3 What Can Go Wrong With Multiple Regression?

Any tool as widely used as multiple regression is bound to be frequently *mis*used. Nowadays, statistical packages are so user-friendly that anyone can perform a multiple regression with a few mouse clicks. As a result, many researchers apply multiple regression to their data with little understanding of the underlying assumptions or the possible pitfalls. Although the review process for scientific journals is supposed to weed out papers with incorrect or misleading statistical methods, it often happens that the referees themselves have insufficient statistical expertise or are simply too rushed to catch the more subtle errors. The upshot is that you need to cast a critical eye on the results of any multiple regression, especially those you run yourself.

Fortunately, the questions that you need to ask are neither extremely technical nor large in number. They do require careful thought, however, which explains why even experts occasionally make mistakes or overlook the obvious. Virtually all the questions have to do with situations where multiple regression is used to make causal inferences. The requirements for a model used only for predictions are much less severe. I'll start with the more important questions and progress to those of lesser importance. Each question will be illustrated with examples drawn from Chapter 2.

3.1. Are Important Independent Variables Left Out of the Model?

In evaluating any regression model, it's just as important to think about what's *not* in the model as what's in it. There are two possible reasons for putting a variable in a regression model:

- You want to know the effect of the variable on the dependent variable
- You want to control for the variable

Obviously, researchers will include variables that are the main focus of their study, but they may not be so careful about including important control variables. (Remember that the multiple regression model makes no distinction between the study variables and the control variables.)

What makes a control variable important? To answer this question, you need to answer two other questions:

- Does the variable have a causal effect on the dependent variable?
- Is the variable *correlated* with those variables whose effects are the focus of the study?

For a control variable to be considered important, the answers to both questions must be "yes." If it has a strong effect on the dependent variable but is unrelated to the independent variables already in the model, there's no need to include it. That's why multiple regression is not essential for most randomized experiments—the randomization process is designed to remove any correlation between the treatment variable and the characteristics of the individuals in the study. Similarly, a potential control variable may be highly correlated with one of the independent variables in the regression model, but if it has no impact on the dependent variable, it can safely be omitted.

What happens when an important control variable is omitted? In the language of statisticians, what happens is *bias*. Simply put, bias means that the estimates of the regression coefficients for variables in the model tend to be either too high or too low. If the bias is large enough, we may conclude that a variable has an effect on the dependent variable when it really doesn't, or we could conclude that it doesn't have an effect when it really does. In the language of social scientists, the conclusions are *spurious*.

These are all matters of degree. If the effect of the omitted variable on the dependent variable is small, then the bias will be small. If the correlation between the omitted variable and the included variable is small, the bias will also be small. The bias is also concentrated in the coefficients for those variables that are correlated with the omitted variable. Here's an example. Suppose you regress income (the dependent variable) on years of schooling and age, as in Chapter 1, and you find a large effect of schooling. A critic might say, "The schooling effect is biased because you didn't control for intelligence. Intelligent people spend more years in school and they also earn more money." On the other hand, we wouldn't expect much bias in the age coefficient because there's little correlation between IQ measures and age.

Now let's consider two of the examples in Chapter 2. Table 2.3 purports to show that living in cities makes people more racially tolerant. After controlling for several other variables, people in larger cities still had higher tolerance scores than those in smaller cities or rural areas. But were all the relevant variables included? Specifically, did the regression model omit any variables that are associated with both tolerance and urban residence? Religion is one possibility. It is well known that Jews tend to be concentrated in urban areas, and there is also a long history of Jewish alliance with organizations promoting racial equality. Perhaps if the regression model included a dummy variable for whether or not the respondent was Jewish, the coefficient for urbanism would no longer be statistically significant. My own guess is that the proportion of Jews in the United States is sufficiently small (only about 3%) that putting this variable in the model would not make a major difference, but you never know for sure until you try.

Table 2.2 is also problematic. In this regression, the author wanted to test the hypothesis that mental patients who believed that most people discriminate against mental patients would be more demoralized. He found a significant coefficient for devaluation-discrimination controlling for four variables: marital status, ethnicity (Hispanics vs. others), hospitalization, and diagnosis (schizophrenia vs. depression). He also reported that three other variables—sex, years of education, and employment status—were originally in the regression equation but were deleted because they were not statistically significant. Again, this is consistent with the principle that you can omit variables if they have no effect on the dependent variable.

Unfortunately, it's not all that difficult to think of other variables that might affect both discrimination beliefs and demoralization. One obvious possibility is actual *experience* of discrimination. Some of these mental patients may have been treated much worse than other patients—by family, friends, coworkers, or mental health practitioners. It wouldn't be surprising if they were more demoralized and also believed more strongly that people discriminated against mental patients. Another possibility is seriousness of illness, which is likely to affect both beliefs and demoralization. The author might claim that he controlled for seriousness by including hospitalization and diagnosis, but these are very crude proxies. I would expect considerable variation in seriousness of illness among hospitalized patients or among schizophrenics. Finally, there are many standard sociodemographic variables, such as age and income, that could affect both variables. In the light of these possibilities, I'm far from convinced that beliefs about discrimination have a causal impact on demoralization.

When you have a clear idea about how a particular omitted variable is related to the dependent variable and an included independent variable, you can often make a reasonable guess about the direction of the bias. For example, as a mental illness becomes more serious, we would expect a patient to become more demoralized and to have stronger beliefs of discrimination. That would produce a spuriously *positive* relationship between those two variables. If this argument is correct, the coefficient for devaluation-discrimination in Table 2.2 is probably an overestimate of the true coefficient.

3.2. Does the Dependent Variable Affect Any of the Independent Variables?

When we estimate a regression model, we often interpret the coefficients as measuring the causal effects of the independent variables on the dependent variable. But what if the "dependent" variable actually affects one or more of the "independent" variables? If it does, the resulting biases can be every bit as serious as those produced by the omission of important variables. This problem—known as *reverse causation*—actually can be worse than the omitted variables problem because

- Every coefficient in the regression model may be biased
- It's hard to design a study that will adequately solve this problem

Unfortunately, there's rarely any information in the data that can help you determine the direction of causation. Instead, decisions about the direction of causation have to be based almost entirely on your knowledge of the phenomenon you're studying. There are, in fact, several different ways to argue against the possibility of reverse causation. If the data come from a randomized experiment, then randomization assures us that the dependent variable isn't influencing who gets the treatment. Often, the time ordering of the variables gives us a pretty clear indication of the causal direction. For example, we usually feel safe in supposing that parents' educational attainment affects the educational attainment of their adult children, not the other way around. Even when there's no time ordering, our knowledge of basic physical and biological processes sometimes gives us a pretty good idea of the causal direction. We feel confident, for example, that a man's height might affect his social prestige, but social prestige couldn't affect height.

With non-experimental data, most applications of regression analysis involve some ambiguity about the direction of causality. In such cases, the causal mechanism could run in either direction, perhaps in both. Equation 2 of Table 2.1 includes "unable to work" as a predictor of self-rated health, but the predominant reason for inability to work is poor health, so the inclusion of this variable could potentially bias other coefficients in the model. It would have been preferable to remove people who were unable to work from the data set before running this regression.

The regression in Table 2.2 is also vulnerable to this criticism. The model presumes that beliefs about discrimination and devaluation affect the demoralization of mental patients, but isn't it equally plausible that demoralized patients are more likely to believe that mental patients are discriminated against? If so, then the coefficient for devaluation-discrimination in Table 2.2 could be highly biased.

Things are somewhat less ambiguous with the urbanism regression reported in Table 2.3. The core hypothesis is that living in a large urban area promotes racial tolerance. That certainly makes sense. One could also argue, however, that more racially tolerant people are more likely to move from non-urban to urban areas and that racially *in*tolerant people are more likely to move from urban to non-urban areas, thereby producing a relationship between urbanism and tolerance. Under this scenario, tolerance causes urbanism. One way to reduce the ambiguity is to focus on residence at age 16 rather than in adulthood, when the questionnaire was administered.

At age 16, people haven't yet had the opportunity to move to places that are more consistent with their attitudes. In fact, tables presented in the original article show that people living in urban areas at age 16 are more racially tolerant in adulthood that those living in non-urban areas at age 16, thereby increasing our confidence that urbanism causes tolerance rather than the reverse.

Table 2.4 purports to show that married couples spend more time together each day if they (a) work fewer hours, (b) work on the same schedule, and (c) have been married longer. What's the direction of causality here? In this case, as in many others, it's helpful to think about how people typically make decisions. Most people are very constrained with respect to the number of hours they work and the scheduling of those hours. Once those constraints are set, they then make decisions about how to spend their time away from work. This reasoning supports the causal ordering assumed by Table 2.4. On the other hand, there are undoubtedly some couples who choose their jobs or their job schedules to accommodate their desires to spend more (or less) time together. So there's still some ambiguity about the causal ordering. This is also true for the apparent effect of length of marriage on time spent together. Maybe it's the other way around: Perhaps couples who spend more time together stay married longer.

In the final example, Table 2.5, we find that the sex composition of occupations appears to affect men's job satisfaction, job-related depression, and job-related self-esteem. In this case, the direction of causality seems relatively unambiguous. If anything, I would expect that men with higher self-esteem are more willing to enter mixed occupations because they feel more confident in less traditional settings, but that would run contrary to the observed results.

It should be clear from these examples that most applications of multiple regression involve some uncertainty about the direction of causality. As with omitted variables, the issue here is one of degree. If we can make an equally persuasive argument for x causing y and for y causing x, then we must be very wary about interpreting a regression with y as the dependent variable and x as the independent variable. On the other hand, if we have a compelling argument for x affecting y but only a weak or contrived argument for y causing x, then we can feel much more secure about interpreting the regression of y on x. Even then, however, some uncertainty usually remains.

3.3. How Well Are the Independent Variables Measured?

As we've repeatedly seen, to control for a variable in a multiple regression, you have to measure that variable and include it in the model. But it's not enough simply to measure a variable—you have to measure it well. To the degree that a variable is afflicted with measurement error, the coefficients for that variable and other variables will be biased. I already hinted at this problem when I discussed Table 2.2, showing the relationship between demoralization and beliefs of discrimination among mental patients. I said that both variables could be affected by seriousness of the illness, which could produce a spurious relationship between them. Two variables in the model—hospitalization and diagnosis—could be taken as measures of seriousness, but only crude measures at best. Within each category of these two variables, there was undoubtedly much variation in seriousness. More generally, whenever an independent variable is measured with error, we can say that for each measured value of that variable, there may be many different values of the true variable. Consequently, the measured variable does not fully control for the true variable.

In a regression with a single independent variable, measurement error in that variable will tend to bias its coefficient toward zero. In models with more than one independent variable, the biases are harder to predict. Usually, however, the coefficients for the variables with a greater degree of measurement error tend to be underestimated (closer to zero), whereas the coefficients for the variables with less measurement error tend to be overestimated (further from zero).

For Equation 1 of Table 2.1, measurement error is not likely to be a serious problem. We can usually do a pretty good job of getting accurate information on years of schooling, sex, race, age, and marital status. However, questions could be raised about the marital status variable, which is a simple dichotomy of married vs. unmarried. Obviously, the unmarried category could be further decomposed into never married, divorced, and widowed. If there are substantial differences in the dependent variable (self-rated health) across these three subcategories, then marital status has not been adequately controlled.

Measurement error is potentially more serious in the study of mental patients (Table 2.2). I already discussed the inadequacy of hospitalization as a measure of seriousness. We can also raise questions about the key independent variable, the 12-item scale measuring beliefs about discrimination and devaluation. "Soft" measures like this one should always be scrutinized for possible measurement error. In defense of the scale, the author reports that it has an estimated *reliability* of .78. Reliability is a number between 0 and 1 that quantifies the degree of measurement error in a variable. The value of .78 tells us that 78% of the variance of this variable is "true" variance and the other 22% of the variance is measurement error. This is OK by most social science standards, but not terrific. If this were the only independent variable in the model, its coefficient would be underestimated by about 22%. This would bias the results *against* the author's hypothesis, so the fact that he *does* find an effect of this variable cannot be explained away by measurement error.

Estimates of reliability are valuable indications of the degree of measurement error, but you should never accept them as absolute truth. First, there are many different ways to estimate reliability, and different methods can produce markedly different answers. Second, conventional methods for estimating reliability tell us only about the *stability* of the variable in question, not how well it measures what we really want it to measure. Self-reports of age, for example, are highly reliable in the sense that people tend to give consistent answers if you ask them their age more than once, but most people would be highly skeptical of a study that used self-reported age as a measure of political conservatism. In short, we need to be concerned about the *validity* of our measurements, not just their reliability. Unfortunately, validity is a much more difficult notion to quantify, both in theory and in practice.

For a final example of measurement error, let's consider the study of gender segregation on the job (Table 2.5). Men were divided into five groups according to how much gender segregation they experienced at work. As the authors admit, "our sample permitted only indirect measurement of the gender mix in each respondent's work setting" (Wharton & Baron, 1987, p. 578). What this means is that they used census data on proportions of workers who were male and female for jobs of that particular type in that particular industry. Thus, they didn't know the actual degree of gender segregation experienced by each man in the sample. As a result, the coefficients they estimated for the various gender categories are probably smaller than if they had actually interviewed each man and asked him about the number of his coworkers who were men and women. Because the coefficients were still statistically significant, the evidence clearly supports the authors' hypotheses.

3.4. Is the Sample Large Enough to Detect Important Effects?

Sample size has a profound effect on tests of statistical significance. With a sample of 60 people, a correlation has to be at least .25 (in magnitude) to be significantly different from zero (at the .05 level). With a sample of 10,000 people, any correlation larger than .02 will be statistically significant. The reason is simple: There's very little information in a small sample, so estimates of correlations are very unreliable. If we get a correlation of .20, there may still be a good chance that the true correlation is zero. Similarly, large samples contain lots of information that allows us to estimate the correlation very precisely. Even if we get a correlation of only .04 in a sample of 10,000, the chances are slim that the true correlation is zero. This means that you should always keep sample size in mind when you're looking at the results of significance tests.

Let's first look at the problem of interpreting results from small samples. In Table 2.4, we saw that most of the independent variables were not significant predictors of the time that married couples spend together. For example, the coefficient for race was not significant, even though it was rather large. Are we justified, then, in concluding that there are no racial differences in the time couples spend together? Absolutely not. The sample consists of 177 couples, only five of which were nonwhite. Although this is not an extremely small sample, neither is it very large. It's quite possible that a sample of 1,000 cases would have provided clear evidence for a race difference. In the absence of such a sample, however, we just don't know whether race matters or not. The general principle is this: In a small sample, statistically significant coefficients should be taken seriously, but a nonsignificant coefficient is extremely weak evidence for the absence of an effect.

Statisticians often describe small samples as having *low power* to test hypotheses. There is another, entirely different problem with small samples that is frequently confused with the issue of power. Most of the test statistics that researchers use (such as *t* tests, *F* tests, and chi-square tests) are only approximations. These approximations are usually quite good when the sample is large but may deteriorate markedly when the sample is small. That means that *p* values calculated for small samples may be only rough approximations of the true *p* values. If the calculated *p* value is .02, the true value might be something like .08. In multiple regression, the *p* values are exact *if* the data satisfy certain demanding distributional assumptions (Chapter 6). That's a big "if," however, and in all other cases the statistics are only approximations.

Although these two problems—low power and poor approximations of test statistics—are entirely distinct, they sometimes become intertwined in practice in an unfortunate way. Because it's difficult to get statistically significant results in small samples, regression analysts are often tempted to raise the level of *p* values that they will accept as statistically significant, say from .05 to .10. But because the *p* values themselves may be poor approximations in small samples, I would argue for exactly the opposite approach. Instead of requiring a *p* value less than .05 for statistical significance, I think a good case can be made for a criterion closer to .01 to allow for the possibility that the *p* value is underestimated.

That brings us to the inevitable question: What's a big sample and what's a small sample? As you may have guessed, there's no clear-cut dividing line. Almost anyone would consider a sample less than 60 to be small, and virtually everyone would agree that a sample of 1,000 or more is large. In between, it depends on a lot of factors that are difficult to quantify, at least in practice.

3.5. Is the Sample So Large That Trivial Effects Are Statistically Significant?

In theory, a bigger sample is always better. With a large sample, you don't have to worry that the *p* values are only approximations, and bigger samples give you more precise estimates of the coefficients. What can be bad about that? In practice, however, large samples can sometimes lead to incorrect conclusions.

In any scientific investigation, there are bound to be some sources of bias. Perhaps the sample wasn't quite random, or maybe the respondents tended to overestimate their income. We do everything we can to minimize those biases, but we can never eliminate them entirely. Consequently, small, artifactual relationships are likely to creep into the data. A large sample is like a very sensitive measuring instrument. It's so sensitive that it detects these artifactual relationships along with the true relationships.

The essence of the problem, then, is that with very large samples, almost any variable you put in a regression model is likely to show up as statistically significant, even if it has no real effect. When a variable has a statistically significant coefficient, we say that its coefficient is unlikely to be zero. In a large sample, the coefficient may be so small that it's not worth serious attention.

In practice, this means that when a regression is run on a large sample, it's not enough to say that a coefficient is statistically significant. You must also determine whether it is *substantively* significant. That is, you must look carefully at the magnitude of the coefficient to see if it is large enough to have theoretical or practical importance.

Suppose, for example, that you run a regression with 10,000 adults in which the dependent variable is annual income (in dollars) and one of the independent variables is a dummy variable for gender, coded 1 for male and 0 for female. You find that the gender coefficient is positive and statistically significant at the .05 level. "Aha!" you say, "This is evidence for sex discrimination." Maybe so, but what if the coefficient is only 45.2, indicating that men, on average, make $45 a year more than women? Although there may be some people who would take this seriously, I can't imagine a reasonable person getting either upset with or intrigued by a difference that small.

In this example, income and gender are such familiar variables that it's fairly easy to get a sense of the substantive significance of the coefficients. However, with many of the variables used in social science research, the metrics are so unfamiliar that it's hard to get a clear notion of what's a big coefficient and what's small. In these situations, standardized coefficients—when available—can be helpful. As we've already seen, standardized coefficients do not depend on the metrics of the variables and therefore can be used to gauge the relative importance of different variables. If a standardized coefficient is very small, say less than .05, it tells us that the variable explains very little of the variation in the dependent variable, even though it may be statistically significant.

In the examples considered in Chapter 2, the largest sample (2,031 cases) was used for the self-reported health regression (Table 2.1). Many of the variables in this table had coefficients with *p* values less than .001, reflecting the large size of the sample. Should we take all these effects seriously? Consider the coefficient for marital status, which has a *p* value less than .01 in Equation 1. Although statistically significant, the difference between married and unmarried people is only about 0.1 on the 5-point scale of self-reported health. This does not seem like a very large effect. The standardized coefficient is only .058, which is pretty low as these things go, so it would probably be unwise to make a big deal out of the effect of marital status on self-rated health.

3.6. Do Some Variables Mediate the Effects of Other Variables?

Does socioeconomic status (SES) affect SAT scores? That seems reasonable. But suppose you run a regression model for a sample of college freshmen with SAT score as the dependent variable and SES and high school GPA as the independent variables. You find that the coefficient for GPA is highly significant but the coefficient for SES is not significant. Are you justified in concluding that SES has no effect on SAT scores? We already saw in Section 3.4 that if the sample is too small, a nonsignificant coefficient is not a sufficient reason for concluding that a variable has no effect on the dependent variable. Even if the sample is not small, there is another reason for being cautious in concluding that a variable has no effect: It's possible that other variables *mediate* the effect of that variable. If those other variables are also included in the regression model, the effect of the variable you're interested in may disappear.

In our hypothetical example, let's suppose that SES has a big effect on GPA, and GPA, in turn, has a big effect on SAT scores. We can represent this by a simple *path diagram*:

$$\text{SES} \rightarrow \text{GPA} \rightarrow \text{SAT}.$$

If this causal structure is correct, then it's appropriate to say that SES affects SAT scores because improving people's SES results in an increase in their SAT scores. But the effect is *indirect*, through improvements in high school scholastic achievement (GPA). We may also say that GPA is an *intervening* variable between SES and SAT. If we put both independent variables in the regression model, the overall effect of SES on SAT is obscured. We may mistakenly conclude that SES has no impact on SAT scores when it may actually be quite important.

The point is *not* that the regression of SAT on both GPA and SES is wrong; rather, the point is that any regression model only estimates the *direct* effect of each variable, controlling for all the other variables in the model. In our example, the direct effect of SES can be interpreted as its effect on SAT scores that *does not* operate through GPA. The regression doesn't tell you anything, however, about possible indirect effects. In many cases, these indirect effects may be of great interest.

How can you tell if a variable has important indirect effects? First, you need to have some clear notion of the causal ordering of the variables in the regression model. In our example, it's apparent that SES affects GPA rather than the reverse. Second, you need more information than you can get from just a single regression model. There are two ways to get this information, but I'll discuss only one of them here. In our example, if we remove GPA from the regression model, then the coefficient for SES is its *total effect* on SAT scores. In general,

$$\text{Total Effect} = \text{Direct Effect} + \text{Indirect Effects.}$$

Because we already have an estimate of the direct effect from the regression of SAT on both SES and GPA, we can calculate the indirect effect by subtracting the direct effect from the total effect. Suppose, for example, that when we regress SAT score on SES alone, the coefficient is 15 (that is, each one-unit increase in SES yields a 15-point increase in SAT score). When we regress SAT score on *both* SES and GPA, the coefficient of SAT declines to 5. Then the direct effect of SES is 5, and the indirect effect (through GPA) is 10.

The general principle is this: If you're interested in the effect of x on y, but there are other variables w and z that may mediate that effect, estimate the regression twice, both with and without the intervening variables w and z. The coefficient of x in the regression without w and z is the total effect of x. The difference in the two coefficients for x represents the indirect effects of x through w and z. Of course, if all you have are published regression results, you may not be able to do these calculations.

A couple of additional points should be keep in mind. First, the inclusion of an intervening variable in a regression model doesn't always eliminate the effect of the variable of interest. If the variable of interest also has a direct effect, then the inclusion of the intervening variable may only reduce its coefficient, not eliminate it. Second, for any causal relationship, it's nearly always possible to think of one or more intervening variables. Thus, what we call a direct effect in one model can potentially be converted into an indirect effect if we add mediating variables.

Knowledge of intervening variables is usually a crucial part of understanding how a causal process works. It's nice to know, for instance, that people in urban areas are more racially tolerant (Table 2.3). But why are they more tolerant? Is it because they have more face-to-face contact with people of different races? Or is it because teachers in urban schools put more stress on racial tolerance? The way to tell is to actually measure these variables and put them in the regression model. If the coefficient for urbanism goes down substantially, that's evidence that these variables mediate the urbanism effect.

Do mediating variables create any problems for the interpretation of the examples in Chapter 2? In the self-rated health regression (Table 2.1), the search for variables mediating the effect of education was a major aim of the study. To accomplish this goal, models were estimated both with and without the potential mediating variables to see how the coefficient for education changed. For the mental patient study (Table 2.2), it seems implausible that patients' beliefs about devaluation and discrimination could affect control variables like diagnosis, hospitalization, and so on, so there's not much cause for concern here. With the urbanism study (Table 2.3), there's a possibility that urban residence could be affecting education (which has a strong affect on tolerance) and income (which has a modest effect in two of the years). Hence, part of the effect of urban residence could be indirect, through income and education.

3.7. Are Some Independent Variables Too Highly Correlated?

I've encountered several people with the mistaken notion that the independent variables in a regression analysis should not be correlated with each other. I'm not sure where they got this idea—maybe because they're called *independent* variables. As we saw in Section 3.1, if the independent variables are all uncorrelated, you don't even need multiple regression: Two-variable regressions or simple correlations are sufficient.

Multiple regression is designed precisely for separating the effects of two or more independent variables on a dependent variable when the independent variables are correlated with one another, but there's a limit to what regression can do. Imagine that you have a sample of AIDS patients and you want to study what factors affect their survival time. You're particularly interested in homosexual orientation (1 = yes, 0 = no) and intravenous drug use (1 = yes, 0 = no). You find that in your sample all the IV drug users are heterosexual and all the nonusers are homosexual. Sorry, but you're out of luck. There's no way you can get separate estimates of the effects of these two variables on survival time, either with multiple regression or with any other statistical method.

This problem goes by the name of *multicollinearity*, and what we've just seen is an extreme case where the two variables are perfectly correlated. If you tried to put both these variables into a regression program, either you would get an error message or the program would kick one variable or the other out of the model.

Multicollinearity doesn't have to be so extreme to cause problems, and, unfortunately, those problems often go undetected. Suppose that in the AIDS example, there were two homosexual IV drug users out of 1,000 patients. The correlation isn't perfect anymore, so you could do the regression and you'd get results. Would the numbers mean anything? Not likely. What would happen is that the estimated standard errors for the regression coefficients would be very large, accurately reflecting the fact that you can't get very precise estimates of the effect of each variable controlling for the other variable. With large standard errors, it's unlikely that either coefficient would be statistically significant. On the other hand, it's entirely possible that if we ran the regression with only the IV drug use variable or, alternatively, only the homosexuality variable, we might find large and highly significant coefficients for both variables. This would tell us that one or both of these variables had a substantial effect on survival time, but we couldn't say which one. If we ran only the model that included both variables, we would incorrectly conclude that *neither* variable had any effect on survival time.

Multicollinearity has other effects, but this one is the most worrisome—the possibility of concluding that two variables have no effect when one or the other of them actually has a strong effect. How highly correlated do two (or more) variables have to be before multicollinearity becomes a problem? That's a fairly technical issue that I'm going to postpone until Chapter 7. You almost certainly have problems if the correlation is above .80, but there may be difficulties that appear well below that value.

Is multicollinearity a problem for any of the examples in Chapter 2? We can't be sure because the authors don't report the correlations among the independent variables. Just examining the variables, I would be surprised if any of them were so highly correlated as to cause errors in interpretation. One possible exception is in the example of marital time. Two of the variables in the regression model are number of children and presence of preschoolers. I would expect these variables to have a correlation of perhaps .50 and possibly larger. The fact that neither of them has a statistically significant coefficient could be a consequence of multicollinearity.

3.8. Is the Sample Biased?

This question needs to be asked for *any* statistical analysis, regardless of the method. There are actually two aspects to this question, however, relating to the notions of *internal* and *external validity*.

With respect to external validity, the question is whether the results for the sample can be generalized to other groups or populations of interest. If the regression is estimated with a simple random sample from some well-defined population, then you are in a good position to generalize from the sample to the population. If the sample is a probability sample but not a simple random sample, the regression may require weighting of some sort, although this is a controversial issue. The study of self-rated health in Section 2.1, the study of urbanism in Section 2.3, and the study of male workers in Section 2.5 were all based on probability samples of the adult U.S. population; this gives us some confidence in the generalizability of the conclusions.

If the sample is merely a convenience sample, then you need to think carefully about whether causal relationships that hold in the sample are likely to hold in other groups in which you may be interested. In the study of mental patients (Section 2.2), the sample was drawn entirely from

> outpatient clinics and inpatient clinics in the same general area [Washington Heights] of New York City. The goal was to select patients in two diagnostic

> categories, major depression and schizophrenia and schizophrenic-like psychiatric disorders. In addition, considerable effort was made to locate cases in their first episode of each of these types of disorders. (Link, 1987, pp. 100-101)

This is clearly not a probability sample: It is restricted to a specific geographic area, and it apparently involved a substantial amount of recruitment effort on the part of the research team. That should raise a number of questions about the generalizability of the results. Are mental patients in this section of New York City likely to be different with respect to the variables or relationships in this study from patients in other geographical areas? How were the clinics selected? Was there a substantial rate of refusal? Answers to these questions could raise doubts about how well the results of this study apply to other situations.

Leaving aside the question of whether the results for the sample apply to other groups as well, we must also consider internal validity—whether the sample selection process can produce erroneous or misleading results for the sample itself. Again, this is usually not a major concern if the sample is a probability sample from some well-defined population, and if the nonresponse rate is low. It should be a concern, however, for convenience samples, or when a high proportion of potential respondents refuse to cooperate or can’t be located. Imagine a study of the relationship between education and income in which people who were inconsistent on these two variables tended not to respond to the survey. That is, what if people with high education and low income, or low education and high income, were disproportionately less likely to return the questionnaire? Such a differential response pattern could easily produce an apparent correlation between education and income when there was really no relationship. In most applications, we don’t expect such differential response patterns, but that doesn’t mean that they don’t occur.

3.9. Are There Any Other Problems to Watch For?

The preceding eight problems are the ones I believe most often lead to serious errors in judging the results of a multiple regression. By no means do they exhaust the possible pitfalls that may arise. Before concluding this chapter, I’ll briefly mention a few others.

Some problems are implicit in the essential features of multiple regression that I discussed in Chapter 1. For example, the regression model assumes that the relationships are all linear. What if they're not? Modest departures from linearity are unlikely to cause really serious inferential errors. The linear model is usually a reasonably good approximation to a wide range of mathematical forms. If the relationship is highly nonlinear, however, we can sometimes be led astray. It may be, for instance, that punishment reduces deviant behavior up to a certain point, but if the punishment becomes too severe, deviant behavior may increase. If we apply a strictly linear model to the data, the positive and negative effects of deviance may cancel out, leaving us with no apparent relationship. Ways of extending the linear model to account for such *curvilinear* relationships will be discussed in Chapter 8.

Also in Chapter 1, we saw that variables measured on an ordinal scale are not strictly appropriate for multiple regression, even though many researchers use them anyway. Most of the time this practice is innocuous, but it can occasionally lead to errors or misinterpretations. The problem is actually quite similar to the discussion of inadequate measurement in Section 3.3. If we want to control for intelligence, for example, but we only have a crude ranking of intelligence, then our statistical control may not be very effective. As a result, we may be led to erroneous conclusions.

I'll mention one last problem that, unfortunately, is very hard to evaluate from published regression results. It sometimes happens that regression results are very strongly influenced by a small number of cases in the sample. These influential cases are usually those that have extreme values on either the independent or the dependent variables. There are graphical and numerical methods for detecting influential observations. One simple approach is to delete them from the analysis and see how much difference it makes. If regression results are highly sensitive to the presence or absence of a few influential cases, then we should be less confident in our results.

Chapter Highlights

1. Leaving important variables out of a regression model can bias the coefficients of other variables and lead to spurious conclusions.

2. Important variables are those that affect the dependent variable and are correlated with the variables that are the focus of the study.
3. If the dependent variable in a regression model has an effect on one or more independent variables, any or all of the regression coefficients may be seriously biased.
4. Non-experimental data rarely tell you anything about the direction of a causal relationship. You must decide the direction based on your prior knowledge of the phenomenon you're studying.
5. Time ordering usually gives us the most important clues about the direction of causality.
6. Measurement error in independent variables leads to bias in the coefficients. Variables with more measurement error tend to have coefficients that are biased toward 0. Variables with little or no measurement error tend to have coefficients that are biased away from 0.
7. The degree of measurement error in a variable is usually quantified by an estimate of its reliability, a number between 0 and 1. A reliability of 1 indicates that the variable is perfectly measured; a reliability of 0 indicates that the variation in the variable is pure measurement error.
8. With small samples, even large regression coefficients may not be statistically significant. In such cases, you are *not* justified in concluding that the variable has no effect—the sample may not have been large enough to detect it.
9. In small samples, the approximations used to calculate p values may not be very accurate, so be cautious in interpreting them.
10. In large samples, even trivial effects may be statistically significant. You need to look carefully at the magnitude of each coefficient to determine whether it is large enough to be substantively interesting. When the measurement scale of the variable is unfamiliar, standardized coefficients can be helpful in evaluating the substantive significance of a regression coefficient.
11. If you're interested in the effect of x on y, but the regression model also includes intervening (mediating) variables w and z, the coefficient for x may be misleadingly small. You have estimated the direct effect of x on y, but you have missed the indirect effects through w and z.

12. If intervening variables *w* and *z* are deleted from the regression model, the coefficient for *x* represents its total effect on *y*. The total effect is the sum of the direct and indirect effects.
13. If two or more independent variables are highly correlated, it's difficult to get good estimates of the effect of each variable controlling for the others. This problem is known as multicollinearity.
14. When two independent variables are highly collinear, it's easy to incorrectly conclude that neither has an effect on the dependent variable.
15. As with any statistical analysis, it's important to consider whether the sample is representative of the intended population. A probability sample is the best way to get a representative sample.
16. If a substantial portion of the intended sample refuses to participate in the study, regression analysis may produce biased estimates.

Questions to Think About

1. Dr. Johnson wants to see if job satisfaction is affected by salary. For a sample of workers in a large corporation, she regresses a measure of job satisfaction on salary, years of schooling, and age. What variables are omitted that might produce bias in the coefficient for salary? (Consider both characteristics of the job and characteristics of the employee.)
2. A psychologist hypothesizes that sleep deprivation is a cause of depression. For a sample of college students, he regresses a measure of depression on hours of sleep in the previous week, plus a large number of control variables. He finds that, as expected, those who get less sleep are more depressed. Can he reasonably conclude the sleep loss causes depression? Why or why not?
3. A criminologist finds that, among college students, there is a strong correlation between frequency of cigarette smoking and frequency of alcohol use. In a regression model, which of these should be the dependent variable and which should be the independent variable?

4. A college admissions director finds that there is a modest correlation between SAT scores prior to admission and GPA at the time of graduation. In a regression model, which of these should be the dependent variable and which should be the independent variable?
5. For the study on sleep and depression described in question 2, the researcher is concerned that both sleep loss and depression may be caused by the level of external stress, leading to a spurious relationship. To control for stress, he asks students “On a scale of 1 to 10, how much stress do you experience in your life?” with 10 being a high level of stress and 1 being very little. When the stress rating is included in the regression model, there is still a significant effect of sleep on depression. Do you see any problems with this procedure?
6. A medical researcher wants to see if zinc lozenges reduce the severity of colds. For a sample of 20 volunteers with colds, she randomly assigns 10 to get the zinc lozenges and 10 to get a placebo. She examines several outcome measures but finds no significant differences between the two groups. Is she justified in concluding that the lozenges are ineffective? Why or why not?
7. For a national probability sample of 15,000 high school seniors seeking college admission, a regression is performed with SAT scores as the dependent variable. There are many independent variables, including an IQ measure, parental income, and a dummy variable for public (0) versus private (1) high school. The coefficient for the dummy variable is positive and statistically significant, with a p value of .001. Is it correct to conclude that private schools do a better job of educating students? Why or why not? What else should be checked?
8. A baseball team statistician wants to know if his team has an advantage in playing home games. He does a regression in which the units of analysis are all the games played by his team in the last season and the dependent variable is number of points scored by his team. Independent variables include a home/away dummy variable, number of hits in the game by his team, number of strikeouts by his team’s pitcher, and the season batting average of the opposing team. The home/away dummy does not have a statistically significant effect. Do you see any problems with this analysis?

9. A survey of computer users asks them “On a scale of 1 to 10, how much do you enjoy working with your computer?” This is the dependent variable in a regression analysis. One of the independent variables is machine type (1 = IBM compatible, 0 = Macintosh or compatible). Another independent variable is system software (1 = Microsoft product, 0 = any other software company). Is there a problem here?