13 Outliers and Ill-Mannered Error

To this point we have generally assumed that error in our basic equation:

DATA = MODEL + ERROR

is well-behaved. In the two previous chapters, we examined the consequences of nonindependent errors and developed an analytic strategy to address the problem. In this chapter, we examine other ways in which error may be ill-mannered. In particular, we consider violations of our original assumption that ε_i is normally distributed with homogeneous variance. These problems are generally not as serious as those associated with nonindependence. However, in many instances the problems caused by non-normality and heterogeneous variance are substantial and, in any case, an examination of the assumptions almost always leads to a better understanding of one's data and the construction of better models.

There are two broad classifications of violations of assumptions about error: systematic and idiosyncratic. As we shall see, the systematic violations are often best identified visually by examining a variety of specialized graphs. Violations sufficient to make us doubt either the estimated model or to question the appropriateness of the probability values for the test statistics usually stand out in these specialized graphs. Our remedy, just as it was for nonindependence, is to transform the data so that we may use our full range of techniques for building models. Idiosyncratic violations arise from wild observations, commonly referred to as *outliers*, that for one reason or another are dramatically inconsistent with the other observations. The presence of outliers frequently causes violations of normality and homogeneous variance assumptions. Hence, we begin by considering methods for detecting and resolving the problems of idiosyncratic outliers and then turn to the visual methods for detecting assumption violations. As a practical matter, it is usually best to address the idiosyncratic outlier issues first, and then consider possible systematic violations of the normality and homogeneous variance assumptions.

OUTLIERS

Outliers are extreme observations that for one reason or another do not belong with the other observations in the data. There are many ways in which outliers can be introduced into the data. The first cause we consider is scientifically uninteresting but quite troublesome and common. Data recording and data entry errors can put wild values into

the data and the predictor variables used in models. The use of computers for data analysis increases the possibility of producing such outliers in our data and at the same time reduces our chances of finding them, unless we look carefully for them. Consider, for example, the effects of entering the heights in inches and the weights in pounds for a number of individuals and then for one individual reversing the numbers for height and weight. If we were doing our calculations with the aid of a calculator, we would likely notice such a reversal. However, when using the computer we are so far removed from our data that we would be unlikely to notice such a mistake.

Throughout we have used the minimization of SSE as the method for estimating model parameters and comparing models. As we noted in Chapter 2, SSE is especially sensitive to outliers. A single outlier can sometimes dramatically alter estimates of group means or regression parameters. Hence, a large data entry error would likely bias the parameter estimates and/or inflate SSE. A cursory examination of the usual output from statistical regression programs would not suggest the presence of such an error. Allowing an outlier to "grab" the model parameters is obviously undesirable. Perhaps more insidious, even if the model is not distorted, is that inflation of SSE substantially increases the difficulty of testing other more complex models by examining changes in SSE. When a few outliers contribute a large proportion of the total error, they overshadow any reductions in error achieved by increasing model complexity. We obviously need statistical techniques for identifying erroneous observations so that they can be removed or corrected in the data.

As an example of the deleterious effects of a data recording error, let us again consider the following two sets of numbers that we examined in Chapter 2:

Set 1:1359 14 $\quad \overline{Y} = 6.4 \quad$ MSE $= s^2 = 26.8$

Set 2:1359 140 $\quad \overline{Y} = 31.6 \quad$ MSE $= s^2 = 3681$

Set 2 is identical to Set 1 except that a data entry error has added an extra digit to 14 to make it 140. As we noted before, this error has a dramatic impact on the mean, which is our estimate of β_0 for the simple model. In this case, the estimate changes from 6.4, which is in the middle of the data values, to 31.6, which is not very representative of any of the data values. The effects on SSE and, consequently, MSE are at least as bad. The inflated MSE makes any hypothesis testing extremely difficult. Examine the 95% confidence interval for β_0 for both sets of numbers:

Set 1: [0, 12.8]

Set 2: [–43.7, 106.9]

The confidence interval for Set 2 is much wider than that for Set 1; in fact, the confidence interval for Set 1 is entirely included within the confidence interval for Set 2. Hence, many hypotheses about β_0 that could be easily rejected for Set 1 will not be rejected for Set 2. The data entry outlier thus not only produces a misleading parameter estimate but also reduces the statistical power of any inferences. The outlier has so inflated the SSE that it is much more difficult to detect any proportional reduction in error produced by using β_0 instead of B_0 in a model for these data. The same would be true for more complex models.

Outliers are, of course, not necessarily erroneous values such as those that result from data entry errors. A second cause of outliers is that the observations are not a single

homogenous set to which a single model will apply, but rather a heterogeneous set of two or more types of observations, one of which is much more frequent. The infrequent cases of the other types will appear as outliers. Discovering that there are really two kinds of things in our data when only one was expected is almost always interesting scientifically. An example is identifying children who thrive and excel despite being raised in adverse environments. Examining other characteristics of extreme observations, especially if they are outliers, may provide clues for building more complex models. In this case, we again need statistical techniques for identifying extreme or outlier observations, not necessarily so that they can be removed or corrected but so that they can be examined with great care to extract their full informational value.

A third cause of outliers is what statisticians refer to as error distributions with "thick tails" in which extreme errors occur with greater frequency than expected for a normal distribution. We examine the normality assumption in greater detail later. For now, it suffices to note that just a few observations from the thick tails of a non-normal error distribution might appear as outliers.

Detecting Outliers

There are three separate outlier questions. Given that our analyses nearly always have predictor variable(s) X_{i1} (X_{i2}, . . ., X_{ip-1}) and always have a data or outcome variable Y_i, two outlier questions are obvious. Is the value for the predictor variable (or the set of predictor values) unusual? Is the value for the outcome variable unusual? The third question pertains to the joint effect of unusual predictor and outcome values by asking about the influence of the observation on joint inferences about all the parameters in the model. That is, does the observation distort or have undue influence on the overall regression model? Each of these three outlier issues is considered in turn after we consider an example, which we use throughout as an illustration.

An outlier example

We now consider a detailed example to demonstrate the serious consequences of even a single outlier and to motivate the detection techniques to be presented later. Figure 13.1 displays the SAT verbal scores and high school ranks for 13 students. The data values on the left side of the figure are the correct values. The values on the right side are identical except that we have simulated an outlier (such as could be caused by a data entry error) by reversing the SAT and high school rank values for the sixth student. Below each dataset are the results of simple regression analyses using high school rank to predict SAT verbal scores. Also given are the values for PRE and F when asking whether high school rank is a useful predictor of SAT verbal scores. Note that PRE and, consequently, F are slightly larger for the dataset that contains the data reversal for the sixth student and the parameter estimates are dramatically different. The intercept changes from about 6 to 97, and the slope changes sign from +0.5 to –0.5. The PRE and F are statistically significant (at $\alpha = .05$) for the dataset with the data entry error, so we would conclude that as high school rank increases, SAT verbal scores decrease! This nonsensical result might cause us to check the data further, but nothing in the basic regression results or in a quick scan of the data in tabular form indicates that a serious problem exists.

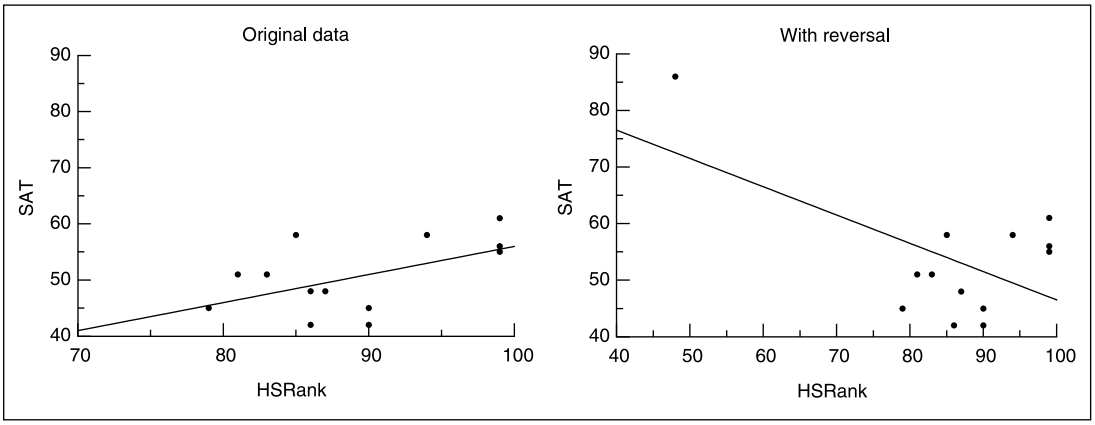
FIGURE 13.1 SAT verbal and high school rank scores with and without data error (reversal of scores for Student 6)

	Original data		*With reversal*	
Student	*SAT*	*HSRank*	*SAT*	*HSRank*
1	42	90	42	90
2	48	87	48	87
3	58	85	58	85
4	45	79	45	79
5	45	90	45	90
6	48	86	86	48
7	51	83	51	83
8	56	99	56	99
9	51	81	51	81
10	58	94	58	94
11	42	86	42	86
12	55	99	55	99
13	61	99	61	99
	$b_0 = 5.95$ $b_1 = 0.50$ PRE = 0.29 $F_{1,11} = 4.53$		$b_0 = 96.55$ $b_1 = -0.50$ PRE = 0.33 $F_{1,11} = 5.35$	

How might we detect the outlier in the second dataset? It is almost always easier to see outliers in graphs than in tables. Hence, the first step is to examine the scatterplot for SAT and HSRank. The scatterplots along with the best-fitting lines for the datasets with and without the reversal of the scores for the sixth observation are displayed in Figure 13.2. The vertical axes in both scatterplots are drawn to the same scale so that the size of an error (i.e., deviation from the least-squares regression line) is comparable.

The scatterplot for the original data in the top half of Figure 13.2 has the pattern we would expect when the model is appropriate. The data values are clustered around a straight line that has a slope of 0.5, as revealed by the regression analysis in Figure 13.1. Note that none of the observations is unusually far from the line or from the other observations. Thus, no outlier is apparent in the top scatterplot. In contrast, an outlier stands out clearly in the bottom scatterplot for the dataset with the reversed observation. Again, none of the observations is particularly far from the bestfitting line (hence the significant value of PRE = .33). However, the outlier observation in the upper left-hand corner of the bottom scatterplot is unusual in three ways. First, the value of this outlying observation on the predictor variable HSRank is unusually small. If we examine the values on the predictor variable for all observations, there is only one large gap in those values–between the predictor value for this unusual observation (48) and the next lowest predictor value (79). Second, the value on the data variable (SAT) for the outlier is unusually large. Again, there is only one large gap between the data value for the unusual observation (86) and the next highest data value (61). Third, if the outlier case were omitted from the analysis, then the best-fitting line would have a positive slope (as in the first scatterplot) rather than a negative slope. This makes the outlier case unusual because it does not appear that omitting any other observation would have nearly as

FIGURE 13.2 Scatterplots and best-fitting lines for data without (left) and with (right) reversal of scores for sixth observation



large an impact on the slope for the regression line. These three ways in which the outlier is unusual suggest that to detect outliers we will want to ask the following three questions for each observation (note: we use the subscript *i* to represent a general observation and subscript *k* to represent an observation being considered as an outlier):

1. Is X_{k1} or, more generally, is the set of predictors $X_{k1}, X_{k2}, \ldots, X_{kp-1}$ unusual?
2. Is Y_k unusual?
3. Would omission of the observation produce a dramatic change in the parameter estimates $b_0, b_1, b_2, \ldots, b_{p-1}$, or equivalently in the predictions $\hat{Y}_i$?

If the answer to any of these questions is yes, then we have an outlier requiring attention. For the HSRank and SAT data it appears as if the answer to all three questions is yes for the artificial outlier we produced by reversing the scores.

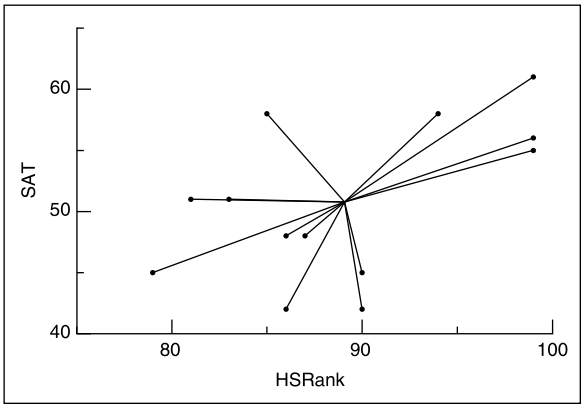
We now consider formal techniques for answering the three outlier questions. The statistical literature abounds with various mathematical indices and graphical procedures for each question. Except in the most unusual circumstances, most of these different procedures give essentially the same answers, so we consider only a small subset of the possible techniques. We have selected those techniques that are most consistent with the general approach to data analysis via model comparisons.

Is the predictor value X_k unusual?

All observations contribute equally to estimating the mean: that is, each observation has a weight of $1/n$. In contrast, each observation does not contribute equally to the estimate of the slope in regression. To gain insight into this issue and its importance, it is useful to reconsider an expression from Chapter 5 for estimating the slope of the best-fitting least-squares line. That is:

$$b_1 = \Sigma w_i \left[\frac{Y_i - \bar{Y}}{X_i - \bar{X}}\right], \text{ where } w_i = \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}$$

FIGURE 13.3 Individual slopes for each data point for original data in Figure 13.1



The term in brackets is the slope "suggested" by each observation or data point—the change in Y between the point and the mean for the variable Y divided by the change in X between the point and the mean for the variable X—that is, the slope of a line between the data point and the mean point. The estimated slope for the regression is simply the weighted average of all the individual slopes suggested by each point. Figure 13.3 illustrates the individual slopes for each data point in the original data of Figure 13.1.

The weight (w_i) given to each data point when calculating the overall slope estimate is based on the unusualness of the data point in terms of the predictor variable X. The further the predictor variable is from the mean, the greater its weight. This makes sense because the slopes of lines with a long horizontal component in Figure 13.3 are unlikely to be affected much by small changes, perhaps caused by error, in the outcome variable Y. Hence, our confidence in the estimated slope is greater for long lines. In contrast, small changes in Y for short lines could dramatically alter the slope. Hence, our confidence in the estimated slope is much less for short lines. One can think of each observation as having a vote on the slope that is to apply to all the observations, but with the votes of some observations counting more. If all the observations are telling more or less the same story, then all the observations ought to be voting for essentially the same slope. In the case of a perfect relationship, all the individual slopes would be identical and would equal the overall slope. It is undesirable for a single observation to have a very large percentage of the total weight, because in that case the slope votes of all the other data points are ignored in calculating the "overall" slope. In that case, the "overall" slope is really a description of only one data point. This occurs when one observation has a very unusual (relative to the other observations) predictor value.

Most modern regression programs report the lever (sometimes unhelpfully referred to as the diagonal of the hat matrix) h_i, which represents the weight or leverage that an observation has in determining the overall model. For simple regression, the lever is defined as:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}$$

For simple regression, the model consists of two parts—the mean (or intercept) and the slope. Each observation has equal weight in the voting for the mean (i.e., $1/n$) and has a weight proportional to the squared deviation of the predictor variable from its mean in the voting for the slope (i.e., w_i). The lever, as defined above, is simply the sum of those two weights.

Unusually high values for the levers are undesirable because they imply that the estimated model in effect only applies to the one or two observations with unusually high levers. It is inappropriate and misleading to report a regression equation as if it applies to all n observations when it actually only applies to the few observations that have particularly high levers.

In judging the magnitude of levers it is useful to note that the sum of the levers necessarily equals the number of parameters, which is two for the case of simple regression. Hence, the average lever equals $2/n$ for simple regression. If a single lever is near 1.0, then it implies that one of the two parameters of the simple regression model is allocated to predict that single observation; this is clearly undesirable. In Figure 13.4 there are columns for the lever values for the original data of Figure 13.1 and for the lever values when there is the data entry error. For a simple regression based on 13 observations, the average lever equals $2/n = .15$. For the original data, which conform to our notions of a typical regression, the levers range around this average from .08 to .25, as expected. However, for the data where we introduced an outlier by reversing the predictor and outcome variables for one observation, the levers range from .08 to .15 except for one extreme value of .76. This implies that of the two parameters in the simple regression, 0.76 of a parameter is effectively allocated to predict a single observation, leaving only 1.24 effective parameters to make the predictions for the other 12 observations. Having such a high proportion of the model focused on a single observation is clearly inconsistent with the goal of having the model describe all the data.

Most researchers know that restricting the range (more properly, restricting the variance) of the predictor variable attenuates the correlation. The converse applies when there is a single outlier with an extreme predictor value: that observation artificially increases the range (variance) and thus inflates the correlation. Allowing one observation to inflate the correlation, or even to create one when otherwise there would be no relationship, obviously increases the chances of making Type I errors—rejecting the null hypothesis when the null hypothesis is in fact correct. Observations with unusual predictor values, assuming they do not also have unusual outcome values, often make Model As appear better than they actually are. In those cases, the story told by the regression model really only pertains to that one observation and it is very misleading to pretend that the regression story applies to all the data.

The generalization of levers from models with one predictor to models with multiple predictors is the same as the generalization of parameter estimates in Chapter 6. If there is no redundancy among the predictor variables, then additional terms for each predictor variable are added to the lever equation. If there is redundancy, then the computations are best left to the computer algorithms. Redundancy also introduces a new wrinkle conceptually for the lever. An observation's predictor values may not be unusual with respect to each predictor variable. However, the pattern across all the predictors might be unusual. For example, if we were using actual height and weight to predict satisfaction with body image among a sample of adolescent girls, a height of 5′ 9″ and a weight

FIGURE 13.4 Outlier indices for the data of Figure 13.1

	Original data			*Data with reversed observation*		
Observation	*Lever*	*RStudent*	*Cook's D*	*Lever*	*RStudent*	*Cook's D*
1	.08	–1.91	0.125	.08	–1.03	0.04
2	.08	–0.31	0.005	.08	–0.53	0.01
3	.11	1.96	0.182	.08	0.38	0.01
4	.26	–0.14	0.004	.10	–1.35	0.10
5	.08	–1.18	0.056	.08	–0.70	0.02
6*	.09	–0.22	0.003	.76	4.63	11.86
7	.14	0.61	0.033	.08	–0.43	0.01
8	.25	0.05	0.001	.15	0.95	0.08
9	.19	0.84	0.086	.09	–0.54	0.02
10	.12	0.89	0.055	.11	0.86	0.05
11	.09	–1.41	0.094	.08	–1.27	0.06
12	.25	–0.15	0.004	.15	0.83	0.06
13	.25	1.09	0.195	.15	1.60	0.21

of 95 pounds would not be particularly unusual, but the combination would be. The generalization of the lever detects this kind of unusualness quite well.

Is the data value Y_k unusual?

If we want to identify those Y_i that are unusual, we must ask "unusual with respect to what?" The obvious answer in the model comparison approach is that we want to find those Y_i that are unusual with respect to the model we are considering. Obviously, the errors:

$$e_i = Y_i - \hat{Y}_i$$

tell us how unusual the ith observation is with respect to the model. If the absolute value of e_i is small, then the model makes a good prediction for that Y_i, so it is not unusual. On the other hand, if e_i is large, then the model makes a bad prediction for that Y_i, so it is unusual. The individual error e_i is often referred to as the *residual* because it is the part of the original observation Y_i that is "left over" after the prediction $\hat{Y}_i$ has been subtracted. Hence, we often refer to an examination of the individual error terms as an "analysis of the residuals."

It is generally difficult to identify outliers by examining the absolute magnitude of e_i for two reasons. First, the importance of an error of a given magnitude is relative to the other errors and the magnitude of the prediction. We therefore need a way to transform the residuals to a common scale on which we can judge small and large values. We might, for instance, standardize them by dividing each one by the root-mean-square error.

A second problem is that there is a paradox in using the standardized residual for identifying unusual values of Y_i. Extreme data values tend "to grab" the model so that the estimated parameters minimize e_i for those extreme values. We are asking whether a particular data value, say Y_k, is unusual with respect to the model, but the extreme Y_k has itself been used to determine the model. If it has seriously altered or biased the parameter estimates in the model, then Y_k might not be unusual in terms of the magnitude

of its residual. Instead, we ought to determine whether Y_k is unusual with respect to a model determined by all the other observations *except* Y_k.

Mathematical statisticians have developed a multitude of transformations of the residuals e_i to solve these two problems in interpreting the magnitude of particular residuals. Most regression programs in computer statistical packages can produce a large variety of these transformed residuals. Some of these transformed residuals only solve the scaling problem, others eliminate the paradox by removing the effect of the *k*th observation when considering the *k*th residual, and some do both. We consider only one of these many residual indices, the *studentized deleted residual*, for three reasons. First, the studentized deleted residual both solves the scaling problem and eliminates the paradox, so there are good theoretical reasons for choosing it as an index of whether Y_k is unusual. Second, the studentized deleted residual has a natural interpretation in the context of the model comparison approach used in this book. We will demonstrate that the square of the studentized deleted residual is simply the *F* for comparing appropriately chosen Models C and A. Third, it is very unlikely that the studentized deleted residual would fail to detect an unusual Y_k that could be detected by any of the other transformed residuals.

We develop the studentized deleted residual by considering a specific model for an outlier. If an observation is so extreme that it is unlike the other observations, then we should be able to reduce the error appreciably by adding a specific parameter to the model just for that one observation. Designating a parameter for a single observation will ensure that there will be no error in our prediction for that observation. The other parameters are thereby freed to describe the remainder of the data. Because all error due to the outlier is eliminated and the other parameters generally provide a better fit to the remaining data, our overall measure of error will inevitably be less. In effect, we will be considering whether the outlier is fundamentally different from the other data observations. We will refer to an original multiple regression model augmented by the addition of a specific parameter for the suspected outlier as an *outlier model*. To screen data for outliers, we will test the outlier model for each observation. That is, we will test one observation at a time to see whether it is so extreme that we should allocate a specific parameter to it.

As with all other models, the question inherent in the outlier model is whether the additional complexity is worth it. Adding a parameter for the suspected outlier to the model will inevitably reduce the error, but will the reduction in error justify the increased complexity of the model? We can use PRE and *F* to answer this question. To be more formal, the model comparison for the outlier model is:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} + \beta_p X_{ip} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

$$H_0 : \beta_p = 0$$

$$\text{where } X_{ip} = \left\{ \begin{array}{ll} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{array} \right\}$$

where Model C is the original regression model and Model A is the outlier model having an additional parameter that estimates the effect of a single observation. If the original model really describes all the observations, then it should not be useful to isolate a single observation with its own parameter.

As an example, consider the sixth observation in the data of Figure 13.1 with the data reversal error. Let $X_i = 1$ if $i = 6$ and 0 otherwise. Then the estimation of the above models reveals:

MODEL A: $\hat{SAT}_i = 6.71 + 0.50HSRank_i + 55.49X_i$

MODEL C: $\hat{SAT}_i = 96.55 - 0.50HSRank_i$

PRE = .68 $F_{1,10} = 21.4$ $p < .01$

Thus, devoting one parameter to the sixth observation or, equivalently, omitting the sixth observation from the data used to fit the model reduces the error by 68%. With 11 degrees of freedom for error left after estimating b_0 and b_1, we would expect that the omission of any one observation would, on average, reduce the error by only about 1/11 = 9%. The very much larger proportional reduction in error obtained by omitting the sixth observation suggests that it is a very unusual observation with respect to the model for the data. This is confirmed by the very large value of $F_{1,10} = 21.4$. Note also that the sign of the coefficient for HSRank changed between Models C and A. Indeed, omitting the sixth observation caused a dramatic difference in the estimation of both the intercept and the coefficient for HSRank. We return to this fact later.

It is interesting to consider what would happen if we were to test the outlier model for an observation that is not an outlier. Suppose, for example, that we tested the outlier model for the first observation in the data with the reversal error in Figure 13.1. Estimating the coefficients in the compact and augmented models yields (where now $X_i = 1$ if $i = 1$ and 0 otherwise):

MODEL A: $\hat{SAT}_i = 95.71 - 0.48HSRank_i - 10.67X_i$

MODEL C: $\hat{SAT}_i = 96.55 - 0.50HSRank_i$

PRE = .096 $F_{1,10} = 1.06$ n.s.

Devoting a special parameter to the first observation is clearly not worth the added complexity. The proportional reduction in error of 9.6% is little more than we would expect by chance (9%), which is confirmed by the low, nonsignificant value of F. Note that in contrast to omitting the sixth observation, omitting the first observation produces minimal changes in b_0 and b_1.

The above results suggest that we can use the outlier model as a means of screening our data for possible outliers. We simply test the outlier model n times, each time letting a different observation be the "odd one out." Although the outlier model is conceptually clear, it would be tedious to specify and test n different models for each regression. Fortunately, this is not necessary because Belsley, Kuh, and Welsch (1980) have shown that F for testing the outlier model for the kth observation is given by:

$$F = \frac{e_k^2(n - \text{PA} - 1)}{\text{SSE}(1 - h_k) - e_k^2}$$

All of the components for computing F for the outlier model are therefore available either from the original regression or from the levers h_k. Most computer programs for multiple regression will provide $t = \sqrt{F}$ as the studentized deleted residual for each observation. The name reflects a different derivation of the same statistic. In the model

comparison approach, it makes sense to think of the studentized deleted residual as a test of the outlier model. The studentized deleted residuals for each observation in our example are listed in Figure 13.4 under its common abbreviation "RStudent." For the original data without an outlier, none of the values would be statistically significant using $\alpha = .05$. However, in the dataset with the reversed observation, the sixth observation has a very large value of RStudent = 4.63 (which squared equals 21.1, identical to the F for testing the outlier model for this observation). If the model is describing all the data, then there should not be such a large value for RStudent. The sixth observation is definitely an outlier.

We have now answered the question of whether an observation's Y value is unusual with respect to a model of the other observations. The studentized deleted residual is a t statistic testing the outlier model. If the resulting t is large, then we conclude that allocating an additional parameter for that particular observation is worthwhile. This conclusion is equivalent to saying that that particular observation's Y value is quite unusual with respect to the other observations. In other words, one model does not apply to all the observations.

We must be careful about using the outlier model to screen the data for outliers. If we examined the studentized deleted residuals for all observations, we would be performing the equivalent of n statistical tests on the same set of observations. If we use $\alpha = .05$ to determine the critical values of PRE and F, or equivalently RStudent, for rejecting the original model in favor of the outlier model, then we have a 5% chance of making a Type I error for *each* of the n model tests we perform. The probability that we do *not* make a Type I error in n trials equals $.95^n$. For our example with 13 observations, the probability that we do not make a Type I error is $.95^{13} = .51$. So the probability of making at least one Type I error when screening the data with the outlier model equals $1 - .51 = .49$. This is unacceptably large. One solution is to use the *Bonferroni inequality*. According to the Bonferroni inequality, the probability of making at least one Type I error will be less than α if the critical value for each test is chosen so that $\alpha' = \alpha/n$. For our example with 13 observations, $\alpha' = .05/13 = .004$, which would imply a critical value for F (t) of about 13 or 14 (3.6 or 3.7). The F (or RStudent) for the sixth observation easily exceeds this more conservative standard and so would be declared a statistically significant outlier.

In general, we do not recommend that the squared studentized deleted residual or the F that results from evaluating the outlier model be used routinely as a formal statistical test unless one has external information questioning the validity or reliability of a particular observation. Instead we suggest that large values of RStudent be used to indicate observations requiring closer scrutiny. An easy rule of thumb applies to RStudent. For reasonably large n, only about 1% of the studentized deleted residuals should be less than –3 or greater than +3. So, if the absolute value of RStudent is > 3, careful attention to that observation is required. Absolute values of studentized deleted residuals of > 4 ought to be extremely rare and even after the Bonferroni adjustment will be statistically significant. Hence, all alarm bells ought to sound. An RStudent that large is a clear indication of an unusual value of Y_k. Using a model for which a studentized deleted residual is > 4 could be very misleading.

Would omission of Y_k dramatically change the predictions?

While considering the lower graph in Figure 13.2 and examining the results of the outlier model applied to the sixth observation, we noted that omitting that observation causes a dramatic change in the estimate of b_1 from –0.5 when the sixth observation is included to +0.5 when it is omitted. It did not appear that omitting any other observation would have an appreciable effect on the estimate b_1 and certainly would not change the sign of the estimate. This suggests that we search for outliers by considering the change in the parameter estimates resulting from the deletion of each observation from the analysis. The deletion of a "typical" observation should have little or no effect on the parameter estimates, but the deletion of an unusual observation that may have "grabbed" the estimates might result in a large change in the parameter estimates. As we noted when defining SSR conceptually, differences in parameter estimates are reflected directly in different predicted values. Hence, to assess whether the set of parameters changes dramatically as a function of one observation, we can compare the predictions for the model based on the entire set of observations to the predictions for the model estimated when one observation is omitted. Let $\hat{Y}_{i,[k]}$ represent the predicted value for the *i*th observation from a model estimated with the *k*th observation omitted. Cook's *D* (1977, 1979) compares the two sets of predictions as:

$$D_k = \frac{\sum_i (\hat{Y}_i - \hat{Y}_{i,[k]})^2}{\text{PA(MSE)}}$$

A large value of Cook's *D* indicates that the model predictions are very different depending on whether the *k*th observation is included or not. For any particular observation that was of concern, it would be relatively simple to compute the two regressions with and without the *k*th observation and then compute the sum of squared differences for the two sets of predictions. However, this again would be tedious as a screening procedure for outliers if we had to perform *n* separate regressions, each time omitting a different observation. Fortunately, as was the case for the outlier model and studentized deleted residuals, this is not necessary because the following equivalent formula for Cook's *D* shows that it can be calculated from basic components available from the original regression. That is:

$$D_k = \frac{e_k^2}{\text{PA(MSE)}}\left[\frac{h_k}{(1 - h_k)^2}\right]$$

For example, for the sixth observation the only quantity we have not determined is MSE, but this is easily obtained as SSE/(n – PA) = 1085.3/11 = 98.66, so:

$$D_6 \frac{13.3^2}{2(98.66)}\left[\frac{.76}{(1 - .76)^2}\right] = 11.83$$

This value (within rounding error) as well as the values of Cook's *D* for all the observations are listed in Figure 13.4. The only very large value of Cook's *D* is the one for the sixth observation in the dataset in which its predictor and outcome values are reversed. As we noted before, that observation grabs the model and dramatically changes the predictions for all the other observations, which is clearly undesirable in a model that purports to be a description of all the data.

There are only informal guidelines for thresholds above which values of Cook's *D* should require attention. One suggested rule is to consider any value greater than 1 or 2 as indicating that an observation requires a careful look. Another suggestion is to look for gaps between the largest values of Cook's *D*. Usually, as in Figure 13.4, a value truly requiring attention is obvious.

Many regression programs will report Cook's *D*, so the above computational formula will seldom be needed. However, the formula is conceptually important because it demonstrates that Cook's *D* is a multiplicative function of the squared error for that observation and the lever for that observation. Thus, the most influential observations in terms of their effects on the parameter estimates will be those having large squared errors (an unusual Y_k value) *and* high leverage (an unusual set of predictor values). Conversely, if either Y_k is not unusual with respect to the model for the other observations (i.e., low value of RStudent) *or* if the set of predictors for the *k*th observation is not unusual with respect to the sets of predictors for the other observations (i.e., a low value of the lever h_k), then Cook's *D* will not be large. But also note that if Y_k is not quite unusual enough to attract our attention and if h_k is also not quite large enough to be considered a high lever, then Cook's *D* can and often will be large enough to require attention.

Dealing with Outliers

Once we have identified outliers in a batch of data (by examining the levers, studentized deleted residuals, and Cook's *D* statistics), we have the problem of what remedial action, if any, should be taken. Although the tools to detect outliers are not particularly controversial, decisions about how to proceed in the face of clearly outlying observations have been a more contentious subject, with divergent and sometimes strong opinions devoted to the subject. If an outlying observation clearly originated because of data recording or coding errors, as in the example we have been using, then everyone would agree that it makes sense to correct the observation if possible, and redo the analysis with the corrected observation included. In many cases, however, it may not be at all clear why a particular observation is an outlier and, as a result, no clear corrective strategy is available for dealing with the problematic observation or observations. In this case, controversies may arise about how one should proceed.

In the early days, before the development of sound statistical procedures for identifying outliers, it was of course somewhat natural to question analyses that left out problematic observations based on seemingly ad hoc rules. One naturally feared, in the absence of clear rules for identifying outliers, that observations might simply have been omitted because they were problematic for the researcher in arguing for a favored hypothesis. Because of such fears, a tradition emerged in psychological research in which it was generally considered unethical to report analyses based on partial datasets that omitted problematic observations.

We believe that the development of sound statistical procedures for identifying outliers, using tools such as the lever, the studentized deleted residual, and Cook's *D*, has rendered this point of view obsolete. In fact, we would argue that it is unethical to include clearly outlying observations that "grab" a reported analysis, so that the resulting conclusions misrepresent the majority of the observations in a dataset. The task of data analysis is to build a story of what the data have to tell. If that story really derives from

only a few overly influential observations, largely ignoring most of the other observations, then that story is a misrepresentation.

At the very least, in the presence of clear outliers, we encourage researchers to analyze their data twice, once including the outliers and once not. If the resulting story is generally the same regardless of the inclusion of the outliers, then it is obviously simpler to include the unusual cases. On the other hand, if the story is dramatically different and the outlier statistics are clearly indicative of one or more problematic cases, then we would recommend reporting the story without the outliers, while clearly indicating the steps that were undertaken for outlier detection. And an interesting question, worthy of further pursuit, is why the unusual cases occurred in the first place and what they may be able to tell us further about the theoretical issues that are being explored. While the majority of observations may have a single story to tell, those few that make for a very different story can often be intriguing cases to follow further, to identify the limits of the story that one has told and how it needs to be expanded for other populations.

SYSTEMATIC VIOLATIONS OF ERROR ASSUMPTIONS

Outliers are observations that are idiosyncratic in a dataset, arising for unknown reasons, and leading to potentially very different stories of what the data have to say. In addition to problems that thus arise idiosyncratically, there may also be systematic tendencies of all the errors in a dataset to violate the assumptions underlying statistical inference. In Chapters 11 and 12 we dealt with the most consequential of these systematic violations, data in which errors or residuals are not independent of each other. The other two assumptions that also may be seriously violated are the assumptions that errors are normally distributed and come from a single distribution, thus having a constant variance.

It is essential to realize that the assumptions of normality and constant variance pertain only to the errors in a given model and not to any of the variables themselves, neither the data variable, Y, nor the predictors, X, in the model. We clearly have extensively used predictors whose values are not normally distributed and there is nothing in all we have done that makes any distributional assumptions about the predictors in a model. As far as the data variable, Y, is concerned, here too there are no assumptions about the distribution of this variable, as long as its errors in the models predicting it are normally distributed with constant variance. To illustrate the distinction, consider the case of a categorical predictor variable, with two levels, used to predict some Y, and there exists a very substantial mean difference between the two groups. In essence, then, our Y variable would have one common distribution that came from two within-group distributions. And if the mean difference between the two groups was substantial enough, the overall distribution of the Y variable, ignoring the two groups, would be far from normal (i.e., bimodal). Yet, within each group the distribution of Y might very well be nicely behaved (i.e., normal in each group, with the same variance). These within-group distributions are the distributions of the errors in the model that predicts the two group means. Sometimes, of course, the Y variable might be distributed in such a way that there is no possible way that its errors are normally distributed. We consider such models, that is, models in which the Y variable is dichotomous, in Chapter 14. The important point is that it is the distribution of the errors that is crucial, not the distribution of Y.

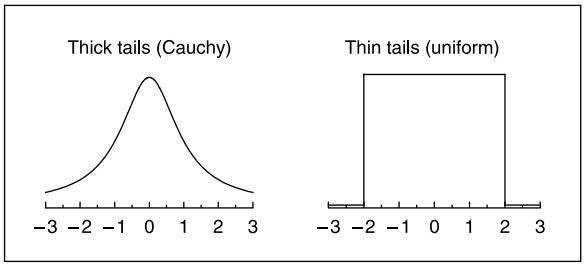
The assumptions that errors are normally distributed with a constant variance are often seen as "robust" assumptions, in the sense that the accepted advice is that violations of the assumptions do not result in major statistical errors. While these assumptions are clearly more robust than the independence assumption considered in the last two chapters, it nevertheless is true that violations of normality and constant variance in many cases can result in inferential errors (both Type I and Type II errors). Therefore it makes sense for the researcher to attend to these issues, particularly in cases where there is good reason to suspect violations—cases that we detail below.

As outlined above, the detection of outliers involves motivated statistics and model comparisons that give clear indexes of the degree to which particular cases might pose analytic problems. The detection of non-normality and heterogeneous variances of errors is a much less precise science and we will largely emphasize data plots and inspections of them to detect violation problems. Although there are statistical tests of departures from normality and of heterogeneous variance, we do not emphasize these because they often involve other assumptions that may be problematic. It is our experience that plots of data will often make clear when these assumptions are violated substantially enough that corrective action is warranted. In this sense, we regard these assumptions as robust: if their violations are likely to cause problems, they should be apparent based on relatively nonexact methods of data inspection.

Diagnosing Non-Normal Error Distributions

As already noted, the assumption that errors are normally distributed is a relatively robust assumption. But the validity of this conclusion depends on the nature of the departure of the distribution of errors from the normal one. Departures of distributions from the normal distribution occur primarily in two ways. On the one hand, the normal distribution is symmetric, with tails of equal thickness both above its central tendency and below it. Other distributions may depart from it because they are skewed in one direction or the other, either with a thicker positive than negative tail or the other way around. On the other hand, distributions may depart from a normal one because they are too peaked, with tails that are too thin, or insufficiently peaked with tails that are too thick. Although it is important to correct error distributions that are skewed compared to the normal one, the most serious problems are encountered when distributions of errors have substantially thicker tails than the normal distribution. Figure 13.5 portrays two distributions. On the left is what is called a Cauchy distribution, resulting from the ratio of two normally

FIGURE 13.5 Cauchy distribution with thick tails (left) and uniform distribution with thin tails (right)



distributed variables. On the right is a uniform distribution. Although the Cauchy distribution looks roughly normal, its tails are actually considerably thicker than those of a normal distribution. On the other hand, the uniform distribution has tails that are considerably thinner than the normal distribution. Because least-squares estimation particularly attends to extreme errors (by minimizing the sum of *squared* errors), thick-tailed distributions, like the Cauchy distribution, are potentially much more problematic than other non-normal distributions such as the uniform one.

In our judgment, the most useful tool for detecting whether the distribution of errors departs from a normal one is the normal quantile-quantile plot, provided by all major software programs in procedures that plot univariate distributions of data.[1] The normal quantile-quantile plot displays the percentile distribution of the data (i.e., the errors) against the percentile distribution that would be expected if the distribution were normally distributed. Thus, the data values are rank-ordered and the percentile of each score in the distribution, based on the number of observations, is computed. These rank-ordered percentile scores for each observation constitute the values on the vertical axis of the normal quantile-quantile plot. Along the other axis are the z scores for each percentile coming from the normal distribution. Thus, each observation is located by a point in the two-dimensional plot, with its value on the vertical axis being its percentile in the actual distribution of errors and its value on the horizontal axis being what its z score would be if the percentiles came from a normal distribution.

To illustrate, the values in Figure 13.6 were generated by sampling from a normal distribution with a mean of 0 and a standard deviation of 10. We may think of them as values for the error for some model. In this sample, the mean of the 25 scores is actually 0.2 and the sample standard deviation is 11.5. The rank-ordered data values or raw scores are given in the far left column of Figure 13.6. In the next column, the value i is given for each score, which is simply a variable that indicates the rank order of the score (hence it goes from 1 to 25). In the third column, the percentile in the sample for each score is given, using the formula $F = (i - .5)/n$. The .5 is included in this formula just to avoid percentiles of 0 and 1.0. These are the scores that are plotted along the vertical axis of the normal quantile-quantile plot for each observation. Then in the final column of Figure 13.6, the z score associated with the given percentile, coming from the normal distribution, is given. Thus, for instance, the third lowest error of –14 in the distribution is the 10th percentile score in the distribution of the data (i.e., $F = (3 - .5)/25 = .10$). And in a normal distribution of data, the z score corresponding to the 10th percentile score is –1.28. The values in this final column are plotted along the horizontal axis of the normal quantile-quantile plot and thus each observation is located by a point in the plot corresponding to its percentile in the sample (vertical axis) and what its z score would be if the distribution were normal (horizontal axis). The resulting normal quantile-quantile plot for these data is given in Figure 13.7.

If the data were perfectly normally distributed they would lie along a straight line in the plot. Since these data were in fact sampled from a normal distribution, the fit of the points to a straight line is exceedingly good. In Figure 13.8 we give four normal quantile-quantile plots for errors sampled from four known distributions. The first row shows another example for the normal distribution, with the original distribution on the left and the normal quantile-quantile plot on the right. The second row depicts a sample from a uniform distribution, with each error value between –2 and +2 being equally likely. In its normal quantile-quantile plot, the plotted points seem to fall along a straight

line in the middle of the distribution but then the slope becomes very flat at the ends. The flat slope at the ends indicates that the most extreme scores in the sample (having the highest and lowest percentiles) are not as extreme as they would be if the data came from a normal distribution. This distribution has thinner tails than data from a normal distribution. The third row contains the Cauchy distribution. Note that in the plot on the left side its tails are thicker than those for the normal distribution in the first row. Here, once again, the slope of the points in the middle of the distribution is as it should be, but now at the ends the slope gets very steep, indicating that the data values at the high and low end of the distribution are more extreme than they would be if they came from a normal distribution of data. Normal quantile-quantile plots that look like this are a cause of serious concern. Finally, the last row in Figure 13.8 depicts errors sampled from a transformed beta distribution, which in this case has a positive skew. As the normal quantile-quantile plot makes clear, this distribution has a thicker positive tail than a normal distribution (the slope of the plot becomes very steep at the upper end) and a thinner negative tail than a normal distribution (the slope becomes very flat at the lower end). Again, it is the thick tail on the positive end that would be cause for concern.

FIGURE 13.6 Empirical errors randomly sampled from a normal distribution with their normal quantiles

Empirical error values	*i*	f_i	*Normal quantile* $z(f_i)$
–21	1	.02	–2.05
–16	2	.06	–1.55
–14	3	.10	–1.28
–12	4	.14	–1.08
–12	5	.18	–0.92
–11	6	.22	–0.77
–10	7	.26	–0.64
–8	8	.30	–0.52
–8	9	.34	–0.41
–6	10	.38	–0.31
1	11	.42	–0.20
1	12	.46	–0.10
2	13	.50	0
4	14	.54	0.10
4	15	.58	0.20
5	16	.62	0.31
5	17	.66	0.41
5	18	.70	0.52
7	19	.74	0.64
7	20	.78	0.77
11	21	.82	0.92
13	22	.86	1.08
15	23	.90	1.28
20	24	.94	1.55
22	25	.98	2.05

FIGURE 13.7 Normal quantile-quantile plots for errors in Figure 13.6

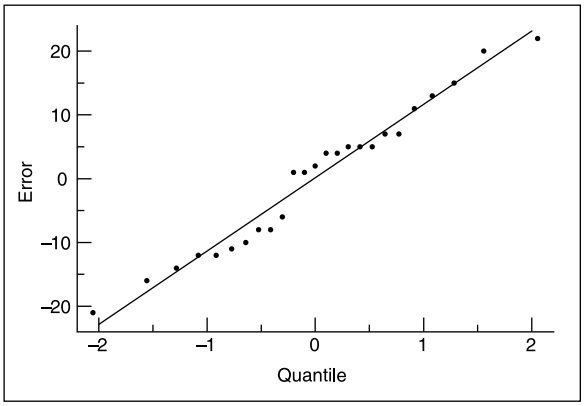
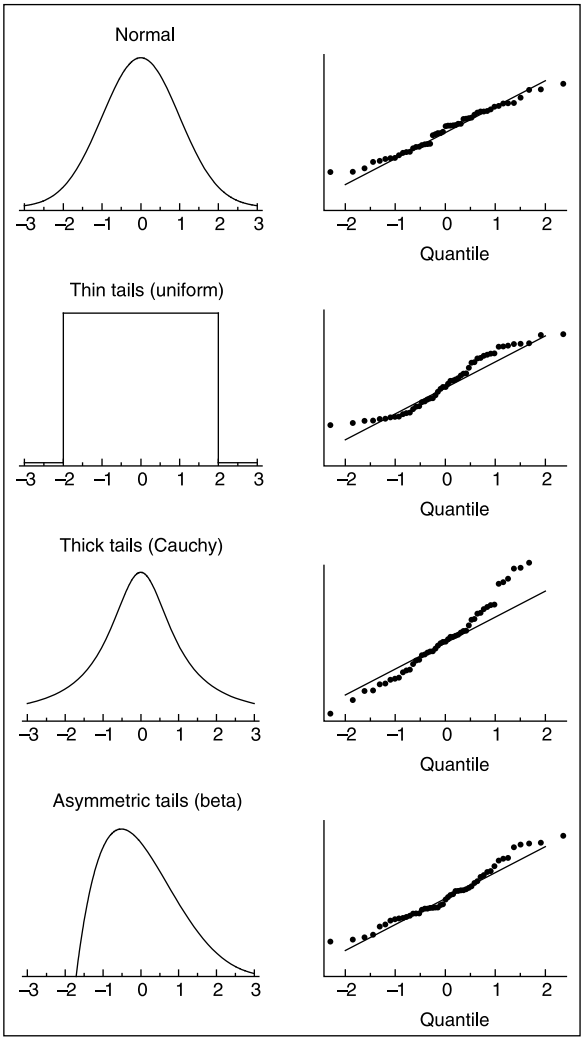


FIGURE 13.8 Normal quantile-quantile plots for errors sampled from normal (first row), uniform (second row), Cauchy (third row), and transformed beta with skew (fourth row) distributions



Diagnosing Errors with Heterogeneous Variances

As was the case when examining whether errors are normally distributed, we have found that graphical inspection methods are generally most useful for examining whether they manifest a constant variance. Most frequently, violations of this assumption occur when the variance of the errors depends systematically on the magnitude of the predicted values, $\hat{Y}_i$. For instance, if we were predicting the sales price of makes of automobiles, then being off by $1000 may be a much more substantial error when the predicted sales price is $20,000 than when the predicted sales price is $100,000.

Accordingly, a useful plot for determining whether errors show heterogeneous variances is to graph the errors as a function of the predicted values, $\hat{Y}_i$. The most common violation of the constant variance assumption in such plots, consistent with the example given in the previous paragraph, is a funnel-shaped plot in which the errors become larger in absolute value (both positive and negative errors) as the predicted values become larger. That is, the errors are more spread out (around zero) at higher predicted values. This is perhaps even easier to see if one plots not the errors themselves as a function of the predicted values, but their absolute value (or more frequently, the square root of their absolute value—such a plot is called a "spread-location plot"). In this plot, if errors are more variable or spread out at higher predicted values, then the absolute values of the errors should become larger at higher predicted values, and the resulting plot will have a generally positive slope.

FIGURE 13.9 Bivariate scatterplots (first row), plots of residuals x predicted values (second row), and spread-location plots (third row) for homogeneous (left) and heterogeneous (right) errors

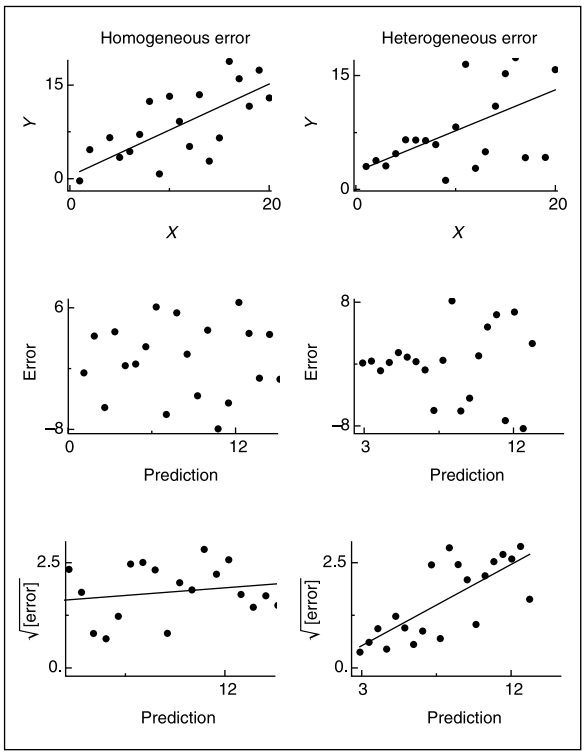


Figure 13.9 illustrates these kinds of plots with data where the errors have a constant variance (on the left) and with data where the errors were generated to be proportional to the size of the predicted values (on the right). The top row of scatterplots is simply the bivariate scatterplot, using a single predictor variable (X) to predict Y. From these plots, it is very hard indeed to see any problems. But things become clearer in the residual by predicted plots in the second row of the figure. On the left, the plotted points show a nice cloud-like structure with no noticeable funnel shape. However, such a funnel shape is much more apparent in the right-side plot where it seems that the errors are more spread out at higher values of the predicted value. The third row presents the spread-location plots. Here, it is very evident that as the predicted value increases, the average absolute error (actually its square root) also increases for the heterogeneous errors in the right column. In contrast, the absolute error for homogeneous errors shows very little increase as the predicted values increase.

While this sort of funnel-shaped pattern is typical, other patterns of Residual × Predicted plots (or spread-location plots) may also be found and would also indicate violations of the constant variance assumption. For instance, when the Y values are proportions, there tends to be less variance in the errors at both very high predicted values (proportions near 1) and very low predicted values (proportions near 0). In such a case, the Error × Predicted value plot may show a diamond shape, with less spread at both ends, and the spread-location plot may show a nonlinear pattern, with higher values of the absolute errors in the mid-range of the predicted values.

Resolving Violations of Normality and Constant Variance

As we suggested at the start of this chapter, it is important to deal with individual problematic observations (i.e., outliers) before addressing the more systematic violations of error assumptions (i.e., normality and constant variance). Once these more systematic problems have been diagnosed, there are various remedial actions that can be taken. In general, we will take an approach that is similar conceptually to our approach for dealing with violations of the independence assumption. There, we computed various transformations of our data variable, referring to them as W. Once we had done these transformations, combining nonindependent observations in various ways, the dependence problem was eliminated and we could proceed with our full repertoire of analytic models developed in the earlier chapters. Here, we take a similar approach. That is, we will recommend various transformations of the data variable, of the general form $W_i = f(Y_i)$, where $f(\ )$ represents some function, to be defined. For different sorts of systematic violations, different functional forms will be required in order to correct the problems in the distribution of the errors. Once an appropriate transformation has been found, we can proceed with the full repertoire of analytic models developed earlier in this book, just as we did in the case of data that violate the independence assumption.

There is an alternative to the approach that we are recommending, which is widely used. It involves using what are called *nonparametric* statistical tests. We do not recommend nonparametric procedures for two primary reasons. First, it is commonly assumed that such tests make no assumptions about the error distributions. This, unfortunately, is not the case. While it is true that they effectively deal with non-normally distributed errors, they routinely make very strong assumptions about

homogeneity of variance. Thus, they do not represent the sort of generic all-purpose solution to assumption violations that they are often judged to be. Second, many nonparametric procedures are equivalent to the parametric ones we have relied upon, done not on the raw data variable but on a particular transformation of the data variable, the rank transformation. Thus, parametric procedures typically analyze a transformed score, $W_i = f(Y_i)$, where the $f(\ )$ is simply the rank transform, ordering the data values and giving the lowest score a value of 1, the next lowest a value of 2, and so forth. In our judgment, other transformations are better than the rank transformation, both because they are likely to preserve more information contained in the raw data and, more importantly, because they actually can bring the data much more in line with normal distributions with constant variance. The rank transformation results in data that are uniformly, rather than normally, distributed. If one recognizes that most nonparametric procedures are simply parametric procedures applied to data that have been rank-transformed, then it becomes obvious that if there are better transformations, better in the sense that they are more likely to deal with problems of non-normality and heterogeneous variances, then these should be used instead of nonparametric statistical procedures.

Nonlinear but monotonic transformations of the data variable can often lead to more normally distributed errors with constant variances. Additionally, data that have been so transformed to alleviate assumption violations may often yield the benefit of simpler models. That is, many times higher order interactions or nonlinear effects in data are eliminated once appropriate transformations have been applied to data that otherwise violate the assumptions we have been considering. And, as always, we prefer models that are more parsimonious.

For certain kinds of data, having certain characteristic distributions, there are widely accepted transformations that generally have the desired effect of correcting assumption violations. For instance, in the case of data variables that are counts (e.g., the number of times someone does something or the number of children in a family), the appropriate transformation is generally a square root transformation, $W_i = \sqrt{Y_i}$. For data that are reaction times, the appropriate transformation is generally the log transformation, $W_i = \log(Y_i)$. For data that are proportions, the logit transformation is often the most appropriate, $W_i = \log [Y_i/(1 - Y_i)]$. And for data that are correlations, Fisher's Z transformation, $W_i = 1/2 \log [(1 + Y_i)/(1 - Y_i)]$, is generally most appropriate. Of course, while these are the routinely applied transformations in these cases, we recommend checking whether the chosen transformation has indeed resulted in errors that have corrected the original distributional problems, again by examining normal quantile-quantile plots and Residual $\times$ Predicted plots, but this time on the residuals from the transformed data variable.

To correct problems of skew in other sorts of distributions, we recommend exploring transformations that derive from what has come to be known as the ladder of powers, $W = Y_i^{\lambda}$, assuming that a constant has been added to Y_i such that all values of Y_i are > 1. If $\lambda = 1$, then of course $W_i = Y_i$ and no transformation is effected. If $\lambda > 1$, then larger values of Y_i become even larger, so that the upper end of the distribution becomes more spread out, relative to the lower end, thus reducing negative skew. And the higher the value of λ, assuming $\lambda > 1$, the more this is true. For values of $\lambda < 1$, the opposite happens: the lower end of the distribution becomes more spread out, relative to the upper end, thus reducing positive skew. Y^0 yields a value of 1 for all cases and so is not used; log (Y_i) is

the transformation that fits in the family of power transformations at that point. Thus, to reduce positive skew, one might start with the square root transformation, $W_i = \sqrt{Y_i} = Y_i^{.5}$. If positive skew remains, then one might try the log transformation. Even more extreme would be the inverse of the square root $W_i = Y_i^{-.5}$, and then the inverse $W_i = Y_i^{-1}$, and so forth. Finding the correct value of λ to reduce skew and render the distribution more like a normal distribution may be a trial and error procedure. One works either up or down the ladder of powers, depending on whether the data variable is positively or negatively skewed, examining the normal quantile-quantile plot for each transformation until the one that yields the most approximately normal distribution is identified.

Importantly, in addition to violating the assumption of normality of errors, skewed error distributions are also routinely characterized by heterogeneous variances. In the case of positive skew, for instance, when dealing with reaction times, it is generally the case that errors are more variable at longer times. Power transformations that reduce skew are thus likely to correct both the non-normality and heterogeneous variance problems.

EXAMPLE

A detailed example will help to elucidate the diagnostic and remedial tools described in this chapter. These data pertain to the relationship between the number of grants in the physical and social sciences awarded by the National Science Foundation (NSF) to various universities. These data were assembled by a colleague whose university's administration was concerned about the low number of NSF grants they received. They had decided to remedy the situation by providing extra resources to departments in the physical sciences. This colleague hoped to convince his administration that it was equally important to support social science departments because there was a strong relationship between the number of grants received in the physical and social sciences at peer institutions. The data analyzed here are the number of NSF grants of each type for the given state's two primary public universities and their self-identified peer institutions. Each university is categorized as the state's flagship or land grant university. Figure 13.10 shows the data and best-fitting regression lines for the flagship and land grant universities. Despite the apparent divergence of the slopes, the difference in slopes (i.e., a test of the interaction between university category and number of social science grants) is not significant: $F(1,15) = 1.9$, PRE = .11, $p = .19$. However, the plot reveals a potential outlier—a flagship university that has by far the greatest number of physical science grants—that may be distorting the regression line for the flagship universities. A visual check of the regression assumptions as well as an outlier analysis is needed.

Figure 13.11 displays the normal quantile plot (on the left) and the spread-location plot (on the right). One point, which corresponds to the apparent outlier in Figure 13.10, is far away from the normality line in the normal quantile plot; its steep slope relative to the other points identifies this as a "thick-tail" normality violation that could have substantial impact on the analysis. The spread-location plot clearly shows that the square root of the absolute value of the residuals is increasing with the size of the predicted values from the regression, thus the assumption of homogeneity of variance is violated. The studentized deleted residual for the unusual observation is 6.78, its lever is .32, and its Cook's D is 1.33, which is about twice as large as the next value of Cook's D.

FIGURE 13.10 Relationship between NSF physical and social science grants at flagship (○) and land grant (●) peer institutions

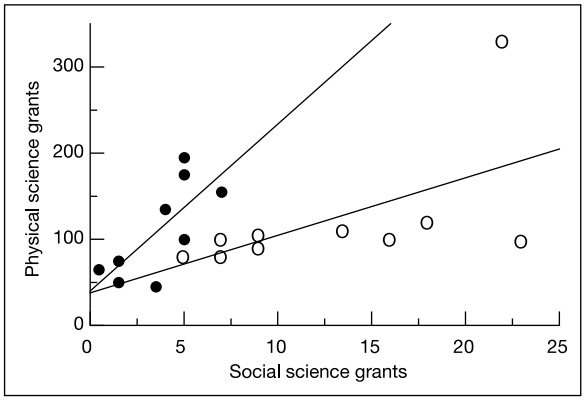
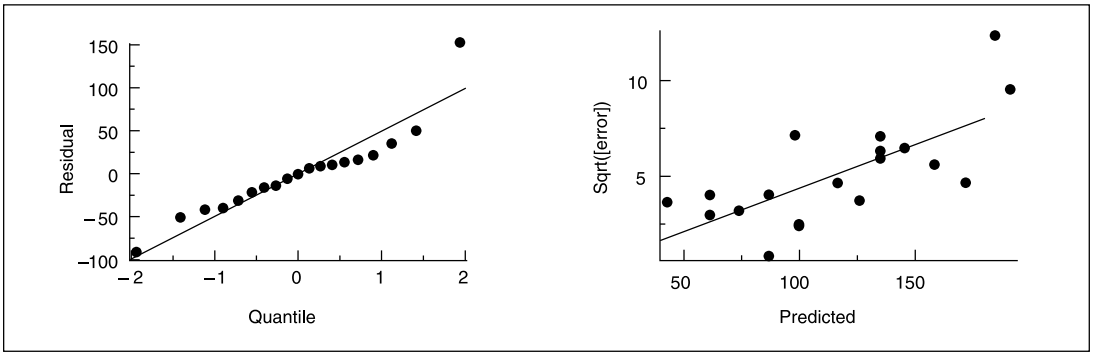


FIGURE 13.11 Normal quantile plot (left) and spread-location plot (right) for grants data of Figure 13.10



Its lever does not identify it as a particularly unusual observation in terms of its predictor values. However, the large studentized deleted residual ($p < .0001$) suggests that this university is telling a very different story about the relationship between the number of physical and social science NSF grants; the large value of Cook's D indicates that it is having a disproportionate effect on the overall model. If the outlier is omitted, the interaction is significant: $F(1,14) = 15.38$, PRE = .53, $p = .0015$. In other words, the story we tell about whether the relationship is different for flagship and land grant universities depends entirely on whether we include this one university. We should not allow one observation to dominate the story we tell about *all* the data! In this particular case, the unusual university has by far the highest total number of grants. And its number of physical science and social science grants fits neither the pattern of the land grant universities nor the pattern of the other flagship universities. In retrospect, this might not be a legitimate peer institution; it may have been included, wishfully, as a peer only because it was in the same intercollegiate athletic conference. For all these reasons, it is appropriate to conduct an analysis with that outlier removed. If the analysis changes appreciably without that university, that of course would not prove any of our post hoc suppositions above. Instead, those suppositions might provide hypotheses to be explored in a larger study including all the state universities.

Removing the one clear outlier does not repair the violation of homogeneity of variance apparent in the right panel of Figure 13.11 (the new graph is not presented here). Hence, a transformation may be appropriate. Also, there are a priori reasons for anticipating the need for a transformation of these data. The scale for the number of grants is not likely to be linear with an underlying scale of institution quality. For example, is the functional difference between, say, 5 and 10 grants equivalent to the difference between, say, 105 and 110 grants? We are likely to judge the second difference to be negligible while considering the first difference to be quite large. Analyzing the raw data implicitly treats these two differences as if they were equal. Also, as suggested above, counts are likely to require a square root (or similar) transformation for models of counts to have errors with homogeneous variance. Although one need not transform on both sides, it seems appropriate in this case to transform both the criterion (number of physical science grants) and the predictor (number of social science grants). Figure 13.12 shows the normal quantile plot of the residuals and the spread-location plot after the square root transformation has been applied with the outlier omitted (it remains an outlier, although not as extreme, after the transformation). Both plots suggest that the analysis of the transformed data without the outlier reasonably satisfies the normality and homogeneity of variance assumptions. Any weaker power transformation (i.e., a power between .5 and 1) leaves a positive slope in the spread-location plot, whereas any stronger transformation (such as the log or the inverse) induces a negative slope. Hence, the square root transformation (power = .5) is best for correcting the variance problems in these data.

An analysis of the transformed data reveals a second problematical observation—a land grant university that has 56 NSF grants in the physical sciences but none in the social sciences. Some land grant universities have remained closer to their roots as agricultural and mechanical universities and so do not have the full complement of social science departments; this university may be such an instance. Fortunately, the story does not change appreciably if this observation is included or omitted. So, we will stop with the analysis of the square root transformed data with the first outlier omitted; this final analysis is depicted in Figure 13.13. There is no evidence for a relationship between the number of physical and social grants at flagship universities (slope = 0.46, $F(1,14) = 0.6$, PRE = .04, $p = .45$) but there is a relationship for land grant universities (slope = 2.4, $F(1,14) = 15.65$, PRE = .53, $p = .0014$). The difference between the two slopes is statistically significant (slope difference = 1.93, $F(1,14) = 5.15$, PRE = .27, $p = .04$). (Omitting the second problematical observation would strengthen the land grant

FIGURE 13.12 Normal quantile and spread-location plots after square root transformation of grants data

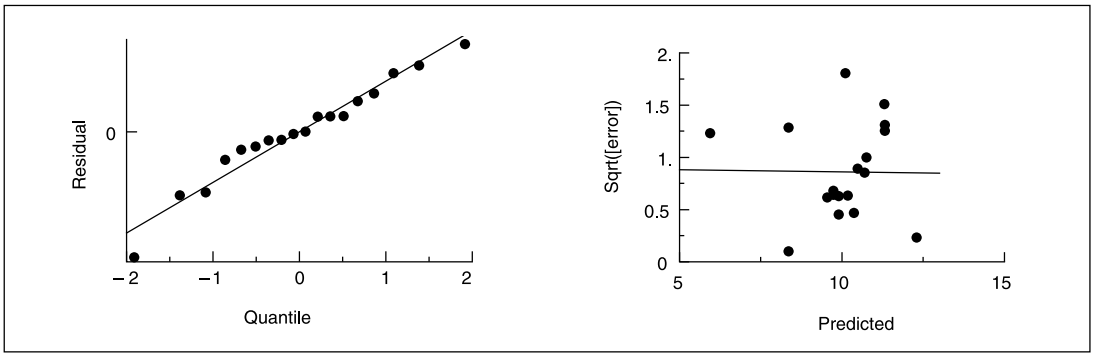
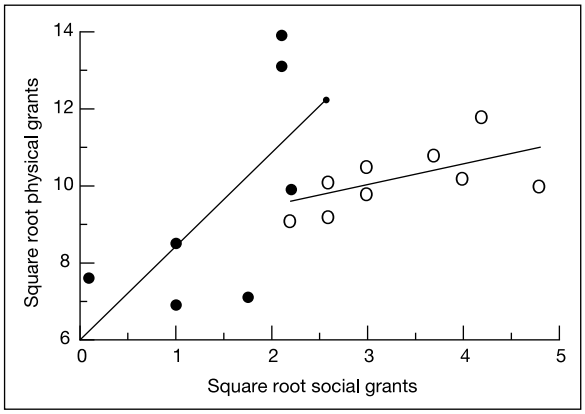


FIGURE 13.13 Final analysis of square-root-transformed data with one outlier omitted for flagship (○) and land grant (●) peer institutions



relationship and enhance the slope difference.) Transforming the original data and omitting the outlier yielded not only an analysis that satisfied the important statistical assumptions but also a clear and consistent story for these data. Including the outlier and not transforming the data produced an analysis that violated the major assumptions underlying the analysis and produced a muddled and inconsistent story for the data.

SUMMARY

In this chapter we have examined two important topics. The first concerns how one identifies and deals with observations that are outliers. We have seen that outliers can be defined in three ways: observations with unusual values on the data or outcome variable, observations with unusual values on the predictors, and observations that are unusual on both. Generally the first sort of outlier results in inflated error terms and Type II errors; the second tends to result in an overabundance of Type I errors; and the third sort is especially pernicious with outlying observations having the potential to seriously bias parameter estimates. We recommend examining the studentized deleted residual to identify outliers of the first sort, the lever to identify outliers of the second sort, and Cook's D to identify outliers of the third sort. Outliers that result from obvious errors in recording or coding data should be corrected if possible. Otherwise, analyses should be conducted without them. One should not let a few pernicious observations dictate the story and distort it from what it would be without them.

The second half of this chapter discussed graphic displays that can be used to detect violations of the assumptions that residuals are normally distributed with a constant variance. These displays include the normal quantile-quantile plot for detecting non-normality and Residual $\times$ Predicted and spread-location plots for detecting heterogenous errors. We concluded by discussing transformations that can be used to deal with the problems of errors that display these assumption violations.

Note

1 But remember that it is the distribution of the errors that one wants to plot, not the distribution of the Y variable.