6 What Are the Assumptions of Multiple Regression?

Most users of multiple regression are aware that the validity of the technique depends on whether certain assumptions are satisfied, but there is enormous confusion about just what those assumptions are, how they might be violated, and what happens if they are violated. There are several reasons for this confusion. One is that there is no single set of assumptions on which everyone agrees. Rather, there are a number of different regression *models* (sets of assumptions), some involving stronger conditions than others. (This is a different use of the term *model* than we have seen before. Previously, model referred to the choice of dependent and independent variables.) Not surprisingly, the models with stronger assumptions usually lead to stronger conclusions. A second source of confusion is that the assumptions are typically expressed in mathematical form, making it difficult for many people to understand their practical implications. Finally, there is a tendency to treat all assumptions as equally important when in fact some are much more critical than others.

Actually, we already considered the most important assumptions in Chapter 3, which dealt with major things that can go wrong with multiple regression. That discussion, however, was very informal and incomplete. This chapter deals with the assumptions in a more systematic fashion.

Why have assumptions at all? Like any technique, least squares multiple regression works well in some situations and poorly in others. The assumptions can be thought of as specifying the conditions under which multiple regression works well. Indeed, under some assumptions it can be shown that least squares regression is at least as good as any other method. To talk intelligently about this,

we first need to discuss some criteria for deciding whether a statistical technique works well or poorly.

6.1. How Should We Assess the Performance of a Statistical Technique?

The best-known standards of performance for a statistical method are *bias* and *efficiency*. In general, we prefer methods that are *unbiased*. An estimation method is unbiased if there is no systematic tendency to produce estimates that are either too high or too low. Implicit in that statement is the notion that there is a "true" value that we are trying to estimate, and we can either overshoot or undershoot that value. Regardless of the estimation method, once we come up with a particular number as our estimate, we have either an underestimate, an overestimate, or the true value. If a method is unbiased, however, we say that "on average" the overestimates and the underestimates balance out.

It is not enough for a method to be unbiased. You wouldn't be very happy with an unbiased scale that was 10 pounds too high on half of the occasions that it's used and 10 pounds too low the other half. That's where efficiency comes in. Efficiency has to do with how much variation there is around the true value. We measure that variation by the standard error. Efficient estimation methods have standard errors that are as small as possible.

There's one other performance issue that we need to consider. Besides getting estimates, we usually want to test hypotheses or construct confidence intervals. To do this, we (or the computer) need to look things up in a distributional table, typically a normal table, a t table, or a chi-square table. If the sample is large, these tables are approximately correct under a very wide range of conditions, but if the sample is small, it may be necessary to make some additional assumptions about the distributions of the variables we are working with.

Now we're ready to consider some assumptions for multiple regression. Essentially, we want to determine the least restrictive set of conditions that would allow us to conclude that multiple regression estimates are unbiased and efficient. In addition, we want to know when the test statistics and confidence intervals are valid. In the remainder of the chapter, I'm going to assume that there is a

dependent variable y and two independent variables x_1 and x_2. Everything I say will readily extend to cases where there are additional independent variables, but having just two will greatly simplify the algebra.

6.2. What Is the Probability Sampling Model?

The first model we'll consider is a very simple one that takes an agnostic view of causality. It also produces only weak conclusions. Suppose we have a large population with three variables y, x_1, and x_2. If we had data for the entire population, we could use least squares to estimate a regression with y as the dependent variable. Let's represent that hypothetical regression in the entire population as

$$\hat{y} = A + B_1x_1 + B_2x_2.$$

Notice that I've used capital A and B to indicate that these are the "true" or population coefficients. Our goal is to get good estimates of these coefficients.

Because we can't afford to study the entire population, we take a *probability sample* with n cases. A probability sample is a sample in which the probability of selecting any possible sample of size n is known or can be calculated. Without going into details, there are basically three types of probability samples: simple random samples, stratified samples, and cluster samples. We can also have various combinations of these three.

For our purposes, it doesn't matter which kind of probability sample we use, as long as *every individual has an equal probability of being chosen*. Once we have such a probability sample, we apply least squares regression to the sample to get

$$\hat{y} = a + b_1x_1 + b_2x_2.$$

That's all there is to the model. Under these conditions, it can be proved that a, b_1, and b_2 are unbiased estimates of A, B_1, and B_2, respectively. But that's as far as we can go. We can't say anything about standard errors or hypothesis tests without introducing stronger assumptions.

The important point here is that least squares regression applied to a probability sample gives reasonable estimates of the least squares regression equation in the population. That shouldn't be

terribly surprising—we use the same principle in estimating means and variances. But that leaves us with a critical question: Is the least squares regression equation for the population something worth estimating? Does it really tell us anything fundamental about the causal relationships among the variables in the equation? To answer that question, we have to shift to a different kind of model, one that directly specifies the process generating the dependent variable.

6.3. What Is the Standard Linear Model?

The next model we'll consider is one that is commonly described in most textbooks on regression (e.g., Chatterjee & Price, 1991; Draper & Smith, 1998; Fox, 1997; Kleinbaum, Kupper, Muller, & Nizam, 1998; McClendon, 1994; Mendenhall & Sincich, 1996). Unlike the probability sampling model, this one says nothing about the relationship between a sample and a population. Instead, we presume that we have data on a set of individuals, labeled $i = 1, \ldots, n$, with measurements on variables y, x_1, and x_2. Then we make some assumptions about how values of y are produced from the values of the x's. Although these assumptions are usually expressed in the form of equations, they embody implicit notions of causal effects of the x's on y. I'll briefly present all five assumptions now and then elaborate on each one of them at some length.

1. Linearity. The dependent variable y is a linear function of the x's, plus a *random disturbance* U:

$$y = A + B_1x_1 + B_2x_2 + U.$$

   What's new here is the disturbance term U. It can be interpreted as a kind of *random noise* that disturbs the relationship between the x's and y. It can also be interpreted as the combined effects of all the causes of y that are not directly included in the equation. Unless we put some restrictions on U, this equation really doesn't say much, so all the remaining assumptions have to do with U.

2. Mean independence. The most important assumption we make about U is that its mean, or average value, does not depend on the x's. More specifically, we assume that the mean of U is always 0.

3. Homoscedasticity (variance independence). The variance of U cannot depend on the x's. It's always the same value, denoted by σ^2.
4. Uncorrelated disturbances. The value of U for any individual in the sample is uncorrelated with the value of U for any other individual.
5. Normal disturbance. U has a normal distribution.

Before examining these assumptions in detail, let's first see what they imply in terms of the performance of least squares.

- Assumptions 1 and 2 guarantee that the least squares estimates a, b_1, and b_2 are unbiased estimates of A, B_1, and B_2, respectively. That's the same result we got from the probability sampling model.
- If we add Assumptions 3 and 4, we find that least squares coefficients are efficient. They have standard errors that are at least as small as those produced by any other unbiased, linear estimation method. This result is captured by the acronym BLUE, which stands for Best Linear Unbiased Estimation method.
- Combined with the other assumptions, the normality assumption (5) implies that a t table can be used validly to calculate p values and confidence intervals.

Now that we've covered the main points of the model, let's look at each of these assumptions in greater detail.

6.4. What Does the Linearity Assumption Mean?

The linear equation in the first assumption tells us how values of y are generated. This equation can be thought of as representing a causal mechanism that can't be directly observed. In particular, the coefficients A, B_1, and B_2 are the "true" parameters that describe that causal mechanism. Our goal is to get good estimates of these parameters.

In thinking about this equation, there are a couple points to remember from Chapter 1. First, we recognize that in most applications of multiple regression, the assumption of linearity will be only approximately true. Second, the form of this equation actually accommodates a wide range of nonlinear relationships that can be introduced by performing some kind of transformation on the x variables.

The disturbance term U is treated as a random variable. Roughly speaking, that means that U has a probability distribution—for every possible value of U, there's a certain probability that that value

will occur. Assumption 5 says that the probability distribution is normal, but it could possibly be something else. It's important to realize that there is a different U for each individual in the data set. Potentially, these could have different probability distributions with different means and variances, but the rest of the assumptions impose considerable uniformity on these distributions.

6.5. What Is Mean Independence?

The assumption of mean independence is a way of saying that the x's are unrelated to the random disturbance U. A stronger assumption would be to say that the x's are *independent* of U, but we don't need an assumption that strong. A weaker assumption is that the x's are uncorrelated with U, but that's not quite strong enough to give us the standard results. We assume that the mean of U is 0 to get unbiased estimates of the intercept A. If we care only about unbiasedness of the B's, we need only assume that the mean of U is some constant value. In particular, the mean doesn't depend on the x's.

The assumption of mean independence is the most critical assumption of all because

- Violations can produce severe bias in the estimates
- There are often reasons to expect violations
- There's no way to test for violations without additional data.

In Chapter 1, we discussed some possible violations of this assumption. Essentially, there are three conditions that lead to violations.

1. Omitted x variables. All causes of y that are not explicitly measured and put in the model are considered to be part of the U term. If any of these omitted variables is correlated with the measured x's, that will produce a correlation between the x's and U, thereby violating the mean independence assumption.
2. Reverse causation. If y has a causal effect on any of the x's, then U will indirectly affect the x's. Consequently, the mean of U will be related to the x's.
3. Measurement error in the x's. If the x's are measured with error, that error becomes part of the disturbance term U. Because the measurement error affects the measured value of the x's, then U must also be related to the x's.

If the data are produced by a randomized experiment, these violations are unlikely to occur. The randomization process ensures that the unmeasured characteristics of the subjects are not related to the treatment variable. Randomization also prevents the dependent variable from affecting the treatment variable.

If you have non-experimental data, violations of the mean independence assumption are always a possibility. If the violation results from omission of variables that are included in your data set, then you can easily correct the problem by putting those variables in your regression equation. In all other cases, there's nothing in the data that will enable you to determine whether or not such violations are present. The only thing you have to go on is your knowledge of the phenomenon you're studying.

As with any statistical assumptions, violations and their consequences are always a matter of degree. A small amount of measurement error in an x variable will produce a small amount of bias in its coefficient. Strong effects of y on the x's will produce large biases in their coefficients.

There are ways of dealing with violations of the mean independence assumption, but they invariably require additional data, additional assumptions, and more complex methods of analysis. For example, biases resulting from measurement error can be corrected if you have external knowledge of the reliability of the variable in question or if you can get multiple indicators of that variable (Hayduk, 1988). In either case, you'll need a special program to incorporate the additional information. For reverse causation, there are *simultaneous equations* methods that are widely used by economists (Greene, 1997; Gujarati, 1995). To use these methods, however, you have to make assumptions that are often as dubious as the assumption of mean independence.

6.6. What Is Homoscedasticity?

The word *homoscedasticity* is derived from a Latin phrase meaning "same variance." The opposite of homoscedasticity is heteroscedasticity. Heteroscedasticity means that the degree of random

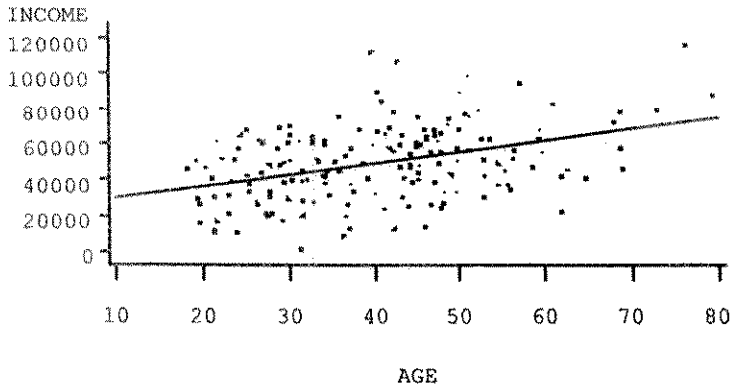

Figure 6.1. Regression of Income on Age With Homoscedasticity

noise in the linear equation varies with the values of the x variables. Homoscedasticity means that the degree of random noise is always the same, regardless of the values of the x variables.

To understand the difference between homo- and heteroscedasticity, there's no substitute for pictures. Figure 6.1 shows a linear regression of income on age for data that were generated under the homoscedasticity assumption. It is apparent that the degree of scatter around the regression line is roughly the same at all ages. There may seem to be a little more variation around age 40 and a little less around age 70, but that's merely an artifact of the greater concentration of people in the middle age levels.

In Figure 6.2, we see a very similar regression line with strong *hetero*scedasticity. The range of variation is very narrow around age 20 but steadily increases with age. The pattern in Figure 6.2 is fairly common for variables like income that are always greater than zero, but heteroscedasticity can come in many other patterns. There could be more variation at lower ages and less variation at higher ages, or the variation could be small at each end and wide in the middle. Anything that departs from a uniform degree of scatter qualifies as heteroscedasticity.

Unlike the assumption of mean independence, the homoscedasticity assumption can be checked readily with the data. For bivariate regression, scatterplots like those in Figure 6.2 can be very informative. For multiple regression, comparable graphs can be produced by plotting the observed value y on the vertical axis and

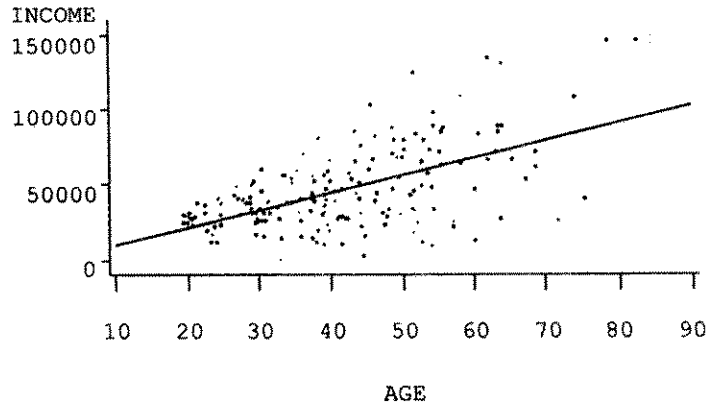

Figure 6.2. Regression of Income on Age With Heteroscedasticity

the predicted value on the horizontal axis. Some regression applications can also perform statistical tests that diagnose the presence of heteroscedasticity.

By itself, heteroscedasticity does not produce any bias in the coefficient estimates, but it does have two notable consequences:

- *Inefficiency.* Least squares estimates no longer have minimum standard errors. That means you can do better using alternative methods. The reason ordinary least squares is not optimal when there is heteroscedasticity is that it gives equal weight to all observations when, in fact, observations with larger disturbance variance contain less information than observations with smaller disturbance variance. The solution is a method known as *weighted least squares* that puts greater weight on the observations with smaller disturbance variance (Gujarati, 1995). Many multiple regression programs will do weighted least squares as an option, but there are too many details for us to consider them here.
- *Biased standard errors.* The standard errors reported by regression programs are only *estimates* of the true standard errors, which cannot be observed directly. If there is heteroscedasticity, these standard error estimates can be seriously biased. That in turn leads to bias in test statistics and confidence intervals.

Of the two problems, bias in the standard errors is more serious because it's more likely to lead to incorrect conclusions. The weighted least squares method can also correct this problem, but a much simpler solution is to use *robust standard errors* that are available as an option in some regression applications. This doesn't change the coefficient estimates and, therefore, doesn't solve the inefficiency

problem, but at least the test statistics will give you reasonably accurate p values. An advantage of this approach over weighted least squares is that it requires fewer assumptions.

Another method that is often used for reducing heteroscedasticity is to transform the dependent variable (Mendenhall & Sincich, 1996). For example, instead of using income as the dependent variable, you could use the logarithm of income. This new dependent variable tends to have much less heteroscedasticity than income itself. If the dependent variable is a count of something, like the number of quarrels that a married couple reports in a month, the square root transformation is sometimes recommended. Such transformations are called *variance stabilizing transformations*. The problem with this approach is that it fundamentally changes the nature of the relationship between the dependent variable and the independent variables. As a result, it can make the coefficients more difficult to interpret. If you don't mind that, variance stabilizing transformations can be a simple and effective solution.

My own experience with heteroscedasticity is that it has to be pretty severe before it leads to serious bias in the standard errors. Although it's certainly worth checking, I wouldn't get overly anxious about it.

6.7. What Are Uncorrelated Disturbances?

As I mentioned earlier, the disturbance term U in the linear equation is actually a different random variable for every individual in the sample. Assumption 4 says that the disturbance variables for any two individuals must be uncorrelated. The best way to understand this assumption is to think about how it might be violated. Remember that U can be thought of as containing all the unmeasured variables that affect the dependent variable y. If one of those unmeasured variables is something that two individuals have in common, the result could be a correlation between their U terms. Suppose, for example, that we have a sample of 200 people obtained by interviewing 100 married couples. The dependent variable is a measure of satisfaction with one's neighborhood. Because the husband and wife will, in the vast majority of cases, share the same neighborhood with all its characteristics, we would certainly expect their disturbance terms to be correlated.

Another way that disturbance terms can be correlated is if the behavior of one person in the sample actually affects the behavior of another person in the same sample. Suppose the sample consists of students in a single high school, and the dependent variable is a measure of educational aspirations. In that case, it's quite plausible that students who are at the top of the social hierarchy will have some impact on the aspirations of those lower down.

The most serious cases of correlated disturbances are likely to arise when the same individuals are measured at multiple points in time. Some panel surveys, for example, interview a sample of people every year for several years in a row, asking them pretty much the same questions each time. As you might expect, people's answers to the same questions are often very highly correlated from one year to the next.

More generally, the issue of correlated disturbances is strongly affected by the sampling design. If we have a simple random sample from a large population, it's unlikely that correlated disturbances will be a problem. The probability that any two individuals in the sample will interact or share a common environment will be small. On the other hand, if the sampling method involves any kind of clustering, where people are chosen in groups rather than as individuals, the possibility of correlated disturbances should be seriously considered.

The general consequences of correlated disturbances are identical to those for heteroscedasticity. Although the coefficients remain unbiased, they will be inefficient—least squares is no longer the optimal method. More seriously, the estimated standard errors will be biased. With heteroscedasticity, the direction of the bias in the standard errors is hard to predict. With correlated disturbances, however, the standard errors are almost always biased downward. That means that the coefficients are less accurately measured than you think they are. It also implies that the test statistics will be biased upward. As a result, there will be a tendency to conclude that relationships exist when they really don't.

Although it is possible to diagnose correlated disturbances by examining the data, there aren't many convenient ways to do it. If the data come in pairs, as with husbands and wives, you can calculate the residuals from the regression and then compute the correlations between husbands' and wives' residuals for the sample of married couples. For more general forms of clustering, the natural statistic to examine would be something called the *intraclass correla-*

tion coefficient (Haggard, 1958), but there are few statistical packages that will calculate it without special programming.

Solutions to the problem of correlated disturbances are quite similar to those for heteroscedasticity. The method of *generalized least squares* will produce optimal estimates of the coefficients and good estimates of the standard errors (Greene, 1997). Unfortunately, generalized least squares is rarely available in standard regression applications. Special programs are necessary, and those programs can be rather complex to use. A simpler approach is to stick with the least squares coefficients but use robust standard errors. Although these standard errors are not currently available in most regression programs, I expect that many applications will add them in the coming years.

6.8. What Is the Normality Assumption?

Much confusion exists about the normality assumption for multiple regression. Many people think that *all* the variables in a regression equation must be normally distributed. Nothing could be further from the truth. The only variable that is assumed to have a normal distribution is the disturbance term U, which is something we can't observe directly. The x variables can have any kind of distribution. Because y is a linear function of both the x's and U, there's no requirement that y be normally distributed either.

Another thing to keep in mind about the normality assumption is that it's probably the least important of the five assumptions. The criteria of unbiasedness and efficiency don't depend at all on this assumption. If the sample is moderately large, we can dispense with the normality assumption entirely. When the sample is small, we need normality of the disturbance term to guarantee that confidence intervals and p values will be accurate, but as the sample gets larger, the *central limit theorem* tells us that these statistics will be good approximations even if U is not normally distributed. How large does the sample have to be for the central limit theorem to apply? That depends on such things as the actual distribution of U and the number of independent variables in the regression equation, but I usually feel comfortable with anything more than 200 cases. Even 100 is probably OK in most circumstances if the number of x variables is small, say less than five.

When you get below 100 cases in the sample, the normality assumption becomes more critical. The natural way to check this assumption is to calculate the residuals from the regression and see if they follow something like a normal distribution. There are well-known tests for determining whether a variable has a normal distribution. Unfortunately, these tests tend to be unreliable in small samples, and that's exactly when you need them.

My advice is to be more conservative in the use and interpretation of p values when the sample is small, to compensate for the possibility that the computed p values may only be rough approximations. If you want to avoid Type I errors (concluding that a variable has an effect when it really doesn't), insist on p values that are smaller than the standard criteria. For example, instead of concluding that a coefficient is significant when the p value is less than .05, do so only when the p value is less than .01. Many researchers do exactly the opposite, however. They use less stringent criteria in small samples than in large samples because it's harder to find significant effects in small samples. In my view, this is a dangerous practice.

6.9. Are There Any Other Assumptions for Multiple Regression?

In some textbooks, you'll find the additional assumption that the x variables are *fixed* rather than random. This means that the values of the x variables don't change from one sample to another. If gender is one of your variables and you have 150 males and 175 females in your sample, the fixed-x condition would require that every possible sample have exactly 150 males and 175 females. This condition is almost never satisfied in non-experimental research. In any kind of practical probability sampling design, the exact distribution of the x variables will vary substantially from one sample to another. The fixed-x condition *can* be satisfied in experimental research because the researcher has control over how many cases fall into each treatment category.

Fortunately, the fixed-x assumption is completely unnecessary. All the standard properties of least squares estimates can be derived without invoking it. So why make this assumption? The main reason is that it greatly simplifies the algebra necessary to prove the prop-

erties of least squares. Because we're accepting those results on faith in this book, there's no need to consider this assumption.

Some textbooks also include the assumption that there is no perfect multicollinearity among the x variables. (An equivalent but obscure way of expressing this condition is to say that the **X** matrix has full rank.) Although this is certainly a necessary condition for computing linear regression estimates, it's really a property of the sample rather than the population or the underlying mechanism generating the data. For that reason, I do not include it in the assumptions, but I do consider multicollinearity in some detail in Chapter 7.

As I mentioned at the beginning of the chapter, you will find somewhat different sets of assumptions discussed in different textbooks. One set of assumptions is the *multivariate normal model*. This model says, quite succinctly, that the observed variables have a multivariate normal distribution. Among other things, this means that every variable in the regression equation is normally distributed. Moreover, every variable is linearly related to every other variable, and those linear equations are homoscedastic. The multivariate normal model implies all the assumptions we have considered here, but it's *much* stronger than necessary to derive the standard properties of least squares estimates.

In some discussions of multiple regression, you'll find the mean independence and homoscedasticity assumptions replaced with the single assumption that x and U are independent. Although this simplifies things a bit, it's more restrictive than necessary. Moreover, it obscures the quite different consequences of violating mean independence and violating homoscedasticity.

Chapter Highlights

1. There are several different models (sets of assumptions) that may be used to justify ordinary least squares regression.
2. We want estimation methods that are unbiased, at least approximately. Unbiased methods have no systematic tendency to underestimate or overestimate the true value.
3. Efficient estimation methods have standard errors that are as small as possible. That means that in repeated sampling, they don't fluctuate much around the true value.

4. If we have a probability sample drawn so that every individual in the population has the same chance of being chosen, then the least squares regression in the sample is an unbiased estimate of the least squares regression in the population.
5. The standard linear model has five assumptions about how values of the dependent variable are generated from values of the independent variables. If all these assumptions are met, ordinary least squares has several desirable properties. The assumptions of linearity and mean independence imply that least squares is unbiased. The additional assumptions of homoscedasticity and uncorrelated errors imply that least squares is efficient. The normality assumption implies that a t table gives valid p values for hypothesis tests.
6. The disturbance term U represents all the unmeasured causes of the dependent variable y. It's assumed to be a random variable, having an associated probability distribution.
7. Mean independence means that the mean of the random disturbance U does not depend on the values of the x variables.
8. Mean independence of U is the most critical assumption of the standard linear model because violations can produce severe bias and there are often reasons to suspect that violations will occur.
9. Violations of mean independence will occur if (a) important x variables are omitted from the regression model, (b) y has a causal effect on one or more x's, or (c) the x's are measured with error.
10. Homoscedasticity means that the degree of random noise in the relationship between y and the x's is always the same.
11. A good way to check for violations of the homoscedasticity assumption is to plot the residuals against the predicted values of y. There should be a uniform degree of scatter, regardless of the predicted values.
12. Heteroscedasticity (violation of homoscedasticity) can make ordinary least squares inefficient and can produce bias in standard error estimates.
13. There are three ways to deal with heteroscedasticity: variance stabilizing transformations, weighted least squares, and corrected standard errors. Although sometimes effective in reduc-

ing heteroscedasticity, variance stabilizing transformations often have other undesirable properties.

14. Weighted least squares is a good method for correcting heteroscedasticity if you are confident about the form of the heteroscedasticity. Otherwise, you're better off using corrected standard errors.
15. The assumption of uncorrelated errors is usually satisfied if you have a simple random sample from a large population, but it's likely to be violated if observations are selected in clusters or if sample units can interact with each other.
16. Correlated errors can cause serious underestimates of standard errors, leading to inflated test statistics. This problem can be solved with a method known as generalized least squares.
17. The normality assumption is the least critical of the five assumptions. It has nothing to do with the independent variables, and it can be safely ignored if the sample is large.

Questions to Think About

1. A researcher tells you that he doesn't want to use multiple regression on his data because it requires too many assumptions. What response would you give?
2. An unbiased estimation method is one that "on average" gives the right answer. Because any specific estimate might be very far from the true value, why do we care about unbiasedness?
3. Suppose we have a simple random sample of students currently enrolled in U.S. colleges. In the sample, we regress GPA on SAT scores, parents' annual income, and hours per week spent studying. What does this regression tell us, if anything, about the complete population of students?
4. Every regression program produces both coefficients and standard errors. What do the standard errors mean? Why do we want them to be small?
5. The first assumption in the standard linear model is that y is a linear function of the x's, plus a random disturbance term U. I claimed that "unless we put some restrictions on U, this equation really doesn't say much." Why does the linearity assumption have little content unless restrictions are placed on U?

6. Chapter 3 states that when "important" variables are omitted from a regression equation, the coefficients of the included variables will be biased. How does the omission of variables lead to a violation of the mean independence assumption?
7. Dr. Hamilton published the results of a multiple regression showing that certain parenting styles lead to higher rates of juvenile delinquency. A critic claims that Hamilton should have checked whether his data satisfy the mean independence assumption. What answer should he give?
8. In later checking, Hamilton's student discovered that there was a substantial amount of heteroscedasticity in the regression of juvenile delinquency on parenting styles. Does this mean Hamilton's published results were invalid? What should he be concerned about, if anything?
9. Professor Long does a regression in which the units of analysis are the 50 U.S. states. Her dependent variable is a measure of support for environmental legislation, and her principal independent variable is the percentage of the state's economy devoted to manufacturing. Should she be concerned about the possibility of correlated errors? Why or why not?
10. Many regression analysts use dummy variables as independent variables, but dummy variables cannot possibly be normally distributed. Is this a problem for regression analysis?
11. For a sample of 5,000 discharged military personnel, Dr. Short wants to do a regression analysis of length of service on several independent variables. Length of service is highly skewed—not at all like a normal distribution. Can he go ahead with the regression analysis?