CHAPTER 12

Regression Discontinuity Designs

What You'll Learn

- Even when experiments are infeasible, there are still some special situations that allow us to estimate causal effects in an unbiased way.
- One such circumstance is when a treatment of interest changes discontinuously at a known threshold. Here a regression discontinuity design may be appropriate.
- Regression discontinuity designs estimate a *local* average treatment effect for units right around the threshold where treatment changes.

Introduction

In chapter 11, we saw some examples of how clever natural experiments can help us learn about causality, even when we can't run an actual experiment. The idea is to look for ways in which the world creates situations where we can make apples-to-apples comparisons without running an experiment. Sometimes, as with charter schools, the world does this through actual randomization. Other times, you have to be a little more clever.

In this chapter, we'll discuss one special situation that can help us generate credible causal estimates—when a treatment of interest changes discontinuously at a known threshold. In the next chapter we'll consider another such situation—when treatment changes over time for some units of observation but not for others.

In chapter 10, we discussed trying to learn about causal relationships by controlling for confounders. We don't typically have much faith in such approaches because it is so hard to measure all of the confounders out there. And if you can't measure something, you can't control for it. However, there are rare situations where we have a lot of information about the assignment of the treatment that may make this plausible. One example is a randomized experiment, the topic of chapter 11. If we know that treatment was assigned randomly, we know there are no confounders. The focus of this chapter is settings in which treatment is assigned according to some sharp rule. In these situations, we might be able to learn about the effect of the treatment using a regression discontinuity design.

Suppose each unit of observation is associated with a score of some sort, and treatment is determined by that score. Units whose score is on one side of a threshold get

the treatment, and units whose score is on the other side of the threshold don't. This sets up a situation where a regression discontinuity design may help you estimate causal effects. Very close to that threshold, units on either side are likely to be similar to one another on average. So a comparison of those two groups (one of whom got treatment and the other didn't) may be very close to apples-to-apples.

Let's be a little more concrete. Suppose that we want to estimate the effect of receiving a merit scholarship to college on future earnings. In general, this is difficult because the kinds of students who receive merit scholarships are probably different in many ways that matter for future earnings—intelligence, ability, ambition, work ethic—from those who do not. And, of course, we can't measure and control for all these differences.

But what if the scholarship was awarded according to a strict scoring rule? A committee generates a score from 0 to 1,000 for every applicant based on GPA, test scores, community service, and extracurricular activities. Everyone with a score of 950 or above gets the scholarship, and everyone below does not. Now, even though nothing is randomized, we might be able to learn about the effect of receiving the scholarship for those applicants who were right around the threshold of 950. How does this work?

Assume that the scholarship committee and the applicants can't precisely manipulate the scores. That is, the students put in effort without knowing exactly where their scores will fall, and the committee honestly evaluates the students also without knowing exactly where the scores will fall. Then, in expectation, the people with scores of 950 are almost identical to those with scores of 949. Nothing is randomized, but there are likely many idiosyncratic factors that could have easily pushed a 949 up to a 950, or vice versa. Had the 949s taken their standardized test on a slightly less stressful day, logged one more hour of community service out of hundreds, gotten one teacher who was a slightly more generous grader in one class, they would have been 950s and won the scholarship. Similarly, had the 950s had one minor, idiosyncratic thing not go their way, they would have been 949s and lost the scholarship. So it seems reasonable to say that, on average, the 949s are essentially the same as the 950s before the scholarship decision is made. And therefore we have something like a natural experiment. The comparison of individuals right around the threshold—some of whom got the scholarship (the 950s) and some of whom did not (the 949s) for essentially random reasons—is apples-to-apples. By comparing the future earnings of these two groups, we can estimate the causal effect of winning a merit scholarship, at least for students with scores close to the threshold.

Here's a more general way to think about this kind of situation. We want to estimate the effect of a binary treatment on some outcome. Treatment assignment is perfectly determined by some third variable (like the score above) that we call the *running variable*. Specifically, if the running variable is above some threshold for a given unit, then that unit receives the treatment ($T = 1$), and if the running variable is below that threshold, that unit does not receive the treatment ($T = 0$). Such a situation might produce data that looks like figure 12.1, with black dots corresponding to treated units and gray dots corresponding to untreated units. In the figure, the threshold is at a value of zero in the running variable.

How can we estimate the effect of the treatment in this kind of situation?

At first glance, it looks like there's not much we can do. The running variable is strongly correlated with the outcome of interest. In the scholarship example, this makes sense because the committee wants to select high-ability people, and, not surprisingly, the criteria they use to create the scores turn out to be highly correlated with future earnings, regardless of whether a student wins the scholarship. The committee uses a

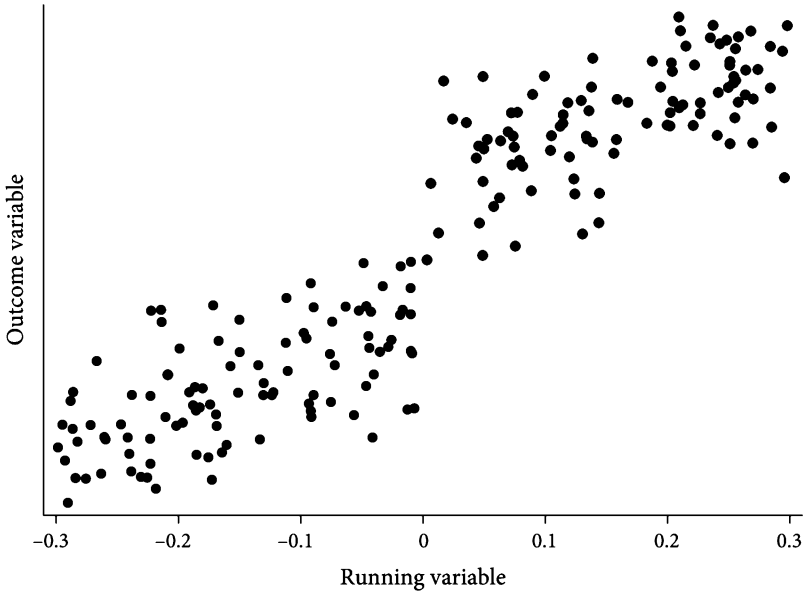

Figure 12.1. Scatter plot with treatment determined by a continuous running variable. Black dots are treated units. Gray dots are untreated units.

cutoff rule, so everyone who receives the scholarship has higher values of the running value than anyone who does not. Clearly, then, if we compare those who do and do not receive treatment, we know that the inputs to the score are confounders. And, because of the cutoff rule, we can't make an apples-to-apples comparison by finding students with the same value of the running variable, some of whom did and some of whom did not receive treatment (i.e., the scholarship). Everyone with the same score has the same treatment status.

But don't give up yet. Let's think more about what we can do here. We can estimate the expected value of the outcome for a given value of the running variable. For units whose score on the running variable is above the threshold, this will tell us the expected outcome with treatment at that value of the running variable. We can estimate this quantity for every value of the running variable all the way down the threshold. Similarly, for units whose score on the running variable is below the threshold, this will tell us the expected outcome without treatment at that value of the running variable. We can estimate this quantity for every value of the running variable all the way up to the threshold. Therefore, right at the threshold, we have estimates of the expected outcome with and without the treatment. The difference between those two values might well be a good estimate of the effect of the treatment, at least for those units with a value of the running variable right at the threshold.

We could estimate this quantity by comparing units on either side of the threshold, all of which have values of the running variable very close to the threshold. This was the idea behind comparing the 949s to the 950s to learn about the effect of merit scholarships. But there are actually somewhat better approaches.

One strategy is to run two regressions of the outcome on the running variable—one for the untreated observations below the threshold and one for the treated observations above the threshold. Then, we can use these two regressions to predict the outcomes

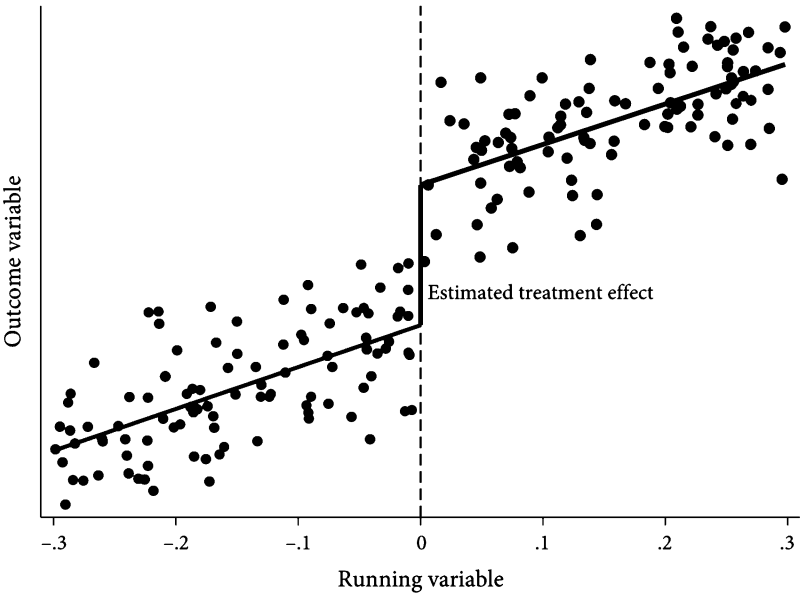

Figure 12.2. The regression discontinuity design estimates the jump in expected outcomes at the threshold, which is the causal effect of the treatment for units at the threshold.

with and without treatment right at the threshold. From these predictions we can estimate the "jump" or "discontinuity" in the outcome as the running variable crosses the threshold. That discontinuity is an estimate of the causal effect of the treatment for units right at the threshold. For this reason, we call this strategy a *regression discontinuity (RD) design*. Figure 12.2 illustrates the idea.

One thing worth emphasizing is the *localness* of the average treatment effect that a regression discontinuity design estimates. It is possible that the average effect of the treatment is different at different values of the running variable, as in figure 12.3. In this figure, both potential outcomes are shown for each unit of observation. For each unit, Y_1 is shown in black, and Y_0 is shown in gray. The actual outcomes that we observe are filled in, and the counterfactual outcomes that we don't observe are hollow. The size of the gap is different at each value of the running variable.

To be more concrete, in our example, the effect of winning a scholarship on future earnings could be different for low- and high-achieving students. The regression discontinuity estimand is the average treatment effect for units with values of the running variable right at the threshold. So, in our example, it estimates the effect of winning a scholarship on the future earnings of students with scores of 950, which might be different from the effect on students with scores of, say, 700. We refer to this estimand as a *local average treatment effect* (LATE). As always, the LATE can differ from the overall average treatment effect in the population. So it is important, when using a regression discontinuity design, to think about whether the quantity estimated is really the one you are interested in.

Regression discontinuity designs are important in a variety of settings. One common application is in estimating the effects of government programs. Many policies change discontinuously at known thresholds. For example, individual-level government benefits are often means-tested, with eligibility determined by whether some continuous

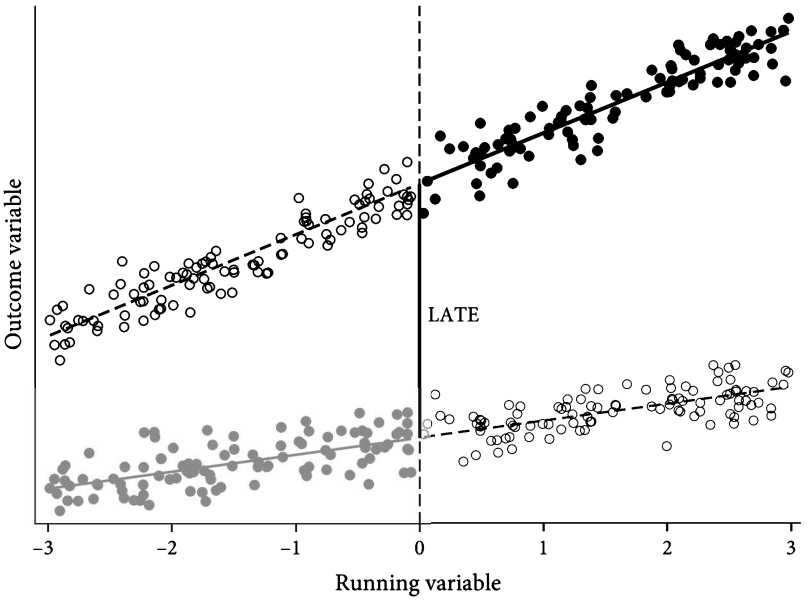

Figure 12.3. A regression discontinuity design estimates the LATE at the threshold. This need not be the overall average treatment effect, as average treatment effects may differ for different values of the running variable

measure of income or poverty is on one or the other side of a threshold. County-level policies are often determined by population thresholds or by the share of residents of a certain type. Regression discontinuity designs provide a straightforward way to estimate the effects of these programs. Furthermore, these designs estimate the effects of the programs for the kinds of people or places about which we care the most—the marginal unit that was just barely eligible or ineligible. So if policy makers are trying to figure out whether they should shrink or expand a particular government program, these regression discontinuity estimates should be highly informative.

How to Implement an RD Design

There are different ways for analysts to go about implementing their own regression discontinuity designs, and there are pros and cons associated with each one.

The simplest approach, as mentioned above, is to just compare the mean outcome for small ranges of the running variable (sometimes called *bins*) on either side of the threshold. For example, we might compare the average earnings for applicants who scored between 950 and 954 to the average earnings for applicants who scored between 945 and 949. For reasons you'll see in a moment, we often call this the *naive* approach.

A clear advantage of the naive approach is its simplicity. What makes it naive is the fact that it is virtually guaranteed to produce biased estimates. Why is this? The running variable is typically correlated with potential outcomes. Why would the committee use the scores to allocate scholarships if they didn't believe the scores corresponded to ability, effort, motivation, or some other factor that is likely correlated with earnings in the future?

Because the running variable is correlated with the potential outcomes of interest, there will always be some baseline difference between the groups just above and just below the threshold. Of course, as the size of the bins being compared (sometimes called the *bandwidth*) shrinks, the bias should shrink, but it will never disappear.

We can already see that one of the important decisions an RD analyst must make is to select a bandwidth. And when they make that decision, they often face a trade-off between reducing bias and improving precision. Smaller bandwidths will generally yield less biased estimates but also less precise estimates because they are using less data.

A potentially less biased alternative to the naive approach is the *local linear* approach. Here, we again select a bandwidth, and for observations within that bandwidth, we run linear regressions of the outcome on the running variable separately on either side of the threshold. We use these estimates to get predicted values of the outcomes with and without treatment right at the threshold, and the differences in those predicted values is our estimate of the effect of the treatment for units at the threshold.

With this approach, we're allowing for the possibility that there is a relationship between the running variable and the outcome, we're allowing that relationship to be different on either side of the threshold, and we're assuming that this relationship is approximately linear (at least for the small window of data that we're analyzing). That is the approach we took in figure 12.2.

To make our lives easier and to obtain an estimate of the standard error, there is a way to implement this local linear approach with a single regression rather than running two separate regressions. First, rescale the running variable so the threshold is zero (i.e., subtract the value of the threshold from the running variable). Second, generate a treatment variable indicating whether an observation is above or below the threshold. Third, generate an interactive variable by multiplying the treatment variable and the rescaled running variable. And lastly, regress the outcome on the treatment, the rescaled running variable, and the interaction of the two for the observations within your bandwidth. The estimated coefficient associated with the treatment provides the estimated discontinuity.

A third common way that people implement RD designs is with polynomial regressions. An analyst might regress the outcome on the treatment, the running variable, and higher-order polynomials (i.e., the running variable to the second power, third power, and so on). This approach accounts for a possible non-linear relationship between the running variable and the outcome. A downside is that data points that are far from the threshold can have a big effect on the estimated discontinuity.

When implementing an RD design, the researcher clearly gets to make a lot of choices, so they have to be careful to avoid the problem of over-comparing and under-reporting. Your particular decisions should depend on your substantive knowledge and beliefs about the relationship between the running variable and the outcome and also how much bias you're willing to accept in exchange for a gain in precision, or vice versa. The best approach is to justify your choices with a combination of theory, substantive knowledge, and data analysis and, perhaps most importantly, show results for different specifications. If your estimates are robust across different bandwidths and specifications, this will lend additional credibility to your results. If your result only holds for one very particular specification, you should be skeptical.

To illustrate how one can explore robustness across bandwidths, figure 12.4 shows an analysis from one of Anthony's papers coauthored with Haritz Garro and Jorg Spenkuch. They hoped to test whether firms benefit from political connections by testing whether a firm's stock price increases when a political candidate to whom the firm

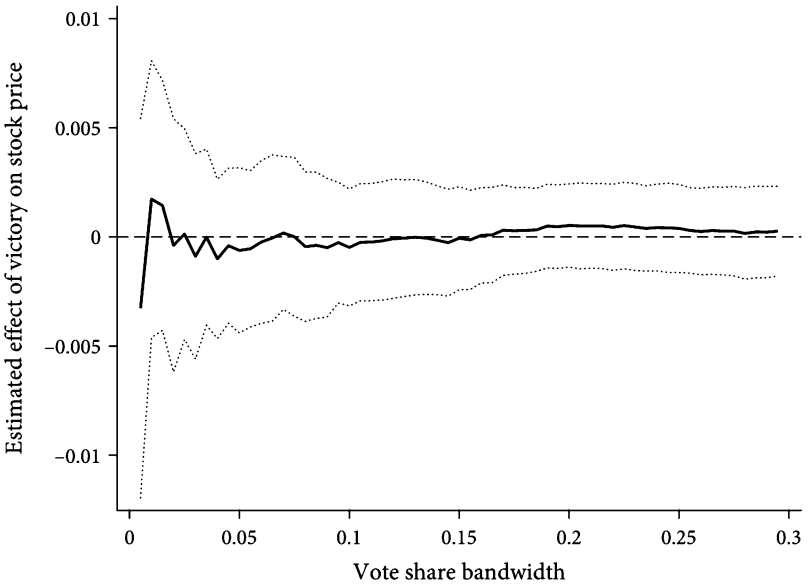

Figure 12.4. Visualizing how an RD estimate (solid) and confidence interval (dotted) depends on the bandwidth.

made a campaign contribution barely wins versus barely loses. So the outcome is a measure of the change in a firm's stock price, the running variable is the vote share of the politically connected candidate, and the treatment is an indicator for whether that candidate won the election.

They use a local linear approach, but they want to make sure that their results are robust to different bandwidths. Figure 12.4 shows the estimated effects along with the upper and lower bounds of the 95 percent confidence interval for sixty different possible bandwidths between 0.5 and 30 percentage points. As we would expect, the confidence intervals are larger and the estimates are more volatile for smaller bandwidths, but the estimates become more precise as the bandwidth increases and more data is included. Fortunately, the estimates are similar for almost all of the bandwidths, which is reassuring. Had the estimate changed meaningfully as the bandwidth increased, that would suggest a trade-off between bias and precision, and we'd have to think further about which estimates we trust more.

Let's think more about how to implement and interpret regression discontinuity designs through an example. The winners and losers of elections are determined solely by vote shares, so if we want to estimate the effects of a certain kind of election result, a regression discontinuity design might be especially useful.

Are Extremists or Moderates More Electable?

Surrounding both the 2016 and 2020 presidential elections, the Democratic party engaged in a heated debate about the electability of extremist versus moderate candidates. In particular, the liberal wing of the party was disappointed by the nominations of Hillary Clinton and Joe Biden, both of whom they perceived as too moderate. The way to win elections, they argued, isn't to appeal to centrist voters. Rather, parties should nominate ideologically pure candidates who can turn out the base. Bernie Sanders, the

argument went, was in a better position to defeat Donald Trump in the general election than either of his more moderate rivals. Sure, there might have been some moderates turned off by some of Sanders's policy proposals. But Sanders would have more than made up for those losses by mobilizing progressives who had lukewarm feelings about Clinton and Biden.

How can we assess whether this argument is right? On the one hand, moderate candidates might persuade more people in the middle to support their party. On the other hand, extremists might mobilize the base. So if you want to maximize the chances that your party wins the general election, whom should you support in the primary election? It is, of course, impossible to say with confidence what would have happened if, counterfactually, Sanders had won the 2016 or 2020 Democratic nomination (remember the fundamental problem of causal inference from chapter 3). But maybe we can say more about what happens, on average, when a party nominates a more extreme versus more moderate candidate.

To try to get a handle on this, let's turn to congressional elections, for which we have a lot more data than we do for presidential elections. At first glance, it looks like the advocates of ideologically pure candidates might be onto something. After all, it sure looks like Congress has a lot of ideological purists in it. If moderation is a winning strategy, why are there so many extremists in office?

For starters, we have to make sure we aren't forgetting the lesson of chapter 4: correlation requires variation. The fact that many congresspeople are ideologically extreme does not imply a positive correlation (to say nothing of a causal relationship) between ideological extremism and electoral success. To ascertain the correlation of interest, we need to compare the electoral fortunes of extremists and moderates. Sure, one possible explanation of the large number of extremists in Congress is that extremism really is correlated with winning. But another is that there are just very few moderates running.

Moreover, it may be misleading to think about extremism and moderation on a national scale. Rather, for the purpose of thinking about electoral strategy, we want to know whether a candidate is extreme or moderate relative to the preferences of their particular electorate or constituency. Sanders is surely an extreme liberal relative to the median voter in the United States. But when he's running to represent Vermont in the Senate, perhaps he's only somewhat left of center. Indeed, maybe many congresspeople appear ideologically moderate relative to their constituencies but ideologically extreme relative to the country as a whole. This could happen if the constituencies are themselves constructed to be ideologically extreme compared to the country—some far to the left and others far to the right. But in this case, you wouldn't want to interpret the presence of lots of ideological extremists as evidence that extremism itself is an effective electoral strategy, because the winning congressional candidates would not have been perceived as ideological extremists by the voters that elected them.

Given these concerns, what we really want to know is not the correlation between ideology and electoral success but the effect of nominating an ideologically extreme candidate on electoral fortunes. To find an unbiased estimate of this, we need to compare how parties do in elections when they nominate an extremist versus a moderate, all else equal. On average, is the party better off running an extremist or a moderate candidate?

Of course, a naive comparison of the correlation between electoral outcomes and ideological extremism of candidates isn't apples-to-apples. Presumably, the times, places, and situations where a party nominates a moderate are different from those where a party nominates an extremist for all sorts of reasons that are consequential for electoral

outcomes. For instance, most likely, liberal Democrats win primaries in more liberal places where the Democratic Party is stronger, and moderate Democrats win in more conservative places where the party is weaker. So if we found that extremists do better in general elections, that wouldn't tell us that parties are better off when they elect extremists. The causal interpretation of that correlation would obviously be confounded. We could try to control for differences across time and place, but we would always be worried that there are still unobservable baseline differences between places nominating extremists and moderates. We can do a better job using a regression discontinuity design.

Major party congressional candidates are selected in primary elections. And election outcomes are determined by a sharp threshold. Suppose we analyze a large sample of primary elections that pitted one extreme candidate against one moderate candidate. The treatment we are interested in is the nomination of an ideologically extreme candidate. We want to know the effect of that treatment on the party's vote share in the general election. To set up the RD, define the running variable as the vote share of the extreme candidate in the primary. If that vote share is below one-half, the party runs the moderate in the general election; if it exceeds one-half, the party runs the extremist. We can now estimate the effect of running an extremist by implementing an RD design, comparing a party's general election outcome when it just barely nominated an extremist in the primaries versus when it just barely nominated a moderate in the primaries.

Andrew Hall did exactly this in a 2015 study. He estimated a large, negative discontinuity in a party's general election results at the threshold. That is, on average, a party that nominates an ideological extremist instead of a moderate significantly decreases its performance in the general election. Despite the predictions of the Sanders supporters, the evidence suggests that nominating extremists, on average, is a bad electoral strategy.

Hall's design is illustrated in figure 12.5. The two lines represent separate linear regressions on each side of the 50 percent threshold. Each small gray circle corresponds to one observation—a party election. The larger, black circles show the average general election vote share for .02-point bins of the winning margin. The large negative discontinuity right at the threshold is the estimated effect on general election vote share of nominating an extremist instead of a moderate for a race where the primary election was evenly split between a moderate and an extremist.

What explains this result? In a follow-up study, Hall and Dan Thompson investigate further. Using a similar regression discontinuity design, they study the effect of nominating an extremist on voter turnout. Interestingly, contrary to the predictions of the Sanders supporters, there's no evidence that extremist candidates turn out the base. Or, rather, nominating an extremist does appear to turn out the base, but the wrong one. When a party runs an extremist candidate, more people from the *other* party turn out to vote in opposition. Therefore, if we had to guess, these results suggest that if Bernie Sanders had won the Democratic primary in 2016 or 2020, he would have performed worse than Clinton and Biden. He likely would have lost some of the centrist voters that preferred Clinton or Biden over Trump *and* likely would have motivated Republican voters to turn out in greater numbers.

Continuity at the Threshold

In order for the regression discontinuity approach to provide an unbiased estimate of the causal relationship, it has to be the case that treatment status changes sharply at the threshold *and* nothing else that matters for outcomes does. If baseline

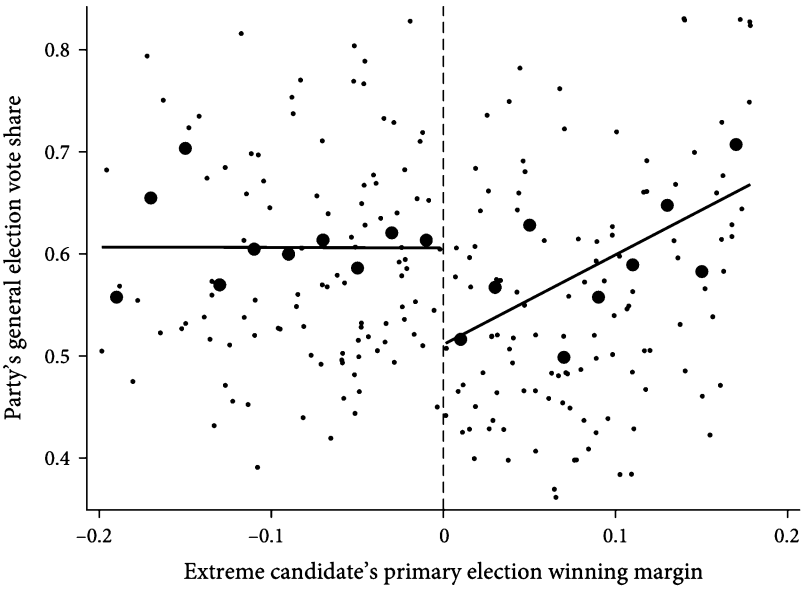

Figure 12.5. The effect of running an extremist on electoral prospects.

characteristics also change discontinuously at the threshold, then any differences in average outcomes right around the threshold could be due to those changes in baseline characteristics rather than treatment. That is, the comparison of treated and untreated units would no longer be apples-to-apples, even right at the threshold, because those two groups would be differentiated by things other than just treatment status. But if average baseline characteristics of the units change continuously (rather than in a discrete jump) as the running variable passes through the threshold, then we can obtain an unbiased estimate of the effect of the treatment for units with a value of the running variable that is right at the threshold because the only thing that will differentiate units just on one or the other side of the threshold, on average, will be their treatment status. We call the requirement that baseline characteristics don't jump at the threshold *continuity at the threshold* (or just *continuity* for short).

Let's see why continuity is crucial. Figure 12.6 illustrates what it looks like if the continuity condition is satisfied. As with figure 12.3, the filled-in dots are data we actually observe. The solid lines plotted through them are the average potential outcome functions (for the relevant value of treatment assignment). The hollow dots are data we don't observe (since we don't ever observe, say, the potential outcome under treatment for a unit with a value of the running value below the cutoff). The dashed lines plotted through them are the average potential outcome functions (again, for the relevant value of treatment assignment). Continuity is satisfied because these average potential outcome functions have no jump. That is, the average potential outcomes under both treatment and no treatment are continuous at the threshold. All that changes at the threshold is that units go from being untreated to treated.

Importantly, if continuity holds, then the gap between the gray and black dots at the threshold is in fact the LATE at that threshold, which is just what we want.

But what if continuity does not hold, so that the potential outcomes look like figure 12.7?

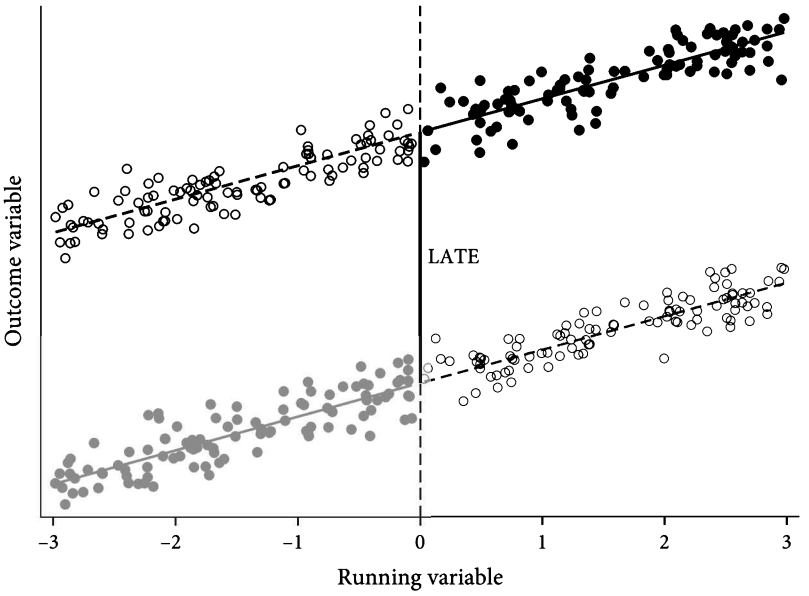

Figure 12.6. A case where average potential outcomes satisfy continuity.

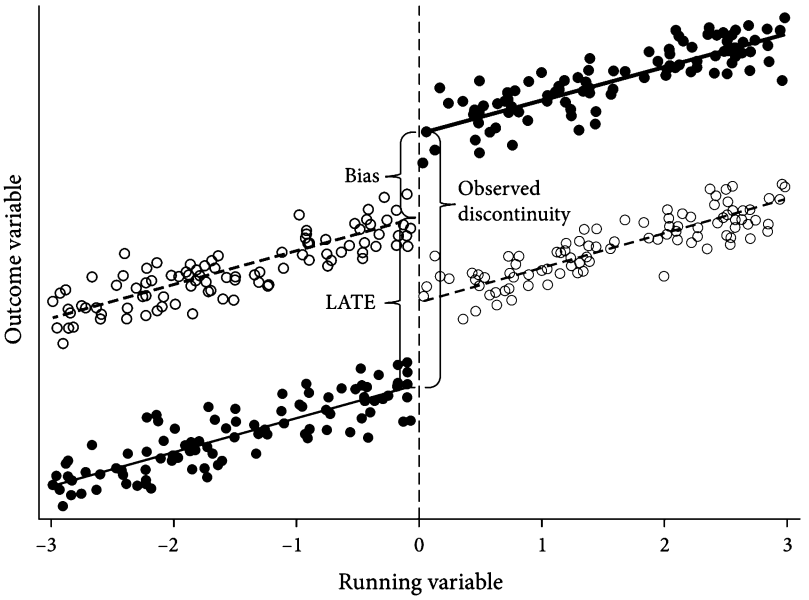

Figure 12.7. A case where average potential outcomes do not satisfy continuity at the threshold.

The true average treatment effect at the threshold is the difference between the filled-in gray dots and the hollow black dots at the threshold. (You could also define it as the difference between the filled-in black dots and the hollow gray dots.) But, right at the threshold, the potential outcomes jump up, even absent a change in treatment. We don't know why, but something besides treatment is changing right at the threshold. As a

consequence, not all of the observed gap—that is, the jump between the filled-in gray and the filled-in black dots—is the result of the change in treatment. Some of it is the result of whatever else is changing. As such, that gap is a biased estimate of the LATE—in this case it is a big over-estimate of the true effect of the treatment—since the gap includes both the effect of the treatment change and also the effect of whatever else is changing. Thus, without continuity at the threshold, the RD will give a biased estimate of the local average treatment effect.

When it comes to implementing an RD design, there are many different paths the analyst can take. However, viewed in the correct light, once a researcher has established that plausibility of continuity of potential outcomes at the threshold, their job is clear. Using the sort of techniques we have already discussed (e.g., regression), they simply have to generate unbiased estimates of two things—the average outcome with and without treatment at the threshold.

To think about when an RD design is appropriate, we want to think about when continuity at the threshold is or is not plausible. It is worth noting that the continuity requirement is less demanding than you might have expected for credibly estimating causal relationships. For instance, it does not require that the treatment is assigned randomly (even by nature). In our scholarship case, we were able to use an RD even though, for every single student, treatment assignment was deterministic (i.e., there was no randomness at all). Continuity also does not require that the outcome be unrelated to the running variable. Again, in our scholarship example, the running variable reflects genuine academic merit and, thus, is positively correlated with future earnings outcomes. Finally, it does not require that units have no control over their value of the running variable or that units have no knowledge of the threshold. In our scholarship example, students could do all sorts of things to affect the running variable (e.g., study harder, do more community service).

So what could go wrong such that continuity does not hold?

Suppose that units have extremely precise control over their value of the running variable such that certain types may cluster just above or just below the threshold. This could potentially be a problem. In our scholarship example, we might worry that more privileged or more ambitious students have better information about the scoring system and can do just enough to exceed the threshold. Or we might worry that the committee has reasons to want to grant scholarships to students with certain characteristics (e.g., children of donors, athletes, particular racial or ethnic groups) and manipulates the scores or the threshold a little bit to get the desired result. In both of these cases, individuals just above the threshold would not be comparable to those just below. Instead they would have been *sorted* (by themselves or others) around the threshold by other baseline characteristics that matter for outcomes. If this is the case, regression discontinuity does not provide an unbiased estimate of the causal effect.

Things can go wrong even without sorting around the threshold, simply because things other than treatment status change at the threshold. Here's a pretty interesting real-world example. In France (and many other countries), a mayor's salary depends on the size of a city's population. For instance, by law, mayoral salaries jump when a city has a population of more than 3,500 residents.

This seems like an opportunity to use an RD to learn about the effect of mayoral salary on all sorts of outcomes. For instance, we might want to know whether cities are better governed or elections are more competitive when mayors are paid more. For either of these outcomes, the treatment of interest is mayoral pay. The running variable is population. And we happen to know that, by law, there is a discontinuous jump in

the treatment as the running variable crosses the 3,500-resident threshold. Surely cities with 3,400 residents and cities with 3,600 residents are similar on average.

It looks good, no? But there's a problem with continuity. It doesn't come from towns strategically determining their populations to change the mayor's salary. It comes from other policies. You see, mayoral salary is not the only feature of city governance that changes by law at the 3,500-resident threshold. Other things that change include the size of the city council, the number of deputy mayors, the electoral rules, the process for considering a budget, gender-parity requirements for the city council, and so on. So any discontinuity in outcomes at the 3,500-resident threshold does not provide an unbiased estimate of the effect of mayoral salary because other characteristics that might matter for those outcomes also change discontinuously at the threshold.

Clearly, then, before interpreting the results from an RD as an unbiased estimate of a causal relationship, it is important to assess the plausibility of the continuity assumption. There are several ways to do this. The most important is to think substantively. The best way to spot possible violations of continuity is to know a lot of details of the situation, so that you can be alert to the potential for sorting, manipulation, or other things changing at the threshold. In our scholarship example, if you had sat in on a committee meeting or had deep knowledge of the kinds of characteristics the committee was under pressure to make sure were well represented among scholarship recipients, you would be in a better position to assess the plausibility of the continuity assumption than if you had no specific substantive knowledge of the situation. There are also other kinds of analyses one can do to help validate the continuity assumption. For instance, an analyst can look directly at measurable pre-treatment characteristics and see whether they seem to have discrete jumps at the threshold. If many measurable characteristics appear continuous at the threshold, we might be more confident that other, unmeasured baseline characteristics are also continuous. One can also look at the distribution of the running variable itself. If we find bunching—that is, significantly more units whose value of the running variable is just above the threshold than just below, or vice versa—then we might be concerned about some manipulation that violates continuity.

Exactly how bad a violation of the continuity assumption is depends on the details of the problem. If there is just a little sorting, or a small discontinuity in baseline characteristics, the RD is biased, but perhaps only a little bit. And if the researcher has a lot of data and, so, can focus on units only extremely close to the threshold, sorting would have to be extremely precise for it to affect the results. For instance, if we are estimating our scholarship RD using data on students with scores in the 940–949 range and students in the 950–959 range, we might be more concerned about sorting than if we have enough data so that we can consider just students with a score of 949 or 950.

Does Continuity Hold in Election RD Designs?

As we discussed earlier in this chapter, elections are a great setting for RD designs since they have a clear running variable and a sharp threshold for winning. Not surprisingly, the election RD has been used in many studies on the effects of elections on outcomes ranging from campaign donations to drug violence to nominating an extremist versus a moderate candidate. So it is important to think clearly about whether the election RD is in fact a good research design.

Let's remember what needs to be true for the election RD to provide an unbiased estimate of a causal relationship. We need for everything else that matters for the outcome under study to be continuous at the threshold. This guarantees that places where

the relevant candidate (e.g., an extremist) just barely won are on average comparable to places where the relevant candidate just barely lost. In any application of the RD approach, including elections, it is always important to ask if this condition is plausible.

And, indeed, some studies have argued that continuity may be violated in some electoral settings. The concerns have to do with manipulation of election results in close elections. For instance, in Hall's study on the effects of nominating an extreme candidate, perhaps the party leadership prefers moderates. If it has ways of intervening (say, by putting pressure on officials responsible for recounts) to nudge close election outcomes, it might do so in favor of moderate candidates. For his study, Hall shows that this does not appear to be the case.

But in another setting, the post-WWII U.S. House of Representatives, some evidence suggests that there may be continuity problems. In the relevant studies, scholars are interested in using the RD to estimate the *incumbency advantage*—How much better does the incumbent party do than the out-party, all else equal? A researcher might compare the probability a Democrat wins an election in situations where a Democrat just barely won or lost the previous election in the hopes of estimating the effect of one election result on subsequent election results. For this to be a valid research design, there must be continuity at the threshold—the probability of the Democrat versus the Republican winning in the next election wouldn't change discontinuously in vote share in the previous election if it weren't for the fact that the previous election result was different. But there is reason to worry this isn't true. In particular, in House elections decided by less than 0.25 percent of the vote, the incumbent party is statistically more likely to win than the challenging party. If this is because parties are able to manipulate close election outcomes, then we might worry that, even very close to the 50 percent threshold, we aren't making an apples-to-apples comparison when we compare future electoral outcomes in places where one party just barely won versus just barely lost. So, what's going on?

Devin Caughey and Jas Sekhon, who wrote a study about this phenomenon, argue that the evidence points to electoral manipulation—incumbents have very precise knowledge of expected vote share and act strategically on or before election day in ways that allow them to win very close elections more than half the time. To believe this, however, you must believe that incumbent candidates can distinguish between situations where they expect their vote percentages to fall between 49.75 and 50.0 versus 50.0 and 50.25. Real-life campaigns appear to have nowhere near this level of precision in their election forecasts. Therefore, strategic campaigning is unlikely to be the explanation. What else could explain the imbalance? Most likely, this is a case of noise producing a false positive, much like Paul the Octopus in chapter 7. When Anthony and four coauthors replicated the same tests that Caughey and Sekhon did, but for twenty different electoral settings across several countries, the postwar U.S. House was the only one for which such an imbalance was present. Thus, we suspect the election RD is in fact a good research design for learning about causal relationships in politics.

Noncompliance and the Fuzzy RD

Thus far, we've talked about using a regression discontinuity design when treatment is completely determined by the running variable and the threshold. When this is the case, we sometimes say we are using a *sharp regression discontinuity design*.

But, just as in experiments, there are sometimes problems of noncompliance in settings that are otherwise suitable for an RD. That is, treatment may be discontinuously

affected by which side of the threshold the running variable is on, but not deterministically. In addition to the compliers, there are some never-takers (units with values of the running variable above the threshold but who are untreated) and there are some always-takers (units with values of the running variable below the threshold but who are nonetheless treated).

When there are such noncompliers, we need to combine the regression discontinuity approach with an instrumental variables (IV) approach of the sort we discussed in chapter 11. We do so by using which side of the threshold the running variable is on as an instrument for treatment assignment. This approach is sometimes called a *fuzzy regression discontinuity design.* To see how fuzzy RD works, let's work through an example.

Bombing in Vietnam

A classic question in counterinsurgency is whether violence by counterinsurgents that kills civilians as well as combatants is productive or counterproductive. Melissa Dell and Pablo Querubin shed some quantitative light on this question in the setting of the U.S. bombing strategy during the Vietnam War.

In Vietnam, the United States engaged in a massive bombing campaign in an attempt to suppress the Viet Cong guerilla forces in the north. Dell and Querubin want to evaluate whether such bombing worked.

One comparison they might make to try to answer that question is whether insurgents were more or less active in the parts of Vietnam that experienced more bombing. But if you think clearly, you'll see that such a comparison is not apples-to-apples. One might, for instance, worry that the United States was more likely to bomb locations where the insurgents were already quite active, in which case there would be a reverse causality problem.

In order to better estimate the effect of bombing, Dell and Querubin use a regression discontinuity design. The history underlying their design is quite amazing.

During the Vietnam War, Secretary of Defense Robert McNamara was obsessed with quantification. McNamara had pioneered the use of quantitative operations research during his time as president of Ford Motor Company. And at the Department of Defense, he surrounded himself with a group of "whiz kids" and a large team of computer scientists, economists, and operations researchers, with the goal of providing precise, scientific, quantitative guidance to war planners and the military.

One of these efforts was the Hamlet Evaluation System (HES). This project collected answers to an enormous battery of monthly and quarterly questions about security, politics, and economics. The data were collected by local U.S. and South Vietnamese personnel who obtained information by visiting hamlets. Question answers were entered by punch card into a mainframe computer, and then a complex algorithm converted them into a continuous score, ranging from 1 to 5, that was supposed to characterize hamlet security. These raw scores, however, were never reported out by the mainframe. No human ever saw them. Instead, the computer rounded the scores to the nearest whole number, so that all the analysts or decision makers ever saw was a grade of A, B, C, D, or E. Better letter grades were understood to correspond to greater hamlet security. These grades helped determine which hamlets should be bombed—with bombing being more often targeted at hamlets receiving worse grades.

Dell and Querubin were able to reconstruct the algorithm and, using declassified data, recover the underlying continuous scores. This set them up for a regression discontinuity design.

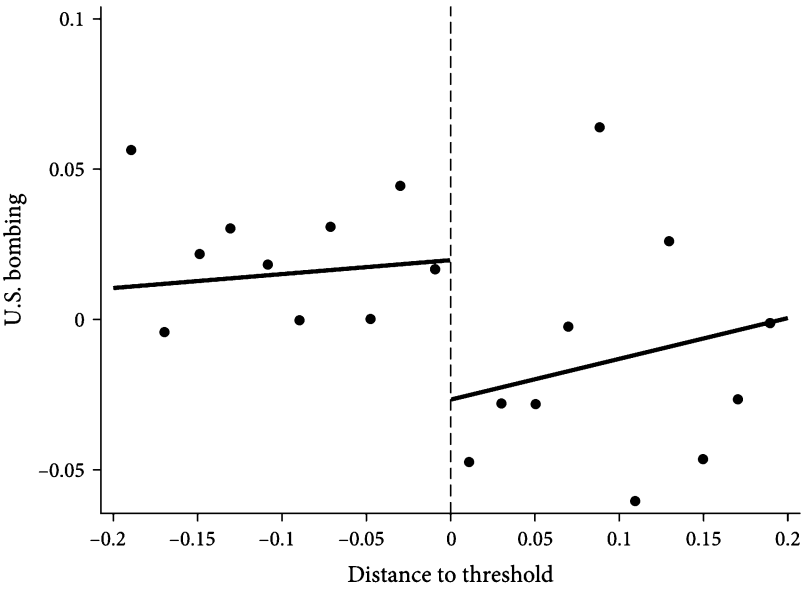

Figure 12.8. Hamlets that just barely received better grades in the Hamlet Evaluation System were bombed less frequently than hamlets that just barely received worse grades.

Think about hamlets with scores in the 1.45–1.55 range. Some of these hamlets ended up with a score just below 1.5 and received an E. Others ended up with a score just above a 1.5 and received a D. But the difference between, say, a 1.49 and a 1.51, on a score created by a complicated (and largely arbitrary) combination of answers to 169 questions is probably pretty arbitrary. So we should expect that the underlying level of Viet Cong activity in these two types of hamlets is the same—that is, we should expect the potential outcomes to be continuous at the threshold.

But treatment—which, here, means being bombed by the United States—changes discontinuously at the threshold. U.S. war planners did not ever see the underlying continuous score. All they saw was the letter grade. And, so, they perceived hamlets that received a D as more secure than hamlets that received an E (and similarly for D vs. C, C vs. B, and B vs. A). As such, they were more likely to bomb the hamlets with lower letter grades.

Figure 12.8 shows that this was the case. The horizontal axis measures the running variable—the distance of the first decimal of a hamlet's score from .5. Hamlets whose value of the running variable is negative (because its score's first decimal was below .5) were rounded down to the nearest letter grade, while those whose value of the running variable is positive were rounded up.

The vertical axis measures the frequency with which a given hamlet was bombed after the scores were tabulated. The gray dots correspond to binned averages of many hamlets with similar values of the running variable. The dark lines correspond to separate regressions on either side of the threshold. The figure shows a discontinuous jump down in the frequency of U.S. bombings at the threshold—hamlets that just barely received better grades were bombed less frequently than hamlets that just barely received worse grades.

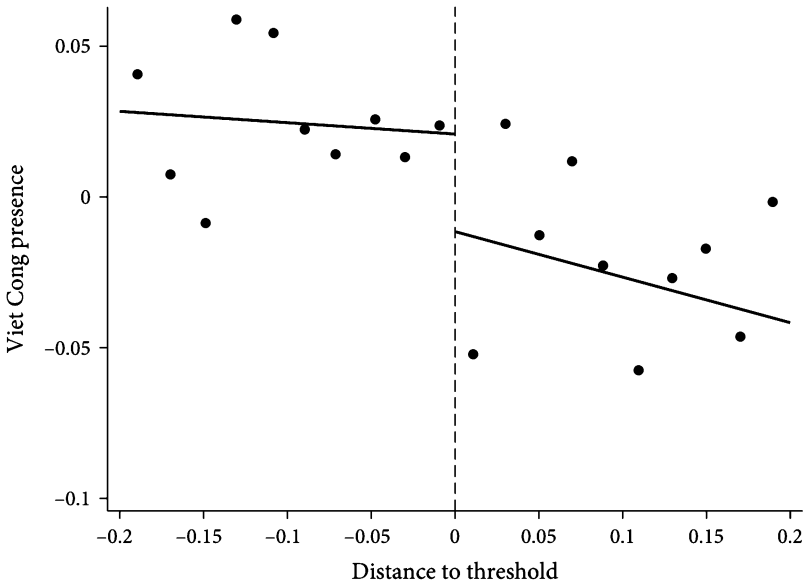

Figure 12.9. Hamlets that experienced more bombing saw more subsequent insurgent activity compared to otherwise similar hamlets that experienced less bombing.

Given this discontinuous change in treatment, it makes sense to use a regression discontinuity design to estimate the effect of bombing on insurgency. Figure 12.9 illustrates the idea. The horizontal axis is the same running variable as above. But now the vertical axis is the outcome of interest—Viet Cong activity in the hamlet following the tabulation of the scores. As the figure shows, indiscriminate bombing appears to have been counter-productive. There is a discontinuous drop in Viet Cong activity at the threshold. This means that hamlets that were bombed more (those to the left of the threshold) experienced more insurgent activity than otherwise similar hamlets that were bombed less.

But notice there is something a little different here from our normal regression discontinuity story. The treatment is not binary (there's a continuum of bombing intensity), and going from a better score to a worse score did not guarantee increased bombing. The security score was only one input to bombing decisions. So it was not the case that treatment went from fully on to fully off at the threshold. That is to say, there was likely noncompliance—hamlets whose treatment status didn't depend on which side of the threshold their score fell.

But we know what to do about noncompliers. As we discussed in chapter 11, we can use an IV approach. Recall, an instrument must satisfy several conditions:

1. **Exogeneity:** The instrument must be randomly assigned or "as if" randomly assigned, allowing us to obtain unbiased estimates of both the first-stage and reduced-form effects.
2. **Exclusion restriction:** All of the reduced-form effect must occur through the treatment. In other words, there is no other pathway for the instrument to influence the outcome except through its effect on the treatment.

3. **Compliers:** There must be some units that receive a different value of the treatment as a result of the instrument.
4. **No defiers:** Whatever the sign of the first-stage effect, there must be no units for whom the instrument affected their treatment value in the opposite direction.

How would we apply an instrumental variables approach here? The idea is to use *which side of the threshold our running variable is on* as the instrument. Let's see that this satisfies the four conditions needed for an instrument.

The whole point of the regression discontinuity design is exogeneity. If potential outcomes are continuous at the threshold, then the RD allows us to obtain an unbiased estimate of both the first stage (the effect of the instrument on bombing, as illustrated in figure 12.8) and the reduced form (the effect of the instrument on Viet Cong activity, as illustrated in figure 12.9).

The exclusion restriction requires that *which side of the threshold the running variable is on* has no effect on Viet Cong activity other than through its effect on bombing. Here there are questions to be asked. For instance, we need to worry about whether these grades were used for any other U.S. military or policy decision making. If so, then the instrument will not satisfy the exclusion restriction.

Dell and Querubin provide two kinds of evidence in support of the plausibility of the exclusion restriction. First, they repeat their RD analysis for lots of other kinds of military operations by both the American and South Vietnamese militaries. They find no evidence of any other kind of military operations changing discontinuously at the threshold. As such, it is unlikely that the effects they find are the result of military actions other than bombing. Second, they review the administrative history of the Hamlet Evaluation System. That review reveals little evidence of the HES scores being used for any other policy decision making. The one exception is a program aimed at driving the Viet Cong out of the least secure hamlets. But that program had ended before the sample period covered by Dell and Querubin's data.

The requirements that there be compliers and no defiers are the most straightforward. It is clear from both the data and the history that the letter grades affected bombing. And it seems unlikely that there were defiers—hamlets that were bombed more because they received a *better* security score. However, unlike in our previous examples, compliance is not so discrete. Different units can change their treatment status in response to the instrument by different amounts.

Given all of this, Dell and Querubin feel justified in employing a fuzzy RD design—using *which side of the threshold a hamlet's security score was on* as an instrument for bombing. In doing so, they are estimating an estimand that is a bit of a mouthful since it reflects the localness of both the RD and the IV. In particular, they are estimating the local average treatment effect of bombing on insurgent activity for hamlets with scores close to the threshold (the LATE from the RD) whose level of bombing is responsive to that score (the CATE from the IV).[1] Doing so, they find that bombing was counterproductive. For such hamlets, going from experiencing no bombing to experiencing the

[1] Further complicating matters, each hamlet is not simply either a complier or not. There is potentially a continuum of compliance whereby the instrument increases bombing in some hamlets by a lot, others by a little, and so on. So instead of thinking about a complier average treatment effect, we actually have to think about a weighted average treatment effect, where each hamlet is weighted according to the extent to which bombing responded to the score in that case.

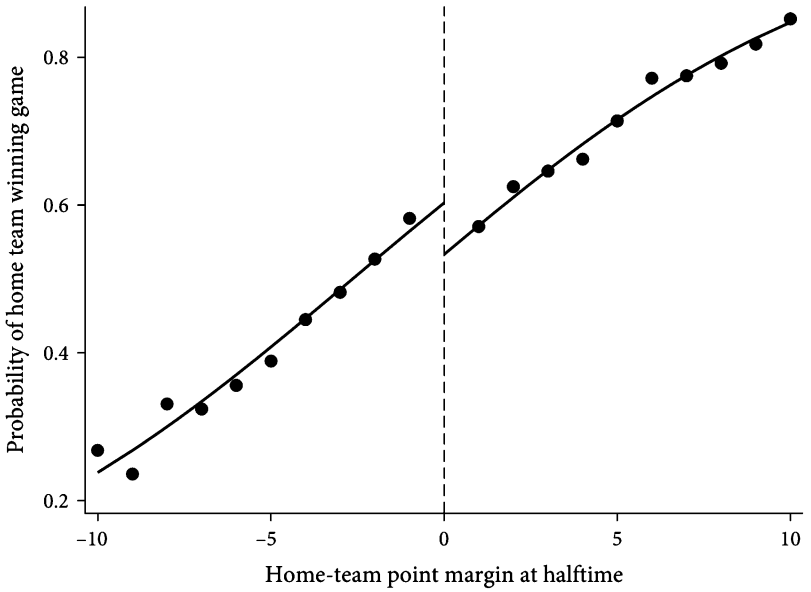

Figure 12.10. The effect of being ahead or behind at half time on winning the game.

average level of bombing increased the probability of Viet Cong activity in the hamlet by 27 percentage points.

Motivation and Success

Let's end with one last, fun example of a regression discontinuity design. Jonah Berger and Devin Pope implement an RD to estimate the effect of psychological motivation on performance. They analyze over eighteen thousand professional basketball games to test whether the motivation of being behind and needing to catch up leads to better performance than the complacency of being ahead and simply needing to hold onto a lead. Their running variable is the point margin of the home team at halftime, and they test whether the probability of ultimately winning a game changes discontinuously as the halftime point margin crosses the threshold of 0, when the home team goes from being just behind to just ahead.

Figure 12.10 shows the results. As we would expect, the point margin at halftime is correlated with the probability of ultimately winning the game. When the home team is 10 points ahead at halftime, they go on to win about 85 percent of the time, but when they're 10 points behind, they only win 25 percent of the time. This makes sense since some teams are better than others—good teams are both more likely to be ahead at the half and more likely to win the game. More interesting, however, is the comparison when the score is almost tied at the half. Presumably, there is very little quality difference, on average, between teams that are ahead or behind by just 1 point at halftime. Yet, the home team is actually more likely to win when they're 1 point behind at halftime than when they're 1 point ahead. Berger and Pope's regression discontinuity shows that being just barely behind increases the probability that the home team wins by 6 percentage points! Maybe those inspirational halftime speeches really do work.

Wrapping Up

When we know that a treatment of interest was determined (at least partly) by a threshold or cutoff, an RD design might allow us to obtain credible estimates of the effect of that treatment at that cutoff.

These situations arise more frequently than you might think. Suppose you're working for a baby food company that asks you to estimate the effect of their television ads. You probably can't convince the marketing department to randomize where they advertise; they want to advertise in places where they are likely to have the biggest effects. But maybe they already decided to air television ads in all media markets where more than 3 percent of households have an infant. This is a perfect opportunity for an RD design. Nothing was randomized, the marketing department did what it wanted to do anyway, but you have an opportunity to learn about the effectiveness of advertising by comparing baby food consumption in places just above and just below that 3 percent threshold.

Another opportunity for us to obtain credible estimates of causal relationships absent any randomization is when treatments change for some units and not others. In these cases a difference-in-differences design may be appropriate, and that's the topic of the next chapter.

Key Terms

- **Running variable:** A variable for which units' treatment status is determined by whether their value of that variable is on one or the other side of some threshold.
- **Regression discontinuity (RD) design:** A research design for estimating a causal effect that estimates the discontinuous jump in an outcome on either side of a threshold that determines treatment assignment.
- **Continuity at the threshold:** The requirement that average potential outcomes do not change discontinuously at the threshold that determines treatment assignment. If continuity at the threshold doesn't hold, then a regression discontinuity design does not provide an unbiased estimate of the local average treatment effect.
- **Sharp RD:** An RD design in which treatment assignment is fully determined by which side of the threshold the running variable is on.
- **Fuzzy RD:** A research design that combines RD and IV. The fuzzy RD is used when treatment assignment is only partially determined by which side of the threshold the running variable is on. The researcher, therefore, uses which side of the threshold the running variable is on as an instrument for treatment assignment. In this setting, continuity at the threshold guarantees that the exogeneity assumption of IV is satisfied. But we still have to worry about the exclusion restriction and the other IV assumptions.

Exercises

12.1 The state of Alaska asks you to estimate the effect of their new automatic voter registration policy on voter turnout. The policy was first implemented in 2017, but they report to you that, unfortunately, they initially didn't have the resources to roll the policy out to everyone in the state. As a result, they initially just applied automatic registration to people who had moved to Alaska

within two years of the date of the policy being implemented, but they haven't yet applied it to people who moved to Alaska before then. They're worried that this might be a limitation for your study, and they apologize that they weren't able to implement the policy for everyone, but they're still hoping that you can help. How would you respond, and how might you go about estimating the effect of automatic voter registration in Alaska?

12.2 The U.S. federal government subsidizes college education for students through Pell Grants. An individual is eligible for a Pell Grant if their family income is less than $50,000 per year.

(a) How could you potentially use this information and implement an RD design to estimate the effect of college attendance on future earnings?
(b) Would this be a sharp or a fuzzy RD design?
(c) What data would you want to have at your disposal?
(d) What is the running variable?
(e) What's the treatment?
(f) What's the instrument (if any)?
(g) What's the outcome?
(h) What assumptions would you have to make in order to obtain credible estimates?

12.3 Download "ChicagoCrimeTemperature2018.csv" and the associated "README.txt," which describes the variables in this data set, at press .princeton.edu/thinking-clearly. This is the same data on crime and temperature in Chicago across different days in 2018 that we examined in chapters 2 and 5. Imagine that the Chicago Police Department implemented a policy in 2018 whereby they stopped patrolling on days when the average temperature was going to be below 32 degrees (and suppose they have really good forecasts so they can very accurately predict, at the beginning of the day, the average temperature for that day). Their logic is that it's less pleasant for police officers to be out on the streets when it's cold, and there's less crime on cold days anyway. Use this (fake) information to estimate the effect of policing on crime.

(a) A helpful first step when implementing an RD design is to generate your own running variable where the threshold of interest is at 0. Rescale the temperature such that the threshold is at 0 by generating a new variable called "runningvariable," which is simply the temperature minus 32.
(b) We'll also need to generate our treatment variable. Generate a variable that takes a value of 1 if policing was in place on that day and 0 if it was not.
(c) It's often helpful to look at our data before conducting formal quantitative analyses. Make a scatter plot with crime on the vertical axis and temperature on the horizontal axis. Focus only on days when the temperature was within 10 degrees of the policy threshold, and draw a line at the threshold. Visually, does it look like there is a discontinuity at the threshold?

(d) There are several different ways to formally implement an RD design. The simplest is to focus on a narrow window around the threshold and simply compare the average outcome on either side. Focusing only on days when the temperature was within 1 degree of the threshold, compute the average number of crimes just above and just below, and compute the difference. Notice that you can (if you'd like) do this in one step with a regression.
(e) What concerns would you have with the naive approach above? Think about the trade-offs you face as you're deciding which bandwidth to select. How does your estimate change if you use a bandwidth of 10 degrees instead of 1 degree? Why?
(f) Another strategy is to use the local linear approach. For days that were less than 5 degrees below the threshold, regress crime on the running variable and compute the predicted value at the threshold. (Hint: Because you rescaled your running variable, this should be given by the intercept.) Do the same thing for days that were less than 5 degrees above the threshold. Compare those two predicted values. (Note that this can also be done with a single regression as described in the text.)
(g) What benefits does this local linear approach have over the naive approach?
(h) You might also consider allowing for a non-linear relationship between the running variable and the outcome. Generate new variables corresponding to the running variable squared and the running variable to the third power. Regress crime on policing, the running variable, the running variable squared, and the running variable to the third power. Only include observations within 10 degrees of the threshold. Interpret the estimated coefficient associated with policing.
(i) What are the pros and cons of this polynomial approach relative to the previous approaches?

Readings and References

The study on corporate returns to campaign contributions is

> Anthony Fowler, Haritz Garro, and Jorg L. Spenkuch. 2015. "Quid Pro Quo? Corporate Returns to Campaign Contributions." *Journal of Politics* 82(3):844–58.

For a discussion of potential violations of continuity in studies of policy changes at population thresholds see

> Andrew C. Eggers, Ronny Freier, Veronica Grembi, and Tommaso Nannicini. 2018. "Regression Discontinuity Designs Based on Population Thresholds: Pitfalls and Solutions." *American Journal of Political Science* 62(1):210–29.

The study on the effects of electing an extremist versus a moderate in a primary election is

Andrew B. Hall. 2015. "What Happens When Extremists Win Primaries?" *American Political Science Review* 109(1):18–42.

The studies on the validity of electoral regression discontinuity designs are

Devin Caughey and Jasjeet S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942–2008." *Political Analysis* 19(4):
385–408.

Andrew C. Eggers, Anthony Fowler, Andrew B. Hall, Jens Hainmueller, and James M. Snyder, Jr. 2015. "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races." *American Journal of Political Science* 59(1):259–74.

The study on U.S. bombing during the Vietnam War is

Melissa Dell and Pablo Querubin. 2018. "Nation Building through Foreign Intervention: Evidence from Discontinuities in Military Strategies." *Quarterly Journal of Economics* 133(2):701–64.

The study on the effect of being behind at halftime in basketball is

Jonah Berger and Devin Pope. 2011. "Can Losing Lead to Winning?" *Management Science* 57(5):817–27.

CHAPTER 13

Difference-in-Differences Designs

What You'll Learn

- Another situation that potentially allows us to estimate causal effects in an unbiased way is when a treatment changes at different times for different units. Here a difference-in-differences design may be appropriate.
- Difference-in-differences designs effectively control for all confounders that don't vary over time, even if they can't be observed or measured.
- Difference-in-differences designs can often be useful as a gut check, a simple way to probe how convincing the evidence for some causal claim really is.

Introduction

Regression discontinuity isn't the only creative research design that lets us get at causality in the absence of an experiment. When some units change treatment status over time but others don't, we may be able to learn about causal relationships using a strategy called difference-in-differences.

The basic idea is pretty simple. Suppose we want to know the effect of a policy. We can find states (or countries or cities or individuals or whatever the relevant unit of observation is) that switched their policies and measure trends in the outcome of interest before and after the policy change. Of course, we may worry that outcomes are systematically changing over time for other reasons. But we can account for that by comparing the change in outcomes for states that changed policy to the change in outcomes for states that did not change policy. If the trends in outcomes for states that did and did not change policy would have been the same if not for the policy change in some states, then we can use the states that did not change policy as a baseline of comparison, to account for the over-time trends. Our estimate of the causal effect of the policy change, then, will come from any change in outcomes in states that did change policy over and above that baseline trend that we estimated from the states that didn't change policy. This is called a *difference-in-differences design* because we first get the differences (or changes) in outcomes over time for states that did and did not change policy. Then we compare the difference in those differences.

Like the regression discontinuity design, the power of the difference-in-differences approach is that it allows us to estimate causal effects even when we can't randomize the treatment or control for every possible confounder. But nothing is for free.

Difference-in-differences designs come with their own requirements. For regression discontinuity, we needed continuity at the threshold. For difference-in-differences, we need the condition we just described: that the trend in outcomes would have been the same on average across units but for the change in treatment that occurred in some units. This condition is often called *parallel trends.*

Parallel Trends

It's worth making sure we are thinking clearly about what the parallel trends requirement really means. As we've said, difference-in-differences estimates are unbiased so long as the trends in outcomes would have been parallel, on average, in the absence of any changes in treatment. In other words, the parallel trends requirement is really about potential outcomes. For a binary treatment, we can think about each unit's outcome with and without treatment in each of two time periods. To capture this idea, let's think about there being potential outcomes for each unit in each time period. We will refer to the two time periods as period I and period II. And let's think about our population being divided into two groups: a group that changes from untreated to treated between the two periods ($\mathcal{UT}$) and a group that remains untreated in both periods ($\mathcal{UU}$).

Label the average potential outcome in group $\mathcal{G}$ in period p under treatment status T as

$$\overline{Y}^{p}_{T,\mathcal{G}}.$$

We observe the outcomes for a sample of the members of each group in each period. Let's start with the group that never changes treatment status ($\mathcal{UU}$). If we just look at the average change in outcome between the two periods, it gives us an estimate of the difference in outcomes in the untreated condition between the two periods:

$$\text{DIFF}_{\mathcal{UU}} = \underbrace{\overline{Y}^{II}_{0,\mathcal{UU}} - \overline{Y}^{I}_{0,\mathcal{UU}}}_{\text{average untreated trend for } \mathcal{UU}} + \text{Noise}_{\mathcal{UU}}$$

The noise comes from the fact that we are looking at a sample.

And analogously for the group that changes treatment status ($\mathcal{UT}$), the average change in outcome between the two periods is

$$\text{DIFF}_{\mathcal{UT}} = \overline{Y}^{II}_{1,\mathcal{UT}} - \overline{Y}^{I}_{0,\mathcal{UT}} + \text{Noise}_{\mathcal{UT}}.$$

The difference-in-differences is, quite literally, the difference between these two differences:

$$\text{Difference-in-Differences} = \text{DIFF}_{\mathcal{UT}} - \text{DIFF}_{\mathcal{UU}}$$

To see where parallel trends comes in, we are going to cleverly rewrite $\text{DIFF}_{\mathcal{UT}}$ by adding and subtracting $\overline{Y}^{II}_{0,\mathcal{UT}}$ from it. You'll recall we did something similar back in chapter 9 in order to understand baseline differences. Just like then, while we know this seems kind of weird, we ask that you trust us for a minute. And, remember, at the very

least it should be clear that we aren't doing any harm since, by adding and subtracting the same term, we are really just adding zero. When we do that, we get

$$\begin{aligned}\text{DIFF}_{\mathcal{UT}} &= \overline{Y}^{II}_{1,\mathcal{UT}} - \overline{Y}^{I}_{0,\mathcal{UT}} \\ &= \underbrace{\left(\overline{Y}^{II}_{1,\mathcal{UT}} - \overline{Y}^{II}_{0,\mathcal{UT}}\right)}_{\text{average treatment effect for } \mathcal{UT} \text{ in } II} + \underbrace{\left(\overline{Y}^{II}_{0,\mathcal{UT}} - \overline{Y}^{I}_{0,\mathcal{UT}}\right)}_{\text{average untreated trend for } \mathcal{UT}} + \text{Noise}_{\mathcal{UT}}.\end{aligned}$$

Once again, our algebra trick was actually pretty cool. We can now see that the over-time difference for group $\mathcal{UT}$ is made up of three things, which correspond to our favorite equation. First, there is the average (period *II*) treatment effect for group $\mathcal{UT}$. In chapter 9 we learned that this was called the ATT, the average treatment effect on the treated units. We can think of this as our estimand. Second, there is the trend in outcomes that would have happened for group $\mathcal{UT}$ even if they had remained untreated. We can think of this as a source of bias that comes from just looking at what happens in the $\mathcal{UT}$ group before and after treatment. And third, there is noise, as always.

With this in hand, we can now rewrite the difference-in-differences in terms of the ATT, the average untreated trends in potential outcomes for both groups, and noise. This will make clear what we are really doing—using the over-time trend in the $\mathcal{UU}$ group to try to eliminate the bias that comes from just looking at $\text{DIFF}_{\mathcal{UT}}$:

$$\begin{aligned}\text{Difference-in-Differences} &= DIFF_{\mathcal{UT}} - DIFF_{\mathcal{UU}} \\ &= \underbrace{\overline{Y}^{II}_{1,\mathcal{UT}} - \overline{Y}^{II}_{0,\mathcal{UT}}}_{ATT} + \underbrace{\underbrace{\overline{Y}^{II}_{0,\mathcal{UT}} - \overline{Y}^{I}_{0,\mathcal{UT}}}_{\text{average untreated trend for } \mathcal{UT}} - \underbrace{\overline{Y}^{II}_{0,\mathcal{UU}} - \overline{Y}^{I}_{0,\mathcal{UU}}}_{\text{average untreated trend for } \mathcal{UU}}}_{\text{difference in average trends}} \\ &\quad + \underbrace{\text{Noise}_{\mathcal{UT}} - \text{Noise}_{\mathcal{UU}}}_{\text{Noise}}\end{aligned}$$

Now we can see what parallel trends really means in terms of potential outcomes and our favorite equation. The difference-in-differences equals the ATT (estimand) plus the difference between the average untreated trends for group $\mathcal{UT}$ and group $\mathcal{UU}$ (bias) plus noise. So when does the difference-in-differences give us an unbiased estimate of the ATT? When the untreated trend is the same for both groups, so that the difference in average trends equals zero.

This is what parallel trends means. The change in average outcome would have been the same in the treated and untreated groups had everyone remained untreated. When this is the case, by subtracting $\text{DIFF}_{\mathcal{UU}}$ from $\text{DIFF}_{\mathcal{UT}}$ we eliminate the over-time trend, leaving an unbiased estimate of the average treatment effect (in period *II*) for units that switched treatment status.

Notice, this notation highlights another subtle point. Difference-in-differences does not quite estimate the ATE. It estimates the average treatment effect for those units who actually change treatment status—that is, the ATT. Whether or not this is a good estimate of the ATE depends on whether treatment effects differ systematically across the units that do and don't switch treatment status. But, in any event, this is a genuine causal effect and, at least for some applications, may in fact be the quantity of interest.

Table 13.1. Fast-food employment in New Jersey and Pennsylvania in 1992.

	January 1992 *NJ and PA* *low minimum wage*	**November 1992** *NJ high minimum wage* *PA low minimum wage*
New Jersey	20.44	21.03
Pennsylvania	23.33	21.17

Two Units and Two Periods

So far, we've been a bit abstract. Let's talk about a concrete example from classic work by David Card and Alan Krueger on the effect of the minimum wage on employment. This example is nice because it shows how difference-in-differences works in its most simple form. There are only two units, two periods, and one change in treatment status for one of the units.

Unemployment and the Minimum Wage

Card and Krueger wanted to know whether a higher minimum wage increased unemployment. Their idea was to exploit the fact that New Jersey raised its minimum wage in early 1992, while Pennsylvania, which borders New Jersey, did not. They collected data on the average number of full-time equivalent employees (FTE) per fast-food restaurant (which tend to pay minimum wage) in both New Jersey and Pennsylvania in January 1992 (before New Jersey raised its minimum wage) and in November 1992 (after New Jersey raised its minimum wage). Their data is summarized in table 13.1.

A first comparison we might think to make to learn about the effect of the minimum wage on employment is the difference between the employment levels in New Jersey and in Pennsylvania in November 1992. After all, by November, New Jersey had a higher minimum wage than Pennsylvania. That comparison shows that Pennsylvania fast-food restaurants employed only 0.14 more people, on average, than New Jersey restaurants, suggesting that a higher minimum wage may have almost no impact on employment.

But that comparison is not apples-to-apples, so we cannot interpret the difference as the effect of raising the minimum wage. New Jersey and Pennsylvania might differ in all sorts of ways that matter for employment besides the minimum wage. For instance, perhaps those two states have different levels of economic prosperity, different tax systems, or differently sized fast-food restaurants. And since, in this comparison, the state and the treatment are perfectly correlated, any such difference between New Jersey and Pennsylvania can be thought of as a confounder.

Another comparison we might make is to look at the change in employment in New Jersey between January and November, since the New Jersey minimum wage changed between these two months. This comparison shows an increase in employment of 0.59 employees per restaurant, suggesting that perhaps raising the minimum wage slightly increased employment. This approach has the advantage of comparing one state to itself, so we no longer need to worry about any cross-state differences. But now we have a new concern. Maybe January and November differ in terms of fast-food employment

Table 13.2. Two comparisons that do not unbiasedly estimate the causal effect of minimum wage increase.

	January 1992 *NJ and PA* *low minimum wage*	**November 1992** *NJ high minimum wage* *PA low minimum wage*	**Difference** *November–January*
NJ	20.44	21.03	0.59 *effect of high minimum wage* *+ over-time trend* *+ noise*
PA	23.33	21.17	
Difference *NJ – PA*		−0.14 *effect of high minimum wage* *+ differences between states* *+ noise*	

for other reasons—for example, because of seasonality or overall changes to the economy over the course of the year. Any such time trends would be a confounder in this comparison. So this comparison also isn't apples-to-apples.

Table 13.2 shows the two differences we've discussed and lays out, in the terms of our favorite equation, why neither gets us an unbiased estimate of the effect of the minimum wage. The difference between employment in November and January in New Jersey is the sum of the effect of the higher minimum wage (estimand), the over-time trend (bias), and noise. The difference between employment in New Jersey and Pennsylvania in November is the sum of the effect of the higher minimum wage (estimand), differences between the states (bias), and noise. So both differences are biased.

But we can do better. Start by thinking about the comparison between New Jersey and Pennsylvania in November. The problem with that comparison is that it reflects both the effect of the higher minimum wage (the estimand) and any systematic differences between New Jersey and Pennsylvania (the bias), plus, as always, noise. But suppose the differences between New Jersey and Pennsylvania aren't changing over time. Then the difference in employment in New Jersey and Pennsylvania in January, when they both have a lower minimum wage, reflects those same across-state differences, but without the effect of the higher minimum wage that New Jersey adopted later in the year. So we can use that employment difference in January to estimate the underlying differences between the two states. And then, subtracting the January difference from the November difference (i.e., finding the difference-in-differences) will leave us with an unbiased estimate of the effect of the higher minimum wage. (Of course, there is different noise in each comparison, so the noise terms don't just cancel.)

The same procedure works if we start from our comparison of New Jersey in November to New Jersey in January. The problem with that comparison is that it reflects both the effect of the higher minimum wage and any other differences between November and January that matter for employment (plus noise). But suppose those over-time

Table 13.3. Difference-in-differences estimate of the effect of minimum wage on fast-food employment.

	January 1992 *NJ and PA* *low minimum wage*	**November 1992** *NJ high minimum wage* *PA low minimum wage*	**Difference** *November–January*
NJ	20.44	21.03	0.59 *effect of high minimum wage* *+ over-time trend* *+ noise*
PA	23.33	21.17	−2.16 *over-time trend* *+ noise*
Difference *NJ − PA*	−2.89 *differences between states* *+ noise*	−0.14 *effect of high minimum wage* *+ differences between states* *+ noise*	**Difference-in-Differences** 0.59 − (−2.16) = −0.14 − (−2.89) = 2.75 *effect of high minimum wage* *+ noise*

trends are the same in New Jersey and Pennsylvania. Then the difference in employment in Pennsylvania between November and January is an estimate of the over-time trend, without any effect of the minimum wage (since Pennsylvania didn't change its minimum wage in 1992). So subtracting the change in employment in Pennsylvania from the change in employment in New Jersey will also leave us with an unbiased estimate of the effect of the higher minimum wage.

As shown in table 13.3, either way we do this calculation, we find the same answer. Surprisingly, the estimate that this procedure leaves us with is that a higher minimum wage appears to increase employment by 2.75 FTE per restaurant. The key is that the Pennsylvania data suggests that there was a big baseline drop in employment from January to November 1992. So the 0.59 FTE increase in NJ was a misleading under-estimate of the true effect being masked by an over-time trend.

Importantly, by calculating the difference-in-differences, we were able to account for systematic differences between the states and this over-time trend, without ever observing what those differences or trends were. This is the power of the difference-in-differences approach.

Of course, this wasn't magic. As we've said, in order for this approach to be valid, we need the parallel trends condition—that the over-time trend in outcomes (and, thus, confounders) would have been the same across units but for the change in treatment status—to hold. But this is typically a less demanding assumption than assuming we've actually controlled for all possible confounders. For instance, in our example, we're not assuming that New Jersey and Pennsylvania are the same (or that we've directly controlled for any differences) absent any differences in minimum wage. We're also not assuming that there are no time trends. Instead, we're assuming that the trends are

parallel: whatever time trends affect employment do so in the same way in both New Jersey and Pennsylvania, at least in expectation.

Difference-in-differences has a lot going for it, and there are a lot of situations where we think this parallel trends condition is quite plausible. This design accounts for all differences between units that don't vary over time that would plague a comparison of the two units in just one time period. It also accounts for all of the time-specific factors that would plague a before-and-after analysis of any one unit. What it does not account for is time-varying differences between units. These are still a problem if they vary in ways that correspond with the treatment. For example, if New Jersey increased its minimum wage because they thought the economy was about to experience a boom relative to neighboring states, then this would be a violation of the parallel trends assumption.

Of course, even if the parallel trends assumption seems conceptually reasonable, just looking at two units is not particularly illuminating. Surely lots of idiosyncratic differences pop up in any two places in any two months, so the noise in the estimate is likely to be large. To do better, we need to extend the intuition we developed in this simple example to situations where we observe more than two units over more than two time periods.

N Units and Two Periods

To start extending our intuition, suppose there are lots of units (e.g., maybe we have data on employment and minimum wage from all fifty states) but still just two time periods. And suppose that some of the units never received the treatment while other units received the treatment in the second but not the first period. We still want to look at changes for units that experience a change in treatment and compare those to changes for units that did not experience a change in treatment. We have three different options for doing so, all of which are algebraically identical and will, thus, provide the same answer:

1. **By hand:** Just as we did in the example above, calculate the average outcome in each period separately for those that never received the treatment and those that got the treatment in the second period and calculate the difference-in-differences by hand.
2. **First differences:** Put the data into a spreadsheet with one row per unit (this is called *wide format*). Calculate the change in the outcome and the change in the treatment for each unit, and regress the former on the latter. The change in treatment will be 0 for the units that never change and 1 for units that do change. So we're just comparing the average change for these two groups.
3. **Fixed effects regression:** Put the data into a spreadsheet with one row per unit period (this is called *long format*). Regress the outcome on the treatment while also including dummy variables for each unit and time period. In this example, we would have a dummy variable that takes a value of 1 if the observation is in period II and 0 if the observation is in period I. We would also have separate dummy variables for each unit. So the dummy variable for unit i would takes the value 1 if the observation involved unit i and 0 if it involved a different unit (there would be one such dummy variable for each unit). We often call these dummy variables *fixed effects*. For instance, if an analyst says they included *state fixed effects* in a regression, they just mean that they included a separate dummy variable for each state. Including these fixed effects ensures that we're removing all average differences between units and all average differences over

time, and once we've done that, the coefficient associated with the treatment variable is just the difference-in-differences.

Let's look at a fun example with multiple units.

Is Watching TV Bad for Kids?

Matthew Gentzkow and Jesse Shapiro were interested in how watching television as a pre-schooler affects future academic performance. The problem, of course, is that how much television a kid watches is affected by all sorts of factors that also affect future school performance. So a simple comparison of TV watchers to non-TV watchers isn't apples-to-apples. To get at the causal relationship more credibly, they used variation in the timing with which TV originally became available in different locations in the United States. We are going to simplify what they did so you can see their basic idea.

Broadcast television first became available in most U.S. cities between the early 1940s and the early 1950s. Happily, in 1965, there was a major study of American schools (called the Coleman Study) that, among other things, recorded standardized test scores for over three hundred thousand 6th and 9th graders. A 9th grader in 1965 was in pre-school in approximately 1955. A 6th grader in 1965 was in pre-school in approximately 1958. Gentzkow and Shapiro use both the over-time rollout of TV and the Coleman data to learn about the effect of pre-school television watching on test scores.

Let's imagine that we have the Coleman data on test scores for the 6th and 9th graders in two types of towns. Towns in group A first got TV in 1953. So they had TV when both the 6th and 9th graders in the Coleman study were in pre-school. Towns in group B didn't get TV until 1956. So they had TV when the 6th graders were in pre-school but not when the 9th graders were. Overall, then, table 13.4 summarizes the way the observed data look.

If you want to learn about the effect of having access to TV as a pre-schooler on future academic achievement, a first comparison you might think to make is to compare the test scores of the 9th graders in the B towns (who couldn't watch TV in pre-school) to the test scores of the 9th graders in the A towns (who could watch TV in pre-school). You could do that by simply subtracting the average test score of a 9th grader in a B town from the average test score of a 9th grader in an A town.

But we already know lots of reasons why we cannot interpret that as an unbiased estimate of the causal effect of having access to TV in pre-school. These two types of towns might be different in all sorts of ways, besides when broadcast TV showed up, that matter for academic performance. For instance, maybe they have different average quality schools, different industries, or what have you. And since, in this example, the type of town and the treatment are perfectly correlated, any such difference between the towns is a confounder.

Another comparison we might make is to look at the difference in test scores in the B towns between the 9th graders and the 6th graders, since the 6th graders had access to TV in pre-school but the 9th graders did not.

This approach has the advantage of holding fixed the type of town, so we no longer need to worry about systematic cross-town differences. But now we have a new concern. Maybe the 9th-grade and 6th-grade cohorts differ in their test performance for other reasons—for example, because 9th graders are older, or because of cohort-specific differences. Any systematic over-time or cohort differences would be a confounder in this comparison.

Table 13.4. TV and test scores data structure.

	9th Graders in 1965 *pre-school in 1955*	**6th Graders in 1965** *pre-school in 1958*
A Towns *TV in 1953*	Avg Test Scores 9A	Avg Test Scores 6A
B Towns *TV in 1956*	Avg Test Scores 9B	Avg Test Scores 6B

Table 13.5. Two comparisons that do not result in unbiased estimates of the effect of TV.

	9th Graders in 1965 *pre-school in 1955*	**6th Graders in 1965** *pre-school in 1958*	**Difference**
A Towns *TV in 1953*	Avg Test Scores 9A	Avg Test Scores 6A	
B Towns *TV in 1956*	Avg Test Scores 9B	Avg Test Scores 6B	6B − 9B *effect of TV* + *cohort differences* + *noise*
Difference	9A − 9B *effect of TV* + *town differences* + *noise*		

Table 13.5 sums up the two ideas we've had thus far and explains why neither gets us a credible estimate of the true effect in terms of our favorite equation.

But, just as with the minimum wage example, we can do better. Start by thinking about the comparison between 9th graders in the two types of towns. The problem with that comparison is that it reflects both the effect of TV exposure and any other systematic differences between the types of towns. But suppose those baseline differences between the types of towns aren't changing over time. Then the difference in academic performance between the 6th graders in the two types of towns, all of whom had access to TV in pre-school, reflects those same cross-town differences, but without the effect of TV. So we can use that difference between the 6th graders to estimate the cross-town differences. And then, subtracting the difference between the 6th graders from the difference between the 9th graders (i.e., calculating the difference-in-differences) will leave us with just the effect of TV exposure in pre-school (plus noise).

The same procedure works if we start from our comparison of 9th graders and 6th graders from the B towns. The problem with that comparison is that it reflects both the effect of exposure to TV in pre-school and any baseline differences between the 6th- and 9th-grade cohorts that matter for academic performance (plus noise). But suppose those over-time or cohort trends are the same in the A towns and B towns. Then the

Table 13.6. How difference-in-differences might give an unbiased estimate of the effect of TV.

	9th Graders in 1965 *pre-school in 1955*	**6th Graders in 1965** *pre-school in 1958*	**Difference**
A Towns *TV in 1953*	Avg Test Scores 9A	Avg Test Scores 6A	6A − 9A *cohort differences* + *noise*
B Towns *TV in 1956*	Avg Test Scores 9B	Avg Test Scores 6B	6B − 9B *effect of TV* + *cohort differences* + *noise*
Difference	9A − 9B *effect of TV* + *town differences* + *noise*	6A − 6B *town differences* + *noise*	**Difference-in-Differences** (6B − 9B) − (6A − 9A) = (9A − 9B) − (6A − 6B) *effect of TV* + *noise*

difference in academic performance between 6th and 9th graders in A towns is an estimate of the over time or cohort trend without any effect of TV (since both sets of kids had access to TV in pre-school in Town A). So subtracting the difference in test scores in the A towns from the difference in test scores in the B towns will again leave us with an unbiased estimate of the effect of pre-school TV exposure.

As shown in table 13.6, either way we do this calculation, we find the same answer.

For those interested in the answer, Gentzkow and Shapiro find evidence that, during the 1950s, having access to TV in pre-school was actually beneficial for average test scores, especially for kids from poorer families. Of course, this was at a time when kids watched shows like *Howdy Doody*. So you might not want to immediately extrapolate to the present day.

More important, for our purposes, is seeing the power of the difference-in-differences approach. By calculating the difference-in-differences, we were able to account for systematic differences between towns and over time (or cohorts), without ever observing what those differences or trends were.

N Units and *N* Periods

Suppose you have more than two periods and suppose that the treatment is changing at different times for different units. What do you do?

Much of the logic from the above discussion still applies. Of course, option 1 above (calculating the difference-in-differences by hand) no longer works. But you can still use option 2 (first differences) or option 3 (fixed effects). However, first differences and fixed effects are no longer mathematically identical and will not necessarily give you the same answers once you move beyond two periods. What's the difference? With first differences, you're regressing period-to-period changes in the outcome on period-to-period changes in the treatment. With fixed effects, you're regressing the outcome on the treatment while controlling for all fixed characteristics of units and time

periods. Both are doing the same basic thing, but they are using slightly different kinds of variation.

Which specification makes more sense depends on the specific context. In general, the fixed effects strategy is more flexible. For instance, it allows you to include additional time-varying control variables in the regression (if necessary), and it also allows you to conduct some helpful diagnostics. Importantly, in both cases, the timing of the effect matters for exactly what you are estimating. In the case of first differences, you are looking for effects that happen immediately after the treatment status changes. If it takes some time for the effect of the treatment to set in, or if the effect size decays or grows over time, you can get misleading estimates. However, complications in the timing of treatment also create complications for interpreting exactly what is being estimated when you use a fixed effects specification. We aren't going to go into these issues in any detail because they are actually the topic of cutting-edge research as of the writing of this book. However, if you go on to do quantitative analysis involving difference-in-differences, you may want to delve more deeply into these questions. We suggest some readings at the end of the chapter.

Even though there can be some complicated technical details, the intuition of difference-in-differences designs should be clear from our examples. And it is an important intuition. If someone shows you that some treatment of interest is correlated with an outcome of interest, you are already skeptical because of what we learned in chapter 9. Difference-in-differences allows you to check whether changes in the treatment are also correlated with changes in the outcome. If they are, then that might be more compelling evidence of a causal relationship. And if they aren't, then the original correlation may have been the result of confounding.

Let's look at an example of a study that uses a fixed effects approach to implementing a difference-in-differences design when there are multiple units changing treatment status at different times.

Contraception and the Gender-Wage Gap

The availability of oral contraceptives, starting in the 1960s, gave women unprecedented control over their reproductive and economic decisions. Understanding the impact of this contraceptive revolution on women's lives is important for understanding the evolution of the modern economy and society.

Of course, if we want to estimate the effects of oral contraception on women's child birth decisions, labor market participation, or wages, we can't simply compare outcomes for women who did and did not use oral contraception. After all, access to health care is affected by things like wealth, education, geography, race, and so on. So such comparisons are sure to be confounded. And no one ran an experiment giving some women access to oral contraceptives while restricting access to others. But this doesn't mean we can't make progress on these causal questions.

In an important paper, Claudia Goldin and Lawrence Katz point out that state policies created a kind of natural experiment. Oral contraceptives first became available in the United States in the late 1950s. However, the legal availability of oral contraceptives to younger women differed across states. In a few states, laws prevented the sale of contraception to unmarried women, and in most states, women under the age of majority needed parental consent before obtaining contraception. Over time, courts and state legislatures gradually removed these restrictions and lowered the age of

majority. Helpfully, for the purposes of causal inference, they moved to do so at different times.

This meant that in the earliest moving states of Alaska and Arkansas, an unmarried, childless woman under the age of twenty-one could obtain oral contraception by 1960. In the latest moving state of Missouri, this wasn't possible until 1976. And for the other states, it was somewhere in between. This is important because women under twenty-one make particularly consequential decisions about when to have children, when to get married, whether to pursue higher education, and so on.

In another influential paper, Martha Bailey uses this variation to implement a difference-in-differences design to estimate the effect of early access to oral contraceptives on when women first have children and whether and to what extent they entered the paid labor force.

The basic idea is straightforward. Imagine four groups of women across two states, Kansas (which allowed younger women access to oral contraceptives in 1970) and Iowa (which didn't allow access until 1973). There are women who were aged eighteen to twenty in the late 1960s in both states; neither of these groups had access to oral contraceptives. And there are women who were aged eighteen to twenty in the early 1970s in each state; the women in Kansas had access to oral contraceptives, while the women in Iowa did not. Thus, we can use the changes in outcomes for the women in Iowa as a baseline of comparison for the changes in outcomes for the women in Kansas to try to estimate the effect of early access to oral contraception for women in Kansas.

Bailey can do better than this simple example, since she has data for women from many age cohorts for all fifty states, and different states changed policy at different times. So she makes use of a fixed effects setup—regressing her outcome measures on a dummy variable for whether a given cohort of women had access to oral contraception when they were aged eighteen to twenty, as well as state fixed effects and cohort fixed effects. This allows her to implement a difference-in-differences design with many units changing treatment status at different times.

Since it's not random which states allowed early access to oral contraceptives first, we should think about parallel trends. Is it reasonable to assume that the trends in childbearing and labor market participation are parallel, on average, across states, and that states did not strategically shift contraceptive rules just as they otherwise expected these outcomes to shift for other reasons? Bailey provides some reasons to think the answer is yes. For instance, she shows that the timing of legal access to contraceptives for younger women is uncorrelated with a wide variety of state characteristics in 1960 that you might expect to influence these outcomes. These include geography, racial composition, average marriage ages, women's education, fertility, poverty, religious composition, unemployment for men or women, wages for men and women, and so on.

Bailey's difference-in-differences results suggest that access to oral contraception at an age when women are making consequential life decisions does in fact have important effects. In particular, she estimates that access to oral contraceptives before age twenty-one reduced the likelihood of becoming a mother before age twenty-two by 14 to 18 percent and increased the likelihood that a woman was participating in the paid labor force in her late twenties by 8 percent. Moreover, women who had access to oral contraception before the age of twenty-one worked about seventy more hours per year in their late twenties. That is, by providing a way to delay and plan childbearing, oral contraception appears to have given women the freedom to pursue longer-term careers and work more.

Useful Diagnostics

As we've said, for difference-in-differences to yield an unbiased estimate of an average treatment effect, we need parallel trends. That is, in the counterfactual world where the treatment did not change, the difference in average outcomes would have stayed the same between the units where the treatment did in fact change and the units where it did not. Since we don't observe that counterfactual world, we can't know if that's true. So a careful analyst always wants to do whatever is possible to probe the plausibility of parallel trends.

One conjecture is that if parallel trends holds, we should see similar trends in outcomes in earlier periods, before any units changed treatment status. We can check these pre-treatment trends (often called *pre-trends*) directly by comparing the trend in outcomes for units that do and do not change treatment status later on. We can also do this in a regression framework by including a *lead treatment* variable—that is, a dummy variable indicating the treatment status in the *next* period. If the trends are indeed parallel prior to the change in treatment, the coefficient on the lead treatment should be zero and the coefficient on the treatment variable should not change when we include that lead treatment variable in the regression.

We can also relax the requirement of parallel trends a bit by allowing for the possibility that different units follow different linear trends over time to see if this changes our results. The specific details for how you implement this are not important for now (you can read about them in a more advanced book). But you can see that there are various strategies for probing a difference-in-differences analysis to see whether parallel trends seem plausible.

Remember that diagnostic tests of this sort are a complement to, not a substitute for, clear thinking. The most important defense of an assumption like parallel trends must be a substantive argument. Why did the treatment change in some units and not in others? Does that reason seem likely to be related to trends in the outcome or independent of trends in the outcome? Can you think of reasons that units might have changed their treatment right as they expected the outcome to change for other reasons? These are critical questions whose answers require deep substantive knowledge of your context, question, and data. Good answers are absolutely essential to assessing how convincing the estimates that come out of a difference-in-differences are.

To get a better sense of how one thinks through questions about parallel trends, let's look at a couple examples.

Do Newspaper Endorsements Affect Voting Decisions?

Newspapers regularly endorse candidates for elected office. Do such endorsements matter?

A study by Jonathan Ladd and Gabriel Lenz attempted to answer that question using a difference-in-differences design with data from the United Kingdom. Their study provides a nice illustration of how to test for parallel pre-trends as a diagnostic for the plausibility of the parallel trends assumption.

During the 1997 general election campaign in the United Kingdom, several newspapers that historcially tended to endorse the Conservative Party unexpectedly endorsed the Labour Party. Ladd and Lenz utilize this rare shift to estimate the effect of newspaper endorsements on vote choice.

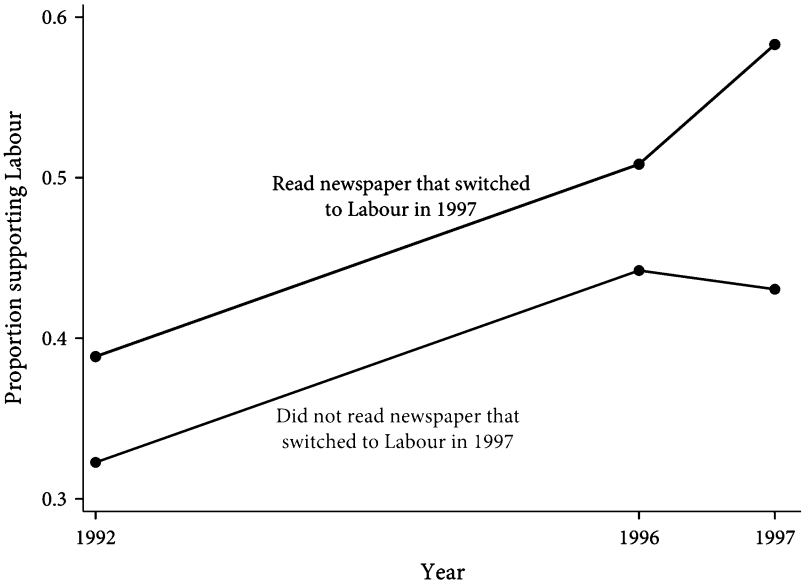

Figure 13.1. Visualizing pre-trends and using difference-in-differences to estimate the effect of newspaper endorsements on vote choices.

Implementing a difference-in-differences design, they compare changes in vote choice for those who regularly read a paper that unexpectedly switched its endorsement to Labour to changes in vote choice for those who did not regularly read one of those papers. Because they had data measuring partisan support of the same British individuals in 1992, 1996, and 1997, they were able to examine the pre-trends to see if they were parallel. If people who did and did not read the papers that switched to Labour between 1996 and 1997 were already trending differently between 1992 and 1996, that would make us worried that the parallel trends assumption is violated (and perhaps we'd worry the newspapers switched because their readers were trending toward Labour). But if these two groups were on similar trends between 1992 and 1996, that would give us more confidence that any resulting difference-in-differences is attributable to the unexpected newspaper endorsement in 1997.

Ladd and Lenz's diagnostics are reassuring, as shown in figure 13.1. People who read a paper that would later switch to endorsing the Labour Party had very similar trends in their level of Labour Party support between 1992 and 1996. But between 1996 and 1997, when the newspapers unexpectedly supported the Labour Party, the voters who read those papers significantly increased their support for the Labour Party relative to those who didn't read those papers. Thus, as long as we don't have reason to believe that other things, besides these surprise endorsements, changed differentially for readers of different newspapers in 1996, we might reasonably interpret the difference-in-differences as an estimate of the causal effect of those endorsements.

Is Obesity Contagious?

Humans are social animals. We live embedded in a complex web of relationships. Increasingly, we are told, our networks define who we are. A growing body of research

claims to measure exactly how our thinking, tastes, and behavior are determined by our social networks.

Perhaps the most well-known of this research is authored by Nicholas Christakis and James Fowler. What is striking about Christakis and Fowler's work is that they find that behaviors and characteristics that many of us think of as profoundly personal—smoking, drinking, happiness, obesity—all appear to be network characteristics. Or, to use their more colorful language, "Obesity is contagious."

In a study of the spread of obesity in social networks, published in the *New England Journal of Medicine*, Christakis and Fowler examine the relationship between a change in a person's weight and changes in their friends', family members', or neighbors' weight. They make these comparisons controlling for personal characteristics like age, gender, and education.

What do they find? The chance that a person becomes obese is 57 percent higher if that person has a friend who becomes obese than if that person does not have a friend who becomes obese. Friendship seems to matter more than familial ties when it comes to weight gain. If a person has a brother or sister who becomes obese, that person's chance of becoming obese increases by 40 percent. If a person's spouse becomes obese, that person's chance of becoming obese increases by 37 percent. Having obese neighbors has no effect. On the basis of these findings, the *New York Times* declared in a front-page article, "The way to avoid becoming fat is to avoid having fat friends." Christakis and Fowler didn't love this interpretation. Instead, the *Times* reported, Christakis suggested, "Why not make friends with a thin person... and let the thin person's behavior influence you and your obese friend?"

It seems indisputable that your behavior is affected by those with whom you interact, that their behavior is affected by those with whom they interact, and so on. In this sense, we are entirely with Christakis and Fowler—we are all influenced by our social networks. But these authors, and many other scientists who study network effects, are making a claim stronger than just the commonsensical observation that our interactions affect how we behave. They are claiming to measure and quantify that effect. How do they claim to do so?

Christakis and Fowler's approach is effectively a difference-in-differences design. They test how changes in one person's obesity correspond to changes in another person's obesity. So if we want to think clearly about whether these are credible estimates of a contagion effect, we need to think about whether we find the assumption of parallel trends plausible.

Recall what parallel trends says here. It requires that, in the counterfactual world where there was no change in treatment (i.e., no one's friends became more or less obese), the trend in outcomes (personal obesity) would have been the same on average among people who in fact experienced a change in treatment (i.e., whose friends' obesity changed) and people who did not experience a change in treatment (i.e., whose friends' obesity did not change). If parallel trends holds, then Christakis and Fowler's difference-in-differences yields an unbiased estimate of the effect of your friends' obesity on your obesity. But if the trends are not parallel, then their estimates are biased, since some of the difference-in-differences they are observing and attributing to network effects would have happened even if your friends' obesity hadn't changed.

One concern often raised about studies of network effects is what medical researchers call *homophily*. People with similar characteristics tend to group together. Suppose you find that people whose friends are smokers are more likely to be smokers themselves. Social network researchers might want to interpret this as evidence that having friends

who smoke causes you to become more likely to smoke. But, for that conclusion to be warranted, we'd have to be comparing apples to apples. That is, other than how their friends behave, people in the social networks of smokers and people in the social networks of non-smokers would have to be essentially the same. If members of networks of smokers were already more likely to be smokers than members of networks of non-smokers, we'd be comparing apples to oranges.

It seems entirely plausible (indeed likely) that people who are members of networks of smokers are, independent of their friends, already more likely to be the sort of people who smoke because of homophily. Smokers might well meet their friends in bars that allow smoking, in the outside area at work or school where people gather to smoke, or in other smoker-friendly environments. Put differently, it might not be that having friends that smoke causes you to smoke. It might be that being a smoker causes you to have friends that smoke. Because people don't choose their social networks randomly, when we compare people in smoking networks to people in non-smoking networks, we aren't comparing apples to apples.

But homophily, alone, is not enough to create a problem for Christakis and Fowler's difference-in-differences design. That design accounts for fixed characteristics of units, such as the possibility that obese people tend to be friends with each other and smokers tend to hang out together. This is one of the great things about difference-in-differences. Their finding is more compelling than just comparing people with more overweight friends to people with fewer overweight friends. They show that when one person *becomes* obese, their friends are also more likely to *become* obese. For homophily to create a problem, it has to be because of a worry about parallel trends not holding. For instance, if people who are on the path to becoming obese (perhaps because they have similar diets, exercise habits, genetic predispositions, cultural pressures, and so on) are more likely to be friends with each other, that would be a violation of parallel trends. And if parallel trends is violated, difference-in-differences doesn't yield an unbiased estimate of the causal effect.

We can't know whether homophily creates violations of parallel trends. But there is some evidence that points toward the possibility that difference-in-differences is not unbiased here. Ethan Cohen-Cole and Jason Fletcher conducted a study of the spread of two individual characteristics—height and acne—in social networks. Using the same difference-in-differences approach that Christakis and Fowler use to argue for the social contagion of divorce, loneliness, happiness, obesity, and many other things, Cohen-Cole and Fletcher find that both height and acne appear contagious in social networks. Knowing what we do about height and acne, it is pretty hard to believe that their spread is actually caused by social interactions within a network. This is Cohen-Cole and Fletcher's point. Height and acne likely don't spread in a social network. Instead, their apparent social contagion almost surely results from violations of parallel trends, perhaps due to homophily. Having friends with acne doesn't give you acne; people at high risk for acne tend to hang out together. The same may well be true for obesity, divorce, happiness, and so on.

To be clear, we're not saying that we think there are no causal network effects. Indeed, we're certain there are. Furthermore, Christakis and Fowler's study is surely more convincing because they compared changes to changes, rather than simply showing that obese people are more likely to be friends with each other. But there are lots of ways in which parallel trends could be violated. So we must be cautious and think clearly about those possibilities before interpreting the results of a difference-in-differences design as an unbiased estimate of the true causal effect.

Difference-in-Differences as Gut Check

Sometimes difference-in-differences analyses can be useful as a way to probe the credibility of a causal claim. Imagine a scenario in which someone estimates the correlation between a treatment and an outcome, perhaps even controlling for some possible confounders. You, thinking clearly about the lessons of chapter 9, might be skeptical that a causal interpretation of this estimate is warranted. Maybe you can think of a bunch of other confounders that aren't observable and, so, can't be controlled for. Even with such arguments, it can be hard to convince people to take your concerns seriously.

But if the data include multiple observations of the same unit, difference-in-differences can provide a useful gut check.[1] If the treatment really has an effect on the outcome, then we should expect a correlation not just between treatment and outcome but between changes in treatment status and changes in outcome. That is, we should expect the relationship between treatment and outcome to still be there in a difference-in-differences analysis.

Even if you find a relationship in the difference-in-differences, you still might not be sure about the causal interpretation. For that, you'd want to think about parallel trends. But if the relationship disappears in the difference-in-differences, then you have bolstered the case for your skepticism. It would seem that differences between units other than the treatment account for the correlation in the data. To see how difference-in-differences can be used for a gut check, let's look at an example.

The Democratic Peace

At least since the philosopher Immanuel Kant wrote *Perpetual Peace*, theorists have argued that democracy leads to peace—or, in its more contemporary formulation, that democracies will be more reluctant to fight one another than they are to fight autocracies or than autocracies are to fight one another. Some argue that this is because democracies share common norms that prevent them from engaging in violence against one another. Others argue that various features of domestic politics constrain democratic leaders from waging war against other democrats.

Empirical scholars have been similarly fascinated by the relationship between democracy and war. And the finding that country pairs (called *dyads*) where both countries are democratic are less likely to fight wars with one another than are dyads where at least one country is not democratic is one of the most important and discussed empirical findings in the literature on international relations.

Let's think a little about that empirical literature and its findings. A first thing scholars have done to try to assess the democratic peace is to simply look at the correlation between democracy and war. We'll start by replicating that approach. Here's how.

We start with a big data set that has an observation for every dyad in every year. So an observation is a dyad-year. We are going to work with data from 1951–1992 because those are the years one of the most famous papers in this literature works with. For each dyad-year, we have a binary variable that indicates whether that dyad had a militarized interstate dispute (MID) in that year. That is our dependent variable. And for each country we have a measure of how democratic it is. We use the Polity score, which you may recall from chapter 2 is a standard measure of the level of democracy. Higher numbers indicate a more democratic country. For estimating the democratic peace, we

[1] It's a gut check because your newly honed clear thinking skills are telling you to always be a bit skeptical.

Table 13.7. The relationship between democratic dyads and war with and without controls and with and without year and dyad fixed effects.

	1	2	3	4
	Dependent Variable = MIDs			
Minimum Level of Democracy in Dyad	−.0082**	−.0066**	.0002	.0005
	(.0016)	(.0016)	(.0017)	(.0017)
Countries Are Contiguous		.0693**	.0002	.0648**
		(.0110)	(.0017)	(.0227)
Log (Capability Ratio)		.0006		.0024
		(.0005)		(.0019)
Minimum 3-Year GDP Growth Rate		−.0001		−.0005**
		(.0002)		(.0002)
Formal Alliance		−.0012		−.0095
		(.0027)		(.0067)
Minimum Trade-GDP Ratio		−.0045**		.0011
		(.0017)		(.0021)
Includes Year Fixed Effects			✓	✓
Includes Dyad Fixed Effects			✓	✓
Observations	93,755	93,755	93,755	93,755
r-squared	.0011	.0289	.2636	.2658

Standard errors are in parentheses. ** indicates statistical significance with $p < .01$.

don't want to know how democratic any one country is. We want to know whether a dyad contains two democracies in a given year. To get at this, we use the lower of the two Polity scores within each dyad. If both countries in a dyad are democratic, then the lower of the two scores will be high. If at least one country in a dyad is not democratic, then the lower of the two scores will be low. We put this variable on a scale from 0 to 1 so we can interpret the coefficients as the estimated effect of going from the lowest to the highest level of democracy. This measure, which we refer to as the *minimal level of democracy in a dyad*, is our treatment variable.

To see the correlation between war and democracy, we regress MIDs on the minimal level of democracy. Figuring out the correct standard errors in this regression is actually a bit tricky, since surely there is correlation between whether, say, France and Germany have a war in a given year and whether England and Germany have a war in that same year. But we aren't going to worry about those issues for the moment.

The first column of table 13.7 shows the results of this regression. We find a statistically significant negative correlation between being a democratic dyad and war. The regression coefficient of −.0082 says that if we compare a dyad where the less democratic country is among the least democratic countries to a dyad where both countries are among the most democratic countries, the probability of there being a war between the two countries in a given year is about eight-tenths of a percentage point lower. Since the overall probability that any given dyad is at war in any given year is only about eight-tenths of a percent to start with, that is an enormous estimated relationship.

Now, we hope that this evidence doesn't convince you there is a causal effect of democracy on war. The lessons about confounders from chapter 9 are still important. And we can think of lots of ways that democracies and autocracies are different that might matter for war.

Scholars are aware of this concern. And the standard approach to addressing it is to try to control for various characteristics of a dyad that correlate with being democratic and with war. For instance, studies commonly control for whether the countries are contiguous, their relative military capabilities, their GDP growth, whether countries are allied, how much countries trade, and so on. Of course, we also shouldn't forget the lessons of chapter 10. Some of these things may be mechanisms by which democracy affects war, rather than confounders, in which case they shouldn't be controlled for. But, to stick close to the literature, in the second column of table 13.7, we control for these variables. As you can see, once we control, the estimated relationship between a democratic dyad and war drops a little bit. But it is still strongly negative and statistically significant.

At this point, many scholars conclude that Kant and other theorists are on to something. There really is a causal effect of being a democratic dyad on going to war. That might be true, but we are certainly entitled to remain skeptical. After all, there are so many features of a dyad that are hard to measure. And any number of them might affect both whether the two countries are democracies and whether they go to war. Indeed, a study by Henry Farber and Joanne Gowa claims that the empirical pattern associated with the democratic peace does not appear in the data prior to World War II precisely because key confounding variables took different values during this earlier period.

Controversies like this are where difference-in-differences can help us. If the theories of the democratic peace are right, then we shouldn't just observe a negative correlation between being a democratic dyad and war. We should observe a change in the likelihood two countries go to war as the dyad becomes more jointly democratic. That is, we should continue to see the correlation we've already observed in a difference-in-differences analysis. If we don't, we have reason to worry about bias—that is, that the estimated correlation reflects the influence of unobserved confounders rather than a true causal effect.

This argument was made in an influential, and controversial, paper by Donald Green, Soo Yeon Kim, and David Yoon. And so, in columns 3 and 4 of table 13.7 we implement a difference-in-differences design for the case of N observations and N time periods. We do so using fixed effect regression, including fixed effects for each dyad and for each year. Column 3 reports the difference-in-differences with no other control variables. Column 4 includes the fixed effects and the controls.

As you can see, once we compare the change in war to the change in whether a dyad is democratic, the correlation disappears. The difference-in-differences finds no meaningful or statistically significant relationship between democracy and war. Our gut check failed. As we've emphasized, this doesn't mean that there is definitely no causal effect. But it does mean that the existing evidence does not make a compelling case for one. By simply checking the difference-in-differences, we come away with a very different picture from the one painted by the simple correlations.

Many scholars who believe in the democratic peace have criticized Green, Kim, and Yoon's argument and the use of difference-in-differences designs to answer questions about international relations. One common critique is that difference-in-differences

ignores most of the variation in the treatment variable, making it hard to find evidence of a relationship.

This is true. The regressions in columns 1 and 2 of table 13.7 make use of a lot of variation in democracy to try to detect a relationship between democracy and war—they use variation over time, variation between dyads, and variation within dyads. The regressions in columns 3 and 4 just use the variation within dyads, holding constant differences between dyads and global changes over time. But some of the variation exploited in columns 1 and 2 is probably not very informative about the causal effect of democracy because there are so many other things that are changing over time and that differ between dyads. So yes, difference-in-differences ignores a lot of the variation and attempts to isolate the variation that is most informative for assessing the effect of democracy on war—namely, the within-dyad variation.

It's also worth noting that this critique would have more bite if the difference-in-differences estimates were far less precise than the other estimates. This would indicate that there is a lot less information about the relationship between democracy and peace in the difference-in-difference estimates. But the estimated standard errors on the minimum level of democracy variable in table 13.7 are only slightly larger in columns 3 and 4 than in columns 1 and 2. It's not as if, in doing the difference-in-differences, we threw up our hands and concluded that we just don't know anything about the relationship between democracy and war. The difference-in-differences design allows us to obtain reasonably precise estimates of the effect of democracy. And those estimates are very close to zero. Furthermore, the difference-in-differences estimates are statistically significantly different from the estimates in columns 1 and 2. So imprecision does not account for the disparate results obtained by these two approaches.

Wrapping Up

We've seen that changes in treatment over time can allow us to more credibly estimate the effects of that treatment using a difference-in-differences design. For this to work we need for the parallel trends condition to hold—it has to be that, had it not been for the change in treatment status, the average outcomes for units that did and did not change treatments would have followed the same trend. There are several useful diagnostic tests to help analysts assess whether this assumption is plausible, but there is no substitute for clear thinking and substantive knowledge.

The last four chapters have been dedicated to methods for obtaining more credible estimates of causal relationships. Estimating causal relationships is a difficult and noble task. But often we want to know more. We aren't satisfied just knowing that the treatment did have an effect. We want to know *why*. The next chapter addresses the important challenge of answering such *why* questions using quantitative evidence.

Key Terms

- **Difference-in-differences:** A research design for estimating causal effects when some units change treatment status over time but others do not.
- **Parallel trends:** The condition that average potential outcomes without treatment follow the same trend in the units that do and do not change treatment status. This says that average outcomes would have followed the same trend had it not been for some unit's changing treatment status. If parallel trends

doesn't hold, difference-in-differences does not provide an unbiased estimate of the ATT.

- **First differences:** A statistical procedure for implementing difference-in-differences. It involves regressing the change in outcome for each unit on the change in treatment for each unit.
- **Wide format:** A way to structure a data set in which each unit is observed multiple times, where each row corresponds to a unique unit.
- **Long format:** A way to structure a data set in which each unit is observed multiple times, where there is a row for each unit in each time period.
- **Fixed effects regression:** A statistical procedure for implementing difference-in-differences. It involves regressing the outcome on the treatment while also including dummy variables (*fixed effects*) for each time period and for each unit.
- **Pre-trends:** The trend in average outcomes before any unit changes treatment status. If pre-trends are not parallel, it is harder to make the case that the parallel trends condition is plausible.
- **Lead treatment variable:** A dummy variable indicating that treatment status in a unit will change in the next time period.

Exercises

13.1 For years, the state of Illinois has administered the Illinois State Aptitude Test (ISAT) to third, fifth, and eighth graders. For much of this time, the test was relatively low stakes—not tied to promotion to the next grade, teacher compensation, school resources, and so on. The stakes changed in 2002, when the ISAT became the test that the Chicago Public Schools used to comply with the federal No Child Left Behind law.

Consider two cohorts of students: students who were fifth graders in 2001 and students who were fifth graders in 2002. Both of these groups of students took the ISAT in third grade when it was low stakes. The students who were in fifth grade in 2001 also took the ISAT in fifth grade when it was low stakes. But the students who were in fifth grade in 2002 took their second ISAT when it was high stakes. Make a two-by-two table showing how we could learn about the average effect of high-stakes testing on student test scores using a difference-in-differences design if we had data on the average test scores of these two cohorts of students when they were fifth and third graders.

13.2 The Nike Vaporfly shoe has been controversial in the world of elite long-distance running because some argue that the shoe provides an unfair advantage to those who use it, and it makes previous records obsolete. Suppose you had data from many different marathons that indicated each runner's time and also which shoes each runner wore. How could you estimate the effect of the Nike Vaporfly? You'd want to be sure to account for the fact that marathon times vary from day to day and course to course. You'd also want to account for the fact that some runners are just better and faster than others.

(a) What analyses would you conduct to separate the effect of shoe technology from other factors, and what assumptions would you have to make?

(b) Do you find those assumptions plausible? Discuss your potential concerns.
(c) Is there anything you can do to address these potential concerns?
(d) Another challenge is that not everyone who starts a marathon finishes it, so you could have attrition in your study. What could you do to address this potential problem?
(e) Could you use the same approach to estimate the effect of a new shoe or glove technology on points scored in professional boxing? Why or why not?

13.3 Suppose we want to estimate the extent to which the policy positions of Democratic and Republican candidates for Congress diverge. In other words, we'd like to know how differently the Democratic and Republican candidate would represent the same set of constituents.

(a) Suppose we measured how conservatively each member of Congress voted on bills and ran a regression of roll-call voting on an indicator for being a Republican. Would this be a satisfying way to estimate divergence? What kinds of bias would you worry about?
(b) Download "CongressionalData.csv" and the associated "README.txt," which describes the variables in this data set, at press.princeton.edu/thinking-clearly. This data set contains information on congressional elections and roll-call voting behavior. Using only the variables available in the provided data set, try to estimate divergence by controlling for confounders. If it helps, you may want to only analyze just one congressional session at a time.
(c) Using the data available, now estimate divergence using a regression discontinuity design. Again, you might find it helpful to focus on just one congressional session at a time.
(d) Finally, estimate divergence using a difference-in-differences design.
(e) Compare and contrast these three different approaches. Which one estimates divergence with the most defensible assumptions? How much do your estimates depend on your design?

13.4 In a study of sex-based discrimination in hiring, Claudia Goldin and Cecilia Rouse study the effect of making auditions for symphony orchestras "blind" by putting candidates behind a screen. The idea is, if the people evaluating the audition can't observe the sex of the person auditioning, they shouldn't be able to discriminate.

It turns out, as Goldin and Rouse document, that different orchestras adopted the practice of using such a screen at different times. Let's think about how we could use that fact to learn about the causal effect of the screens. (We'll talk through a somewhat different empirical approach than the one Goldin and Rouse use.)

(a) Suppose for each orchestra and each year you observed the share of new hires for that orchestra who were women and whether or not that orchestra used a screen in its audition. If you just pooled together all of your data and regressed share of women on using a screen, would you feel comfortable giving the output of that regression a causal interpretation. Why or why not?

(b) Suppose, instead, you wanted to use a difference-in-differences design with this data. What regression would you run?
(c) Describe the assumptions that would have to be true for this to give you an unbiased estimate of a causal effect. (Don't just say "parallel trends"; describe what would have to be true about the world for parallel trends to hold.)
(d) Does this assumption seem plausible to you? What kinds of concerns would you have?

Readings and References

The study on the effect of increasing the minimum wage in New Jersey is

David Card and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84(4):772–93.

The study on television and academic performance is

Matthew Gentzkow and Jesse M. Shapiro. 2008. "Preschool Television Viewing and Adolescent Test Scores: Historical Evidence from the Coleman Study." *Quarterly Journal of Economics* 71(3):279–323.

If you want to learn more about the complications of difference-in-differences when there are N units and N periods, have a look at

Andrew Goodman-Bacon. 2018. "Difference-in-Differences with Variation in Treatment Timing." NBER Working Paper No. 25018.

Kosuke Imai and In Song Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63(2):467–90.

The two studies on oral contraception that we mentioned are

Claudia Goldin and Lawrence F. Katz. 2002. "The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions." *Journal of Political Economy* 110(4):730–70.

Martha J. Bailey. 2006. "More Power to the Pill: The Impact of Contraceptive Freedom on Women's Life Cycle Labor Supply." *Quarterly Journal of Economics* 121(1): 289–320.

The study of newspaper endorsements in the United Kingdom is

Jonathan McDonald Ladd and Gabriel S. Lenz. 2009. "Exploiting a Rare Communication Shift to Document the Persuasive Power of the News Media." *American Journal of Political Science* 53(2):394–410.

The studies of the contagiousness of obesity and of acne and height are

Nicholas A. Christakis and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network over 32 Years." *New England Journal of Medicine* 373:370–79.

Ethan Cohen-Cole and Jason Feltcher. 2009. "Detecting Implausible Social Network Effects in Acne, Height, and Headaches: Longitudinal Analysis." *British Medical Journal* 338(7685):28–31.

There is a ton of work on the democratic peace. A Google Scholar search will turn up many interesting theoretical arguments. The papers we mentioned are

Henry S. Farber and Joanne Gowa. 1991. "Common Interests or Common Polities? Reinterpreting the Democratic Peace." *Journal of Politics* 59(2):393–417.

Donald P. Green, Soo Yeon Kim, and David H. Yoon. 2001. "Dirty Pool." *International Organization* 55(2):441–68.

The argument that it was appropriate to include fixed effects in regressions probing the democratic peace was sufficiently controversial at the time that the journal editors invited several other prominent social scientists to comment on the piece in the same issue of the journal.

The study of orchestra auditions discussed in exercise 4 is

Claudia Goldin and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90(4):715–41.