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.

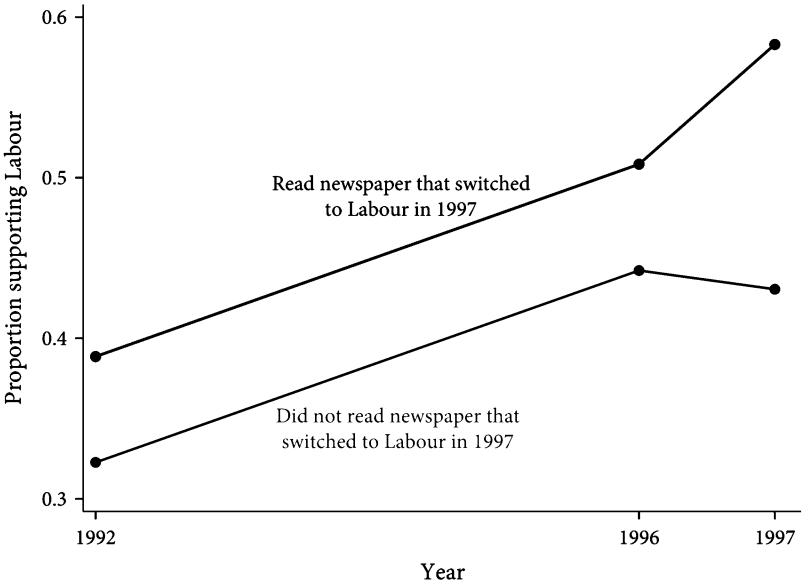

Figure 13.1. Visualizing pre-trends and using difference-in-differences to estimate the effect of newspaper endorsements on vote choices.

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.