CHAPTER 3

Causation: What Is It and What Is It Good For?

What You'll Learn

- A causal effect is a change in some feature of the world that would result from a change to some other feature of the world.
- Assessing causal relationships is crucial for policy and decision making.
- *"What effect did this have on the outcome?"* is a more conceptually clear question than *"What caused the outcome?"*
- Causal relationships are about comparisons of *counterfactual* worlds. As a result, they are fundamentally unobservable. But, in certain situations, we can learn about them from data.

Introduction

As we saw in chapter 2, knowledge of correlations is useful for many purposes. Among the most important, but also most vexing, purposes is learning about causal relationships.

We make claims about causal knowledge all the time. I did poorly on the test because I didn't get enough sleep. Going to college will improve my future job prospects. A political candidate lost an election because of an attack ad. Violent crime is down because of a new policing strategy.

Thinking clearly about whether a causal relationship exists is perhaps the most important conceptual challenge for learning to use information to make better decisions. This is because causal knowledge is the key to understanding how your decisions and actions affect the world around you. If you propose a new tax policy, test-prep strategy, exercise plan, or advertising campaign, you're doing so not because you think it is correlated with better outcomes. Rather, you must believe that enacting your proposal will actually cause better outcomes.

Our goal in this chapter is to clarify exactly what we mean when we talk about causal relationships. Causality is a deep and perplexing topic to which much attention has been paid by scholars from many different fields. We won't be able to resolve all the thorny philosophical questions here. Instead we've set more modest goals. First, we want to make sure we are all on the same page by defining how we will use causal language for the duration of this book. Then we will explain why the notion of causality we adopt is a particularly valuable one. Finally, we will discuss some other approaches to talking

about causality and explain why, from our point of view, they are less helpful than the one we adopt.

What Is Causation?

When we talk about causation, we're talking about the effect of one thing on another. In non-technical terms, a *causal effect* is a change in some feature of the world that would result from a change to some other feature of the world. So, for instance, we would say that the tax rate has a causal effect on government revenue if changing the tax rate would lead to a change in government revenue.

We've defined the notion of an effect in non-technical terms, so you might not have noticed that we actually slipped in a bit of philosophy. What do we mean by *would result*? After all, the world is as it is. Where did this change in some other feature of the world come from?

That's a good question. In fact, our definition of a causal effect relies on a thought experiment about which we need to be explicit. Let's start with an example.

The movie star Gwyneth Paltrow runs a company called Goop that promotes stickers, called Body Vibes, that are supposed to promote health, wellness, *and* good skin. Here's what the Goop website says about Body Vibes:

> Human bodies operate at an ideal energetic frequency, but everyday stresses and anxiety can throw off our internal balance, depleting our energy reserves and weakening our immune systems. Body Vibes stickers come pre-programmed to an ideal frequency, allowing them to target imbalances. While you're wearing them—close to your heart, on your left shoulder or arm—they'll fill in the deficiencies in your reserves, creating a calming effect, smoothing out both physical tension and anxiety. The founders, both aestheticians, also say they help clear skin by reducing inflammation and boosting cell turnover.

Suppose you paid the required six dollars per sticker because you really want clear skin. But then your friends started making fun of you for being a sucker. In defending yourself, you'd want to claim that Body Vibes really do have an effect on the clarity of your skin. But what, exactly, would you mean by that claim?

Here's a way of thinking about this. Imagine an alternative world where, at the exact moment you went to stick on your Body Vibes stickers, unbeknownst to you, one of your friends replaced them with identical-looking stickers that cost ten cents instead of six dollars, but which hadn't been "pre-programmed to an ideal frequency." If your skin clarity would be worse in that alternative world, then we would say that Body Vibes have a positive effect on your skin clarity. If your skin clarity would be the same in that alternative world, we'd have to conclude that Body Vibes don't have the claimed effect on skin clarity And if your skin clarity would actually be better in that alternative world, we'd conclude Body Vibes have a negative effect.

We can extend this thought experiment. There's nothing particularly special about the real world. Once we're already thinking about one alternative world, we might as well think about two. For instance, we could think about the effect of ten-cent stickers compared to magical crystals, even if you've never tried either of those approaches to skin care. We just have to compare two make-believe worlds: one where your friends secretly stuck stickers on your upper left shoulder near your heart, and another where

they snuck crystals into your pockets. These kinds of comparisons are called *counterfactual* thought experiments because at least one of the worlds we are comparing isn't the real, factual world—it's in our imaginations. The comparison of outcomes in such a thought experiment is a *counterfactual comparison.*

We can now make sense of the phrase *would result* in our definition of a causal effect. It refers to a counterfactual comparison between the outcome in the actual world and the outcome in a counterfactual world that is identical to the actual world up until the point where the feature of the world claimed to have a causal effect is changed.

This idea of counterfactuals is philosophically subtle. So, to help us make sure we are thinking clearly, we are going to introduce a mathematical framework for representing counterfactuals called *potential outcomes.* Using the potential outcomes framework requires some notation, but it isn't too complicated. And once you master the notation, you will have a much deeper understanding of what causality really is. So let's give it a shot.

Potential Outcomes and Counterfactuals

We are interested in the effect of some *treatment* (say, Body Vibes) on some outcome (say, skin health). Let's call the treatment T. It is a binary variable, taking a value of 0 or 1. If $T = 1$ for some person, that means the person received the Body Vibes treatment. If $T = 0$ for some person, that means the person didn't receive the Body Vibes treatment. We sometimes say that a unit (here, a person) with $T = 1$ is *treated* and a unit with $T = 0$ is *untreated,* although it's often arbitrary what we call treated and what we call untreated (e.g., we could just as easily talk about the effect of *not* wearing Body Vibes).

Similarly, let's refer to the outcome we are interested in as Y. In our example, Y describes a person's skin health. In a metaphysical sense, there is some level of skin health that each individual would have had if they'd used Body Vibes and some level of skin health they would have had if they hadn't used Body Vibes. These are that person's *potential outcomes.* However, at any given moment, we only ever get to observe one of these—each person is either using or not using Body Vibes. Nonetheless, thinking about both potential outcomes helps us to think clearly about counterfactuals:

$$Y_{1i} = \text{outcome for unit } i \text{ if } T = 1$$

$$Y_{0i} = \text{outcome for unit } i \text{ if } T = 0$$

The effect of wearing Body Vibes on person i's skin health is just the difference in i's skin health with and without Body Vibes. In our potential outcomes notation, it is

$$\text{Effect of Body Vibes on } i\text{'s Skin Health} = Y_{1i} - Y_{0i}.$$

Table 3.1 makes this more concrete. We observe ten individuals. For each individual, we observe whether they received Body Vibes and whether their skin is clear. If person i received Body Vibes, their treatment status is $T_i = 1$; if they did not, their treatment status is $T_i = 0$. And if person i had treatment status T, we write their outcome as $Y_{Ti} = 1$ if their skin is clear and $Y_{Ti} = 0$ if their skin is not clear.

The actual outcome for each individual is bold in the table. Individuals 1–5 received Body Vibes, so their actual outcome is Y_{1i}. The table also tells us what these individuals' outcomes would have been if they hadn't received Body Vibes, Y_{0i}. However, in

Table 3.1. Potential outcomes for skin health with and without Body Vibes. For each individual, the actual outcome that we can observe is in bold type. The counterfactual outcome that we do not observe is in regular type.

		Skin Health with Body Vibes Y_{1i}	Skin Health without Body Vibes Y_{0i}	Treatment Effect for Individual i $Y_{1i} - Y_{0i}$
Receive Body Vibes	**Individual 1**	**1**	1	0
	Individual 2	**0**	0	0
	Individual 3	**0**	0	0
	Individual 4	**1**	1	0
	Individual 5	**1**	1	0
Don't Receive Body Vibes	**Individual 6**	0	**0**	0
	Individual 7	0	**0**	0
	Individual 8	1	**1**	0
	Individual 9	1	**1**	0
	Individual 10	0	**0**	0

the actual world, no one can observe these counterfactual outcomes, since they don't actually occur. Individuals 6–10 do not receive Body Vibes. So their actual outcome is Y_{0i}. Again, although the table tells us what their outcomes would have been if they'd received Body Vibes, Y_{1i}, these counterfactual outcomes are not observed in the actual world.

Because the table tells us the potential outcomes in the actual and counterfactual worlds, we can find the treatment effect of Body Vibes for each individual by calculating $Y_{1i} - Y_{0i}$. Doing so reveals that Body Vibes don't actually have any effect on the skin health of any individual. Individuals 1, 4, 5, 8, and 9 all have clear skin. But for all of these individuals, that would be true whether or not they received Body Vibes. Individuals 2, 3, 6, 7, and 10 all have unclear skin. Again, however, this would be true with or without Body Vibes. Importantly, as we will come back to later, this absence of a causal effect can't actually be observed in the world because we only observe the actual outcome for each individual, not the potential outcome in the counterfactual world where they had a different treatment status.

We say that causality is about counterfactual comparisons because we can only observe, at most, one of the two quantities, Y_{1i} or Y_{0i}, for any individual at any particular point in time. This means that we can't directly measure the effect of wearing Body Vibes on an individual's skin health. We suspect this fact is key to their business model.

What Is Causation Good For?

Knowledge of causation is necessary for understanding the consequences of an action that changes some feature of the world. In particular, to weigh the costs and benefits of a decision, you need to know how your action will affect the outcomes you care about.

For instance, you can't possibly know if it is a good idea to spend money on a drug to treat heart disease without knowing about a causal relationship—whether the drug reduces the risk of heart disease. The same goes for many decisions. When you are deciding whether or not to intervene in the world in some way—with a policy, an exercise plan, a parenting strategy, a new kind of online learning, or what have you—you want to know how the intervention *affects* the outcomes you care about.

While the examples we've discussed are easily understood in terms of counterfactual comparisons, sometimes thinking in terms of counterfactuals can seem vexing or confusing. In the next sections, we explore some of these issues.

The Fundamental Problem of Causal Inference

In our discussion of table 3.1 we nodded toward an important issue—causal effects as we've defined them can never, ever be directly observed. Everyone either receives Body Vibes or doesn't receive Body Vibes. So you only observe one potential outcome for each person. But the causal effect is the difference in a person's potential outcomes. This inherent unobservability of causal effects is called the *fundamental problem of causal inference*. Let's see exactly why we can't observe causal effects and what that implies for our ability to learn about causality.

The effect of going to college on your income is the difference in your income in a world in which you go to college versus a world in which you are the same up until the college decision but you don't go to college. At least one of those worlds is counterfactual. You can't both go to college and not go to college. That is, you have two potential outcomes—Y_{college} and $Y_{\text{no college}}$. But you have only one *actual* outcome: either you went to college or you didn't. Given this, we can never observe the effect of going to college on your income since we only observe your income in the actual world, not the counterfactual world.

The fundamental problem of causal inference, then, is that, at any given time, we only observe any given unit of analysis (e.g., a person, basketball team, or country) in one state of affairs. So we can't observe the effect on that unit of being in that state of affairs versus some other state of affairs, because all the other states of affairs are counterfactual. We can't know $Y_{\text{college}} - Y_{\text{no college}}$ for you, because we only observe one of the two values. We saw this fact earlier, in table 3.1, where we noticed that we could only observe the actual outcome for each individual; the other potential outcome was counterfactual.

So how do we make progress on answering causal questions if effects are fundamentally unobservable? Fortunately, there are lots of situations where we don't necessarily need to know the effect for every individual unit of analysis. Instead, we want to know the average effect across lots of individuals.

Suppose, for instance, that the Food and Drug Administration (FDA) is deciding whether to approve a new drug. To learn about the health effects of the drug, scientists conduct a randomized trial, assigning some people to take the drug (the treated group) and other people to take a placebo (the untreated group). Because of the fundamental problem of causal inference, the scientists can't observe the effect of taking the drug on any individual. Each person is either taking the drug or not. But by comparing the average health outcomes for people in the untreated group to the average health outcomes for people in the treated group, they can assess the average effect of the drug. (We'll talk a lot more about how this works in parts 2 and 3.) Doing so allows the scientists

to answer what turns out to be the key causal question for the FDA's decision: If we approve the new drug, how will health change in the population on average?

Drug approval is one setting in which knowledge about average effects is sufficient to inform the key decisions. But there are some settings where this is not the case and the fundamental problem of causal inference constitutes a real challenge. For instance, assessing legal liability involves what's called the *but-for* test. The test requires answering questions like "Would a harm to Anthony not have happened but for Ethan's actions?" The fundamental problem of causal inference says we can never know for sure, since the world in which Ethan did not take his action is counterfactual, so we don't know what happens to Anthony in that world. Instead, what we've just said, and will cover in much more detail in the rest of the book, is that there are methods for answering a slightly different question like "On average, when people take actions of the sort Ethan took, does it tend to cause harm to other people?" A convincing answer to that latter question may or may not be compelling in a court that wants to answer the former.

Part of clear thinking about causal relationships involves admitting that sometimes we cannot answer certain questions with complete confidence, even when those questions are very important.

Conceptual Issues

Causality is a deep and difficult topic. The counterfactual definition of causality doesn't provide all the answers. But it can help us think more clearly about some thorny conceptual issues. Let's talk through a few of these.

What Is the Cause?

One frustration people sometimes feel with regard to the counterfactual approach is that some of the causal questions that we are accustomed to asking appear incoherent within the counterfactual framework. Think of questions like the following: Why did housing prices tank during the latest financial crisis? Why did the Chicago Blackhawks win the Stanley Cup? What caused World War I? Questions of causal attribution like these are common. But when causation is defined in terms of counterfactual comparisons, they don't make a ton of sense.

Let's think about World War I. A common claim is that World War I was caused by the assassination in 1914 of Archduke Ferdinand, the heir to the throne of Austria-Hungary. The assassins were part of a movement that wanted Serbia to take control over the southern Balkans, including Bosnia and Herzegovina, which Austria-Hungary had annexed in 1908. The government of Austria-Hungary responded to the assassination with the July Ultimatum, a list of demands so onerous they were certain to be rejected by the Serbian government. When the ultimatum was rejected, Austria-Hungary declared war on Serbia, leading Russia to mobilize its army to defend Serbia. In response, Germany (an ally of Austria-Hungary) declared war on Russia, France (an ally of Russia) declared war on Germany, and the whole mess cascaded into World War I. Thus, the claim goes, the assassination of Archduke Ferdinand caused World War I.

Now, there is a sense in which this claim is perfectly simple to think about in our framework. We can ask, In the counterfactual world in which Ferdinand was not assassinated, would World War I still have occurred? If World War I would not have occurred in that counterfactual world, then it seems right to say that the assassination had an effect on war breaking out. But that is a far cry from saying that the assassination of

the archduke was *the* cause of the war. Surely, there are many factors that, had they been different, would have prevented World War I from being fought. Sure, had Archduke Ferdinand not been assassinated, maybe the war wouldn't have been fought. But also, had Austria-Hungary not annexed Bosnia and Herzegovina, perhaps Ferdinand would have never been assassinated and the war would have never been fought, so the annexation was just as much a cause as the assassination. Similarly, had the Serbian government complied with the July Ultimatum, perhaps the war would have been avoided, so the noncompliance with the ultimatum was also a cause. And to further illustrate how many such causes there are, had some fish-like creature in the Paleozoic Era swam left instead of right, perhaps the human race as we know it would not exist, and again, World War I would have never been fought. Or, to take an example with some historical gravitas, the seventeenth-century French mathematician Blaise Pascal, reflecting on Mark Antony's attraction to a long proboscis, quipped, "Cleopatra's nose, had it been shorter, the whole face of the world would have been changed."[1] This led James Fearon, in an essay on counterfactual reasoning, to ask, "Does this imply that the gene controlling the length of Cleopatra's nose was a cause of World War I?" As you can see, then, the problem isn't that it is false that the assassination of Archduke Ferdinand caused World War I. Rather, since so many factors appear to have caused World War I, talk of one single cause seems pointless and misguided.

Once we start thinking about counterfactuals, it becomes pretty clear that things have lots of causes. That makes it hard to answer "What is *the* cause" questions. Instead, it pushes us to ask "Was this a cause" or "Did this have an effect" questions. This is perhaps disappointing.

One thought you might have, in response, is that surely some causes of a phenomenon are more important or more proximate than others. If that is true, perhaps we can still talk about the *important* or the *proximate* causes of World War I. How might we do this?

An approach that some philosophers advocate goes something like this. Imagine all the counterfactual worlds in which World War I did not occur. Some of these counterfactual worlds are very different from the actual world—for instance, World War I probably doesn't occur in many counterfactual worlds in which there is no gravity. Others are quite similar to the actual world—perhaps World War I doesn't occur in a world identical to ours through June 27, 1914, but in which Archduke Ferdinand overslept on June 28. We learn about the proximate causes of World War I by comparing the actual world to the counterfactual world in which World War I did not occur that is most similar to the actual world. This kind of analysis may allow us to give reasonable-sounding answers to "What is *the* cause" questions without abandoning our definition of causation based on counterfactual comparisons. For instance, it seems reasonable to think that the assassination of Archduke Ferdinand is a more proximate cause of World War I than is Cleopatra's nose, the laws of gravity, or the whims of Paleozoic fish.

There is certainly something to this approach. But, that said, it is often hard to assess the importance or proximity of one cause versus another in a principled way. If you know a bit of history, you surely can come up with other causes of World War I that seem equally proximate. For instance, many scholars have argued that early-twentieth-

[1] Antony and Cleopatra's love affair had major repercussions for world history. For instance, historians generally believe that the end of the Roman Republic and the establishment of the Roman Empire were ensured when Antony and Cleopatra were defeated by Octavian (later, Emperor Augustus) at the Battle of Actium. Had this not occurred, who knows how differently the rest of western history might have played out?

century military doctrines favoring offensive over defensive strategies played a role in causing World War I. Is the world in which a slightly different military doctrine was adopted more proximate to our world than the one in which Archduke Ferdinand was not assassinated? For that matter, is the world in which one Paleozoic fish took a different turn really such a large leap? It's hard to say.

To see the problem in a somewhat less lofty and perhaps more familiar setting, consider an NCAA Division III women's basketball game between the Chicago Maroons (where some of our star students are also star athletes) and the Emory Eagles. Suppose the Maroons are trailing the Eagles by one point, and the Maroons have just enough time left to take one final shot. They make it, winning the game by one point (in basketball, field goals are worth at least two points). The next day, the *Chicago Maroon* newspaper will fixate on that last shot, and the reporter might even write that the last shot was *the* reason the Maroons won.[2] But think about this counterfactually for a moment. Dozens of shots throughout the game were pivotal. Plausibly, every shot the Maroons made was pivotal—in a counterfactual world in which they missed that shot and everything else played out as it did in the actual world, they would have lost instead of won. Similarly, every shot the Eagles missed was pivotal—in a counterfactual world in which they made it and everything else played out as it actually did, they would have won instead. So what's so special about that last shot? One possibility is that everyone knew that the final shot would be pivotal when it was taken. But very few other causes meet this criterion, certainly not the assassination of Archduke Ferdinand. So, in our view, there is no obvious reason to think that the last shot was a more important cause of the Maroons' victory than the other shots. Instead, we think this example illustrates a basic, if frustrating, fact of life: individual events can have many equally important and consequential causes.

Another surprising fact about the counterfactual approach is that, at least in principle, it's possible for some event to have no causes at all. Suppose that the authors of this book concoct the perfect crime. We both shoot and kill our sworn enemy at the same time, knowing that either bullet would be fatal on its own. When questioned, Anthony says, "Clearly, I can't be charged with a crime. My actions had no effect whatsoever. Had I not fired my gun, the victim would still have died." And similarly, Ethan retorts, "I could not have possibly caused the victim's death either. Had I not shot my gun, he would have still died." While the justice system might not be impressed by our defense, the counterfactual logic is sound. Some events may be the result of a confluence of factors whereby no single factor could have changed the outcome. This theoretical possibility is yet another reason that it might not make much sense to ask questions like "What caused World War I?" It could well be that, for all the factors we like to talk about, taking away any one of them would in fact not have sufficed to prevent the war.

Causality and Counterexamples

One common skeptical reaction to evidence showing the existence of an average effect is to point to counterexamples. Perhaps you've had an experience like the following at a family gathering. You read a study showing that, on average, flu shots reduce the risk of contracting the flu. You mention this over Thanksgiving dinner, encouraging

[2] We know it's confusing that the basketball players are the Maroons, the newspaper is the *Maroon*, and probably neither sports teams nor newspapers should be named after a color. Our university is typically not known for athletics or branding.

your loved ones to get the vaccine. But your vaccine-skeptic relative says, "I don't know, I got the flu shot last year and I still got the flu." Many people nod and agree, perhaps pointing out that their friend so-and-so also got the flu shot and still got sick.

The intuition behind this kind of objection-by-way-of-counterexample is something like this: "If flu shots really prevent the flu, then no one who got a flu shot would get the flu. Thus, my one counterexample means the vaccine doesn't work."

This argument does not reflect clear thinking. The evidence says that the flu shot caused flu risk to go down, averaging across lots of people, each with their unique biology, level of flu exposure, environment, and so on. It doesn't say that it eliminated flu risk for each and every individual. But to get flu risk to go down on average, the flu shot must have prevented the flu (i.e., had a causal effect) for at least some people. We just don't know exactly which ones experienced the effect.

Let's think about this in our potential outcomes notation. Think of the potential outcomes as whether or not you get the flu. We'll say $Y = 1$ if you stayed healthy and $Y = 0$ if you got the flu. And think of the treatment as whether you got the flu shot, with $T = 1$ meaning you got the shot and $T = 0$ meaning you didn't.

Maybe there are three different kinds of people—call them the *always sick*, the *never sick*, and the *vaccine responders*. The always sick and the never sick have potential outcomes that don't respond to treatment. The always sick get the flu regardless of whether they get the flu shot, and the never sick never get the flu. In our notation,

$$Y_{1,\text{always sick}} = 0 \qquad Y_{0,\text{always sick}} = 0$$

and

$$Y_{1,\text{never sick}} = 1 \qquad Y_{0,\text{never sick}} = 1$$

But the vaccine responders are different; they get the flu if they don't get the shot, and they don't get the flu if they do get the shot:

$$Y_{1,\text{vaccine responder}} = 1 \qquad Y_{0,\text{vaccine responder}} = 0$$

In a population made up of these three groups of people, getting the flu shot reduces the probability you will get the flu. That is, on average, the treatment effect is positive. You don't know which group you are in. There is a chance you are a vaccine responder. So getting a flu shot reduces your probability of getting sick.

Let's see this in an example. Suppose there are 10 individuals. Individuals 1–5 get the flu shot, while individuals 6–10 don't. Individuals 1, 3, 4, 5, and 8 are always-sick types, so they get the flu. Individuals 5, 6, 7, and 10 are never-sick types, so they stay healthy. Individuals 2 and 9 are vaccine responders. Individual 2 gets the flu shot, so she stays healthy. But individual 9 does not get the flu shot, so he gets sick.

Table 3.2 shows potential outcomes and treatment effects. As we can see, not everyone in this population has a positive treatment effect. But the average of the treatment effects across these 10 individuals is $\frac{2}{10}$ because two of the ten are vaccine responders. So, for any individual, not knowing which type of person they are, there is a 20 percent chance that taking the flu shot will prevent them from getting the flu.

Importantly, pointing to one counterexample is neither here nor there with respect to such evidence. Perhaps your unlucky relative was a person, like individual 1, 3, or 4, whose confluence of circumstances were such that the flu shot didn't have an effect (i.e., they were an always sick). That doesn't mean it didn't have an effect for other people.

Table 3.2. Potential outcomes for flu with and without the flu shot. For each individual, the actual outcome that we can observe is in bold type. The counterfactual outcome that we do not observe is in regular type.

		Health with Flu Shot Y_{1i}	Health without Flu Shot Y_{0i}	Treatment Effect for Individual i $Y_{1i} - Y_{0i}$
Flu Shot	**Individual 1** (always sick)	**0**	0	0
	Individual 2 (vaccine responder)	**1**	0	1
	Individual 3 (always sick)	**0**	0	0
	Individual 4 (always sick)	**0**	0	0
	Individual 5 (never sick)	**1**	1	0
No Flu Shot	**Individual 6** (never sick)	1	**1**	0
	Individual 7 (never sick)	1	**1**	0
	Individual 8 (always sick)	0	**0**	0
	Individual 9 (vaccine responder)	1	**0**	1
	Individual 10 (never sick)	1	**1**	0

And it doesn't even mean that the flu shot won't prevent the flu for that same relative next year or that it won't help you. Absent any further information about which group they are in, any individual's best guess is that the flu shot will reduce their chances of contracting the flu since it does so on average. And we haven't even discussed the more complicated issue that outcomes aren't actually binary, so the shot may have a causal effect on the severity of the flu.

Of course, the possibility that effects are different for different people presents another set of important conceptual challenges. We might be able to detect such *heterogeneous treatment effects*, especially if they correspond with observable categories (e.g., men versus women, older versus younger, healthy versus sick). To identify such heterogeneous effects, we could run a separate experiment for each group, which would tell us the average effect for each group rather than for the whole population. But what if effects differ across people for complicated or obscure reasons that might never occur to us? Then, when we go to look at the effect of some intervention, it is very important to keep in mind that we are learning about an average effect. Some people may have effects much larger than the average. Others may have effects much smaller than the average. Indeed, some people may have no effect at all or an effect in the opposite direction from the average. If we don't know the source of this heterogeneity, all we will

be able to say is something about the average, which, as we've discussed, may still be valuable.

Causality and the Law

As we briefly mentioned previously, one place where philosophical questions about causality become of serious practical import is in the law. Administering justice requires assigning blame and assessing liability. If we want to know whether, say, Ethan should be held liable for some harm suffered by Anthony, surely we need to know whether Ethan's actions caused that harm. But, as we've just discussed, talking about causes in this way is conceptually fraught. Many things, from the behavior of a Paleozoic fish to Ethan's alleged negligence, may have had a causal effect on the harm Anthony suffered. Is the fish liable too?

The law is aware of the philosophical conundrum. But it must ultimately come up with some pragmatic resolution that allows judges and lawyers to get on with the business of administering justice. Here's, roughly, where it comes down.

In the Common Law, causality is thought of in terms of two conditions that are closely related to things we've talked about. These are referred to as *cause-in-fact* and *proximate causality*.

Cause-in-fact is essentially counterfactual causality. Whether Ethan's actions are a cause-in-fact of Anthony's suffering is determined by whether Anthony wouldn't have suffered *but for* Ethan's actions.

Of course, as you already know, a counterfactual standard like the but-for test isn't very stringent. World War I wouldn't have happened but for a Paleozoic fish turning the wrong direction. Does that mean we should blame the poor fish for World War I?

The law's answer is no. The fish is off the hook, so to speak. This is where proximity comes in. For there to be liability, the law requires that some cause-in-fact be close enough in the causal chain. This thought is also familiar—for instance, from our argument that the assassination of Archduke Ferdinand is a more proximate cause of World War I than is Cleopatra's nose.

So an assessment of legal causality might go something like this. Suppose you order food delivery and the delivery person drives recklessly, crashing into your neighbor's car. Are you liable for your neighbor's suffering? It is plausible that, but for your decision to order delivery, the delivery person wouldn't have been in the area and your neighbor's car wouldn't have been hit. So your actions are probably a cause-in-fact of your neighbor's suffering. But there are many steps in the causal chain between your actions and the car crash, all of which are out of your hands. So the law would not find you liable for the damage to your neighbor's car.

Of course, as we've discussed, knowing exactly how to apply the conditions of cause-in-fact and proximate causality is tricky. To apply the but-for test, we have to know what the right counterfactual world is. And defining how close is close enough for a proximity test is a fraught problem, full of judgment calls. All of which is to say that these questions about causality are vexing and of great practical importance.

Can Causality Run Backward in Time?

One common intuition is that causality must run forward in time. That is, an event that happens now can have an effect on events that happen in the future. But surely, the thought goes, events that happen in the future can't affect events in the past. Indeed,

one common strategy for trying to establish a causal relationship is to show that the supposed cause typically occurs prior to the supposed effect.

Let's check this intuition by thinking about birthday cards. Here's a correlation that we hope is true in the world: the number of birthday cards that get mailed to you in a given week is strongly correlated with it being within a week of your birthday. That is, many more birthday cards are mailed to you in the week before your birthday than in any other week of the year.

Now, although correlation need not imply causation, we suspect that there is a causal relationship here but not the one that's implied by thinking of causal relationships as running forward in time. Receiving birthday cards does not cause your birthday to occur. In a counterfactual world in which those cards were sent at a different time, or even in a counterfactual world in which greeting cards cease to exist, your birthday will still occur on the date you were born. Instead, you might say the causal relationship runs backward in time. Your birthday exerts an effect on the sending of birthday cards. In the counterfactual world in which your birthday occurs in a different month, you will be sent fewer birthday cards in the week preceding your birthday in this world. Thus, on our counterfactual definition, your birthday exerts a causal effect on birthday cards. Causality appears to run backward in time.

There are objections to this line of argument. For instance, one might argue that it isn't your future birthday, but anticipation of that birthday, that exerts a causal effect on the sending of birthday cards. If we changed people's beliefs about whether your birthday is coming up, we'd change their card-sending behavior. But if we changed your actual birthday, without a change in their beliefs, the cards would still be sent. On this argument, causality is operating forward in time, in the intuitive way.

Even that need not be the end of the argument. After all, where did the anticipation of your birthday come from? It presumably came from the fact of your actual birthday. If we changed the fact of your actual birthday in the future, we'd change people's anticipation of your birthday now (which would, in turn, change their card-sending behavior). Perhaps we are back to causality running backward in time. Or perhaps not. Is it really the changing of your birthday in the future that affects people's anticipation today? Or is it telling them about the change in your future birthday, in which case we are right back to causality running forward in time.

As you can no doubt tell by this point, we aren't going to solve this issue here. But we do want you to see two things clearly. First, evidence that one thing occurred before another is not, on its own, convincing evidence that the one caused the other. Second, whether or not you think causality can or cannot run backward in time, we can always define the causal effects in terms of a counterfactual.

Does Causality Require a Physical Connection?

Another intuition many people share is that causation necessarily has to do with physical connection—a view that we'll refer to as *physicalism*. One billiard ball affects another by bumping into it. Maybe such physical connections always underlie causal relationships.

While, of course, there are many examples of causal effects that occur through physical connection, there are good arguments to suggest such physical connection is not required. Think of a person who is deterred from robbing a bank by worry about imprisonment. Such a person's behavior is affected by the existence of the police, the courts, the penal code, and the prison system. The criminal justice system affects whether this person commits a crime, even though there is no physical connection between them.

Indeed, think of our previous discussion of the effect of birthdays on the sending of birthday cards. Birthdays aren't a physical thing in the world at all. It is hard to see what it would even mean for the causal relationship between birthdays and the sending of birthdays cards to occur through physical connection.

A defender of physicalism might say that with enough creativity, we can describe the effect of the criminal justice system on crime in purely physical terms. Perhaps the past arrest and conviction of people who committed crimes led reporters to write about this activity in newspapers, which led the person in question to read about these arrests in the newspaper, which, through a complicated sequence of light hitting the person's eyeballs, led to lots of chemical and electrical connections in that person's brain, which deterred them from committing a crime. You could do a similar exercise for birthdays and birthday cards.

Again, we aren't going to provide a definitive answer. There may be reasonable arguments on both sides of the physicalism debate. The important point is that we can think about counterfactually defined causal relationships that do not depend on anything like the simple, commonsense kind of physical connections suggested by the billiard ball example.

Causation Need Not Imply Correlation

We've agreed that correlation need not imply causation. But, perhaps more surprisingly, causation also need not imply correlation and certainly not correlation in the expected direction. There are many situations in which some feature of the world has (say) a *negative* effect on some other feature of the world, but those two features of the world are *positively* correlated (or vice versa).

You'd probably find a strong, positive correlation between the number of firefighters who have recently visited a house and the amount of fire damage to that house. But if we had to guess, we'd suspect that firefighters, on average, reduce fire damage. In other words, if fewer firefighters had visited, we suspect there would be even more fire damage.

So why is the correlation positive? Firefighters tend to visit houses that are on fire. So, although firefighters reduce fire damage to some degree, the houses that have been visited by firefighters tend to have more fire damage. Hence, not only should one not conclude from a correlation that there must be a causal relationship, but one also should not assume that just because a causal relationship exists, the correlations found in the world will correspond to those causal relationships in some straightforward way.

Wrapping Up

Understanding whether a causal relationship exists is one of the fundamental goals of quantitative analysis. But, if we are going to do that, we need to think clearly about what causality means.

We believe that the best way to conceptualize causality is through a thought experiment involving counterfactuals. A treatment has a causal effect on an outcome if the outcome would have been different had the treatment been different. Of course, in the actual world, the treatment was what it was. We can't observe the counterfactual world in which the treatment was different in order to figure out if the outcome would have been different. This is the fundamental problem of causal inference.

The fact that causal effects are unobservable doesn't mean data analysis cannot help us learn about them. In particular, we can learn about the average effect in some population, even though we can't observe any of the individual effects directly.

Doing so involves making careful use of quantitative knowledge about things like correlations. In part 2 we turn to a more detailed discussion of how we establish and quantify correlations. This will set us up to be able to think clearly in part 3 about estimating causal effects.

Key Terms

- **Causal effect:** Informally, the change in some feature of the world that would result from a change to some other feature of the world. Formally, the difference in the potential outcomes for some unit under two different treatment statuses.
- **Body Vibes:** Stickers that a company called Goop claims cause clear skin. The authors of this book do not endorse Body Vibes, mainly because we will be releasing our own competitor: Brain Vibes. One sticker applied to the temple causes clear thinking.
- **Counterfactual comparison:** A comparison of things in two different worlds or states of affairs, at least one of which does not actually exist.
- **Treatment:** Terminology we use to describe any intervention in the world. We usually use this terminology when we are thinking about the causal effect of the treatment, so we want to know what happens with and without the treatment. Importantly, although it sounds like medical terminology, *treatment* as we use it can refer to *anything* that happens in the world that might have an effect on something else.
- **Potential outcomes framework:** A mathematical framework for representing counterfactuals.
- **Potential outcome:** The potential outcome for some unit under some treatment status is the outcome that unit would experience under that (possibly counterfactual) treatment status.
- **Fundamental problem of causal inference:** This refers to the fact that, since we only observe any given unit in one treatment status at any one time, we can never directly observe the causal effect of a treatment.
- **Heterogeneous treatment effects:** When the effect of a treatment is not the same for every unit of observation (as in the case of flu shots and virtually every other interesting example of a causal relationship), we say that the treatment effects are heterogeneous. Sometimes we're still interested in the average effect even though we know the treatment effects are heterogeneous, and sometimes we want to explicitly study the nature of the heterogeneity. (In contrast, when discussing the unlikely possibility that treatment effects are the same for every unit, we would refer to *homogeneous* treatment effects.)

Exercises

3.1 Sarah says that she is hungry. John hands her a piece of pizza. Sarah eats the pizza and then declares that she is no longer hungry.

(a) The fundamental problem of causal inference seems to say that you can't know that Sarah eating the pizza had a causal effect on her no longer being hungry. Is that right? Explain.

(b) Do you think you nonetheless have good reasons to believe that eating the pizza had an effect on Sarah no longer being hungry? Explain why or why not.
(c) Do you have good reasons for believing that John handing Sarah the pizza had a causal effect on her no longer being hungry? In your assessment, are the reasons to believe John's actions had a causal effect better or worse than the reasons to believe Sarah eating the piece of pizza had a causal effect?

3.2 A government is considering making alcohol consumption illegal as part of a public health campaign. Let's think of making alcohol illegal as the treatment T. Write $T = 1$ if the government makes alcohol illegal and $T = 0$ if the government leaves alcohol legal.

We will think of a binary outcome for each person: either they drink alcohol or they do not. If person i drinks at treatment status T, we write her potential outcome as $Y_{Ti} = 1$, and if she doesn't drink, we write it as $Y_{Ti} = 0$.

Suppose the society is made up of three groups: the always drinkers, the legal drinkers, and the never drinkers. The always drinkers will drink whether or not alcohol is legal. The legal drinkers will drink if and only if alcohol is legal. The never drinkers won't drink whether or not alcohol is legal.

(a) Write down, in potential outcomes notation and as a number (0 or 1), each of the two potential outcomes for each of the three groups.
(b) Write down, in both potential outcomes notation and as a number (0 or 1), the causal effect of making alcohol illegal on drinking for each of the three groups.
(c) Is there an effect, on average, of banning alcohol in this society?
(d) Suppose you are out to lunch with some friends and one of them says, "My uncle lives in a place where they banned alcohol and all of his friends kept drinking, so I don't think the ban does anything." Explain, in terms of our example, why this isn't a convincing argument.

3.3 The Republican National Committee (RNC) has hired three consultants and asked them to figure out the cause of their loss in the 2020 presidential election. The first consultant says that they didn't do enough television advertising. The second consultant reports that they should have encouraged more of their supporters to vote rather than criticizing voting by mail. The third consultant concludes that Donald Trump should have done a better job responding to the COVID-19 pandemic and should have shown more compassion on the campaign trail. Confused by the apparently conflicting information, the RNC hires you, a quantitative analyst, to adjudicate between these three possibilities. What would you tell them? How would you proceed?

3.4 In the 2016 U.S. Open golf tournament, Dustin Johnson was leading the tournament in the final round, and his ball was resting on the fifth green. While preparing for his upcoming putt, he tapped his putter on the ground next to the ball and the ball moved. The rules at the time stated that if we were highly certain that a player caused his ball to move, even if it was inadvertent, he or she should incur a penalty. Because you're an expert on causation, the rules

officials call you in to evaluate the situation. The officials make the following arguments. Please provide your expert response to each one.

(a) Johnson couldn't have possibly caused the ball to move, because he (and his putter) never touched it.
(b) Johnson shouldn't receive a penalty because the true cause of the ball moving was the greenskeeper. Had the greenskeeper not cut and rolled the greens so much that morning, the ball wouldn't have moved.
(c) An empirically minded official went out to the same green, placed a ball down, tapped his putter on the ground next to the ball, and it didn't move. Therefore, Johnson's actions couldn't have caused the ball to move.
(d) One official was watching the incident up close and says he's virtually certain that if Johnson had not tapped his putter next to the ball, it wouldn't have moved. Therefore, he caused it to move and should receive a penalty.

Readings and References

You can read about Body Vibes on the Goop website. We last accessed it on June 15, 2020. http://goop.com/wearable-stickers-that-promote-healing-really/

The quote from Blaise Pascal on Cleopatra's nose is from his seventeenth-century collection entitled *Pensées*.

The essay about counterfactual reasoning discussing the gene controlling the length of Cleopatra's nose is

James D. Fearon. 2011. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43(2):169–195.

If you'd like to read more about the counterfactual definition of causality, potential outcomes, and surrounding discussions and debates, have a look at these:

David Lewis. 1973. "Causation." *Journal of Philosophy* 70:556–67.

Paul W. Holland. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–60.

Stephen Mumford and Rani Lill Anjum. 2014. *Causality: A Very Short Introduction.* Oxford University Press.

There is also a nice entry by Peter Menzies and Helen Beebee in the *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/entries/causation-counterfactual/.

CHAPTER 9

Why Correlation Doesn't Imply Causation

What You'll Learn

- Correlation does not necessarily imply causation.
- There are two key reasons why an observed correlation might be a biased estimate of a causal relationship: confounders and reverse causation.
- If we think clearly, we can sometimes sign this bias.
- There is an important distinction between confounders and mechanisms.

Introduction

As we discussed in chapters 2 and 3, information about correlations and information about causal relationships are useful for different purposes. Knowledge of correlations, on its own, can help us describe the world and forecast the presence of certain features of the world on the basis of the presence of other features of the world. Knowledge of causal relationships is particularly valuable for decision making because it can tell us how the actions we take will affect the world. Remember our definition of causation from chapter 3. A *causal effect* is a change in some feature of the world that would result from a change to some other feature of the world.

So when we say that some action has a causal effect on some outcome, we're asserting that the outcome would be different in a counterfactual world in which the action was different. Knowing the effects of our actions allows us to anticipate and weigh their costs and benefits.

From a pragmatic perspective, this distinction is why the maxim "Correlation doesn't imply causation" is so important. If we mistakenly take knowledge of a correlation as implying knowledge of a causal relationship, we might end up making big mistakes, taking actions because of a misguided belief about how those actions will affect outcomes we care about. In this chapter, we are going to learn to think clearly about the difference between correlation and causation, discuss the sources of bias that can make correlations unreliable estimates of causal effects, and start considering what that means for how we learn about causal effects.

To give you a sense of how important this topic is, let's talk through an example where high-stakes decisions are being made about how to deploy resources and where we can disentangle correlation from causation with some confidence. The example concerns the topic of charter schools in the United States.

Charter Schools

In his heart-rending movie *Waiting for Superman*, David Guggenheim tells the story of several young children from poor families. In each case, a child is enrolled in a substandard public school. And, in each case, the parents are working hard to get their child into a charter school (or, in some cases, a magnet school).

Charter schools are operated at public expense but independently of the public school system. Some charter schools are run by not-for-profit organizations and others by for-profit corporations. The idea behind the charter school movement is to encourage innovation and choice. Charter schools are free from some of the constraints (e.g., union contracts, legacy curricula) that public schools face. Hence, the argument goes, they can innovate in curriculum, teacher incentives, and the like in ways that regular public schools cannot. And, because they have to compete for students, they will be motivated to come up with new and potentially better approaches to education. Whether charter schools in fact succeed in improving educational outcomes is a matter of heated debate.

In many areas, there are more kids applying to attend charter schools than can be accommodated. By law, when a charter school is oversubscribed it must admit students by random lottery. Families apply to the school, and after that, luck determines which kids get the coveted spots. As the movie powerfully illustrates, the odds are stacked against the children. Some of the charter schools have hundreds of applicants for a couple dozen spots.

During the course of the movie we are told a slew of facts about the performance of the charter schools to which the students are seeking admission. Compared to public schools with socioeconomically similar populations of students, these charter schools have better test scores, higher graduation rates, less crime, and so on. Indeed, the charter schools featured in the movie perform much better than public schools with respect to virtually every measurable outcome.

As the movie ends, we discover that few of the students we've been following were admitted to the school of their choice. Instead, they will be enrolled in "failure factories" where, we are left to believe, in all likelihood, their potential will be wasted.

But is this the right inference? Does getting into a charter school really improve a child's educational outcomes? There is more at stake here than our feelings about the kids in the movie. Over the past several decades, charter schools have emerged as one of the dominant approaches to school reform in the United States. Expansion of charter schools as an alternative to traditional public schools has received bipartisan support—it was, for instance, pushed aggressively by both the Bush and Obama administrations. The share of public school students attending charter schools has risen from less than 1 percent in 1999 to more than 6 percent in 2021. But critics raise concerns about the possibility that this expansion has led to a decline in resources available to traditional public schools, perhaps harming the students enrolled in those schools. So we'd really like to know whether charter schools have a positive effect on academic outcomes for students.

Here's what we know. There is definitely a correlation between charter school attendance and academic performance. Within a city, on average, low-income students who go to charter schools have better educational outcomes than low-income students who go to traditional public schools.

As an example, consider the Preuss School, a charter school created by the University of California at San Diego. The Preuss School serves low-income middle and high

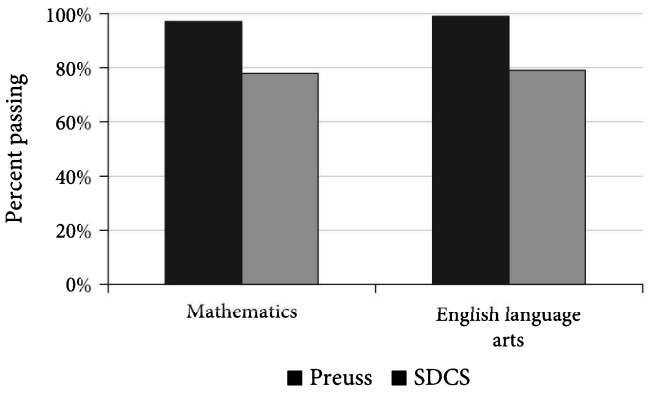


Figure 9.1. Standardized test scores at Preuss School and San Diego City Schools (SDCS).

school students from all over San Diego County. By all accounts, it is a remarkable school, sending almost 100 percent of its students, who typically come from families with no history of college attendance, on to higher education. The school has garnered praise from many sources, including *Newsweek* magazine, which named Preuss the "top transformative high school" in the country multiple times.

And, indeed, it is certainly the case that students at the Preuss School perform much better than their peers at the San Diego public schools. For instance, look at the data in figure 9.1, showing the difference in the percentage of students who pass standardized tests in math and English at Preuss versus the San Diego City Schools (SDCS).

Just like in the stories from *Waiting for Superman*, from these data, it sure looks like the Preuss School is having a huge impact on the academic performance of its students.

But before we jump to conclusions, let's think about this a little more clearly. These data show a positive correlation between going to Preuss and academic performance. But does the fact that students going to Preuss (or other charter schools) perform better academically than students at public schools imply that going to a charter school is *causing* them to perform better? That is, in the counterfactual world in which some other kids go to Preuss and these kids go to the public schools, would those other kids perform better academically and the current Preuss students perform worse? Does the correlation imply causation? That's what we need to know if we are going to figure out whether investing resources in charter schools is a good decision.

Of course, it could well be that charter schools are genuinely causing students to perform better, meaning that in the counterfactual world where the charter school students had attended regular public schools, they would have performed worse and their replacements from the public school would have performed better. But another possible explanation, as *Waiting for Superman* so eloquently illustrates, is that the students and families who choose charter schools are themselves different from the average public school student in important ways. That is, perhaps the explanation for the correlation isn't a better school, but better students. If that is the case, then, in the counterfactual world in which these better students had not gone to a charter school, can we be so sure that they wouldn't still have outperformed their peers?

Ask yourself, Under what circumstances are economically disadvantaged parents likely to sign up their child for the lottery to get into a charter school? Two circumstances occur to us. First, if the parents believe their child is particularly talented, they

might be particularly motivated to get their child into a school with a good reputation. Second, if the parents themselves are particularly invested in their child's education, they might be more likely to do the work necessary to get that child into the lottery.

Natural talent and parental involvement are themselves pretty important determinants of student achievement. Suppose that the pool of students in the charter school lotteries (and, hence, at those schools) are, on average, more talented and come from families more interested in education than the population at large. Then, even if the charter schools themselves have no effect on the performance of their students, those students would nonetheless outperform the general population simply by virtue of their greater ability and more supportive family. Put differently, if all children went to the exact same school, the children who are currently at the charter schools would still be above average because they are more talented and have more dedicated parents.

Remember the question we care about: Can sending a child to a charter school be expected to improve the performance of that child relative to what they would have achieved at their local public school? The discussion above shows that we can't know the answer to this crucial question by comparing the performance of students currently in charter schools versus traditional public schools. We'd be comparing a group of particularly talented, ambitious kids from highly dedicated families to the general population. How would we know whether differences between these groups arose because of the effects of charter schools, because of underlying differences between groups, or both? In simpler terms, we'd be comparing apples to oranges.

To determine whether the charter schools are actually a cause of their students' excellent outcomes, we need to make a comparison that comes closer to capturing the counterfactual nature of causality. To do this, we need a way to compare apples to apples. The ideal question we'd like to answer is something like this: If everything else about two children were identical, would the child who went to the charter school do better than the child who went elsewhere? We obviously can't answer this question. But we can get closer by trying to answer something like this: If everything else about two groups of children was identical on average, would the kids who went to charter schools do better on average than kids who went to public schools?

To take a shot at this latter question, we have to move beyond just comparing charter school kids to all other public school kids. We do so by narrowing our focus to just the children who tried to get into charter schools. All of those children were promising enough or had families dedicated enough to apply to a charter school. But, as a result of the admissions lottery, some lucky students got into the charter school and others did not. Since the lottery was random, the pool of lottery winners and the pool of lottery losers should have the same characteristics, on average (that is, if we ran the lottery over and over again, the kids winning the lottery would be no more or less motivated or talented than those who lost). So we can learn a lot more about the actual effect of attending a charter school by comparing the academic performance of those who entered and won the admissions lottery to those who entered but lost the lottery. If the positive correlation between charter school attendance and academic performance is still there in this narrower comparison, we will feel much more secure in giving it a causal interpretation because now we're comparing apples to apples.

This comparison has been done for many charter schools. Let's start by looking at what happens when you make that comparison for the Preuss School. We don't have data on this comparison for the same standardized test as in figure 9.1, but we do for another important standardized test, shown in figure 9.2.

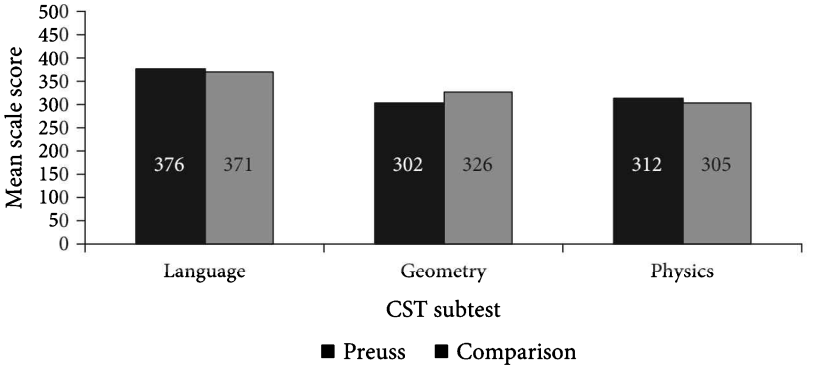


Figure 9.2. Comparing standardized test scores of kids who did and did not win the Preuss School admission lottery.

This comparison demonstrates the importance of comparing apples to apples. Yes, the students at the Preuss School tend to outperform the students in the San Diego City Schools as a whole. But when lottery-winning students are compared to lottery-losing students, the correlation disappears—there is no performance difference.

Similar findings emerge from studies of many other charter schools. To be sure, different studies find different things. Researchers in Boston find that kids admitted to charter schools run by the Knowledge is Power Program do better than kids who applied to but lost the lottery. But our sense is that the following, from another study of school choice programs in San Diego, is more typical of the literature:

> In the vast majority of cases, we found no evidence that winners and losers of a given lottery fared differently in these achievement tests one to three years after the admissions lottery was conducted. We interpret this to mean that winning a lottery neither helps nor hurts achievement growth.

Think about what this means. When we make an apples-to-apples comparison, the high-performing charter schools appear to have little to no effect on student performance. Most of the apparent effects of these schools come from the fact that the students who enter the charter school lotteries are already academically different from the average student. Those students would have done better than average anyway. Separating correlation from causation in this way may change your views about how we should spend education resources.

Thinking Clearly about Potential Outcomes

When can a correlation between two variables be plausibly interpreted as compelling evidence of a causal relationship? We just saw an example of how we might mistakenly think a correlation indicates causation and how it can really matter for decision making. But let's try to be a bit more systematic about why correlations aren't always evidence for causation, so we can think more clearly about when we do and don't have a credible estimate of a causal relationship.

Remember, we define causal relationships counterfactually. In chapter 3, we introduced the notion of potential outcomes, which helps us think more clearly about such counterfactuals.

Let's suppose we're trying to estimate the effect of going to a charter school on academic performance, as measured by standardized test scores. So the *outcome* of interest is standardized test scores, and the *treatment* of interest is attending the charter school.

Represent the outcome, standardized test scores, with Y. And represent the treatment, going to the charter school, with a binary variable T. If $T = 1$ for some individual, that means they attended the charter school. If $T = 0$ for some individual, that means they attended a public school. We sometimes say that a unit with $T = 1$ is *treated* and a unit with $T = 0$ is *untreated*, although it's often arbitrary which groups are labeled *treated* versus *untreated* (e.g., we could similarly talk about the effect of attending traditional public schools versus charter schools).

In a metaphysical sense, for each individual there is some standardized test score that they would have gotten had they gone to the charter school and some standardized test score that they would have gotten had they not gone to the charter school. However, we only ever get to observe one of these. Nonetheless, having notation for each of these potential outcomes helps us think clearly about counterfactuals:

$$Y_{1i} = \text{outcome for unit } i \text{ if } T = 1$$

$$Y_{0i} = \text{outcome for unit } i \text{ if } T = 0$$

Using this notation, the effect of going to the charter school on person i's test scores is

$$\text{Effect of Charter School on } i\text{'s Test Scores} = Y_{1i} - Y_{0i}.$$

We say that causality is about counterfactual comparisons because we can only observe, at most, one of the two quantities—Y_{1i} or Y_{0i}—for any individual at a particular point in time. This means that we can't directly measure the effect of going to the charter school on an individual.

Maybe we can instead hope to estimate the average effect of going to a charter school across a bunch of individuals in some population of interest. For whatever population we are interested in, let's define notation for the average test score *if everyone went to the charter school* and the average test score *if everyone went to the public school* as follows:

$$\overline{Y}_1 = \text{average outcome if all units had } T = 1$$

$$\overline{Y}_0 = \text{average outcome if all units had } T = 0$$

With this notation, we can now think about the *average treatment effect* (ATE):

$$\text{ATE} = \overline{Y}_1 - \overline{Y}_0$$

Of course, we can't directly observe this average effect any more than we can observe the effect of charter schools on an individual. We never see all units both treated *and* untreated. Indeed, at any given time, each unit is only one or the other. But we can try to estimate the ATE.

A first thing we might do to try to estimate the average treatment effect is simply look at the correlation—comparing the average test scores of students who go to the charter school (treated) to the average test scores of students who go to the public schools (untreated).

Start by thinking of our population as divided into two groups, those who went to charter schools ($\mathcal{T}$) and those who went to public schools ($\mathcal{U}$). Denote the average test scores in each of these groups by

$$\overline{Y}_{1\mathcal{T}} = \text{average outcome among units with } T = 1$$

$$\overline{Y}_{0\mathcal{U}} = \text{average outcome among units with } T = 0.$$

We will refer to the difference in the average test scores between these two groups as the *population difference in means*. (Remember that *mean* is usually just another word for *average*, and in the context of this book, the two terms are used interchangeably.) It is just

$$\text{Population Difference in Means} = \overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{U}}.$$

Of course, we might not observe the whole population; we might observe just a sample. For instance, perhaps we only observe the students from one particular charter school. So the difference in average test scores that we observe in our sample is equal to the difference in average test scores among students who go to charter and public schools in the whole population plus some noise. So, we have

$$\text{Sample Difference in Means} = \underbrace{\overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{U}}}_{\text{Population Difference in Means}} + \text{Noise},$$

which is just a measure of the correlation between standardized test scores and attending a charter school in our sample.

Of course, we want to know the average *effect* of going to the charter school, not just the correlation. To start thinking about the difference between these, it helps to introduce two more concepts—the *average treatment effect on the treated* (ATT) and the *average treatment effect on the untreated* (ATU). The ATT is the average effect of going to the charter school among those students who in fact went to the charter school. That is,

$$\text{ATT} = \overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{T}}.$$

And the ATU is the average effect of going to the charter school among those students who in fact went to the public schools. That is,

$$\text{ATU} = \overline{Y}_{1\mathcal{U}} - \overline{Y}_{0\mathcal{U}}.$$

Notice two things. First, the ATE is just a weighted average of the ATT and the ATU, where the weights depend on how many kids are in each group.[1] Second, just like the ATE, the ATT and ATU are both fundamentally unobservable. We don't observe the test scores that students who go to the charter school would have made had they gone to the public schools ($\overline{Y}_{0\mathcal{T}}$). And we don't observe the test scores

[1] A weighted average is just an average where we put different weights on different items. For example, suppose 75 percent of the population is treated and 25 percent of the population is untreated; then the ATE is the weighted average of the ATT and the ATU with 75 percent of the weight on the ATT. That is,

$$\text{ATE} = \frac{75 \cdot \text{ATT} + 25 \cdot \text{ATU}}{75 + 25} = .75 \cdot \text{ATT} + .25 \cdot \text{ATU}$$

students who go to the public schools would have made had they gone to the charter school ($\overline{Y}_{1\mathcal{U}}$).

Okay, now that we have all that notation, we should be able to think clearly about the difference between correlation and causation. To start doing so, let's compare the difference in means that we in fact observe (which is our measure of the correlation) to the ATT—the effect of going to charter schools among students who went to charter schools. This will help us build some intuition that we will then be able to apply to thinking about the comparison of the difference in means to the ATU and, ultimately, the ATE.

We are going to want to get back to working with our favorite equation:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

That is, we want to find a way to write

$$\text{Sample Difference in Means} = ATT + \text{Bias} + \text{Noise}.$$

How do we do this?

Let's start by remembering, from above, that

$$\text{Sample Difference in Means} = \underbrace{\overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{U}}}_{\text{Population Difference in Means}} + \text{Noise}.$$

Now, we are going to cleverly rewrite the population difference in means by adding and subtracting $\overline{Y}_{0\mathcal{T}}$ from it. We know that seems weird. But, trust us, it's going to help. And, for now, it should at least be clear that we aren't doing any harm since, by adding and subtracting the same term, we are really just adding zero. Anyway, when we do that, we get

$$\text{Sample Difference in Means} = \overbrace{\overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{T}}}^{\text{ATT}} + \overbrace{\overline{Y}_{0\mathcal{T}} - \overline{Y}_{0\mathcal{U}}}^{\text{Bias}_{\text{ATT}}} + \text{Noise},$$

where we've subscripted *Bias* with ATT to indicate that this is the bias we get when using the difference in means to estimate the ATT.

Our algebraic trick was actually pretty cool, right? By adding and subtracting the same term, we were able to write things in terms of our favorite equation. The sample difference in means (estimate) is equal to the ATT (estimand) plus a bias term plus noise!

But what exactly does that bias term say? If we are trying to estimate the effect of going to a charter school, our comparison of test scores among students who did and did not go to the charter school is biased if we expect that those two groups of students would have made different average scores on their standardized tests even in the counterfactual world where they all went to the public schools (i.e., $\overline{Y}_{0\mathcal{T}} - \overline{Y}_{0\mathcal{U}} \neq 0$). When this is the case, we say the two groups have *baseline differences*.

So far, we've seen how the difference in mean test scores between charter public school students might be a biased estimate of the true effect of charter school attendance on those students who attend charter schools (the ATT). We could do a similar

analysis for the effect of charter school attendance on those students who attend public schools (the ATU):

$$\text{Sample Difference in Means} = \overbrace{\overline{Y}_{1\mathcal{U}} - \overline{Y}_{0\mathcal{U}}}^{\text{ATU}} + \overbrace{\overline{Y}_{1\mathcal{T}} - \overline{Y}_{1\mathcal{U}}}^{\text{Bias}_{\text{ATU}}} + \text{Noise}$$

Here we find a similar bias, but now the baseline differences we are worried about have to do with differential outcomes between the treated and untreated groups if both groups were to receive treatment (i.e., $\overline{Y}_{1\mathcal{T}} - \overline{Y}_{1\mathcal{U}} \neq 0$). Since the overall average treatment effect (ATE) is itself just a weighted average of the ATT and the ATU, the bias associated with using the difference in means to estimate the ATE comes from both of these kinds of baseline differences.

Think back to the difference in academic performance between students at the Preuss School (treated) and the students at the San Diego City Schools (untreated). We were concerned that the relationship might not be causal because, say, the students at Preuss were more academically talented, on average, than the students at the San Diego City Schools. If the Preuss students are in fact more academically talented, then there are baseline differences between the two groups of students—a difference in academic performance would exist between the treated and untreated students even if all students in both groups attended the same school (i.e., $\overline{Y}_{0\mathcal{T}} - \overline{Y}_{0\mathcal{U}} > 0$ and $\overline{Y}_{1\mathcal{T}} - \overline{Y}_{1\mathcal{U}} > 0$). Because this comparison is so clearly not apples-to-apples, we can't be certain that the difference in average performance between the two groups is evidence of an effect of the Preuss School. Even if the ATE was zero, we would still expect to find a positive difference in means. This is exactly what it means to say correlation doesn't imply causation.

The lottery was convincing evidence precisely because it randomized people into treated and untreated. Randomization guarantees that, on average, the two groups are the same with respect to potential outcomes. That is, if we ran the randomization over and over again, on average the two groups would have the same baseline outcomes. (Of course, for any one run of the randomization, there could still be non-causal differences in academic performance between the two groups, just because of sampling variation or other kinds of noise.) Hence, a difference in average outcomes between lottery winners and lottery losers is an unbiased estimate of the causal effect of the school.

When talking about causality, we often use language that evokes experiments, as we have here by discussing treatment. We do so because experiments provide a clear way to think about inferring causality from a correlation. If there is experimental randomization into treatment, then there are no systematic baseline differences between the treated and untreated groups.

Importantly, though, in many circumstances where we are interested in causality, we don't actually get to run an experiment. Instead, some people get the treatment and others do not, for reasons that we are not in charge of. In those circumstances, we have to be very careful about interpreting a correlation between outcome and treatment status as an estimate of the causal relationship. As we saw, the positive correlation between test scores (outcome) and going to the Preuss School (treatment) in the overall population was not in fact indicative of a causal relationship. The reason is because, out there in the world, social processes gave rise to baseline differences between the treated and untreated groups.

Sources of Bias

To take the proper care in interpreting correlations, we need to be able to think clearly about when there will be systematic baseline differences, because it is these baseline differences that give rise to bias. There are two main sources of such differences: *confounders* and *reverse causality*. Understanding these is a big step toward being able to think about when you can and cannot learn something credible about causality from a correlation.

Confounders

A *confounder* is a feature of the world that satisfies two conditions:

1. It has an effect on treatment status.
2. It has an effect on the outcome over and above the effect it has through its effect on treatment status.

Confounders create baseline differences and, thus, bias. Suppose some feature of the world makes people more likely to receive treatment. And suppose it also makes people more likely to have a particular outcome. Then, because of the confounder, there will be a correlation between that outcome and treatment, for reasons separate from any actual effect of the treatment. Hence, if there are such confounders (and we haven't done anything to account for them, which we will discuss in coming chapters), then it is a mistake to interpret a correlation between an outcome and a treatment as an unbiased estimate of the causal effect of the treatment.

To be a little more concrete, remember our concern that the correlation between going to the Preuss School and academic achievement was not convincing evidence of a causal relationship. That concern was about baseline differences that resulted from more academically talented kids being more likely to seek out (or have families that seek out) the Preuss School. Another way of expressing that same concern is that the underlying academic talent of a kid is a confounder. Academic talent has an effect on treatment status—kids who are more academically talented are more likely to be in the treatment group (i.e., go to Preuss). And academic talent has an effect on outcomes over and above the effect it has through its effect on treatment status—more academically gifted kids are going to do better on tests for reasons over and above the fact that they seek out better schools. Looking at the lottery winners and losers helped to tease out causality because it broke the link between talent (the confounder) and going to Preuss (the treatment).

Consider one further example. Many studies show a strong negative correlation between a country's economic productivity and whether it experiences civil war. There are reasons to think that there could be a causal relationship underlying that correlation—for instance, perhaps when the economy is doing better, people have better lives and, thus, are less likely to be willing to mobilize to fight. But before interpreting the correlation as causal, one needs to think about whether there are potential confounders. One confounder you might worry about has to do with politics. A democratic political system, by incentivizing the government to adopt better public policy, is likely to positively affect a country's economy. Moreover, by giving people a non-violent means to express various grievances, democracy may also directly reduce the risk of civil war. Hence, democracy is a potential confounder. And, so, we are not justified in interpreting the

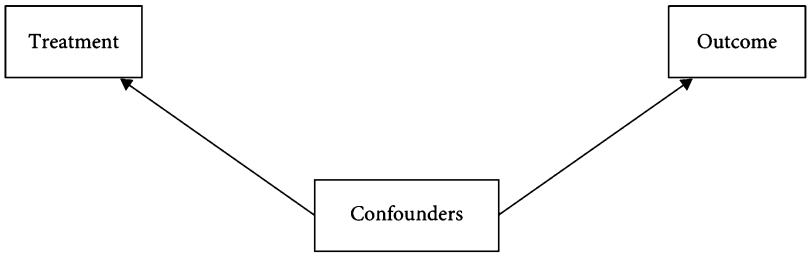


Figure 9.3. Confounders have an effect on treatment status and an independent effect on outcome.

correlation between economic prosperity and civil war risk as an unbiased estimate of the causal relationship.

So the first step in assessing whether a correlation is evidence of a causal relationship is to ask yourself whether there are any confounders. The schematic picture in figure 9.3 might help you remember to do so. The question the figure asks is, For any given treatment and outcome, are there any factors that you suspect belong in the confounders box? To fit in the box, two things must be true. First, the arrow from confounder to treatment has to make sense—that is, you must believe the confounder might exert an effect on the treatment. And second, the arrow from the confounder to outcome has to make sense—that is, you must believe the confounder might exert an effect on the outcome that doesn't run through the treatment. If you can fill in factors that satisfy both these conditions, then you have a reasonable concern about confounders and should be wary of giving a causal interpretation to the correlation between treatment status and outcome.

Reverse Causality

The second source of bias we need to worry about is reverse causality. There is *reverse causality* if the outcome affects treatment status. Reverse causality creates baseline differences because, if an outcome affects whether or not a unit receives treatment, there will be systematic differences in outcomes between the treated and untreated groups that are not due to the effect of the treatment.

Consider, again, our example of the negative correlation between the state of a country's economy and civil war. We've already seen that there could be confounders underlying this relationship. But there might also be reverse causality. For instance, during the course of fighting a civil war, infrastructure is destroyed, production is disrupted, and people are killed. All of these effects of civil war directly reduce economic prosperity. Hence, a negative correlation between a measure of economic prosperity and civil war might reflect the effect of war on the economy, rather than the effect of the economy on war. The potential for such reverse causality is yet another reason that a causal interpretation of this correlation is not justified.

The schematic picture in figure 9.4 is a way of helping to remind yourself to check for reverse causality before interpreting a correlation as causal. The question the figure asks is, For any given treatment and outcome, are there potential sources of reverse causality? That is, can we think of reasons that a causal arrow might run from the outcome to the treatment? If so, then you have a legitimate concern about reverse causality and should be wary of giving a causal interpretation to the correlation between treatment and outcome.

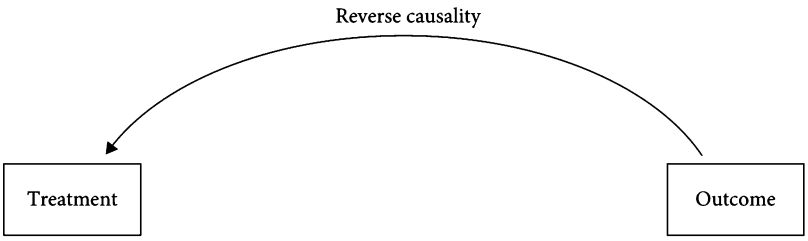


Figure 9.4. Reverse causality is when the outcome affects treatment status.

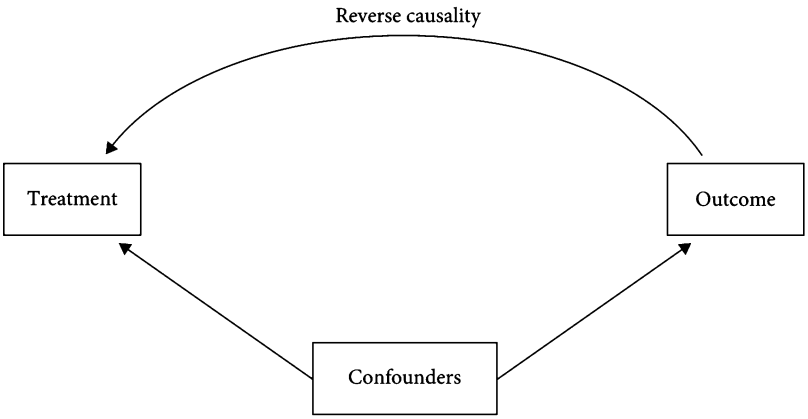


Figure 9.5. Confounders and reverse causality—two key sources of bias for estimating causal relationships.

In general, if someone shows you a correlation between an outcome of interest and a treatment of interest, without some additional information and investigation, you might have no way of knowing whether that correlation arose because the treatment affects the outcome, the outcome affects the treatment, confounders affect both the treatment and the outcome, or some combination of all of these possibilities.

Figure 9.5 provides an overall schematic for thinking about the two sources of bias we've discussed. Now that we've got the conceptual material summarized in figure 9.5 firmly in hand, let's get some practice thinking clearly about correlation versus causation, confounders, and reverse causality by talking through a couple examples in some detail.

The 10,000-Hour Rule, Revisited

You probably don't know the name Dan McGlaughlin, but he got quite a bit of press back in 2010. In April of that year, McGlaughlin quit his job as a photographer to pursue the dream of playing professional golf. He planned to practice golf for at least thirty hours every week for more than six years, until he had put in 10,000 hours of deliberate practice. He believed that by the end of those 10,000 hours he would be an expert golfer, ready to qualify for the PGA Tour. It hasn't quite worked out as McGlaughlin planned. He didn't make the PGA Tour but did open what appears to be a pretty cool artisanal soda company.

McGlaughlin's quixotic plan should sound familiar. He took Malcolm Gladwell's 10,000-hour rule, which we discussed back in chapter 4, to its (il)logical extreme. Talent, the argument goes, is secondary; great success is all about putting in those 10,000 hours. So any of us can achieve virtually anything, even a career in professional sports, if we just commit ourselves to 10,000 hours of really serious practice.

As we've already discussed, Gladwell's evidence for the 10,000-hour rule—even just as a statement about a correlation—is dubious because of lack of variation. But McGlaughlin did not hang his hat entirely on this evidence. He was also inspired by research done by the psychologist K. Anders Ericsson of Florida State University.

Ericsson argues that the key to super-high performance at anything is deliberate practice. Once a certain level of expertise has been achieved at some task, he claims, people's performance tends to plateau even if they keep gaining experience or general practice. The only way to continue improving at that point is through deliberate practice—working on exercises specifically targeted to particular aspects of performance. The more deliberate the practice is, the better the performance will be. What distinguishes true masters of an activity from good, but not great, experts, is the amount of time devoted to deliberate practice.

The 10,000-hour idea derives from a seminal study of expert musicians by Ericsson and collaborators. Unlike Gladwell, Ericsson does have variation. He studied violin students at an elite music school in Berlin. All of the students in the study were expert violinists. But they could nonetheless be distinguished in terms of quality. Ericsson asked the faculty to identify three groups—the very best violinists who were likely to go on to careers as soloists or in major orchestras, good violinists who were less likely to have successful performance careers, and the weakest group of violinists, who would likely become teachers. (We are trying not to take offense.)

The violinists were interviewed about their history of practice—age at which they started, hours per week, the type of activities they engaged in during practice, level of concentration, and so on. They were also asked to keep diaries recording their practice habits. Armed with this information, the researchers were able to compare the practice behaviors of the different groups of violinists. The finding: The best violinists had practiced at least 10,000 hours by the time they were eighteen, while the less accomplished violinists in the second group had only practiced about 7,500 hours, and the future teachers in the third group had practiced only about 5,000 hours. Moreover, the best violinists were distinguished by spending a greater share of their practice time deliberately. For example, they spent more time on difficult tasks designed to improve performance rather than simply playing enjoyable pieces that they had already mastered. Similar studies report analogous findings for chess players, athletes, and others. So the data show a positive correlation between deliberate practice and high performance.

Given this evidence, it looks like both the 10,000-hour rule and the focus on deliberate practice might not be so far-fetched. The highest-performing experts, in an array of fields, don't seem to be distinguished by measurable physical characteristics. The key appears to be that the best performers are those who practice the longest number of hours and in the most focused way. So maybe practicing deliberately for 10,000 hours really can make you world-class. Maybe Dan McGlaughlin had a pretty good shot at becoming the next Tiger Woods.

But before you stop reading to go become a professional golfer, let's think a bit more. Ericsson didn't make Gladwell's mistake. He had variation and so established a correlation between deliberate practice and achievement. But that doesn't mean the correlation

reflects a causal effect. To reach that conclusion, we need to think about confounders and reverse causality.

Here's one possible concern. Suppose innate natural talent really is important. Imagine two kids, both of whom love to play the violin. One kid is more innately talented than the other. They both practice hard, putting in many hours. The hard work pays off and both progress rapidly. But over time the talent differential starts to manifest itself. The more talented kid masters difficult pieces of music more quickly and with greater precision. She receives more accolades and performance opportunities than the less talented kid.

Time progresses and the two kids become teenagers. New opportunities and distractions—dating, sports—arise. Each teen has to decide how much time and energy to continue to devote to violin practice. The more talented of the two teens finds that every time she devotes a day to violin, she masters new skills and repertoire. This progress and achievement is inspiring. It creates a positive feedback loop whereby practice leads to success, which inspires further practice and greater focus. So she continues to devote herself to deliberate violin practice, achieving those magic 10,000 hours by the time she is eighteen.

The less talented of the two teens also progresses each time he devotes a day to the violin, but he does so more slowly and with less proficiency. A piece that takes the more talented teen a week to master might take him a month. Even then, he plays it with less technical accuracy and musicality. His achievements are slower to come and met with fewer accolades. This lack of progress is frustrating. Met with less positive feedback, he finds practicing less rewarding. As a result, he still loves and continues to work hard at violin, but as new opportunities come up, he is more likely to take a few hours or a day off of violin practice to pursue them. And even while practicing, perhaps he is less focused because he has other things to think about. By the time he is eighteen, he's put in only three-quarters as much practice time as his more talented friend, and less of it is deliberate.

Two young people like those we've just described could easily have ended up music school classmates in Ericsson's study. The more talented of the two would have been identified by the faculty as one of the best violinists, while the less talented would have been identified as good but not great. Comparing them, Ericsson would have found, as he did, that the stronger of the two violinists put in 10,000 hours of deliberate practice, while the weaker put in only about 7,500 hours of less deliberate practice.

From this comparison, Ericsson concludes that practice caused the difference in their success. But as we've seen, this causal interpretation of the correlation is not warranted. As highlighted in figure 9.6, innate talent could well be a confounder—affecting the amount of deliberate practice (treatment) and having a direct effect on performance (outcome) over and above its effect through practice. Differences in talent cause baseline differences in achievement.

Of course, in the story we told, it's not as if practice has no effect. Success is surely influenced by talent, practice, and the combination of the two (the most innately gifted person in the world could not become a great violinist without practice). But in our hypothetical example, the more talented student would likely still be a better violinist than her classmate even if she had only practiced for 7,500 hours, and the less talented student would still be worse than his classmate even if he had forced himself to practice for 10,000 hours.

The extent to which the correlation reflects a causal relationship versus the bias from a talent differential is important here. To see why, think back to Dan McGlaughlin.

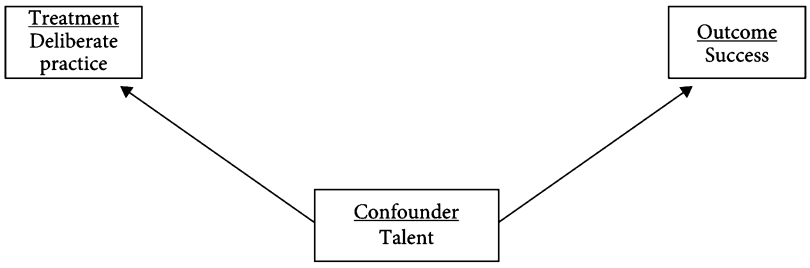


Figure 9.6. Is the correlation between deliberate practice and performance an unbiased estimate of the causal relationship?

Before quitting his job, McGlaughlin hadn't shown any of the signs of a world-class golfer. That fact might help explain why he hadn't previously put in thousands of hours of deliberate practice. If a significant portion of the correlation between practice and success is due to both of them being affected by innate talent rather than being the causal effect of practice, then McGlaughlin's efforts were unlikely to yield the desired outcome. Profoundly talented people practice a lot and have great success. That doesn't mean that a person who lacks profound talent but forces himself to practice a lot will find that same success. And so, perhaps, he should have anticipated that his journey would end as it did, in an artisanal soda company rather than a PGA Tour card. (For what it is worth, Ethan thinks running an artisanal soda company sounds *way* more fun than playing professional golf. Anthony disagrees.)

Diet Soda

Speaking of fizzy beverages, at the time of this writing, there is a near consensus among nutrition experts that diet soda is bad for you. Studies by experts in respected scientific journals have linked diet soda consumption to a range of health problems including obesity, diabetes, and heart attacks.

Curiously, despite all of the purportedly hard evidence on the dangers of diet soda, scientists don't yet have a compelling explanation (aside from the adverse dental consequences of diet soda—acid is bad for your teeth). They have been racking their brains to explain why beverages with virtually no calories are somehow making people overweight.

Several theories have been put forward. One explanation is that diet soda has *chemicals*, which might be bad for us. Of course, everything we consume has chemicals, so this isn't much of an explanation, and it just kicks the can of soda one more step down the road. Another explanation is that diet soda confuses your body and makes it somehow want more calories. After consuming diet soda, the story might go, your brain expects you to receive some calories from this sweet beverage, and when it doesn't, it urges you to raid your pantry for cookies and chips. In this way, your brain is like a child who is told they're about to get some candy only to have it revoked at the last minute. A third explanation is that diet soda (and presumably anything sweet) desensitizes your taste buds, meaning that you need to eat more and more sugary foods to get your fix.

We aren't nutritionists, but none of these explanations sounds overwhelmingly compelling to our untrained ears. Indeed, we could have imagined similarly convincing stories for why the effect should go in the opposite direction. Diet soda might allow

someone with a sweet tooth to enjoy a refreshing treat without consuming extra calories. And diet soda might even trick your brain into thinking you've ingested calories and therefore speed up metabolism, which could be good for health and weight management. As we say, we're not experts, but it seems at least as plausible that diet soda is good for health, especially as a substitute for sugary beverages. So why have experts so strongly agreed that diet soda is bad for your health?

We have scoured the studies, and the extent of the evidence seems to be the following. There is a negative correlation between drinking diet soda and health outcomes. People who drink diet soda are more likely to be obese, have diabetes, and suffer from a range of other health problems than are people who do not drink any kind of sweet beverages.

Before agreeing with the nutritionists that this correlation reflects a genuine causal effect of diet soda on health, we should think about whether there are confounders or reverse causality.

What if, for example, snacking makes people more likely to both drink soda (because the soda goes well with the snacks) and, for reasons unrelated to soda, more likely to be obese? Then snacking would be a confounder. Or perhaps it's reverse causality—what if obesity or diabetes makes people more likely to drink diet soda? Presumably, if you like soda and become diabetic, you'll switch to diet soda. Similarly, we ourselves could imagine switching from diet soda to sugary beverages if only we were healthier. Clearly, confounders and reverse causality are serious concerns, and we should not treat the correlation between diet soda and health outcomes as a credible estimate of the causal effect.

How Different Are Confounders and Reverse Causality?

While we are thinking about confounders and reverse causality, it is worth pausing to reflect on how they relate to one another. Often a problem that appears to be about reverse causality can also be thought of in terms of confounders, where the relevant confounder is simply the anticipated outcome.

To see what we mean, think back again to our example of the negative correlation between the economy and civil war risk. We've seen that there are both confounders and reverse causality that invalidate a causal interpretation of this correlation. Consider one more problem. Suppose that, for a variety of reasons (e.g., lack of democracy, ethnic divisions, nearby civil wars), people believe some country is at high risk for a civil war. This risk of civil war might deter investment in the country, lead to capital flight, cause a brain drain, and so on. In this way, anticipation of a future civil war can cause the country to have a weaker economy. You could think of this as a case of reverse causality: civil war risk causes economic weakness. But it may be more clarifying to think of it as a case of confounding, where the confounders are whatever factors lead people to believe the country is at high risk of civil war. Those factors cause economic weakness by deterring investment and causing brain drain. And, presumably, they lead people to believe the country is at high risk of civil war precisely because they exert an independent effect on civil war occurring.

Let's consider another example—campaign spending.

Campaign Spending

Political candidates spend huge amounts of time raising money for their campaigns. Members of Congress, for example, often spend several hours per day in a call center

phoning wealthy constituents and asking them to help fund their next reelection effort. (It turns out that being in Congress isn't a particularly glamorous job.)

Of course, politicians do this because they believe that campaign dollars are essential for their electoral prospects. And campaign consultants constantly advise candidates about how much they should be spending on television ads, digital ads, direct mail, and personal voter outreach. Electoral campaigns are clearly a big business predicated on the notion that candidates can improve their chances of success by raising and spending more money.

Given the scale of campaign spending, political scientists have devoted a lot of time and effort to estimating the returns on these efforts. Can spending on advertising really influence election results? And are those effects big enough to justify the millions of dollars donated to finance campaigns and the thousands upon thousands of hours spent raising those dollars?

One of the earliest and most influential studies of campaign spending was conducted by Gary Jacobson in 1978. Jacobson concludes that campaign spending seems to significantly help challengers' electoral prospects but has little benefit for incumbents. Indeed, campaign spending by incumbents might even be counterproductive, hurting their electoral fortunes!

What is Jacobson's evidence for this claim that campaign spending helps challengers but not incumbents? Challenger spending is strongly positively correlated with challengers' vote shares. But incumbent spending is negatively correlated with incumbents' vote shares.

One explanation for these correlations, Jacobson speculates, is that incumbents typically raise and spend more money than challengers. Maybe some initial amount of spending at the levels that we typically see for challengers helps a candidate to obtain name recognition and persuade voters. But perhaps too much spending from an already well-known incumbent annoys and turns off potential supporters. On this account, incumbents are making systematic mistakes, both in spending their time raising money and in spending that money once they've raised it.

Of course, the comparisons underlying these correlations may not be apples-to-apples. We need to think about confounders and reverse causality.

One big concern along these lines has to do with electoral strength. Which kinds of challengers tend to be able to raise and spend lots of money? Presumably, popular challengers with a real shot at victory. It is those electorally strong challengers that donors are likely to be willing to invest in. But, of course, strong challengers are those who were expecting to do well in the election even before they raised the money—perhaps they are charismatic, well-known, or particularly talented. So it would be a mistake to interpret the positive correlation between challenger spending and electoral performance as purely causal. It, at least in part, reflects baseline differences in electoral strength between challengers who can and can't raise a lot of money.

The thing we want you to notice in this example is that you can think of the problem of electoral strength as one of reverse causality or as one of confounding. Thought of as reverse causality, you might describe it as follows: "When a challenger is going to do well, she can raise and spend more on her campaign." Thought of as a confounder, you might describe it as follows: "When a challenger has characteristics that make her competitive, this affects both her ability to raise and spend money and how well she does in the election." Both sentences describe the same concern, just framed slightly differently.

A similar argument holds for incumbents. In general, although they spend and raise a lot of money, most incumbents in U.S. elections are electorally pretty safe. The ones

who really need to exert a lot of effort raising and spending money are those who are electorally vulnerable. So we might expect exactly the opposite relationship for incumbents as for challengers. Incumbents spend a lot of money not when they are strong but when they are weak. And, again, you can view this problem in terms of reverse causality—"Electorally weak incumbents spend more money"—or in terms of confounders—"Characteristics that weaken incumbents, making the race competitive, separately cause them to spend more money and lead to worse than average electoral outcomes."

Subsequent studies using randomized experiments and other clever approaches to try to tease out the causal relationship generally suggest that campaign spending does have positive effects for both challengers *and* incumbents, although the substantive size of those estimated effects is typically small. A campaign might have to spend hundreds of dollars to swing a single vote, which means that meaningfully influencing the outcome of a large election through campaign donations is typically unaffordable. For example, consider a gubernatorial or senatorial race in a large U.S. state. Even in a race thought to be very close, the outcome will likely be decided by hundreds of thousands of votes. This means that if donors wanted to influence the outcome of the election, they would have to spend tens of millions of dollars and hope that their spending does not trigger an offsetting response from supporters of the opponent. Because of this, even the very largest donors have likely swung very few elections.

As you can see, there isn't a ton at stake as to whether we think about such cases as reverse causality or as confounders. What really matters is that we interrogate correlations for possible baseline differences, whether from confounders or reverse causality, and if there are baseline differences, that we show proper caution before interpreting a correlation as implying causation.

Signing the Bias

When there are confounders or reverse causality, the correlation between treatment and outcome is not an unbiased estimate of the true causal relationship of interest (whether the ATE, ATT, ATU, or other causal quantities that we'll discuss in later chapters). But sometimes we can make some progress on learning about causality by asking whether the correlation over- or under-estimates the causal effect.

Let's think back to our favorite equation, this time written in terms of causal inference:

$$\text{Observed Correlation (Estimate)} = \text{True Causal Effect (Estimand)} + \text{Bias} + \text{Noise}$$

Suppose the observed correlation between administering some medical treatment and survival rates following a stroke is positive. But also suppose there are confounders that you have not accounted for, so there is bias in your estimate of the true causal effect. If you have reason to believe that the bias is positive, then the observed correlation is an *over-estimate* of the true causal effect of the treatment. This means that you can't be confident, on the basis of the observed positive correlation, that the treatment does anything at all. Even if the true causal effect is zero, you would observe a positive correlation on average due entirely to confounders creating positive bias.

But now suppose you have reason to believe that the bias is negative instead of positive. In this case, the observed correlation is an *under-estimate* of the true causal effect of the treatment. So, if you are confident that the observed correlation is positive, you

should be even more confident that the true causal effect is positive. And this can be useful to know. For instance, suppose administering the treatment would be a good idea (given its various costs) even if the true effect was equal to the observed correlation. Then the fact that the observed correlation is an under-estimate of the true effect suggests that you should administer the treatment.

Note, of course, you could still end up being wrong because of noise. Even if you are under-*estimating* the true causal effect, on average, that doesn't mean that any one estimate is in fact lower than the true effect. It just means that your estimates will be lower than the true effect on average.

Because this kind of thinking about the sign of the bias in an estimate can sometimes be valuable, it is useful to spend a little time thinking conceptually about when confounders imply that an observed correlation over-estimates the true effect and when they imply that an observed correlation under-estimates the true effect.

Start with our discussion of the relationship between votes and the campaign spending of challengers, where we worried that electoral strength was a confounder. Does this confounder tend to make the correlation between votes and campaign spending an over- or under-estimate of the causal effect? It seems likely that electoral strength has a positive effect on both fundraising and votes for challengers. So some of the extra votes received by high-spending challengers are actually the result of their electoral strength rather than an effect of the spending. As such, we should expect the correlation between spending and votes to be an over-estimate of the true effect.

To see another example, let's return to our discussion of the positive correlation between attending a charter school and standardized test scores. There, we said, one possible confounder is that students who go to the trouble of applying to a charter school may on average be more academically gifted than the general student population. And the fact that those students are more academically gifted or motivated may have a direct effect on their test scores.

If this story is right, does this confounder tend to make the correlation between charter school attendance and standardized test scores an over- or under-estimate of the true effect? Let's think about it. Being academically gifted has a positive effect on the likelihood a student goes to a charter school. And it also has a positive effect on test scores. That means part of the observed positive relationship between going to a charter school and test scores is the result of differential academic talent. Hence, this confounder is pushing the observed correlation to be an *over-estimate* of the true effect. That is, the bias in our favorite equation is positive.

It is straightforward that the same would be true if we had a confounder that negatively affected both attending a charter school and test scores. Indeed, this is simply the same case but with relabeling. If we think of the confounder as "lack of academic talent" instead of "academic talent," then that confounder has a negative effect on the treatment and outcome but still, obviously, leads the observed correlation to be an over-estimate of the true effect. Thus, as illustrated in figure 9.7, if you have a confounder that has the same sign effect on both treatment and outcome (whether negative or positive), then failing to account for this confounder will create positive bias. In such circumstances, the observed correlation will tend to be larger than the true effect.

Now let's think about a confounder that has differently signed effects on treatment and outcome. For instance, suppose students from poor neighborhoods are more motivated to apply to charter schools (perhaps because their local public schools are underfunded), but are also expected to do worse academically because of challenges in their living environment. This, again, is a confounder—it exerts an effect on both

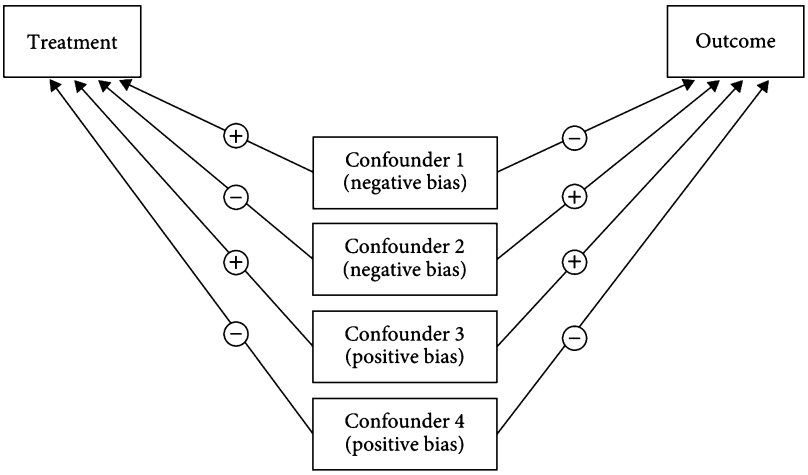


Figure 9.7. Signing the bias from a confounder.

the treatment (whether or not a student attends a charter school) and the outcome (performance on standardized tests). But unlike the case of academic talent (which positively affects the treatment and the potential outcomes), this confounder will create negative bias. Hence, the observed correlation between charter school attendance and standardized test scores is an under-estimate, rather than an over-estimate, of the true effect.

Why is this the case? In our new story, living in a poor neighborhood has a positive effect on the likelihood a student goes to a charter school. And it has a negative effect on test scores. That means the observed correlation between going to a charter school and test scores reflects the fact that the kids at charter schools over-represent poor neighborhoods, relative to the full population. This fact tends to lower test scores for charter school students for reasons having nothing to do with the effect of the charter school. If charter schools and public schools had the same proportion of kids from wealthier and poorer neighborhoods, the positive correlation between charter school attendance and test scores would be even more positive. Hence, this confounder is pushing the observed correlation to be an *under-estimate* of the true effect.

It is again straightforward that the same would be true if we had a confounder that negatively affected the likelihood of attending a charter school and positively affected the outcome. Thus, as illustrated in figure 9.7, if you have a confounder that has one sign effect on treatment and the opposite sign effect on outcome, this confounder creates negative bias. In such circumstances, the observed correlation will tend to be smaller than the true effect.

Signing the bias is even easier in the case of reverse causality. The outcome is, by definition, positively related to itself. So, if the outcome also has a positive effect on the treatment, the bias is positive. This means the observed correlation is an over-estimate of the true causal effect. And if the outcome has a negative effect on the treatment, the bias is negative, so the observed correlation is an under-estimate.

In addition to simply signing the bias, if we had a lot more information, we might be able to say something about the magnitude of the bias. Under some assumptions, the bias induced by a confounder is simply the effect of the confounder on the outcome multiplied by a measure of the correlation between the confounder and the treatment

(measured by the coefficient you would get from regressing the confounder on the treatment).

As we'll see in chapter 10, if we have data on this confounder, we can try to remove this bias by controlling. But if we don't have that data, one could still make some guesses about the extent to which the confounder affects the outcome and is correlated with the treatment in order to gauge the extent of the bias.

The discussion above illustrates that we can learn something about causal effects even from biased estimates. It's not as if we have to throw away all our analyses just because there might be confounders, and if we have good guesses about the direction and magnitude of the biases, then we might still be able to learn a lot. But often, it's difficult to know how much an observed correlation is the result of bias, which is why simple correlations are not our preferred approach for learning about causal relationships. Less naive and more informative approaches to causal inference are the focus of the subsequent chapters.

A related approach to learning about causal effects from potentially biased correlations is to work in reverse. Instead of inferring how big an effect is by making guesses about the magnitude of the bias, we can start with the assumption that the true effect is zero and then ask how big the bias would have to be to explain an observed correlation. If the extent of that bias is implausibly large, then we can conclude that the effect probably is not zero. This kind of analysis is often referred to as *sensitivity analysis*. We won't discuss the details in this book, but as a general rule of thumb, it's good to think about sources of bias, their likely signs, their likely magnitudes, and what that implies for the effect you are trying to estimate.

With an understanding of different sources of bias and their likely signs, you can more deeply understand why correlation is not necessarily evidence of a causal relationship. The true effect could be zero, but the observed correlation could have emerged because of confounding or reverse causation. Similarly, as we discussed in chapter 3, causation need not imply correlation. Even if some treatment has a large, positive effect, confounding or reverse causality could create a large, negative bias. This could lead to an observed correlation that is small, zero, or even negative (as in the case of campaign spending and votes for incumbents), despite the positive treatment effect. So, not only does correlation not necessarily imply causation. Causation does not necessarily imply correlation.

With all of this in mind, let's think through a more extended example.

Contraception and HIV

One of the greatest public health scourges of our time is the spread of HIV and AIDS in Africa. Researchers have worked hard to determine why these diseases are spreading so quickly and to try to stem the tide. One hypothesis that has received attention from scholars and public health officials alike is that the use of hormonal contraception by women may increase the risk of HIV transmission by inducing changes in the immune system or body tissue.

In a 2012 study in *The Lancet Infectious Diseases*, researchers presented evidence supporting this hypothesis. The researchers analyzed data on more than 3,500 couples in which one partner was infected with HIV and the other was not. They had data on a variety of self-reported behaviors—for example, condom use, other sexual partners—and on whether the woman received hormonal contraception from the clinic that was conducting the study. The data also reported whether the non-infected partner

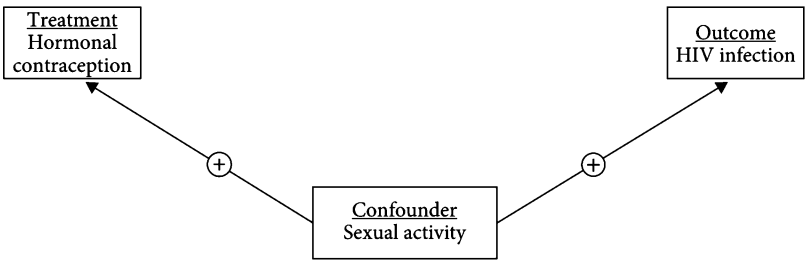


Figure 9.8. Underlying level of sexual activity tends to make the correlation between hormonal contraception use and HIV transmission an over-estimate of the true causal relationship.

contracted HIV over the course of a year or two. Finally, for those partners who did contract HIV, genetic screening provided information on whether it was transmitted partner to partner or from some third source.

There were two big findings. First, HIV-negative women who used hormonal contraception were twice as likely to acquire HIV from their infected male partners as were HIV-negative women who did not use hormonal contraception. Second, HIV-infected women who used hormonal contraception were twice as likely to transmit HIV to their HIV-negative male partners as were HIV-infected women who did not use hormonal contraception. These results held true controlling for self-reported condom use. (We'll talk more about what *controlling* means in the next chapter.) From these findings, the authors, the *New York Times*, National Public Radio, and many other sources reported that hormonal contraception likely increases the risk of HIV transmission.

This study was a major improvement over existing studies on this critically important issue. But it was a long way from comparing apples to apples. What might be going wrong?

The biggest worry is the possibility of confounders—women who take hormonal contraception are different from women who don't in lots of unmeasured ways, some of which may also be relevant for HIV transmission risk. If this is the case, then the observed correlation between hormonal contraception use and HIV transmission may be a biased estimate of the true causal relationship.

One concern is that women who intend to be more sexually active might also be more likely to use hormonal contraception. The researchers who authored the *Lancet* study were not able to randomly assign some women to take hormonal contraception and other women not to. Women received hormonal contraception if they wanted it. Sexual activity is a risk factor for HIV transmission. So, independent of anything else, more sexually active women are at greater risk of HIV transmission. If the women who are taking hormonal contraception are systematically engaging in more sexual activity, they will have higher transmission rates, even if the contraceptives themselves are playing no direct biological role.

In which direction would this confounder bias the estimates? As highlighted in figure 9.8, the thought is that sexual activity increases the use of hormonal contraception and also increases HIV transmission for reasons unrelated to contraception. So the bias is positive. As such, this confounder tends to make the observed correlation an over-estimate of the true causal relationship between hormonal contraception and HIV transmission.

The *Lancet* authors are aware of these types of concerns and make some attempts to address them. In particular women were asked about past sexual behavior and condom

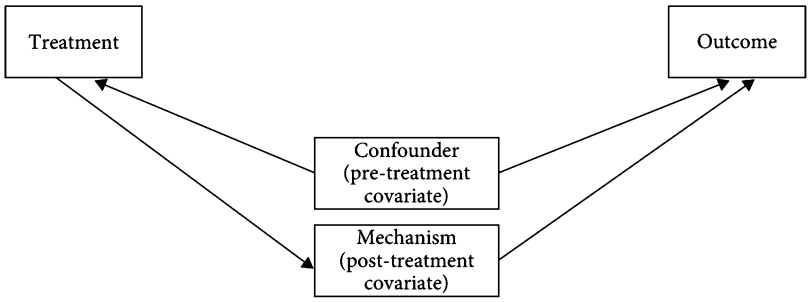


Figure 9.9. The distinction between confounders and mechanisms.

use. But self-reported behavior is notoriously unreliable, especially for sensitive topics like sexual activity and condom use.

Mechanisms versus Confounders

It is easy to get a little confused about what is and what is not a confounder. One particularly common error is to mistake the mechanisms by which a treatment affects an outcome for confounders. A *mechanism* (sometimes also called a *mediator*) is some feature of the world that the treatment affects, which then, in turn, affects the outcome. So a mechanism, rather than being a confounder, is part of the way that the treatment has its effect on the outcome.

For instance, one way that a charter school might cause students to get better test scores than they would if they went to their local public school is by providing more advanced placement (AP) classes that better prepare students for tests. When looking at a correlation that says charter school students perform better than public school students on standardized tests, it is tempting to say, "Yeah, but that is confounded by the fact that those charter school students had access to more advanced placement classes." But this isn't right.

Remember, a confounder is not simply a feature of the world that is correlated with treatment and outcome (which, in this story, AP classes are). It is a feature of the world that *affects* both treatment and outcome. But, in our story, access to AP classes doesn't affect whether a student goes to a charter school (treatment). Rather, it is affected by the student going to a charter school and then, in turn, affects the student's performance on standardized tests. Thus, access to AP classes is not a confounder; it is one of the mechanisms by which charter schools improve test scores. We sometimes describe confounders as *pre-treatment covariates*—that is, variables that were correlated with treatment and outcome before the treatment occurred—and describe mechanisms as *post-treatment covariates*—that is, variables that become correlated with treatment and outcome after treatment occurs. Figure 9.9 illustrates the distinction (note the direction of the arrows).

As we say, it is easy to get confused about these issues. So, let's talk through a couple examples.

Suppose that a medical study of middle-aged men finds that those who take statins are less likely to die of heart attacks. You note that those men who take statins are on average wealthier and have lower cholesterol. Which of these is a confounder and which might be a mechanism? Think about it for a moment before we tell you the answer.

Let's start with wealth. Remember, when assessing whether some feature of the world is a potential confounder, you need to ask whether it could affect both treatment and outcome. So we ask two questions:

1. Could a man's wealth affect whether he takes statins? Surely the answer is yes. Wealthier men are, presumably, better able to afford medication and also probably more likely to see a doctor who would prescribe that medication to them.
2. Could a man's wealth affect his risk of dying from heart disease? Again, the answer is yes. Wealthier men might be better able to afford heart-healthy lifestyles (e.g., joining a gym) and are more likely to get swift access to health care in the event of a heart attack.

Thus, we should worry that wealth is a confounder here.

What about lower cholesterol? Medical evidence suggests that higher cholesterol might affect the likelihood of having a heart attack (although it's hard to tease out the causal effect). But does cholesterol affect whether or not a person takes statins? Here, we might need a little more information—in particular, when exactly the cholesterol levels were measured.

If cholesterol was measured before the person started taking statins, then it is a good candidate for a confounder. After all, people typically choose to take statins when they have high cholesterol. (Using your skills from the previous section on "signing the bias," does this confounder make you think the study under- or over-estimates the efficacy of statins?)

But if cholesterol levels were measured after the person started taking statins, then it is a mechanism. We suspect that one of the ways that statins might reduce the risk of heart disease is by lowering cholesterol. If this is true, and if we randomly assigned some people to take statins and others not to, we would expect the ones who took the statins to have lower cholesterol (and lower risk of heart disease). This difference in cholesterol levels isn't a problem for inferring the efficacy of statins; rather, it is a mechanism by which that efficacy is achieved.

Here's another example. Suppose we are interested in whether a good economy helps reduce the risk of civil war. We find that there is indeed a negative correlation between per capita income and the frequency with which a country experiences civil war. But we also note that democracy is positively correlated with per capita income and negatively correlated with civil war risk. Should we think of democracy as a confounder or a mechanism in this case?

This is a tricky one. You can certainly see how democracy might be a confounder. Having a democratic form of government might improve the quality of governance. And good governance might cause a country's economy to grow. Moreover, being a democracy might give people non-violent ways to resolve political disputes, thereby directly reducing the risk of civil war. In this story, democracy is a confounder, since it has a direct causal effect on both treatment (per capita income) and outcome (civil war).

But you can also see how democracy might be a mechanism. Perhaps as countries become richer, citizens become more informed, better educated, more able to take actions for their own benefit, and so on. In this way, having a higher per capita income might directly increase the probability that a country becomes a democracy. And then, for the reasons already stated, democracy might decrease the risk of civil war. In this story, rather than being a confounder, democracy is part of the mechanism by which higher per capita income reduces civil war risk.

As this example highlights, the distinction between a confounder and a mechanism is important, but not always cut-and-dry. For now, it is important to see the distinction at a conceptual level, even if in many real-world scenarios you are not always sure whether some factor is the one or the other. We will return to this theme in the next chapter, when we talk about the benefits and limitations of *controlling*.

Thinking Clearly about Bias and Noise

We'd like to pause to make sure you don't forget the lessons from part 2—about assessing whether a relationship exists—just because we've now turned our attention to thinking about causal questions. In this spirit, let's think about the questions you should ask yourself when someone shows you a correlation and interprets it as an estimate of some causal relationship.

First, are we actually observing a correlation? Recall from chapter 4 that people often think they have measured a correlation when they haven't because they didn't collect data with variation in one of the key variables. So, for instance, you need to make sure that they didn't just look at instances when the outcome of interest occurred or the purported treatment was always present. If they made this kind of mistake, you can't even know from the data presented whether the variables are correlated, let alone related causally.

Second, does the estimated correlation reflect a genuine relationship in the world? For example, suppose someone shows you that peanut butter consumption is correlated with appendicitis in a sample of 100 people—within that sample, people who ate more peanut butter were more likely to get appendicitis. You might ask yourself a series of questions. Is the correlation statistically distinguishable from the null hypothesis of no correlation? Why do they only have data on 100 people? Did they collect the data with the goal of measuring this particular correlation? Would they have told you about this finding if they had found no correlation? If you're worried about p-hacking or p-screening, then you might be skeptical that there's actually a correlation between peanut butter and appendicitis in the broader population, and you'd want to collect an independent sample of data to see if the correlation persists in that new sample. If it doesn't, you should worry that the true estimand (the correlation in the population) is zero and that they found a positive correlation in their 100-person sample because of noise.

Third, is this correlation convincing evidence of a causal relationship? You'd want to ask whether they're comparing apples to apples—are there confounders or reverse causation that biases the estimated correlation away from the true causal relationship? For example, if someone shows you that ice cream consumption is correlated with sunburns across days of the year, you'd probably believe that they've identified a genuine correlation. If they collected a new sample, they'd probably continue to find a strong correlation between ice cream and sunburns. But that doesn't mean the correlation constitutes evidence that ice cream causes sunburns. It might. Maybe eating ice cream inspires people to go outside. But a far more likely explanation is that sunshine increases both ice cream consumption (for reasons unrelated to sunburns) and sunburns (for reasons unrelated to ice cream).

To help us put all of this together, let's return to the special case of our favorite equation when we are doing causal inference:

$$\text{Observed Correlation (Estimate)} = \text{True Causal Effect (Estimand)} + \text{Bias} + \text{Noise}$$

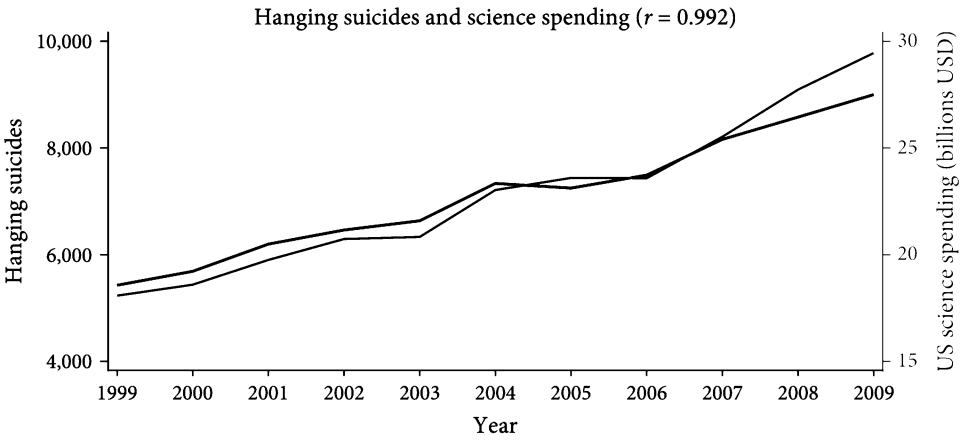


Figure 9.10. The strong correlation over time between suicides by hanging and government spending on science.

There are two kinds of ways an estimated correlation can deviate from the causal effect of interest. First, there could be noise. *Noise* here refers to idiosyncratic factors that affect our estimate. This could come from sampling variation in cases where you care about a population but you only have data on a sample of the whole. Or noise could come from other idiosyncratic variation in your variables of interest that are independent of any kind of causal connection (e.g., you might measure the variables with error). We might think that, since the noise is zero on average, we can just ignore it. But the fact that the noise is zero on average doesn't mean it is zero in any particular sample. And furthermore, in the presence of *p*-hacking and *p*-screening, even the average noise won't be zero. This was the focus of chapter 7. Second, in addition to noise, there could be bias—that is, confounders or reverse causation that makes the estimate different from the estimand on average, which is the focus of this chapter.

When confronted with a correlation that is presented as evidence of causation, it helps to consider all three factors—a true causal effect, bias, and noise—and try to think through the role each plays in explaining the correlation. Of course, it is often the case that an estimate reflects some combination of all three.

In some cases, it's tricky to separate bias and noise or even to think about them in a conceptually clear way. Let's see some examples of this. Tyler Vigen's book *Spurious Correlations* identifies pairs of trends over time that happen to correspond with one another, even though there's no good reason to think that those two trends are causally or logically connected in any way. The term *spurious correlation* is certainly apt, although we tend to avoid it because it doesn't clarify whether the person using the term thinks the correlation arose because of bias or noise.

Figure 9.10 illustrates one of Vigen's examples. It shows the correlation over time between suicides by hanging and government spending on science in the United States. Although it's not presented in a conventional way, this figure shows a positive correlation. If you think of each year as a unit of observation, it's clear that years with more hanging suicides than usual also have higher-than-average spending on science. In fact, the correlation coefficient (r) is .992, essentially the strongest correlation one can find without making up the data.

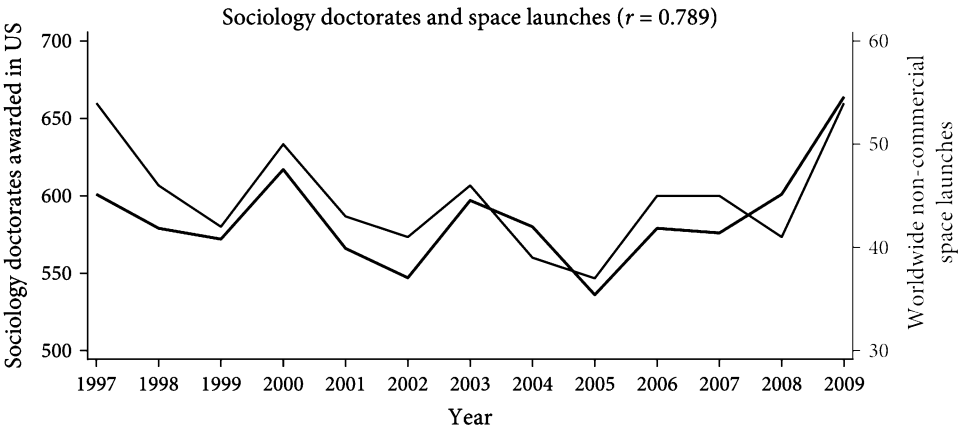


Figure 9.11. The quirky correlation over time between sociology doctorates and space launches.

What's going on here? Is this correlation attributable to a true causal effect of science spending on suicides, to bias, or to noise? It's theoretically possible, but very unlikely, that science spending has a large, positive effect on suicide by hanging (or vice versa). Noise certainly seems like a plausible explanation. If you look at enough variables, you're bound to find two of them that happen to correspond by chance, and we know that this is exactly what Vigen did. He checked for correlations over time for many variables and selectively reported the correlations that were significant.

But maybe it's also bias. What's an example of a confounder here? Could there be a variable that affects both hanging suicides and also science spending? One potential confounder is population. Over this period (1999–2009), the U.S. population grew steadily from about 279 million to 307 million. And population growth could plausibly increase both suicides and science spending.

To explore whether bias or noise is the more important explanation for the observed correlation, it might help to think about whether you expect this correlation to also hold for years before 1999 and after 2009. If you suspect that this correlation would likely hold more generally outside this sample of data, then it can't just be noise. Alternatively, if you think that this correlation is just a fluke, unlikely to hold outside the short period for which Vigen collected data, then it's just noise, due to neither a causal relationship nor bias.

Let's take a look at another couple examples. Figure 9.11 shows the correlation over time between sociology doctorates awarded in the United States and worldwide non-commercial space launches. Again, there's a strong correlation. Furthermore, it's not so easy to simply attribute the result to population growth (or something else changing over time) because the correlation is not driven by the two variables generally increasing over time. On average, space launches and sociology doctorates aren't increasing or decreasing, but the years with more space launches also tend to be years with more sociology doctorates.

We're pretty comfortable chalking this one up to noise. There's idiosyncratic variation from year to year in space launches and sociology doctorates, and they happened to line up during this period. But we suspect that if we looked at the next thirteen years of data, the correlation would be close to zero.

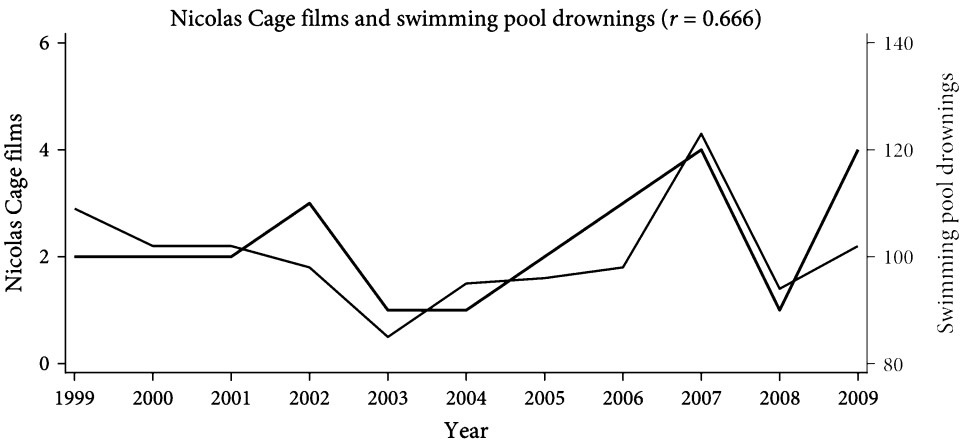


Figure 9.12. The metaphysically challenging correlation between Nicolas Cage movies and swimming pool drownings.

Finally, figure 9.12 shows the correlation over time between the number of movies in which Nicolas Cage starred and swimming pool drownings. This one feels like another straightforward case of noise. There's surely no causal connection, and there's also probably no compelling confounder. And, as with sociology and space launches, we're willing to bet that this correlation won't continue to hold in future years.

However, the Cage-drowning correlation poses a different conceptual conundrum. Suppose that this analysis included all of the years during which Nicolas Cage was acting and all of the years during which people had swimming pools (this is obviously not the case, but just imagine with us). If Nicolas Cage no longer made movies and people no longer had swimming pools, we couldn't assess the correlation between these two variables in some future period. So how could we think about whether this correlation was the result of noise? Furthermore, what would it even mean to say that this correlation was the result of noise if we had all the data there was to have on Nicolas Cage movies and swimming pool drownings? If you have observed the entire population (here, of Nicholas Cage movies), there is no sampling variation.

One way to resolve this puzzle is to make the metaphysical leap we discussed back in chapter 6 when we talked about statistical inference when we have data for the whole population. Sure, there's an observed correlation between Nicolas Cage movies and swimming pool drownings in this world, but that's just a small sample of a broader population of alternative, hypothetical worlds that might have been. Those worlds are just like our own, but all of the idiosyncratic, unrelated factors happen to play out differently. Do we have any reason to expect that Nicolas Cage movies would be correlated with swimming pool drownings in those worlds? If the answer is *no*, we might say that the correlation we observed is just noise, even though we have all the data there is to have about Nicolas Cage and swimming pool drownings.

Wrapping Up

We've seen that a correlation is often a biased estimate of a causal relationship because of confounders or reverse causality. This is what we mean when we say that correlation does not imply causation.

If we know what confounders to look out for, and if we can measure them, can we correct the bias and obtain a better estimate of the causal relationship? How to do so is the topic of chapter 10.

Key Terms

- **Causal effect:** The change in some feature of the world that would result from a change to some other feature of the world.
- **Average Treatment Effect (ATE):** The difference in average outcome comparing two counterfactual scenarios—one where everyone in the population is treated and one where everyone in the population is untreated.
- **Average Treatment Effect on the Treated (ATT):** The difference in average outcome comparing the scenario where everyone in the subgroup of people who in fact received treatment is treated and the counterfactual scenario where everyone in that subgroup is untreated.
- **Average Treatment Effect on the Untreated (ATU):** The difference in average outcome comparing the counterfactual scenario where everyone in the subgroup of people who did not receive treatment is treated and the scenario where everyone in that subgroup is untreated.
- **Difference in means:** The difference in average outcome comparing the subgroup of people who in fact received treatment to the subgroup of people who in fact did not receive treatment.
- **Baseline differences:** Differences in the average potential outcome between two groups (e.g., the treated and untreated groups), even when those two groups have the same treatment status.
- **Confounder:** A feature of the world that (1) has an effect on treatment status and (2) has an effect on the potential outcome over and above the effect it has through its effect on treatment status.
- **Reverse causality:** When the outcome affects treatment status.
- **Over-estimate:** When the bias is positive, so that the estimate is larger than the true effect in expectation.
- **Under-estimate:** When the bias is negative, so that the estimate is smaller than the true effect in expectation.
- **Mechanism (or mediator):** A feature of the world that the treatment affects, which then, in turn, affects the outcome.
- **Pre-treatment covariate:** A variable that is correlated with treatment and outcome before the treatment occurs.
- **Post-treatment covariate:** A variable that becomes correlated with treatment and outcome after treatment occurs.

Exercises

9.1 At the end of our discussion of violent and non-violent resistance in chapter 1 we asked you the following:

> Why might the fact that there are more government crack-downs following violent protests than non-violent protests *not* mean that switching from violence to non-violence will reduce the risk of crack-downs?

We promised that you would be able to give a compelling answer by the end of this chapter. So, please identify at least one reason why the fact that violent protests are more often met with a government crack-down than non-violent protests is not compelling evidence that the use of violent protest tactics causes government crack-downs.

9.2 Let's think about over-estimates and under-estimates in two of our examples.

(a) In our discussion of violin practice, we noted that a musician with greater talent might both practice more and play the instrument better for reasons having nothing to do with how much she practices. Does this suggest that the correlation between practice and playing quality is an over-estimate or an under-estimate of the true effect of practice on playing?
(b) In our discussion of campaign spending, we argued that incumbents are likely to spend heavily on their campaigns when they are electorally weak. Does this suggest that the observed lack of (or even negative) correlation between campaign spending and electoral performance of incumbents is an over-estimate or an under-estimate of the true effect of spending on votes?

9.3 Ethan was once at a meeting where he was briefed on the ways in which data analytics can improve universities' operations. The example the presenter was most excited about was from a data analytics team in a major research university's development (which is jargon for *fundraising*) department. The data analytics team had discovered the following correlation by analyzing years of data: alumni who donate to the university six years in a row are way more likely to be lifelong givers than are alumni who only donate five years in a row.

The presenter was excited because, in their view, this finding from the analytics team suggested a clear strategy to improve fundraising and alumni engagement. In particular, on the basis of this analysis, they had decided to make a major push to encourage alumni who had already given for five years in a row to give a sixth—the idea being that the evidence of a correlation between giving for six years and giving in the future suggested that giving in that sixth year had a big causal effect on future giving, so resources spent encouraging five-year givers to become six-year givers were being put to the best possible use.

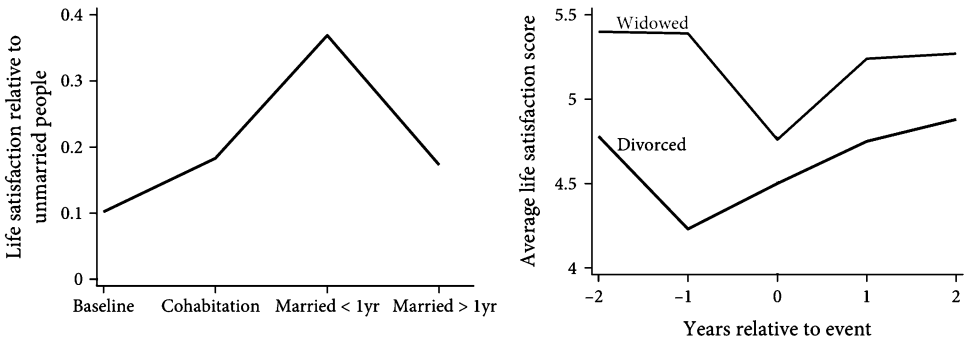
Provide two arguments, using the clear thinking skills you acquired in this chapter, to explain why this might not be a good plan.

9.4 Shortly after Harvard psychologist Daniel Gilbert's book, *Stumbling on Happiness*, was released, he was on TV, where he informed Stephen Colbert that "marriage is one of the best investments you can make in happiness." That advice implicitly rests on a causal claim: marriage causes happiness.

Much recent research documents a positive correlation between marriage and happiness. But is the relationship causal?

(a) Provide an argument for why the correlation between marriage and happiness might be the result of reverse causation (happiness causing marriage, rather than the other way around).
(b) Identify two confounders that you think might make a causal interpretation of the correlation between marriage and happiness problematic. For each, explain why you believe the confounder might affect both treatment (being married) and outcome (happiness).
(c) Sign the bias for each of the confounders you identified. Having done so, explain whether each tends to make the observed correlation between marriage and happiness an over- or under-estimate of the true causal effect.
(d) A study by Anke Zimmermann and Richard Easterlin follows people from up to four years prior to their first marriage through several years after getting married. The basic finding is illustrated in the left-hand panel of the figure on this page, which shows the life satisfaction of people who got married during the study period relative to those who never got married during the study period. As we go from left to right, we see how the life satisfaction of a person changes over time as they first cohabitate with a partner, then get married, and continue that marriage for more than a year.

    i. Compare the life satisfaction of people who have been married for a while to that of people who are not married but are living with their partner. Do you find this evidence supportive of or contrary to Gilbert's advice?
    ii. Identify a confounder that this comparison suggests may have existed in the original correlation.



Life satisfaction and marriage.

(e) A study by Jonathan Gardner and Andrew Oswald also follows individuals over time but asks a different question. It considers what happens to people's happiness when marriages end. The study looks at two ways a marriage might end: divorce or death of a spouse. The results are summarized in the right-hand panel of the figure.

The horizontal axis shows years relative to an important event (divorce or widowhood) at time 0. The vertical axis shows life satisfaction. Life satisfaction is shown in black for those who became divorced and in gray for those who became widowed.

i. Notice the initial difference in life satisfaction between those who became widowed and those who got divorced, even before the event occurred. Does this difference make you more or less confident in Gilbert's causal interpretation? Why?
ii. Now consider the widows and widowers (gray line). How does their happiness change before, during, and after the year in which their spouses passed away? Does this make you more or less confident in Gilbert's causal interpretation? What does this comparison make you think might be going on in Gilbert's original correlation?

9.5 Download "HouseElectionsSpending2018.csv" and the associated "README.txt," which describes the variables in this data set, at press.princeton.edu/thinking-clearly.

(a) Run a linear regression that finds the relationship between incumbent vote share and incumbent spending. (Note: This may require you to recode some of the variables in the data set or generate your own variables that better suit your goal.)

i. Is the correlation positive or negative?
ii. According to this data, do incumbents who spend more do better or worse?
iii. Interpret the magnitude and direction of the correlation between incumbent spending and incumbent vote share.

(b) Do the same as above for challengers.
(c) Let's think about whether the regressions you've run constitute compelling evidence of the effect of campaign spending of vote shares.

i. Identify three confounders you are worried about.
ii. Do you have any variables in this data set that measure those confounders? If so, identify a variable that might plausibly measure a confounder that is in the data set.
iii. Using linear regression, assess whether incumbent spending and challenger spending (the treatments) are in fact correlated with one of the potential confounders measured in the data set.

9.6 Find an example of a researcher, journalist, policy maker, or analyst who you believe has made an error by wrongly interpreting a correlation as credible evidence of a causal relationship. Your example should not be closely related to any example discussed in class or in the readings. Explain the evidence presented, and explain why you think this correlation is not persuasive evidence of the purported causal relationship. Discuss the likely direction of the bias. As

a bonus exercise, continue to think about your example as you read through the next four chapters. Can you think of a better way to more credibly estimate the causal relationship of interest?

Readings and References

The study of the Preuss School is

Larry McClure, Betsy Strick, Rachel Jacob-Almeida, and Christopher Reichher. 2005. The Preuss School at UCSD. Research report of The Center for Research on Educational Equity, Assessment and Teaching Excellence. create.ucsd.edu/_files/publications/PreussReportDecember2005.pdf.

The study on the Knowledge is Power Program is

Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management* 31(4):837–60.

The quote about many null findings in the literature studying the effects of charter schools is from

Julian R. Betts, Lorien A. Rice, Andrew C. Zau, Y. Emily Tang, Cory R. Koedel. 2006. *Does School Choice Work?: Effects on Student Integration and Achievement.* Public Policy Institute of California.

The study of practice and skill among violinists is

K. Anders Ericsson, Ralf T. Krampe, and Clemens Tesch-Römer. 1993. "The Role of Deliberate Practice in the Acquisition of Expert Performance." *Psychological Science* 100(3):363–406.

The study of hormonal contraception and HIV is

Renee Heffron, Deborah Donnell, Helen Rees, and Connie Celum. 2012. "Use of Hormonal Contraceptives and Risk of HIV-1 Transmission: A Prospective Cohort Study." *The Lancet Infectious Diseases* 12(1):19–26.

The study examining the correlation between electoral success and campaign spending for incumbents and challengers is

Gary C. Jacobson. 1978. "The Effects of Campaign Spending in Congressional Elections." *American Political Science Review* 72(2):469–91.

We discussed several examples drawn from:

Tyler Vigen. 2015. *Spurious Correlations: Correlation Does Not Equal Causation.* Hachette Books.

We discussed three studies of happiness in exercise 4. You can find a general discussion of happiness research in

Daniel Gilbert. 2007. *Stumbling on Happiness.* Vintage.

The study of happiness before and after marriage is

Anke C. Zimmermann and Richard A. Easterlin. 2006. "Happily Ever After? Cohabitation, Marriage, Divorce, and Happiness in Germany." *Population and Development Review* 32(3):511–28.

The study of happiness before and after the ending of a marriage is

Jonathan Gardner and Andrew J. Oswald. 2006. "Do Divorcing Couples Become Happier by Breaking Up?" *Statistics in Society* 169(2):319–36.