E}(Y_1)$. Thus,

$$\mathbb{E}(Y|D = 1) - \mathbb{E}(Y|D = 0) = \mathbb{E}(Y_1|D = 1) - \mathbb{E}(Y_0|D = 0) = \mathbb{E}(Y_1 - Y_0) = \text{ATE}$$

since $Y = (1 - D)Y_0 + DY_1$. Because students in this hypothetical experiment are randomly assigned to high schools, we don't need to condition on *anything* to identify the average treatment effect D on Y: a simple comparison of means suffices. But what would happen if we nevertheless *did* choose to condition on something?

Given that she is a university senior lecturer, it will come as no surprise that Gwynaeth attended university herself. Let $X = 1$ if a person attended university and zero otherwise. Should we condition on X to estimate the ATE in our hypothetical experiment? *Absolutely not!* College attendance X is an **intermediate outcome** aka a **bad control**. Because D causes X as well as Y, the treatment D is no longer randomly assigned if we *condition* on X. In other words, conditioning on X introduces selection bias that was not present unconditionally. We will examine this in two ways: first intuitively using a simple stylized model, and then mathematically, building on our earlier derivations. Consider

the following stylized model:

(i) Two factors increase a person's wage: knowledge K and innate ability A.

(ii) Attending a bilingual high school increases K more than attending a regular one.

(iii) The top 30% of people in the population distribution of $(K + A)$ attend university.

Because D was randomly assigned it is independent of A. This is no longer true, however, conditional on X. First consider the group of people from our experiment who attended university ($X = 1$). Among them, those who didn't attend a bilingual high school ($D = 0$) will have *higher* average ability than those with did ($D = 1$). Why is this the case? Our second assumption was that those who didn't attend a bilingual school end up with a lower value of K, on average, than those who did. Thus, for them to make it into the top 30% of $(K + A)$ requires a *higher* value of A. Putting it another way, if you did attend a bilingual school, then you can make in into the top 30% of $(K+A)$ with a lower value of A. Because those with $(D = 1, X = 1)$ have lower ability than those with $(D = 0, X = 1)$ and lower ability implies lower wages,

$$\mathbb{E}[Y|D = 1, X = 1] - \mathbb{E}[Y|D = 0, X = 1] < \mathbb{E}[Y_1|X = 1] - \mathbb{E}[Y_0|X = 1] = \text{ATE}(X = 1).$$

A similar argument shows that, that among those who did not attend university, those with $D = 1$ will have lower average ability than those with $D = 0$.[4] It follows that

$$\mathbb{E}[Y|D = 1, X = 0] - \mathbb{E}[Y|D = 0, X = 0] < \mathbb{E}[Y_1|X = 0] - \mathbb{E}[Y_0|X = 0] = \text{ATE}(X = 0).$$

In this simple model, conditioning on university attendance would lead us to *understate* the true treatment effect. Now that we understand the basic intuition, we'll take a more mathematical look at the problem of a bad control. The following lemma and discussion are a special case of Wooldridge (2005).

Lemma 4.1. *Let X be a binary RV and suppose that $\mathbb{E}(Y_j) = \mathbb{E}(Y_j|D)$ for $j = 0, 1$. If $\mathbb{E}(Y_j|X, D) = \mathbb{E}(Y_j|X)$ for $j = 0, 1$ then at least one of the following must hold:*

(i) $X \perp\!\!\!\perp D$

(ii) $\mathbb{E}(Y_j|X) = \mathbb{E}(Y_j)$ *for* $j = 0, 1$

Proof of Lemma 4.1. Since Y_j is mean independent of D for $j = 0, 1$ and X is binary, the law of iterated expectations gives

$$\begin{aligned}\mathbb{E}(Y_1) &= \mathbb{E}(Y_1|D) = \mathbb{E}_{X|D}\left[\mathbb{E}(Y_1|D, X)\right] \\ &= \mathbb{E}(Y_1|D, X = 0)\mathbb{P}(X = 0|D) + \mathbb{E}(Y_1|D, X = 1)\mathbb{P}(X = 1|D)\end{aligned}$$

[4] If you did not make it into the top 30% of the distribution of $(K + A)$ in spite of receiving the extra boost to K that comes from $D = 1$, then you must have had a low value of A.

and similarly for Y_0. Further imposing that (Y_0, Y_1) are mean independent of D given X

$$\mathbb{E}(Y_0) = \mathbb{E}(Y_0|X=0)\mathbb{P}(X=0|D=d) + \mathbb{E}(Y_0|X=1)\mathbb{P}(X=1|D=d) \tag{4.3}$$
$$\mathbb{E}(Y_1) = \mathbb{E}(Y_1|X=0)\mathbb{P}(X=0|D=d) + \mathbb{E}(Y_1|X=1)\mathbb{P}(X=1|D=d). \tag{4.4}$$

The left-hand sides of (4.3) and (4.4) do not depend on the value d that the treatment D takes on. Thus, to avoid a contradiction between $\mathbb{E}(Y_j|D) = \mathbb{E}(Y_j)$ and $\mathbb{E}(Y_j|X, D) = \mathbb{E}(Y_j|X)$, the RHS *cannot* depend on d either. There are only two ways that this is possible. The first is if $X \perp\!\!\!\perp D$ so that $\mathbb{P}(X = x|D = d) = \mathbb{P}(X = x)$ and

$$\mathbb{E}(Y_0) = \mathbb{E}(Y_0|X=0)\mathbb{P}(X=0) + \mathbb{E}(Y_0|X=1)\mathbb{P}(X=1)$$
$$\mathbb{E}(Y_1) = \mathbb{E}(Y_1|X=0)\mathbb{P}(X=0) + \mathbb{E}(Y_1|X=1)\mathbb{P}(X=1).$$

If X and D are *dependent*, then the only way that the RHS (4.3) and (4.4) could *not* involve d is if $\mathbb{E}(Y_1|X = 0) = \mathbb{E}(Y_1|X = 1) = \mathbb{E}(Y_1)$ and similarly for Y_0, so that

$$\mathbb{E}(Y_0|X=0)\mathbb{P}(X=0|D=d) + \mathbb{E}(Y_0|X=1)\mathbb{P}(X=1|D=d) = \mathbb{E}(Y_0)$$
$$\mathbb{E}(Y_1|X=0)\mathbb{P}(X=0|D=d) + \mathbb{E}(Y_1|X=1)\mathbb{P}(X=1|D=d) = \mathbb{E}(Y_1)$$

since $\mathbb{P}(X = 0|D = d) + \mathbb{P}(X = 1|D = d) = 1$ for any value of d. □

Lemma 4.1 tells us that if treatment is randomly assigned, then any covariate X that is both related to treatment *and* affects the average potential potential outcomes is *necessarily* a bad control. Given that D is mean independent of (Y_0, Y_1), such an X cannot satisfy $\mathbb{E}(Y_j|X, D) = \mathbb{E}(Y_j|X)$, the selection on observables assumption. This means that we *cannot* identify the ATE by conditioning on X and using, for example, regression adjustment or propensity score weighting. In our example from above, college attendance (X) was both affected by attending a bilingual high school (D) and in turn affected wages. Given that D was randomly assigned, the lemma shows that college attendance is a bad control in the wages and high-school experiment.

Lemma 4.1 does *not* say that conditioning on a covariate that is related to D and Y is always bad. Indeed the whole point of this chapter is to try to eliminate selection bias by finding covariates that *are* related to D and Y. The lemma concerns a setting where we have already solved the selection problem by randomly assigning D. It tells us when conditioning on X would *introduce* selection bias that was not there to begin with. This may strike you as odd: why would we bother to condition on X if we already knew that the treatment had been randomly assigned? There are two answers to this question. First, it is fairly common in practice for researchers to condition on covariates when analyzing experimental data, either to estimate conditional ATEs for people with different characteristics or to reduce the variance of their overall ATE estimator by "projecting

out" sources of noise in Y. Lemma 4.1 tells us that this is perfectly fine *provided that these covariates were measured before assigning the treatment*: because D is randomly assigned, we know that any *pre-existing* characteristics of individuals, e.g. sex or age, will be independent of treatment and hence cannot be bad controls.

Second, the reasoning used in our proof of Lemma 4.1 also applies to settings in which D is not randomly assigned. Suppose that we have a set of "good controls" $\mathbf{W}$ that satisfy Assumption 4.1, i.e. (Y_0, Y_1) are mean independent of D given $\boldsymbol{W}$. Now suppose that we are considering adding an *additional* binary variable X to our set of controls. We should only add X if (Y_0, Y_1) are mean independent of D given the *full* set of controls $(\boldsymbol{W}, X)$. Suppose this is the case. Then, by iterated expectations and our two of mean independence assumptions,

$$\begin{aligned}\mathbb{E}(Y_1|\boldsymbol{W}) &= \mathbb{E}(Y_1|\boldsymbol{W}, D) = \mathbb{E}_{X|\boldsymbol{W},D}\left[\mathbb{E}(Y_1|\boldsymbol{W}, D, X)\right] \\ &= \mathbb{E}(Y_1|\boldsymbol{W}, D, X=0)\mathbb{P}(X=0|\boldsymbol{W}, D) + \mathbb{E}(Y_1|\boldsymbol{W}, D, X=1)\mathbb{P}(X=1|\boldsymbol{W}, D) \\ &= \mathbb{E}(Y_1|\boldsymbol{W}, X=0)\mathbb{P}(X=0|\boldsymbol{W}, D) + \mathbb{E}(Y_1|\boldsymbol{W}, X=1)\mathbb{P}(X=1|\boldsymbol{W}, D)\end{aligned}$$

and similarly for Y_0, yielding

$$\begin{aligned}\mathbb{E}(Y_0|\boldsymbol{W}) &= \mathbb{E}(Y_0|\boldsymbol{W}, X=0)\mathbb{P}(X=0|\boldsymbol{W}, D) + \mathbb{E}(Y_0|\boldsymbol{W}, X=1)\mathbb{P}(X=1|\boldsymbol{W}, D) \\ \mathbb{E}(Y_1|\boldsymbol{W}) &= \mathbb{E}(Y_1|\boldsymbol{W}, X=0)\mathbb{P}(X=0|\boldsymbol{W}, D) + \mathbb{E}(Y_1|\boldsymbol{W}, X=1)\mathbb{P}(X=1|\boldsymbol{W}, D).\end{aligned}$$

The left hand sides of these equations do not depend on D. By reasoning similar to that used in the proof of Lemma 4.1, the only way that the right hand sides could *not* depend on D is if either $X \perp\!\!\!\perp D|\boldsymbol{W}$ or $\mathbb{E}(Y_j|\boldsymbol{W}, X) = \mathbb{E}(Y_j|\boldsymbol{W})$. If X is determined after D, it is unlikely that the first of these conditions hold. If X satisfies the second condition, then conditioning on it is completely irrelevant in any case: it will neither help us to identify ATEs, conditional or unconditional, nor will it improve the precision of our estimates.

Chapter 5

The Local Average Treatment Effects (LATE) Model

In section 1.8 we showed that an OLS regression of Y on D does not in general identify the ATE: selection bias is the norm rather than the exception in social science. One possible solution, considered in chapter 4 is to make the *selection on observables* assumption. Under this assumption, conditioning on observed characteristics $\boldsymbol{X}$ suffices to break any dependence between D and (Y_0, Y_1). Selection on observables, however, is a very strong assumption. How likely is it that we truly observe *all* the factors that create dependence between D and (Y_0, Y_1)? As an alternative to selection on observables, this chapter considers the use of an *instrumental variable* Z to identify causal effects. For simplicity we focus on the case where Z, like D, is binary. We first review the "textbook" homogeneous effects IV model before addressing our key question for this chapter: what does an instrumental variable identify in a world of heterogeneous effects?

5.1 Instrumental Variables with Homogeneous Effects

Suppose that $Y = \alpha + \beta D + U$, where (Y, D) are observed random variables, U is an unobserved random variable, and (α, β) are unknown constants. Under standard conditions, the OLS estimator for β converges in probability to

$$\beta_{OLS} = \frac{\text{Cov}(D, Y)}{\text{Var}(D)} = \frac{\beta \text{Cov}(D, D) + \text{Cov}(D, U)}{\text{Var}(D)} = \beta + \frac{\text{Cov}(D, U)}{\text{Var}(D)}$$

which does not equal β unless $\text{Cov}(D, U) = 0$. Suppose we have an instrumental variable Z such that $\text{Cov}(Z, U) = 0$ (exogeneity) and $\text{Cov}(Z, D) \neq 0$ (relevance). Then, under standard conditions, the instrumental variables (IV) estimator of β converges to

$$\beta_{IV} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, D)} = \frac{\beta \text{Cov}(Z, D) + \text{Cov}(Z, U)}{\text{Cov}(Z, D)} = \beta + \frac{\text{Cov}(Z, U)}{\text{Cov}(Z, D)} = \beta.$$

What do we make of $Y = \alpha + \beta D + U$ in light of our discussion of potential outcomes from above? This is a **homogeneous treatment effects** model. In other words the model (implicitly) assumes that the treatment effect is the same for everyone. We can express this model in the potential outcomes notation from above as follows. To find Y_0, set $D = 0$; to find Y_1, set $D = 1$. This gives

$$Y_0 \equiv \alpha + U, \quad Y_1 \equiv \alpha + \beta + U \implies Y_1 - Y_0 = \beta$$

so the ATE equals the *constant* β. But what if the assumption of homogeneous treatment effects is incorrect? Does IV still identify a meaningful causal quantity? To answer this question, we will drop the assumption that $Y = \alpha + \beta D + U$ where (α, β) are constants and study the behavior of the **IV estimand** $\beta_{IV} \equiv \text{Cov}(Z, Y)/\text{Cov}(Z, D)$ under heterogeneous treatment effects.

When D and Z are both binary, as we will assume they are in these notes, the IV estimand can be written in a simpler form. Applying Lemma 1.1 to the numerator and denominator of β_{IV}

$$\beta_{IV} = \frac{p(1-p)\left[\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)\right]}{p(1-p)\left[\mathbb{E}(D|Z=1) - \mathbb{E}(D|Z=0)\right]} = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(D|Z=1) - \mathbb{E}(D|Z=0)} \tag{5.1}$$

where $p \equiv \mathbb{P}(Z = 1)$. The rightmost fraction in (5.1) is often called the **Wald estimand**. Substituting sample means for population expectations gives the **Wald estimator**, a convenient shorthand for $\widehat{\beta}_{IV}$ in the binary-treatment/binary-instrument case.

5.2 Non-compliance and the Intent-to-Treat Effect

Loosely speaking, an instrument is a variable Z that only affects Y through its affect on D. In the preceding section we assumed that Z was **relevant**, $\text{Cov}(Z, D) \neq 0$, and **exogenous**, $\text{Cov}(Z, U) = 0$. These two assumptions sufficed to identify the ATE under homogeneous treatment effects. Because it only involves D and Z, the relevance assumption is unchanged in a heterogeneous effects model, as we will see below. In contrast, the exogeneity assumption $\text{Cov}(Z, U) = 0$ explicitly involves the additive error term U from the homogeneous effects model. We will need to find alternative assumptions to take its place if we wish to allow different people to have different treatment effects. Before we can state these assumptions, however, we need to develop some terminology and definitions that are specific to the heterogeneous effects case.

While IV methods are routinely applied to observational datasets, the definitions we need to introduce are easiest understood by considering an experimental example. The influential "Moving to Opportunity" (MTO) intervention from the mid 1990s offered vouchers to families living in high-poverty neighborhoods that would allow them

to relocate to low-poverty areas. A number of influential recent papers in economics have studied the causal effect of the MTO intervention, e.g. on labor market outcomes later in life. Although vouchers were assigned at random, for obvious reasons families could neither be compelled to move if they received one nor prevented from moving if they did not. Nearly 50% of the families offered vouchers through the MOT intervention chose to remain in their original neighborhoods while 20% of those not offered a voucher nevertheless moved to a low-poverty neighborhood. This phenomenon is called **non-compliance**: subjects in social experiments can only be offered treatment rather than compelled to take it up.

Let $Z = 1$ for families who were offered a voucher and $D = 1$ for those that moved to a low-poverty neighborhood. Non-compliance means that Z may not equal D. Crucially, while Z was randomly assigned, D was not: families *chose* whether or not to move, and those who did likely differed in many ways from those who did not. For this reason, a näive comparison of $\mathbb{E}[Y|D = 1]$ against $\mathbb{E}[Y|D = 0]$ will be polluted by selection bias, as detailed in section 1.8. But what about conditioning on Z rather than D? The so-called **intent-to-treat** estimand does exactly this, regressing Y on Z rather than D:

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

Notice that the ITT is *precisely* the same thing as the numerator of the Wald estimand. Because Z was randomly assigned, the ITT identifies the ATE of Z on Y: the average causal effect of being *offered* a voucher on labor market outcomes.[1]

There is nothing wrong with the ITT. It answers a perfectly well-defined causal question and is immune to selection bias. At the same time, the intervention was called *moving* to opportunity for a reason: we don't expect that being offered a voucher, on its own, could affect labor market outcomes for families that did not, in fact, relocate. In other words, it is the causal effect of D that truly interests us, not the causal effect of Z. Under **perfect compliance**, $Z = D$ so the ITT equals the ATE. If everyone who is offered a voucher moves, and everyone who is not offered a voucher stays put, then the effect of being offered a voucher is the same as the effect of actually moving. By driving a wedge between Z and D, however, non-compliance causes the ITT and ATE to differ. Recall that 50% of families offered a voucher ($Z = 1$) chose not to move ($D = 0$). Since $Y = (1 - D)Y_0 + DY_1$, we have

$$\begin{aligned}
\mathbb{E}[Y|Z = 1] &= \mathbb{E}_{D|Z=1}\left[\mathbb{E}(Y|Z = 1, D)\right] \\
&= \mathbb{E}\left[Y_0|Z = 1, D = 0\right]\mathbb{P}(D = 0|Z = 1) + \mathbb{E}\left[Y_1|Z = 1, D = 1\right]\mathbb{P}(D = 1|Z = 1) \\
&= \frac{1}{2} \times \mathbb{E}(Y_0|Z = 1, D = 0) + \frac{1}{2} \times \mathbb{E}(Y_1|Z = 1, D = 1).
\end{aligned}$$

[1] See section 2.3, changing the notation so that Z replaces D.

This calculation shows that $\mathbb{E}[Y|Z=1]$ contains a mixture of Y_0 and Y_1 with mixing weights that depend on the extent of non-compliance. In effect, families that choose not to move despite being offered a voucher dilute $\mathbb{E}[Y|Z=1]$ so that it averages over the outcomes of both treated and untreated families. Non-compliance causes similar challenges for interpreting $\mathbb{E}[Y|Z=0]$. Since 20% of families who were not offered a voucher chose to relocate,

$$\begin{aligned}\mathbb{E}[Y|Z=0] &= \mathbb{E}_{D|Z=0}\left[\mathbb{E}(Y|Z=0,D)\right] \\ &= \mathbb{E}\left[Y_0|Z=0,D=0\right]\mathbb{P}(D=0|Z=0) + \mathbb{E}\left[Y_1|Z=0,D=1\right]\mathbb{P}(D=1|Z=0) \\ &= \frac{4}{5}\times\mathbb{E}(Y_0|Z=0,D=0) + \frac{1}{5}\times\mathbb{E}(Y_1|Z=0,D=1).\end{aligned}$$

Again we obtain a mixture of Y_0 and Y_1 that depends on the extent of non-compliance. None of these calculations in any way contradicts our earlier claim that the ITT identifies the average causal effect of Z on Y. They merely show the challenges involved in trying to relate this quantity to the causal effect of interest, that of D on Y.

5.3 Compliers, Defiers, and Friends

In the previous section we showed that $\mathbb{E}[Y|Z=0]$ and $\mathbb{E}[Y|Z=1]$ each contain a mixture of Y_0 and Y_1, making it difficult to relate the causal effect of being offered treatment, the ITT, to that of actually receiving it. But there is a second and more subtle challenge lurking here. Notice that $\mathbb{E}(Y|Z=1)$ involves $\mathbb{E}(Y_0|Z=1,D=0)$ while $\mathbb{E}(Y|Z=0)$ involves $\mathbb{E}(Y_0|Z=0,D=0)$. While these are both averages of Y_0, they average over *different kinds of families*. Families who did not choose to move despite being offered a voucher $(Z=1,D=0)$ are probably very different from those that didn't move but weren't offered a voucher $(Z=0,D=0)$.[2] Perhaps the group with $Z=1$ and $D=0$ is much poorer, making it harder for them to move even with a voucher. Or perhaps they are less ambitious, and less willing to take advantage of the opportunity presented by moving to a new neighborhood. Whatever the reasons behind their choice not to move, we would likely expect their values of Y_0 to differ systematically from those of other families. Similarly, $\mathbb{E}[Y|Z=0]$ involves $\mathbb{E}[Y_1|Z=0,D=1]$ while $\mathbb{E}[Y|Z=1]$ involves $\mathbb{E}[Y_1|Z=1,D=1]$. Both are averages of Y_1 but, again, for different groups of families. Those that chose to move despite not being offered a voucher $(Z=0,D=1)$ may be richer, making it easier for them to move. They might also be more ambitious given their willingness to "move to opportunity" without being offered a voucher.

To make it easier to discuss these different types of families, and as a lead-in to our derivations for IV under heterogeneous effects, we introduce the notion of **compliance**

[2] Remember that Z was randomly assigned, but D is something that families could *choose*.

types.[3] The idea is to consider every possible rule that a family could use to decide whether or not to move, depending on whether they were offered a voucher. Mathematically, this involves listing every function that maps Z to D. Because both are binary, there are four possibilities. We give each a name corresponding to the "type" (T) of family that would adopt it:

$$\begin{aligned} \text{Never-taker:} \quad & T = n \iff D(Z) = 0 \\ \text{Always-taker:} \quad & T = a \iff D(Z) = 1 \\ \text{Complier:} \quad & T = c \iff D(Z) = Z \\ \text{Defier:} \quad & T = d \iff D(Z) = (1 - Z). \end{aligned}$$

In the MTO example, never-takers are the families that would *never* choose leave their current neighborhood, regardless of whether they are offered a voucher. Similarly, always-takers are families that would *always* choose to move to a lower-poverty area, regardless of whether they are offered a voucher. As their name suggests, compliers are the families that *comply* with their treatment offer: they move when given a voucher and stay put otherwise. Defiers are the most exotic of the four compliance types. These are families with a decidedly contrarian streak: they will *only* move to a low-poverty neighborhood when they are *not* offered a voucher.[4]

There is a crucial point about compliance types that is easy to miss: they are defined relative to a *particular* instrumental variable Z. Consider three different instruments intended to encourage families to move to low-poverty neighborhoods: Z_1 is an offer to provide help looking for a new apartment, Z_2 is a voucher worth \$500 per month towards rent for families that move, and Z_3 is title to a house, completely free of charge, in a low-poverty area. Someone who is a never-taker for Z_1 could easily be a complier for Z_2, and someone who is a never-taker for Z_2 could easily be a complier for Z_3. Our intuition is that Z_1 provides the weakest inducement to move while Z_3 provides the strongest. This point will be especially important below, where we will show that IV identifies the ATE *for compliers* when treatment effects can vary across individuals (or families).

5.4 The Local Average Treatment Effects Model

Having defined the four compliance types—always-takers, never-takers, compliers, and defiers—we can now state the assumptions that take the place of instrument exogeneity when treatment effects may be heterogeneous. The first is **unconfounded type**.

[3] Much of the literature refers to these as the "LATE Principal Strata." The meaning of LATE will be clarified below, but *principal stratum* is exceedingly vague, so I prefer to avoid the term.

[4] Our definition of "defier" assumes that $Z = 1$ should make families *more* likely to move. If the reverse is true, simply replace Z with $\widetilde{Z} = 1 - Z$.

Assumption 5.1 (Unconfounded Type)**.** *For all compliance types* $t \in \{a, c, n, d\}$

$$\mathbb{P}(T = t) = \mathbb{P}(T = t|Z = 0) = \mathbb{P}(T = t|Z = 1).$$

Assumption 5.1 says that knowing the value of a person's instrumental variable tells us nothing about her type. If Z is randomly assigned, as in an experiment like MTO example, this assumption holds automatically: a person's type can be viewed as a pre-existing characteristic just like age, sex, or income. In examples where Z arises from observational data, Assumption 5.1 is a substantive assumption that could in principle fail to hold. The second assumption is that there are **no defiers**.

Assumption 5.2 (No Defiers)**.** $\mathbb{P}(T = d) = 0$.

Assumption 5.2 says that there are no defiers in the population: everyone is a never-taker, a complier, or an always-taker. If we view Z as an inducement that lowers the cost of taking up the treatment, this assumption entails that the instrument shifts everyone in the *same direction* or not at all. In the MTO example, this requires that offering a family a voucher cannot make them less likely to move. For this reason the no defiers assumption is sometimes called **monotonicity**. Assumption 5.2 is natural in a model of rational choice, particularly if Z is a straightforward cost shifter such as the rent voucher from the MTO example. Nevertheless, it is still a restriction. Under the unconfounded type and no-defiers assumptions we can calculate the fraction of each compliance type in the population using the results of the following lemma.

Lemma 5.1. *Under Assumptions 5.1 and 5.2,*

$$\begin{aligned}
\mathbb{P}(D = 1|Z = 1) &= \mathbb{P}(T = a) + \mathbb{P}(T = c) \\
\mathbb{P}(D = 0|Z = 1) &= \mathbb{P}(T = n) \\
\mathbb{P}(D = 1|Z = 0) &= \mathbb{P}(T = a) \\
\mathbb{P}(D = 0|Z = 0) &= \mathbb{P}(T = c) + \mathbb{P}(T = n)
\end{aligned}$$

Notice that Lemma 5.1 directly identifies the share of never-takers and always-takers in the population: $\mathbb{P}(D = 0|Z = 1)$ equals $\mathbb{P}(T = n)$ while $\mathbb{P}(D = 1|Z = 0)$ equals $\mathbb{P}(T = a)$. To obtain the share of compliers, we take differences:

$$\mathbb{P}(T = c) = \mathbb{P}(D = 1|Z = 1) - \mathbb{P}(D = 1|Z = 0). \tag{5.2}$$

Notice that the right-hand side of (5.2) is *precisely* the denominator of the Wald estimand from (5.1): since D is binary, $\mathbb{E}(D|Z) = \mathbb{P}(D = 1|Z)$. To keep us from getting bogged down in book-keeping, I defer the proof of Lemma 5.1 to section 5.6. The intuition behind the proof, however, is fairly straightforward. Consider $\mathbb{P}(D = 0|Z = 1)$. Because

we have assumed that there are no defiers (Assumption 5.2), a family that is offered a voucher ($Z = 1$) but doesn't move ($D = 0$) is a never-taker. Since we have assumed that the fraction of each type is the same among those who are offered a voucher and those who are not (Assumption 5.1), $\mathbb{P}(D = 0|Z = 1)$ tells us the overall fraction of never-takers in the population. The remaining equations follow by similar reasoning: see section 5.6. The third and final assumption that replaces the more familiar instrument invalidity assumption from section 5.1 is called **mean exclusion**.

Assumption 5.3 (Mean Exclusion)**.** $\mathbb{E}[Y_j|Z = z, T = t] = \mathbb{E}[Y_j|T = t]$ *for all* (j, z, t).

In words, Assumption 5.3 says that the average values of (Y_0, Y_1) for any type t of family does not depend on whether they were offered a voucher. Mean exclusion does *not* require that Y is unaffected by Z. Because $Y = (1 - D)Y_0 + DY_1$, the instrument will necessarily affect Y if it affects D. What Assumption 5.3 entails is that Z has no *direct* effect on Y: it may shift D but leaves Y_0 and Y_1 unchanged. In the MTO example, this requires that merely being offered a voucher has no effect on labor market outcomes for a family that does not choose to move. Similarly, it requires that not being offered a voucher has no effect on labor market outcomes for a family that does choose to move. Combining Assumptions 5.1, 5.2, and 5.3 we can derive a lemma that will ultimately allow us to relate the ITT to the causal effect of D on Y.

Lemma 5.2. *Under Assumptions 5.1, 5.2, and 5.3.*

$$\begin{aligned}
\mathbb{E}[Y|D = 1, Z = 1] &= \frac{\mathbb{P}(T = a)\mathbb{E}[Y_1|T = a] + \mathbb{P}(T = c)\mathbb{E}[Y_1|T = c]}{\mathbb{P}(T = a) + \mathbb{P}(T = c)} \\
\mathbb{E}[Y|D = 0, Z = 1] &= \mathbb{E}[Y_0|T = n] \\
\mathbb{E}[Y|D = 1, Z = 0] &= \mathbb{E}[Y_1|T = a] \\
\mathbb{E}[Y|D = 0, Z = 0] &= \frac{\mathbb{P}(T = n)\mathbb{E}[Y_0|T = n] + \mathbb{P}(T = c)\mathbb{E}[Y_0|T = c]}{\mathbb{P}(T = n) + \mathbb{P}(T = c)}
\end{aligned}$$

Lemma 5.2 relates the means of the observed outcome Y given D and Z to the means of the potential outcomes (Y_0, Y_1) for different types of families. In the MTO intervention, for example, the average value of Y for families that were not offered a voucher but moved to a low-poverty neighborhood nonetheless, $\mathbb{E}[Y|D = 1, Z = 0]$, identifies the average value of Y_1 for *always-takers*. Similarly, the average value of Y for families that were offered a voucher but chose not to move, $\mathbb{E}[Y|D = 0, Z = 1]$, identifies the average value of Y_0 for *never-takers*. So far we have used Assumptions 5.1, 5.2, and 5.3 to derive two lemmas. Lemma 5.1 tells us something about the denominator of the Wald estimand. And since $\mathbb{E}[Y|Z] = \mathbb{E}_{D|Z}[\mathbb{E}(Y|D, Z)]$ by the law of iterated expectations, Lemma 5.2 tells us something about the numerator of the Wald estimator, i.e. the ITT. Before combining them, we require one futher assumption.

Assumption 5.4 (Existence of Compliers)**.** $\mathbb{P}(T = c) > 0$

By (5.2), Assumption 5.4 is equivalent to $\mathbb{P}(D = 1|Z = 1) \neq \mathbb{P}(D = 1|Z = 0)$ under Assumptions 5.1 and 5.2. To put it another way, the *lack* of a first-stage relationship between Z and D would indicate that there are *no compliers* in the population. We cannot carry out IV without compliers: notice that the Wald estimator is undefined if $\mathbb{E}(D|Z = 1) = \mathbb{E}(D|Z = 0)$. Because D is binary, $\text{Cov}(Z, D) = p(1 - p)\left[\mathbb{E}(D|Z = 1) - \mathbb{E}(D|Z = 0)\right]$. Thus, Assumption 5.4 is equivalent to the instrument relevance condition in the textbook homogeneous effects IV case from section 5.1. We are finally ready to answer our main question for the chapter: what does IV identify in a world of heterogeneous effects?

Theorem 5.1. *Under Assumptions 5.1–5.4,*

$$\frac{\mathbb{E}(Y|Z = 1) - \mathbb{E}(Y|Z = 0)}{\mathbb{E}(D|Z = 1) - \mathbb{E}(D|Z = 0)} = \mathbb{E}\left[Y_1 - Y_0|T = c\right]$$

Proof of Theorem 5.1. To begin, consider the numerator of the Wald Estimand. By the Law of Iterated Expectations,

$$\begin{aligned}\mathbb{E}\left(Y|Z = z\right) = \mathbb{E}_{D|Z=z}\left[\mathbb{E}\left(Y|D, Z = z\right)\right] = \mathbb{E}(Y|D = 0, Z = z)\mathbb{P}(D = 0|Z = z) \\ + \mathbb{E}(Y|D = 1, Z = z)\mathbb{P}(D = 1|Z = z).\end{aligned} \tag{5.3}$$

Now that we have expressed $\mathbb{E}(Y|Z)$ in terms of $\mathbb{E}(Y|D, Z)$ and $\mathbb{P}(D|Z)$, we can substitute our results from Lemma 5.1 and Lemma 5.2 into (5.3) to relate the numerator of the Wald Estimand to the average potential outcomes $\mathbb{E}(Y_j|T = t)$ of the different types and their prevalence $\mathbb{P}(T = t)$ in the population. In particular:

$$\begin{aligned}\mathbb{E}(Y|Z = 1) &= \mathbb{E}(Y|D = 0, Z = 1)\mathbb{P}(D = 0|Z = 1) + \mathbb{E}(Y|D = 1, Z = 1)\mathbb{P}(D = 1|Z = 1) \\ &= \mathbb{P}(T = n)\mathbb{E}(Y_0|T = n) + \left[\mathbb{P}(T = a)\mathbb{E}(Y_1|T = a) + \mathbb{P}(T = c)\mathbb{E}(Y_1|T = c)\right]\end{aligned}$$

and similarly for $Z = 0$,

$$\begin{aligned}\mathbb{E}(Y|Z = 0) &= \mathbb{E}(Y|D = 0, Z = 0)\mathbb{P}(D = 0|Z = 0) + \mathbb{E}(Y|D = 1, Z = 0)\mathbb{P}(D = 1|Z = 0) \\ &= \left[\mathbb{P}(T = n)\mathbb{E}(Y_0|T = n) + \mathbb{P}(T = c)\mathbb{E}(Y_0|T = c)\right] + \mathbb{P}(T = a)\mathbb{E}(Y_1|T = a).\end{aligned}$$

Taking the differences of these two expressions, we find that

$$\mathbb{E}(Y|Z = 1) - \mathbb{E}(Y|Z = 0) = \mathbb{P}(T = c)\mathbb{E}(Y_1 - Y_0|T = c) \tag{5.4}$$

since $\mathbb{E}(Y_1|T = c) - \mathbb{E}(Y_0|T = c) = \mathbb{E}(Y_1 - Y_0|T = c)$ by the Linearity of Conditional Expectation. Now consider the denominator of the Wald Estimand. Since D is binary,

$\mathbb{E}(D|Z=z) = \mathbb{P}(D=1|Z=z)$. Thus, by Lemma 5.1 we obtain

$$\mathbb{E}(D|Z=1) - \mathbb{E}(D|Z=0) = \mathbb{P}(D=1|Z=1) - \mathbb{P}(D=1|Z=0) = \mathbb{P}(T=c). \quad (5.5)$$

Since $\mathbb{P}(T=c) \neq 0$ we can divide (5.4) by (5.5), completing the proof. □

Theorem 5.1 shows that IV does not in general identify the ATE in a heterogenous treatment effects setting. Instead, it identifies the average treatment effect *for compliers*. In the MTO intervention, for example, IV would identify the average causal effect of moving to a low-poverty neighborhood for the kind of family that could be *induced* to move by offering them a voucher. This average effect for compliers is typically called the **local average treatment effect** or **LATE** for short. Accordingly, Theorem 5.1 is sometimes called the LATE Theorem and Assumptions 5.1–5.4 the LATE assumptions. We discuss the interpretation of LATE in section 5.5 below.

5.5 Who are the compliers? Why should we care?

As shown in section 5.4, IV does *not* identify the ATE in a world of heterogeneous treatment effects. Instead it identifies the LATE: an average treatment effect for a particular *subset* of individuals, namely the compliers. So who are these compliers, and why should we care about them? The short answer is: we don't know, and it is unclear whether we should. The key point to recognize is this: IV came first, and LATE came second. No one sat down and asked "how can I recover the average treatment effect for the compliers?" Instead the question was "suppose I run IV in a world of heterogeneous effects. Is there any way to give an interpretation to the result?" The answer, as we have seen, is *yes* but the interpretation is somewhat strained for several reasons. First, we cannot point to any individual in the sample and say "this is a complier." A person with $(Z=1, D=1)$ could be a complier or an always-taker; a person with $(Z=0, D=0)$ could be a complier or a never-taker. Second, as discussed in section 5.5 above, compliance is only defined relative to a *particular* instrumental variable. In the MTO intervention, families were offered a rent voucher that could be used to move to a low-poverty area. The LATE is specific not only to the fact that the IV was a rent voucher—as opposed, say, to an offer of assistance searching for an apartment—but also to the precise *amount* of the voucher. A \$500 per month rent voucher, for example, identifies a different LATE from a \$600 per month rent voucher if there are any families that would choose to move when offered the larger amount but not the smaller one.

While we cannot identify individual compliers, it is possible to say something about the so-called **compliant sub-population**, i.e. the population of compliers. First, as seen from (5.2), the denominator of the Wald estimator tells us the fraction of compliers in the population. If this fraction is large, the fact that IV identifies a LATE rather than

the ATE is less worrying: when nearly everyone is a complier, the two causal effects will likely be quite similar. Second, we can estimate the *average* characteristics of compliers. Suppose, for example, that we wanted to determine whether compliers are more likely to be women. Let W be a dummy variable that equals one if a person is female, zero otherwise. Then $\mathbb{P}(W = 1|T = c)$ is the share of women among compliers. By Bayes' Theorem, we obtain

$$\mathbb{P}(W = 1|T = c) = \frac{\mathbb{P}(T = c|W = 1)\mathbb{P}(W = 1)}{\mathbb{P}(T = c)} = \frac{\mathbb{P}(T = c|W = 1)\mathbb{P}(W = 1)}{\mathbb{E}(D|Z = 1) - \mathbb{E}(D|Z = 0)} \tag{5.6}$$

using $\mathbb{P}(T = c) = \mathbb{E}(D|Z = 1) - \mathbb{E}(D|Z = 0)$ from Lemma 5.1. Now, $\mathbb{P}(W = 1)$ is simply the share of women in the population but what about $\mathbb{P}(T = c|W = 1)$? By an argument nearly identical to the proof of Lemma 5.1, only with additional conditioning on $W = 1$, we can show that

$$\mathbb{P}(T = c|W = 1) = \mathbb{E}(D|Z = 1, W = 1) - \mathbb{E}(D|Z = 0, W = 1). \tag{5.7}$$

Combining (5.6) and (5.7)

$$\mathbb{P}(W = 1|T = c) = \mathbb{P}(W = 1)\left[\frac{\mathbb{E}(D|Z = 1, W = 1) - \mathbb{E}(D|Z = 0, W = 1)}{\mathbb{E}(D|Z = 1) - \mathbb{E}(D|Z = 0)}\right].$$

Since all of the quantities on the right-hand-side of this equation are observable, even though we cannot tell *which* women are compliers, we can nevertheless identify the share of women among compliers.

5.6 LATE Appendix: Proofs

Proof of Lemma 5.1. By the Law of Total Probability

$$\begin{aligned}\mathbb{P}(D = d|Z = z) &= \sum_{t\in\{a,c,d,n\}} \mathbb{P}(D = d|Z = z, T = t)\mathbb{P}(T = t|Z = z) \\ &= \sum_{t\in\{a,c,n\}} \mathbb{P}(D = d|Z = z, T = t)\mathbb{P}(T = t)\end{aligned}$$

since $\mathbb{P}(T = t|Z = z) = P(T = t)$ by Assumption 5.1 and $\mathbb{P}(T = d) = 0$ by Assumption 5.2. The key to the rest of the argument is that D is *completely determined* by Z and T: if I know that your offer is $Z = z$ and and your type is $T = t$, then I know with certainty what your take-up decision will be. It follows that $\mathbb{P}(D = d|Z = z, T = t)$ is either zero or one, depending on the values of (d, z, t).

Suppose first that $Z = 1$ but $D = 0$. Because you did not take up treatment, you cannot be an always-taker. Moreover, because $D \neq Z$ you cannot be a complier. Hence

$\mathbb{P}(D=0|Z=1,T=c)=\mathbb{P}(D=0|Z=1,T=a)=0$. If you were a never-taker, then given $Z=1$ you would indeed have $D=0$: $\mathbb{P}(D=0|Z=1,T=n)=1$. Therefore, we see that:

$$\begin{aligned}\mathbb{P}(D=0|Z=1) &= 0\times\mathbb{P}(T=a)+0\times\mathbb{P}(T=c)+1\times\mathbb{P}(T=n)\\ &= \mathbb{P}(T=n).\end{aligned}$$

Now suppose that $Z=0$ but $D=1$. Because you took up treatment, you cannot be a never-taker. Moreover, because $Z\neq D$ you cannot be a complier. As a result, we see that $\mathbb{P}(D=1|Z=0,T=c)=\mathbb{P}(D=1|Z=0,T=n)=0$. If you were an always-taker, you would indeed have $D=1$: $\mathbb{P}(D=1|Z=1,T=a)=1$. Hence,

$$\begin{aligned}\mathbb{P}(D=1|Z=0) &= 1\times\mathbb{P}(T=a)+0\times\mathbb{P}(T=c)+0\times\mathbb{P}(T=n)\\ &= \mathbb{P}(T=a).\end{aligned}$$

Next suppose that $Z=1$ and $D=1$. Because you took treatment, we know that you cannot be a never-taker: $\mathbb{P}(D=1|Z=1,T=n)=0$. You could, however, be a complier or an always-taker: $\mathbb{P}(D=1|Z=1,T=c)=\mathbb{P}(D=1|Z=1,T=a)=1$. Hence,

$$\begin{aligned}\mathbb{P}(D=1|Z=1) &= 1\times\mathbb{P}(T=a)+1\times\mathbb{P}(T=c)+0\times\mathbb{P}(T=n)\\ &= \mathbb{P}(T=a)+\mathbb{P}(T=c).\end{aligned}$$

Finally, suppose that $Z=0$ and $D=0$. Because you did not take up treatment, you cannot be an always-taker: $\mathbb{P}(D=0|Z=0,T=a)=0$. You could, however, be a never-taker or a complier: $\mathbb{P}(D=0|Z=0,T=n)=\mathbb{P}(D=0|Z=0,T=c)=1$. Hence,

$$\begin{aligned}\mathbb{P}(D=0|Z=0) &= 0\times\mathbb{P}(T=a)+1\times\mathbb{P}(T=c)+1\times\mathbb{P}(T=n)\\ &= \mathbb{P}(T=c)+\mathbb{P}(T=n)\end{aligned}$$

completing the proof. □

Proof of Lemma 5.2. By the Law of Iterated Expectations,

$$\begin{aligned}\mathbb{E}[Y|D=d,Z=z] &= \mathbb{E}_{T|(D=d,Z=z)}\left[\mathbb{E}\left(Y|D=d,Z=z,T\right)\right]\\ &= \sum_{t\in\{a,c,n\}}\mathbb{E}\left(Y|D=d,Z=z,T=t\right)\mathbb{P}(T=t|D=d,Z=z)\end{aligned}\tag{5.8}$$

since $\mathbb{P}(T=d)=0$ by Assumption 5.2. In the proof of Lemma 5.1 above, we examined the probabilities $\mathbb{P}(D=d|Z=z,T=t)$ in detail, arguing that they must be either zero or one, depending on whether the take-up and offer combination $(D=d,Z=z)$

is compatible with the type $T = t$. In contrast, the present argument involves $\mathbb{P}(T = t|D = d, Z = z)$. Fortunately, the two probabilities are related by the conditional version of Bayes' Theorem. In particular, using the fact that $\mathbb{P}(T = t|Z = z) = \mathbb{P}(T = t)$ by Assumption 5.1, we have

$$\mathbb{P}(T = t|D = d, Z = z) = \frac{\mathbb{P}(D = d|Z = z, T = t)\mathbb{P}(T = t)}{\mathbb{P}(D = d|Z = z)}. \tag{5.9}$$

While it may appear that we have made things more complicated rather than less, the preceding equality is actually very useful: in Lemma 5.1 we have already shown that the denominator is a sum of type probabilities $\mathbb{P}(T = t)$. Moreover, as argued above, the first term in the numerator is either zero or one depending on the values of (d, t, z).

Before we can combine all of these ingredients, however, we first need to take a closer look at the expectation $\mathbb{E}(Y|D = d, Z = z, T = t)$. Recall from (1.2) that $Y = DY_1 + (1 - D)Y_0$. Taking conditional expectations of this equality, we have

$$\begin{aligned}\mathbb{E}[Y|D = 0, Z = z, T = t] &= \mathbb{E}[Y_0|D = 0, Z = z, T = t] \\ \mathbb{E}[Y|D = 1, Z = z, T = t] &= \mathbb{E}[Y_1|D = 1, Z = z, T = t]\end{aligned}$$

A key idea in our proof of Lemma 5.1 above was that, given the way we have defined T, knowledge of a person's type t and her treatment offer z immediately implies her take-up decision d. In other words, D is a *function* of Z and T. For this reason, conditioning on D *in addition* to Z and T is *redundant*: a person's take-up decision cannot provide us with any further information given that we already know her treatment offer and type. It follows that we can drop D from the conditioning set in the preceding pair of equalities, yielding

$$\begin{aligned}\mathbb{E}[Y|D = 0, Z = z, T = t] &= \mathbb{E}[Y_0|Z = z, T = t] \\ \mathbb{E}[Y|D = 1, Z = z, T = t] &= \mathbb{E}[Y_1|Z = z, T = t].\end{aligned}$$

There is one more simplification that we can apply to the expressions for $\mathbb{E}[Y|D, Z, T]$. Assumption 5.3 states that the conditional mean of Y_0 and Y_1 does not depend on Z after conditioning on T: in other words the average potential outcomes for any type of individual (a, c, n) are unaffected by her treatment offer. Imposing this restriction, the preceding equalities become

$$\mathbb{E}[Y|D = 0, Z = z, T = t] = \mathbb{E}[Y_0|T = t] \tag{5.10}$$

$$\mathbb{E}[Y|D = 1, Z = z, T = t] = \mathbb{E}[Y_1|T = t]. \tag{5.11}$$

The remainder of the argument, though admittedly somewhat tedious, is just algebra.

For each pair of values (d, z) we first substitute either (5.10) or (5.11) into (5.8), depending on whether $D = 0$ or $D = 1$. We then substitute Lemma 5.1 into (5.9), and the result into (5.8).

Consider first $\mathbb{E}[Y|D = 0, Z = 1]$. Recall from the proof of Lemma 5.1 above that the probability $\mathbb{P}(D = 0|Z = 1, T = t)$ equals one for $T = n$ and zero for all other types $T = t$. (If you do not take when offered, you must be either a never-taker or defier, but we have assumed that there are no defiers.) By Equation 5.9, this implies that $\mathbb{P}(T = t|D = 0, Z = 1)$ equals zero for any $T \neq n$. Hence, substituting Equation 5.9 and (5.10) into (5.8), we obtain

$$\mathbb{E}[Y|D = 0, Z = 1] = \mathbb{E}[Y_0|T = n] \frac{\mathbb{P}(T = n)}{\mathbb{P}(D = 0|Z = 1)}$$

but since $\mathbb{P}(D = 0|Z = 1) = \mathbb{P}(T = n)$ by Lemma 5.1, the numerator and denominator cancel, leaving us with $\mathbb{E}[Y|D = 0, Z = 1] = \mathbb{E}[Y_0|T = n]$. Nearly identical reasoning for the case in which $(D = 1, Z = 0)$ gives $\mathbb{E}[Y|D = 1, Z = 0] = \mathbb{E}[Y_1|T = a]$, using the fact that anyone who takes up treatment when *not* offered must be an always-taker, given that we have assumed that there are no defiers.

Now consider $\mathbb{E}[Y|D = 1, Z = 1]$. Recall from the proof of Lemma 5.1 above that the probability $\mathbb{P}(D = 1|Z = 1, T = t)$ equals one for $T = c$ and $T = a$ but zero for $T = n$. By Equation 5.9, this implies that $\mathbb{P}(T = n|D = 0, Z = 1)$ equals zero. Hence, substituting Equation 5.9 and (5.10) into (5.8), we obtain

$$\mathbb{E}[Y|D = 1, Z = 1] = \mathbb{E}[Y_1|T = a] \frac{\mathbb{P}(T = a)}{\mathbb{P}(D = 1|Z = 1)} + \mathbb{E}[Y_1|T = c] \frac{\mathbb{P}(T = c)}{\mathbb{P}(D = 1|Z = 1)}.$$

The desired result follows since $\mathbb{P}(D = 1|Z = 1) = \mathbb{P}(T = a) + \mathbb{P}(T = c)$ by Lemma 5.1. A nearly identical argument gives the required expression for $\mathbb{E}[Y|D = 0, Z = 0]$, with never-takers replacing always-takers. □

Chapter 6

Testing the LATE Assumptions

A common criticism of instrumental variables approaches is that they merely replace one untestable assumption–selection on observables–with another, instrument exogeneity. There's something to this argument. On the one hand, it's easy to find published papers that use dubious instruments to produce even more dubious results. On the other hand, all causal inference relies on assumptions. And as we will see in this chapter, it's not *quite* correct to say that instrument exogeneity is untestable. Unlike the textbook, just-identified, homogeneous effects IV model, The LATE model introduced in chapter 5 *does* have testable implications. We can use them both to screen out particularly bad instruments and to gain a deeper understanding of the LATE model.

6.1 Instrument Exogeneity in the Textbook IV Model

Recall the "textbook" homogeneous effects model from section 5.1: $Y = \alpha + \beta D + U$ where $\text{Cov}(D, U)$ may not be zero. In this model, the treatment effect β is homogeneous: $\beta = Y_1 - Y_0$ is *constant*, so everyone has the same treatment effect. Given a single instrument Z, the IV estimand is

$$\beta_{\text{IV}} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, D)} = \frac{\beta\text{Cov}(Z, D) + \text{Cov}(Z, U)}{\text{Cov}(Z, D)} = \beta + \frac{\text{Cov}(Z, U)}{\text{Cov}(Z, D)}.$$

In order for β_{IV} to equal the treatment effect of interest, β, the instrument Z must be **exogenous**: $\text{Cov}(Z, U) = 0$. In order for the IV estimand to *exist* in the first place, Z must also be a **relevant**: $\text{Cov}(Z, D) \neq 0$. If Z is both relevant and exogenous, we say that it is a **valid instrument**.

Since D and Z are observed, instrument relevance is testable. We can simply regress D on Z and carry out an F-test for the significance of the regression. From the perspective of identification, $\text{Cov}(D, Z)$ can be as close to zero as we like. So long as it is not *exactly zero* the IV estimand is well-defined. In practice however, IV estimation and inference will

go haywire if the first stage relationship between D and Z is *weak*, $|\text{Cov}(Z, D)|$ is small, a problem called **weak instruments**. We will not discuss this issue further here, but it is important for you to be aware of, regardless of whether your interests are primarily theoretical or applied.

Unlike (Y, D, Z), the error term U is unobserved so we cannot estimate $\text{Cov}(Z, U)$. Wait a minute, you might say, couldn't we use the IV residuals $\widehat{U}_i \equiv (Y_i - \bar{Y}) - \widehat{\beta}_{\text{IV}}(D_i - \bar{D})$ to "stand in" for U and check whether they are correlated with Z? Unfortunately this approach can never tell us anything about $\text{Cov}(Z, U)$. To see why, suppose that Z is in fact an *endogenous instrument*, i.e. that $\text{Cov}(Z, U) \neq 0$. In this case the IV estimand is still perfectly well-defined, it simply doesn't equal β:

$$\beta_{IV} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, D)} = \beta + \frac{\text{Cov}(Z, U)}{\text{Cov}(Z, D)}, \quad \alpha_{IV} = \mathbb{E}(Y) - \beta_{IV}\mathbb{E}(D).$$

Now, let V be the IV *residual*: $V \equiv Y - \alpha_{IV} - \beta_{IV} D$. Note that V is *only* equal to U if Z is a valid instrument, because this is the only way that we can have $\beta_{IV} = \beta$ and $\alpha_{IV} = \alpha$. Using our definition of V, we can calculate $\text{Cov}(Z, V)$ as follows:

$$\begin{aligned}\text{Cov}(Z, V) &= \text{Cov}(Z, Y - \alpha_{IV} - \beta_{IV} D) = \text{Cov}(Z, Y) - \beta_{IV}\text{Cov}(Z, D) \\ &= \text{Cov}(Z, Y) - \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, D)}\text{Cov}(Z, D) = 0.\end{aligned}$$

In other words, Z is always perfectly uncorrelated with the IV residual V *by construction*, regardless of whether Z is correlated with the structural error U.

Without further information, there is no way to test instrument exogeneity in this model. So where might we obtain additional information? One possibility is to consider multiple instruments, as described in the next section. Another is to incorporate *prior information* about other features of the model. When presenting IV results, applied researchers often discuss the likely direction of selection bias and the extent of measurement error in their regressor of interest. By combining this information with the observed data, it is sometimes possible to conclude that the proposed instrumental variable cannot be exogenous, even in a just-identified, homogeneous linear model like the one described above. See DiTraglia and García-Jimeno (2021) for details.

6.2 Multiple Instruments and Over-identification

You may recall from your earlier econometrics training that it is possible to test the *joint exogeneity* of multiple instrumental variables in a linear, homogeneous effects model.[1] This is called a **test of over-identifying restrictions**, and is a special case of the J-

[1] In case you're rusty on the terminology: "just-identified" means that there are as many instrumental variables as endogenous regressors; "over-identified" means that there are more.

test from the theory of generalized method of moments (GMM) estimation. The basic idea is as follows. Let $Y = \alpha + \beta D + U$ as above, but now suppose that we have two relevant instruments Z_1 and Z_2, i.e. $\text{Cov}(Z_1, D) \neq 0$ and $\text{Cov}(Z_2, D) \neq 0$. Now define two IV estimands: one that uses Z_1 to instrument for D and another that uses Z_2, namely

$$\beta_{IV}^{(1)} \equiv \frac{\text{Cov}(Z_1, Y)}{\text{Cov}(Z_1, D)} = \beta + \frac{\text{Cov}(Z_1, U)}{\text{Cov}(Z_1, D)}, \quad \beta_{IV}^{(2)} \equiv \frac{\text{Cov}(Z_2, Y)}{\text{Cov}(Z_2, D)} = \beta + \frac{\text{Cov}(Z_2, U)}{\text{Cov}(Z_2, D)}.$$

Taking differences of the two estimands, we obtain

$$\beta_{IV}^{(1)} - \beta_{IV}^{(2)} = \frac{\text{Cov}(Z_1, U)}{\text{Cov}(Z_1, D)} - \frac{\text{Cov}(Z_2, U)}{\text{Cov}(Z_2, D)}.$$

If both Z_1 and Z_2 are exogenous, then $\text{Cov}(Z_1, U) = \text{Cov}(Z_2, U) = 0$ implying that $\beta_{IV}^{(1)} = \beta_{IV}^{(2)}$. Therefore, if $\beta_{IV}^{(1)}$ and $\beta_{IV}^{(2)}$ are *not equal* then at least one of the instruments (Z_1, Z_2) must be endogenous. While it is formulated in a slightly different way, a test of over-identifying restrictions exploits this basic intuition to evaluate the *joint* null hypothesis that both instruments are valid: $\text{Cov}(Z_1, U) = \text{Cov}(Z_2, U) = 0$. This example concerns two instruments in a model with a single endogenous regressor, but the same idea applies whenever there are more instruments than endogenous regressors or, more generally, when there are more moment conditions than parameters.

As we have seen, the basic idea behind a test of overidentifying restrictions is that two different instruments should both identify the same parameter, namely β. If they disagree, this indicates that one of our assumptions must be incorrect. The reasoning from above relies crucially on the assumption of homogeneous effects: $Y = \alpha + \beta D + U$ posits the same treatment effect β for everyone in the population. Unfortunately this logic does not carry over to a LATE setting. In a world of heterogeneous treatment effects, Theorem 5.1 shows that the Wald estimand recovers the average treatment effect for compliers, the subset of people who people who only take treatment when offered. Crucially, compliance is only defined *relative to a particular instrument*. If two researchers study the same question using two different instruments, there is no reason to suppose that their LATEs will coincide. For this reason, comparing LATEs across instruments does not provide a test of the LATE assumptions.

Consider two researchers, Alice and Bob, studying the effect of quitting smoking on birthweight. Each researcher recruits a random sample of subjects from the same population—newly-pregnant British women between the ages of 18 and 30 who regularly smoke cigarettes. Alice and Bob follow identical research protocols. In each study half of the subjects are offered an inducement $(Z = 1)$ intended to encourage quitting smoking, while the other half are not $(Z = 0)$. The researchers then record which mothers successfully quit smoking $(D = 1)$ and which do not $(D = 0)$ along with the birthweight of their babies (Y). The only difference between the two studies is the choice of Z. Alice's

instrument is a counseling session that teaches the dangers of smoking during pregnancy. Accordingly, Alice's LATE is the effect of quitting smoking on birthweight for the sub-population of mothers who can be induced to quit by attending a counseling session. In contrast Bob's instrument is an offer to pay £500 to those who successfully quit smoking. His LATE is the effect of quitting smoking on birthweight for the sub-population of mothers who can be induced to quit by offering to pay £500. Because these subpopulations almost certainly differ, it will come as no surprise that the LATEs could as well.

6.3 We can sometimes reject the LATE model.

In the preceding section, we saw that the over-identifying restrictions test can't help us when treatment effects are heterogeneous. The LATE model has "extra degrees of freedom" in that it allows $(Y_1 - Y_0)$ to vary across individuals. This means that two equally valid instruments could yield different IV estimands. Unlike the just-identified, linear, homogeneous IV model from above, however, the LATE model from chapter 5 has testable implications beyond $\text{Cov}(D, Z) \neq 0$. Huber and Mellace (2015) show that Assumptions 5.1–5.3 from chapter 5 imply four inequalities of the form

$$\theta_1 \leq 0, \quad \theta_2 \leq 0, \quad \theta_3 \leq 0, \quad \theta_4 \leq 0 \tag{6.1}$$

where $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \theta_3, \theta_4)$ is a vector of parameters that we will define in a moment. The key point for now is that $\boldsymbol{\theta}$ can be *directly calculated* from observations of (Y, D, Z). The inequalities from (6.1) provide a test of the LATE restrictions. If any of the elements of $\boldsymbol{\theta}$ is positive then we know that at least one of assumptions 5.1–5.3 must be false. In practice, we would compare estimated parameters $\widehat{\boldsymbol{\theta}}$ with appropriate standard errors to discern whether any observed violation of the inequalities is statistically significant. While any violation of (6.1) constitutes a violation of the LATE assumptions, not all violations of the LATE assumptions will lead to a violation of (6.1). As such, the testable implications from Huber and Mellace (2015) are **necessary but not sufficient** for the validity of the LATE model.

6.3.1 Describing the Inequalities

In this section we'll define the parameters $\theta_1, \theta_2, \theta_3$, and θ_4 introduced above and *describe* the associated inequalities. In the next section we'll show why they hold. The inequalities from (6.1) are best considered in pairs. The first two, $\theta_1 \leq 0$ and $\theta_2 \leq 0$, arise from the following lemma.

Lemma 6.1. *Let F_{11} be the conditional CDF of $Y|(D=1,Z=1)$ and define*

$$y_q \equiv F_{11}^{-1}(q), \quad y_{1-q} \equiv F_{11}^{-1}(1-q), \quad q \equiv \frac{\mathbb{P}(D=1|Z=0)}{\mathbb{P}(D=1|Z=1)}.$$

Then, under Assumptions 5.1–5.3,

$$\mathbb{E}(Y|D=1,Z=1,Y\leq y_q) \leq \mathbb{E}(Y|D=1,Z=0) \leq \mathbb{E}(Y|D=1,Z=1,Y\geq y_{1-q}).$$

Lemma 6.1 contains two inequalities, both of which involve expectations of Y conditional on $D=1$. By Lemma 5.2, $\mathbb{E}(Y|D=1,Z=0) = \mathbb{E}(Y_1|T=a)$ so we have

$$\mathbb{E}(Y|D=1,Z=1,Y\leq y_q) \leq \mathbb{E}(Y_1|T=a) \leq \mathbb{E}(Y|D=1,Z=1,Y\geq y_{1-q}). \tag{6.2}$$

In other words, Lemma 6.1 provides upper and lower bounds for the average value of Y_1 among always-takers. The lower bound equals the mean of the bottom $q\times 100\%$ of the distribution of $Y|(D=1,Z=1)$ and the upper bound equals the mean of the top $q\times 100\%$. It's natural to ask: why bother? After all, $\mathbb{E}(Y|D=1,Z=0)$ point identifies $\mathbb{E}(Y_1|T=a)$ based on the observed data (Y,D,Z). The crucial observation is that the bounds from Lemma 6.1 rely on information that was *not used* to construct (5.1), the Wald estimand. As such, they function like "overidentifying restrictions." If the LATE assumptions are correct, $\mathbb{E}(Y|D=1,Z=0)$ indeed identifies $\mathbb{E}(Y_1|T=a)$ and hence must lie within the bounds. If it does not, at least one of our assumptions must be false. To express Lemma 6.1 in the form $\theta_1\leq 0$ and $\theta_2\leq 0$, subtract the right-hand side of each from the left-hand side, defining

$$\begin{aligned}\theta_1 &\equiv \mathbb{E}(Y|D=1,Z=1,Y\leq y_q) - \mathbb{E}(Y|D=1,Z=0)\\ \theta_2 &\equiv \mathbb{E}(Y|D=1,Z=0) - \mathbb{E}(Y|D=1,Z=1,Y\geq y_{1-q}).\end{aligned}$$

From these definitions we see that **at most one** of the bounds $\theta_1\leq 0$ and $\theta_2\leq 0$ can be violated in any given example. If $\mathbb{E}(Y|D=1,Z=0)$ does not lie within the bounds from Lemma 6.1, then it either exceeds the upper bound, in which case it satisfies the lower bound, or it falls short of the lower bound, in which case it satisfies the upper bound. The following lemma gives analogous bounds for $\mathbb{E}(Y|D=0,Z=1)$.

Lemma 6.2. *Let F_{00} be the conditional CDF of $Y|(D=0,Z=0)$ and define*

$$y_r \equiv F_{00}^{-1}(r), \quad y_{1-r} \equiv F_{00}^{-1}(1-r), \quad r \equiv \frac{\mathbb{P}(D=0|Z=1)}{\mathbb{P}(D=0|Z=0)}.$$

Then, under Assumptions 5.1–5.3,

$$\mathbb{E}(Y|D=0,Z=0,Y\leq y_r) \leq \mathbb{E}(Y|D=0,Z=1) \leq \mathbb{E}(Y|D=0,Z=0,Y\geq y_{1-r}).$$

Lemma 6.2 contains two inequalities, both of which involve expectations of Y conditional on $D = 0$. By Lemma 5.2, $\mathbb{E}(Y|D = 0, Z = 0) = \mathbb{E}(Y_0|T = n)$ so we have

$$\mathbb{E}(Y|D = 0, Z = 0, Y \leq y_r) \leq \mathbb{E}(Y_0|T = n) \leq \mathbb{E}(Y|D = 0, Z = 0, Y \geq y_{1-r}). \tag{6.3}$$

In other words, Lemma 6.1 provides upper and lower bounds for the average value of Y_0 among never-takers. The lower bound equals the mean of the bottom $r \times 100\%$ of the distribution of $Y|(D = 1, Z = 1)$ and the upper bound equals the mean of the top $r \times 100\%$. If the LATE assumptions are correct, $\mathbb{E}(Y|D = 0, Z = 1)$ indeed identifies $\mathbb{E}(Y_0|T = n)$ and hence must lie within the bounds. If it does not, at least one of our assumptions must be false. We convert Lemma 6.1 into the pair of inequalities $\theta_3 \leq 0$ and $\theta_4 \leq 0$ by defining

$$\begin{aligned}
\theta_3 &\equiv \mathbb{E}(Y|D = 0, Z = 0, Y \leq y_r) - \mathbb{E}(Y|D = 0, Z = 1) \\
\theta_4 &\equiv \mathbb{E}(Y|D = 0, Z = 1) - \mathbb{E}(Y|D = 0, Z = 0, Y \geq y_{1-r}).
\end{aligned}$$

At most one of the inequalities $\theta_3 \leq 0$ and $\theta_4 \leq 0$ can be violated in a given example. For convenience, we collect Lemmas 6.1 and 6.2 in the following result.

Theorem 6.1. *Under Assumptions 5.1–5.3,*

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} \equiv \begin{bmatrix} \mathbb{E}(Y|D = 1, Z = 1, Y \leq y_q) - \mathbb{E}(Y|D = 1, Z = 0) \\ \mathbb{E}(Y|D = 1, Z = 0) - \mathbb{E}(Y|D = 1, Z = 1, Y \geq y_{1-q}) \\ \mathbb{E}(Y|D = 0, Z = 0, Y \leq y_r) - \mathbb{E}(Y|D = 0, Z = 1) \\ \mathbb{E}(Y|D = 0, Z = 1) - \mathbb{E}(Y|D = 0, Z = 0, Y \geq y_{1-r}) \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{6.4}$$

where y_q, y_{1-q} are as defined in Lemma 6.1 and y_r, y_{1-r} as defined in Lemma 6.2.

6.3.2 Deriving the Bounds

We'll begin by deriving (6.2). The derivation of (6.3) is nearly identical, so we'll only discuss it briefly at the end of this section. By Lemma 5.1, the probability q defined in Lemma 6.1 can be written as

$$q \equiv \frac{\mathbb{P}(D = 1|Z = 0)}{\mathbb{P}(D = 1|Z = 1)} = \frac{\mathbb{P}(T = a)}{\mathbb{P}(T = a) + \mathbb{P}(T = c)}. \tag{6.5}$$

Thus, q equals the share of always takers among those with $(D = 1, Z = 1)$ and $(1 - q)$ equals the share of compliers in the same group. Using this notation we can re-write the expression for $\mathbb{E}(Y|D = 1, Z = 1)$ from above as

$$\mathbb{E}\left(Y|D = 1, Z = 1\right) = (1 - q)\mathbb{E}\left(Y_1|T = c\right) + q\mathbb{E}\left(Y_1|T = a\right). \tag{6.6}$$

Thus, Assumptions 5.1–5.3 imply that $\mathbb{E}(Y|D=1,Z=1)$ is a **mixture** of $\mathbb{E}(Y_1|T=a)$ and $\mathbb{E}(Y_1|T=c)$ with mixing probability q. Notice that the left-hand side of (6.6) conditions on $Z=1$ while the right-hand side does not. This is because the average value of Y_1 for any compliance "type" is unaffected by Z, provided that Assumption 5.3 holds. In words: the instrument has **no direct effect** on the outcome. Again, anyone with $(D=1,Z=1)$ must either be an always-taker or a complier. If we knew *which* individuals in this group were always-takers, we could directly test Assumption 5.3 by comparing $\mathbb{E}(Y|Z=1,T=a)$ against $\mathbb{E}(Y|D=1,Z=0)$. Alas we cannot tell whether a given individual with $(D=1,Z=1)$ is an always-taker or a complier. What we can say, using the mixture idea from above, is that $q \times 100\%$ of the people in this group are always-takers and the rest are compliers.

Let's pick up this idea and run with it. Define the shorthand

$$\begin{aligned}
F(y) &\equiv \mathbb{P}(Y \leq y|D=1,Z=1) = \mathbb{P}(Y_1 \leq y|T \in \{a,c\},Z=1) \\
G(y) &\equiv \mathbb{P}(Y_1 \leq y|T=c,Z=1) \\
H(y) &\equiv \mathbb{P}(Y_1 \leq y|T=a,Z=1).
\end{aligned}$$

By the law of total probability and Assumptions 5.1–5.2, it follows that

$$\begin{aligned}
F(y) &= \mathbb{P}(T=c|T \in \{a,c\},Z=1)G(y) + \mathbb{P}(T=a|T \in \{a,c\},Z=1)H(y) \\
&= \mathbb{P}(T=a|T \in \{a,c\})G(y) + \mathbb{P}(T=c|T \in \{a,c\})H(y) \\
&= \frac{\mathbb{P}(T=c)}{\mathbb{P}(T \in \{a,c\})}G(y) + \frac{\mathbb{P}(T=a)}{\mathbb{P}(T \in \{a,c\})}H(y).
\end{aligned}$$

Substituting our definition of q from (6.5) gives

$$F = (1-q)G(y) + qH(y). \tag{6.7}$$

What (6.6) tells us about average outcomes for compliers and always-takers, (6.7) tells us about the corresponding *distributions* of Y_1. While neither G nor H observed–both depend on T, the unobserved compliance type–we know that they "mix together" in proportions $(1-q)$ and q to form the observed distribution F. Here is the crucial point: while H is defined conditional on $Z=1$, its *mean* must equal $\mathbb{E}(Y_1|T=a)$ under Assumption 5.3. Thus, our task is to find all possible values for the mean of H that are consistent with (6.7), F, and q. The result will be (6.2).

It's taken a fair amount of work to get to this point, but we've managed to reduce our problem to a completely abstract probability puzzle: if $F=(1-q)G+qH$ where q and F are known, what are all possible values for the mean of H? This is a question that crops up in a variety of econometric problems that have nothing to do with testing

the LATE assumptions.[2] As such, it's worth looking at the details. Solving (6.7),

$$H(y) = \left(\frac{1}{q}\right) F(y) - \left(\frac{1-q}{q}\right) G(y).$$

While q and $F(y)$ are known, we have no information about $G(y)$. But in order for it to be a valid CDF, it must lie between zero and one. Substituting $G(y) = 0$ and $G(y) = 1$ gives the following bounds for $H(y)$

$$\frac{F(y)}{q} - \frac{1-q}{q} \leq H(y) \leq \frac{F(y)}{q}.$$

Finally, since $H(y)$ must *itself* lie between zero and one, we obtain

$$\max\left\{0, \frac{F(y)}{q} - \frac{1-q}{q}\right\} \equiv \overline{H}(y) \leq H(y) \leq \underline{H}(y) \equiv \min\left\{1, \frac{F(y)}{q}\right\}. \tag{6.8}$$

Notice that $\underline{H}$ and $\overline{H}$ are *themselves* CDFs. Both are non-decreasing, approach zero as $y \to -\infty$, and approach one as $y \to \infty$. What's more, $\overline{H}$ first-order **stochastically dominates** H which in turn first-order stochastically dominates $\underline{H}$.[3] It follows that

$$\underbrace{\int_{\mathbb{R}} y\underline{H}(dy)}_{\underline{\mu}} \leq \underbrace{\int_{\mathbb{R}} yH(dy)}_{\mu} \leq \underbrace{\int_{\mathbb{R}} y\overline{H}(dy)}_{\overline{\mu}}. \tag{6.9}$$

In other words: the mean μ of H, a distribution we do not know, *must lie between* the mean $\underline{\mu}$ of $\underline{H}$ and the mean $\overline{\mu}$ of $\overline{H}$, two distributions that we do know! While I will not prove this here, these are in fact the best possible bounds for the mean of H based on the information provided.[4]

In fact, (6.9) is *precisely* the same thing as (6.2). To make this clearer, let's make a simplifying assumption: suppose that F is a continuous, strictly increasing CDF with probability density function f. This will allow us to write (6.9) in a more evocative form by working out the probability density functions that correspond to $\overline{H}$ and $\underline{H}$. To do this, we differentiate with respect to y while keeping track of the behavior of the min and max functions. Consider first $\overline{H}(y)$. This function equals zero until $F(y)$ exceeds $(1-q)$; thereafter it equals $[F(y) - (1-q)]/q$. Solving $F(y) = (1-q)$ for y, we see that $\overline{H}(y)$ is positive and strictly increasing for any y greater than $F^{-1}(1-q)$. Since

$$\frac{d}{dy}\left[\frac{F(y) - (1-q)}{q}\right] = \frac{F'(y)}{q} = \frac{f(y)}{q}$$

[2] See for example DiTraglia and Garcia-Jimeno (2019).

[3] Let X_1 and X_2 be random variables with CDFs F_1 and F_2. We say that X_1 first-order stochastically dominates X_2 if $F_1(x) \leq F_2(x)$ for all x or equivalently if $F_1^{-1}(x) \geq F_2^{-1}(x)$ for all x. Intuitively: any quantile of X_1, e.g. the median, is *at least as large* as the corresponding quantile of X_2.

[4] See Horowitz and Manski (1995) for more details.

it follows that the density function $\overline{h}(y)$ corresponding to $\overline{H}(y)$ is

$$\overline{h}(y) = 1\left\{y > F^{-1}(1-q)\right\}\frac{f(y)}{q}.$$

Written this way, we see that $\overline{h}$ is simply f truncated to the interval $[y_{1-q}, \infty)$ where y_{1-q} denotes the $(1-q)\times 100$ percentile of f. In other words, $\overline{h}$ is the density formed by **keeping only the top** $q \times 100\%$ of f and rescaling the result so it integrates to one. Now, $\underline{H}(y) \equiv \min\{1, F(y)/q\}$ equals $F(y)/q$ up to the point at which $F(y) = q$; for larger values of y it equals one. Hence, the density $\underline{h}(y)$ corresponding to $\underline{H}(y)$ is

$$\underline{h}(y) = 1\left\{y < F^{-1}(q)\right\}\frac{f(y)}{q}.$$

So we see that $\underline{h}$ is simply f truncated to the interval $(-\infty, y_q]$, where y_q denotes the $q \times 100$ percentile of f. In other words, $\underline{h}$ is the desnsity formed by **keeping only the bottom** $q \times 100\%$ of f and rescaling the result so it integrates to one. Finally, re-expressing (6.9) in terms of densities rather than CDFs, we obtain

$$\underline{\mu} \equiv \int_{-\infty}^{F^{-1}(q)} \frac{y}{q} f(y)\, dy \leq \mu \leq \int_{F^{-1}(1-q)}^{\infty} \frac{y}{q} f(y)\, dy \equiv \overline{\mu}. \tag{6.10}$$

Now it's clear that $\underline{\mu}$ and $\overline{\mu}$ are precisely the lower and upper bounds from (6.2).

To make things more concrete, let's look at an example. Figure 6.1 illustrates (6.10) in an example where $q = 0.4$ and f is a mixture of normals. The middle panel shows the density f with its bottom 40% shaded in red and top 40% shaded in blue. The top panel depicts $\underline{h}$, constructed by "cutting out" the bottom 40% of f and rescaling so that the result integrates to one The mean of this distribution is $\underline{\mu} = -2.69$. The bottom panel depicts $\overline{h}$, constructed by "cutting out" the top 40% of f and rescaling analogously. The mean of this distribution is $\overline{\mu} = 1.4$. These are our bounds for μ, the mean of H. From the figure we see that, all else equal, the larger the value of q the tighter the bounds. For $q > 0.5$, the red and blue shaded regions in the middle panel *overlap*. As q approaches one, they eventually *coincide*.

We haven't said a word about compliers or never-takers for nearly two pages! But in fact there's a simple and intuitive story behind everything we've just discussed. Recall that that F was defined as the CDF of $Y|(D = 1, Z = 1)$, H as the CDF of $Y_1|(T = a, Z = 1)$, and q as the share of always-takers among those with $(D = 1, Z = 1)$. In the example from Figure 6.1, 40% of the people who make up f are always-takers. Since we don't know where they fall in the distribution, we consider the two most extreme possibilities. At one extreme they could all be packed together in the **bottom** 40% of f; at the other, they could all be packed together in the **top** 40%. Accordingly, our

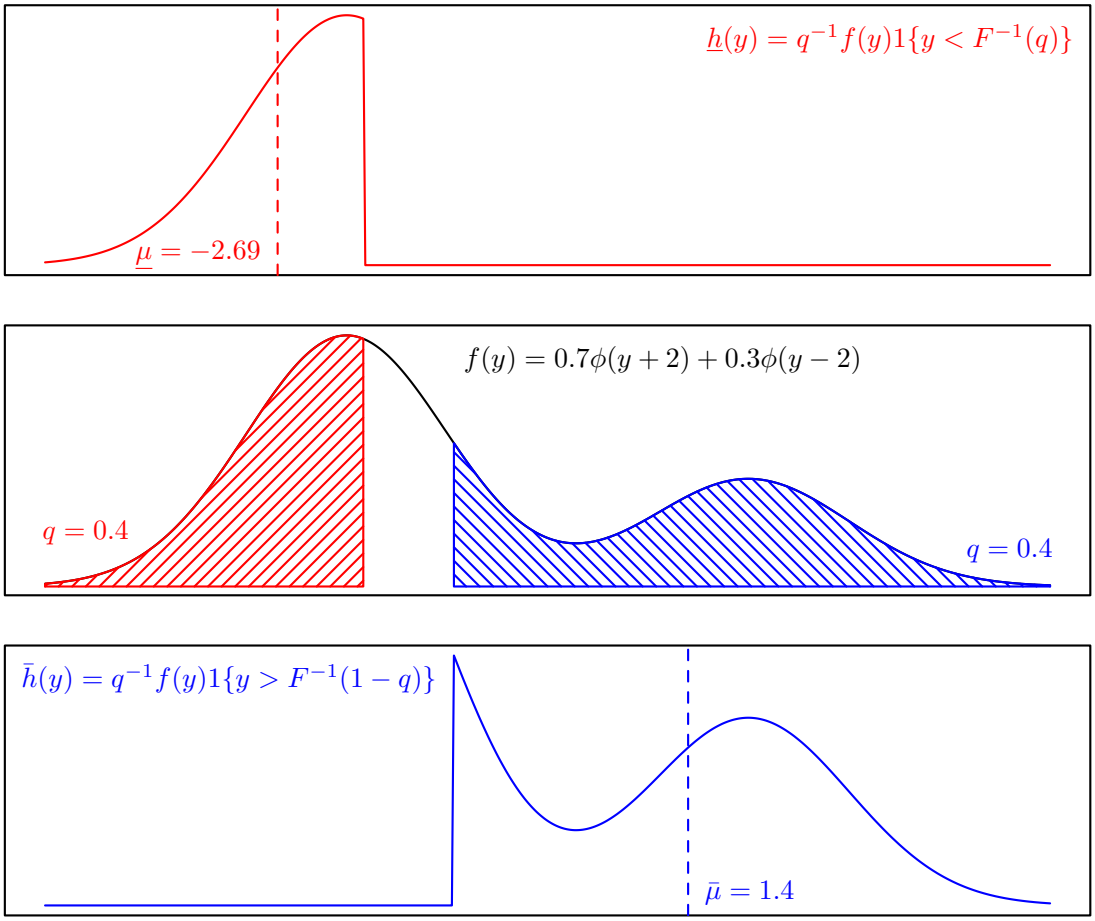


Figure 6.1: A numerical example illustrating (6.10). The middle panel depicts the density f, a mixture of standard normals with means -2 and 2 and corresponding mixing probabilities 0.7 and 0.3. The density $\underline{h}$ is constructed from the bottom 40% of f (top panel, red) while $\overline{h}$ is constructed form the top 40% of f (bottom panel, blue). Dashed lines indicate the means of these truncated distributions.

bounds are simply the average of the top and bottom 40% of f. The intuition and the mathematics agree, as they should! The larger the value of q, the smaller the share of compliers. All else equal, this yields *tighter bounds* at the cost of making the LATE a less interesting parameter, since it applies to only a very small subset of the population.

A nearly identical argument can be applied to derive (6.3). Here we rely upon the fact that $Y|(D=0,Z=0)$ is a mixture of Y_0 for compliers and never-takers while

$$r \equiv \frac{\mathbb{P}(D=0|Z=1)}{\mathbb{P}(D=0|Z=0)} = \frac{\mathbb{P}(T=n)}{\mathbb{P}(T=n)+\mathbb{P}(T=c)}$$

is the share of never-takers among those with $(D=1,Z=1)$, by Lemma 5.1. Adopting the shorthand $F(y) = \mathbb{P}(Y \leq y|D=0,Z=0)$, $G(y) \equiv \mathbb{P}(Y_0 \leq |T=c,Z=0)$, and $H(y) \equiv \mathbb{P}(Y_0|T=n,Z=0)$, the rest of the argument proceeds almost exactly as above.

6.3.3 Testing a Stronger Version of the LATE Model

The discussion from above follows Huber and Mellace (2015). Two closely-related papers that propose methods for testing the LATE model are Kitagawa (2015) and Mourifié and Wan (2017). The key distinction between these papers is the model that they test. Whereas Huber and Mellace (2015) derive testable implications of *precisely* the same assumptions that we used to identify the LATE in chapter 5, Kitagawa (2015) and Mourifié and Wan (2017) test a model that strengthens Assumption 5.3 from conditional mean independence to *full independence*.[5] While full independence is more than we need to point identify the LATE, it is difficult to think of an applied example in which mean independence is plausible but full independence is not. The advantage of assuming full independence is that it leads to additional testable implications of the LATE model, at the cost of somewhat greater complexity in implementation and exposition.

[5]See chapter 2 for an explanation of the difference.

Chapter 7

Regression Discontinuity

Selection bias arises when people are free to choose whether or not they are treated. An idealized randomized controlled trial eliminates this bias by removing the element of choice: subjects are *compelled* to take the treatment or to refrain from doing so. Yet experiments are not the only situations in which people's freedom to choose their own treatment is restricted. "Naturally occurring" constraints on self-selection are common and, under the right conditions, can provide a powerful tool for causal inference using observational data. Regression discontinuity methods exploit the existence of an administrative or legal cutoff that either completely or partially determines whether a person is treated. In the **sharp** regression discontinuity design, everyone on one side of the cutoff is treated and everyone on the other side is not, so the treatment variable "jumps" from zero to one at the cutoff. In a **fuzzy** regression discontinuity design, a person's *probability* of treatment jumps at the cutoff. Either way, the basic idea is to compare those who are slightly above the cutoff to those who are slightly below. In the sharp design we compare mean outcomes for data near the threshold; in the fuzzy design we calculate the Wald estimate in the same region. This chapter draws mainly on Hahn et al. (2001). For a detailed survey of regression discontinuity methods, see Lee and Lemieux (2010).

In the discussion below, we will use some notation that has not appeared in earlier chapters. First, $\mathbb{1}\{A\}$ denotes the **indicator function** of the event A. If A occurs, this function equals one; otherwise it equals zero. Second, $\lim_{x \downarrow c} f(x)$ and $\lim_{x \uparrow c} f(x)$ denote the **one-sided limits** of a function f "from above" and "from below," respectively. Some references call these the limits "from the right" and "from the left." We will tacitly assume that both one-sided limits exist, so that if f has a discontinuity at c, it is a **jump discontinuity**, also known as a discontinuity of the first kind. This means that $\lim_{x \downarrow c} f(x) \neq \lim_{x \uparrow c} f(x)$.

7.1 The Sharp Regression Discontinuity Design

The best way to understand the sharp regression discontinuity design, or sharp RD for short, is with an example. The following is based on Sekhri (2020). Quite unlike the situation in the US, the most prestigious colleges in India are public. Given the choice, effectively everyone would choose to attend a public rather than a private institution. Admissions to public colleges are based on a threshold rule: anyone whose senior secondary school exam score exceeds the appropriate cutoff is admitted and anyone whose score falls below this threshold is rejected. The cutoffs vary from year-to-year and subject-to-subject, but crucially applicants do not know the cutoff and all applications are evaluated blind. This situation is depicted in Figure 7.1. If your senior secondary school exam score X exceeds the admissions cutoff c, then you attend a public college ($D = 1$); if it does not, then you attend a less-prestigious private college ($D = 0$). In RD parlance, the variable X that determines a person's treatment status is called the **running variable**.

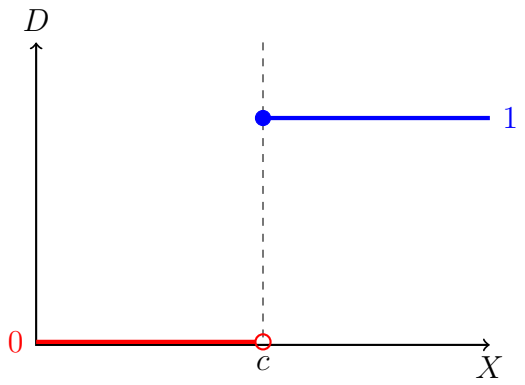


Figure 7.1: Treatments in the sharp regression discontinuity design. Anyone with $X \geq c$ is treated ($D = 1$) and no one with $X < c$ is treated ($D = 0$).

The key insight behind sharp RD is that students with exam scores just below the admissions cutoff are effectively identical to those with scores just above the cutoff. Whether you score 59.9% versus 60% on a exam is basically down to luck, but moving from just below the cutoff to just above it causes your treatment to jump discontinuously from zero to one. Hence, if we restrict our attention to students whose exam scores fall within a small window around the cutoff, it's *as if* some of them had been randomly assigned to attend an elite public college ($D = 1$) while the rest had been randomly assigned to attend a less-prestigious public college ($D = 0$). We formalize as follows.

Assumption 7.1 (Sharp RD Design)**.** $D = \mathbb{1}\{X \geq c\}$ *where* X *is an observed covariate, and* c *is a known threshold.*

Assumption 7.1 states that treatment is completely determined by the running variable X and a cutoff c that is known to the researcher. This is the situation depicted in

Figure 7.1. Note that this *implies* selection on observables (Assumption 4.1). We have:

$$\mathbb{E}[Y_j|D,X] = \mathbb{E}[Y_j|\mathbb{1}\{X \geq c\}, X] = \mathbb{E}[Y_j|X], \quad j = 0,1$$

because $\mathbb{1}\{X \geq c\}$ is a measurable function of X. So why can't we just use the methods from chapter 4 to identify the ATE? The basic idea from chapter 4 was to compare people with the same value of X but different values of D. For this to be possible, we required not just selection on observables, but *overlap*: Assumption 4.2. Under Assumption 7.1, however, overlap fails completely: because anyone with $X \geq c$ is treated and no one with $X < c$ is treated, we cannot compare people with different values of D but the same value of X. This is why we need a different approach.

Assumption 7.2 (Continuity of Conditional Means)**.** $\mathbb{E}[Y_0|X = x]$ *and* $\mathbb{E}[Y_1|X = x]$ *are both continuous functions of* x *at the point* $x = c$.

Assumption 7.2 formalizes the idea that students whose test scores fall in a sufficiently small neighborhood around the admissions cutoff are "effectively identical." Say that Y is wage. We would expect that both Y_0 and Y_1 depend on X. Regardless of whether you attend a public or private college, your wage is likely related to your secondary school test score. Assumption 7.2 allows the potential outcomes to be related to X. What it rules out is a "jump" in Y_0 or Y_1 as X moves from just below to just above the cutoff c. How could such a jump occur? Suppose that it were possible for students to "precisely manipulate" their test scores. Perhaps I know both my own likely score and the cutoff. If I'm a highly-motivated person and my expected score falls just below what I know to be the cutoff, then perhaps I might choose to work extremely hard in the weeks before the exam to boost my score. Or if I'm a highly unscrupulous person, perhaps I might try to cheat in some way. Either of these possibilities would lead to a systematic difference between people with X just below c and those with X just above c. If being highly-motivated, or highly unscrupulous, affects wages later in life, this would lead to a violation of Assumption 7.2. In the context of Sekhri (2020), it is unlikely that students can "precisely manipulate" their secondary school test scores in this way because admissions thresholds vary both year-to-year and subject-to-subject. The fact that diligence causes higher test scores as well as later-life outcomes does *not* constitute a violation of Assumption 7.2 unless diligence jumps discontinuously at the admissions cutoff. In effect, we require that D is the *only thing* that jumps at c.

Under Assumption 7.2, we can identify the conditional ATE at $X = c$ by studying the observed conditional mean function $\mathbb{E}[Y|X = x]$ close to the cutoff. Figure 7.2 gives the visual intuition. Since anyone with $X < c$ is untreated, $\mathbb{E}[Y|X = x]$ equals $\mathbb{E}[Y_0|X = x]$ for $x < c$. Similarly, since anyone with $X \geq c$ is treated, $\mathbb{E}[Y|X = x]$ equals $\mathbb{E}[Y_1|X = x]$ when $x \geq c$. By assumption, neither $\mathbb{E}[Y_0|X = x]$ nor $\mathbb{E}[Y_1|X = c]$ has a discontinuity at $X = c$. Hence, if $\mathbb{E}[Y|X = x]$ *does* have a discontinuity at this point, there must be a

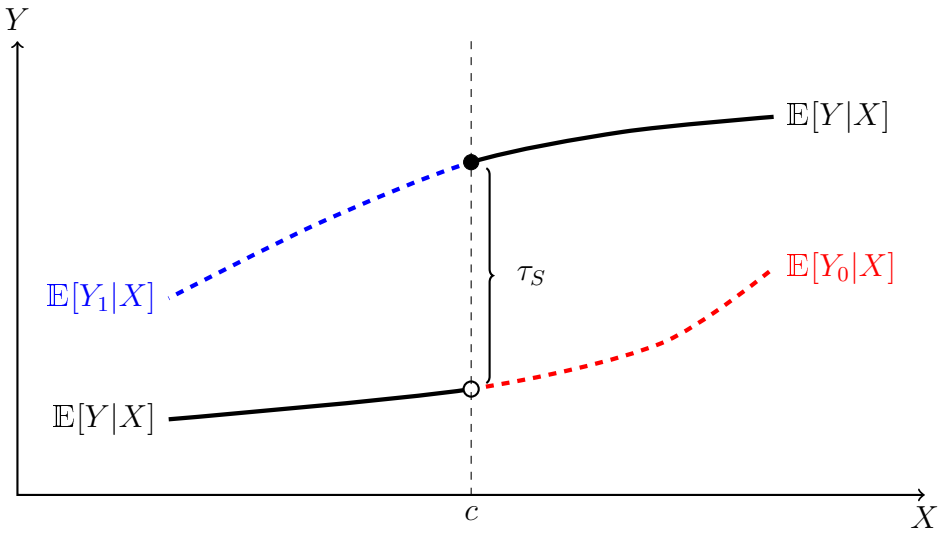


Figure 7.2: Outcomes in the sharp regression discontinuity design. For $X < c$ we observe $\mathbb{E}[Y|X] = \mathbb{E}[Y_0|X]$; for $X \geq c$ we observe $\mathbb{E}[Y|X] = \mathbb{E}[Y_1|X]$. The "jump" in $\mathbb{E}[Y|X]$ at c is the conditional ATE when $X = c$, namely $\tau_S \equiv \mathbb{E}[Y_1 - Y_0|X = c]$.

difference between $\mathbb{E}[Y_1|X = c]$ and $\mathbb{E}[Y_0|X = c]$. This difference, τ_S, is the conditional ATE at $X = c$. In the example from Sekhri (2020) discussed above, τ_S is the causal effect of attending an elite public institution on later-life wages for a certain kind of person: someone whose secondary school exam results are *close to the admissions cutoff*. This is a well-defined causal effect, but it may differ from the overall average treatment effect.

Theorem 7.1 (Sharp Regression Discontinuity)**.** *Under Assumptions 7.1 and 7.2,*

$$\tau_S \equiv \mathbb{E}[Y_1 - Y_0|X = c] = \lim_{x \downarrow c} \mathbb{E}[Y|X = x] - \lim_{x \uparrow c} \mathbb{E}[Y|X = x].$$

Proof of Theorem 7.1. By Assumption 7.1, we can re-write (1.2) as

$$Y = (1 - D)Y_0 + DY_1 = \mathbb{1}\{X < c\} Y_0 + \mathbb{1}\{X \geq c\} Y_1.$$

Taking conditional expectations of both sides, it follows that

$$\begin{aligned} \mathbb{E}[Y|X = x] &= \mathbb{E}\left[\mathbb{1}\{X < c\} Y_0 + \mathbb{1}\{X \geq c\} Y_1 | X = x\right] \\ &= \mathbb{1}\{x < c\} \mathbb{E}[Y_0|X = x] + \mathbb{1}\{x \geq c\} \mathbb{E}[Y_1|X = x]. \end{aligned}$$

Under Assumption 7.2, both $\mathbb{E}[Y_0|X = x]$ and $\mathbb{E}[Y_1|X = x]$ are continuous at c. Hence,

$$\begin{aligned} \lim_{x \uparrow c} \mathbb{E}[Y_0|X = x] &= \lim_{x \downarrow c} \mathbb{E}[Y_0|X = x] = \mathbb{E}[Y_0|X = c] \\ \lim_{x \uparrow c} \mathbb{E}[Y_1|X = x] &= \lim_{x \downarrow c} \mathbb{E}[Y_1|X = x] = \mathbb{E}[Y_1|X = c]. \end{aligned}$$

Now, since $\lim_{x \uparrow c} \mathbb{1}\{x < c\} = 1$ while $\lim_{x \uparrow c} \mathbb{1}\{x \geq c\} = 0$,

$$\begin{aligned}\lim_{x \uparrow c} \mathbb{E}[Y|X = x] &= \lim_{x \uparrow c} \mathbb{1}\{x < c\} \lim_{x \uparrow c} \mathbb{E}[Y_0|X = x] + \lim_{x \uparrow c} \mathbb{1}\{x \geq c\} \lim_{x \uparrow c} \mathbb{E}[Y_1|X = x] \\ &= \mathbb{E}[Y_0|X = c].\end{aligned}$$

Similarly, since $\lim_{x \downarrow c} \mathbb{1}\{x \geq c\} = 1$ while $\lim_{x \downarrow c} \mathbb{1}\{x < c\} = 0$,

$$\begin{aligned}\lim_{x \downarrow c} \mathbb{E}[Y|X = x] &= \lim_{x \downarrow c} \mathbb{1}\{x < c\} \lim_{x \downarrow c} \mathbb{E}[Y_0|X = x] + \lim_{x \downarrow c} \mathbb{1}\{x \geq c\} \lim_{x \downarrow c} \mathbb{E}[Y_1|X = x] \\ &= \mathbb{E}[Y_1|X = c].\end{aligned}$$

The result follows by subtracting the equality for $\lim_{x \uparrow c} \mathbb{E}[Y|X = x]$ from that for $\lim_{x \downarrow c} \mathbb{E}[Y|X = x]$. □

7.2 The Fuzzy Regression Discontinuity Design

Sometimes an administrative or legal cutoff strongly influences who is treated without completely determining D. Jacob and Lefgren (2004) provide an interesting example studying the causal effect of two remedial education policies in the Chicago Public Schools: *summer school*, requiring students to extend their school year through August, and *grade retention*, requiring students to repeat a year of school. Between 1996 and 1997, a policy was introduced to end the practice of "social promotion," in which students are automatically advanced to the next school grade with no regard to their academic performance. The policy set achievement thresholds for third-grade, sixth-grade, and eighth-grade students. For these students to advance to the next grade, they were required to score sufficiently well on standardized tests of reading and mathematics. As written, the policy stated that students who exceeded the threshold in June would automatically advance to the next grade. Those who did not would be made to attend summer school and retested in August. Any student who fell below the threshold in the August retest would be required to repeat the previous grade.

So much for the intent of the policy. How was it actually administered in practice? In fact, test scores did *not* perfectly determine which students were assigned to remedial programs. Around 3% of students who fell below the June test threshold were given waivers that exempted them from summer school while roughly 14% of students who completed summer school but failed to exceed the required threshold in the August retest were exempted from repeating a grade. Moreover, some students were held back a grade despite exceeding the June admissions threshold. To accommodate examples like this one in which Assumption 7.1 fails, the fuzzy RD design replaces it with the following.

Assumption 7.3 (Fuzzy RD Design)**.**

$$\lim_{x \downarrow c} \mathbb{P}(D = 1|X = x) \neq \lim_{x \uparrow c} \mathbb{P}(D = 1|X = x)$$

where X is an observed covariate, and c is a known threshold.

Assumption 7.3 requires a jump in the *probability of treatment* at the cutoff c. Since D is binary, an equivalent and slightly more compact way of writing this assumption is

$$\lim_{x \downarrow c} \mathbb{E}[D|X = x] \neq \lim_{x \uparrow c} \mathbb{E}[D|X = x].$$

In the example of Jacob and Lefgren (2004), barely passing the August retest means that you will almost certainly be advanced to the next grade while barely failing means that you will very likely be held back. Thus Assumption 7.3 is satisfied.

In Jacob and Lefgren (2004), the probability of treatment *falls* when test scores X exceed the cutoff c. For consistency with our treatment of sharp RD from above, however, the discussion below tacitly assumes that $\mathbb{P}(D = 1|X)$ jumps *upwards* at c. To accommodate the remedial education example in this context, we could simply re-define X and c. If S is a student's test score, 100 is the maximum possible score, and 40 is the cutoff, for example, we would set $X = 100 - S$ and $c = 60$.

With this in mind, the rough intuition behind fuzzy RD is as follows. Because X does not completely determine D, some people with $X \geq c$ have $D = 0$ and some people with $X < c$ have $D = 1$. For this reason, both $\lim_{x \downarrow c} \mathbb{E}[Y|X = x]$ and $\lim_{x \uparrow c} \mathbb{E}[Y|X = x]$ contain a *mixture* of Y_0 and Y_1, much like $E[Y|Z = 1]$ and $E[Y|Z = 0]$ in the ITT from (5.2). The Wald estimand from (5.1) "magnifies" the ITT by dividing it by the IV first-stage. Analogously, the fuzzy RD estimand τ_F divides the sharp RD estimand τ_S from Theorem 7.1 by the equivalent fuzzy RD "first-stage," in particular

$$\tau_F \equiv \frac{\lim_{x \downarrow c} \mathbb{E}[Y|X = x] - \lim_{x \uparrow c} \mathbb{E}[Y|X = x]}{\lim_{x \downarrow c} \mathbb{E}[D|X = x] - \lim_{x \uparrow c} \mathbb{E}[D|X = x]}. \tag{7.1}$$

Assumption 7.3 is effectively the "IV relevance" condition: it ensures that the denominator of (7.1) does not equal zero. The question remains: what if any *causal interpretation* can we give to τ_F? The answer depends on exactly what we are willing to assume in addition to Assumption 7.3. We consider two possibilities. In the first, we make a relatively strong assumption under which τ_F equals the conditional ATE at $X = c$, namely $\mathbb{E}[Y_1 - Y_0|X = c]$, just as in the sharp RD case. In the second, we make a weaker but more plausible assumption under which τ_F is effectively a local average treatment effect.

Assumption 7.4. *There is some $\varepsilon > 0$ such that for $|x - c| < \varepsilon$,*

$$\mathbb{E}[Y_1 - Y_0|D, X = x] = \mathbb{E}[Y_1 - Y_0|X = x].$$

Assumption 7.4 states that the treatment effect $(Y_1 - Y_0)$ is mean-independent of D conditional on $X = x$ when x is sufficiently close to the threshold. This is very similar to selection on observables (Assumption 4.1). The main difference is that Assumption 7.4 is only required for X in a small neighborhood of c while the selection on observables assumption was supposed to hold for all covariate values. Under this condition, a relatively straightforward argument shows that $\tau_F = \tau_S$, defined in Theorem 7.1.

Theorem 7.2. *Under Assumptions 7.2–7.4, the fuzzy RD estimand τ_F from (7.1) satisfies $\tau_F = \mathbb{E}[Y_1 - Y_0|X = c]$.*

Proof of Theorem 7.2. By (1.2), we have

$$\mathbb{E}[Y|X] = \mathbb{E}\left[Y_0 + D(Y_1 - Y_0)|X\right] = \mathbb{E}[Y_0|X] + \mathbb{E}\left[D(Y_1 - Y_0)|X\right].$$

Moreover, for $|x - c| < \varepsilon$, Assumption 7.4 gives

$$\begin{aligned}\mathbb{E}\left[D(Y_1 - Y_0)|X = x\right] &= \mathbb{E}_{D|X=x}\left\{D\mathbb{E}\left[Y_1 - Y_0|X = x, D\right]\right\} = \mathbb{E}_{D|X=x}\left\{D\mathbb{E}\left[Y_1 - Y_0|X = x\right]\right\} \\ &= \mathbb{E}[D|X = x]\mathbb{E}\left[Y_1 - Y_0|X = x\right]\end{aligned}$$

by iterated expectations. Combining the two preceding equalities and taking the limit from above, it follows that

$$\begin{aligned}\lim_{x\downarrow c}\mathbb{E}[Y|X = x] &= \lim_{x\downarrow c}\mathbb{E}[Y_0|X = x] + \lim_{x\downarrow c}\mathbb{E}[D|X = x]\lim_{x\downarrow c}\left\{\mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x]\right\} \\ &= \mathbb{E}[Y_0|X = c] + \lim_{x\downarrow c}\mathbb{E}[D|X = x]\mathbb{E}[Y_1 - Y_0|X = c]\end{aligned}$$

by Assumption 7.2. Similarly, taking the limit from below,

$$\begin{aligned}\lim_{x\uparrow c}\mathbb{E}[Y|X = x] &= \lim_{x\uparrow c}\mathbb{E}[Y_0|X = x] + \lim_{x\uparrow c}\mathbb{E}[D|X = x]\lim_{x\uparrow c}\left\{\mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x]\right\} \\ &= \mathbb{E}[Y_0|X = c] + \lim_{x\uparrow c}\mathbb{E}[D|X = x]\mathbb{E}[Y_1 - Y_0|X = c].\end{aligned}$$

The result follows by subtracting these expressions and solving for $\mathbb{E}[Y_1 - Y_0|X = c]$. □

Assumption 7.4 rules out *selection on gains* near the threshold: it requires that individuals do not choose their treatment based on knowledge of their treatment effect $(Y_1 - Y_0)$. In the example from Jacob and Lefgren (2004) this would require there to be no relation between being granted an exemption from remedial education and the causal effect of remedial education, given a student's test score. This seems implausible. Indeed, Jacob and Lefgren (2004) found that students who received waivers and exemptions were systematically different in observable characteristics from those who did not. Giving τ_F a meaningful causal interpretation while allowing for selection on gains is more complicated. To do so, we view D as a *deterministic function* of x. For any possible value x that

X could take on, $D(x)$ tells us whether someone would take the treatment, $D(x) = 1$, or not, $D(x) = 0$. Although somewhat more complicated, this idea is very similar to the way that we viewed D as a deterministic function of a person's treatment offer z when defining the compliance types in section 5.5 above. A person for whom $D(x)$ switches from 0 to 1 as x crosses the threshold c is a fuzzy RD *complier*. In the remedial education example, compliers are students who would receive remedial education if and only if their test scores warrant it, according to policy set down by the Chicago Public Schools. Under appropriate assumptions, τ_F turns out to identify the local average treatment effect of remedial education for *this type of student*. In particularly, we impose the following.

Assumption 7.5 (No Defiers)**.** *There is some $\varepsilon > 0$ such that $D(c-h) \leq D(c+h)$ for all $0 < h < \varepsilon$.*

Assumption 7.6 (Exclusion Restriction)**.** *$(Y_1 - Y_0)$ and $D(x)$ are jointly independent of X near c.*

Assumption 7.5 is analogous to Assumption 5.2 in that it rules out *defiers*: people whose treatment status moves in the "wrong direction" as X crosses the threshold. In our running example, defiers would be students who only receive remedial education when their test scores *disqualify them* from doing so. Assumption 7.6 is effectively Assumption 5.1 and Assumption 5.3 rolled into one. This condition is a bit tricky to interpret. In essence it states that, provided that we are sufficiently close to the cutoff, both a person's compliance type, as determined by $D(x)$, and her treatment effect $(Y_1 - Y_0)$ are independent of her test score X. Note that this assumption *allows* selection on gains: there is no requirement that $D(x)$ be independent of $(Y_0 - Y_1)$. Using these conditions, we obtain the following.

Theorem 7.3. *Suppose that $\mathbb{E}[Y_0|X = x]$ is a continuous function of x at $x = c$. Then, under Assumptions 7.5 and 7.6, the fuzzy RD estimand τ_F from (7.1) satisfies*

$$\tau_F = \lim_{h \downarrow 0} \mathbb{E}[Y_1 - Y_0 | D(c+h) - D(c-h) = 1].$$

Proof of Theorem 7.3. By (1.2),

$$\mathbb{E}[Y|X] = \mathbb{E}\left[Y_0 + D(Y_1 - Y_0)|X\right] = \mathbb{E}[Y_0|X] + \mathbb{E}\left[D(Y_1 - Y_0)|X\right]$$

and hence

$$\lim_{x \downarrow c} \mathbb{E}[Y|X = x] = \lim_{x \downarrow c} \mathbb{E}[Y_0|X = x] + \lim_{x \downarrow c} \mathbb{E}[D(Y_1 - Y_0)|X = x]$$
$$\lim_{x \uparrow c} \mathbb{E}[Y|X = x] = \lim_{x \uparrow c} \mathbb{E}[Y_0|X = x] + \lim_{x \uparrow c} \mathbb{E}[D(Y_1 - Y_0)|X = x].$$

Since $\mathbb{E}[Y_0|X=x]$ is continuous at c, its limit from the right at this point equals its limit from the left. Thus, subtracting the two preceding equalities

$$\begin{aligned}\lim_{x\downarrow c}\mathbb{E}[Y|X=x]-\lim_{x\uparrow c}\mathbb{E}[Y|X=x] &= \lim_{x\downarrow c}\mathbb{E}[D(Y_1-Y_0)|X=x]-\lim_{x\uparrow c}\mathbb{E}[D(Y_1-Y_0)|X=x] \\ &= \lim_{h\downarrow 0}\left\{\mathbb{E}\left[D(Y_1-Y_0)|X=c+h\right]-\mathbb{E}\left[D(Y_1-Y_0)|X=c-h\right]\right\}.\end{aligned}$$

Now, for $0<h<\varepsilon$ Assumption 7.6 gives

$$\begin{aligned}&\mathbb{E}[D(Y_1-Y_0)|X=c+h]-\mathbb{E}[D(Y_1-Y_0)|X=c-h] \\ &\quad=\mathbb{E}\left[D(c+h)(Y_1-Y_0)|X=c+h\right]-\mathbb{E}\left[D(c-h)(Y_1-Y_0)|X=c-h\right] \\ &\quad=\mathbb{E}\left[D(c+h)(Y_1-Y_0)\right]-\mathbb{E}\left[D(c-h)(Y_1-Y_0)\right] \\ &\quad=\mathbb{E}\left[\left\{D(c+h)-D(c-h)\right\}(Y_1-Y_0)\right] \\ &\quad=\mathbb{E}\left[Y_1-Y_0|D(c+h)-D(c-h)=1\right]\mathbb{P}\left[D(c+h)-D(c-h)=1\right] \\ &\quad=\mathbb{E}\left[Y_1-Y_0|D(c+h)-D(c-h)=1\right]\mathbb{E}\left[D(c+h)-D(c-h)\right]\end{aligned} \tag{7.2}$$

since $[D(c+h)-D(c-h)]\in\{0,1\}$ by Assumption 7.5. Similarly,

$$\begin{aligned}\mathbb{E}[D|X=c+h]-\mathbb{E}[D|X=c-h] &= \mathbb{E}[D(c+h)|X=c+h]-\mathbb{E}[D(c-h)|X=c-h] \\ &= \mathbb{E}[D(c+h)-D(c-h)].\end{aligned}$$

The result follows by substituting the preceding into the final equality of (7.2), since

$$\lim_{x\downarrow c}\mathbb{E}[D|X=x]-\lim_{x\uparrow c}\mathbb{E}[D|X=x]=\lim_{h\downarrow 0}\mathbb{E}[D(c+h)-D(c-h)]. \qquad \square$$

Like the assumptions under which it was proved, the result of Theorem 7.3 is somewhat difficult to interpret at first glance. What does it mean to condition on the event $\{D(c+h)-D(c-h)=1\}$ in the limit as h approaches zero from the right? Recall that the individuals for whom $D(x)$ changes from zero to one as x crosses the threshold are *precisely* the fuzzy RD compliers. It follows that Theorem 7.3 identifies a causal effect for this sub-group of individuals. For the compliers, the fuzzy RD is really a *sharp* RD: these are precisely the individuals who take the treatment if and only if $X\geq c$. Accordingly, $\lim_{h\downarrow 0}\mathbb{E}[Y_1-Y_0|D(c+h)-D(c-h)=1]$ represents the average treatment effect for compliers at $X=c$.

Chapter 8

Difference-in-differences

Thus far we've focused exclusively on methods for causal inference that can be applied to *cross-section data*, a large number of individuals observed at one point in time. Panel data, repeated observations of the same individuals over time, open up a new range of possibilities for learning cause-and-effect from observational data. In this chapter we'll examine one of them: the **difference-in-differences** (DiD) approach. When we only had a single time period to worry about, we used subscripts we used subscripts to represent potential outcomes: (Y_0, Y_1). In this chapter we have the added complication of different time periods in addition to different potential outcomes, so we need to adopt a new convention. Below we use subscripts to denote *time periods* and parentheses to denote potential outcomes. Thus, $Y_t(d)$ is the potential outcome at time t when treatment status equals $d \in \{0, 1\}$. Remember it like this **P**arentheses = **P**otential Outcomes.

8.1 Adding a Time Dimension

Adding a time dimension to our problem doesn't introduce any mathematical complications but it does present some new conceptual hurdles. Before discussing DiD and causal identification let's begin by seeing how the problem has changed and making clear what we hope to achieve.

To keep life simple, we'll focus on a *two period* model. The outcome Y_t is observed for two time periods: $t \in \{\text{Before}, \text{After}\}$.[1] Some individuals are *treated* between these two time periods: $D = 1$. Others are untreated: $D = 0$. "Before" refers to the time period before anyone has been treated, while "After" refers to the time period after some individuals have received the treatment. Each individual has a pair of potential outcome *time series*, namely $\big(Y_{\text{Before}}(0), Y_{\text{After}}(0)\big)$ and $\big(Y_{\text{Before}}(1), Y_{\text{After}}(1)\big)$. If Alice is treated, she experiences $Y_{\text{Before}}(1)$ followed by $Y_{\text{After}}(1)$; if she is untreated, she experiences $Y_{\text{Before}}(0)$

[1] Many references write these as Y_0 and Y_1. I use the "Before" and "After" convention to avoid confusing time periods with potential outcomes and to make it clear that treatment takes place *between* the time periods in which we observe outcomes.

followed by $Y_{\text{After}}(0)$. In other words,

$$Y_{\text{Before}} = (1 - D)Y_{\text{Before}}(0) + DY_{\text{Before}}(1) \tag{8.1}$$

$$Y_{\text{After}} = (1 - D)Y_{\text{After}}(0) + DY_{\text{After}}(1) \tag{8.2}$$

Notice that the treatment indicator D lacks a time subscript. This is because the treatment takes place *between* the periods "Before" and "After." In our simple model no one is treated in the first period, but some people are treated before the second period. As such, $Y_{\text{Before}}(1)$ does *not* refer to the potential outcome if a person is treated in the first time period. Instead, it refers to the potential outcome in the first period if a person is *eventually treated*, i.e. treated between the two time periods.

That last sentence may strike you as needlessly complicated. Why on earth do we need to distinguish between $Y_{\text{Before}}(0)$ and $Y_{\text{Before}}(1)$? Or to put it another way: how can a person's outcome in the first period depend on a treatment that she can only obtain after this period has ended? The answer to this question is **anticipation**: if I know that I will be treated tomorrow, this may lead me to change my behavior today in ways that affect today's outcome. The fact that I am *eventually* treated could affect my potential outcome before I actually receive the treatment. This phenomenon is sometimes called an "Ashenfelter dip," after Ashenfelter (1978) who found that "all of the trainee [treatment] groups suffered unpredicted earnings declines in the year prior to training," in a study of the effects of a government training program.[2] Figure 8.1 depicts this situation for a hypothetical individual, Alice. Let Y be wage. In the figure, Alice experiences an upward earnings trajectory when treated, the blue line from $Y_{\text{Before}}(1)$ to $Y_{\text{After}}(1)$, and a downward earnings trajectory when untreated, the red line from $Y_{\text{Before}}(0)$ to $Y_{\text{After}}(0)$. She also experiences an Ashenfelter dip, in that her pre-treatment earnings are *lower* if she receives treatment than if she does not: $Y_{\text{Before}}(1) < Y_{\text{Before}}(1)$.

Our goal is to learn the average effect of **treatment on the treated** (TOT) in the *second* time period, namely:

$$\text{TOT} \equiv \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{After}}(0)|D = 1]. \tag{8.3}$$

We focus on the second period because we're not interested in learning the *anticipation effect* of the treatment but rather the *causal effect*: we want to know its effect on the future rather than its effect on the past! We focus on the *treated* sub-population, because these are the only people for whom we can carry out a before-and-after comparison: we observe them once before they have received the treatment and again afterwards. Returning to Figure 8.1, the quantity $\Delta \equiv Y_{\text{After}}(1) - Y_{\text{Before}}(0)$ is is Alice's causal effect.

[2] Ashenfelter's dip is great with tortilla chips. For more discussion of this phenomenon, but sadly no recipe, see Heckman and Smith (1999).

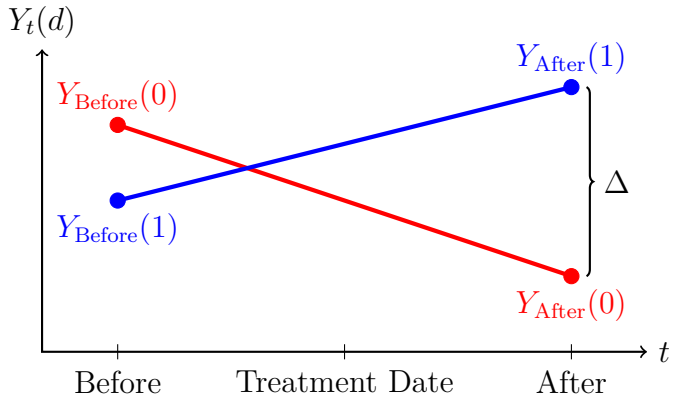


Figure 8.1: Potential time series for a particular individual in a two-period model, depicting an *Ashenfelter dip*. The treatment effect of interest is $\Delta \equiv Y_{\text{After}}(1) - Y_{\text{After}}(0)$.

The TOT is the average of Δ for individuals in the population who eventually receive the treatment.

While adding a time dimension is extremely helpful, the fundamental problem of causal inference remains: **we can never observe both $Y_{\text{After}}(1)$ and $Y_{\text{After}}(0)$ for the same person**. In earlier chapters, we addressed this problem by using *between-person* comparisons. The selection-on-observables approach, for example, constructs two groups of people who are "comparable" in terms of observed covariates X: one that was treated and one that was not. Similarly, the regression discontinuity approach compares people who are just above a threshold, hence treated, to people who are just below, hence untreated, in the hope that the arbitrariness of the cutoff itself makes the two groups comparable. The new idea in this chapter is to use the time dimension to make *within-person* comparisons. The following two sections discuss some simple assumptions under which we can use this approach to identify the TOT.

8.2 The Before-and-After Design

Let's make two very strong assumptions. First we'll rule out Ashenfelter dips, by assuming that there is *no anticipation* of the treatment. This means that Alice's potential earnings in the first period are *the same* regardless of whether she eventually receives the treatment. Second we'll assume that there's no trend in untreated potential outcomes: the path from $Y_{\text{Before}}(0)$ to $Y_{\text{After}}(0)$ is *flat*. This means that Alice will have the *same* earnings in period two as she did in period one if she does not receive the treatment. Figure 8.2 modifies our earlier diagram to incorporate these assumptions, producing a neat little right triangle with vertices $Y_{\text{Before}}(1), Y_{\text{After}}(1)$, and $Y_{\text{After}}(0)$.

This little triangle turns out to solve all of our problems. While we can't observe $Y_{\text{After}}(0)$ and $Y_{\text{After}}(1)$ for the same person, we can observe $Y_{\text{After}}(1)$ and $Y_{\text{Before}}(1)$ for someone who receives the treatment. These are the two endpoints of the blue line in the

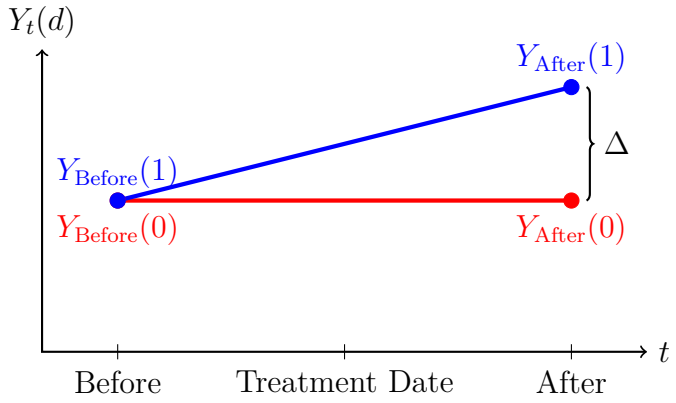


Figure 8.2: Potential time series for a particular individual in a two-period model in which the assumptions of the Before-and-after design are satisfied. The treatment effect of interest is $\Delta \equiv Y_{\text{After}}(1) - Y_{\text{After}}(0)$.

figure. But since $Y_{\text{Before}}(1)$ equals $Y_{\text{Before}}(0)$ which in turn equals $Y_{\text{After}}(0)$ we see that $\Delta = Y_{\text{After}}(1) - Y_{\text{Before}}(1)$. And since these are *precisely* the outcomes that we observe for someone with $D = 1$, it follows that $\mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 1] = \text{TOT}$. This is the **before-and-after design**, a *within-person* comparison of observed outcomes for the treated sub-population: after minus before. And while it makes the pictures easier to read, we don't actually need $Y_{\text{After}}(0)$, $Y_{\text{Before}}(0)$, and $Y_{\text{Before}}(1)$ to be *equal*. We merely need them to be equal *on average*. Assumptions 8.1 and 8.2 make this precise; and Theorem 8.1 shows what these assumptions buy us.

Assumption 8.1 (No Anticipation)**.** $\mathbb{E}\left[Y_{Before}(1) - Y_{Before}(0)\middle| D = 1\right] = 0$.

Assumption 8.2 (No Trend)**.** $\mathbb{E}\left[Y_{After}(0) - Y_{Before}(0)\middle| D = 1\right] = 0$.

Theorem 8.1. *Under Assumption 8.1 and Assumption 8.2, the before-and-after estimand identifies the TOT:*

$$\mathbb{E}[Y_{After} - Y_{Before}|D = 1] = \mathbb{E}\left[Y_{After}(1) - Y_{After}(0)\middle| D = 1\right].$$

Proof. Taking the difference of average *observed* outcomes for the treated gives

$$\text{BA} \equiv \mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 1] = \mathbb{E}[Y_{\text{After}}(1)|D = 1] - \mathbb{E}[Y_{\text{Before}}(1)|D = 1].$$

But since

$$\mathbb{E}[Y_{\text{Before}}(1)|D = 1] = \mathbb{E}[Y_{\text{Before}}(0)|D = 1] = \mathbb{E}[Y_{\text{After}}(0)|D = 1]$$

by Assumptions 8.1 and 8.2, we obtain

$$\text{BA} = \mathbb{E}[Y_{\text{Before}}(1)|D = 1] - \mathbb{E}[Y_{\text{After}}(0)|D = 1] = \text{TOT}. \qquad \square$$

The assumptions required for the before-and-after design to identify the TOT are strong, especially Assumption 8.2. While there are some situations where the "no trends" assumption makes sense, e.g. over a very short time horizon and for a process that follows a random walk, there are many others where it doesn't. Would we really expect Alice's earnings to be the same on average in period two as in period one if she doesn't received the treatment? A year from now the economy may be in a recession. And if Alice spent that year employed, in one year's time she has an extra year of work experience. Here is a useful way to think about the before-and-after design: it provides sufficient conditions for a purely *within-person* comparison to identify the causal effect we hope to learn. If we want to weaken these conditions, we need something more than a within-person comparison.

8.3 The Difference-in-Differences Design

The difference-in-differences design builds on the idea of within-person comparisons introduced in the before-and-after design, but *combines this* with a between-person comparison similar to that used in earlier chapters. We continue to rule out anticipation effects, Assumption 8.1. But rather than assuming that there is *no trend* in the untreated potential outcomes, Assumption 8.2, we instead assume that trend that does exist is *the same* for the treated and untreated populations. In this way, the untreated population serves as a control group *for the trend*.

Figure 8.3 gives the intuition. The treatment effect is $\Delta \equiv Y_{\text{After}}(1) - Y_{\text{After}}(0)$ but we can't observe the two potential outcomes needed to compute this quantity for the same person. For the treated we observe $Y_{\text{After}}(1)$ and $Y_{\text{Before}}(1)$, allowing us to calculate $\Delta_1 \equiv Y_{\text{After}}(1) - Y_{\text{After}}(0)$. This is precisely the *within-person* comparison that we used in the before-and-after design. But since $Y_{\text{After}}(0) \neq Y_{\text{Before}}(0)$, we see that $\Delta_1 \neq \Delta$. Here's where the between person comparison comes in. For an *untreated* person, we observe $Y_{\text{After}}(0)$ and $Y_{\text{Before}}(0)$, allowing us to calculate $\Delta_0 \equiv Y_{\text{After}}(0) - Y_{\text{Before}}(0)$. This is the *trend* in untreated potential outcomes, the quantity that Assumption 8.2 assumed was equal to zero. In Figure 8.3 it clearly does not equal zero. But suppose that this trend were *the same* for treated and untreated people. Then we could use the value of Δ_0 computed from the untreated as a "stand-in" for the value of Δ_0 for the treated. And since $Y_{\text{Before}}(1) = Y_{\text{Before}}(0)$, Assumption 8.1, $\Delta = \Delta_1 - \Delta_0$.

So as its name suggests, the DiD approach is based on taking *double differences*. First we compute the before-and-after average of Y for the treated and untreated individuals in our sample: $\mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 1]$ and $\mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 0]$. The DiD estimand equals the difference of *these differences*, namely

$$\mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 1] - \mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 0].$$

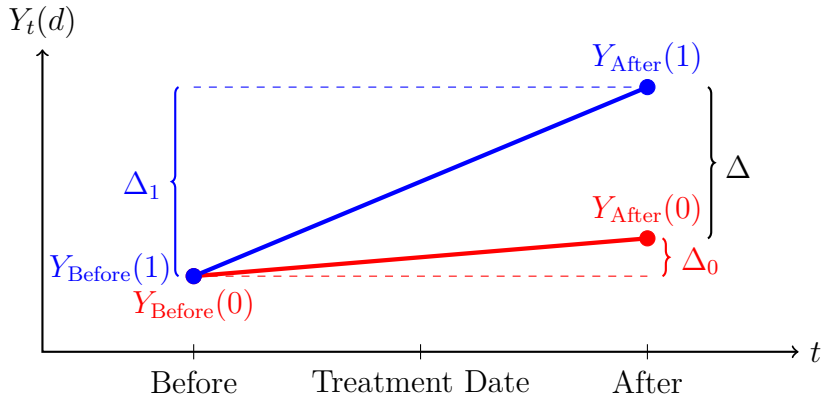


Figure 8.3: Schematic of the difference-in-differences design. The treatment effect of interest is $\Delta \equiv Y_{\text{After}}(1) - Y_{\text{After}}(0)$. We observe $\Delta_0 = Y_{\text{After}}(0) - Y_{\text{Before}}(0)$ for the untreated and $\Delta_1 = Y_{\text{After}}(1) - Y_{\text{Before}}(1)$ for the treated. Since $Y_{\text{Before}}(1) = Y_{\text{Before}}(0)$, no anticipation, $\Delta = \Delta_1 - \Delta_0$. The parallel trends assumption implies that Δ_0 is the same on average for the treated and untreated.

For this approach to identify the TOT, we need two assumptions. First is Assumption 8.1, described above. Second is a new assumption called **parallel trends**.

Assumption 8.3 (Parallel Trends)**.**

$$\mathbb{E}\left[Y_{After}(0) - Y_{Before}(0) \middle| D = 1\right] = \mathbb{E}\left[Y_{After}(0) - Y_{Before}(0) \middle| D = 0\right]$$

Notice that Assumption 8.3 doesn't require that the trend in untreated potential outcomes is *identical* across treated and untreated individuals, merely that it's the same on average. This assumption both allows and restricts selection on unobservables. If *time-invariant* unobservables drive selection into treatment, Assumption 8.3 holds. What the parallel trends assumption rules out is selection based on time-varying unobservables, aka "transitory shocks."

Theorem 8.2. *Under Assumption 8.3 and Assumption 8.1, the difference-in-differences estimand identifies the TOT:*

$$\mathbb{E}[Y_{After} - Y_{Before}|D = 1] - \mathbb{E}[Y_{After} - Y_{Before}|D = 0] = \mathbb{E}\left[Y_{After}(1) - Y_{After}(0) \middle| D = 1\right]$$

In Figure 8.1, $\Delta \equiv Y_{\text{After}}(1) - Y_{\text{After}}(0)$ so DiD identifies the average value of Δ over the subpopulation of individuals who choose to take the treatment.

Proof of Theorem 8.2. Let θ denote the DiD estimand. By definition,

$$\theta \equiv \mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 1] - \mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 0] \tag{8.4}$$

and by (8.2),

$$\mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D=1] = \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(1)|D=1] \tag{8.5}$$
$$\mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D=0] = \mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0)|D=0]. \tag{8.6}$$

Now, by Assumption 8.1

$$\mathbb{E}[Y_{\text{Before}}(1)|D=1] = \mathbb{E}[Y_{\text{Before}}(0)|D=1].$$

Substituting this equality into (8.5), it follows that

$$\mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D=1] = \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(0)|D=1]. \tag{8.7}$$

Therefore, substituting (8.6) and (8.7) into (8.4), we obtain

$$\begin{aligned}\theta &= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(0)|D=1] - \mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0)|D=0] \\ &= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(0)|D=1] - \mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0)|D=1] \\ &= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{After}}(0)|D=1]\end{aligned}$$

where the second equality follows from Assumption 8.3. □

8.4 A Regression Interpretation of DiD

In our simple two-period model, computing the DiD estimator is easy. Simply replace population expectations with sample means and take differences:

$$\widehat{\text{DiD}} = \left(\bar{Y}_{\text{After, Treated}} - \bar{Y}_{\text{Before, Treated}}\right) - \left(\bar{Y}_{\text{After, Untreated}} - \bar{Y}_{\text{Before, Untreated}}\right).$$

An equivalent way of obtaining the same result is by running a linear regression of the observed outcome on a treatment dummy D_i a time dummy $\text{After}_t = \mathbb{1}(t = \text{After})$ and their interaction, namely

$$Y_{it} = \alpha + \beta D_i + \gamma \text{After}_t + \delta\left(D_i \times \text{After}_t\right) + U_{it}, \quad t \in \{\text{Before}, \text{After}\}.$$

The coefficient on the interaction, δ, is the DiD estimand. This approach has several advantages. First it simplifies the process of computing standard errors: we simply obtain them from our usual regression output, possibly adjusting for heteroskedasticity or clustering. Second, it allows us to introduce *control regressors*. Suppose an observed, time-varying covariate X_{it} is responsible for a violation of the parallel trends assumption:

X_{it} affects Y_{it} and is correlated with treatment. By including X_{it} in our regression, e.g.

$$Y_{it} = \alpha + \beta D_i + \gamma \text{After}_t + \delta \left(D_i \times \text{After}_t \right) + X_{it}'\theta + U_{it}.$$

we may be able to salvage the DiD approach in a situation where the parallel trends assumption does *not* hold unconditionally.

8.5 What if the treatment is anticipated?

The *crucial* assumption for DiD is parallel trends: Assumption 8.3. But, as in the before-and-after design, we also assumed that there is no anticipation of the treatment, Assumption 8.1. What happens if we relax this assumption? To find out, we'll first define an additional causal effect: the **anticipation effect for the treated** (AET)

$$\text{AET} \equiv \mathbb{E}[Y_{\text{Before}}(1) - Y_{\text{Before}}(0) | D = 1].$$

The AET quantifies the effect of *future* treatment on *current* outcomes. For example, if a criminal learns that harsher sentencing regimes will be put in place next year and decides to carry out a robbery *now* in response to this policy change, the AET would pick this up. As shown in the following two results, relaxing the no anticipation assumption changes the causal effect identified by the DiD and BA designs. Rather than identifying the TOT, they identify the *difference* TOT − AET. Depending on the application, this may or may not be an interesting quantity but it is at least a meaningful one: the "future effect" of treatment net of any anticipated effect in the present period.

Theorem 8.3. *Under Assumption 8.2* $\mathbb{E}[Y_{After} - Y_{Before} | D = 1] = TOT - AET$.

Proof. By the definition of the before-and-after estimand and (8.2),

$$\text{BA} \equiv \mathbb{E}[Y_{\text{After}} - Y_{\text{Before}} | D = 1] = \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(1) | D = 1].$$

But by Assumption 8.2, $\mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0) | D = 1] = 0$. Therefore,

$$\begin{aligned}
\text{BA} &= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(1) | D = 1] - 0 \\
&= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(1) | D = 1] - \mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0) | D = 1] \\
&= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{After}}(0) | D = 1] - \mathbb{E}[Y_{\text{Before}}(1) - Y_{\text{Before}}(0) | D = 1] \\
&= \text{TOT} - \text{AET}.
\end{aligned}$$

□

Theorem 8.4. *Under Assumption 8.3, the difference-in-differences estimand identifies*

the difference of the TOT and AET:

$$\mathbb{E}[Y_{After} - Y_{Before}|D = 1] - \mathbb{E}[Y_{After} - Y_{Before}|D = 0] = TOT - AET.$$

Proof. By the definition of the DiD estimand and (8.2),

$$\begin{aligned}\text{DiD} &\equiv \mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 1] - \mathbb{E}[Y_{\text{After}} - Y_{\text{Before}}|D = 0] \\ &= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(1)|D = 1] - \mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0)|D = 0].\end{aligned}$$

But by Assumption 8.3,

$$\mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0)|D = 0] = \mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0)|D = 1].$$

Substituting this in the expression for DiD, we obtain

$$\begin{aligned}\text{DiD} &= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{Before}}(1)|D = 1] - \mathbb{E}[Y_{\text{After}}(0) - Y_{\text{Before}}(0)|D = 1] \\ &= \mathbb{E}[Y_{\text{After}}(1) - Y_{\text{After}}(0)|D = 1] - \mathbb{E}[Y_{\text{Before}}(1) - Y_{\text{Before}}(0)|D = 1] \\ &= \text{TOT} - \text{AE}.\end{aligned}$$

□

Bibliography

Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. The Review of Economics and Statistics , 47–57.

Deaton, A., Cartwright, N., 2018. Reflections on randomized control trials. Social science and medicine. 210, 86–90.

DiTraglia, F.J., Garcia-Jimeno, C., 2019. Identifying the effect of a mis-classified, binary, endogenous regressor. Journal of Econometrics 209, 376–390.

DiTraglia, F.J., García-Jimeno, C., 2021. A framework for eliciting, incorporating, and disciplining identification beliefs in linear models. Journal of Business & Economic Statistics 39, 1038–1053.

Fan, Y., Park, S.S., 2010. Sharp bounds on the distribution of treatment effects and their statistical inference. Econometric Theory 26, 931–951.

Hahn, J., Todd, P., Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69, 201–209.

Heckman, J.J., Smith, J.A., 1999. The pre-programme earnings dip and the determinants of participation in a social programme. implications for simple programme evaluation strategies. The Economic Journal 109, 313–348.

Horowitz, J.L., Manski, C.F., 1995. Identification and robustness with contaminated and corrupted data. Econometrica: Journal of the Econometric Society , 281–302.

Huber, M., Mellace, G., 2015. Testing instrument validity for late identification based on inequality moment constraints. Review of Economics and Statistics 97, 398–411.

Jacob, B.A., Lefgren, L., 2004. Remedial education and student achievement: A regression-discontinuity analysis. Review of economics and statistics 86, 226–244.

Kam, C.D., Palmer, C.L., 2008. Reconsidering the effects of education on political participation. The Journal of Politics 70, 612–631.

Kitagawa, T., 2015. A test for instrument validity. Econometrica 83, 2043–2063.

Lee, D.S., Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of economic literature 48, 281–355.

Manski, C.F., 2003. Partial identification of probability distributions. Springer.

Mourifié, I., Wan, Y., 2017. Testing local average treatment effect assumptions. Review of Economics and Statistics 99, 305–313.

Sekhri, S., 2020. Prestige matters: Wage premium and value addition in elite colleges. American Economic Journal: Applied Economics 12, 207–225.

Wooldridge, J.M., 2005. Violating ignorability of treatment by controlling for too many factors. Econometric Theory , 1026–1028.