9 Multiple Regression: The Basics

OVERVIEW

Despite what we have learned in the preceding chapters on hypothesis testing and statistical significance, we have not yet crossed all four of our hurdles for establishing causal relationships. Recall that all of the techniques we have learned in Chapters 8 and 9 are simply bivariate, X- and Y-type analyses. But, to fully assess whether X *causes* Y, we need to control for other possible causes of Y, which we have not yet done. In this chapter, we show how multiple regression – which is an extension of the two-variable model we covered in Chapter 9 – does exactly that. We explicitly connect the formulae that we include to the key issues of research design that tie the entire book together. We also discuss some of the problems in multiple regression models when key causes of the dependent variable are omitted, which ties this chapter to the fundamental principles presented in Chapters 3 and 4. Lastly, we will incorporate an example from the political science literature that uses multiple regression to evaluate causal relationships.

9.1 MODELING MULTIVARIATE REALITY

From the very outset of this book, we have emphasized that almost all interesting phenomena in social reality have more than one cause. And yet most of our theories are simply bivariate in nature.

We have shown you (in Chapter 4) that there are distinct methods for dealing with the nature of reality in our designs for social research. If we are fortunate enough to be able to conduct an experiment, then the process of randomly assigning our participants to treatment groups will automatically "control for" those other possible causes that are not a part of our theory.

But in observational research – which represents the vast majority of political science research – there is no automatic control for the other possible causes of our dependent variable; we have to control for them statistically. The main way that social scientists accomplish this is through multiple regression. The math in this model is an extension of the math involved in the two-variable regression model you just learned in Chapter 9.

9.2 THE POPULATION REGRESSION FUNCTION

We can generalize the population regression model that we learned in Chapter 9,

$$\text{bivariate population regression model: } Y_i = \alpha + \beta X_i + u_i,$$

to include more than one systematic cause of Y, which we have been calling Z throughout this book:

$$\text{multiple population regression model: } Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_i.$$

The interpretation of the slope coefficients in the three-variable model is similar to interpreting bivariate coefficients, with one very important difference. In both, the coefficient in front of the variable X (β in the two-variable model, β_1 in the multiple regression model) represents the “rise-over-run” effect of X on Y. In the multiple regression case, though, β_1 actually represents the effect of X on Y *while holding constant the effects of* Z. If this distinction sounds important, it is. We show how these differences arise in the next section.

9.3 FROM TWO-VARIABLE TO MULTIPLE REGRESSION

Recall from Chapter 9 that the formula for a two-variable regression line (in a sample) is

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i.$$

And recall that, to understand the nature of the effect that X has on Y, the estimated coefficient $\hat{\beta}$ tells us, on average, how many units of change in Y we should expect given a one-unit increase in X. The formula for $\hat{\beta}$ in the two-variable model, as we learned in Chapter 9, is

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Given that our goal is to control for the effects of some third variable, Z, how exactly is that accomplished in regression equations? If a scatter plot in two dimensions (X and Y) suggests the formula for a *line*, then adding a third dimension suggests the formula for a *plane*. And the formula for that plane is

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i.$$

That might seem deceptively simple. A formula representing a plane simply adds the additional $\beta_2 Z_i$ term to the formula for a line.[1]

Pay attention to how the notation has changed. In the two-variable formula for a line, there were no numeric subscripts for the β coefficient – because, well, there was only one of them. But now we have two independent variables, X and Z, that help to explain the variation in Y, and therefore we have two different coefficients β, and so we subscript them β_1 and β_2 to be clear that the values of these two effects are different from one another.[2]

The key message from this chapter is that, in the preceding formula, the coefficient β_1 represents more than the effect of X on Y; in the multiple regression formula, it represents *the effect of X on Y while controlling for the effect of Z*. Simultaneously, the coefficient β_2 represents *the effect of Z on Y while controlling for the effect of X*. And in observational research, this is the key to crossing our fourth causal hurdle that we introduced all the way back in Chapter 3.

How is it the case that the coefficient for β_1 actually controls for Z? After all, β_1 is not connected to Z in the formula; it is, quite obviously, connected to X. The first thing to realize here is that the preceding multiple regression formula for β_1 is different from the two-variable formula for β from Chapter 9. (We'll get to the formula shortly.) The key consequence of this is that the value of β derived from the two-variable formula, representing the effect of X on Y, will almost always be different – perhaps only trivially different, or perhaps wildly different – from the value of β_1 derived from the multiple regression formula, representing the effect of X on Y while controlling for the effects of Z.

But how does β_1 control for the effects of Z? Let's assume that X and Z are correlated. They need not be related in a *causal* sense, and they need not be related *strongly*. They simply have to be related to one another – that is, for this example, their covariance is not exactly equal to zero. Now, assuming that they are related somehow, we can write their relationship just like that of a two-variable regression model:

$$X_i = \hat{\alpha}' + \hat{\beta}' Z_i + \hat{e}_i.$$

[1] All of the subsequent math about adding one more independent variable, Z, generalizes quite easily to adding still more independent variables. We use the three-variable case for ease of illustration.

[2] In many other textbooks on regression analysis, just as we distinguish between β_1 and β_2, the authors choose to label their independent variables X_1, X_2, and so forth. We have consistently used the notation of X, Y, and Z to emphasize the concept of controlling for other variables while examining the relationship between an independent and a dependent variable of theoretical interest. Therefore we will stick with this notation throughout this chapter.

Note some notational differences here. Instead of the parameters $\hat{\alpha}$ and $\hat{\beta}$, we are calling the estimated parameters $\hat{\alpha}'$ and $\hat{\beta}'$ just so you are aware that their values will be different from the $\hat{\alpha}$ and $\hat{\beta}$ estimates in previous equations. And note also that the residuals, which we labeled $\hat{u}_i$ in previous equations, are now labeled $\hat{e}_i$ here.

If we use Z to predict X, then the predicted value of X (or $\hat{X}$) based on Z is simply

$$\hat{X}_i = \hat{\alpha}' + \hat{\beta}' Z_i,$$

which is just the preceding equation, but without the error term, because it is expected (on average) to be zero. Now, we can just substitute the left-hand side of the preceding equation into the previous equation and get

$$X_i = \hat{X}_i + \hat{e}_i$$

or, equivalently,

$$\hat{e}_i = X_i - \hat{X}_i.$$

These $\hat{e}_i$, then, are the exact equivalents of the residuals from the two-variable regression of Y on X that you learned from Chapter 9. So their interpretation is identical, too. That being the case, the $\hat{e}_i$ are the portion of the variation in X that Z cannot explain. (The portion of X that Z *can* explain is the predicted portion – the $\hat{X}_i$.)

So what have we done here? We have just documented the relationship between Z and X and partitioned the variation in X into two parts – the portion that Z *can* explain (the $\hat{X}_i$) and the portion that Z *cannot* explain (the $\hat{e}_i$). Hold this thought.

We can do the exact same thing for the relationship between Z and Y that we just did for the relationship between Z and X. The process will look quite similar, with a bit of different notation to distinguish the processes. So we can model Y as a function of Z in the following way:

$$Y_i = \hat{\alpha}^* + \hat{\beta}^* Z_i + \hat{v}_i.$$

Here, the estimated slope is $\hat{\beta}^*$ and the error term is represented by $\hat{v}_i$.

Just as we did with Z and X, if we use Z to predict Y, then the predicted value of Y (or $\hat{Y}$) (which we will label $\hat{Y}^*$) based on Z is simply

$$\hat{Y}_i^* = \hat{\alpha}^* + \hat{\beta}^* Z_i,$$

which, as before, is identical to the preceding equation, but without the error term, because the residuals are expected (on average) to be zero. And again, as before, we can substitute the left-hand side of the preceding equation into the previous equation, and get

$$Y_i = \hat{Y}_i^* + \hat{v}_i$$

or, equivalently,

$$\hat{v}_i = Y_i - \hat{Y}_i^*.$$

These $\hat{v}_i$, then, are interpreted in an identical way to that of the preceding $\hat{e}_i$. They represent the portion of the variation in Y that Z cannot explain. (The portion of Y that Z *can* explain is the predicted portion – the $\hat{Y}_i^*$.)

Now what has this accomplished? We have just documented the relationship between Z and Y and partitioned the variation in Y into two parts – the portion that Z *can* explain and the portion that Z *cannot* explain.

So we have now let Z try to explain X and found the residuals (the $\hat{e}_i$ values); similarly, we have also now let Z try to explain Y, and found those residuals as well (the $\hat{v}_i$ values). Now back to our three-variable regression model that we have seen before, with Y as the dependent variable, and X and Z as the independent variables:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{u}_i.$$

The formula for $\hat{\beta}_1$, representing the effect of X on Y while controlling for Z, is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{e}_i \hat{v}_i}{\sum_{i=1}^{n} \hat{e}_i^2}.$$

Now, we know what $\hat{e}_i$ and $\hat{v}_i$ are from the previous equations. So, substituting, we get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \hat{X}_i)(Y_i - \hat{Y}_i^*)}{\sum_{i=1}^{n} (X_i - \hat{X}_i)^2}.$$

Pay careful attention to this formula. The "hatted" components in these expressions are from the two-variable regressions involving Z that we previously learned about. The key components of the formula for the effect of X on Y, while controlling for Z, are the $\hat{e}_i$ and $\hat{v}_i$, which, as we just learned, are the portions of X and Y (respectively) that Z cannot account for. And that is how, in the multiple regression model, the parameter β_1, which represents the effects of X on Y, *controls for* the effects of Z. How? Because the only components of X and Y that it uses are components that Z cannot account for – that is, the $\hat{e}_i$ and $\hat{v}_i$.

Comparing this formula for estimating β_1 with the two-variable formula for estimating β is very revealing. Instead of using the factors $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ in the numerator, which were the components of the *two-variable* regression of Y on X from Chapter 8, in the multiple regression formula that controls for Z the factors in the numerator are $(X_i - \hat{X}_i)$ and $(Y_i - \hat{Y}_i^*)$, where, again, the hatted portions represent X as predicted by Z and Y as predicted by Z.

Note something else in the comparison of the two-variable formula for estimating β and the multiple regression formula for estimating β_1. The result of $\hat{\beta}$ in the two-variable regression of Y and X and $\hat{\beta}_1$ in the three-variable regression of Y on X while controlling for Z will be different almost all the time. In fact, it is quite rare – though mathematically possible in theory – that those two values will be identical.[3]

And the formula for estimating β_2, likewise, represents the effects of Z on Y while controlling for the effects of X. These processes, in fact, happen simultaneously.

It's been a good number of chapters – five of them, to be precise – between the first moment when we discussed the importance of controlling for Z and the point, just above, when we showed you precisely how to do it. The fourth causal hurdle has never been too far from front-and-center since Chapter 3, and now you know the process of crossing it.

Don't get too optimistic too quickly, though. As we noted, the three-variable setup we just mentioned can easily be generalized to more than three variables. But the formula for estimating β_1 controls only for the effects of the Z variable that are included in the regression equation. It does not control for other variables that are not measured and not included in the model. And what happens when we fail to include a relevant cause of Y in our regression model? Bad things. Those bad things will come into focus a bit later in this chapter. Next, we turn to the issues of how to interpret regression tables using our running example of U.S. presidential elections.

9.4 INTERPRETING MULTIPLE REGRESSION

For an illustration of how to interpret multiple regression coefficients, let's return to our example from Chapter 8, in which we showed you the results of a regression of U.S. presidential election results on the previous year's growth rate in the U.S. economy (see Figure 8.4). The model we estimated, you will recall, was Vote $= \alpha + (\beta \times \mathit{Growth})$, and the estimated coefficients there were $\hat{\alpha} = 51.51$ and $\hat{\beta} = 0.62$. For the purposes of this example, we need to drop the observation from the presidential election of 1876. Doing this changes our estimates slightly so that $\hat{\alpha} = 51.67$ and $\hat{\beta} = 0.65$.[4] Those results are in column A of Table 9.1.

[3] Later in this chapter, you will see that there are two situations in which the two-variable and multiple regression parameter estimates of β will be the same.

[4] We had to drop 1876 because Fair's data do not include a measure for the new variable that we are adding in this example, "Good News," for that year. As we discuss in more detail in Section 12.4.1, when making comparisons across different models of the same data, it is extremely important to have exactly the same cases.

Table 9.1. Three regression models of U.S. presidential elections

	A	B	C
Growth	0.65*	—	0.57*
	(0.16)	—	(0.15)
Good News	—	0.92*	0.67*
	—	(0.33)	(0.28)
Intercept	51.67*	47.29*	48.24*
	(0.86)	(1.94)	(1.64)
R^2	0.36	0.20	0.46
n	33	33	33

Notes: The dependent variable is the percentage of the two major parties' vote for the incumbent party's candidate. Standard errors are in parentheses.
$* = p < 0.05$ (two-tailed t-test).

In column A, you see the parameter estimates for the annual growth rate in the U.S. economy (in the row labeled "Growth"), and the standard error of that estimated slope, 0.16. In the row labeled "Intercept," you see the estimated y-intercept for that regression, 51.67, and its associated standard error, 0.86. Both parameter estimates are statistically significant, as indicated by the asterisk and the note at the bottom of the table.

Recall that the interpretation of the slope coefficient in a two-variable regression indicates that, for every one-unit increase in the independent variable, we expect to see β units of change in the dependent variable. In the current context, $\hat{\beta} = 0.65$ means that for every extra one percentage point in growth rate in the U.S. economy, we expect to see, on average, an extra 0.65% in the vote percentage for the incumbent party in presidential elections.

But recall our admonition, throughout this book, about being too quick to interpret any bivariate analysis as evidence of a causal relationship. We have not shown, in column A of Table 9.1, that higher growth rates in the economy *cause* incumbent-party vote totals to be higher. To be sure, the evidence in column A is consistent with a causal connection, but it does not *prove* it. Why not? Because we have not controlled for other possible causes of election outcomes. Surely there are other causes, in addition to how the economy has (or has not) grown in the last year, of how well the incumbent party will fare in a presidential election. Indeed, we can even imagine other *economic* causes that might bolster our statistical explanation of presidential elections.[5]

Consider the fact that the growth variable accounts for economic growth over the past year. But perhaps the public rewards or punishes the incumbent party for *sustained* economic growth over the long run. In particular, it does not necessarily make sense for the public to reelect a party that has presided over three years of subpar growth in the economy but a fourth year with solid growth. And yet, with our single measure of growth, we are assuming – rather unrealistically – that the public would pay attention to the growth rate only in the past year. Surely the public does pay attention to recent growth, but the public might also pay heed to growth over the long run.

In column B of Table 9.1, we estimate another two-variable regression model, this time using the number of consecutive quarters of strong economic growth leading up to the presidential election – the variable is labeled "Good News" – as our independent variable.[6] (Incumbent-Party Vote Share remains our dependent variable.) In the row labeled "Good News," we see that the parameter estimate is 0.92, which means that, on average, for every additional consecutive quarter of good economic news, we expect to see a 0.92% increase in incumbent-party vote share. The coefficient is statistically significant.

Our separate two-variable regressions each show a relationship between the independent variable in the particular model and incumbent-party vote shares. But none of the parameter estimates in columns A or B controls for the other independent variable. We rectify that situation in column C, in which we estimate the effects of both the Growth and Good News variables on vote shares simultaneously.

Compare column C with columns A and B. In the row labeled "Good News," we see that the estimated parameter of $\hat{\beta} = 0.67$ indicates that, for every extra quarter of a year with strong growth rates, the incumbent party should expect to see an additional 0.67% of the national vote share, *while controlling for the effects of Growth*. Note the additional clause in the interpretation as well as the emphasis that we place on it. Multiple regression coefficients always represent the effects of a one-point increase in that particular independent variable on the dependent variable, *while controlling for the effects of all other independent variables in the model*. The higher the

[5] And, of course, we can imagine variables relating to success or failure in foreign policy, for example, as other, noneconomic causes of election outcomes.

[6] Fair's operationalization of this variable is "the number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2 percent."

number of quarters of continuous strong growth in the economy, the higher the incumbent-party vote share should be in the next election, controlling for the previous year's growth rate.

But, critical to this chapter's focus on multiple regression, notice in column C how including the "Good News" variable changes the estimated effect of the "Growth" variable from an estimated 0.65 in column A to 0.57 in column C. The effect in column C is different because it *controls for the effects of Good News*. That is, when the effects of long-running economic expansions are controlled for, the effects of short-term growth falls a bit. The effect is still quite strong and is still statistically significant, but it is more modest once the effects of long-term growth are taken into account.[7] Note also that the R^2 statistic rises from .36 in column A to .46 in column C, which means that adding the "Good News" variable increased the proportion of the variance of our dependent variable that we have explained by 10%.[8]

In this particular example, the whole emphasis on controlling for other causes might seem like much ado about nothing. After all, comparing the three columns in Table 9.1 did not change our interpretation of whether short-term growth rates affect incumbent-party fortunes at the polls. But we didn't know this until we tested for the effects of long-term growth. And later in this chapter, we will see an example in which controlling for new causes of the dependent variable substantially changes our interpretations about causal relationships. We should be clear about one other thing regarding Table 9.1: Despite controlling for another variable, we still have a ways to go before we can say that we've controlled for all other possible causes of the dependent variable. As a result, we should be cautious about

[7] And we can likewise compare the bivariate effect of Good News on Vote shares in column B with the multivariate results in column C, noting that the effect of Good News, in the multivariate context, appears to have fallen by approximately one-fourth.

[8] It is important to be cautious when reporting contributions to R^2 statistics by individual independent variables, and this table provides a good example of why this is the case. If we were to estimate Model A first and C second, we might be tempted to conclude that Growth explains 36% of Vote and Good News explains 10%. But if we estimated Model B and then C, we might be tempted to conclude that Growth explains 26% of Vote and Good News explains 20%. Actually, both of these sets of conclusions are faulty. The R^2 is always a measure of the overall fit of the model to the dependent variable. So, all that we can say about the R^2 for Model C is that Growth, Good News, and the intercept term together explain 46% of the variation in Vote. So, although we can talk about how the addition or subtraction of a particular variable to a model increases or decreases the model's R^2, we should not be tempted to attribute particular values of R^2 to specific independent variables. If we examine Figure 9.1, we can get some intuition on why this is the case. The R^2 statistic for the model represented in this figure is $\frac{f+d+b}{a+f+d+b}$. It is the presence of area d that confounds our ability to make definitive statements about the contribution of individual variables to R^2.

interpreting those results as proof of causality. However, as we continue to add independent variables to our regression model, we inch closer and closer to saying that we've controlled for every other possible cause that comes to mind. Recall that, all the way back in Chapter 1, we noted that one of the "rules of the road" of the scientific enterprise is to always be willing to consider new evidence. New evidence – in the form of controlling for other independent variables – can change our inferences about whether any particular independent variable is causally related to the dependent variable.

9.5 WHICH EFFECT IS "BIGGEST"?

In the preceding analysis, we might be tempted to look at the coefficients in column C of Table 9.1 for Growth (0.57) and for Good News (0.67) and conclude that the effect for Good News is roughly one-third larger than the effect for Growth. As tempting as such a conclusion might be, it must be avoided for one critical reason: The two independent variables are measured in different metrics, which makes that comparison misleading. Short-run growth rates are measured in a different metric – ranging from negative numbers for times during which the economy shrunk, all the way through stronger periods during which growth exceeded 5% per year – than are the number of quarters of consecutive strong growth – which ranges from 0 in the data set through 10. That makes comparing the coefficients misleading.

Because the coefficients in Table 9.1 each exist in the native metric of each variable, they are referred to as **unstandardized coefficients**. Although they are normally not comparable, there is a rather simple method to remove the metric of each variable to make them comparable with one another. As you might imagine, such coefficients, because they are on a standardized metric, are referred to as **standardized coefficients**. We compute them, quite simply, by taking the unstandardized coefficients and taking out the metrics – in the forms of the standard deviations – of both the independent and dependent variables:

$$\hat{\beta}_{\text{Std}} = \hat{\beta}\frac{s_X}{s_Y},$$

where $\hat{\beta}_{\text{Std}}$ is the standardized regression coefficient, $\hat{\beta}$ is the unstandardized coefficient (as in Table 9.1), and s_X and s_Y are the standard deviations of X and Y, respectively. The interpretation of the standardized coefficients changes, not surprisingly. Whereas the unstandardized coefficients represent the expected change in Y given a one-unit increase in X, the standardized coefficients represent the expected *standard deviation* change in

Y given a *one-standard-deviation* increase in X. Now, because all parameter estimates are in the same units – that is, the standard deviations – they become comparable.

Implementing this formula for the unstandardized coefficients in column C of Table 9.1 produces the following results. First, for Growth,

$$\hat{\beta}_{\text{Std}} = 0.5704788 \left(\frac{5.496239}{6.01748} \right) = 0.52.$$

Next, for Good News,

$$\hat{\beta}_{\text{Std}} = 0.673269 \left(\frac{2.952272}{6.01748} \right) = 0.33.$$

These coefficients would be interpreted as follows: For a one-standard-deviation increase in Growth, on average, we expect a 0.52-standard-deviation increase in the incumbent-party vote share, controlling for the effect of Good News. And for a one-standard-deviation increase in Good News, we expect to see, on average, a 0.33-standard-deviation increase in the incumbent-party vote shares, controlling for the effect of Growth. Note how, when looking at the unstandardized coefficients, we might have mistakenly thought that the effect of Good News was larger than the effect of Growth. But the standardized coefficients (correctly) tell the opposite story: The estimated effect of Growth is just over 150% of the size of the effect of Good News.[9]

9.6 STATISTICAL AND SUBSTANTIVE SIGNIFICANCE

Related to the admonition about which effect is "biggest," consider the following, seemingly simpler, question: Are the effects found in column C of Table 9.1 "big"? A tempting answer to that question is "Well of course they're big. Both coefficients are statistically significant. Therefore, they're big."

That logic, although perhaps appealing, is faulty. Recall the discussion from Chapter 6 (specifically, Subsection 6.4.2) on the effects of sample size on the magnitude of the standard error of the mean. And we noted the same

[9] Some objections have been raised about the use of standardized coefficients (King 1986). From a technical perspective, because standard deviations can differ across samples, this makes the results of standardized coefficients particularly sample specific. Additionally, and from a broader perspective, one-unit or one-standard-deviation shifts in different independent variables have different substantive meanings regardless of the metrics in which the variables are measured. We might therefore logically conclude that there isn't much use in trying to figure out which effect is biggest.

effects of sample size on the magnitude of the standard error of our regression coefficients (specifically, Section 8.4). What this means is that, even if the strength of the relationship (as measured by our coefficient estimates) remains constant, by merely increasing our sample size we can affect the statistical significance of those coefficients. Why? Because statistical significance is determined by a t-test (see Subsection 8.4.7) in which the standard error is in the denominator of that quotient. What you can remember is that larger sample sizes will shrink standard errors and therefore make finding statistically significant relationships more likely.[10] It is also apparent from Appendix B that, when the number of degrees of freedom is greater, it is easier to achieve statistical significance.

We hope that you can see that arbitrarily increasing the size of a sample, and therefore finding statistically significant relationships, does not in any way make an effect "bigger" or even "big." Recall, such changes to the standard errors have no bearing on the rise-over-run nature of the slope coefficients themselves.

How, then, should you judge whether an effect of one variable on another is "big?" One way is to use the method just described – using standardized coefficients. By placing the variances of X and Y on the same metric, it is possible to come to a judgment about how big an effect is. This is particularly helpful when the independent variables X and Z, or the dependent variable Y, or both, are measured in metrics that are unfamiliar or artificial.

When the metrics of the variables in a regression analysis are intuitive and well known, however, rendering a judgment about whether an effect is large or small becomes something of a matter of interpretation. For example, in Chapter 11, we will see an example relating the effects of changes in the unemployment rate (X) on a president's approval rating (Y). It is very simple to interpret that a slope coefficient of, say, −1.51, means that, for every additional point of unemployment, we expect approval to go down by 1.51 points, controlling for other factors in the model. Is that effect large, small, or moderate? There is something of a judgment call to be made here, but at least in this case, the metrics of both X and Y are quite familiar; most people with even an elementary familiarity with politics will need no explanation as to what unemployment rates mean or what approval polls mean. Independent of the statistical significance of that estimate – which, you should note, we have not mentioned here – discussions of this sort represent attempts to judge the **substantive significance** of a

[10] To be certain, it's not always possible to increase sample sizes, and, even when possible, it is nearly always costly to do so. The research situations in which increasing sample size is most likely, albeit still expensive, is in mass-based survey research.

coefficient estimate. Substantive significance is more difficult to judge than statistical significance because there are no numeric formulae for making such judgments. Instead, substantive significance is a judgment call about whether or not statistically significant relationships are "large" or "small" in terms of their real-world impact.

From time to time we will see a "large" parameter estimate that is not statistically significant. Although it is tempting to describe such a result as substantively significant, it is not. We can understand this by thinking about what it means for a particular result to be statistically significant. As we discussed in Chapter 8, in most cases we are testing the null hypothesis that the population parameter is equal to zero. In such cases, even when we have a large parameter estimate, if it is statistically insignificant this means that it is not statistically distinguishable from zero. Therefore a parameter estimate can be substantively significant only when it is also statistically significant.

9.7 WHAT HAPPENS WHEN WE FAIL TO CONTROL FOR Z?

Controlling for the effects of other possible causes of our dependent variable Y, we have maintained, is critical to making the correct causal inferences. Some of you might be wondering something like the following: "How does omitting Z from a regression model affect my inference of whether X causes Y? Z isn't X, and Z isn't Y, so why should omitting Z matter?"

Consider the following three-variable regression model involving our now-familiar trio of X, Y, and Z:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_i.$$

And assume, for the moment, that this is the *correct* model of reality. That is, the only systematic causes of Y are X and Z; and, to some degree, Y is also influenced by some random error component, u.

Now let's assume that, instead of estimating this correct model, we fail to estimate the effects of Z. That is, we estimate

$$Y_i = \alpha + \beta_1^* X_i + u_i^*.$$

As we previously hinted, the value of β_1 in the correct, three-variable equation and the value of β_1^* will not be identical under most circumstances. (We'll see the exceptions in a moment.) And that, right there, should be enough to raise red flags. For, if we know that the three-variable model is the *correct* model – and what that means, of course, is that the estimated value of β_1 that we obtain from the data will be equal to the true population value – and if we know that β_1 will not be equal to β_1^*, then there is a problem with the estimated value of β_1^*. That problem is a statistical problem called **bias**, which means that the expected value of the parameter estimate that we obtain from a sample will not be equal to the true population parameter. The specific type of bias that results from the failure to include a variable that belongs in our regression model is called **omitted-variables bias**.

Let's get specific about the nature of omitted-variables bias. If, instead of estimating the true three-variable model, we estimate the incorrect two-variable model, the formula for the slope β_1^* will be

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Notice that this is simply the bivariate formula for the effect of X on Y. (Of course, the model we just estimated is a bivariate model, in spite of the fact that we know that Z, as well as X, affects Y.) But because we know that Z *should* be in the model, and we know from Chapter 8 that regression lines travel through the mean values of each variable, we can figure out that the following is true:

$$(Y_i - \bar{Y}) = \beta_1(X_i - \bar{X}) + \beta_2(Z_i - \bar{Z}) + (u_i - \bar{u}).$$

We can do this because we know that the plane will travel through each variable's mean.

Now notice that the left-hand side of the preceding equation – the $(Y_i - \bar{Y})$ – is identical to one portion of the numerator of the slope for $\hat{\beta}_1^*$. Therefore we can substitute the right-hand side of the preceding equation – yes, that entire mess – into the numerator of the formula for $\hat{\beta}_1^*$.

The resulting math isn't anything that is beyond your skills in algebra, but it is cumbersome, so we won't derive it here. After a few lines of multiplying and reducing, though, the formula for $\hat{\beta}_1^*$ will reduce to

$$E(\hat{\beta}_1^*) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

This might seem like a mouthful – a fact that's rather hard to deny – but there is a very important message in it. What the equation says is that the estimated effect of X on Y, $\hat{\beta}_1^*$, in which we do not include the effects of Z on Y (but should have), will be equal to the true β_1 – that is, the effect with Z taken into account – plus a bundle of other stuff. That other stuff, strictly speaking, is bias. And because this bias came about as a result of omitting a variable (Z) that should have been in the model, this type of bias is known as omitted-variables bias.

Obviously, we'd like the expected value of our $\hat{\beta}_1^*$ (estimated without Z) to equal the true β_1 (as if we had estimated the equation with Z). And if the product on the right-hand side of the "+" sign in the preceding equation equals zero, it will. When will that happen?[11] In two circumstances, neither of which is particularly likely. First, $\hat{\beta}_1^* = \beta_1$ if $\beta_2 = 0$. Second, $\hat{\beta}_1^* = \beta_1$ if the large quotient at the end of the equation – the $\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ – is equal to zero. What is that quotient? It should look familiar; in fact, it is the bivariate slope parameter of a regression of Z on X.

In the first of these two special circumstances, the bias term will equal zero if and only if the effect of Z on Y – that is, the parameter β_2 – is zero. Okay, so it's safe to omit an independent variable from a regression equation if it has no effect on the dependent variable. (If that seems obvious to you, good.) The second circumstance is a bit more interesting: It's safe to omit an independent variable Z from an equation if it is entirely unrelated to the other independent variable X. Of course, if we omit Z in such circumstances, we'll still be deprived of understanding how Z affects Y; but at least, so long as Z and X are absolutely unrelated, omitting Z will not adversely affect our estimate of the effect of X on Y.[12]

We emphasize that this second condition is unlikely to occur in practice. Therefore, if Z affects Y, and Z and X are related, then if we omit Z from our model, our bias term will not equal zero. In the end, omitting Z will cause us to misestimate the effect of X on Y.

This result has many practical implications. Foremost among them is the fact that, even if you aren't interested theoretically in the connection between Z and Y, you need to control for it, statistically, in order to get an unbiased estimate of the impact of X, which is the focus of the theoretical investigation.

That might seem unfair, but it's true. If we estimate a regression model that omits an independent variable (Z) that belongs in the model, then the effects of that Z will somehow work their way into the parameter estimates for the independent variable that we do estimate (X) and pollute our estimate of the effect of X on Y.

The preceding equation also suggests when the magnitude of the bias is likely to be large and when it is likely to be small. If either or both of the components of the bias term [β_2 and $\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$] are *close to* zero, then the bias is likely to be small (because the bias term is the product of both components); but if both are likely to be large, then the bias is likely to be quite large.

[11] To be very clear, for a mathematical product to equal zero, either one or both of the components must be zero.

[12] Omitting Z from our regression model also drives down the R^2 statistic.

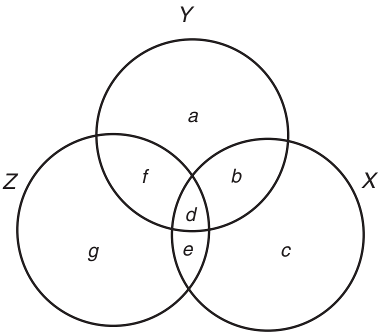

Figure 9.1. Venn diagram in which X, Y, and Z are correlated.

Moreover, the equation also suggests the likely *direction* of the bias. All we have said thus far is that the coefficient $\hat{\beta}_1^*$ will be biased – that is, it will not equal its true value. But will it be too large or too small? If we have good guesses about the values of β_2 and the correlation between X and Z – that is, whether or not they are positive or negative – then we can suspect the direction of the bias. For example, suppose that β_1, β_2, and the correlation between X and Z are all positive. That means that our estimated coefficient $\hat{\beta}_1^*$ will be larger than it is supposed to be, because a positive number plus the product of two positive numbers will be a still-larger positive number. And so on.[13]

To better understand the importance of controlling for other possible causes of the dependent variable and the importance of the relationship (or the lack of one) between X and Z, consider the following graphical illustrations. In Figure 9.1, we represent the total variation of Y, X, and Z each with a circle.[14] The covariation between any of these two variables – or among all three – is represented by the places where the circles overlap. Thus, in the figure, the total variation in Y is represented as the sum of the area $a+b+d+f$. The covariation between Y and X is represented by the area $b+d$.

Note in the figure, though, that the variable Z is related to both Y and X (because the circle for Z overlaps with both Y and X). In particular, the relationship between Y and Z is accounted for by the area $f+d$, and the relationship between Z and X is accounted for by the area $d+e$. As we have already seen, d is also a portion of the relationship between Y and X. If, hypothetically, we erased the circle for Z from the figure, we would (incorrectly) attribute all of the area $b+d$ to X, when in fact the d portion of the variation in Y is shared by *both* X and Z. This is why, when Z is related to both X and Y, if we fail to control for Z, we will end up with a biased estimate of X's effect on Y.

Consider the alternative scenario, in which both X and Z affect Y, but X and Z are completely unrelated to one another. That scenario is portrayed

[13] With more than two independent variables, it becomes more complex to figure out the direction of the bias.

[14] Recall from Chapter 8 how we introduced Venn diagrams to represent variation (the circles) and covariation (the overlapping portion of the circles).

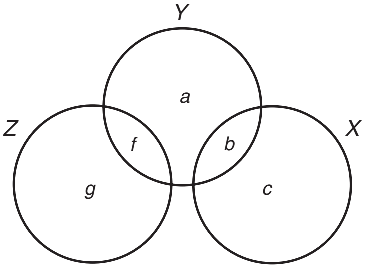

Figure 9.2. Venn diagram in which X and Z are correlated with Y, but not with each other.

graphically in Figure 9.2. There, the circles for both X and Z overlap with the circle for Y, but they do not overlap at all with one another. In that case – which, we have noted, is unlikely in applied research – we can safely omit consideration of Z when considering the effects of X on Y. In that figure, the relationship between X and Y – the area b – is unaffected by the presence (or absence) of Z in the model.[15]

9.7.1 An Additional Minimal Mathematical Requirement in Multiple Regression

We outlined a set of assumptions and minimal mathematical requirements for the two-variable regression model in Chapter 9. In multiple regression, all of these assumptions are made and all of the same minimal mathematical requirements remain in place. In addition to those, however, we need to add one more minimal mathematical requirement to be able to estimate our multiple regression models: It must be the case that *there is no exact linear relationship* between any two or more of our independent variables (which we have called X and Z). This is also called the assumption of no **perfect multicollinearity** (by which we mean that X and Z cannot be *perfectly* collinear, with a correlation coefficient of $r = 1.0$).

What does it mean to say that X and Z cannot exist in an exact linear relationship? Refer back to Figure 9.1. If X and Z had an *exact* linear relationship, instead of having some degree of overlap – that is, some imperfect degree of correlation – the circles would be exactly on top of one another. In such cases, it is literally impossible to estimate the regression model, as separating out the effects of X on Y from the effects of Z on Y is impossible.

This is not to say that we must assume that X and Z are entirely uncorrelated with one another (as in Figure 9.2). In fact, in almost all applications, X and Z will have some degree of correlation between them. Things become complicated only as that correlation approaches 1.0; and when it hits 1.0, the regression model will fail to be estimable with both X and Z as independent variables. In Chapter 10 we will discuss these issues further.

[15] For identical reasons, we could safely estimate the effect of Z on Y – the area f – without considering the effect of X.

9.8 AN EXAMPLE FROM THE LITERATURE: COMPETING THEORIES OF HOW POLITICS AFFECTS INTERNATIONAL TRADE

What are the forces that affect international trade? Economists have long noted that there are economic forces that shape the extent to which two nations trade with one another.[16] The size of each nation's economy, the physical distance between them, and the overall level of development have all been investigated as economic causes of trade.[17] But in addition to economic forces, does politics help to shape international trade?

Morrow, Siverson, and Tabares (1998) investigate three competing (and perhaps complementary) political explanations for the extent to which two nations engage in international trade. The first theory is that states with friendly relations are more likely to trade with one another than are states engaged in conflict. Conflict, in this sense, need not be militarized disputes (though it may be).[18] Conflict, they argue, can dampen trade in several ways. First, interstate conflict can sometimes produce embargoes (or prohibitions on trade). Second, conflict can reduce trade by raising the risks for firms that wish to engage in cross-border trading.

The second theory is that trade will be higher when both nations are democracies and lower when one (or both) is an autocracy.[19] Because democracies have more open political and judicial systems, trade should be higher between democracies because firms in one country will have greater assurance that any trade disputes will be resolved openly and fairly in courts to which they have access. In contrast, firms in a democratic state may be more reluctant to trade with nondemocratic countries, because it is less certain how any disagreements will be resolved. In addition, firms may be wary of trading with nondemocracies for fear of having their assets seized by the foreign government. In short, trading with an autocratic government should raise the perceived risks of international trade.

The third theory is that states that are in an alliance with one another are more likely to trade with one another than are states that are not in

[16] Theories of trade and, indeed, many theories about other aspects of international trade are usually developed with pairs of nations in mind. Thus all of the relevant variables, like trade, are measured in terms of pairs of nations, which are often referred to as "dyads" by international relations scholars. The resulting **dyadic data** sets are often quite large because they encompass each relevant pair of nations.

[17] Such models are charmingly referred to as "gravity models," because, according to these theories, the forces driving trade resemble the forces that determine gravitational attraction between two physical objects.

[18] See Pollins (1989) for an extended discussion of this theory.

[19] See Dixon and Moon (1993) for an elaboration of this theory.

Table 9.2. Excerpts from Morrow, Siverson, and Tabares's table on the political causes of international trade

	A	B	C	D
Peaceful relations	1.12*	—	—	1.45*
	(0.22)	—	—	(0.37)
Democratic partners	—	1.18*	—	1.22*
	—	(0.12)	—	(0.13)
Alliance partners	—	—	0.29*	−0.50*
	—	—	(0.03)	(0.16)
GNP of exporter	0.67*	0.57*	0.68*	0.56*
	(0.07)	(0.07)	(0.07)	(0.08)
R^2	0.77	0.78	0.77	0.78
n	2631	2631	2631	2631

Note: Standard errors are in parentheses.
$* = p < 0.05$.
Other variables were estimated as a part of the regression model but were excluded from this table for ease of presentation.

such an alliance.[20] For states that are not in an alliance, one nation may be reluctant to trade with another nation if the first thinks that the gains from trade may be used to arm itself for future conflict. In contrast, states in an alliance stand to gain from each other's increased wealth as a result of trade.

To test these theories, Morrow, Siverson, and Tabares look at trade among all of the major powers in the international system – the United States, Britain, France, Germany, Russia, and Italy – during most of the twentieth century. They consider each pair of states – called *dyads* – separately and examine exports to each country on an annual basis.[21] Their dependent variable is the amount of exports in every dyadic relationship in each year.

Table 9.2 shows excerpts from the analysis of Morrow, Siverson, and Tabares.[22] In column A, they show that, as the first theory predicts,

[20] See Gowa (1989) and Gowa and Mansfield (1993) for an extended discussion, including distinctions between bipolar and multipolar organizations of the international system.

[21] This research design is often referred to as a time-series cross-section design, because it contains both variation between units and variation across time. In this sense, it is a hybrid of the two types of quasi-experiments discussed in Chapter 3.

[22] Interpreting the precise magnitudes of the parameter estimates is a bit tricky in this case, because the independent variables were all transformed by use of natural logarithms.

increases in interstate peace are associated with higher amounts of trade between countries, controlling for economic factors. In addition, the larger the economy in general, the more trade there is. (This finding is consistent across all estimation equations.) The results in column B indicate that pairs of democracies trade at higher rates than do pairs involving at least one nondemocracy. Finally, the results in column C show that trade is higher between alliance partners than between states that are not in an alliance with one another. All of these effects are statistically significant.

So far, each of the theories received at least some support. But, as you can tell from looking at the table, the results in columns A through C do not control for the other explanations. That is, we have yet to see results of a fully multivariate model, in which the theories can compete for explanatory power. That situation is rectified in column D, in which all three political variables are entered in the same regression model. There, we see that the effects of reduced hostility between states is actually enhanced in the multivariate context – compare the coefficient of 1.12 with the multivariate 1.45. Similarly, the effects of democratic trading partners remains almost unchanged in the fully multivariate framework. However, the effect of alliances changes. Before controlling for conflict and democracy, the effect of alliances was (as expected) positive and statistically significant. However, in column D, in which we control for conflict and democracy, the effect flips signs and is now *negative* (and statistically significant), which means that, when we control for these factors, states in an alliance are less (not more) likely to trade with one another.

The article by Morrow, Siverson, and Tabares represents a case in which synthesizing several competing explanations for the same phenomenon – international trade – produces surprising findings. By using a data set that allowed them to test all three theories simultaneously, Morrow, Siverson, and Tabares were able to sort out which theories received support and which did not.

9.9 IMPLICATIONS

What are the implications of this chapter? The key take-home point of this chapter – that failing to control for all relevant independent variables will often lead to mistaken causal inferences for the variables that do make it into our models – applies in several contexts. If you are reading a research article in one of your other classes, and it shows a regression analysis between two variables, but fails to control for the effects of some other possible cause of the dependent variable, then you have some reason to be skeptical about the reported findings. In particular, if you can think of another independent variable that is likely to be related to *both* the independent variable and the dependent variable, then the relationship that the article does show that fails to control for that variable is likely to be plagued with bias. And if that's the case, then there is substantial reason to doubt the findings. The findings *might* be right, but you can't know that from the evidence presented in the article; in particular, you'd need to control for the omitted variable to know for sure.

But this critical issue isn't just encountered in research articles. When you read a news article from your favorite media web site that reports a relationship between some presumed cause and some presumed effect – news articles don't usually talk about "independent variables" or "dependent variables" – but fails to account for some other cause that you can imagine might be related to both the independent and dependent variables, then you have reason to doubt the conclusions.

It might be tempting to react to omitted-variables bias by saying, "Omitted-variables bias is such a potentially serious problem that I don't want to use regression analysis." That would be a mistake. In fact, the logic of omitted-variables bias applies to any type of research, no matter what type of statistical technique used – in fact, no matter whether the research is qualitative or quantitative.

Sometimes, as we have seen, controlling for other causes of the dependent variable changes the discovered effects only at the margins. That happens on occasion in applied research. At other times, however, failure to control for a relevant cause of the dependent variable can have serious consequences for our causal inferences about the real world.

In Chapters 10 and 11, we present you with some crucial extensions of the multiple regression model that you are likely to encounter when consuming or conducting research.

CONCEPTS INTRODUCED IN THIS CHAPTER

- bias – a statistical problem that occurs when the expected value of the parameter estimate that we obtain from a sample will not be equal to the true population parameter.
- dyadic data – data that reflect the characteristics of pairs of spatial units and/or the relationships between them.
- omitted-variables bias – the specific type of bias that results from the failure to include a variable that belongs in our regression model.
- perfect multicollinearity – when there is an exact linear relationship between any two or more of a regression model's independent variables.

Table 9.3. Bias in $\hat{\beta}_1$ when the true population model is $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_i$ but we leave out Z

β_2	$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	Resulting bias in $\hat{\beta}_1$
0	+	?
0	–	?
+	0	?
–	0	?
+	+	?
–	–	?
+	–	?
–	+	?

- standardized coefficients – regression coefficients such that the rise-over-run interpretation is expressed in standard-deviation units of each variable.
- substantive significance – a judgment call about whether or not statistically significant relationships are "large" or "small" in terms of their real-world impact.
- unstandardized coefficients – regression coefficients such that the rise-over-run interpretation is expressed in the original metric of each variable.

EXERCISES

1. Identify an article from a prominent web site that reports a causal relationship between two variables. Can you think of another variable that is related to both the independent variable and the dependent variable? Print and turn in a copy of the article with your answers.
2. In Exercise 1, estimate the direction of the bias resulting from omitting the third variable.
3. Fill in the values in the third column of Table 9.3.
4. In your own research you have found evidence from a bivariate regression model that supports your theory that your independent variable X_i is positively related to your dependent variable Y_i (the slope parameter for X_i was statistically significant and positive when you estimated a bivariate regression model). You go to a research presentation in which other researchers present a theory that their independent variable Z_i is negatively related to their dependent variable Y_i. They report the results from a bivariate regression model in which the slope parameter for Z_i was statistically significant and negative. Your Y_i and their