16
Detecting and Managing Irregularities

This chapter addresses the topic of *regression diagnostics*. Diagnostic statistics are useful for identifying cases in an analysis that are "irregular" in some way. We introduce *leverage*, *distance*, and *influence* as measures of irregularity and discuss how irregular cases may distort a regression analysis and so are worth identifying prior to interpretation of the results. We describe how diagnostic statistics can be used for testing whether the assumptions of linear regression analysis are met, introduce some ways of dealing with assumption violations, and discuss how violations may affect the validity of the inferences one makes using regression analysis.

All too many investigators discover clerical errors in their data, such as inputting a person's age or a response to a question on a survey incorrectly, only after they have already spent hours on their data analysis and have perhaps reached conclusions that are hard to erase from their minds. Or after publication, a critic may point out that the researcher's main conclusion depended entirely on one research participant who was very unusual and perhaps should not have even been included. Some statistical techniques designed for avoiding mishaps like these are the topic of this chapter. We discuss some methods of detecting cases that are somehow "irregular," which we define later in a number of ways. We talk about what to do when they are detected and methods you might consider employing if you are worried about the effects such irregularities may have on the quality of the inferences you report. We provide only a rough overview of these topics, which can be quite complicated. A more extensive treatment of some of the topics we discuss and others we don't can be found in Berry (1993), Fox (1991), and Kaufman (2013), among others.

16.1 Regression Diagnostics

In the evolution of regression analysis, diagnostic statistics are relatively new, having been developed mostly since 1975 or so. These statistics have several purposes. First, they can help to detect clerical errors, such as inputting a person's height as 720 inches rather than 72, which can seriously distort an entire analysis if not caught. Second, they can detect violations of the secondary assumptions of homoscedasticity and normality. Third, they can be used to examine data that are suspect for some reason, such as questionnaire results from someone who appeared not to understand directions, to determine whether those data are irregular in some way.

Diagnostic statistics can also be used to identify cases whose presence in the analysis are greatly influencing the results. For this reason, they can easily be misused. For instance, using some of the statistics and methods in this chapter, a clinical psychologist could find that three people in a study, if deleted from the analysis, could improve the apparent effectiveness of a therapeutic method he or she developed. This discovery could lead the psychologist to look at the files of these patients and find some rationale for excluding them. But any tool can be misused, and diagnostic statistics are an important part of regression analysis. The best protection against misuse is to require authors to explain in detail the reasons for deleting any cases and the ways in which those deletions affected the major conclusions. Although you can be faulted for your decision to exclude cases, you can't be accused of misconduct or unethical behavior if you are open about what you have done.

Diagnostic statistics may also occasionally detect violation of the primary assumption of linearity. But intuition suggests that they would not be nearly as powerful for that purpose as the methods discussed in Chapter 12, and our own analyses confirm that conjecture. For example, in a small-scale simulation study, we found that a test on the regression coefficient for X^2 to detect curvilinearity correctly detected real nonlinearity 98% of the time, while an approach we describe in section 16.2.4 detected nonlinearity only 33% of the time.

One of the best ways of detecting irregularities is to search for cases that are "extreme" in one sense or another. Such cases are often called *outliers*, though we confine that term to a particular type of extreme case.

16.1.1 Shortcomings of Eyeballing the Data

When computers were in their infancy, one of the major arguments given against their use in data analysis was that computer analysis made it easier to overlook extreme or unusual cases, even when they reflect obvious errors such as adult human weights of 16 or 1,600 pounds. Most likely, this is a clerical error of some kind that should be fixed before data collection. One simple way of catching extreme cases such as this is to scan the data file with your eyes, just looking for things that seem amiss. This is easy to do if the data file is small, but with large data sets with many variables, such "eyeballing" of the data may miss important irregularities.

Statistical computer programs quickly met the objection mentioned above by making it easy to identify the highest and lowest score on every variable, so that such extreme cases can be called to the investigator's attention. We recommend that prior to conducting an analysis, you ask your computer program to print the smallest and largest values of every variable in the data. Doing this would condense information about extreme cases for all the variables into one small output and make it easy to detect problems, such as someone whose weight is 1,600 pounds or who is −4.5 years old. Such values in the data are likely to show up as the minimum or maximum value for the variable. If you see something like this, fix it or otherwise investigate the source of the problem. Maybe you or your research assistant simply mistyped a weight when entering the data. Or if the data were collected by a computer program, maybe there is a bug in the program that generates incorrect data in certain circumstances.

Today's computer programs allow us to go far beyond this basic step. Using statistics discussed in this chapter we can detect irregularities that could never be discovered by eyeballing the data or looking at maximums and minimums. For instance, suppose your data file contains information about employees at a particular company, and the records for one employee in your data include the following information:

- Present salary: $30,000
- Hours worked per week: 20
- Starting salary: $20,000
- Hours worked per week on starting: 40
- Number of years worked: 2

This case is very unusual, though it isn't obvious how unless you think carefully about it. Notice that the employee earns the equivalent of a full time employee (40 hours per week) who makes $60,000 per year. But only 2 years ago, when the person was working full time, he or she was making only $20,000, so the employee's salary is three times what it was only 2 years ago. Most people don't get such large raises so quickly. Such an unusual case may represent a clerical error or some other factor worth checking. But ignoring any one of the five entries in this person's file would make the case appear normal. For instance, if "number of years worked" were not shown, we might assume it was 10 or 20 instead of 2, and the case would appear normal. A similar argument can be made about any of the other entries. Only when all five entries are considered together is the case identified as unusual. But if these five entries were scattered among 20 or 30 other entries about the same employee, it is highly unlikely that eyeballing of the data matrix would reveal anything amiss. Nor would this case likely be brought to our attention if we look only at the minimum and maximum values across all the employees on all five of these variables. Some regression diagnostic statistics can easily detect such cases.

16.1.2 Types of Extreme Cases

A case can be extreme or otherwise noteworthy in three major ways, all of which can be quantified. A case has high *leverage* if its pattern of regressor scores (ignoring Y) puts it far from most or all other cases. Speaking a bit loosely, cases with the highest *distance* are those whose vertical distance from the regression surface is greatest. *Influence* measures how much a case's presence in the analysis actually moves the regression surface. As we see later, there is a sense in which influence is the product of leverage and distance, so high influence requires both high leverage and high distance.

The distinctions among distance, leverage, and influence are illustrated most easily in simple regression. Consider the data set in Figure 16.1. Suppose the sample contains only the 37 cases represented with a solid square. If you regressed Y on X for only these 37 cases, the resulting model would be $\hat{Y} = 4.0 + 0.0X$. Now suppose you added case A to the data, denoted with a hollow square in the figure, bringing the sample size to 38. This case is extreme in the distribution of X. But if case A is included in the analysis, the regression model is unaffected; it is denoted with the thin, solid black line, and its equation is identical: $\hat{Y} = 4.0 + 0.0X$. This case is high in leverage, low in distance, and low in influence.

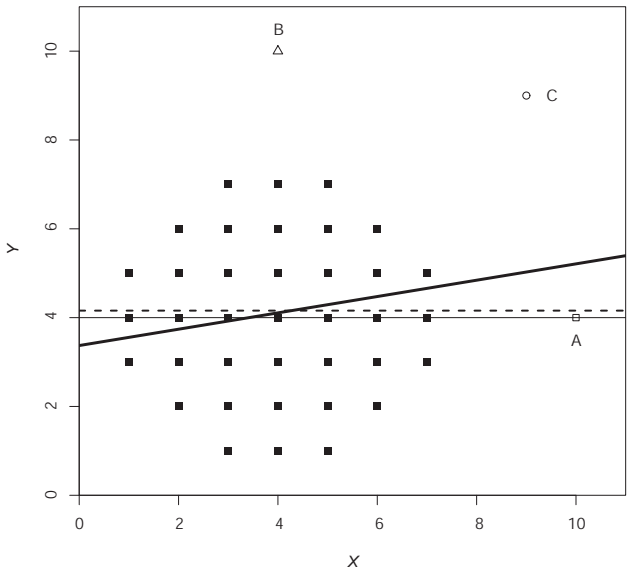

FIGURE 16.1. The influence of adding one case (represented with a hollow triangle, circle, or square) to a regression model containing 37 cases (represented with solid squares).

But now suppose you added only case B, denoted with the the hollow triangle. It is unusual on Y but quite ordinary on X. If it were included in the analysis, $\hat{Y} = 4.160 + 0.0X$, depicted with the dashed line. The regression constant has changed slightly, but the regression coefficient for X has not changed at all. This case is low in leverage, high in distance, and low in influence.

Finally, suppose you added only case C, denoted with the hollow circle. When it is included in the analysis, $\hat{Y} = 3.372 + 0.184X$, represented in the figure with the solid dashed line. Case C is high in leverage, high in distance, and high in influence. If leverage is potential to influence, then case C has realized that potential, whereas case A has not.

Users of regression analysis often focus on residuals when looking for extreme or influential cases, paying close attention to cases with large residuals (i.e., large distance). But this example shows that residuals are not necessarily the best way to identify influential cases, because cases that influence a regression analysis can "hide" by shrinking their own residual. Notice that case C pulls the regression line toward it, cutting its residual by

about 20%. Better would be some kind of a statistic that quantifies a case's residual relative to what it would be if the case weren't in the analysis. There is a measure of this, and we discuss it.

Complicating your learning of regression diagnostics, terminology is not standard, and there are many statistics in the literature that are not aptly named. For example, there is a measure called *Cook's distance* that we discuss later, but it is really a measure of influence rather than distance. Another statistic, called *Mahalanobis distance*, is a measure of leverage, because it quantifies how unusual a case's pattern of regressor scores is.

Cases with high leverage are here called *leverage points*. A case high in leverage has the potential to be influential, but it may not be. Cases high in distance are here called *outliers*, though other writers often use this term to describe any kind of extreme or unusual case. Cases high in influence are here called *influential cases* or *influential points*.

Leverage differs qualitatively from distance, in that cases extreme in distance can invalidate statistical inference in regression. But extreme leverage violates none of the standard assumptions of regression, because regression analysis makes no assumption about the distribution of regressors.[1] But high-leverage cases can affect power and precision of estimation. Consider a single dichotomous regressor such as sex. If a sample includes 90 men and 10 women, the difference between men and women on Y is going to be estimated with less precision than if the sample includes 50 men and 50 women. But the 10 women are going to be much higher in leverage than the 90 men.

16.1.3 Quantifying Leverage, Distance, and Influence

There is a variety of different ways that leverage, distance, and influence can be measured, depending on how you think about these concepts, and we talk about only some of them. They are all interrelated in one way or another.

Leverage. We start first with leverage, which we defined earlier as the atypicalness of a case's pattern of values on the regressors in the model. A case in a data set may have quite ordinary values on the individual regressors, but its combination of regressor values might be quite unusual. For instance, being 55 and being pregnant each are not particularly unusual

[1]There is a common misconception that regression analysis assumes normally distributed regressors. This is not true. We have seen that dichotomous variables can be used as regressors, and ANOVA is just a special case of regression analysis with dichotomous regressors. But dichotomous variables are by definition not normal.

if you were to randomly sample people from a broad population, but if pregnancy and age were both regressors in a regression model, then a 55-year-old pregnant woman would have high leverage since this combination would be very unusual in almost any sample (except perhaps a sample from a population of older pregnant women).

Consider a single variable X_1 used as a regressor in a simple regression model $\hat{Y} = b_0 + b_1 X_1$. Quantify the discrepancy between X_1 and $\overline{X}_1$ for case i in standard deviations of X_1, and then square this result. This is just the squared standardized value of X_1 for case i:

$$Z^2_{X_{1i}} = \left(\frac{X_{1i} - \overline{X}_1}{s_{X_1}} \right)^2 \tag{16.1}$$

The farther case i's X_1 value is from the mean of X_1, the larger is $Z^2_{X_{1i}}$. $Z^2_{X_{1i}}$ cannot be negative, and it will be zero only if $X_{1i} = \overline{X}_1$. $Z^2_{X_{1i}}$ is known as *the Mahalanobis distance* for case i, which we denote as MD_i, but it is really a measure of leverage rather than distance as we have defined the terms. Note that although we have introduced this statistic in the context of a simple regression model, Y is not used in its computation at all.

As defined in equation 16.1, we might call MD_i *univariate* Mahalonobis distance, because it is a measure of case i's atypicalness on a single variable. But Mahalanobis distance can be defined more generally in a multivariate form that considers a case's atypicalness on a set of regressors. Suppose we have a second variable X_2, and we want to calculate case i's atypicalness in its pattern of values on X_1 and X_2 considered jointly. You might think we could just calculate $Z^2_{X_{2i}}$ for X_2 in a comparable way and then add it to $Z^2_{X_{1i}}$ to get a multivariate Mahalanobis distance that considers both X_1 and X_2. The trouble with this reasoning is that if X_1 and X_2 are correlated, this sum would contain some redundancy. The stronger the correlation between X_1 and X_2, the more likely a case is to be atypical on both, and this sum would double-count part of the discrepancy. A preferred multivariate measure would quantify case i's atypicalness on X_2 accounting or adjusting for the correlation between X_1 and X_2.

Recall from section 2.4.2 that the residuals in a regression are uncorrelated with the regressor or regressors. Later, in section 3.2.2, we showed that if we regress X_2 on X_1, then the residuals from this regression, $X_{2.1}$, are uncorrelated with X_1, making $X_{2.1}$ a measure of X_2 that has been purified of its linear relationship with X_1. With these residuals calculated, we can then quantify how atypical case i is on $X_{2.1}$, the part of X_2 independent of

X_1, using the same logic as above. Case i's atypicalness on X_2 controlling for X_1 is

$$Z^2_{X_{2.1_i}} = \left(\frac{X_{2.1_i} - \overline{X}_{2.1}}{s_{X_{2.1}}} \right)^2 \tag{16.2}$$

but because $\overline{X}_{2.1} = 0$ (residuals always have a mean of zero), equation 16.2 simplifies slightly to

$$Z^2_{X_{2.1_i}} = \left(\frac{X_{2.1_i}}{s_{X_{2.1}}} \right)^2$$

The farther case i's $X_{2.1}$ value is from zero, regardless of sign, the larger is $Z^2_{X_{2.1_i}}$. Now we can add $Z^2_{X_{1i}}$ and $Z^2_{X_{2.1_i}}$ to get MD_i, the Mahalanobis distance for case i on the set of regressors X_1 and X_2. The larger MD_i, the more atypical is case i's pattern of values of X_1 and X_2.

You might wonder what would happen if we reversed the order of computations above, starting first with X_2 and then generating the residuals from regressing X_1 on X_2 to generate $X_{1.2}$. It turns out that this doesn't matter. MD_i will be the same. We can then further extend this logic to k regressors by adding successive values of Z_{ji} to those that come before, i.e., $Z^2_{X_{3.12_i}}$, $Z^2_{X_{4.123_i}}$, and so forth. The resulting MD_i calculated as the sum of all these k values of Z^2 will not be affected by the order of the partialing process.

MD_i will tend to be large for cases that are more distant from the center of a multivariate space defined by the joint distribution of the k regressors. But when statisticians use the term *leverage* in regression analysis, they are often not talking about MD_i but rather a different statistic h_i, which is often labeled case i's *hat value*. It is difficult to talk about the computation of h_i without using matrix algebra, so we refer interested readers to Appendix D where we provide the formula. It turns out that h_i is perfectly linearly correlated with MD_i. That is, their correlation across all N cases is exactly 1. So we know the case highest on MD_i is also highest on h_i, the case that is second highest on MD_i is second highest on h_i, and so forth. Unlike MD_i, which has no upper bound, h_i is always between $1/N$ and 1. Furthermore, $\overline{h} = (k + 1)/N$. From now on, whenever we make specific references to "leverage" in computations, we are referring to h and not MD. As we will see, h appears in the computation of many regression diagnostics, so it has more value in regression diagnostics analysis than MD.

In large data sets with more than a couple of regressors, there will often be one or two cases with large values of MD_i and h_i that stand out in the distribution relative to others. But in small samples, or when considering a

small number of regressors, it would not be uncommon to find several cases with large values. For instance, from the weight loss data set in Table 3.1, the largest MD_i calculated using exercise frequency and food intake is 2.500 and the largest h_i is 0.378. But four of the 10 cases have these largest values. So really they aren't atypical at all. Thus, these aren't perfect measures of atypicalness, but they generally are sensitive to the concept as most people would think about it.

Distance. Distance measures how far case i's Y value deviates from $\hat{Y}_i$. Cases with extreme distance are *outliers*. Such cases are more important than leverage cases because a sufficiently extreme outlier represents a violation of at least one of the standard assumptions of regression, while leverage points do not. An outlier may or may not have high leverage.

Outliers can be the result of clerical errors, so it is always worth checking that first. Assuming any outliers found are legitimate values, they may suggest revisions to the model are needed. For instance, if 80% of the cases in a sample were women and 20% were men, and most of the outliers were men, this might mean you need different models for men and women, and the predominance of women in the sample forces the model to fit the women's data. Thus, developing separate models for men and women, perhaps through the methods discussed in Chapters 13 and 14, may be appropriate. Or if there are too few men to develop a separate model for men, a large number of male outliers may suggest that they be excluded from the sample and that the conclusions of the model be applied only to women.

The most obvious measure of distance is a case's residual $e_i = Y_i - \hat{Y}_i$, but residuals can be refined. Cases with high leverage tend to pull the regression surface toward them more than other cases do, thereby shrinking their own residual. So residuals can be adjusted for the case's leverage. We can define a leverage corrected residual as $e_i/\sqrt{(1-h_i)}$. Leverage-corrected residuals are rarely actually used, but an interesting fact is that the square of a leverage-corrected residual equals the amount $SS_{residual}$ would drop if case i were excluded from the analysis.

The expected value of the squared leverage-corrected residual is ${}_T\text{Var}(Y.X)$, which is estimated by $MS_{residual}$. So leverage corrected residuals can be standardized by dividing them by $\sqrt{MS_{residual}}$, as

$$str_i = \frac{e_i}{\sqrt{(1-h_i)MS_{residual}}} \tag{16.3}$$

In large samples, residuals transformed by equation 16.3 are normally distributed with a mean of zero and a standard deviation of one. An additional transformation

$$tr_i = str_i \sqrt{\frac{df_{residual} - 1}{df_{residual} - str_i^2}}$$

results in residuals that are exactly t-distributed with $df_{residual} - 1$ degrees of freedom. We will refer to these as *t-residuals*.[2] Because t-residuals are exactly t-distributed, they are useful for testing some of the standard assumptions of regression, as discussed in section 16.2. The transformation of str_i to tr_i does not change the relative ordering of the cases on these measures of distance. Their rank correlation will be 1.

Earlier we said that cases with high leverage tend to pull the regression surface toward them more than cases with low leverage, thereby shrinking their own residuals. We also just said that tr_i quantifies distance for case i in reference to its $\hat{Y}_i$ when it is excluded from the analysis. It turns out h_i has a similar interpretation. Define e_i as case i's ordinary residual $Y_i - \hat{Y}_i$ and define ${}_de_i$ as $Y_i - \hat{Y}_{i,\text{not}\,i}$, where $\hat{Y}_{i,\text{not}\,i}$ is defined as in section 7.2.3, as case i's estimate of Y derived from the model estimated without case i. It turns out that

$$h_i = \frac{{}_de_i - e_i}{{}_de_i}$$

In words, h_i equals the proportion by which case i lowers its own residual by pulling the regression surface (i.e., the model that produces $\hat{Y}$ for all cases) toward itself. Consider, for instance, a case with a residual of 6, meaning its Y is 6 points above its $\hat{Y}$. If that case's residual would be 8 points above its $\hat{Y}$ if it were excluded from the analysis, then that point's h_i is $(8 - 6)/8 = 0.25$ since inclusion of the point has pulled the regression surface 25% of the way toward the point. Thus, the highest possible value of h_i is 1. The lowest possible value is $1/N$; a case exactly at the mean on all regressors but above or below the mean on Y will not change any regression coefficients, but it will pull the entire regression surface up or down $1/N$th of the point's distance from the surface's previous location. This may be the simplest single definition of h_i, but it can't be considered the primary definition, because it obscures the important fact that h_i is computed without reference to Y. That is, h_i is determined entirely by the

[2]Terminology is inconsistent in the literature and computer software. What we call t-residuals other authors and some statistics programs call *studentized* residuals. A distinction is also made by some authors between *internally* studentized residuals and *externally* studentized residuals. In our notation, these are str_i and tr_i, respectively. SPSS produces something it calls *standardized* residuals, but these are something different still.

regressors, not by Y. It also reduces to an indeterminate form when ${}_d e_i = 0$, but h_i is just as precisely defined for such cases as for any other case. But h_i is not really a measure of distance, even though we have included this manner of defining h_i in this section.

Influence. The influence of a case is quantified by the extent to which its inclusion changes the regression solution or some aspect of it, such as the estimates it generates for Y. It is the cases that most change the regression surface by their inclusion in the analysis that we are most concerned about and wish to identify for further scrutiny. There are many ways one can measure influence. We restrict our discussion here to how the inclusion of case i changes $\hat{Y}$ for all cases or how b_j is changed by the inclusion of case i. But these aren't the only ways of quantifying influence; a case could have little influence on a regression coefficient or $\hat{Y}$, but its presence in a model could greatly influence R or $SE(b_j)$, for instance.

The standard measure of a case's influence on the regression surface was suggested by Cook (1977), and is here denoted $Cook_i$. This measure is inappropriately named *Cook's distance*; *Cook's influence* would be better. $Cook_i$ is proportional to the sum of squared changes in values of $\hat{Y}$ across all cases when case i is deleted from the analysis. To be precise, let d_{ij} denote the change in the value of case j's residual when the residuals are rederived after case i is deleted from the analysis. Then

$$Cook_i = \frac{\Sigma_{j=1}^{N} d_{ij}^2}{k \times MS_{residual}}$$

where k is the number of regressors. Thus, $Cook_i$ is a measure of the amount values of $\hat{Y}$ move when case i is deleted from the analysis. It can be thought of as the product of a particular measure of distance and a particular measure of leverage. The key formula is

$$Cook_i = str_i^2 \times \frac{h_i}{(1 - h_i)(k + 1)}$$

As discussed earlier, str_i ranks cases in the same order as tr_i, the best measure of distance from the regression surface. And all the rest of the right side is a measure of leverage in that it ranks cases in the same order as h_i.

Some have stated that $Cook_i$ is distributed as F with $k + 1$ and $N - k - 1$ degrees of freedom. But, in fact, the mean of an F distribution is always over 1, and values of $Cook_i$ are rarely found as high as 1. Also, the standard

assumptions of regression do not require any particular distribution for h_i, but h_i has a major effect on $Cook_i$, so no general rule can be stated for the distribution of $Cook_i$.

In multiple regression we can distinguish between *total influence* and *partial influence*. Whereas $Cook_i$ measures the influence of case i on the entire regression model, as manifested by what it generates for $\hat{Y}$ for each and every case, partial influence measures a case's influence on a specific regression coefficient b_j. If, say, 10 regressors include nine covariates and one independent variable X_1, then we may be more concerned about cases that substantially affect b_1 than about cases with high total influence. Thus, if your focus is on a specific regressor j, you may be particularly concerned about identifying cases that have a lot of influence on that specific b_j, but care little or not at all about how any case influences any of the other $k-1$ regression coefficients or the regression constant.

A statistic called $dfbeta_i$ quantifies how much case i influences a specific regression coefficient. In a regression model with k regressors, there are $k+1$ *dfbeta* values for each case, one for each regression coefficient and one for the constant. We will denote the $dfbeta_i$ for regressor j as $DB(b_j)_i$. It is defined as

$$DB(b_j)_i = b_j - b_{j,\text{not}\,i}$$

where $b_{j,\text{not}\,i}$ is b_j when case i is excluded from the analysis. For instance, if $b_j = 1$ but $b_{j,\text{not}\,i} = 0.25$, then $DB(b_j)_i = 0.75$, meaning that including case i in the analysis raises b_j by 0.75. Large values of $DB(b_j)_i$ relative to other cases suggests that case i is having a big effect on the estimate of the X_j's partial relationship with Y. It can be shown that

$$DB(b_j)_i = \frac{e_i ce_{ij}}{N(1-h_i)\text{Var}(X_j)Tol_j}$$

where ce_{ij} is the residual for case i in the crosswise regression predicting X_j from the other regressors.

16.1.4 Using Diagnostic Statistics

The analysis of regression diagnostics is as much art as science. The ultimate objective is to flag any cases in the data that are unusual or extreme in some fashion for closer scrutiny. Some authors provide rules of thumb for deciding whether a certain diagnostic statistic is too large or offer ways of testing hypotheses about whether a certain diagnostic is larger than you would expect to observe by chance. These hypothesis tests and rules of

thumb make assumptions about the distribution of various diagnostics, but extreme cases may make those assumptions less tenable. So, with one exception mentioned here and later in section 16.2, we recommend instead a descriptive and holistic approach in which you look at the distribution of each of the diagnostics, notice those that really stand out as unusual relative to others, and see if there are some cases in the data that seem to consistently come to your attention using various diagnostics.

We illustrate this approach using the data set in Table 16.1. The 12 cases in the data represent two groups coded $X_1 = 0$ and $X_1 = 1$, such as an experimental and a control condition, along with two numerical variables, X_2 and X_3. The diagnostic statistics in Table 16.1 are generated from a regression estimating Y from X_1, X_2, and X_3.

As already mentioned, one of the first uses of diagnostic statistics is to identify clerical errors or other problems that may have occurred at the data entry or data generation stage of the research. We discussed the use of leverage for this purpose, as cases with an unusual pattern of scores on the regressors will often show up as high in leverage. A leverage measure such as h_i can be useful for the identification of such errors and supplement what can be learned by looking at the minimum and maximum values. We provided an example of how the minimum and maximum values may fail to detect a case with an unusual pattern of values in section 16.1.1.

The data in Table 16.1 provide another illustration of how simple eyeballing of the data or the use of maximum and minimum values can fail to uncover extreme cases. In these data, the third case is highest in leverage, with values of MD_i and h_i of 7.157 and 0.734, respectively. These values are no less than 75% larger than the corresponding statistics for the case with the next highest leverage. But only very careful examination shows that the value of 11 for X_2 is unusual not by itself but in relation to its X_3 value. Notice that all cases with relatively small values of X_3 also have values of X_2 that are relatively small, and that this is true regardless of whether X_1 is 0 or 1. But not so for case 3, which has quite a large value of X_2 even though this case's X_3 value is relatively small. So it doesn't fit the pattern of the association between X_2 and X_3. Yet examining case 3's values of X_1, X_2, and X_3 individually reveals nothing extreme or unordinary about this case, and 11 is neither the maximum nor the minimum value of X_2 in the data, so looking at the minimums and maximums would not flag this case as unusual. Measures of leverage have flagged this case as worthy of further attention. If these were your data, you might take a look at the data collection records to see whether X_2 was entered incorrectly for this case

TABLE 16.1. Various Measures of Leverage, Distance, and Influence

i	X_1	X_2	X_3	Y	$\hat{Y}$	e	${}_de$	MD	h	str	tr	$Cook$	$DB(b_0)$	$DB(b_1)$	$DB(b_2)$	$DB(b_3)$
1	0	1	3	8	9.523	−1.523	−2.497	3.374	0.390	−0.932	−0.923	0.139	−0.942	0.610	0.121	−0.051
2	0	3	4	13	10.204	2.796	3.800	1.991	0.264	1.558	1.746	0.218	1.200	−0.790	−0.111	0.035
3	0	11	5	17	16.994	0.006	0.022	7.157	0.734	0.005	0.005	0.000	0.004	0.003	0.002	−0.002
4	0	7	8	7	8.857	−1.857	−2.377	1.488	0.219	−1.004	−1.005	0.071	−0.290	0.605	0.064	−0.085
5	0	9	8	10	10.893	−0.893	−1.096	1.117	0.185	−0.473	−0.449	0.013	−0.105	0.173	−0.011	0.000
6	0	12	12	10	8.528	1.472	2.349	3.192	0.374	0.889	0.876	0.118	−0.200	−0.594	−0.014	0.104
7	1	2	5	10	10.664	−0.664	−0.911	2.069	0.271	−0.372	−0.351	0.013	−0.164	−0.152	0.016	0.008
8	1	4	6	15	11.345	3.655	4.637	1.413	0.212	1.968	2.563	0.260	0.551	0.945	0.007	−0.091
9	1	3	4	11	13.037	−2.037	−3.048	2.733	0.332	−1.191	−1.228	0.176	−0.617	−0.839	−0.057	0.154
10	1	7	12	6	6.270	−0.270	−0.444	3.389	0.391	−0.165	−0.155	0.004	0.059	0.005	0.022	−0.033
11	1	9	10	11	11.016	−0.016	−0.020	1.395	0.210	−0.009	−0.008	0.000	0.002	−0.004	0.000	0.000
12	1	13	14	9	9.669	−0.669	−1.149	3.681	0.418	−0.419	−0.396	0.032	0.324	−0.182	−0.029	−0.018

or otherwise examine your measurement system to see if something went awry.

We might worry that case 3, because of its unusual pattern of values on the regressors, may distort the regression surface in some way. Diagnostic statistics can help identify whether this is so for case 3, or perhaps for some other case in the data. Starting first with distance, cases with a large discrepancy between Y and $\hat{Y}$ can suggest a violation of one of the assumptions of regression, such as normality or homoscedasticity. We recommend the use of the t-residual as the best measure of distance rather than relying on str_i or e_i. In section 16.2 we discuss a way of using the t-residuals for testing whether one of the assumptions of regression has been violated. For now, notice that case 3's t-residual is not particularly large in absolute value. We might be more concerned about case 8, with a t-residual of 2.563. You would expect only 1 in 29 cases in a regression analysis to have a t-residual this large or larger in absolute value if the assumptions of regression have been met. So in a sample of only 12 cases, this residual stands out as potentially unusual or uncommon to observe. But as will be seen in section 16.2.4, we would want to correct this probability for the fact that we have looked at 12 residuals rather than just 1 before claiming we have violated an assumption. This should remind you of the multiple test problem discussed in Chapter 11.

Remember that MD_i and h_i measures the atypicality of a case i's pattern of regressor values. Neither of these statistics is calculated in reference to Y. It could be that the large residual observed for case 8 reflects some kind of data entry error for Y. This would be worth checking. You could also calculate MD_i or h_i while treating Y as if it were a regressor. This could be accomplished by requesting your computer to produce one of these leverage measures when regressing some other variable in the data set on X_1, X_2, X_3, and Y. The dependent variable could even be a set of random numbers since the dependent variable is not used in the computation of leverage. When we did so, we found that case 8's leverage was not particularly large (though it was the second largest out of 12, it didn't stand out much from many of the other cases), thereby reducing our concern that its large t-residual is due to a clerical or computational error of some kind.

A case can be influential in that it changes $\hat{Y}$ a lot for all cases in the data, or it could be influential in its effect on one or more of the regression coefficients. The former is measured with $Cook_i$ and the latter with $DB(b_j)_i$. Observe that case 3, our case with the highest leverage, has a tiny $Cook$ value. Notice as well that the regression coefficients and regression constant, as

measured by the $DB(b_j)$ statistics, are barely affected at all by the inclusion of case 3. It has very little influence. The inclusion of case 8 (the case with the largest distance as well) has the biggest influence in shifting all cases' $\hat{Y}$ values around, because it has the largest value of *Cook*. Observe as well that it has the largest $DB(b_1)$ in absolute value. Its value of $DB(b_1) = 0.945$ means that b_1 is 0.945 larger than it would be if this case were excluded from the analysis. With the case included, $b_1 = 2.832$, which means that if this case were excluded, $b_1 = 1.887$. If X_1 coded a treatment or control condition, then including this case makes the adjusted mean difference in Y between the groups 0.945 units larger than it otherwise would be. But note that this value of $DB(b_1)$ is not particularly large relative to some of the other cases. Observe that cases 2 and 7 have values of $DB(b_1)$ that are not much smaller than 0.945 in absolute value. And whether case 8 is included or excluded does not influence whether we claim a statistically significant partial association between X_1 and Y in these data.

16.1.5 Generating Regression Diagnostics with Computer Software

Most good regression programs have options for saving and displaying various regression diagnostics for examination and analysis. Different programs use different labels in the code for generating the same statistic, so take a close look at your program's manual to make sure that you understand what is being generated.

The SPSS command below will generate all the regression diagnostics we have discussed in this chapter.

```
regression/dep=y/method=enter x1 x2 x3/
    save pred resid dresid sresid sdresid cook mahal leverage dfbeta.
```

The options following the **save** command produce, respectively, $\hat{Y}_i$, e_i, ${}_d e_i$, str_i, tr_i, $Cook_i$, MD_i, $h_i - (1/N)$, and $DB(b_j)_i$. These diagnostics are inserted into the data file, though not in this order. Note that SPSS produces something called the "centered leverage" rather than h_i. To convert centered leverage to h_i, add $1/N$ to the centered leverage. SPSS labels some of these diagnostics differently than we have. For instance, what we are calling the t-residual, SPSS calls the "studentized deleted residual."

The SAS code below accomplishes something similar:

```
proc reg data=chap16;
    model y=x1 x2 x3/influence;
    output out=ch16diag p=pred r=resid student=str rstudent=t
    cookd=cook h=h;run;
proc print data=ch16diag;run;
```

This code produces a new file (named "ch16diag" in the code above) containing values for each case for all regressors and Y_i as well as $\hat{Y}_i$, e_i str_i, tr_i, $Cook_i$, and h_i, and prints these values on the screen. The influence option following the model command outputs (though does not save) $DB(b_j)_i$ values, though these are expressed in standardized form, meaning standard errors from the estimate of b_j. See the SAS documentation for guidance.

STATA can also generate diagnostic statistics from a regression analysis. For instance, the code below generates $\hat{Y}_i$, e_i, h_i, str_i, tr_i, and standardized $DB(b_j)_i$. The text prior to the comma provides a variable name for the diagnostics saved into the data file. The **list** command prints the diagnostics on the screen.

```
regress y x1 x2 x3
predict pred,xb
predict resid,residuals
predict h,hat
predict str,rstandard
predict tr,rstudent
predict dbb1,dfbeta(x1)
predict dbb2,dfbeta(x2)
predict dbb3,dfbeta(x3)
list pred resid h str t dbb1 dbb2 dbb3
```

The RLM macro described in Appendix A will produce all the diagnostics discussed in this chapter, except for the *dfbeta* values, by adding the **diagnose=1** option to the RLM command. The diagnose option also generates output showing the minimum and maximum values of the regressors and the outcome, $\hat{Y}$, and a few of these diagnostics. See the documentation in Appendix A.

16.2 Detecting Assumption Violations

In Chapter 4 we introduced the assumptions of linearity, normality, and homoscedasticity. In this section we describe some approaches to detecting

violations of these assumptions. These assumptions can be tested individually or they can be tested as a set, though testing them as a set provides only the vague conclusion that an assumption is violated without specifying which one.

16.2.1 Detecting Nonlinearity

Under the assumption of linearity, the expected value of the errors in estimation of Y for any combination of regressors is zero. Residuals can be used to determine whether the linearity assumption is violated, but none of the methods based on a residual analysis that you will find described here or in other books is likely to be as good at detecting nonlinearity as the methods discussed in Chapter 12.

In section 2.4.4 we provide an example of a nonlinear relationship, depicted in Figure 2.7 and replicated here in this section in Figure 16.2, panel A. The best-fitting line of the form $\hat{Y} = b_0 + b_1 X$ is found superimposed on the scatterplot. Notice that for both relatively large and relatively small values of X, the residuals are predominantly negative, but for moderate values of X, the residuals are predominantly positive. Figure 16.2, panel B, depicts the t-residuals generated from $\hat{Y} = 3.289 - 0.220X$, the best-fitting linear regression line, against X (the solid line in Figure 16.2, panel A). Notice the obvious pattern, with negative residuals for extreme values of X and positive residual in the middle of X. This kind of pattern, with residuals that are systematically positive or negative in certain ranges of the regressor, suggests that the relationship between X and Y is not well described as linear. Figure 16.2, panel C, is a comparable plot of t-residuals from the quadratic model $\hat{Y} = 1.254 + 1.587X - 0.359X^2$. The quadratic model itself is depicted with the dotted line in Figure 16.2, panel A. In the scatterplot of t-residuals against X, there appears to be no systematic tendency for residuals to be positive or negative in certain ranges of X, suggesting that any nonlinearity that does exist in the relationship between X and Y is well described by the quadratic model.

For models with more than one regressor, comparable plots of residuals, such as those in Figure 16.2, can be generated with $\hat{Y}$ on the X-axis. Alternatively, a residual scatterplot can be used to check for evidence of partial nonlinearity. For instance, if you are concerned that the partial relationship between X_1 and Y is nonlinear when you control for X_2, you can regress Y on X_1 and X_2, generate the residuals from this regression, and then plot the residuals against X_1, looking for evidence of nonlinearity in the plot.

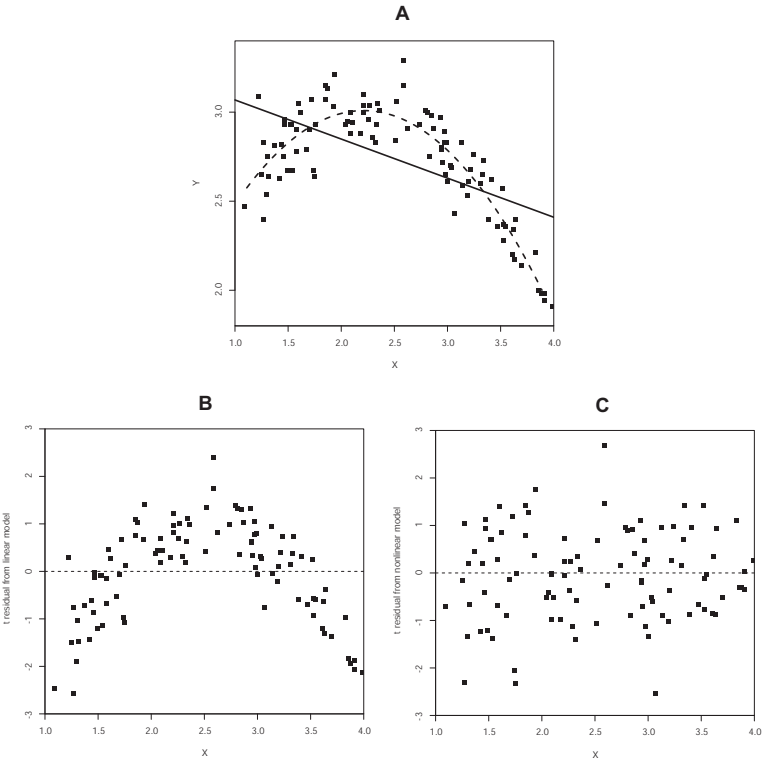

FIGURE 16.2. A nonlinear relationship (panel A) and the *t*-residuals from a model without (panel B) and with (panel C) the square of *X* as a regressor.

Intuition tells us that the conclusions we reach with an "eyeball" test of nonlinearity should be treated with a grain of salt. Looking at scatterplots such as these will tend to reveal only obvious nonlinearity, such as in this example. More subtle nonlinear relationships, such as a shallow curve, are not likely to be detected with the eye. There is also the possibility that your brain may detect a pattern in what is really just a random dispersion of the residuals in the plot. Systematic tests of nonlinearity described in Chapter 12 not only are superior, for they may detect nonlinearity we may not see, but also protect us from misinterpreting random variation as nonlinearity.

16.2.2 Detecting Non-Normality

Regression analysis assumes that the conditional distributions of Y are normal or, equivalently, that the errors in estimation of Y are normally distributed conditioned on the regressors. Some authors recommend constructing a histogram of the residuals (either e_i or tr_i) and eyeballing the histogram to see if you can detect evidence of non-normality. The two problems with this approach are just as described in the section on detecting nonlinearity—that we often see non-normality that really is just random variation, or we fail to see real non-normality when it exists. The eye is good at detecting only obvious non-normality, just as it is good at detecting only obvious nonlinearity. The second problem is that a histogram of the residuals reflects only the marginal distribution of the errors in estimation, ignoring the conditioning that is part of the assumption. The counter to this concern is that if the marginal distribution of the errors in estimation is non-normal, mostly likely so too is one or more of the conditional distributions.

There are formal tests of non-normality of the errors in estimation that one could apply. But they can detect non-normality that is trivial and not likely to affect the accuracy of the inferences one is making with a regression analysis. In Chapter 12 we discussed various transformations that can be used to reduce nonlinearity in relationships that also can have the effect of reducing non-normality in errors in estimation. But they carry with them the disadvantage that transformed metrics may be harder to interpret, and it can be perceived by potential critics as arbitrary and used in an attempt to make results cleaner than they actually are.

Our perspective is that unless you see clear evidence of fairly extreme non-normality in the residuals and have ruled out the existence of clerical errors and highly influential cases using the methods discussed in section 16.1.3, don't worry too much about all but extreme violations of normality. It turns out the normality assumption is one of the least important of the assumptions of regression for most of the widespread uses. You might also consider verifying that your results replicate when using one of the methods we discuss in section 16.3 that make weaker assumptions about the errors in estimation. But if the non-normality is inherent in the system of measurement of Y, such as the result of using a single-item ordinal response scale (e.g., *strongly disagree*, *disagree*, *agree*, *strongly agree*) or small counts of things (e.g., how many televisions a person has), consider learning about one of the methods discussed in Chapter 18 designed for the modeling of

ordinal, discrete, or count outcomes, which are non-normal by definition or turn out to be so in most applications.

16.2.3 Detecting Heteroscedasticity

In most simple terms, homoscedasticity means that the conditional distributions of Y have equal variances. The assumption is most easily described in the context of simple regression and states that ${}_T\text{Var}(Y.X)$ is the same regardless of X. Because the conditional distribution of Y is centered around $\hat{Y}$, the assumption can also be expressed in terms of the variance of the errors in estimation ${}_T\text{Var}(e.X)$. Figure 16.3, panel A, depicts a sample of 500 cases from a population regression model $\hat{Y} = 5 + 0.25X$ with homoscedastic errors. As can be seen, there is no apparent pattern in the distribution of the residuals or, alternatively, the conditional distribution of Y given X. The residuals appear roughly equally dispersed around the regression line. It appears that the dispersion of Y given X is the same regardless of X.

In the description above, as well as what follows below, we can replace X with $\hat{Y}$, which, of course, is a linear combination of k values of X_j, the regressors in the model. That is, the assumption pertains to the conditional distribution of Y for the linear combination of k values of X_j that is $\hat{Y}$.

Violation of this assumption is known as *heteroscedasticity*. The most common type of heteroscedasticity occurs when ${}_T\text{Var}(Y.X)$, the true conditional variance of Y given X, is largest for the highest or lowest values of some regressor or combination of regressors, a situation we could call *ordinary* heteroscedasticity. Figure 16.3, panel B, depicts such a situation, where the variability of Y and therefore e_i is larger for higher values of X or $\hat{Y}$. Two alternative forms of heteroscedasticity are *butterfly* heteroscedasticity, as in Figure 16.3, panel C, and *inverse butterfly* heteroscedasticity, as in Figure 16.3, panel D. In butterfly heteroscedasticity, the conditional distribution of Y is larger at more extreme values of X or $\hat{Y}$, and in inverse butterfly heteroscedasticity, variability in Y is largest in the middle of the distribution of X or $\hat{Y}$.

In Figure 16.3 we place Y and X on the axes of the figures. But you could replace the Ys with residuals to produce partial scatterplots (see, e.g., Figures 3.10 and 3.12). When testing the significance of the regression coefficient for X_j or producing confidence intervals for ${}_Tb_j$, we would assume *partial homoscedasticity*, meaning that the variance of the errors in the estimation of Y when controlling for all regressors but X_j is uncorrelated with X_j when holding all other regressors constant.

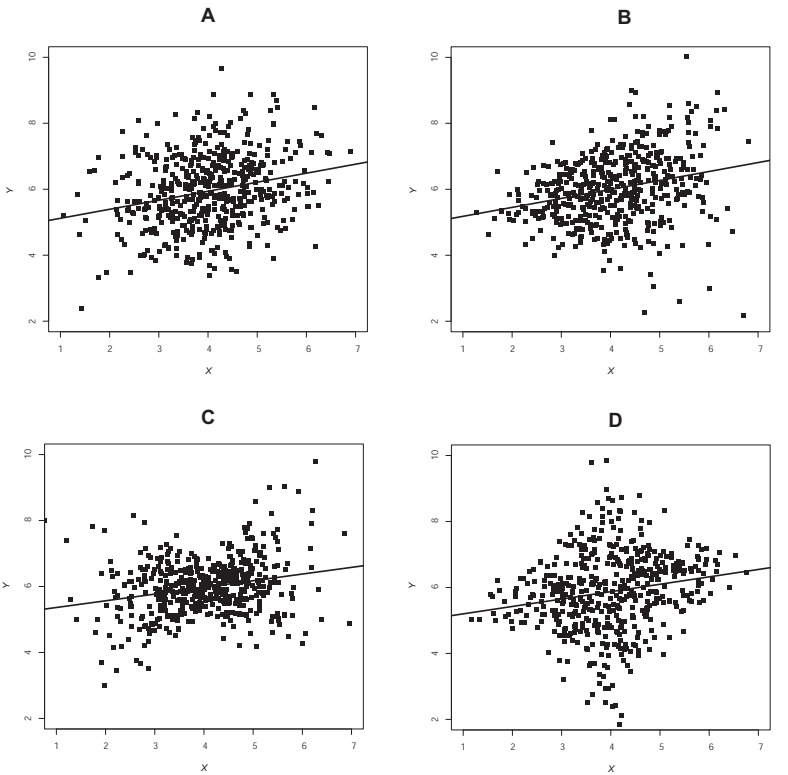

FIGURE 16.3. Scatterplots and the linear regression of Y on X reflecting homoscedasticity (panel A), ordinary heteroscedasticity (panel B), butterfly heteroscedasticity (panel C), and inverse butterfly heteroscedasticity (panel D).

Heteroscedasticity can occur in a number of ways. One way is the existence of an interaction involving one regressor and another variable that may or may not be one of the regressors in the model. For instance, we could imagine in Figure 16.3, panels B or C, drawing two lines relating X to Y, one for group A and another for group B, that differ in slope but are opposite in sign. By ignoring the existence of two subpopulations, each with a different relationship between X and Y, and estimating a single regression coefficient for X can produce a pattern of residuals or conditional distributions of Y that look like those in panel B or C. Including an interaction in the model can eliminate heteroscedasticity.

A second situation that can produce heteroscedasticity occurs when the population of interest is composed of two subpopulations, with one ranging more widely than the other on all the variables. For instance, consider natives born in a specific country and immigrants not born there. The immigrants will typically be from many nations and continents, while the natives by definition are from one. The immigrant subpopulation is likely to be more heterogeneous on many variables than the native subpopulation. When subpopulation A is more heterogeneous than subpopulation B on all variables in the analysis, then the extremes of each regressor will be dominated by group A, which also has a greater variance on Y than group B. This will produce butterfly heteroscedasticity. It can be shown that when a population consists of two equal-size bivariate normal populations, and ${}_TS_X$ and ${}_TS_Y$ are twice as large in one subpopulation as in the other but the simple regression coefficients relating Y to X are equal, ${}_TSE(b_1)$ is 17% larger than the value calculated from the regression formula ordinarily used to calculate $SE(b_1)$. This does not sound like much, but if the actual standard error is 17% larger than what your calculations show, then the probability of finding a significant association between X and Y when there is no real association is nearly twice as high as the α-level being used for the test.

A third situation that can produce heteroscedasticity occurs when Y is measured with less random error for certain cases than others that differ on the regressors. We discuss random measurement error in section 17.2. Suffice it to say now that random measurement error tends to increase a variable's variance relative to what it would be if that variable were measured without random error. If people who score higher (or lower) on X have their Y measured with more error, heteroscedasticity is the result.

Heteroscedasticity can also result when modeling discrete count outcomes using ordinary least squares regression. If your dependent variable were something like the number of times a person donated to political candidates in the last year, Y would be dominated by zeros and 1s, a few 2s, fewer 3s, and so forth. In a least squares linear model of a count Y such as this, the conditional variance of Y is typically positively correlated with its expected value. That means the variance of the errors in estimation will tend to be larger for people who are estimated by the model to donate more often.

Heteroscedasticity does not bias regression coefficients. Rather, heteroscedasticity exerts its influence on inference in regression analysis primarily through its effects on the estimates of the standard errors of the regression coefficients. Ordinary and butterfly heteroscedasticity tend to

result in estimates of standard errors that are too small. This produces confidence intervals that are too narrow and hypothesis tests for regression coefficients and ${}_{T}R$ that are invalid. Inverse butterfly heteroscedasticity tends to result in estimates of standard errors that are too large. This produces confidence intervals that are too wide and hypothesis tests that are lower in power than they otherwise would be if homoscedasticity were met.

When a regressor is dichotomous, we can talk about the conditional variance of Y in each of the two groups. Heteroscedasticity has its biggest effect on the standard error for the regression coefficient for a dichotomous regressor when the groups are different in size. When the smaller group is more variable on Y, the standard error for the dichotomous regressor tends to be underestimated, but when the smaller group is less variable on Y, the standard error tends to be overestimated.

Given that the quality of our inferences in regression analysis are dependent on the quality of our estimates of standard errors (since standard errors determine confidence interval width and p-values), it is worth testing for its existence so you can make an informed decision about how to proceed. There are many tests of heteroscedasticity that have been described in the regression analysis literature (e.g., Breusch & Pagan, 1979), and you may be familiar with some from the ANOVA literature, such as Levene's test. These generally require some belief about the nature of the heteroscedasticity (e.g., the variance in Y increases with X) or they make assumptions, putting you in the awkward predicament of wondering whether the assumptions of your test of assumptions are met.

Rather than describing these tests, of which there are several, we provide a fairly simple method of testing for ordinary and butterfly heteroscedasticity that can be conducted with any regression program that allows you to generate and save t-residuals, as most do. The test relies on the fact that under the standard assumptions of regression, $E(tr_i^2)$, the variance of the t-residuals, tr_i, is identical for all values on all regressors. So a significant association between tr_i^2 and any regressor or set of regressors is evidence for heteroscedasticity, and we can test for heteroscedasticity by testing the independence of tr_i^2 from the regressors.

The form of this test we advocate requires *normalizing* tr_i^2, which forces its distribution to one approximately normal in form. This process involves replacing the values of tr_i^2 with their rank position in the distribution, such that the smallest squared t-residual gets a value of 1, the next smallest a value of 2, and so forth, up to N. The rank order of ties can be determined

arbitrarily, or they can each be assigned the mean rank for which they are tied. Most statistical programs have a command for replacing scores with their rank position in the distribution.

With these ranks derived, divide them all by $N+1$ and then replace these with the value from the standard normal distribution that cuts off the lower $100 \times [\text{rank}/(N+1)]\%$ of the normal distribution from the rest. These values can be found with the help of Appendix C, or they can be derived by your software. For example, if $N = 19$, then dividing the ranks 1 through 19 by 20 yields .05, .10, .15, and so forth, up to 0.95. From the standard normal distribution, these convert Z-scores of $-1.645, -1.282, -1.036, \ldots, 1.645$. These resulting normalized values or Z-scores are known as Van der Waerden scores.

These Z-scores are roughly normally distributed, and under the assumption of homoscedasticity they are independent of all the regressors. So to test for ordinary heteroscedasticity, we regress these Z-scores on all the regressors in the model and test the significance of the multiple correlation. If R in this regression is statistically significant, then the homoscedasticity assumption is violated. If you are particularly interested in certain regressors, you would look at the t-statistic for those regression coefficients in this regression with the Z scores as the dependent variable. A significant regression coefficient implies partial heteroscedasticity.

This approach only tests for ordinary heteroscedasticity. You could also test for butterfly or inverse butterfly heteroscedasticity by including the squares of numerical regressors in this model at the same time. A nonsignificant R would suggest no violation of the homoscedasticity assumption, whereas a significant R could mean either ordinary, butterfly, or inverse heteroscedasticity. But you could test collectively for any butterfly or inverse butterfly heteroscedasticity by testing all squared terms as a set using the method described in section 5.3.3, or you could test for heteroscedasticity due to a specific variable by testing the significance of the set defined as that variable's unsquared and squared terms. If a variable's squared term is nonsignificant, you could drop it and reestimate the model, examining the partial regression coefficient for that regressor as a test of ordinary partial heteroscedasticity, while allowing for butterfly or inverse butterfly heteroscedasticity involving other regressors that still have their squared terms in the regression.

We illustrate by testing for heteroscedasticity in the self-censorship analysis from section 10.2.4. Recall in that example we estimated a person's willingness to self-censor from his or her age and shyness. Age was a mul-

ticategorical variable with four ordinal age categories (Generation X, Generation Y, baby boomer, pre-baby boomer). In the code below we assume that three indicator variables coding age cohort are already constructed and held in variables *d1*, *d2*, and *d3*. The SPSS code below generates and normalizes the squared *t*-residuals and regresses these normalized residuals on age cohort, shyness, and the square of shyness.

```
regression/dep=wtsc/method=enter d1 d2 d3 shy/save sdresid.
compute trsq=sdr_1*sdr_1.
rank variables=trsq.
compute rtrsq=rtrsq/462.
compute z=idf.normal(rtrsq,0,1).
compute shysq=shy*shy.
regression/dep=z/method=enter d1 d2 d3 shy shysq.
```

In STATA, use

```
regress wtsc d1 d2 d3 shy
predict tr,rstudent
gen trsq=tr*tr
egen rtrsq=rank(trsq)
replace rtrsq=rtrsq/462
gen z=invnormal(rtrsq)
gen shysq=shy*shy
regress z d1 d2 d3 shy shysq
```

The SAS code below does the same analysis, assuming that the data file containing the regressors and the *t*-residuals are in a file named "ch16diag." SAS has a special procedure built into PROC RANK for generating Van der Waerden scores, which the code below utilizes.

```
data ch16diag;set ch16diag;trsq=t*t;shysq=shy*shy;run;
proc rank data=ch16diag normal=vw ties=mean out=ch16diag;
    var trsq;run;
proc reg data=ch16diag;
model trsq=d1 d2 d3 shy shysq;run;
```

From this analysis, $R = 0.135$, $F(5, 455) = 1.695$, $p = .134$, meaning we fail to reject the assumption of homoscedasticity. But this omnibus test doesn't preclude the possibility of partial heteroscedasticity. The regression coefficient for the square of shyness was not statistically significant, meaning no butterfly heteroscedasticity involving shyness. When the squared term

was removed and the model reestimated, none of the regression coefficients were statistically significant, nor was $R, F(4, 456) = 2.104, p = .079$. We also looked for evidence of partial heteroscedasticity involving age by adding the three indicator variables coding age to the model that already contained shyness (only the linear term). The increase in R was not statistically significant, $F(3, 456) = 1.744, p = .157$. Combined, these analyses support the conclusion that the homoscedasticity assumption is met.

In the next section we describe a test on the whole set of standard assumptions that can be performed by applying a Bonferroni correction to the p-value of the highest t-residual. But that test is not nearly as powerful at detecting heteroscedasticity as the test we just described. In 1,000 bivariate samples of size 50 from artificial populations with butterfly heteroscedasticity, the test we describe next failed to discover the problem in 371 samples, while the test described in this section failed in only six samples.

16.2.4 Testing Assumptions as a Set

In the prior pages we described some methods for examining the plausibility of the assumptions of linear regression analysis. We can conduct a more general test of the null hypothesis that none of the assumptions is violated against the alternative that at least one is violated. Perhaps the simplest method for detecting a violation of this set of assumptions relies on the distribution of t-residuals. As discussed in section 16.1.3, these follow an exact t-distribution under the standard assumptions of regression. Using the $t(df_{residual})$ distribution, one can derive a two-tailed p-value for tr_i.

The p-value for each t-residual is sometimes misinterpreted as testing the null hypothesis that case i falls on the true regression line. If that were so, then the proportion of significant t-residuals would approach 1 as N increases since almost no cases in fact fall exactly on the true regression line. But if the standard assumptions hold, we expect only 5% of the t-residuals to be significant at the .05 level, no matter how large the sample. The hypothesis tested using the p-value for each t-residual is actually that Y_i falls within a normal distribution of scores around the regression line. But because the number of residuals is N, a Bonferroni correction should be applied to the p-value for each tr_i to compensate for the fact that we are doing N hypothesis tests in search of something statistically significant. So the largest t-residual in absolute value is considered statistically significant only if its significance level is below some chosen α-level, such as 0.05, even after being multiplied by N. A statistically significant residual after

this Bonferroni correction suggests that at least one of the assumptions of regression is violated without specifying which one.

Most statistical packages have a command for generating the *p*-value for a *t*-statistic, so this test is fairly easy to implement once you have generated the *t*-residuals as discussed in section 16.1.5. Section 11.2.5 provides SPSS, SAS, and STATA code for generating *p* from *t*. This test is implemented in the RLM macro described in Appendix A. It provides output containing the largest *t*-residual and its Bonferroni-corrected *p*-value.

Another test of the standard assumptions of regression does not rely so heavily on individual *t*-residuals and may be considerably more powerful for detecting any violation that affects many residuals somewhat without affecting any single residual too greatly. In this test we pick some arbitrary probability, count the number of *t*-residuals that are statistically significant at this level (without a Bonferroni correction), and use the binomial distribution to test whether this number is greater than would be expected by chance. For instance, in a sample of 50 cases, by chance we would expect five *t*-residuals to be statistically significant with a *p*-value of no greater than .10. If we observe 11 such residuals, the binomial distribution tells us that that the probability of observing so many is only .0094; this indicates that at least one of the standard assumptions must be violated. The binomial test is not perfectly accurate for this use, for it assumes that the *N* residuals are statistically independent, and they are not quite independent. But in tests we have run, the error is small.

16.2.5 What about Nonindependence?

We have not yet addressed the assumption of independence. The assumption of independence pertains to the size of the errors in estimation—that there is no relationship between the error in the estimation of *Y* for case *i* and the error in estimation of *Y* for case *j*. This assumption can be harder to test than other assumptions and is probably routinely violated. Nonindependence can creep into a study in all kinds of ways if you aren't careful about your sampling, study design, and data collection procedure.

Nonindependence can have various effects on statistics from a regression analysis, but its effect on standard errors is one of the bigger concerns. Research shows that violation of the independence assumption can result in standard errors for regression coefficients that are too large or too small, but in most circumstances the result will be underestimation. As a result, confidence intervals will be too narrow and *p*-values inappropriately small when this assumption is violated.

To understand why, consider two studies identical in purpose but different in method. Suppose you are interested in comparing men and women in their attitudes toward a controversial social topic, such as gun control. You decide to ask 200 people their attitudes by randomly visiting houses in a city and asking the opinions of everyone at home at the time you visit. Because some houses have more than one person in the home, you won't need to visit 200 houses, but this doesn't change the fact that you will still end up talking to 200 people. Once you have talked to 200 people, you can then conduct a test comparing the attitudes of the men and women you ask.

Now consider a variation on this method, where you actually visit 200 houses because you decided to talk to only one of the people living at each house you visit. This may take more time than the variation of this study just described, but at the end you'll have 200 responses, and you can compare the responses of men and women, just as in the prior version.

In both variants of this study, $N = 200$. But the latter study contains more information about how men and women differ, because the responses of the men and women are more likely to be independent, with the caveat we describe later. Its *effective sample size* is 200 or nearly so, but the former study's effective sample size would be much smaller than 200. In the former study, people living together are likely to have similar attitudes, because we know that people influence each other, they selectively sort themselves into social groups based on similarity in beliefs, and they are more likely to be attracted to and partner with people who are like themselves. So if you were to regress a person's attitude about gun control on an indicator variable coding sex in order to test for sex differences, the errors in estimation of Y are not likely to be independent between people living in the same house. But this is not a problem in the second version of the study, because you have data from only one person in each house. The consequence is that we would expect the standard error for sex to be smaller in the first version of the study, because it is treating the 200 people as if they are providing independent information about variability between people in their attitudes. Although 200 people were asked about their attitudes, we don't have 200 independent measurements of those attitudes.

The problem with the analysis from the first version of the study is not easily fixed after the fact without relying on more complicated regression methods. Although you could include a set of indicator variables to code the house a person lives in, this would consume many degrees of freedom and could drastically lower the power of hypothesis tests.

Earlier we said that the version of the study based on only a single person interviewed at a selected house is more likely to produce independent responses than if everyone at the house were interviewed. This is true, but even then, nonindependence may exist. For instance, people living on the same street may know each other, talk to each other, and influence each other's attitudes. So two people living on the same street may give nonindependent responses even if they don't live in the same house. Or maybe people who are politically liberal are more likely to live on Equality Street, whereas politically conservative people are more likely to live on Liberty Street. Even if no one talks to his or her neighbors, the errors in estimation of a person's response may be related to errors in estimation for people living on the same street.

Or suppose you were to randomly call 500 people living throughout the United States to provide data on some variable of interest. This is common in survey research and public opinion polling. Such a sampling plan might seem like the epitome of a method that would satisfy the independence assumption. But people living in the same state or city might be more similar to each other on the variable you are measuring than people living in different states. Technically, this is a violation of independence, although researchers rarely do or even think much about it. And it is common in experimental research to collect data from people in groups. For instance, perhaps you are presenting stimuli to people on a computer screen, and to save time, you recruit five people at a time and sit them in front of different computers in the same room to collect data from them at the same time. But are their responses likely to be independent? Perhaps, but suppose that the dependent variable is affected by the temperature of the room. If the temperature of the room fluctuates from day to day or even hour to hour, this can produce nonindependence in the errors of estimation of Y between subsets of people in the room at the same time their data were collected. Obviously, if these people are allowed to interact during the study, this can also produce nonindependence, especially if they talk about the study itself, their responses to the questions, and so forth, as the data are being collected.

Although you may not be able to completely avoid or eliminate nonindependence, you can at least be conscious of its possible existence and try to reduce it through choices made about sampling and study design. After the fact, it is hard to eliminate unless you have a good idea of where it comes from. Of course, some methods you are already familiar with are designed with nonindependence in mind. An example is the paired-

samples *t*-test, which is designed for comparing the means of Y among people who are "matched" and hence nonindependent. There are some tests of independence that can be used for certain types of sampling and research designs, and there are special analytical methods that are well suited to modeling data that are likely to be nonindependent in some way, such as *multilevel modeling*. For a discussion of some of these methods and nonindependence more generally, as well as ways of quantifying nonindependence, see Griffin and Gonzales (1995), Grawitch and Munz (2004), Kenny and Judd (1986), Kenny, Mannetti, Pierro, Livi, and Kashy (2002), Luke (2004), O'Connor (2004), and Raudenbush and Bryk (2002).

16.3 Dealing with Irregularities

Neither heteroscedasticity nor non-normality affects the expected values of b_0, b_j, and $MS_{residual}$, so these statistics provide unbiased estimates of ${}_T b_0$, ${}_T b_j$, and ${}_T\text{Var}(Y.X)$ even in the presence of these conditions. But hypothesis tests and confidence intervals can be invalidated by violations of any of the standard assumptions. Thus, you typically should do something about cases suggesting violations of the standard assumptions.

But what do you do? There are many exceptions, but generally your four options are correction, transformation, elimination, and robustification. They are normally considered in that order. *Correction* refers simply to the correction of clerical errors. *Transformation* means applying a logarithmic or other transformation to a variable—either a regressor or the dependent variable—so that the case is no longer so extreme. *Elimination* means eliminating the case from the sample. *Robustification* means replacing the regression analysis by an alternate method less sensitive to extreme cases. Correction of clerical errors needs no discussion here, and transformations were discussed in Chapter 12. In the rest of this section, we discuss elimination and robustification.

When you eliminate a case simply because it is extreme in some sense, you are essentially adding a major qualification to your conclusions. You are admitting that the conclusions apply only to the subpopulation defined as the population of cases that exclude extreme cases like the one or ones you eliminated. At least four questions are left unanswered: (1) how the studied subpopulation differs from the rest of the population, (2) how large the included and excluded subpopulations are, (3) how the independent variables relate to the dependent variable in the excluded subpopulation, and (4) whether these relationships in the excluded subpopulation might

be so large as to make the relationships in the studied subpopulation irrelevant. Nevertheless, an extreme score may be the major available clue that a participant in a study did not understand the directions he or she was given in a survey or experiment, or that the experimental manipulation was done improperly for that one participant, or may in other ways provide a defensible reason for discarding the participant's data. Thus, elimination may be a reasonable choice. This is especially true if post hoc examination of the case reveals something odd about it—for instance, evidence that a person did not understand experimental directions. But for the reasons mentioned, elimination may sometimes be a reasonable choice even without such evidence.

There are two general types of robust approach. One set of approaches uses alternative methods for estimating the regression coefficients. The other uses ordinary formulas for the regression coefficients but some alternative method for calculating significance levels or estimates of standard errors. The former approaches essentially give less weight to outliers. This raises fundamental questions about the purpose of the regression. After all, down-weighting an outlier can lead to a regression solution that fails to represent adequately the fact that such outliers do occasionally occur. So we shall consider only the second approach, in which the investigator uses ordinary regression formulas to derive the best-fitting model, but employs an alternative method to find standard errors, confidence intervals, or p-values.

We consider four methods: heteroscedasticity-consistent standard errors, the jackknife, bootstrapping, and permutation tests. All of these are practical only with computers, but with the right software they take anywhere from a few seconds to a few minutes on an ordinary personal computer. None of these are panaceas for problems produced by various irregularities such as assumption violations. Even these methods are nonrobust in certain circumstances too numerous and complicated to outline here. Each has variants we do not describe to deal with some of the weaknesses of other variants. The point of our discussion below is to outline a bit about how these methods work, not to describe all the forms they take or offer recommendations as to the specific circumstances in which you might choose to use them. Each of these methods has been heavily studied. General overviews can be found in Edgington (1995), Efron and Tibshirani (1993), Good (2001), Lunneborg (2000), and Rodgers (1999).

16.3.1 Heteroscedasticity-Consistent Standard Errors

The formula for the standard error of a regression coefficient in section 4.4.3 that is implemented in most regression analysis programs assumes homogeneity in the variance of the errors in estimation. This assumption justifies the use of $MS_{residual}$ in the numerator of the formula as an estimate of the conditional variance of Y, which is assumed to be equal for all combinations of regressors.

There is a family of *heteroscedasticity-consistent* (HC) standard error estimators for the regression coefficients that do not require this assumption. They are known as *sandwich estimators* in the statistics literature, because their formulas in matrix algebra look like a sandwich, with the matrix of values on the regressors as the "bread" and the residuals, usually squared and possibly weighted in some fashion by each case's leverage, serving as the "meat." They are called HC estimators, because unlike the usual OLS standard error estimator, which is biased and does not converge with increased sample size to the proper value when the homoscedasticity-assumption is violated, the HC estimators approach the correct value with increasing sample size even in the presence of heteroscedasticity. In statistics, the converging of an estimator to its correct value with increasing sample size is a property called *consistency*.

Use of one of these standard errors does not require modifying the mathematics to estimate the regression coefficients. Rather, the usual standard error estimator is simply replaced with a HC standard error estimator. There are many forms HC estimators take, the earliest frequently attributed to White (1980) and often called the White or Huber–White estimator and denoted HC0. This early version has been improved into forms labeled HC1, HC2, HC3, and HC4. They defy nonmathematical description. We offer the formula for HC3 in matrix algebra form in Appendix D. Otherwise, see Cribaro-Neto (2004), Hayes and Cai (2007), and Long and Ervin (2000) for details about their computation and examples of application.

When heteroscedasticity is a concern, one of these estimators can provide more solid footing. But Long and Ervin (2000) make a case for the regular use of one of these standard errors even when the homoscedasticity assumption is met. This is because they tend to perform better when the homoscedasticity assumption is violated, regardless of the form heteroscedasticity takes, than the standard error estimator that assumes homoscedasticity. Research shows HC3 and HC4 tend to work best. Importantly, these standard error estimators work well even when the homoscedasticity assumption is reasonable. Given that these estimators are easy to compute

and are even available in some software packages (all these HC estimators are available in the RLM macro for SPSS and SAS described in Appendix A; STATA and SAS offer several of them as well), perhaps one day researchers will rid themselves of the homoscedasticity assumption and use one of these estimators for inference in regression analysis as a matter of routine.

16.3.2 The Jackknife

The jackknife, or *jackknifing*, was given its name by J. W. Tukey on the grounds that it may not be the very best tool for anything at all, but it's a serviceable tool in a great many situations. To jackknife a statistic or a test, divide the sample into g groups of equal size, where g is at least 10. In fact, in practice g is frequently set to N, so each "group" contains only one case. Then compute the statistic of interest after deleting group 1 from the sample; then add group 1 back in, delete group 2, recompute the statistic; then add group 2 back in, delete group 3, recompute the statistic; and continue in this manner through all g groups. At the end of this process, you will have g estimates of the statistic of interest. The standard deviation of these g estimates can be used to compute the standard error of the original statistic. Inference can then proceed in the usual way, by constructing a confidence interval using this jackknife estimate of the standard error. Or you could divide the observed statistic by this standard error and generate a p-value for testing the null hypothesis that the corresponding parameter equals zero using the normal distribution.

16.3.3 Bootstrapping

Like the jackknife method, the bootstrap method has been suggested for inference for virtually any statistic. It is based on a simple idea documented in Efron and Tibshirani (1993). If we make absolutely no assumptions about the nature of the population distributions of the variables measured, then the distribution of the measurements in the sample is in every respect the best estimate of the population distribution. That is, if our sample size is 50, then our best assumption-free estimate of the population distribution of the variables measured is that 1/50th of the cases are exactly like case 1, another 1/50th are exactly like case 2, and so on. We then draw, say, B independent random "bootstrap samples" of size 50 from this imaginary population, where B is some large number. This sampling of the original data is done with replacement, so that the bootstrap sample data set does not just reproduce the original data. We then compute the statistic(s) of

interest in each of these bootstrap samples, giving us B estimates of the corresponding parameter.

In one version of the bootstrap method, we then calculate the standard deviation of all B values of each statistic and use that as the estimated standard error of that statistic. As with the jackknife, the normal distribution is then ordinarily used to test hypotheses about the statistic or to find a confidence interval. B does not need to be particularly large when using bootstrapping in this way. Usually 100 or 200 bootstrap samples will do.

In the other version of the bootstrap we never compute a standard error but base our inferences on the number of bootstrap samples yielding statistics in various ranges. This requires a larger value of B—at least 1,000, but more is better. For example, if b_j is positive in the original sample, the proportion of the bootstrap estimates of b_j that yield negative values of b_j can serve as the significance level for testing the null hypothesis ${}_T b_j > 0$. Alternatively, a confidence interval for ${}_T b_j$ can be constructed by using the percentiles of the distribution of B values of b_j. For instance, for a 95% confidence interval, the lower and upper endpoints are defined as the 2.5th and 97.5th percentiles of the distribution of B bootstrap estimates of b_j.

16.3.4 Permutation Tests

Consider a simple correlation r_{XY} based on a sample of size N. Suppose we were to take the N measurements of Y and randomly match them with the N values of X and then recompute r_{XY}. Imagine doing this 999 times, so we have 1,000 values of r_{XY} including the original one. Suppose we find that the original correlation is the 28th-highest of all 1,000 values. We can then say that if these X scores had been matched randomly with these Ys, the probability is only 28/1,000, or .028, that the original correlation r_{XY} would have ranked so high. This value .028 is the one-tailed significance level p for the obtained correlation; it is a *permutation* or *randomization test* of random association. If we ignored the sign of r_{XY} both in the original data and in all 999 recomputed correlations, then a two tailed p-value is the proportion of the absolute values of the 1,000 correlations that are at least as large as the original absolute correlation.

In this example we held constant the order of measurements on X and randomly reassigned values of Y to those X values. In multiple regression we can hold constant the entire matrix of regressor scores, rescramble the order of the Ys many times, and recompute R and all values of b_j each time for construction of p-values using the same approach as in the simpler example.

Rescrambling the Ys themselves is actually not as powerful as an alternative method. To see why, suppose b_1 is high positive, and one person has extremely high measurements on X_1 and Y, but this person's measurement on Y is about what we would predict from his or her high measurement on X_1. This high Y will increase the variation across all the rescramblings of every b_j. This is as it should be for b_1, but it will also be true for every other b_j tested. So in testing the unique contribution of any regressor X_j, the most powerful procedure will generally be to use the portion of Y independent of all regressors except X_j. This means we should use a different column of residuals for each X_j, and still another column for testing R. Thus, we should altogether use $(k + 1)$ different columns of Y-residuals when constructing permutation tests for partial regression coefficients.

16.4 Inference without Random Sampling

In section 6.1.3 we mentioned briefly that valid statistical inferences may be drawn without random sampling, and even without either random sampling or random assignment. In an example presented there, we pondered a statistical test about the change from one decade to another in the proportion of female professors hired by a particular college. Or suppose a club of 50 local businesspeople contains 30 retailers and 20 others. If 25 of the retailers but only 10 of the others vote to change the bylaws, it is valid to perform a 2×2 test of independence in a cross-tabulation to test for a nonchance association between vote and type of business. But, again, there is no hint of either random assignment or random sampling from a broader population. When used in this way, tests of association test the null hypothesis of random association—the hypothesis that the association observed between two variables is caused solely by chance.

Both these examples could be instances of nonsampling, because there is no sampling at all. In the first example, we might study every professor ever hired by the college, or in the second example, the entire membership of the business club. But it is often difficult to distinguish between nonsampling and nonrandom sampling. For instance, in the second example we might think of the local club as a nonrandom sample of the population of members of other business groups in that city or in the nation. The distinction between nonsampling and nonrandom sampling is unnecessary, as well as ambiguous since the types of conclusions we can draw are much the same under both conditions. So the important distinction we must make is between the presence and absence of random sampling.

Nonrandom sampling and nonsampling are very common in both large-scale and small-scale research. On a small scale, suppose an experimenter posts an ad asking for volunteers to serve as participants in an experiment, and uses the first 20 people who sign up. Those participants are not a random sample from any broader population. But if the experimenter assigns the 20 subjects randomly to conditions, then the experiment has random assignment without random sampling. On a larger scale, many behavioral scientists study the entire population of interest: Analysts at the Educational Testing Service have data from all students who take College Board tests, workers at the American Association of Medical Colleges have data on every applicant to an American medical school, census analysts have data from virtually the entire U.S. population, and so on.

Frick (1998) and Mook (1987) discuss how it is inappropriate to put random sampling on a pedestal, thereby condemning all studies that fail to include it inferior in some way. But others have argued that studies that don't include a random sampling component are "pseudoscientific" (Potter, Cooper, & Dupagne, 1993). We agree with the former perspective. Random sampling certainly has a role to play in the kind of inferences we can make. But as Frick (1998) notes, we should distinguish between inferences about *process* and inferences about *populations*. Most researchers care about *process inference*: what is the process that generates the data and the obtained result? They often care less or not at all about *population inference*: does the result obtained reflect what would have been found if the entire population could have been included in the study? Of course, some people care very much about population inference. Public opinion pollsters who generate poll results you read about in the news are an example. Their business is founded on the importance of solid population inference. But most researchers have different research goals than the typical pollster has.

When a significant association between two variables is found under random sampling, it establishes both the *replicability* and *meaningfulness* of the association. We say an association is meaningful if valid hypothesis tests indicate that chance may be excluded from a list of the possible causes of the association. We say the association is replicable if we can have a certain confidence that a nonzero association will be observed again under specifiable conditions, such as drawing a large second sample from the same population. Finding a statistically significant association under nonrandom sampling establishes the association as meaningful, though not necessarily replicable. This at least allows us to speculate on the causes of

the association, as in the previous examples concerning the college's hiring practices or the business club.

When there is random *assignment* without random sampling, as in the example involving the signup sheet, we can go beyond such speculation. Then the existence of a causal relation can be demonstrated, though its generality or replicability is still unknown. In particular, if scores of a treatment group are significantly above those of a control group, then you have shown that the treatment increases at least some scores. This can be a finding of some interest if the dependent variable is a trait thought to be wholly beyond control, such as baldness—or if the independent variable is thought to be imperceptible, such as infrared light or messages flashed on a screen too fast to be seen consciously.

Conclusions of this sort can sometimes be generalized to a broader population, even without random sampling. This is possible if it is assumed that causation is unidirectional, meaning that exposure to the treatment condition rather than the control does not lower anyone's score on the dependent variable. Then, even without random sampling, we have shown that the treatment increases the population mean merely by demonstrating that it raises at least some scores in the population but doesn't lower any scores.

16.5 Keeping the Diagnostic Analysis Manageable

At the level we have now reached in regression analysis, it may be clear that statistical analysis is as much art as science, and not a set of mechanical do and do-not rules. But some general suggestions on the conduct of diagnostic analysis should be helpful.

We saw in Chapters 12, 13, and 14 that curvilinearity and interaction can distort analyses that ignore them, and the same is true of the various kinds of irregularities considered in this chapter. Thus, all these chapters concern potential complications. When should you check for them? You cannot do everything at once. There is no "right" order of checking for these complications, any more than there is a right order of checking for problems when you buy a used car. But there are three reasons for normally applying diagnostic methods before checking for unanticipated curvilinearity or interaction. First, diagnostic methods can uncover clerical errors, and such errors clearly should be detected as early as possible. Second, at least the basic diagnostic methods are easier, and it is always sensible to do easier things first. Third, experience suggests that diagnostic methods uncover

complications more often than do tests for curvilinearity and interaction, and you want to find any problems as soon as possible.

A diagnostic analysis always concerns a particular regression, so the first step in a diagnostic analysis is to identify the regression analysis you would conduct if there were no irregularities. The diagnostic analysis should focus on that regression.

The next step is to choose particular diagnostic methods and tests. We have described measures or tests for many types of irregularities involving leverage, distance, influence, partial influence, and several kinds of heteroscedasticity. And examination of partial influence and partial heteroscedasticity can be done for each regressor. Thus, the number of possible analyses may be large. You should not try to use every one of these tools in every possible analysis. Rather, you should focus on the three major goals of the diagnostic analysis: to check for clerical errors, to examine previously suspect cases, and to test the standard assumptions of regression. To check for clerical errors, check the cases with the highest scores on overall leverage, distance, and influence. Previously suspect cases should be checked primarily for excessive influence—either total or in part—for the most important regressors.

To test the standard assumptions of regression, nearly every analysis should include the Bonferroni-corrected test on the highest t-residual. In addition, tests for ordinary and butterfly heteroscedasticity described in section 16.2.3 should be routinely conducted. The exception might be if you choose to use a heteroscedasticity-consistent standard error estimator for inference, but even then, it isn't a bad idea to test for heteroscedasticity, because its detection could reveal things about the model that could be modified, such as including a missing interaction. And if the major focus of the analysis is on a single regression coefficient b_j, then pay special attention to things that might affect the quality of the estimate of b_j or inference about ${}_T b_j$. For any of these tests, absence of significance does not prove the assumptions hold, but at least violations of the assumptions have been given a chance to show themselves.

16.6 Chapter Summary

Regression diagnostics are used to detect unusual or irregular cases in a data set and to test the assumptions of regression. Before taking any regression analysis at face value, it is important to examine the data for any irregularities, such as impossibly large or small values of regressors

or the dependent variable or strange combinations of regressor values. Often these represent clerical errors or other data collection problems, and they should be fixed. But unusual cases may be hard to detect by merely eyeballing the data in search of something strange. Diagnostic statistics that measure leverage—the atypicality of a pattern of regressor values—can be helpful in this task.

If a case's value of Y is very far from $\hat{Y}$—the case's distance—this may represent a violation of one or more of the assumptions of regression analysis. The residuals, after a transformation, can be used to test whether the assumptions of normality or heteroscedasticity have been violated using one or more of the methods discussed in this chapter. Often, an assumption violation will have no deleterious effects on the quality of the resulting inference and conclusions reached, but you can never be sure, so it is worth looking for assumption violations so you can make an informed decision on what to do about it.

A case can also be highly influential, meaning that its presence in the analysis is having a large effect on the regression results. Measures of influence introduced in this chapter quantify the amount that the inclusion of a case affects the estimates of Y for all cases in data, as well as how a case changes the regression coefficients when it is included relative to when it is excluded from the analysis. These influence measures should be examined and appropriate action taken if a case appears to be distorting a regression analysis, especially if its inclusion seems to work in favor of a hypothesis you are advocating or claim is supported in the data. The decision to include or exclude a case from an analysis should not be taken lightly and needs to be justified. Most important is that you are open with consumers of your research about what you have done.

Assumption violations can affect the validity of the inferences reached with regression analysis or lower the power of hypothesis tests. It is worth examining how robust one's regression analysis is to assumption violations by employing an alternative method, such as bootstrapping or the use of heterscedasticity-consistent standard errors, to see if your conclusions change using one of these alternative methods. This should certainly be done when you have evidence that one or more of the assumptions has been violated, but even if you don't, evidence that an alternative method of inference does not change one's findings can be comforting to both yourself and consumers of your research.