CHAPTER 10

Controlling for Confounders

What You'll Learn

- If we can observe a confounder, we can control for it and mitigate the bias arising from it.
- The most common way to control for a confounder is by including it in a regression, although there are other approaches.
- Through graphs and simple examples, you will develop an intuitive understanding of how this works.
- Controlling is not magic. It doesn't remove bias arising from unobserved confounders or reverse causation.
- We should typically control for confounders but not mechanisms.

Introduction

In chapter 9, we saw that confounders are a big problem when we're trying to learn about causal relationships from correlations. Here, we are going to talk about the first line of defense against confounders, controlling.

You likely have heard people talk about controlling before, but what does it really mean? *Controlling* involves using statistical techniques to find the correlation between two variables, holding the value of other variables constant. The easiest way to begin to understand the idea is through some examples.

Party Whipping in Congress

A not terribly surprising fact about the United States Congress is that Republicans are more likely to vote in a conservative manner than Democrats. One way of measuring this, quantitatively, is through the scores given to each congressional representative by the right-leaning interest group the American Conservative Union (ACU). Each year, the ACU chooses twenty-five important bills and gives each congressperson a score between 0 and 100 based on how they voted on those bills. Since the ACU leans right, a higher score indicates a more conservative voting record.

We can back up the claim that being a Republican is correlated with a conservative voting record by checking whether Republicans have higher ACU scores than Democrats on average. Table 10.1, based on data from the House of Representatives

Table 10.1. Comparing the voting records of Republicans and Democrats in the U.S. Congress.

	Average ACU Score
Republicans	83
Democrats	19
Difference	64

in 1997, shows that they do. Democratic congressional representatives have an average ACU score of 19, while Republicans have an average ACU score of 83. On average, Republicans vote 64 ACU points more conservatively than do Democrats.

This data indicates that Republican and Democratic congressional representatives vote quite differently. What might explain such polarization?

One idea advanced by many political scientists is that party pressure causes the divergence in legislative voting behavior. Parties have lots of tools at their disposal to pressure rank-and-file members to vote the party line. Perhaps most important among these tools is help with fundraising for reelection campaigns.

But before we interpret the correlation between party membership and voting record as evidence for the effect of party discipline, we should consider possible confounders. A confounder, in this case, is some other feature of the world that affects both congressional representatives' party membership and their voting records.

As illustrated in figure 10.1, ideology is one obvious candidate for a confounder. The Republican party has a conservative reputation. The Democratic party has a liberal reputation. Hence, a conservative may be more likely to run as a Republican and a liberal may be more likely to run as a Democrat. Moreover, a politician's personal ideological leanings may well influence how they vote on legislation once in Congress. If people sort into the parties according to ideology in this way, there is reason to think that Republican representatives would vote more conservatively and Democratic representatives would vote more liberally, even if the parties exercised no discipline. So personal ideology is plausibly a confounder. In light of this, it would be a mistake to interpret the correlation between party membership and voting record as an unbiased estimate of the causal impact of party discipline on the voting behavior of representatives.

In order to address this potential confounder, we would like to control for it. In its simplest form, controlling for ideology simply means looking at the correlation between party membership and voting record, holding personal ideology constant. To do so, we first need a measure of personal ideology. Fortunately, we have a plausible candidate.

In 1996, a non-partisan organization called Project Vote Smart administered a survey, the National Political Awareness Test (NPAT), to congressional candidates. The survey asked candidates their views on a wide array of issues. From their answers, Project Vote Smart then generated a liberal to conservative ranking. Seventy-six percent of candidates responded to the survey, so we have a measure of the political ideology of a large number of congressional representatives.[1]

To control for personal ideology in our analysis of the relationship between party membership and voting record, we simply compare the voting records of Democrats

[1]Response rates declined considerably in subsequent elections, which explains why we're showing data from the late 1990s, even though it precedes the birth of many of our readers.

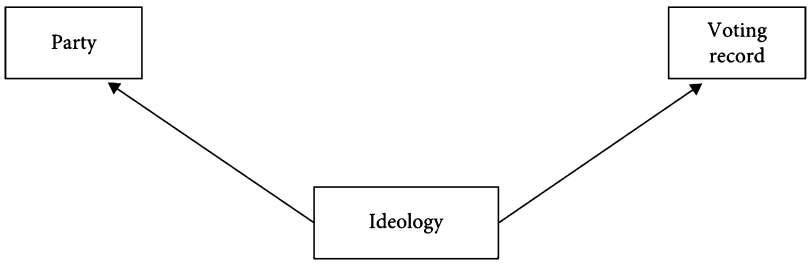

Figure 10.1. Ideology affects what party a politician joins and how that politician votes when in office. Hence, it is a confounder.

Table 10.2. NPAT scores controlling for party.

		Liberal ⟵ NPAT Percentile ⟶ Conservative				
		1–20	*21–40*	*41–60*	*61–80*	*81–100*
Republicans	*Avg ACU Score*	n/a	44	68	86	94
	# of People	0	4	45	69	69
Democrats	*Avg ACU Score*	10	18	41	96	84
	# of People	70	66	24	1	1
Difference in Average ACU Score		n/a	26	27	−10	10

and Republicans with similar NPAT scores. If the NPAT is doing a good job of measuring personal ideology, these comparisons will tell us about the difference in voting records of Republicans and Democrats, holding personal ideology constant (or, controlling for personal ideology).

Table 10.2 sorts congressional representatives into five bins, based on their NPAT scores. The left-most bin has representatives with the most liberal ideology according to their NPAT answers. The bins become progressively more conservative as we move to the right.

Looking at the data broken down in this way, a few things immediately jump out. First, and most importantly, in no column is the difference between Republican and Democratic voting records anywhere close to the 64-point difference we found before controlling for personal ideology. This suggests that personal ideology was an important confounder in that correlation—a large portion of the difference in voting record between Democrats and Republicans was due to the members of those two parties having different underlying personal preferences about policy rather than to party pressure. The reason, of course, is as we said earlier. More conservative people tend to become Republicans and more liberal people tend to become Democrats. This fact is reflected in the observation that the number of people in each cell is increasing in conservatism for Republicans and decreasing in conservatism for Democrats (i.e., almost all the Republicans are in the NPAT 41st–100th percentile and almost all the Democrats are in the NPAT 1st–60th percentile).

Second, within a party, as you move across the ideological bins, average ACU scores are, for the most part, going up. There is one exception—Democrats in the 61st–80th percentile vote more conservatively than Democrats in the 81st–100th percentile—but this comparison is not especially informative because it involves the comparison of only two people, since there are so few ideologically conservative Democrats.

Third, the difference between average Republican and average Democratic voting records varies across the columns. That is, the correlation between party and voting depends on ideology. This is fine. But often we want a single, overall measure of the correlation between partisanship and voting record controlling for personal ideology, rather than an ideology-by-ideology measure. To get that single number, we will need to take some sort of a weighted average of the differences from the various columns. But how do we decide what weights to give to each column?

As we start to think about the correct weights, note that there is clearly one column that is more informative than the others about the different voting behavior of Democrats and Republicans with similar personal ideologies—the column for the NPAT 41st–60th percentile. In each of the other columns, there are either very few Republicans or very few Democrats. But in the 41st–60th percentile column there are a large number of representatives from both parties. This isn't surprising—the place to look for ideological overlap across the parties is in the ideological center. So we probably want our weighted average to put a lot of weight on that column.

More generally, it is useful to think back to chapter 5, where we learned about how ordinary least squares (OLS) *regression* fits a line to data to minimize the sum of squared errors. (When we refer to regression in this chapter, we will always be referring to OLS regression.) OLS is one principled way to choose weights for the five columns. So consider the following regression:

$$\begin{aligned}\text{ACU Rating} = \alpha + \beta_1 \cdot \text{Republican} + \beta_2 \cdot \text{NPAT}_{21-40} + \beta_3 \cdot \text{NPAT}_{41-60} \\ + \beta_4 \cdot \text{NPAT}_{61-80} + \beta_5 \cdot \text{NPAT}_{81-100} + \varepsilon\end{aligned}$$

In this regression, the unit of analysis is an individual representative. The variable *ACU Rating* is an individual representative's ACU score. The variable *Republican* is what we call a *dummy variable*: it takes the value 1 if the representative is a member of the Republican party and the value 0 if the representative is a member of the Democratic party. The various *NPAT* variables are also dummy variables, taking a value of 1 if the representative is in the relevant percentile range and a value of 0 otherwise.[2] The greek letter ε (*epsilon*) represents the error.

The coefficient β_1 in this regression gives us the weighted average we have been talking about—that is, β_1 is the correlation between ACU score and being a Republican, controlling for personal ideology (as measured by NPAT percentile). We will also get estimates for the coefficients on the four included NPAT categories and the intercept (α). These also have interpretations. However, we are running the regression because we are interested in the correlation between ACU score and Republicanism controlling for ideology, so we focus on β_1.

[2]Since everyone is in one of the five NPAT categories, one of them must be omitted. Here, we have omitted the 1st–20th percentile. This is analogous to the fact that we can't include both a Democrat and a Republican variable in the regression when every member is either one or the other. We couldn't separately identify the effect of being a Democrat and the effect of being a Republican, so we just include a Republican variable and interpret the coefficient as the effect of being a Republican versus being a Democrat.

Running this regression on our data yields an estimate of β_1, which we label $\hat{\beta}_1$, equal to 24. (It is an estimate because our data is a sample drawn from the population of all congressional representatives, so the observed correlation also reflects noise.) Not surprisingly, this is very close to the difference between average Republican ACU score and average Democratic ACU score in the column corresponding to the 41st–60th percentile, which, as we said, is where almost all the information is. The regression, of course, puts a little weight on the other columns, dragging the estimate down from 27 to 24. But that column is basically telling us the answer.

Having controlled for ideology, we still probably don't have a terribly credible estimate of the causal effect of party discipline on the voting records of congressional representatives. This is because there could be many other confounders beyond personal ideology. That is, within an NPAT bin, there may be lots of other factors that lead some people to become Democrats and others to become Republicans that also have an independent effect on their voting behavior in Congress. For instance, even holding fixed personal ideology, Democrats may tend to represent districts with more liberal voters and Republicans may tend to represent districts with more conservative voters. If politicians choose how to vote on bills with an eye toward how their voters will react, then these differences in constituencies are yet another confounder. We are sure you can think of others.

As the list of confounders grows, making a table that breaks down the data into all the different possible cells becomes more difficult and unwieldy. But, as long as you can measure the potential confounders, you can control for them in a regression. Doing so will always get you an estimate of β_1 reflecting the weighted average of the various cells in that (imagined) big table that minimizes the sum of squared errors. Given this, regression will be our most important tool for controlling for confounders. Therefore, it is useful to have a better understanding of exactly how controlling with regression works.

A Note on Heterogeneous Treatment Effects

As we discussed in chapter 3, for almost all interesting examples of causal relationships, the effects of interest are heterogeneous—that is, they're not the same for every unit of observation. This was true in our flu shot example, where the flu shot prevented some people from getting the flu who otherwise would have, but didn't prevent other people from getting the flu, either because they weren't going to get it in the first place or because they were and the flu shot didn't work for them. It's probably also true for the above example about party effects on voting. To the extent that parties affect roll-call voting by members of Congress, this effect is probably not the same for every member of Congress. Perhaps some members of Congress are strong ideologues who will vote the same way regardless of any party pressure, so that there is no treatment effect. Perhaps others depend on their party's support for reelection and would do whatever their party leaders asked, so that there is a strong treatment effect. And perhaps others are somewhere in between.

It's important to think clearly about such heterogeneity when controlling because, as our discussion around table 10.2 showed, once we start controlling for confounders, we're no longer estimating the average effect of the treatment across all units. In our example, to estimate the relationship between party and voting, controlling for ideology, we put more weight on members of Congress with moderate ideologies. This is

because there isn't much variation in party among members of Congress with extreme ideologies—basically, all strong conservatives are Republicans and all strong liberals are Democrats. If the effect of party membership is different for ideological moderates than it is for ideological extremists, we have to acknowledge that we're focusing on the former effect.

This acknowledgement raises a thorny problem. If controlling for a potential confounder meaningfully changes our causal estimate, this could be a sign that the estimate without controlling was biased and that controlling reduced that bias. This is to the good. But it could also be a sign that there are heterogeneous treatment effects, and we've changed the subset of units for which we're estimating the average effect. To the extent that the estimand we really care about is the average treatment effect across all units, this could be to the bad.

These challenges will arise for other methods beyond controlling that we'll discuss later in the book. We will refer back to this discussion when relevant. Sometimes we'll say that instead of estimating the average treatment effect (ATE) as our estimand, we can only estimate a local average treatment effect (LATE) as our estimand, where *local* refers to the subset of units for which we can generate a credible estimate. When treatment effects are heterogeneous across units, the LATE need not be the same as the ATE. So if the ATE is the estimand we really care about, we need to think clearly about the extent to which estimates of the LATE may or may not be informative about the ATE. But, as the economist Guido Imbens says of situations where we can only credibly estimate a local average treatment effect, "Better LATE than nothing."

The Anatomy of a Regression

The key ingredients in any regression for causal inference are

- the *dependent variable* (also called the *outcome variable*),
- the *treatment variable*, and
- a set of *control variables*.

The dependent variable is the outcome you are trying to understand. The treatment variable is the feature of the world whose effect on the dependent variable you are trying to estimate. And the control variables are potential confounders that you are including in the regression to reduce bias.

In the simple case where there is only one control variable, we write the regression equation as

$$Y = \alpha + \beta \cdot T + \gamma \cdot X + \varepsilon \tag{10.1}$$

where Y is the dependent variable, T is the treatment variable, and X is the control variable. The regression parameters (i.e., the quantities we'd like to estimate) are the intercept α, the effect of the treatment β, and the "effect" of the control variable γ. There is also an error term, ε, reflecting the fact that units differ from their predicted outcome for idiosyncratic reasons.

There's nothing in the regression equation that distinguishes the treatment variable from the control variable. This distinction is conceptual and is driven by the question you are trying to answer. If you want to know the effect of party on voting, controlling for ideology, the party variable is your treatment and NPAT is your control. But if you had wanted to know the effect of ideology on voting, controlling for party, this would be reversed.

This is also why the word *effect* is in scare quotes above when referring to the effect of the control variable. Often, we don't actually care about the regression parameter associated with a control variable (here, γ). What's important is that β is an effect of interest, and we are going to try to estimate it in an unbiased way.

One way to read Equation 10.1 is to take it literally. We can pretend that we know the data-generating process. Each individual i's outcome (Y_i) equals a common intercept (α) plus $\beta \cdot T_i$ plus $\gamma \cdot X_i$ plus idiosyncratic factors (ϵ_i). Another way to read the equation is to acknowledge that we don't know the data-generating process, but nonetheless, we'd like to estimate β—the average linear relationship between Y and T, controlling for X.

As we noted in chapter 5 (though we didn't quite put it this way), whatever the data-generating process, OLS regression always gives us the *best linear approximation to the conditional expectation function* (BLACEF). So we don't have to pretend to know the data-generating process in order to run a regression. If there are no baseline differences across values of T after controlling for X, then the BLACEF corresponds to the average effect of T on Y. In this case, knowing β is very valuable.

Just as in our discussion from chapter 5, when we run this regression, we get estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ by computing the values of α, β, and γ that minimize the sum of the squared errors. Let's see what that means.

For any arbitrary values of the regression parameters—say α', β', and γ'—the associated prediction of Y_i for an individual i is

$$\alpha' + \beta' \cdot T_i + \gamma' \cdot X_i.$$

Let's label the idiosyncratic errors associated with this regression ε'. For each observation i, they are the actual outcome minus the predicted outcome:

$$\varepsilon'_i = Y_i - (\alpha' + \beta' \cdot T_i + \gamma' \cdot X_i)$$

The OLS estimates—$\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$—are the particular values of the regression parameters that minimize the sum of the square of these errors. Our computer can compute them really quickly.

Suppose we know that once we control for X, there are no other omitted confounders. So the regression of Y on T and X gives an unbiased estimate of the effect of T on Y. One question we might ask is how biased our results would have been if we failed to control for X.

It turns out, we can answer that question. Call Equation 10.1 above the *long regression* because it includes X. Now suppose we ran the following *short regression* instead:

$$Y = \alpha^S + \beta^S \cdot T + \varepsilon^S \tag{10.2}$$

The superscript S here indicates that we are talking about the short regression. Importantly, there's no guarantee that β^S from the short regression will be the same as β from the long regression. In fact, they won't be the same if X is a confounder.

We can quantify the bias associated with failing to include X in the regression. Consider a regression that treats the control variable (X) as a dependent variable and regresses it on the treatment (T):

$$X = \tau + \pi \cdot T + \xi$$

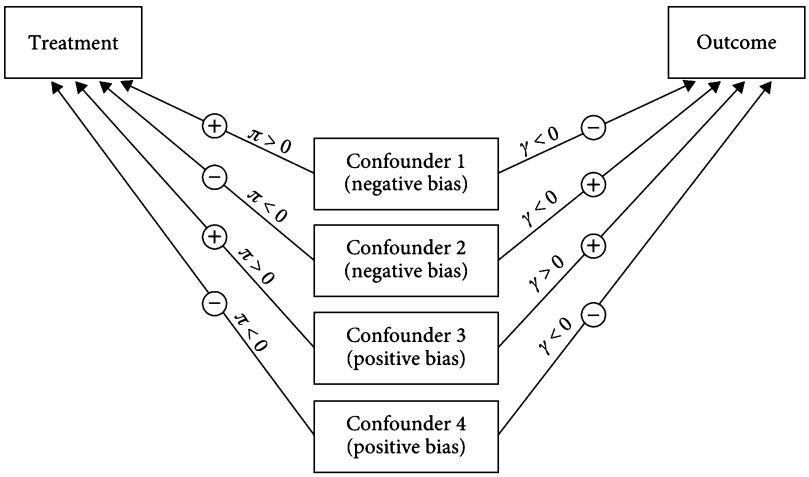

Figure 10.2. The omitted variables bias formula tells us how to sign the bias from an omitted confounder.

You'll notice that we've used different Greek letters for the regression parameters here. We now call the intercept τ (the Greek letter *tau*), the coefficient on the treatment π (the Greek letter *pi*), and the error ξ (the Greek letter *xi*). We did this for a couple reasons.

First, and most importantly, we didn't want to use the same letters here that had different meanings above. The parameter π, here, describes the correlation between T and X—it's the slope relating changes in T to changes in X. We didn't want you to confuse that with the two β's we've seen in this section (β and β^S, each of which describes some version of the relationship beween the treatment and the outcome Y). Second, we also don't want you to think that there's something special about certain Greek letters. It's not the case that α must always represent the intercept, β the coefficient on the treatment, and so on. After all, these are just symbols. We would like you to be able to look at the equation and figure out what the constant is, what the coefficient on the treatment is, what the error is, and so on, even if someone uses completely different symbols than the ones we use.

The bias from excluding X from the regression of the outcome on the treatment is $\beta^S - \beta$. It turns out, this bias is equal to $\pi \cdot \gamma$. That is,

$$\text{Bias} = \beta^S - \beta = \pi \cdot \gamma.$$

We sometimes call this the *omitted variable bias* formula.

What this formula tells us is that the short regression gives a biased estimate of the effect of the treatment on the outcome if the control variable is correlated with the treatment variable (so that $\pi \neq 0$) and the control variable influences the outcome variable (so that $\gamma \neq 0$).

If we can't observe X, we can't control for it by including it in the regression. But the omitted variable bias formula gives us a way to think about the direction and extent of the bias. Indeed, the omitted variable bias formula formalizes our ideas from chapter 9 about how to sign the bias, as summarized in figure 10.2, which repeats figure 9.7 but points out that the regression parameters π and γ directly measure the relationships relevant for determining the sign of the bias.

Table 10.3. The omitted variable bias formula helps us think about whether failing to control for a confounder results in an over- or under-estimate of the causal effect.

	Omitted Variable Positively Correlated with Treatment $\pi > 0$	**Omitted Variable Negatively Correlated with Treatment** $\pi < 0$
Omitted Variable Positively Correlated with Outcome $\gamma > 0$	Positive bias $\pi \cdot \gamma > 0$	Negative bias $\pi \cdot \gamma < 0$
Omitted Variable Negatively Correlated with Outcome $\gamma < 0$	Negative bias $\pi \cdot \gamma < 0$	Positive bias $\pi \cdot \gamma > 0$

If there's an unobserved confounder that we suspect is positively related to both T (so $\pi > 0$) and Y (so $\gamma > 0$), then the omitted variable bias formula tells us that $\beta^S - \beta > 0$, so we are over-estimating the effect of T. The same is true if the confounder is negatively related to both T and Y (so that π and γ are both negative)—again, the bias is positive and we are getting an over-estimate. If the confounder is positively related to T but negatively related to Y (so $\pi > 0$ and $\gamma < 0$) or vice versa (so $\pi < 0$ and $\gamma > 0$), the bias is negative and we under-estimate the effect of T. This is summarized in table 10.3.

How Does Regression Control?

We've seen that controlling for a variable (X) can change the coefficient describing the relationship between some other variable of interest (T) and an outcome variable (Y). In particular, controlling for X will change the estimated relationship between T and Y if X is correlated with T and has an independent relationship with Y. Here's one way to think, graphically, about what the regression is doing when we control for a variable.

Suppose we want to know the effect of height on income, in which case both our outcome and treatment variables of interest are continuous (they can, in principle, take an infinite and uncountable number of possible values). Figure 10.3 shows some data on income and height from the National Longitudinal Survey conducted by the U.S. Bureau of Labor Statistics. A representative sample of U.S. residents born between 1980 and 1984 were asked about their heights and their incomes in 2014, when they were between the ages of 34 and 38.

To allow for easier visualization, we grouped respondents by height and gender, so every dot in figure 10.3 corresponds to a group of fifteen or more individuals of the same gender and height (measured in inches). The figure plots the average income of each group, measured in thousands of dollars above $20,000, and the average height, measured in feet above 5 feet. (You'll see in a moment why we scaled our variables in such an unusual way.) The hollow dots correspond to groups of men and the solid dots correspond to groups of women.

Visually, we see a strong, positive correlation between height and income. What would we get if we ran a regression of income on height with this data, ignoring gender?

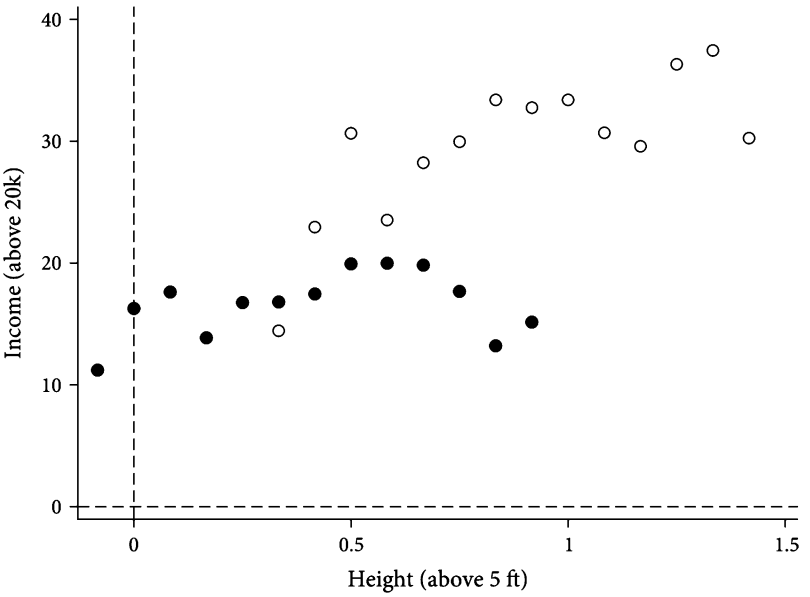

Figure 10.3. Income and height among 34- to 38-year-old Americans in 2014.

As we've seen previously, this would simply involve finding the line that best fits the data. Figure 10.4 plots that line. Indeed, the best-fitting line has a strong positive slope, indicating that, on average, taller people earn higher incomes.

To be a little more precise, the regression finds the line that best fits the data by identifying the values of α and β that minimize the sum of squared errors in the following equation:

$$\text{Income} = \alpha + \beta \cdot \text{Height} + \varepsilon$$

These two values are illustrated by figure 10.5. The height of the line when Height = 0 (i.e., when a person is 5 feet tall) is $\hat{\alpha}$, and the slope of the line is $\hat{\beta}$. For this particular data set, we estimate a slope of about 14.8. On average, people who are one foot taller earn an extra $14,800 of income per year!

Of course, before we draw a causal interpretation from this regression coefficient, we should think about confounders. Gender is one possibility. Men are, on average, taller than women. And we suspect that men, on average, earn higher incomes than women for reasons unrelated to height. (This could be the result of gender discrimination in labor markets or other societal factors. Although the reasons are, of course, very important, we don't need to know them in order to control for gender as a confounder.) Indeed, we can see in the picture that women do seem to have lower heights and lower incomes, on average. So gender is a confounder that we might want to control for in this regression.

One way we could start addressing this concern would be to run separate regressions for men and women:

$$\text{Income} = \alpha^M + \beta^M \cdot \text{Height} + \varepsilon^M$$

$$\text{Income} = \alpha^W + \beta^W \cdot \text{Height} + \varepsilon^W$$

If we did that, we would fit two regression lines, as shown in figure 10.6.

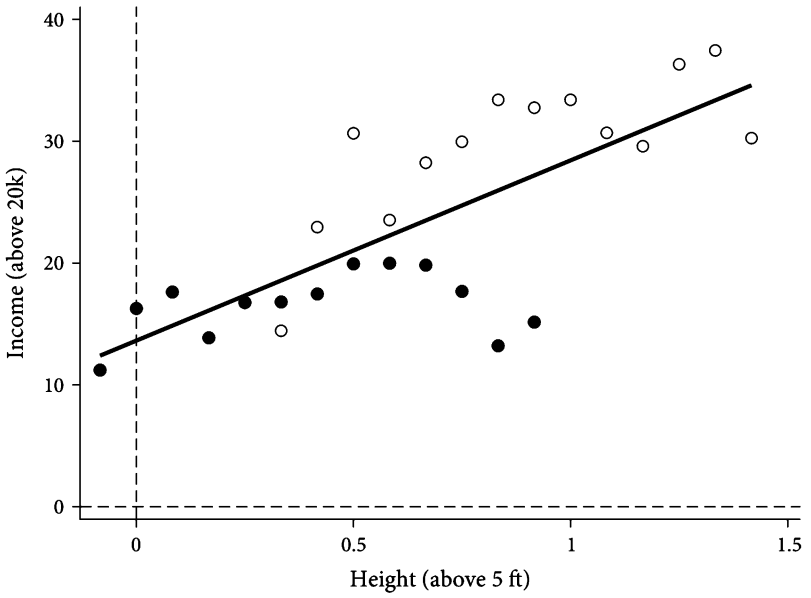

Figure 10.4. Regressing income on height.

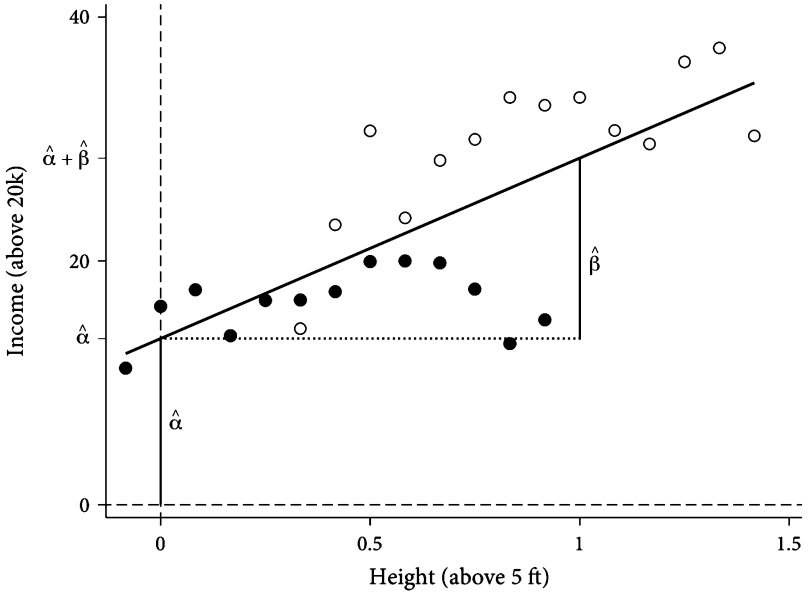

Figure 10.5. Graphical interpretation of regression coefficients.

The separate regression lines for men and women are shown in gray, while the previous regression line that pooled everyone together is still shown in black. Interestingly, the correlation between height and income is smaller within each gender than it is across the population as a whole. That is, both $\hat{\beta}^W$ and $\hat{\beta}^M$ are smaller than $\hat{\beta}$ from our earlier regression. Also notice that the slope is greater for men than it is for women, $\hat{\beta}^M > \hat{\beta}_W$.

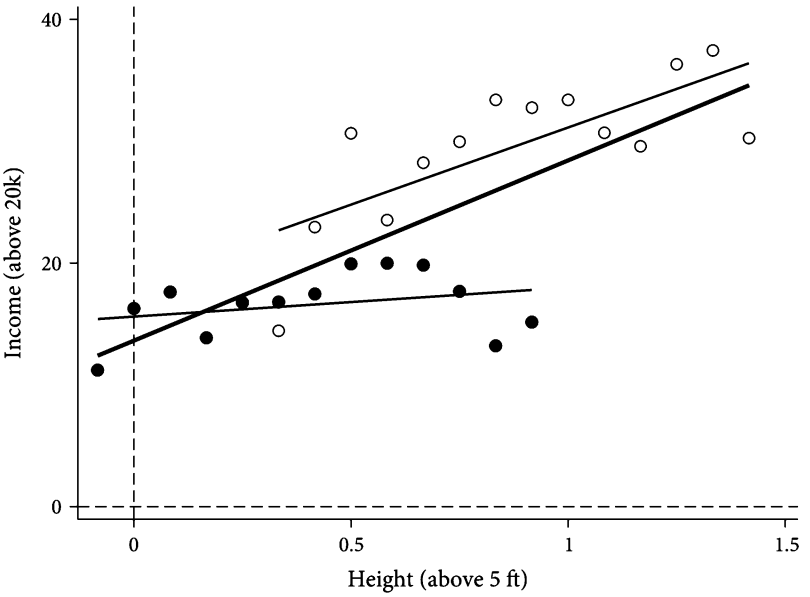

Figure 10.6. Representing both the regression line for the pooled data (black line) as well as separate regression lines for men and women (gray lines).

This procedure of splitting the data and running separate regressions tells us the correlation between income and height separately for men and women. Thinking back to our congressional politics example, this is analogous to the cells at the bottom of table 10.2, which told us the difference in average ACU score between Republicans and Democrats for each bin of NPAT scores.

While the separate correlations are good to know, just as in the congressional politics example, we might want to have one summary estimate of the correlation between income and height, controlling for gender. That number will be a weighted average of the slopes of the two gray lines in figure 10.11 (just as in the congressional politics example, where the single number was a weighted average of the individual differences at the bottom of table 10.2.). But we need to know how to assign the weights.

The most straightforward way to do this is to run a regression of income on both height and gender. The regression equation would look like this:

$$\text{Income} = \alpha + \beta \cdot \text{Height} + \gamma \cdot \text{Male} + \varepsilon$$

Graphically, how will this regression separately estimate α, β, and γ? Instead of finding one line that best fits the data, we can think of finding two lines that best fit the data—one for men and one for women. But unlike when we ran separate regressions, we now constrain those two lines to have the same slope ($\hat{\beta}$). Figure 10.7 shows how those two lines look if we do that and compares them to the lines we got when we ran separate regressions for men and women.

Notice that the slope of the two black lines is identical, by construction. And the slope is somewhere in between the slope for those we got running the two separate regressions (the gray lines). That is, it is a weighted average of the two. Figure 10.8

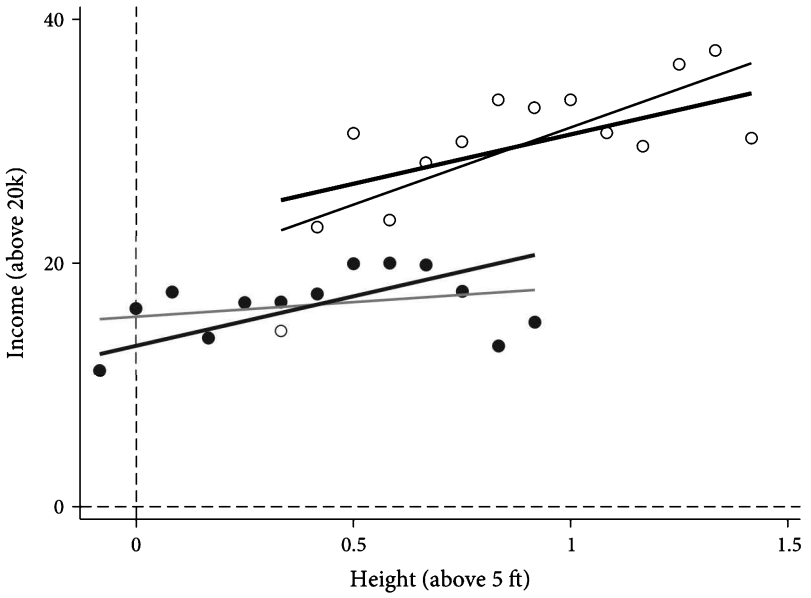

Figure 10.7. Representing the regression where we control for gender by including it in the regression of income on height (black lines) as well as separate regression lines for men and women (gray lines).

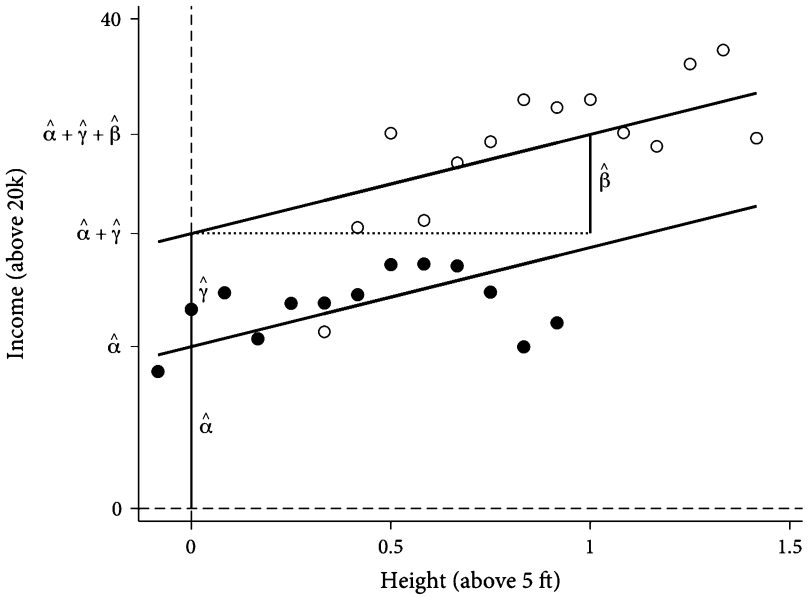

Figure 10.8. Regression coefficients when controlling for gender in the regression of income on height.

shows how, having estimated these two parallel lines that best fit the data, we've also estimated the regression parameters.

The intercept of the line for women (Male = 0) is $\hat{\alpha}$. The distance between the two lines is $\hat{\gamma}$. And the slope of the two lines is $\hat{\beta}$. Put differently, $\hat{\alpha}$ is the predicted income for women who are 5 feet tall; $\hat{\gamma}$ is the predicted difference in income between men

and women of the same height; and $\hat{\beta}$ is the average relationship between height and income, controlling for gender.

Not surprisingly, controlling for gender has significant implications for the estimated relationship between height and income. Instead of 14.8, our new estimate for the slope is about 8.1. The change is due to the fact that gender is a confounder—it affects both height and income. Using our tools for signing the bias from chapter 9, we know that if the confounder is positively correlated with both treatment and outcome, as is the case here, it creates positive bias. Since 14.8 was an over-estimate of the true effect of height on income, when we control for gender, we get a smaller estimate.

It is worth noting that controlling for gender affects not only our estimate of the relationship between income and height but also the precision of that estimate, although the direction of that effect is theoretically ambiguous. On the one hand, adding a control that is correlated with the outcome reduces the residual variation in that outcome, which improves precision. On the other hand, adding a control that is correlated with the treatment reduces the residual variation in the treatment, which increases the uncertainty of our estimates. Whether controlling for a confounder improves or harms precision will depend on the relative impact of those two forces.

Given the discussion above, it might be tempting to add additional control variables to your regression, not for the purpose of reducing bias but with the goal of improving precision. Indeed, if you can find pre-treatment variables that are strongly correlated with the outcome but not the treatment; including them in a regression will tend to improve the precision of your estimates. However, if you keep trying control variables until you get a statistically significant estimate, that's p-hacking, and it's a bad idea.

Since we've talked about the analogy between what we've just done and our congressional politics example, let's revisit that example in a regression framework. Notice, in this case, the treatment (Republican or Democrat) is binary, but the potential confounder (ideology) is measured continuously by the NPAT score.

Again, start with a scatter plot, this time of the American Conservative Union rating on the vertical axis and the NPAT conservative score on the horizontal axis. In figure 10.9, the hollow dots correspond to Democrats and the solid dots correspond to Republicans.

Since the treatment is binary, we can start with a simple comparison of the average ACU rating for Republicans and for Democrats. Consider the following regression equation:

$$\text{ACU Rating} = \alpha + \beta \cdot \text{Republican} + \varepsilon$$

To minimize the sum of squared errors, the coefficient $\hat{\alpha}$ equals the average ACU rating for a Democrat (Republican = 0) and the coefficient $\hat{\beta}$ is the difference between the average ACU rating for a Republican and for a Democrat. Thus, as we've already seen, $\hat{\alpha} = 20$ and $\hat{\beta} = 84 - 20 = 64$. This is illustrated in figure 10.10, where the horizontal lines correspond to the average ACU ratings for Democrats and Republicans.

Our concern, of course, is that personal political ideology is a confounder in this regression, so that $\hat{\beta}$ does not estimate the true effect of party on congressional voting behavior. So we would like to control for personal ideology. We will do so using the NPAT Conservativeness score, measured on the horizontal axis.

In our income-height example, we were concerned about a confounder that was measured as a binary variable—gender. Our first step in building intuition about controlling was to consider running the original regression (income and height) separately for each

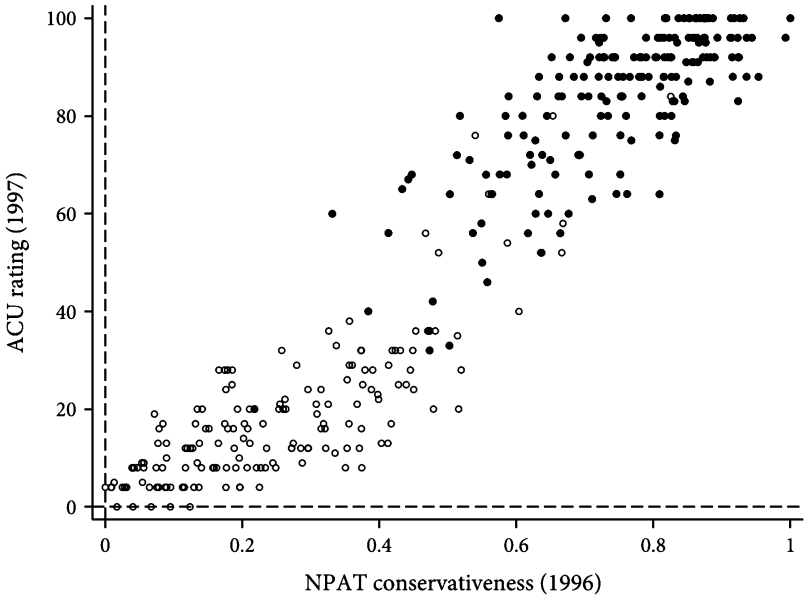

Figure 10.9. ACU score and NPAT conservative score.

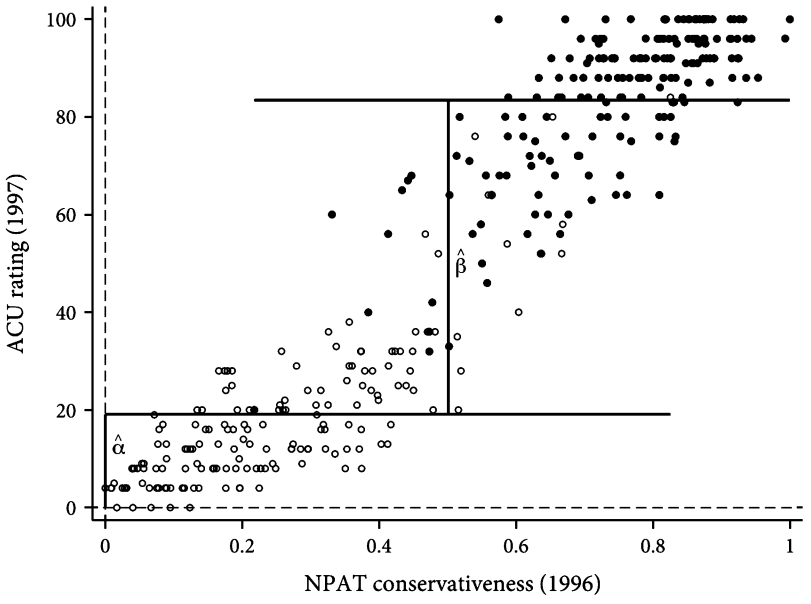

Figure 10.10. Coefficients in regression of ACU score on party.

value of the confounder. Then we saw that the final regression coefficient on height controlling for gender was a weighted average of the slopes of these two separately estimated regression lines.

Here, in our congressional politics example, because our confounder is continuous, we cannot do a separate regression for each value of the confounder. But we can

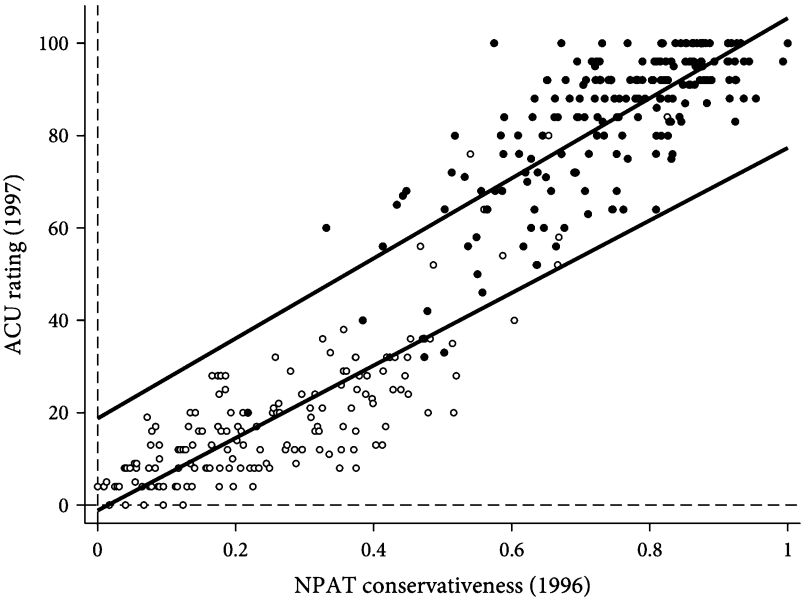

Figure 10.11. Controlling for NPAT score (ideology) in the relationship between ACU score and party by running two separate regressions, one for each party.

do something similar: run a regression of ACU Rating on NPAT Conservativeness separately for Democrats and for Republicans (the superscripts P on the regression coefficients refer to the idea that this is the regression for party P):

$$\text{ACU Rating} = \alpha^P + \gamma^P \cdot \text{NPAT Conservativeness} + \varepsilon^P$$

That gives us two regression lines, one for Republicans and one for Democrats, as in figure 10.11.

For each value of NPAT Conservativeness, the predicted ACU Rating of a Republican with that NPAT Conservativeness score is

$$\hat{\alpha}^R + \hat{\gamma}^R \cdot \text{NPAT Conservativeness}$$

And for each value of NPAT Conservativeness, the predicted ACU Rating of a Democrat with that NPAT Conservativeness score is:

$$\hat{\alpha}^D + \hat{\gamma}^D \cdot \text{NPAT Conservativeness}$$

This means that at any given value of NPAT Conservativeness, the gap between the two lines is the difference in predicted ACU Rating between Republicans and Democrats with that NPAT score. Hence, this regression allows us to get a continuous analogue of our earlier binary comparison. It tells us, for each value of NPAT Conservativeness, what the predicted difference in mean ACU Rating is between Republicans and Democrats.

But we aren't done. As before, the goal is to get a single measure of the relationship between ACU Rating and party membership, controlling for NPAT Conservativeness.

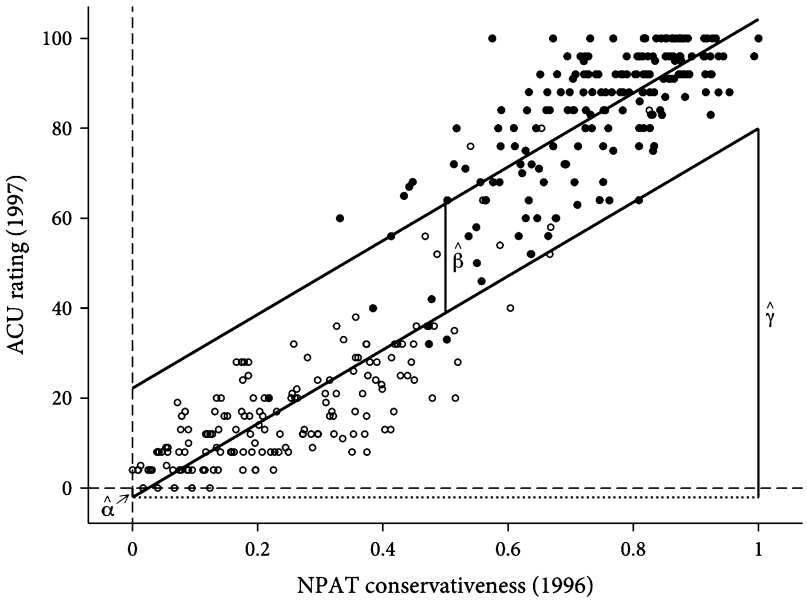

Figure 10.12. Controlling for NPAT score (ideology) in a regression of ACU score on party.

Currently, we have a separate measure of that relationship for each value of NPAT Conservativeness. The final step in controlling, then, is to use regression to create the weighted average of these differences that minimizes the sum of squared errors. We do so with the following regression:

$$\text{ACU Rating} = \alpha + \beta \cdot \text{Republican} + \gamma \cdot \text{NPAT Conservativeness} + \varepsilon$$

Figure 10.12 illustrates this regression. The parameter $\hat{\alpha}$ tells us the average ACU Rating of a Democrat with NPAT Conservativeness of 0. The parameter $\hat{\gamma}$ tells us the slope of the relationship between ACU Rating and NPAT Conservativeness. Importantly, unlike our previous two regressions, where $\hat{\gamma}^R$ and $\hat{\gamma}^D$ were different, this regression imposes that the slope of the relationship between ACU Rating and NPAT Conservativeness be the same for both parties. Hence, this slope $\hat{\gamma}$ is a weighted average of $\hat{\gamma}^R$ and $\hat{\gamma}^D$. Finally, the coefficient $\hat{\beta}$ is the gap between the two lines. This gap is constant across NPAT Conservativeness scores because we forced $\hat{\gamma}$ to be the same for both parties, making the lines parallel. Hence, $\hat{\beta}$ estimates the average difference in ACU Rating between Republicans and Democrats controlling for NPAT Conservativeness.

Controlling and Causation

While controlling allows you to mitigate or remove the biases arising from specific confounders that you are able to measure and include in your regression, we are typically still skeptical in most cases that controlling alone allows us to uncover unbiased estimates of causal relationships. Remember from chapter 9 that if we want to interpret a correlation as an unbiased estimate of a causal effect, we must believe that there are no baseline differences between the treated and untreated units. In other

words, our comparison has to plausibly be apples-to-apples. If we regress Y on T and X (and other possible confounders), we're still making a similar assertion if we give the coefficient on T a causal interpretation. We are saying that, other than the set of variables we controlled for, we believe there are no confounders in the relationship between Y and T and no reverse causality. Put differently, for controlling to give an unbiased estimate of a causal effect, we must control for *all* the confounders.

In our experience, it is hard to find situations (other than randomized experiments, which we will discuss in chapter 11) where it feels plausible that there are really no omitted confounders. Typically, even if the analyst controls for lots of things, you can think of other potential confounders that are either unobservable or unmeasured in the data and, thus, can't be controlled for. For instance, ask yourself whether you can think of any potential confounders beyond gender in the relationship between income and height. The answer is, of course, yes, including economic, biological, cultural, health, and other characteristics. For instance, wealthy parents might provide their children with better nutrition, which might make them taller, and might also help their children in other ways that allow them to earn higher incomes. It is hard to imagine that you would be able to measure and control for all possible confounders.

Reverse causation is another reason we don't generally think controlling for confounders can uncover causal relationships on its own. In chapter 9, we talked about how both confounders and reverse causation can prevent us from making an apples-to-apples comparison. The idea of controlling is to try to account for confounders as best we can, but if there is reverse causation, meaning the outcome affects the treatment, there's no amount of controlling that can make that problem go away.

Let us give you an example.

Is Social Media Bad for You?

There is widespread concern that exposure to social media is bad for people. And, indeed, many studies show a negative correlation between social media usage and various measures of a person's subjective well-being and mental health.

Of course, that correlation may not reflect a causal effect of social media on well-being. For instance, there might be reverse causality—perhaps people who are sad, lonely, or distressed spend more time on social media than people who are happier or more socially connected. Or there might be confounders—perhaps socioeconomic status, education, or geography affect both social media usage and subjective well-being.

A first thing you might think to do to get at an estimate of the causal relationship is to control for some of these confounders. How well will that controlling strategy do at estimating the true causal effect?

There is a study that can provide some insight into that question. A group of scholars interested in the effects of social media ran an experiment. They first identified a large group of Facebook users willing to participate in their study. (The participants didn't know what the study was about.) From each of these people they elicited measures of subjective well-being, Facebook usage, and how much they'd have to be paid to turn off Facebook for a month. Then the experimenters randomly selected some of these people and in fact paid them to turn off Facebook for a month (which they were able to monitor). The others didn't turn off Facebook but continued to be part of the study as a control group. The researchers then measured subjective well-being again at the end to see whether turning off Facebook had changed the subjective sense of well-being for

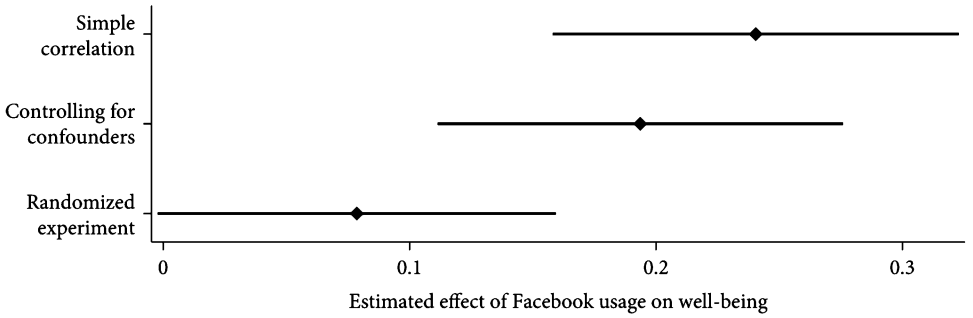

Figure 10.13. Estimates of relationship between Facebook usage and subjective well-being.

those in the treatment group, relative to those in the control group who did not turn off Facebook.

The nice thing about this study, for our purposes, is that the experiment, by randomly assigning Facebook usage, gives an unbiased estimate of the effect of Facebook usage. The researchers also asked about Facebook usage and subjective well-being at the start of the study, so by comparing levels of well-being across people with different levels of Facebook usage at the beginning of the study, they could also replicate the simple correlation reported in earlier studies. Moreover, the researchers, also observed a bunch of details about the individuals in their study and so could control for some potential confounders—for example, income, age, sex, education, race, political affiliation.

If they were able to control for all confounders, then the estimate of the relationship between Facebook usage and subjective well-being from controlling and from the experiment would be the same in expectation. So, by comparing the simple correlation, the correlation controlling for these potential confounders, and the experimental estimate, we can start to get a sense of how well controlling does, in this setting, at recovering the true causal effect.

Figure 10.13 shows the simple correlation, the correlation controlling for potential confounders, and the estimate from the experiment (each surrounded by a 95% confidence interval). All are measured in units of average Facebook use per day. As you can see, the simple correlation gives the biggest estimate of Facebook's negative relationship with subjective well-being. Controlling for potential confounders reduces that estimate a bit. But the experimental estimate is about one-third the size of the estimates from the simple correlation and about one-half the size of the estimate controlling for potential confounders. This suggests that the strategy of controlling, here, still leaves us with a substantial over-estimate of the true effect.

Reading a Regression Table

You've seen graphical representations of regression. But when you run a regression on your computer or see regression results discussed in a report, they are often presented in the form of a table. So it is worthwhile to be able to understand and interpret the different parts of a regression table. We first saw regression tables in chapter 5, but we now know enough to go into more detail.

Let's go back to our analysis of the relationship between congressional voting records and party. In that setting, we ran three regressions.

First, we regressed our measure of roll-call voting on party without controlling for anything:

$$\text{ACU Rating} = \alpha + \beta \cdot \text{Republican} + \varepsilon$$

Second, we controlled for ideology by including indicators for different ranges of the NPAT score:

$$\text{ACU Rating} = \alpha + \beta_1 \cdot \text{Republican} + \beta_2 \cdot \text{NPAT}_{21-40} + \beta_3 \cdot \text{NPAT}_{41-60} + \beta_4 \cdot \text{NPAT}_{61-80} + \beta_5 \cdot \text{NPAT}_{81-100} + \varepsilon$$

This is what we did to find the correct weighted average in our discussion surrounding table 10.2.

Third, we controlled for ideology by including the continuous NPAT variable:

$$\text{ACU Rating} = \alpha + \beta \cdot \text{Republican} + \gamma \cdot \text{NPAT Conservativeness} + \varepsilon$$

Each column of table 10.4 presents the results from one of these three regressions.

Let's talk about how to read this table. The first column simply contains labels. The second column shows the results from our first regression: ACU Rating on Republican, controlling for nothing. The third column shows the results from our second regression: ACU Rating on Republican, controlling for NPAT category. The fourth column shows the results from our third regression: ACU Rating on Republican, controlling for the continuous NPAT Conservativeness score.

Along the first row, we see the name of our dependent variable. For these regressions, this is always ACU Rating. For each of our regressions, the row labeled *Republican* shows three pieces of information. The top number is our estimate of the coefficient on Republican in our regression. The bottom number in parentheses is the standard error on that estimate. And the stars indicate whether the result is statistically significantly different from zero (and at what level). Looking across this row, we see that in the first regression, the coefficient on Republican is 64.32. But once we control for NPAT score, it drops dramatically. If we control with NPAT categories, it drops to 23.74. And if we control with the continuous NPAT Conservativeness score, it is 24.28. (Not surprisingly, it doesn't much matter exactly how we control for ideology.)

Going down the table, we then get the coefficient estimates, standard errors, and statistical significance for each of our control variables. This is why the next five rows are blank in the second column—we didn't control for anything in that regression. In the third column, the four rows associated with the NPAT categories are filled in, but the row associated with the NPAT Conservativeness score is blank. And in the fourth column, the NPAT category rows are blank, but NPAT Conservativeness is filled in. For all three regressions, the row called *Constant* is filled in. This is the estimate of the intercept ($\hat{\alpha}$) from that regression.

The table contains two more pieces of information. For each regression, the table tells us how many observations there were in the data. Here, the answer is 349, reflecting the number of congresspeople who filled out the NPAT survey in 1997.

And, for each regression, the table reports the r-squared statistic. Recall from chapter 2 that this is the proportion of variation in one variable that can be predicted by variation in other variables. So a value of .93 in our final regression says that, within the sample of data we have, you can predict 93 percent of the variation in a congressperson's ACU rating using party and NPAT score.

Table 10.4. Relationship between ACU rating and party.

Variables	ACU Rating	ACU Rating	ACU Rating
Republican	64.32**	23.74**	24.28**
	(1.71)	(2.25)	(1.98)
$NPAT_{21-40}$		8.01**	
		(1.76)	
$NPAT_{41-60}$		32.74**	
		(2.29)	
$NPAT_{61-80}$		52.27**	
		(2.83)	
$NPAT_{81-100}$		59.77**	
		(2.83)	
NPAT Conservatism			82.05**
			(3.44)
Constant	19.09**	10.29**	−2.10
	(1.25)	(1.24)	(1.18)
Observations	349	349	349
r-squared	.80	.92	.93

Standard errors in parentheses. $^{**}p < .01$

While that sounds pretty good, we urge you not to over-interpret the r-squared statistic. In fact, when we run regressions, we often don't even report it. Typically, our goal is not to predict or model the variation in our dependent variable. It is to learn whether our key treatment variable matters for our outcome. For that, what we are really interested in is the coefficient estimate on that variable. Moreover, getting a high r-squared statistic, on its own, isn't very meaningful. One easy way to successfully predict a lot of the variation in your data is to just include lots of control variables in your regression. But this doesn't mean you've understood what is going on at all. Think back to our discussion of overfitting in chapter 5. Just because you fit the data really well (which is all that a high r-squared means) by including lots of variables doesn't mean you can do a good job predicting what the outcome will be when you look at observations not from your data set. And in some cases, you can have a reliable, unbiased estimate of your quantity of interest even though your r-squared is low.

Controlling for Confounders versus Mechanisms

Thinking clearly about controlling gets dicier when there is some variable that affects both the treatment and the outcome but is also affected by the treatment. What can we do in this situation? That variable is a confounder: it affects both the treatment and the outcome. So it seems that we should control for it. But as we discussed in chapter 9, that variable is also a mechanism: it is affected by the treatment and affects the outcome. So it seems that we should not control for it, since it is part of the pathway by which the treatment affects the outcome. What are we to do?

To make this conundrum more concrete, let's return to our example from chapter 9 where we were interested in the effect of per capita income on civil war. On the one hand, democracies might implement better policies that improve income and might also provide better opportunities for non-violent expression of political grievances, which might directly affect civil war risk. From this perspective, whether a country is a democracy or not is a confounder—that is, a pre-treatment covariate—and, thus, should be controlled for. On the other hand, perhaps as a country becomes richer, its citizens become more informed and start demanding greater democracy, which then reduces the likelihood they turn to civil war. From this perspective, democracy is one of the mechanisms by which GDP affects the likelihood of civil war—that is, a post-treatment covariate—and, thus, should not be controlled for.

There really isn't a solution in such situations. You are damned if you do and damned if you don't—which means that you aren't in a position to learn much about the causal relationship you are interested in. To do that, you'd need a more creative approach, which will be the subject of the next several chapters.

There Is No Magic

People would really like to believe that they can estimate causal relationships by just controlling for all the confounders. And they'd like you to believe it too. But, as we've just discussed, in many important settings you will only believe there are no omitted confounders if you aren't thinking clearly. And so, sometimes, people will use mathematical jargon, cool-sounding statistical methods, complicated computer programs, and other technical wizardry to try to get you to not think clearly. It is important to not be fooled. No matter what fancy techniques an analyst uses, if the fundamental strategy is to control for the confounders, and if there are plausible confounders that are either unobservable or unmeasured in the data, then they can't possibly have controlled for them. Computers aren't magical. They can control for observable confounders. But they cannot make unobservable confounders observable.

To see what we mean, consider the case of a perfectly fine and useful statistical technique called *matching*. Here's the idea. Suppose you have a continuous variable, X, that you'd like to control for. You could match each treated unit to whatever untreated unit has the most similar value of X. Then, you might compute a difference of means (or run a regression) on that matched data set in order to estimate the effect of T on Y. This is called *nearest neighbor matching*.

So, in our congressional politics example, you would start by matching each Democratic congressperson to the Republican congressperson with the most similar NPAT score. Then, in this matched sample, you would compute the difference in average ACU score between matched pairs of Republicans and Democrats, which is another way of estimating the relationship between ACU score and political party, controlling for NPAT score.

If you had multiple variables you wanted to control for, you'd have to define some summary measure of how similar any two observations are across those variables. There are lots of strategies for doing this. Some of them get pretty fancy in terms of computation, which can make it hard to keep thinking clearly. You must try to keep your wits about you.

Matching has some advantages over regression as a technique for controlling. One nice feature of matching is that it allows for more flexibility in the way in which the control variable might influence the outcome of interest. For instance, whereas regression assumes the relationship is linear, matching makes no such assumption. Matching also

has disadvantages relative to regression. One downside of matching is that it's often less precise than regression because you're using less information. Another downside is that matching estimates can be biased because the best match for a treated observation will have, in expectation, a higher value of X if, for example, X is positively correlated with T. There are statistical solutions to that problem as well, which can again start looking pretty technical and fancy.

Matching, like regression, is a good statistical technique for controlling. We have no objection to it. We get concerned because analysts sometimes like to present some very technical matching algorithm and then say things like "Matching creates an experiment-like comparison of units that differ in the treatment but are otherwise the same." Such claims are an attempt to blind you with science. Matching is just a tool for controlling. It creates no more of an experiment-like comparison than does a regression that includes control variables. That is to say, it controls for the variables that were observed and matched on—nothing more. Your computer, no matter how fancy the statistical algorithm, can't make the unobservables observable. Because that would be magic. And there is no magic.

Wrapping Up

Controlling is a way to account for confounders and obtain better, less biased estimates of causal relationships. There are lots of different ways to control, but they're all fundamentally trying to do the same thing—generate more credible estimates by comparing treated and untreated units with similar values of other observable pre-treatment covariates.

While controlling is a valuable tool, it's not a silver bullet. In most interesting cases, there will still be unobservable confounders that we can't control for, reverse causation, or variables that are part confounder and part mechanism. So even when researchers have controlled for lots of potential confounders, we should still worry about biased estimates.

If controlling is typically an unconvincing strategy for estimating causal relationships, what can we do that would be more convincing? One way—perhaps the only way—to ensure unbiased estimates is to randomize the treatment yourself. Therefore, the next chapter focuses on the so-called gold standard for causal inference—the randomized experiment.

Key Terms

- **Controlling:** Using a statistical technique to find the correlation between two variables, holding the value of other variables constant.
- **Dummy variable:** A variable that indicates whether a given unit has some particular characteristic, taking a value of 1 if the unit has that characteristic and 0 if the unit does not.
- **Dependent or Outcome variable:** The variable in your data corresponding to the feature of the world that you are trying to understand or explain with your regression.
- **Treatment variable:** The variable in your data corresponding to the feature of the world whose effect on the dependent variable you are trying to estimate.
- **Control variable:** A variable in your data that you include in your statistical analysis in an attempt to reduce bias in your estimate of a causal effect.

- **Omitted variables bias:** The bias resulting from failing to control for some confounder when attempting to estimate a causal effect.
- **Local average treatment effect (LATE):** The average treatment effect for some specific subset of the population.

Exercises

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

(a) Run a regression of incumbent vote share (your dependent variable) on both incumbent spending and challenger spending.

i. Note that if challenger spending is positively correlated with higher challenger vote shares, it must be negatively correlated with incumbent vote share. In light of this, how should we interpret the estimated coefficients associated with your independent variables?

ii. Are the results you obtained different from those you obtained when you ran separate regressions of incumbent vote share on incumbent spending and incumbent vote share on challenger spending in chapter 9? Why or why not?

(b) Let's add some controls to your regression in an attempt to obtain more reliable estimates of the effect of campaign spending. As you may know, 2018 was a good year for Democrats in House elections.

i. Is the overall good performance of Democrats in 2018 a potential confounder in your regression?

ii. Create a new variable indicating whether the incumbent is a Republican—call it *republicanincumbent*. It should take a value of 1 if the incumbent is a Republican and a value of 0 if the incumbent is a Democrat.

iii. Re-run your regression, but include that variable as a control.

iv. Interpret the estimated coefficient associated with your new republicanincumbent variable.

v. Does including this control variable meaningfully change your estimated coefficients of interest (i.e., the coefficients on incumbent and challenger spending)? Why or why not, do you think?

(c) Now add in a control for the vote share that the incumbent's party received in that district in the 2016 presidential election.

i. What kind of concern might including this control variable address?

ii. Interpret the estimated coefficient associated with this control variable.

iii. Does including this control variable meaningfully change your estimated coefficients of interest (i.e., the coefficients on incumbent and challenger spending)? Why or why not, do you think?

10.2 Produce a regression table that shows the results of each of the regressions from exercise 1, along with the number of observations and the r-squared.

10.3 In chapter 2, we discussed a study that finds a correlation between taking advanced math classes in high school and college completion, which the researchers presented as evidence of a causal relationship. Of course, we might worry that the kinds of students who take advanced math courses are different from those who do not, so the authors of the study run regressions that control for gender, socioeconomic status, race, cognitive ability test scores, and eighth-grade reading and math scores.

(a) Do these control variables assuage your concerns about potential confounders?
(b) Even after controlling for these background variables, name a potential omitted confounder that concerns you. What is the likely direction of the bias associated with this potential confounder?

Readings and References

For more details on controlling, as well as more details on experiments, instrumental variables, difference-in-differences, and regression discontinuity (topics we will cover in the next three chapters), we recommend

> Joshua Angrist and Jorg-Steffen Pischke. 2014. *Mastering 'Metrics*. Princeton University Press.

For more information on political polarization, including details on increasing polarization in the U.S. Congress over the past seven or so decades, we recommend

> Nolan McCarty. 2019. *Polarization: What Everyone Needs to Know*. Oxford University Press.

For more on the LATE versus the ATE, including a defense of credible estimates of a LATE, see

> Guido W. Imbens. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48(2):399–423.

The study on Facebook and subjective well-being is

> Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The Welfare Effects of Social Media." *American Economic Review* 110(3):629–76.