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

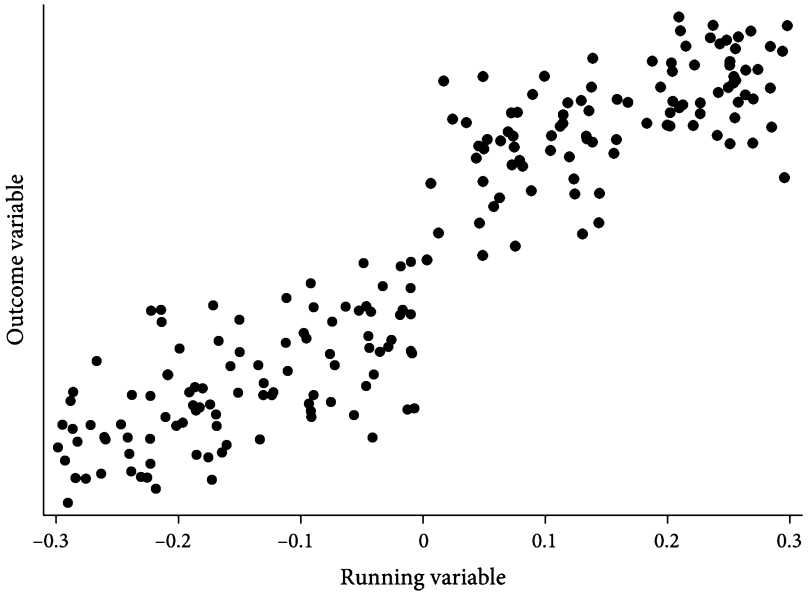

Figure 12.1. Scatter plot with treatment determined by a continuous running variable. Black dots are treated units. Gray dots are untreated units.

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

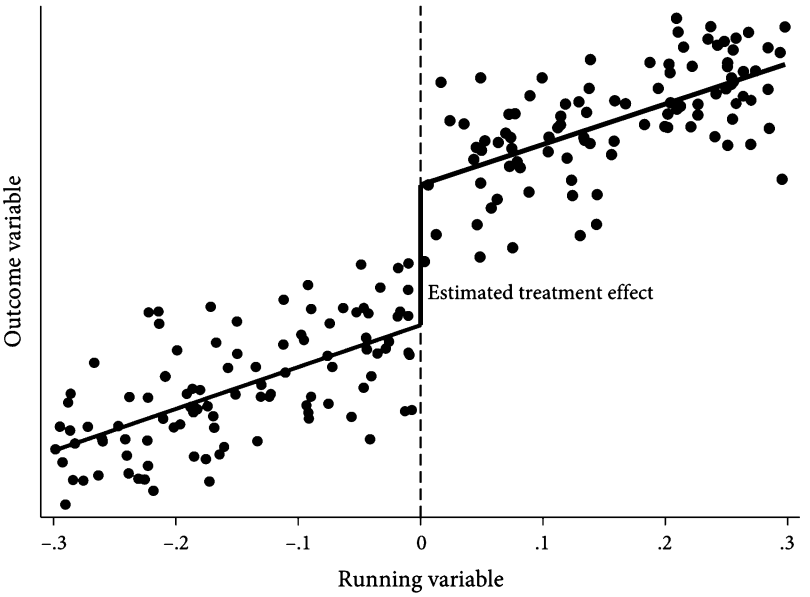

Figure 12.2. The regression discontinuity design estimates the jump in expected outcomes at the threshold, which is the causal effect of the treatment for units at the threshold.

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

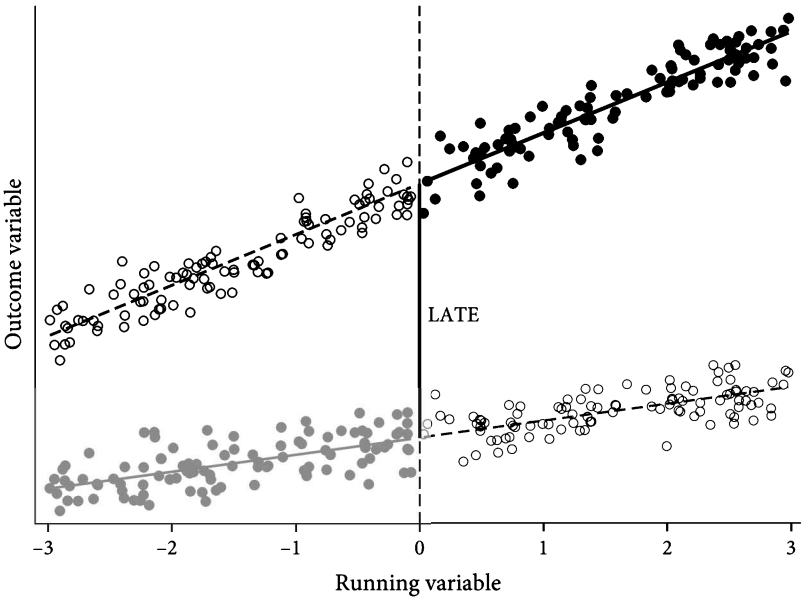

Figure 12.3. A regression discontinuity design estimates the LATE at the threshold. This need not be the overall average treatment effect, as average treatment effects may differ for different values of the running variable

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

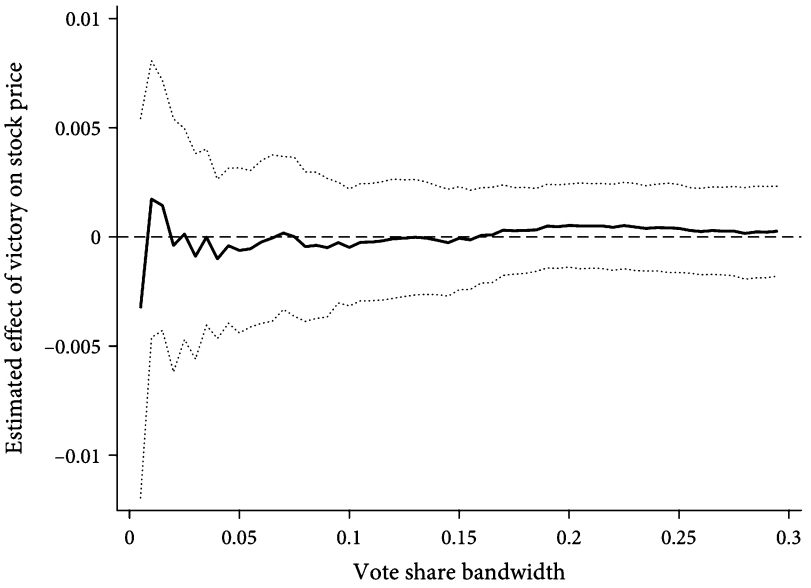

Figure 12.4. Visualizing how an RD estimate (solid) and confidence interval (dotted) depends on the bandwidth.

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

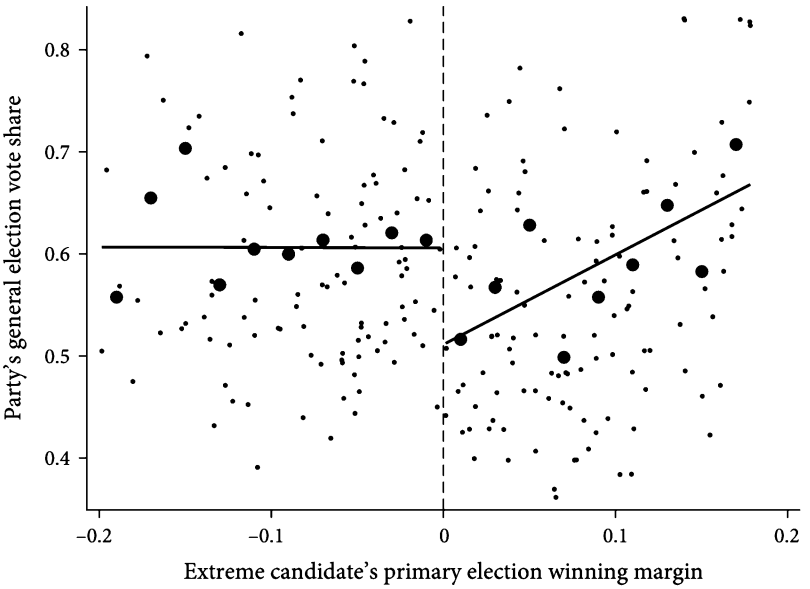

Figure 12.5. The effect of running an extremist on electoral prospects.

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?

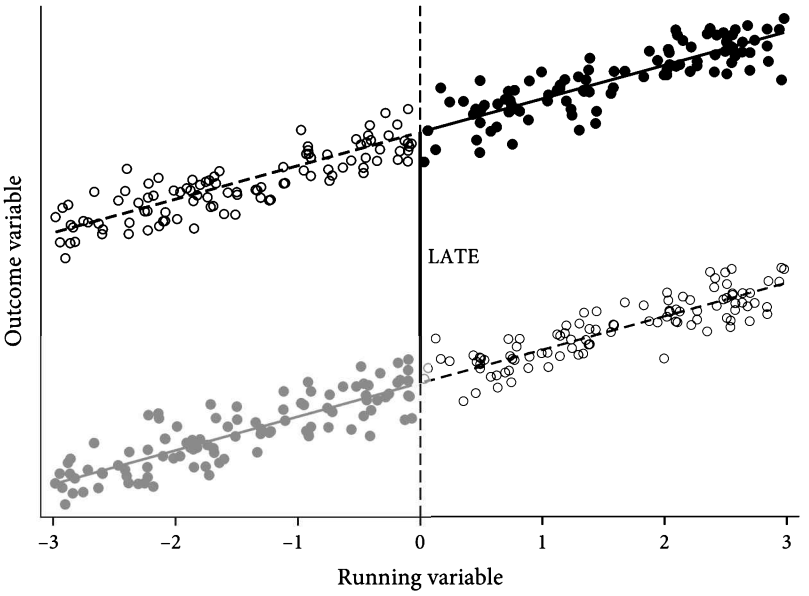

Figure 12.6. A case where average potential outcomes satisfy continuity.

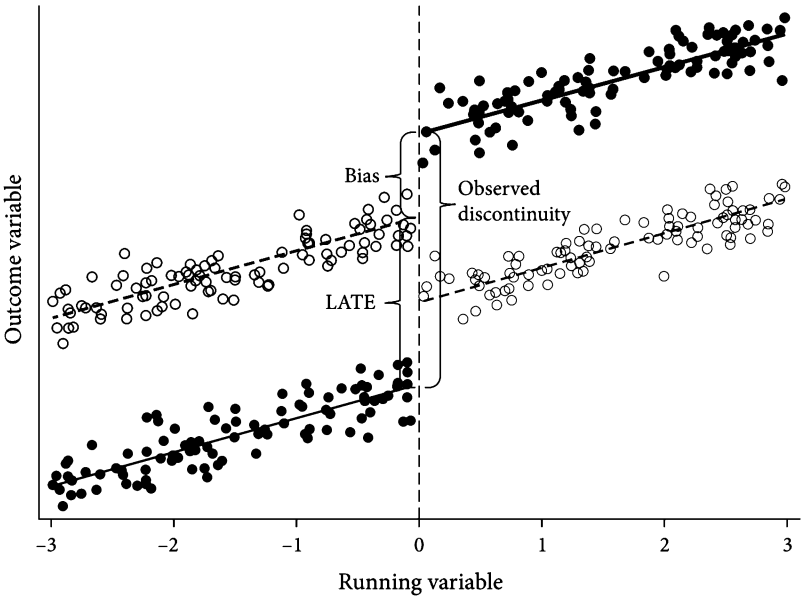

Figure 12.7. A case where average potential outcomes do not satisfy continuity at the threshold.

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.

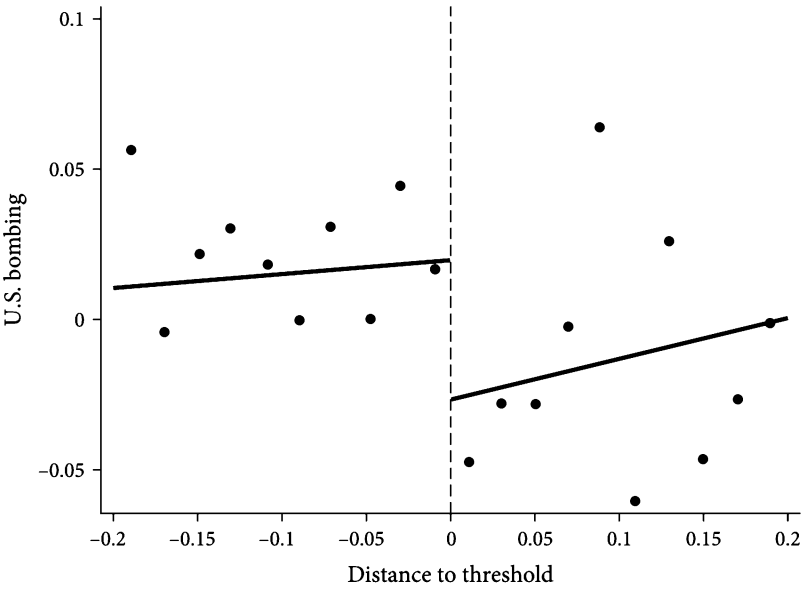

Figure 12.8. Hamlets that just barely received better grades in the Hamlet Evaluation System were bombed less frequently than hamlets that just barely received worse grades.

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.

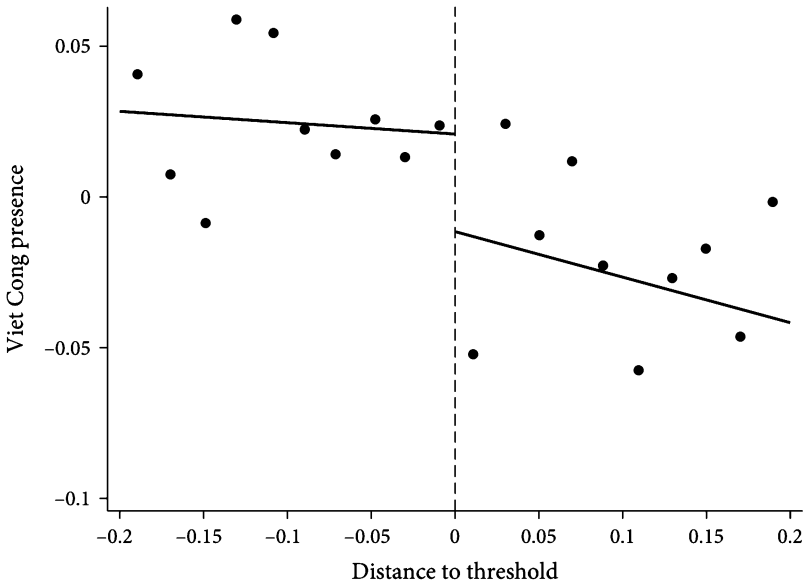

Figure 12.9. Hamlets that experienced more bombing saw more subsequent insurgent activity compared to otherwise similar hamlets that experienced less bombing.

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.

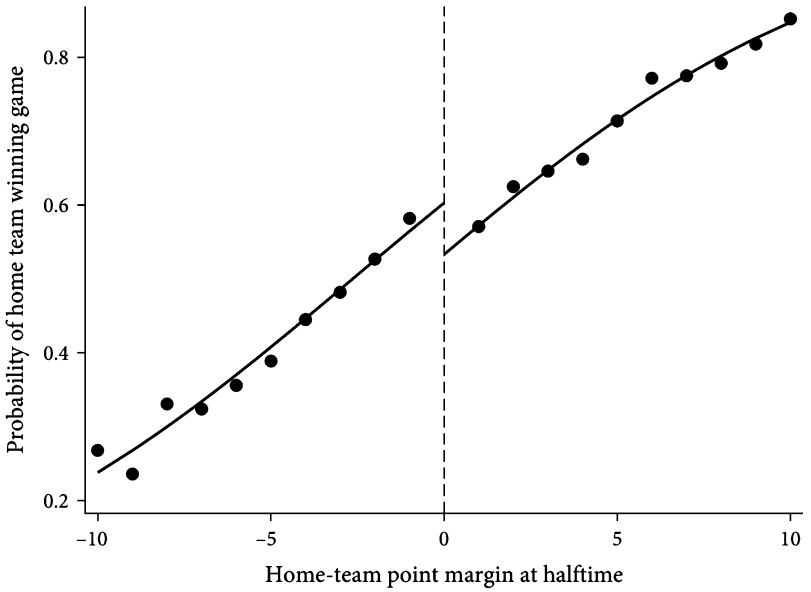

Figure 12.10. The effect of being ahead or behind at half time on winning the game.

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.