Lecture Notes on Treatment Effects

(or *Completely Innocuous Econometrics*)

Francis J. DiTraglia

University of Oxford

This Version: 2019-11-25 14:07:39

Abstract

These are lecture notes to accompany weeks 7 and 8 of of the second-year MPhil topics course *Advanced Econometrics 1* at Oxford. Depending on time constraints, the lectures may not cover all of the material included in these notes. If in doubt, feel free to ask me which material you are responsible for. If you spot any typos, I would be very grateful if you could point them out. My email address is: francis.ditraglia@economics.ox.ac.uk.

Contents

Chapter 1

Introduction

In this chapter we set the stage for the material to come, introducing the basic problem of causal inference, developing notation for later use, and reviewing some important mathematical facts that will be used below.

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 you may have gathered from your exposure to panel data methods. 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. 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 Rubin Causal Model (RCM). 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 an 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 constitutes a very active area of research in the treatment effects literature.

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.

These notes 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 distribution $(Y_0, Y_1, D, \boldsymbol{W})$ 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 now allow use to reconstruct both of her potential outcomes.

1.5 Identification Versus Estimation

These notes mainly focus on the problem of **identifying** causal effects rather than that of estimating them. 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 would be sufficient to uniquely pin 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 Our goal: identify the Average Treatment Effect

When treatment effects are heterogeneous, every person in the population could have her own, unique causal effect: $(y_{i1} - y_{01})$. 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 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

[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 these random variables.

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, there is at least the possibility that we could 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 it depends on a feature of the joint distribution of (Y_0, Y_1)—namely the covariance—the variance of the distribution of treatment effects 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? We would love to be able to answer these questions, but unfortunately we cannot. The ATE thus represents not an ideal measure of the effect of D on Y, but the *best we can manage* given the fundamental problem of causal inference.

1.7 Quantile Treatment Effects

If you have studied quantile regression, you may have encountered the term *quantile treatment effect*. Don't let the name fool you: the fact that this quantity is called a treatment effect does not mean that it has a genuine causal interpretation. 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 typically 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. Without further assumptions, $\delta(\tau)$ lacks a causal interpretation. Giving it one requires the so-called **rank invariance** assumption. This condition requires that if Alice occupies the τth quantile of the Y_0 distribution, then she also occupies the τth quantile of the Y_1 distribution. 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. It is difficult to think of real-world examples in which rank invariance is likely to hold. For this reason we focus on identifying the ATE in the remainder of these notes.

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 $\text{ATE} \equiv \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$, 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 regressing 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.*

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

Since D is binary, $\text{Var}(D) = \mathbb{P}(D=1)\,[1-\mathbb{P}(D=1)]$. 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

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

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 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]

The second term in (1.4) measures the difference in average values of Y_0 between the treated and the untreated. In the university and income example, this measures the average **difference in outside options** between those who chose to graduate from university and those who did not.[4] If higher ability people are more likely to graduate from university ($D=1$) and also have a higher-paying outside option, say because ability

[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.

[4]Some authors call the second term in (1.4) the "selection bias." In contrast I reserve this phrase for the *overall* difference between β_{OLS} and the ATE that arises when people are free to choose their treatments.

has a direct effect on income, the second term in (1.4) will be *positive*. Thus, even if the TOT is 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,

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

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 $\text{Cov}(X, W)$,

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

□

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).

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

Definition 1.2 (Covariance and Conditional Covariance).

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

Lemma 1.7 (Shortcut Rule for Variance and Covariance).

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

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

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

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 $f(y|x) = f(x)$ 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 $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ be random vectors. If $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$, then*

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

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

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

Proof. The general case follows as a corollary of Proposition 2.1. Here we will consider the scalar rather than vector case and assume that X, Y, Z are continuous random

[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).

variables. Results for discrete RVs follow by 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

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

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

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

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**: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ implies that $\mathbb{E}[\boldsymbol{XY}] = \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}]$, $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \mathbb{E}[\boldsymbol{Y}]$, and $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}]$. A good exercise would be to prove these implications for yourself in the scalar, continuous RV case.

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 "$\boldsymbol{Z}$," this would imply that $\mathbb{E}(Y_0|D) = \mathbb{E}(Y_0)$ and $\mathbb{E}(Y_1|D)$. 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

[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.

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 thoughtful recent critique, see Deaton & Cartwright (2018).

2.4 The Axioms of Conditional Independence

Now that we understand that 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.

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 $\mathbb{E}(\mathbb{1}\{A_X\}|Y,Z) = \mathbb{E}(\mathbb{1}\{A_X\}|Z)$ 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 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

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

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 Additional 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 $\mathbb{E}(\mathbb{1}\{A_X\}|Y, Z) = \mathbb{E}(\mathbb{1}\{A_X\}|Z)$, $\mathbb{P}$-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,* $\mathbb{E}[f(X)|Y, Z] = \mathbb{E}[f(X)|Z]$

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

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

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,

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

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,

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

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). $\square$

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). $\square$

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). $\square$