10
Treatment Effects

For Whom the Effect Holds

For most of this book, and indeed in the title, we have stuck to the fiction that there is such a thing as *the effect*. As though a treatment could possibly have a single effect—the same impact on literally everybody! That might be plausible in, say, physics. But in social science, everything affects everyone differently.[1]

To give a very simple example, consider a drug designed to reduce the rate of cervical cancer. This drug might be very effective! Perhaps it reduces the rate of cervical cancer by half... *for people with a cervix*. For people without a cervix, we can be pretty certain that the drug has absolutely no effect on the rate of cervical cancer.[2]

So at the very least, the drug has two effects—one for people with a cervix, and one for people without. But we don't need to stop there. Even if we just focus on people with a cervix, maybe the drug is highly effective for some people and not very effective for others. Something to do with body chemistry, or age, or dietary habits, who knows? The point is we might have a whole bunch of effects! Whenever we have a treatment effect that varies across a population (i.e., all the time), we can call that a *heterogeneous treatment effect*.

We can actually think of each individual has having their *own* treatment effect. Maybe the drug reduces the rate by 1% for you and 0% for me and .343% for the woman who lives next door to me.

We're all unique, with different circumstances, lives, physiologies, and responses to the world. Why would we start with an assumption that any two of us would be affected in exactly the same way? It's more of a convenience than anything.

So what can we make of the idea that we have heterogeneous treatment effects?

One thing we can try to do is to estimate those heterogeneous

[1] If we'd just all become frictionless spheres, social science would be way easier. Downhill travel, too.

[2] Ugh, no fair.

Heterogeneous treatment effect. An effect of a treatment on an outcome for which the effect itself varies across the population.

treatment effects! Instead of just estimating one effect, we can estimate a *distribution* of effects and try to predict, for a given person with a given set of attributes, what their effect might be.

This is a valid goal, and it is something that people try to do! This idea is behind concepts you might have heard of like "personalized medicine." It's also one thing that machine-learning types tend to focus on when they get into the area of causal inference.[3]

[3] If this kind of thing interests you, I recommend that you go looking for anything and everything that the duo of Susan Athey and Guido Imbens have worked on together.

However, in addition to being a valid goal and the subject of some extremely cool work, it also gets highly technical very quickly. So in this chapter, we will instead focus on the other thing we can do with the concept of heterogeneous treatment effects: ask "if effects are so heterogeneous, then what exactly are we identifying anyway?"

After all, we've established in the rest of this book that we can identify causal effects if we perform the right set of adjustments. But whose causal effects are those? How can we tell?

It turns out that, if we've done our methods right, what we get is some sort of average of the individual treatment effects. However, it's often not the kind of average where everyone gets counted equally.

Different Averages

WHAT WE HAVE is the concept that each person has their own treatment effect. That means that we can think of there as being a *distribution* of treatment effects. This works just like any other distribution of a variable, like back in Chapter 3. The only difference is that we don't actually observe the treatment effects in our data. So this is a theoretical distribution of some sort.

Treatment effect distribution. The distribution of the individual effect of treatment across the sample or population.

And like any typical distribution, we can describe features of it, like the mean.

The mean of the treatment effect distribution is called, for reasons that should be pretty obvious, the *average treatment effect*. The average treatment effect, often referred to as the ATE, is in many cases what we'd *like* to estimate. It has an obvious interpretation—if you impose the treatment on everyone, then this is the change the average individual will see. If the average treatment effect of taking up gardening as a hobby is an increase of 100 calories eaten per day, then if everyone takes up gardening, some people will see an increase of less than 100 calories, some will see more, but on average it will be 100 calories extra per person.

Average treatment effect. The mean of the treatment effect distribution.

However, estimating the average treatment effect is not always feasible, or in some cases even desirable.

Let's use the cervical cancer drug as an example. In truth, the drug will reduce Terry's chances of cervical cancer by 2 percentage points

and Angela's by 1 percentage point, but Andrew and Mark don't have cervices so it will reduce their chances by 0. The average treatment effect is $(.02 + .01 + 0 + 0) = .0075$, or .75 percentage points.

Now, despite your repeated pleas to the drug company, they refuse to test the drug on people without cervices, since they're pretty darn sure it won't do anything. They get a whole bunch of people like Terry and Angela and run a randomized experiment of the drug. They find that the drug reduces the chances of cervical cancer by, on average, $(.02 + .01) = .015$, or 1.5 percentage points.

That's not a *wrong answer* but it's definitely not the average treatment effect among the population.[4] So if it's not the population average treatment effect, what is it? We will want to keep in our back pocket some ideas of other kinds of treatment effect averages we might go for or might identify.

[4] It *is* the average treatment effect among their sample, but we certainly wouldn't want to take that effect and assume it works for Andrew or Mark!

There are lots and lots and lots of different kinds of treatment effect averages,[5] but only a few important ones we really need to worry about. They fall into two main categories: (1) treatment effect averages where we only count the treatment effects of *some* people but not others, i.e. treatment effect averages conditional on something, and (2) treatment effect averages where we count everyone, but we count some individuals more than others.[6]

[5] I even have one of my own! It's called SLATE and it's not very widely used but it's super duper cool and the way it works is hey where are you going?

[6] Technically, (1) is just a special case of (2) where some people count 100% and other people count 0%. But conceptually it's easier to keep them separate.

What happens when we isolate the average effect for just a certain group of people? And how might we do it?

To answer this question, let's make some fake data. This will be handy because it will allow us to see what is usually invisible—what the treatment effect is for each person.

Once we have our fake data, we will be able to: (a) discuss how we can take an average of just some of the people, and (b) give an example of how we could design a study to *get* that average.

Table 10.1: Fake Data For Four Individuals

Name	Gender	Outcome Without Treatment	Outcome With Treatment	Treatment Effect
Alfred	Male	1	2	1
Brianna	Female	1	5	4
Chizue	Female	2	5	3
Diego	Male	2	4	2

We can see from Table 10.1 that these four individuals have different effects of the outcome. Keep in mind these are counterfactuals—we can't possibly see someone both treated and untreated. The table just describes what we *would see* under treatment or no treatment. If nobody were treated, then Alfred and Brianna would have an outcome of 1, and Chizue and Diego would have an outcome of 2. But with

treatment, Alfred jumps by 1, Brianna by 2, Chizue by 3, and Diego by 4. The average treatment effect is $(1 + 4 + 3 + 2)/4 = 2.5$.

One common way we get an average effect for only a certain group is to literally pick a certain group. Notice in Table 10.1 that we have men and women. Let's say we run an experiment but only recruit men in our experiment for whatever reason.[7] So we get a bunch of guys like Alfred and a bunch of guys like Diego and we ran domly assign them to get treatment or not. Our data ends up looking like Table 10.2.

[7] Perhaps we are a labor economist from the 1980s, or a biologist using mice from the... very recent past.

Name	Treated	Outcome
Alfreds	Treated	2
Alfreds	Untreated	1
Diegos	Treated	4
Diegos	Untreated	2

Table 10.2: Men-Only Experiment

Then, using Table 10.2, we calculate the effect. We find that the treated people on average had an outcome of $(2 + 4)/2 = 3$, and the untreated had $(1 + 2)/2 = 1.5$ and conclude that the treatment has an effect of $3 - 1.5 = 1.5$. This is, of course, the average of Alfred's and Diego's treatment effect, $(1 + 2)/2 = 1.5$. So we have an average treatment effect *among men*, or an average treatment effect *conditional* on being a man.

Conditional average treatment effect. An average treatment effect conditional on the value of a variable.

Again, this isn't a *wrong answer*. It just represents only a certain group and not the whole population. It's only a wrong answer if we think it applies to everyone.

Another common way in which the average effect is taken among just one group is based on who gets treated. Based on the research design, we might end up with the *average treatment on the treated* (ATT) or the *average treatment on the untreated* (ATUT), which averages the treatment effects among those who *actually* got treated (or not).

Average treatment on the treated. The average treatment effect among those who actually received treatment.

Average treatment on the untreated. The average treatment effect among those who did not actually receive treatment.

To see how this works, imagine that we can't randomize anything ourselves, but we happen to observe that Alfred and Chizue get treated, but Brianna and Diego did not. We do our due diligence of drawing out the diagram and notice that being assigned to treatment is unrelated to the outcome.[8] So we're identified! Great.

[8] Knowing the secret counterfactuals that we do, we can see that the average outcome *if treatment had never happened* is exactly $(1 + 2)/2 = 1.5$ for both the treated and untreated groups. In other words, there are no back doors between treatment and outcome. The differences arise only because of treatment!

Name	Treated	Outcome
Alfred	Treated	2
Brianna	Untreated	1
Chizue	Treated	5
Diego	Untreated	2

Table 10.3: Assigning Alfred and Chizue to Treatment

What do we get in our actual data? We can see in Table 10.3 that we get an average of $(2 + 5)/2 = 3.5$ among the treated people, and $(1 + 2)/2 = 1.5$ among the untreated people, giving us an effect of $3.5 - 1.5 = 2$. This also happens to be the average of Alfred's and Chizue's treatment effects, $(1 + 3)/2 = 2$. In other words, we've taken the average treatment effect among just the people who actually got treated. ATT![9]

It's a bit harder to imagine how we might get the average treatment effect among the *untreated* (ATUT). And indeed this one doesn't show up a lot. But the basic idea is that you take what you know about the treatment effect distribution and try to construct one for the untreated group.

For example, say we get a sample of 1000 Alfreds and 1000 Briannas, where 400 Alfreds and 600 Briannas have been assigned to treatment on a basically-random basis, leaving 600 Alfreds and 400 Briannas untreated.

The average outcome for treated people will be $(400 \times 2 + 600 \times 5)/1000 = 3.8$, and for untreated people will be 1. However, we can run our analysis an extra two times, once just on Alfreds and once just on Briannas, and find that the average treatment effect conditional on being Alfred appears to be 1, and the average treatment effect conditional on being Brianna appears to be 4. Since we know that there are 600 untreated Alfreds and 400 untreated Briannas, we can work out that the average treatment on the untreated is $(600 \times 1 + 400 \times 4)/1000 = 2.2$. ATUT!

One other way in which a treatment effect can focus on just a particular group is with the *marginal treatment effect*. The marginal treatment effect is the treatment effect of a person who is *just on the margin* of either being treated or not treated. This is a handy concept if the question you're trying to answer is "should we treat more people?" I won't go too much into the marginal treatment effect here, as actually getting one can be a bit tricky. But it's good to know the idea is out there.

[9] You can imagine how the ATT might crop up a lot. After all, we only see people getting treated if they're... actually treated. It's almost hard to imagine how we could get anything else. How can we possibly ever get the average treatment effect, rather than the ATT, if we can't see what the untreated people are like when treated? Well, it comes down to setting up conditions where we can expect that the treatment effect is the same in treated and untreated groups. In this example, they clearly aren't! But if we truly randomized over a large group of people, there's no reason to believe the treated and untreated groups would have different effect distributions, so we'd have an ATE.

Marginal treatment effect. The treatment effect of the next person who would get treatment if treatment rates expanded.

Instead of focusing our average just on a group of people, what if we include everyone, but perhaps weight some people more than others? We can generically think of these as being called "weighted average treatment effects."

In general, a weighted average is a lot like a mean. Let's go back to the average treatment effect—that was just a mean. The mean of 1, 2, 3, and 4 is $(1 + 2 + 3 + 4)/4 = 2.5$, as you'll recall from our fake data, reproduced below. Now, let's not change that calculation, but just recognize that $1 = 1 \times 1$, $2 = 2 \times 1$, and so on.

Now the mean is $(1\times1+2\times1+3\times1+4\times1)/(1+1+1+1) = 2.5$. Here,

Weighted average treatment effect. An average of individual treatment effects where different individuals count more than others in the average. Each individual has a "weight."

Name	Gender	Outcome Without Treatment	Outcome With Treatment	Treatment Effect
Alfred	Male	1	2	1
Brianna	Female	1	5	4
Chizue	Female	2	5	3
Diego	Male	2	4	2

Table 10.4: Fake Data For Four Individuals

everyone's number is getting multiplied by 1, and that's the same 1 for everybody.

BUT WHAT IF PEOPLE GOT *DIFFERENT* NUMBERS BESIDES 0 AND 1? Continuing with the same example, let's say for some reason that we think Brianna should count twice as much as everyone else, and Diego should count half as much. Now our weighted average treatment effect is $(1 \times 1 + 4 \times 2 + 3 \times 1 + 2 \times .5)/(1 + 2 + 1 + .5) = 2.89$.

Of course, in general we aren't going to just decide that some people should count more and weight them up![10] There's going to be something about the design that weights some people more than others.

[10] Unless you're using survey weights—that's a whole other story.

A common way this shows up is as *variance-weighted* average treatment effects. Statistics is all about variation. And the relationship between Y and X is a lot easier to see if X moves around a whole lot! If you don't see a lot of change in X, then it's hard to tell whether changes in Y are related to changes in X because, well, *what* changes in X are we supposed to look for exactly? What's the relationship between living on Earth and your upper-body strength? Statistics can't help there, because pretty much everybody we can sample lives on Earth. We don't see a lot of people living elsewhere, so we can't observe how it makes them different to live elsewhere.

Variance-weighted average treatment effect. A treatment effect average where the kinds of people with lots of variation in treatment left after closing back doors are counted more heavily.

As a result, if some kinds of people have a lot of *variation in treatment* while others don't, our estimate may weight the treatment effect of of those with variation in treatment more heavily, simply because we can see them both with and without treatment a lot.

Let's say that we get a sample of 1,000 Briannas and 1,000 Diegos. For whatever reason, half of all Briannas have ended up getting treatment, but 90% of Diegos have. So our data looks like Table 10.5.

Name	N	Treated	Outcome
Brianna	500	Treated	5
Brianna	500	Untreated	1
Diego	900	Treated	4
Diego	100	Untreated	2

Table 10.5: Briannas and Diegos get Treatment at Different Rates

Now, we can't just compare the treated and untreated groups because we have a back door! "Being a Brianna / Being a Diego" is related both to whether you're treated, and to the outcome (notice that their outcomes would be different if nobody got treated). So we want to close that back door. One way we can do that is by subtracting out mean differences between Brianna and Diego, both for the outcome and the treatment.

When we do this, and reevaluate the treatment effect, we get an effect of 3.47.[11] This is closer to Brianna's treatment effect of 4 than to Diego's treatment effect of 2. We're weighting Brianna more heavily. Specifically, we are weighting her by the variance in her treatment. The variance in treatment among Briannas is $.5 \times .5 = .25$.[12]. The variance in treatment among Diegos is $.9 \times .1 = .09$. The weighted average, then, is $(.25 \times 4 + .09 \times 2)/(.25 + .09) = 3.47$.

Our estimate of 3.47 is closer to Brianna's effect (4) than Deigo's (2) because we see a lot of her both treated and untreated, whereas Diego is mostly treated. Less variation in treatment means we can see the effect of that variation less! Note also that Diego counts less even though we see a lot of treated Diegos—this isn't the average treatment on the treated. We know we're getting a variance-weighted average treatment effect rather than the average treatment on the treated, because if we were getting ATT, we'd be closer to Diego and farther from Brianna.

Weighted average treatment effects pop up a lot whenever we start closing back doors. When we close back doors, we shut out certain forms of variation in the treatment. The people who really count are the ones who have a lot of variation left!

Variance-weighted treatment effects of course aren't the only kind of weighted average treatment effect. For example, if you close back doors by *selecting a sample* where the treated and untreated groups have similar values of variables on back door paths (i.e., picking untreated observations to *match* the treated observations), you end up with *distribution-weighted* average treatment effects, where individuals with really common values of the variables you're matching on are weighted more heavily.

Another form of weighted treatment effects that pops up often is based on how *responsive* treatment is.

In Chapter 9 we discussed the different ways that we can isolate just *part* of the variation in treatment. We either focus just on the part of the data in which treatment is determined exogenously (like running an experiment, and only including data from the experiment in your analysis) or use some source of exogenous variation to *predict* treatment, and then use those predictions instead of your actual data

[11] The math to get here gets a little sticky, although you can refer to the Conditional Conditional Means section of Chapter 4, or to Chapter 15. But basically, we subtract Brianna's outcome average of 3 from her outcomes, giving treated Briannas a 2 outcome and Untreated Briannas a -2 outcome, and her 50% treatment from her treatments, giving treated Briannas a ".5 treatment" and untreated Briannas a "-.5 treatment". Similarly, treated/untreated Diegos get .2/-1.8 for outcome and .1/-.9 for treatment. Fitting a straight line on what we have left tells us that a one-unit change in treatment gets a 3.47 change in outcome.

[12] The variance of a binary variable is always (probability it's 1)×(probability it's 0)—that's worth remembering!

on treatment.

Of course, heterogeneous treatment effects don't *only* apply to the effect of treatment on an outcome. They can *also* apply to the effect of exogenous variation on treatment!

For example, suppose you're running a random experiment about diet where the treatment is having to eat 100 fewer calories per day than you normally would, and the outcome is your weight. Some people have pretty good willpower and control over their diet. If you tell them to eat less, they can do that. If you tell them to keep doing what they normally do, they can do that too.

Other people have less willpower (or less interest in satisfying a researcher).[13] They might only eat 90 fewer calories per day when told to eat 100 less. Or 50. Or 5. Or 0. Maybe a few people will be disappointed by being assigned to the "continue as normal" treatment and will cut their calories anyway.

[13] Or it's the middle of a pandemic and the Hot Cheetos are *right there in the pantry.*

So for some people, being assigned to treatment makes them eat 100 fewer calories. For some people it's 90, or 50, 0, or 10 *more* calories, or whatever. Heterogeneous treatment effects, but this time for the effect of treatment assignment on treatment, rather than the effect of treatment on outcome!

Naturally, if we limit our data to just the people in our experiment and look at the impact of the experiment, it's going to give us strange results.

When this happens—we have exogenous variation but not everybody followed it, we limit our data to just the people in our experiment, and we look at the relationship between treatment *assignment* and the outcome— what we get is called the *intent-to-treat* estimate.[14] Intent-to-treat is basically the effect *of assigning treatment*, although not the effect of treatment itself, since not everybody follows the assignment.

[14] More broadly, when we have *exogenous variation* of some sort driving treatment, and we look directly at the relationship between that exogenous variation and the outcome.

Intent to treat. The average treatment effect *of assigning treatment*, which is not necessarily the same as the average treatment effect *of treatment*.

Intent-to-treat gives us the average treatment effect of assignment, which is usually not what we want.[15] What does it give us for the effect of treatment? It's not *exactly* a weighted average treatment effect at that point. It does weight each person's treatment effect by *the proportion of their treatment effect they received*.[16] So if you got enough treatment to get 50% of its effects, you get a weight of .5. This weighting makes a lot of sense—if you get the full treatment, we see the full effect of your treatment when we start adding up differences. If you don't get the treatment you were assigned to, we still include you in our addition, but it couldn't have had an effect so you get a 0.[17]

[15] Unless we're going to use that same assignment in the real world. If I'm using "a policy that forces insurers to cover therapy" to understand the effect of therapy on depression, maybe I *do* want to know the effect of that policy, rather than the effect of therapy itself, since I have more control as a policymaker over that policy than I do over therapy.

[16] In most cases, this is just "actually got the treatment" or "didn't" so it's just 0 and 1.

[17] All of this applies even if treatment isn't 0/1! In those cases the weights are "how much more treatment you got."

The thing that makes it not exactly a weighted treatment effect is that instead of dividing by the sum of the weights, you divide by the number of individuals. In a weighted average treatment effect, a

weight of 0 (you didn't respond to assignment at all) wouldn't affect the weighted average treatment effect. But in intent-to-treat, someone with a weight of zero has no effect on the numerator, but they do affect the denominator, bringing the effect closer to 0.

Name	Gender	Outcome Without Treatment	Outcome With Treatment	Treatment Effect
Alfred	Male	1	2	1
Brianna	Female	1	5	4
Chizue	Female	2	5	3
Diego	Male	2	4	2

Table 10.6: Fake Data For Four Individuals

Returning to our fake data once more, if we recruited two Chizues and two Diegos and treated one of each, but Chizue went along with assignment while Diego decided never to receive treatment, then in the treatment-assigned group we'd see Chizue's 5 and Diego's 2 (since Diego was never actually treated), and in the treatment-not-assigned group we'd see Chizue's 2 and Diego's 2. The calculated effect would be $(3.5 - 2) = 1.5$. This is also $(3 \times 1 + 3 \times 1 + 2 \times 0 + 2 \times 0)/(1 + 1 + 1 + 1) = 1.5$, or the effect of the two Chizues weighted by 1 (since they receive full treatment when assigned) plus the effect of the two Diegos weighted by 0 (since they never receive any treatment), divided by the number of people (4).

What if we take the other approach to finding front doors, where we use some source of exogenous variation to *predict* treatment, and then use those predictions instead of your actual data on treatment?

This turns out to do something very similar to the intent-to-treat. However, because this approach doesn't just say "were you assigned treatment or not?" but rather "how much more treatment do we think you got due to assignment?" we can now replace that "number of people" denominator with a "how much more treatment was there?" denominator.

Since "how much more treatment" was also our weight in the numerator, we're back to an actual weighted average treatment effect! Specifically, the weights are how much additional treatment each individual would get if assigned to treatment. We call this one the *local average treatment effect* (LATE).

Local average treatment effect. A weighted average treatment effect where the weights are *how much more treatment* that individual would get if assigned to treatment.

For example, let's go back to Chizue and Diego, and Diego not going along with his treatment assignment. We look at assignment and at treatment, and notice that being assigned to treatment only seems to increase treatment rates by 50% (in the not-assigned group, nobody is treated; in the assigned group, 50% are treated). Based on that prediction, we expect to see only half of the treatment effect, and we can get back to the full treatment effect by dividing by .5.

This gives us an effect estimate of $(3.5 - 2)/.5 = 3$. We can also get this 3 from $(3 \times 1 + 3 \times 1 + 2 \times 0 + 2 \times 0)/(1 + 1 + 0 + 0) = 3$, which is the 3 effect of the two Chizues, each with a weight of 1 (since assignment increases their treatment from 0 to 1), and the 2 effect of the two Diegos, each with a weight of 0 (since assignment doesn't affect their treatment).

In other words, the LATE is a weighted average treatment effect where you get weighted more heavily the more strongly you respond to exogenous variation.[18] This is kind of a strange concept—why would we want to weight people who respond to irrelevant exogenous variation more strongly? Well, maybe we don't. But the LATE still looms large because it happens to be the weighted average treatment effect that pops up in a lot of research designs. Maybe not what you want, but what you get.

And along those lines, what *do* you get? How do we know, for a given research design, which of these treatment effect estimates we will end up with?

[18] It is common in an econometrics class to hear that the LATE is "the average treatment effect among those who respond to assignment" and you might hear those who respond called "compliers." However, this is a simplification. If one person responds fully to assignment and another only has half a response, the LATE will not average them equally, even though both are compliers. It will weight the full-response person twice as much as the half-response person.

I Just Want an ATE, It Would Make Me Feel Great, What Do I Get?

By this point we know that there are far more ways to get a single representative treatment effect than just averaging them (to get the average treatment effect), due to the fact that we have heterogeneous treatment effects. We can get the treatment effect just for certain groups, we can weight some individuals more heavily than others, we can weight people based on how the treatment was assigned.

Now, usually (not always), what we want is the average treatment effect—the effect we'd see on average if we took a single individual and applied the treatment to them.[19] The reason we bring up most of those *other* treatment effects at all is that we don't always get what we want!

The treatment effect you get isn't necessarily a choice you make. It's a consequence of the research design you have. And since there aren't usually multiple available research designs that you can use to answer a given question, you're often stuck with the treatment effect average you get.

So for a given research design, which one do we get?

[19] Why might we not always want this? It depends what question we're trying to answer. If we want to know "what was the actual effect of this historical policy?" then we might want to know what effect treatment had on the people it actually treated (ATT). If we want to know "what would be the effect if we treated more people?" we might want the treatment on the untreated (ATUT) or the marginal treatment effect. If we want to know "is this more effective for men or women?" we would want some conditional treatment effects. And so on.

The treatment effect we get is almost entirely determined by the source of treatment variation we use. That's pretty much it. Ask where the variation in your treatment is coming from,[20] and you'll have a pretty good idea whose treatment effects you

[20] After removing any variation you choose to remove by controlling for things, etc.

are averaging, and who is being weighted more heavily.

We've already discussed one example of this. If we perform a randomized experiment, then we will be ignoring everyone who isn't in our experiment. *The only treatment variation we are allowing is among the people in our sample—any variation outside our sample is ignored.* If our sample isn't representative of the broader population,[21] then we will be getting the average treatment effect *conditional on being in our sample*, a conditional average treatment effect.

[21] And thus doesn't have the same average treatment effect as the broader population.

Let's take another example. Let's say we're interested in the effect of being sent to traffic school on your future driving performance. Let's also say that we know there are only two reasons anyone goes to traffic school: making a terrible driving mistake, or having *someone else* make a terrible driving mistake that you are somehow punished for. This gives us the diagram in Figure 10.1.

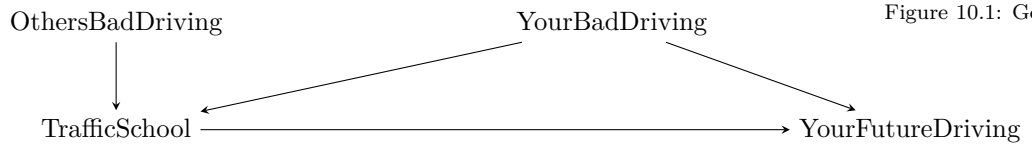

Figure 10.1: Going to Driving School

Recognizing the clear TrafficSchool ← YourBadDriving → YourFutureDriving back door, we decide to identify the effect by measuring and controlling for your own bad driving skills.

This will identify the effect, but it will also shut out any variation in TrafficSchool that's driven by YourBadDriving. So imagine two people, Rodney and Richard. Rodney has a 50% chance of not going to TrafficSchool, a 10% chance of going because of someone else's bad driving, and a 40% chance of going because of his own bad driving. Richard has a 50% chance of not going to TrafficSchool, a 30% chance of going because of someone else's bad driving, and a 20% chance of going because of his own bad driving.

We're tossing out that 40% for Rodney and 20% for Richard chances of going because of their own bad driving. There's only a 10% chance that Rodney goes to TrafficSchool *for the reason we still allow to count*, and similarly a 30% chance for Richard. That means there's *more remaining variation in treatment for Richard than for Rodney*, so Richard's treatment effect will be weighted more heavily than Rodney's will. A weighted average treatment effect!

FOLLOWING THIS LOGIC—WHICH TREATMENT VARIATION DO WE ALLOW TO COUNT—will tell us just about every time which treatment effect we're about to get.

We can go a little bit further and apply this logic ahead of time to develop some rules of thumb. These are just shortcuts to applying

that same logic, but they're often easier to think about, and they work most of the time.

Rule of thumb 1: If you have true randomization in a representative sample and don't need to do any adjustment, you have an average treatment effect (ATE).

Rule of thumb 2: If you have true randomization only within a certain group, and you isolate that group so you can take advantage of that randomization, you have a conditional average treatment effect.

Rule of thumb 3: If you know that some variation in treatment is connected to back doors and so you close those back doors, using only the remaining variation, you have a weighted average treatment effect—variance-weighted if you're subtracting out explained variation, or weighted by how representative the observations are if you're picking a subsample of the data or picking control observations by matching them with treated observations.

Rule of thumb 4: If we are identifying our effect by assuming that the untreated group is what the treated group would look like if they hadn't been treated, but *not* assuming they both have the same treatment effects, then we have the average treatment on the treated (ATT).

Rule of thumb 5: If part of the variation in treatment is driven by an exogenous variable, and we isolate just the part driven by that exogenous variable, then we have a local average treatment effect (LATE).

Who Cares?

IT SEEMS ALMOST BEYOND THE POINT to worry too much about which kind of treatment effect average we get, doesn't it? After all, we've gone to all the work of identifying the effect in the first place. And each of these are averages of the actual treatment effects. Why should it matter?

We should care because we're interested in understanding causal relationships in the world!

The reason for paying attention to treatment effect averages (and which ones we are getting) is very clear if the reason we care about causal effects is that *we want to know what will happen if we intervene.*

Think way back to when we were defining causality back in Chapter 6—one way we talked about it was in the form of intervention. If we were to intervene to change the value of X, and Y changes as a result, then X causes Y.

This approach to causality is one reason why we care about getting

causal effects in the first place. It's useful! If we know that X causes Y, then if we want to improve Y, we can change X! If aspirin reduces headaches, and you have a headache, then take an aspirin. We know what will happen because we've established the causal relationship.

Bringing in treatment effect averages changes considerably *what we can infer about what will happen* based on *the estimates we get in our analysis.*

For example, let's say we suspect that the presence of lead in the drinking water has resulted in increased crime.[22] If we find evidence that lead in the drinking water *does* cause crime to rise, what would we use that information to do? Probably get the lead out of the drinking water, right?

[22] Which it might well do! See for example Reyes (2007).

However, what if it doesn't reduce crime for everyone? Let's say we found a number of localities that won government grants, awarded at random, to clean up the lead in their water. But among the localities that applied for the grants, there was no change in crime rates that followed. Perhaps their crime rates were already very low, or only localities with lead levels already too low to have an effect were the ones who applied for the grant.

In the case of this study, we got an average treatment effect conditional on being in the study. That conditional average treatment effect misrepresented the average treatment effect that we would get if we reduced lead levels in *everyone's* drinking water. If we don't pay attention to which treatment effect average we're getting, we might erroneously think that the effect is zero for everyone.

This can go the other way too, where we estimate an average treatment effect but don't want that! For example, imagine you develop a new (and you think better) vaccine for the measles. You study your new vaccine with an experiment in the United States. And because you want to get a really representative average effect, you do a very careful job randomly recruiting everyone into your study, sampling people from all walks of life completely at random. For simplicity, let's assume nobody refuses being in your study.

This approach—selecting people completely at random and nobody opting out of the study—will give us an average treatment effect (at least in the United States).

Then you get the results back and you're shocked! The vaccine reduces the chances of measles, but only by a few tenths of a percent.

Well, that's probably because in the United States, north of 90% of people already have a measles vaccine, so your vaccine won't do much extra for them. What you wanted was the average treatment effect conditional on not already having had a measles vaccine.[23]

[23] Strangely, this does not count as an average treatment on the untreated.

In general, what you want is to think about *what intervention would look like*, whether it will be in the form of a policy that could be considered (changing how vaccinations occur, reducing lead for everyone, etc.) or in understanding how the world works (wages are going up for group X; what is the effect of wages on home ownership among group X?).

Once we know what intervention looks like, we want a treatment effect average that will match it. Planning to apply treatment to everyone, or at random? The average treatment effect is what you want! Just to a particular group? The conditional average treatment effect for that group. Wanting to expand an already-popular treatment to more people? Probably want the average treatment on the untreated or a marginal treatment effect. Planning to continue a policy that people opt into? Average treatment on the treated!

Understanding not just the overall effect, but who that effect is for, really fills in the gaps on making information from causal inference useful.

Treatment Effect Glossary

We've talked about a whole lot of different kinds of treatment effects. Let's remind ourselves what they are.

Average Treatment Effect. The average treatment effect across the population.

Average Treatment on the Treated. The average treatment effect among those who actually received the treatment in your study.

Average Treatment on the Untreated. The average treatment effect among those who did not actually receive the treatment in your study.

Conditional Average Treatment Effect. The average treatment effect among those with certain values of certain variables (for example, the average treatment effect among women).

Heterogeneous Treatment Effect. A treatment effect that differs from individual to individual.

Intent-to-Treat. The average treatment effect of assigning treatment, in a context where not everyone who is assigned to receive treatment receives it (and vice versa).

Local Average Treatment Effect. A weighted average treatment effect where the weights are based on how much more treatment an individual would get if assigned to treatment than if they weren't assigned to treatment.

Marginal Treatment Effect. The treatment effect of the next individual that would be treated if treatment were expanded.

Weighted Average Treatment Effect. A treatment effect aver-

age where each individual's treatment effect is weighted differently.

Variance-Weighted Average Treatment Effect. A treatment effect average where each individual's treatment effect is weighted based on how much variation there is in their treatment variable, after closing back doors.

Chapter Problems

1. The glossary is just above this section. But ignore it for a moment. Define *in your own words* each of the following terms:

   (a) Conditional average treatment effect

   (b) Average treatment on the treated

   (c) Average treatment on the untreated

2. Provide an example of a treatment effect that you would expect to be highly heterogeneous, and explain why you think it is likely to be heterogeneous.

3. Consider the data below in Table 10.7 that shows the hypothetical treatment effect of cognitive behavioral therapy on depression for six participants. For the sake of this example, the six participants represent the population of interest.

Case	Age	Gender	Effect
A	15	Man	7
B	40	Woman	3
C	30	Woman	7
D	20	Non-binary	8
E	15	Man	7
F	25	Woman	4

Table 10.7: Six Hypothetical Treatment Effects

   (a) What is the overall average treatment effect for the population?

   (b) What is the average treatment effect for Women?

   (c) If nearly all Non-binary people get treated, and about half of all Women get treated, and we control for the differences between Women and Non-binary people, what *kind* of treatment effect average will we get, and what can we say about the numerical estimate we'll get?

   (d) If we assume that, in the absence of treatment, everyone would have had the same outcome, and also only teenagers (18 or younger) ever receive treatment, and we compare treated people to control people, what *kind* of treatment effect average will we get, and what can we say about the numerical estimate we'll get?

4. Give an example where the average treatment effect on the treated would be more useful to consider than the overall average treatment effect, and explain why.

5. Which of the following describes the average treatment effect of assigning treatment, whether or not treatment is actually received?

   (a) Local average treatment effect

   (b) Average treatment on the treated

   (c) Intent-to-treat

   (d) Variance-weighted average treatment effect

6. On weighted treatment effects:

   (a) Describe what a variance-weighted treatment effect is

   (b) Describe what a distribution-weighted treatment effect is

   (c) Under what conditions/research designs would we get each of these?

7. Suppose you are conducting an experiment to see whether pricing cookies at $1.99 versus $2 affects the decision to purchase the cookies. The population of interest is all adults in the United States. You recruit people from your university to participate and randomize them to either see cookies priced as $1.99 or $2, then write down whether they purchased cookies. What kind of average treatment effect can you identify from this experiment?

8. For each of the following identification strategies, what kind of treatment effect(s) is feasible to identify? Provide rationales for your answers.

   (a) A randomized experiment using a representative sample

   (b) True randomization within only a certain demographic group

   (c) Closing back door paths connected to variation in treatment

   (d) Isolating the part of the variation in treatment variable that is driven by an exogenous variable

   (e) The control group is comparable to the treatment group,but treatment effects may be different across these groups