10

Outliers and Multicollinearity: Diagnosing and Solving Regression Problems II

10.1 INTRODUCTION

In Chapters 2, 3, and 5 through 9 we focused on presenting multiple regression/correlation analysis as a general data analytic system. We have illuminated the flexibility and power of this analytical tool to answer a wide variety of research questions of interest to behavioral scientists. Our presentation has progressed from simple to complex models, from linear to nonlinear and interactive relationships, and from quantitative to qualitative to combinations of quantitative and qualitative IVs. The sole exception to this progression was in Chapter 4. There we considered problems that arise from the violation of the assumptions underlying multiple regression analysis. We considered graphical and statistical methods of detecting such violations and methods of solving these problems when they are detected. These procedures help researchers gain a fuller understanding of their data so that they do not report misleading results. The present chapter continues this theme, considering two problems in regression analysis that were not considered in Chapter 4.

First is the problem of *outliers*—one or more atypical data points that do not fit with the rest of the data. Outliers may represent data that are contaminated in some way (e.g., a recording error; an error in the experimental procedure). Or, they may represent an accurate observation of a rare case (e.g., a 12-year-old college student). Whatever the source of the outliers, they can in some cases have a profound impact on the estimates of the regression coefficients and their standard errors, as well as on the estimate of the overall prediction, R^2. We present graphical and statistical methods of detecting outliers and remedial approaches that may be taken when outliers are discovered. These methods become particularly important as the number of variables in the data set increases. They help researchers avoid reporting misleading results when outliers are present in the data.

Second is the problem of *multicollinearity*. This problem occurs in data sets in which one (or more) of the IVs is highly correlated with the other IVs in the regression equation. The estimate of the regression coefficient B_i for this correlated predictor will be very unreliable because little unique information is available from which to estimate its value—the regression coefficient will have a *very* large standard error. Although the estimate of the value of the regression coefficient B_i will on average be equal to the value in the population, its confidence interval will be so large as to make the estimate of little or no value. The regression coefficient will also

often become more difficult to interpret. We present methods of detecting multicollinearity and methods for addressing this problem when it does occur.

10.2 OUTLIERS: INTRODUCTION AND ILLUSTRATION

We turn first to the problem of outliers, one or more atypical data points that do not fit with the rest of the data. We begin with the presumption that the data being analyzed have been carefully entered. Ideally, the full data set has been entered a second time using a data entry program that cross checks the two sets of entries and identifies errors. Checks have been performed for out of range values (e.g., a score of 8 on a 1 to 5 scale) and logical inconsistencies (e.g., a person who reports no lifetime alcohol consumption who also reports he consumed five drinks during the past week). Yet, even under conditions in which the data set has been thoroughly cleaned and checked, errors, unusual cases, or both may be present. For example, Cleveland (1993) presents evidence indicating that there were serious undetected errors even in a classic data set that had been repeatedly analyzed (Immer, Hayes, & Powers, 1934; presented as an illustrative example, for example, by R. A. Fisher in his classic *Design of Experiments*, 1971). Hoaglin and Velleman (1995) present a case study showing that analytic teams that did not perform adequate checks for outliers overlooked errors in a large data set, produced incorrect regression results, and reached seriously flawed conclusions. On a more successful note, climatologists checking for outliers discovered an anomalous observation of too low a reading for ozone levels in the upper atmosphere over Antarctica. Further study of this outlier led to the discovery of the Antarctic ozone hole, which has raised world concerns about loss

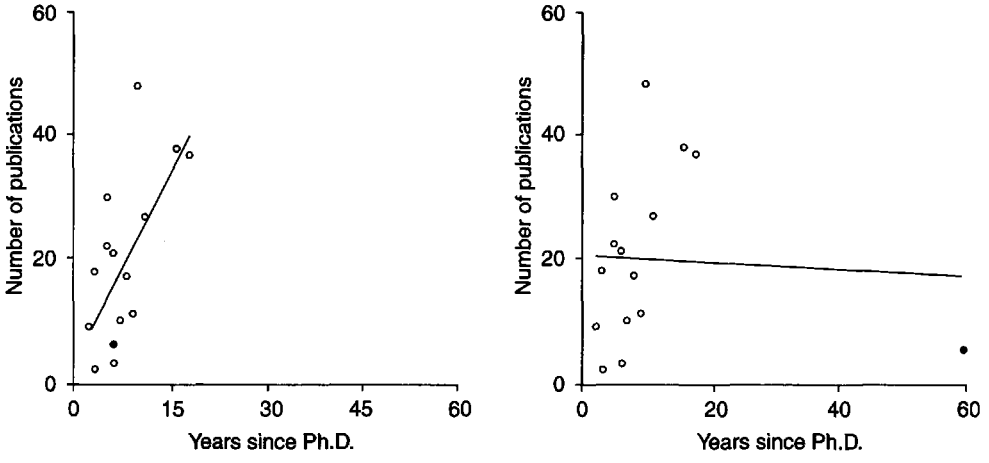
TABLE 10.2.1
Years Since Ph.D. and Number of Publications: Data


CH10EX01

	A. Original data set		B. Data set containing outlier for case 6	
Case	*X* Years since Ph.D.	*Y* Number of publications	*X* Years since Ph.D.	*Y* Number of publications
1	3	18	3	18
2	6	3	6	3
3	3	2	3	2
4	8	17	8	17
5	9	11	9	11
6	6	6	**60**	6
7	16	38	16	38
8	10	48	10	48
9	2	9	2	9
10	5	22	5	22
11	5	30	5	30
12	6	21	6	21
13	7	10	7	10
14	11	27	11	27
15	18	37	18	37

Note: The data in A are from Table 2.2.2. The data in B are identical to those in A except that the observation on *X* for case 6 has been replaced by an outlier. The outlier in B is highlighted in boldface type.

(A) Original data set ($n = 15$). (B) Data set containing outlier for case 6 ($n = 15$).



	Estimate	*SE*	*t*	*p*		Estimate	*SE*	*t*	*p*
B_0	4.73	5.59	0.85	*ns*	B_0	20.61	4.78	4.31	<.001
B_1	1.98	0.63	3.14	<.01	B_1	−0.06	0.27	−0.22	*ns*
$MS_{residual} = 117.04$					$MS_{residual} = 204.97$				
$R^2 = .43.$					$R^2 = .004.$				

Note: Years since Ph.D. is shown on the abscissa (x axis). Number of publications is shown on the ordinate (y axis). Case 6 is denoted by a • in each plot. The best fitting linear regression line is superimposed in each plot. The results of the regression analysis are presented below each part.

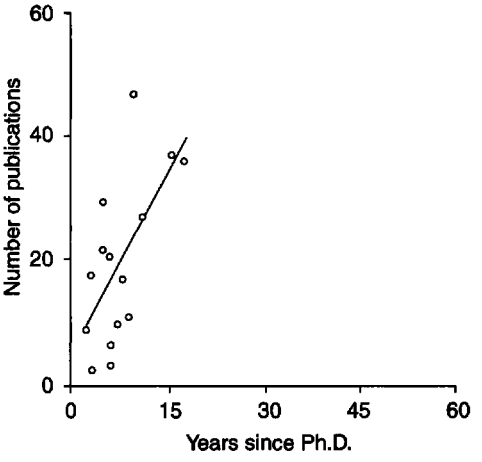
FIGURE 10.2.1 Plot of years since Ph.D. vs. number of publications.

of ozone in the upper atmosphere.

To provide a basis for consideration of outliers, let us reconsider the faculty salary example originally presented in Chapter 2. Suppose we had found an observation that indicated that a faculty member received his Ph.D. 60 years ago! This observation may represent an error in the data (e.g., a transcription error in which 6 is mistakenly recorded as 60). Alternatively, such a data point might represent an accurate observation of a rare case—a faculty member in his 80s who is still holding a full-time position.[1] When outliers are present, regression analyses may produce results that strongly reflect a small number of atypical cases rather than the general relationship observed in the rest of the data. Even one outlier in a data set can produce a dramatic result (e.g., an interaction that disappears when the outlier is removed). Thus, a researcher may report exciting and unexpected "new results" only to discover later that they cannot be replicated because they were produced by an outlier. On the other hand, important predicted results may not be detected because they are masked by outliers. If the outliers are detected and appropriate remedial actions are taken, then the important predicted effect will emerge. The impact of outliers will typically decrease as sample size increases. However, under conditions in which there is substantial multicollinearity or the regression equation contains interactions or power polynomials, outliers may have dramatic effects even in moderate or large samples.

[1] In 2000, at least one major university had an active professor who was still teaching full time at the age of 100.

(C) Data set deleting case 6 ($n = 14$).



	Estimate	*SE*	*t*	*p*
B_0	6.00	5.71	1.05	*ns*
B_1	1.92	0.63	3.03	.01

$MS_{residual} = 116.60$.
$R^2 = .43$.

Note: Years since Ph.D. is shown on the abscissa (x axis). Number of publications is shown on the ordinate (y axis). Case 6 is deleted. The best fitting linear regression line is superimposed. The results of the regression analysis are presented for each part.

FIGURE 10.2.1 Continued.

Table 10.2.1A reproduces the original 15 cases of the faculty salary data set originally presented in Chapter 2. Corresponding to these data is Fig. 10.2.1(A), which depicts the regression line and the results of the regression analyses based on these 15 cases. The regression estimate is $\hat{Y} = 1.98X + 4.73$, where $\hat{Y}$ is the predicted number of publications and X is years since Ph.D. For simplicity in our initial presentation, we will presume that the original data set is correct and study what happens to the results of our regression analysis if one data point is replaced by an outlier.

Let us start with the case just presented. The original case with $X = 6$ is replaced by a single outlying case with $X = 60$ years. Table 10.2.1B illustrates this situation. All of the entries in the right panel are identical except that for subject 6, the number of years since the Ph.D. is 60 instead of 6 (in boldface type). How would this affect our results? Figure 10.2.1(B) shows the regression line and the results of the regression analysis based on this altered data set. The results change dramatically: The relationship between years and publications has disappeared, the R^2 dropping from .43 in the original data set to .00 in the data set containing the single outlier. For comparison purposes, Fig. 10.2.1(C) shows the results of another analysis, which we consider in Section 10.3. In this third analysis, outlying case 6 is dropped from the data set and the regression analysis is recomputed. The results are similar but not identical to the analysis presented in Fig. 10.2.1(A) with the original data. As this example dramatically illustrates, a single outlier can potentially have a major impact on the results of a regression analysis, especially with a small sample.

10.3 DETECTING OUTLIERS: REGRESSION DIAGNOSTICS

In the example presented in Fig. 10.2.1(B), the outlier (case 6) was easy to detect. The errant data point was far from the rest and could be detected by visual inspection of the raw data or the scatterplot of X (years) by Y (publications). Scatterplot matrices, presented in Section 4.2, can also be very useful in identifying outliers when there is more than one independent variable. In cases when there are more than one or two IVs, some outliers may be difficult to identify by such visual inspection. We encourage analysts to supplement such visual inspection with the use of specialized statistics known as regression diagnostics which can greatly aid in the detection of outliers. Regression diagnostics are *case statistics*, meaning there will be one value of each diagnostic statistic for each of the n cases in the data set. A sample of 150 cases will produce 150 values of each diagnostic statistic, one representing each case in the data set. Regression diagnostic statistics are used to examine three characteristics of potentially errant data points. The first is *leverage*: How unusual is the case in terms of its values on the IVs? The second is the *discrepancy* (or distance[2]) between the predicted and observed values on the outcome variable (Y). The third is *influence*, which reflects the amount that the regression coefficients would change if the outlier were removed from the data set. Conceptually, influence represents the product of leverage and discrepancy. Each of these characteristics should be examined, as they identify different aspects of errant data points.

In this section, we present definitions and simple worked examples to convey a conceptual understanding of the meaning of each major diagnostic statistic. We illustrate with the two data sets introduced in Section 10.2: (1) The original 15-case data set from Chapter 2 and (2) the same data set, in which 60 is used as the value for years since Ph.D. for case 6. The comparison of the results for the two data sets highlights how relatively extreme values of the diagnostic statistics may be obtained when there is an outlier in the data set, here case 6. For convenience in our initial presentation, the value of the outlier in data set (2) was chosen to produce extreme values on each of the measures of leverage, discrepancy, and influence. This need not be the case: Outliers may produce high values on leverage, but not discrepancy—or high values on discrepancy, but not leverage. We return to this issue and consider these possibilities at this end of this section. Each of the diagnostic statistics provides somewhat different information that is useful in identifying and understanding the effects of potentially errant points in the data. Remedial actions that may be taken when outliers are detected follow in Section 10.4.

10.3.1 Extremity on the Independant Variables: Leverage

Leverage reflects only the case's standing on the set of IVs. For each case, leverage tells us how far the observed values for the case are from the mean values on the set of IVs. When there is only one IV, leverage can be determined as

$$\text{leverage} = h_{ii} = \frac{1}{n} + \frac{(X_i - M_X)^2}{\sum x^2}, \tag{10.3.1}$$

where h_{ii} is the leverage for case i, n is the number of cases, X_i is the score for case i on the predictor variable, M_X is the mean of X, and Σx^2 is the sum over the n cases of the squared deviations of X_i from the mean. If case i has a score at the value of M_X, then the second term of equation 1 will be 0 and h_{ii} will have the minimum possible value of $1/n$. As case i's score on X gets further and further from M_X, h_{ii} increases in size. The maximum value of h_{ii} is 1.0.

[2]To avoid confusion, we use the term *discrepancy* to represent the difference between the observed and predicted value of Y for specified values on each of the predictors. The term *distance* is associated with indices of leverage, discrepancy, and influence in the regression diagnostics literature.

The mean value of leverage for the n cases in the sample is $M_{h_{ii}} = (k + 1)/n$, where k is the number of IVs.

To illustrate, let us return to the original 15 cases from Table 10.2.1A. Consider first case 4, with 8 years since Ph.D., a value very close to the mean of 7.67. The value of Σx^2 is 293.33. Substituting into Eq. (10.3.1), we find

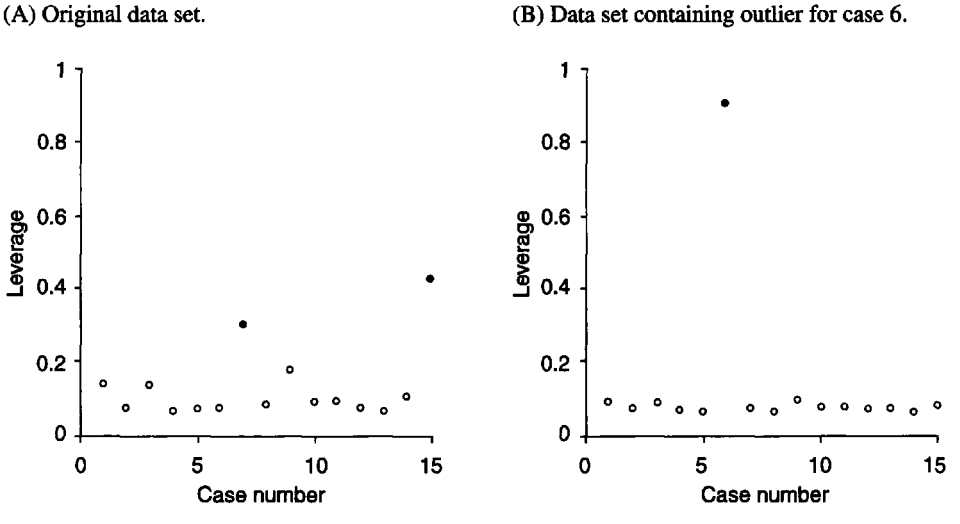
$$h_{ii} = \frac{1}{15} + \frac{(8 - 7.67)^2}{293.33} = .0670,$$

which is very close to the minimum value of $1/n$ for h_{ii}, here $1/15 = .0667$. In contrast, consider case 15, with 18 years since Ph.D., the most extreme value of X in the original data set. Substituting case 15's value of 18 into Eq. (10.3.1) yields

$$h_{ii} = \frac{1}{15} + \frac{(18 - 7.67)^2}{293.33} = .43,$$

a considerably larger value. Not surprisingly, this value of leverage is larger than the mean leverage value for our sample, $M_{h_{ii}} = (1 + 1)/15 = .13$.

Figure 10.3.1 is an index plot that displays the values of leverage for each of the cases in the original data set. Index plots provide a convenient method of displaying the value of regression diagnostic statistics in small and moderate sized data sets by displaying the value of the diagnostic statistic on the ordinate (vertical or y axis) and the case number on the abscissa (horizontal or x axis). Index plots make it easy to identify those cases that have particularly extreme values of the diagnostic statistic. Figure 10.3.1(A) displays the leverage values for the original data set in which $X = 6$ for case 6. Figure 10.3.1(B) displays the leverage values for the data set that includes the outlier ($X = 60$ for case 6). The highest values of leverage for each data set are highlighted (•) in the figure.



Note: The case number for each participant is shown on the abscissa (x axis). The value of leverage (h_{ii}) is shown on the ordinate (y axis). Cases with relatively high values of leverage are indicated by • in each panel. In Fig. 10.3.1(A), which contains the original data, cases 7 ($h_{ii} = .30$) and 15 ($h_{ii} = .43$) have somewhat higher leverage values than the other points since their years since Ph.D. are the most extreme in the data set (case 7, $X = 16$; case 15, $X = 18$). In Fig. 10.3.1(B), which contains the outlier, case 6 ($X = 60$; $h_{ii} = .90$) has an extremely high value for leverage that differs dramatically from the values for leverage of the other cases.

FIGURE 10.3.1 Index plot of leverage vs. case number.

Cases further from the mean of the single IV have a greater *potential* to influence the results of the regression equation. The leverage values identify these cases. Whether a particular case actually influences the regression coefficients or R^2 also depends on the discrepancy between the observed and predicted values of Y for that case, $Y_i - \hat{Y}_i$.

The basic idea of leverage generalizes directly to regression models in which there is more than one IV. We are now interested in how far case i's score on each of the k independent variables, $X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ik}$, is from the centroid of the independent variables. Recall that the *centroid* is the point corresponding to the mean of the independent variables, $M_1, M_2, M_3, \ldots, M_k$. Conceptually, the sum of each case's squared deviations from the sample's means across the IVs is "adjusted" by the correlation between each pair of IVs. Because the algebraic expressions become complex,[3] statistical packages are used to compute the value of h_{ii} for each case in the sample.

As an illustration, consider the two-predictor regression equation presented in Section 3.2. Years since Ph.D. (X_1) and number of publications (X_2) are the independent variables, and salary (Y) is the dependent variable. The data for the 15 cases are presented in Chapter 3 in Table 3.2.1. Figure 10.3.2 presents a scatterplot of the two independent variables. In this scatterplot the centroid of the X_1X_2 space, $M_{X_1} = 7.67, M_{X_2} = 19.93$, is indicated by the symbol $\times$. We have identified four of the cases in the scatterplot by case number. Note that case 4 ($X_1 = 8, X_2 = 17, h_{ii} = .08$) and case 12 ($X_1 = 6, X_2 = 21, h_{ii} = .09$) are located close to the centroid and have values of h_{ii} that are only slightly higher than the minimum leverage value $1/n = .07$. In contrast, case 8 ($X_1 = 10, X_2 = 48, h_{ii} = .45$) and case 15 ($X_1 = 18, X_2 = 37, h_{ii} = .44$) are located at a greater distance from the centroid as is indicated by the substantially higher leverage values. The standard statistical packages include options that compute h_{ii} for all cases in the data set. Note that leverage is based *only* on the IVs in the regression model. Changing only the DV in a regression equation will not affect the leverage values.

Centered Leverage Values

Caution must be employed in interpreting leverage values because some statistical packages[4] calculate a centered index of leverage which we will term h_{ii}^*. The centered index h_{ii}^* may be expressed in terms of the unstandardized index, h_{ii}, as

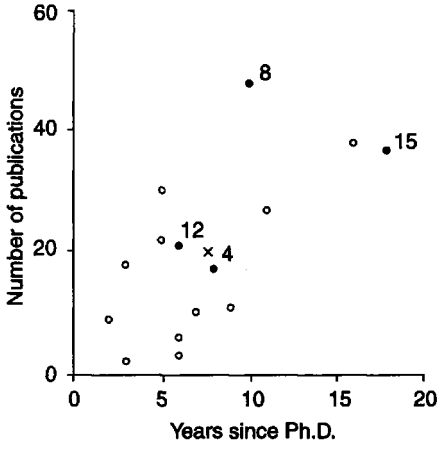
$$h_{ii}^* = h_{ii} - \frac{1}{n}.$$

The minimum possible value of h_{ii}^* is 0, and the maximum value is $1 - 1/n$. To illustrate, we calculated that $h_{ii} = .43$ for case 15 in Table 10.2.1A, so $h_{ii}^* = .43 - 1/15 = .43 - .07 = .36$.

---

[3]For readers familiar with matrix algebra, a simple matrix algebra expression is available to calculate leverage. The hat matrix is defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. $\mathbf{H}$ is an $n \times n$ matrix and $\mathbf{X}$ is the $n \times (k + 1)$ augmented matrix of the predictor scores. This matrix has a 1 in the first column of the matrix for each case. The next k columns of the matrix contain each case's (participant's) scores on the k IVs (see Appendix 1 for an overview of matrix algebra). The main diagonal of $\mathbf{H}$ contains the leverage values. Thus, the leverage for case i, h_{ii}, is the value for the ith row and ith column of $\mathbf{H}$. It is this diagonal value that gives the leverage its identification as h_{ii}.

[4]SAS and SYSTAT programs report h_{ii} whereas SPSS reports the corresponding centered value h_{ii}^*. h_{ii} is labeled in SAS output as "Hat Diag H" and SYSTAT output as "leverage." h_{ii}^* is labeled in SPSS output as "lever." Conceptually, h_{ii} has a value of $1/n$ for a case at the centroid of the IVs because the case can still potentially affect the value of the intercept. h_{ii}^* excludes consideration of the cases potential effect on the intercept as would occur if both the IVs and DV were centered.



Note: Data set is from Table 3.2.1. Values of years since Ph.D. (X_1) are shown on the abscissa. Values of number of publications (X_2) are shown on the ordinate. The open and filled in circles depict the values of X_1 and X_2 for the 15 cases. × is the centroid ($M_{X_1} = 7.67; M_{Y_1} = 19.93$) of the X_1X_2 space. The four filled circles (with their case numbers) are the four cases presented in the text.

FIGURE 10.3.2 Scatterplot of time since Ph.D. vs. number of publications.

Guidelines for Identifying Cases With High Leverage Values

Two general approaches have been suggested for identifying cases with high leverage. The first approach is to plot the distribution of the h_{ii} values and to identify a very small number of cases with leverage values that are *substantially* higher than those of the other cases. Index plots are typically used in small or moderate sized samples; histograms with a large number of bins or stem and leaf displays are often used in large samples. Figure 10.3.1(B) presents an index plot of leverage by case number for the data set with the outlying case 6 ($X = 60$). The leverage of case 6 sharply stands out from the other cases.

The second approach is to examine leverage values that fall above rough rule of thumb cutoff values. Different authors have proposed different rule of thumb cutoff values. So that the cutoff values we present are consistent across the various regression diagnostic measures, we will present the guidelines of Belsley, Kuh, and Welsch (1980) that identify approximately the most extreme 5% of the cases when all of the predictors are normally distributed. For leverage, they proposed that values of h_{ii} greater than $2M_h = 2(k+1)/n$ be considered to have high leverage when both the number of predictors and the number of cases are large. For small samples, a more stringent cutoff of about $3M_h = 3(k+1)/n$ is sometimes recommended to avoid identifying too many points for examination. In the present small sample ($n = 15$) case, values greater than (3)(.13) = .39 might be selected for possible examination. For the centered measure of leverage, h_{ii}^*, the cutoffs will be lower. Since $h_{ii}^* = h_{ii} - (1/n)$, the corresponding cutoffs for the centered measure h_{ii}^* will be $2k/n$ in large samples and $3k/n$ in small samples.

Belsley, Kuh, and Welsch's guidelines identify a *minimum* threshold at which it may be worthwhile to identify cases for examination. In large samples, the use of the Belsley, Kuh, and Welsch guidelines will nearly always identify far too many cases. In practice, we encourage

analysts to identify for examination only a *very small* number of cases that have the highest leverage values. In particular, those cases for which there is a *large* gap in the value of leverage from the remainder (i.e., unusual values) should be carefully checked for accuracy. When no cases exceed Belsley, Kuh, and Welsch's rule of thumb cutoffs, special checking of cases with relatively high leverage values is not indicated.[5] Chatterjee and Hadi (1988, Chapter 4) present a full discussion of possible cutoff values for leverage.

Mahalanobis Distance

A measure that is closely related to leverage is reported by some statistical packages (e.g., SPSS). This measure, known as Mahalanobis distance, is a measure of the distance between the specific case's values on the predictor variables and the centroid of the IVs. Weisberg (1985) points out that Mahalanobis distance can be expressed as $(n-1)h_{ii}^{*}$. Thus, Mahalanobis distance provides the same information as leverage, but will have different rule of thumb cutoffs. J. P. Stevens (1984) presents more information on Mahalanobis distance and conventional cutoff scores for the interpretation of this measure.

10.3.2 Extremity on *Y*: Discrepancy

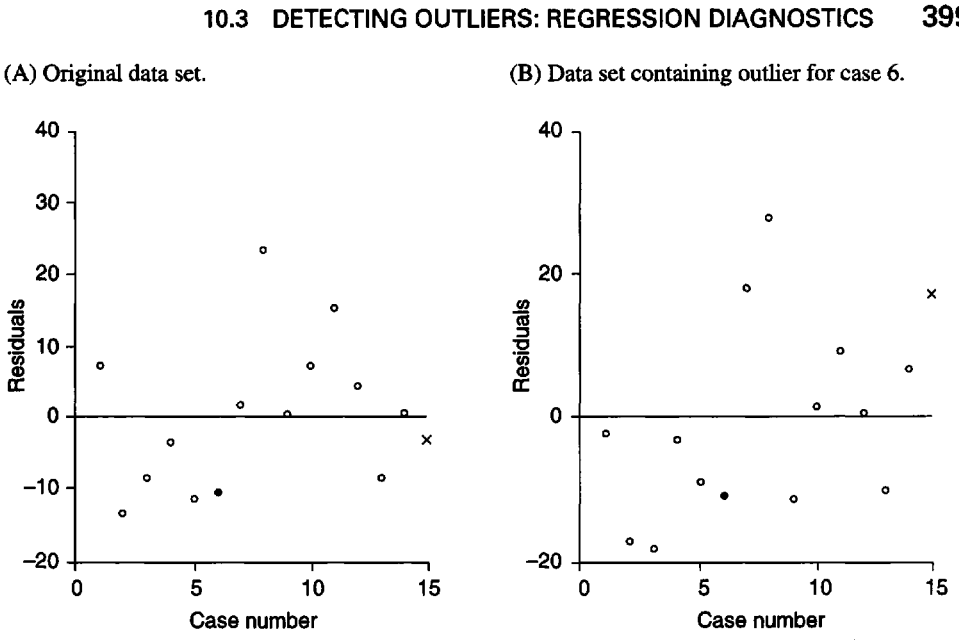
A second set of statistics measures the *discrepancy* or distance between the predicted and observed values on Y. In Chapters 2 and 3, we saw that the raw residual for case i, $e_i = Y_i - \hat{Y}_i$, typically provides an excellent measure of this discrepancy. However, reconsider the regression lines and the data presented in Fig. 10.2.1(A) and (B). Figure 10.3.3(A) displays the raw residuals for the original data and Fig. 10.3.3(B) displays the residuals from the data containing the outlier. As can be seen, the discrepancy between case 6 (marked by •) and the regression line in Fig. 10.3.3(B) is smaller than the discrepancy for several of the other cases. In essence, the outlying point has pulled the regression line toward it to improve the overall fit. Other diagnostic statistics are needed that are less influenced by this problem. Two are commonly calculated by statistical packages, internally studentized residuals and externally studentized residuals. Externally studentized residuals will nearly always be the preferred measure of discrepancy.

Internally Studentized Residuals

Internally studentized residuals address one of two problems associated with raw residuals. As a case's scores on the predictor get further from the centroid, the estimate of the value of the residual for that case gets more precise (Behnken & Draper, 1972). The expected variance of the residual for case i can be expressed as:

(10.3.2) $$\text{variance}(e_i) = \text{MS}_{\text{residual}}(1 - h_{ii}).$$

[5] Whenever humans transcribe and key data into the computer, errors leading to incorrect values should be presumed to occur in a small proportion of the cases. As noted in Section 10.2, a variety of checks on the accuracy should always be performed on the data prior to any analysis. Examination of leverage values provides an important additional check because other procedures do not identify multivariate outliers.



Note: Values of the raw residuals are presented on the ordinate. Case number is presented on the abscissa. The horizontal line in each panel represents a value of 0 for the raw residual. The highlighted point (•) is case 6. Note that the magnitude of this residual is not particularly large in either part because the outlying point in (B) pulls the regression line toward itself. In contrast, case 15 (years since Ph.D. = 18; number of publications = 37) is marked by × in each panel. In (A), this case has a small negative residual (−3.42). In (B), the regression line has been pulled away from this point by the outlier, so the residual is now positive and much larger in magnitude (+17.47).

FIGURE 10.3.3 Index plot of residuals vs. case number.

Recall from Section 3.6 that $MS_{residual}$ is the estimate of overall variance of the residuals around the regression line $= (1 - R^2)(\Sigma y^2)/(n - k - 1)$. h_{ii} is the leverage of case i. The standard deviation of the residual for case i is then

$$sd_{e_i} = \sqrt{MS_{residual}(1 - h_{ii})}.$$

The internally studentized residual takes the precision of the estimate of the residual into account. The internally studentized residual is the ratio of the size of the residual for case i to the standard deviation of the residual for case i,

(10.3.3) $$\text{internally studentized residual}_i = \frac{e_i}{sd_{e_i}}.$$

The magnitude of the internally studentized residual ranges between 0 and $\sqrt{n - k - 1}$ (Gray and Woodall, 1994). Unfortunately, internally studentized residuals do not follow a standard statistical distribution (the numerator and denominator in Eq. 10.3.3 are not independent), so they can *not* be interpreted using normal curve or t tables.

Externally Studentized Residuals

Externally studentized residuals directly address a second issue associated with outliers. Recall that the outlier can pull the regression line toward itself as we saw in Fig. 10.2.1(B). Externally studentized residuals address this issue by considering what would happen if the outlying case were deleted from the data set.

In Fig. 10.2.1(B) we saw that case 6 was an outlier. Suppose that we deleted case 6 from the data set and recalculated the regression equation based only on the other $n - 1 = 14$ cases. The results of this analysis are presented in Fig. 10.2.1(C). With case 6 deleted, the new regression equation is $\hat{Y}_{i(i)} = 1.92X_i + 6.00$. The notation $\hat{Y}_{i(i)}$ indicates that we are calculating the predicted value for case i, but with case i deleted from the data set. The outlier contributes substantially to the estimate of the variance of the residuals around the regression line, $MS_{residual}$. $MS_{residual(i)}$ for the new regression equation with case 6 deleted is 116.6, whereas $MS_{residual}$ for the full 15 cases (including the outlier, case 6) is 204.0.

Using the new regression equation with case 6 deleted, we calculate the predicted value for case 6 based on this new regression equation with case 6 deleted: $\hat{Y}_{i(i)} = 1.92(60) + 6.00 = 121.05$. We define the deleted residual d_i as the difference between the original Y observation for case i and the predicted value for case i based on the data set with case i deleted:

$$d_i = Y_i - \hat{Y}_{i(i)}.$$

In our present example, $d_i = 6 - 121.05 = -115.05$. For purposes of comparison, we calculated the raw residual based on all 15 cases in the data set, $e_i = Y_i - \hat{Y}_i = 6 - 17.00 = -11.00$. The greater magnitude of the deleted residual than of the raw residual helps highlight case 6 as an outlier. Case 6 can no longer hide itself by drawing the regression line toward itself.

The externally studentized residual draws on this idea of deletion of case i to remove its influence. The externally studentized residual for case i, t_i, is calculated as follows:

(10.3.4)
$$t_i = \frac{d_i}{SE_{d_i}}.$$

Paralleling the general form of Eq. (10.3.3) for the internally studentized residual, the numerator is now the *deleted* residual for case i, and the denominator is the standard error of the *deleted* residual for case i. Most sources attempt to simplify Eq. (10.3.4). The deleted residual d_i can also be computed from the raw residual e_i:

$$d_i = \frac{e_i}{1 - h_{ii}}.$$

The standard error of the deleted residual for case i can also be expressed as

$$SE_{d_i} = \sqrt{\frac{MS_{residual(i)}}{1 - h_{ii}}}.$$

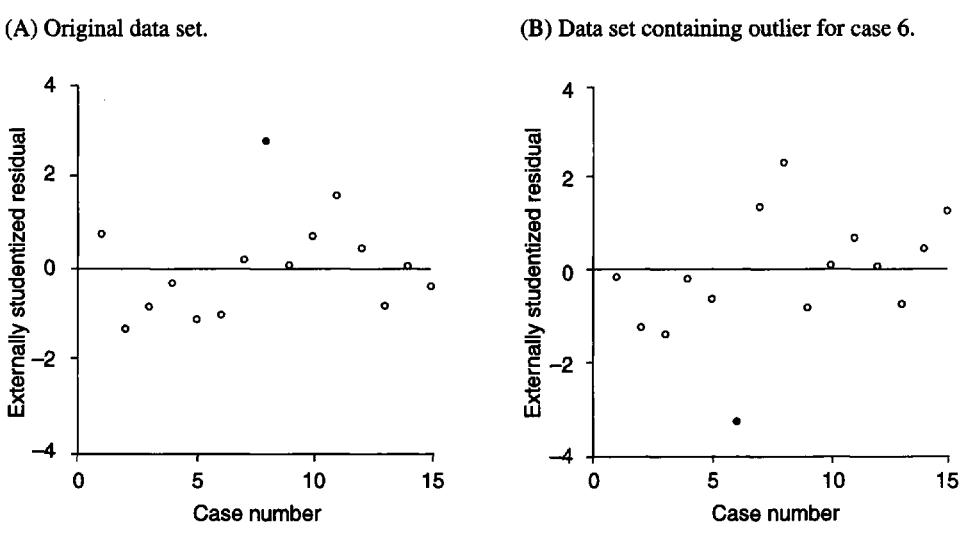
If these values are substituted into Eq. (10.3.4) and the resulting expression is simplified, the internally studentized residual t_i can be expressed in terms of the following equation:

(10.3.5)
$$t_i = \frac{e_i}{\sqrt{MS_{residual(i)}(1 - h_{ii})}}.$$

Here, e_i is the raw residual, $MS_{residual(i)}$ is the mean square residual with case i deleted from the data, and h_{ii} is the leverage for case i. When years has a value of 60 for case 6:

$$t_i = \frac{-11.00}{\sqrt{116.60(1 - .90)}} = -3.29$$

Standard statistical packages will compute the externally studentized residual for all cases in the data set. In Fig. 10.3.4, we present index plots of the externally studentized residuals.



Note: Case numbers are shown on the abscissa. Values of the externally studentized residual (t_i) are shown on the ordinate. The horizontal line represents a value of 0 for t_i. In (A), the value of t_i for case 8 is larger in magnitude (2.80) than any of the other points. In (B) the value of t_i for case 6, the outlier, is larger in magnitude (−3.29) than any of the other points. The case with largest magnitude of t_i is identified with the symbol • in each panel.

FIGURE 10.3.4 Index plot of externally studentized residuals (t_i) vs. case number.

In Fig. 10.3.4(A) we present the plot of t_i for the original 15 cases in which $X = 6$ for case 6; in Fig. 10.3.4(B) we present the plot of t_i for the 15 cases including the outlier—case 6 has a value of $X = 60$. In Fig. 10.3.4(B), the magnitude of t_i for case 6 (denoted by •, $t_i = -3.29$) is the most extreme value for the cases in the data set.

Guidelines for Identifying Cases with High Discrepancy Values

The externally studentized residual is the preferred statistic to use to identify cases whose Y values are highly discrepant from their predicted values. As with h_{ii}, one good strategy is to use an index plot like that in Fig. 10.3.4 to identify a very small number of cases that have the most extreme values of t_i in the data set for examination. Outliers for which there are *large* gaps in the value of t_i (ignoring sign) from the remainder of the cases merit particular attention.

Alternatively, recommendations have been made for cutoff values for t_i. If the regression model fits the data, the externally studentized residuals will follow a t distribution with $df = n - k - 1$. About 5% of the cases are expected to be greater than about 2.0 in magnitude for moderate to large sample sizes. Therefore some authors recommend that a value of ±2.0 should be chosen as a cutoff for selecting cases to examine. However, once again the use of this cutoff can result in far too many cases that would need to be examined in large samples, even if there are no real outliers in the data. For example, if $n = 1000$, about 50 cases (5%) would be selected, a very large number for individual attention. Consequently, many data analysts use a higher cutoff score (e.g., ±3.0, ±3.5, ±4.0) in larger samples. Once again, both large positive and large negative values of externally studentized residuals indicate a point that is discrepant from the rest.

Beckman and Cook (1983) have suggested a procedure for testing the significance of the *largest* studentized residual. They propose that the Bonferroni procedure be used to adjust the level of α based on the number of cases in the sample (i.e., the number of cases that can potentially be tested). The value chosen should be α/n. For example, in the present sample of 15 cases with $\alpha = .05$, two tailed, $\alpha/n = .05/15 = .0033$. For $df = n - k - 1 = 13$, the

critical two tailed value of t for $\alpha = .0033$ can be found using extensive statistical tables or computed using standard statistical packages. Here, the exact critical value of 3.23 is less than the magnitude of the $t_i = -3.29$ for the largest outlier, case 6, so we would conclude that the observed Y value for case 6 showed a statistically significant discrepancy from its predicted value. Note that the case with the highest discrepancy in the original data set is case 8 (see Fig. 10.3.4A). Its value of $t_i = 2.80$ does not exceed the Bonferroni adjusted critical value of 3.23.

A note on terminology. Our terminology for the internally studentized and externally studentized residuals follows that of Cook and Weisberg (1982). However, considerable confusion is created in this area because authors have failed to use consistent terminology in referring to these statistics. The internally studentized residual has been given terms such as the standardized residual and studentized residual; the externally studentized residual has been given terms such as the studentized residual and the studentized deleted residual. The internally studentized residual is labeled "SRESID" in SPSS and "Student Residual" in SAS. The externally studentized residual is labeled "SDRESID" in SPSS, "RStudent" in SAS, and "Student" in SYSTAT output. When consulting other sources or referring to the output of other computer programs, researchers should take care to be sure they understand which statistic is being reported.

10.3.3 Influence on the Regression Estimates

Measures of influence combine information from measures of leverage and discrepancy to inform us about how the regression equation would change if case i were removed from the data set. Two types of measures of influence are commonly considered. First, global measures of influence (*DFFITS*, Cook's *D*) provide information about how case i affects overall characteristics of the regression equation. Second, specific measures of influence (*DFBETAS*) provide information about how case i affects each individual B. Generally, both global and specific measures of influence should be examined.

Global Measures of Influence

Standard statistical packages report one or both of two global measures of influence, $DFFITS_i$ (Belsley, Kuh, and Welsch, 1980) or Cook's D_i (Cook, 1977). Like the externally studentized residual, both are deletion statistics that compare aspects of the regression equations when case i is included versus is not included in the data set. The two global measures of influence are very closely related; analysts may use the measure they prefer as the two measures provide redundant information.

DFFITS$_i$. The first global measure of influence is $DFFITS_i$, which is defined as

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MS}_{\text{residual}(i)} h_{ii}}}, \tag{10.3.6}$$

where $\hat{Y}_{i(i)}$ is the predicted value of Y if case i were deleted from the data set. The numerator of Eq. (10.3.6), sometimes termed *DFFIT*, tells us how much the predicted value for case i would change in the raw score units of Y if case i were deleted from the data set. The denominator serves to standardize this value so that $DFFITS_i$ estimates the number of standard deviations by which $\hat{Y}_i$, the predicted value for case i, would change if case i were deleted from the data set. *DFFITS* stands for "difference in fit, standardized."

To illustrate conceptually how $DFFITS_i$ is calculated, consider once again the data set with the outlier presented in Table 10.2.1B. As shown in Fig. 10.2.1(B), the regression equation

with all 15 cases (including case 6) included is $\hat{Y}_i = -0.06X_i + 20.61$. For case 6, the predicted value is $\hat{Y}_i = -(0.06)(60) + 20.61 = 17.00$. As we saw in Section 10.2.2, when case 6 is dropped from the data set and the regression equation is computed based on the remaining 14 cases, the new regression equation is $\hat{Y}_{i(i)} = 1.92X + 6.00$, so $\hat{Y}_{i(i)} = 121.05$. Thus, the change in the predicted value of Y that results from deleting case i from the data set is $\hat{Y}_i - \hat{Y}_{i(i)} = -104.05$—an enormous difference in the predicted number of publications for this faculty member! Recall from Section 10.3.2 that with case 6 deleted $MS_{residual(i)} = 116.60$ and from Fig. 10.3.1(B) that h_{ii} for case 6 is .90. Substituting into Eq. (10.3.7), we find that for case 6,

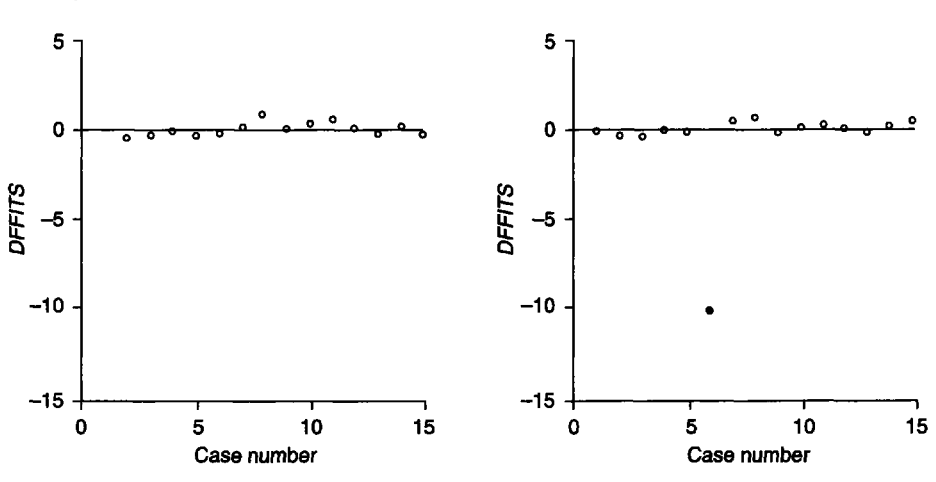
$$DFFITS_i = \frac{17.00 - 121.05}{\sqrt{(116.60)(.90)}} = -10.13,$$

a change of over 10 standard deviations. In Fig. 10.3.5, we present an index plot of the values of $DFFITS_i$ for (a) the original data set and (b) the data set containing the outlying value for case 6. As can be seen, in Fig. 10.3.5(B), in which case 6 is an outlier, the value of $DFFITS_i$ for this case differs greatly from the values for the other cases.

Earlier we noted that measures of influence can be thought of as reflecting the product of leverage and discrepancy. Another expression for $DFFITS_i$ that is algebraically equivalent to Eq. (10.3.6) clearly shows this relationship:

$$DFFITS_i = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}. \tag{10.3.7}$$

The first term in this equation is t_i, the externally studentized residual for case i. The second term is a function of the leverage for case i, h_{ii}. As values of t_i and h_{ii} both increase in magnitude, the magnitude of $DFFITS_i$ will also increase indicating the case has a larger influence on the results of the regression analysis. $DFFITS_i$ has its minimum magnitude of 0 when the deletion of case i has no effect on the predicted value of Y, $\hat{Y}_i$. $DFFITS_i = 0$ when case i falls exactly on the regression line so that $\hat{Y}_i$ will not change when case i is deleted. Cases at the centroid



Note: Values of $DFFITS_i$ are presented on the ordinate. Case number is presented on the abscissa. The highlighted point in (B) corresponds to case 6, the outlier. The horizontal line in each panel represents a value of 0 for $DFFITS_i$.

FIGURE 10.3.5 Index plot of $DFFITS_i$ vs. case number.

of the sample can still have some influence[6] because the minimum value of h_{ii} is $1/n$. The sign of $DFFITS_i$ will be positive when $Y_i > \hat{Y}_{i(i)}$ and negative when $Y_i < \hat{Y}_{i(i)}$. Most standard statistical packages will compute $DFFITS_i$ for all cases in the data set.

***Cook's* D_i.** An alternative measure of the global influence of case i on the results of the regression equation known as Cook's D_i is also reported by statistical packages. Cook's D_i can be expressed as

$$\text{Cook's } D_i = \frac{\sum(\hat{Y} - \hat{Y}_{(i)})^2}{(k+1)\text{MS}_{\text{residual}}}. \tag{10.3.8}$$

Thus, Cook's D_i compares the predicted value of Y with case i included and deleted for all cases in the data set. These differences are squared and then summed. The denominator serves to standardize the value. Cook's D_i ranges upward from its potential minimum value of 0 with higher numbers indicating the case has a larger influence on the results of the regression analysis. Unlike $DFFITS_i$, Cook's D_i will always be ≥ 0; it cannot be negative.

$DFFITS_i$ and Cook's D_i are closely related measures. Cook and Weisberg (1982) have shown that Cook's D_i and $DFFITS_i$ have the following mathematical relationship

$$\text{Cook's } D_i = \frac{(DFFITS)_i^2 \text{MS}_{\text{residual}(i)}}{(k+1)\text{MS}_{\text{residual}}}.$$

Since the values of $\text{MS}_{\text{residual}(i)}$ and $\text{MS}_{\text{residual}}$ will be very similar except in those small data sets in which case i has an extreme discrepancy relative to the other cases, this relationship can typically be approximated as

$$\text{Cook's } D_i \approx \frac{DFFITS_i^2}{(k+1)}.$$

Guidelines for Identifying Cases with High Global Influence

$DFFITS_i$ and Cook's D_i can be viewed as interchangeable statistics. Either measure can be used to provide information about the global influence of case i. Once again, one good strategy in small to moderate sized samples is to use an index plot—either $DFFITS_i$ or Cook's D_i is plotted against case number. The analyst identifies a very small number of cases that have the most extreme values as being potentially influential. Those cases that have large gaps in the value of $DFFITS_i$ or Cook's D_i relative to other cases deserve particular scrutiny.

Alternatively, rule of thumb cutoffs may be used. For $DFFITS_i$ a conventional cutoff is that cases with magnitudes (ignoring sign) of $DFFITS_i > 1$ in small or medium sized data sets or $> 2\sqrt{(k+1)/n}$ in large data sets be flagged as potentially influential observations. For Cook's D_i a value of 1.0 or the critical value of the F distribution at $\alpha = .50$ with $df = (k+1, n-k-1)$ is used. For example, in the present case with 1 IV, if Cook's D_i exceeded $F(2, 13) = 0.73$, the 50th percentile of the F distribution, the case would be flagged as influential. We will provide further discussion of guidelines following the next section.

A Measure of Influence on a Specific Regression Coefficient

$DFBETAS_{ij}$ is a second type of influence statistic that is very important when the researcher's interest focuses on specific regression coefficients within the equation. Once again, it is a deletion statistic that compares regression coefficients when case i is included versus not included in the sample.

---

[6]When cases fall at the centroid, they can still affect the regression intercept.

To provide a simple illustration of when $DFBETAS_{ij}$ would provide a useful measure, suppose a researcher is interested in the relationship between IQ and children's school performance. This researcher might also include parent's income in the regression equation, performance $= B_1 Income + B_2 IQ + B_0$. Here the researcher's interest is not in parents' income per se, but to control for the effects of parents' income in understanding the relationship between IQ and performance. $DFBETAS_{ij}$ provides information about the effect of case i on the specific regression coefficient(s) of interest, here B_2, that is, $DFBETAS_{i2}$.

$DFBETAS_{ij}$ for case i is defined for regression coefficient B_j as follows:

(10.3.9)
$$DFBETAS_{ij} = \frac{B_j - B_{j(i)}}{SE_{B_{j(i)}}}.$$

We see in this equation that the numerator is the difference between the B_j calculated with all cases in the data set and the $B_{j(i)}$ calculated after case i is deleted. The denominator is the SE of $B_{j(i)}$, calculated after case i is deleted. The calculation of the standard error is complex when there is more than one predictor,[7] but this calculation is performed by standard statistical packages. The division serves to standardize $DFBETAS_{ij}$, facilitating a common interpretation of the influence of case i across each of the regression coefficients. Each *case* will have $(k+1)$ $DFBETAS_{ij}$ associated with it, one corresponding to each of the regression coefficients in the equation including the intercept.

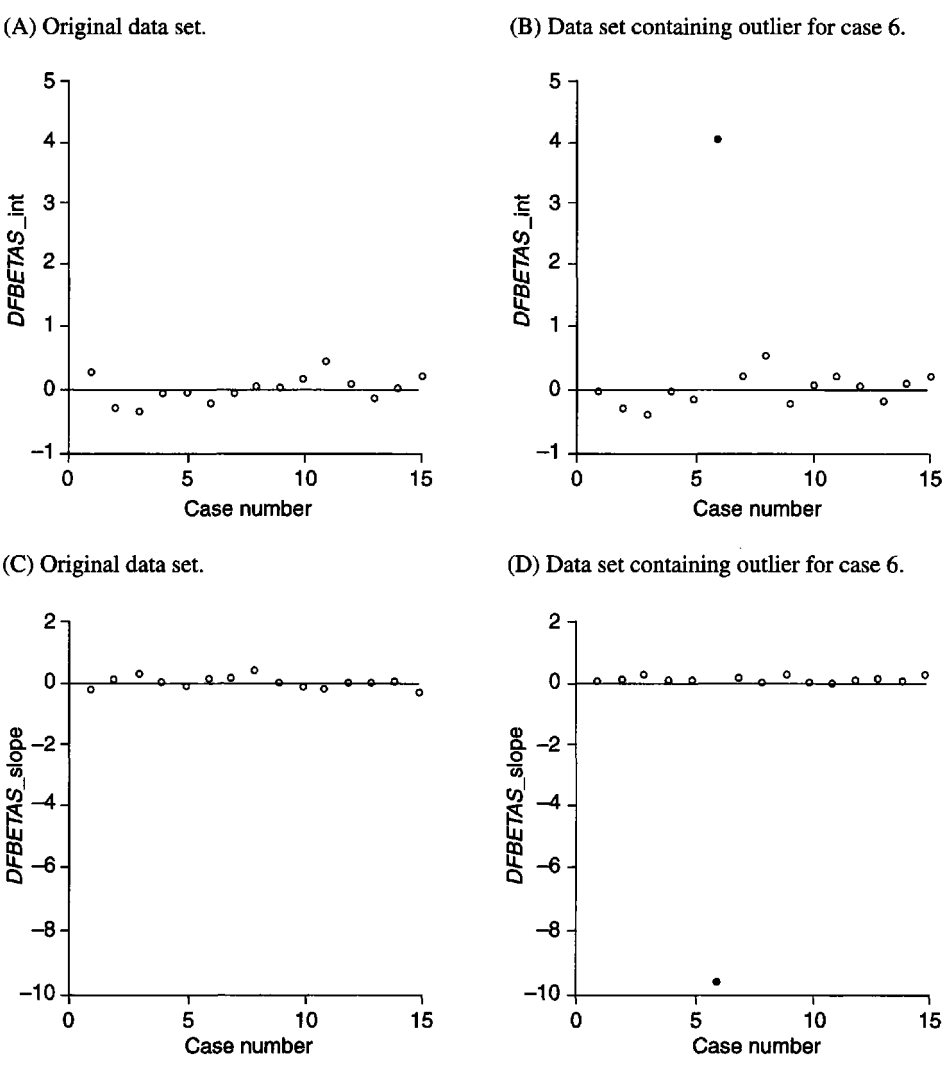
To illustrate the interpretation of $DFBETAS_{ij}$, consider once again the data set containing the outlier presented in Table 10.2.1B. For case 6, which is the outlier, $DFBETAS_{ij} = 4.05$ for the intercept B_0, and $DFBETAS_{ij} = -9.75$ for the slope B_1. The sign of $DFBETAS_{ij}$ indicates whether the *inclusion* of case i leads to an increase or decrease in the corresponding regression coefficient. For case 6, we see that its inclusion leads to an increase in B_0, but a decrease in B_1. The magnitude of $DFBETAS_{ij}$ describes the magnitude of the change with higher values indicating greater change. Figure 10.3.6(A) and (B) provide an index plot of $DFBETAS_{ij}$ for the intercept, B_0. Figure 10.3.6(A) displays these values based on the original data for case 6 $(X = 6)$; Fig. 10.3.6(B) displays the values for the data including the outlier for case 6 $(X = 60)$. Figure 10.3.6(C) and (D) present index plots of $DFBETAS_{ij}$ for the slope. Figure 10.3.6(C) displays the values based on the original data; Fig. 10.3.6(D) displays the values for the data including the outlier for case 6. Case 6 is highlighted in each panel. As can be seen, no extreme values of $DFBETAS_{ij}$ are observed for the original data in either Fig. 10.3.6(A) for the intercept or Fig. 10.3.6(C) for the slope. In contrast, in Fig. 10.3.6(B) and Fig. 10.3.6(D) case 6 is far from the values of the other cases for both the intercept and the slope.

Guidelines for Identifying Cases With High Influence on Specific Regression Coefficients

For small to moderate sized samples, it is useful to construct a separate index plot for each regression coefficient of $DFBETAS_{ij}$ against the case number. Any cases that have large values of $DFBETAS_{ij}$ relative to the remaining cases have high influence on the regression coefficient B_j. Only those few cases with the most extreme values are studied. Histograms with a large number of bins or stem and leaf displays can be used with large samples.

For researchers who prefer rule of thumb guidelines, cases having $DFBETAS_{ij} > \pm 1$ for small or moderate sized data sets or $DFBETAS_{ij} > \pm 2/\sqrt{n}$ for large data sets are considered

[7] The standard error does not have a simple algebraic expression when there is more than one IV. The matrix formula is $SE_{B_{j(i)}} = \sqrt{\mathrm{MS}_{\mathrm{residual}(i)}(\mathbf{X}'\mathbf{X})^{-1}_{jj}}$. For B_j, the term in the jth row and jth column (on the diagonal) of the inverse of the $(\mathbf{X}'\mathbf{X})$ matrix is used as the value of $(\mathbf{X}'\mathbf{X})^{-1}_{jj}$.



Note: In each panel, values of $DFBETAS_{ij}$ are presented on the ordinate. Case numbers are presented on the abscissa. The values of $DFBETAS_{ij}$ presented in Figs. 10.3.6(A) and (B) are for the intercept, B_0. The values of $DFBETAS_{ij}$ presented in Figs. 10.3.6(C) and (D) are for the slope, B_1. The highlighted points correspond to case 6, the outlier. The horizontal line represents a value of 0 for $DFBETAS_{ij}$ in each panel.

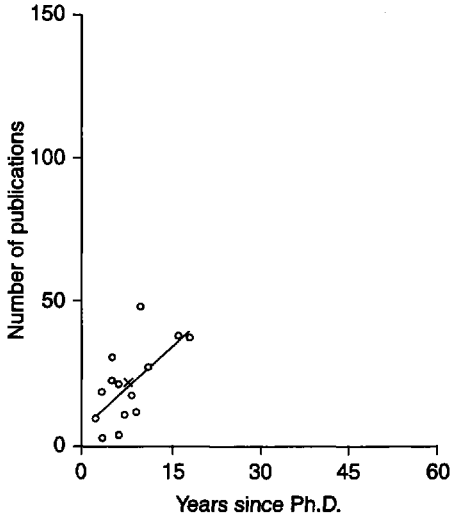
FIGURE 10.3.6 (A), (B): Index plot of $DFBETAS_{ij}$ vs. case number: intercept. (C), (D): Index plot of $DFBETAS_{ij}$ vs. case number: slope.

to be influential. In the present illustration involving a small sample ($n = 15$), the value of $DFBETAS_{ij}$ for the intercept B_0 and the slope B_1 both far exceed the rule of thumb cutoff of 1 for our outlying case 6.

10.3.4 Location of Outlying Points and Diagnostic Statistics

In the example using case 6 (with $X = 60$) and the outlying point that we have used throughout this section, the measures of leverage, discrepancy, and influence for case 6 were all extreme in value. However, leverage and discrepancy measure two distinct properties of outliers; they

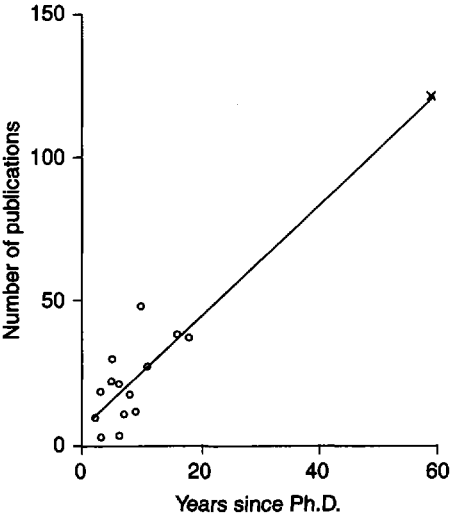
(A) At the mean of X, mean of Y.



$\hat{Y} = 1.92X + 6.00.$
$R^2 = .43.$
For case 6 (added point):

h_{ii}	$= 0.067.$
t_i	$= 0.00.$
$DFFITS_i$	$= 0.00.$
$DFBETAS_{ij}$ for B_0	$= 0.00.$
$DFBETAS_{ij}$ for B_1	$= 0.00.$

(B) At extreme X, extreme Y (on original regression line).



$\hat{Y} = 1.92X + 6.00.$
$R^2 = .88.$
For case 6 (added point):

h_{ii}	$= 0.90.$
t_i	$= 0.00.$
$DFFITS_i$	$= 0.00.$
$DFBETAS_{ij}$ for B_0	$= 0.00.$
$DFBETAS_{ij}$ for B_1	$= 0.00.$

FIGURE 10.3.7 Effect of adding a single data point at various locations.

are not necessarily related. Recall also that influence can be conceptually thought of as the product of leverage and discrepancy (see Eq. 10.3.7). Cases with high values of influence will typically have at least moderately high values of both leverage and discrepancy.

To illustrate these ideas, we will use the 14 cases presented in Fig. 10.2.1(C). These are the 14 cases included in the original data set with case 6 deleted. In Fig. 10.3.7 we try adding different values of a single case to this data set and observe what happens to the regression equation and the diagnostic statistics. For the 14 original cases, the regression equation is $\hat{Y} = 1.92X + 6.00, R^2 = .43, M_X = 7.79, M_Y = 20.93.$

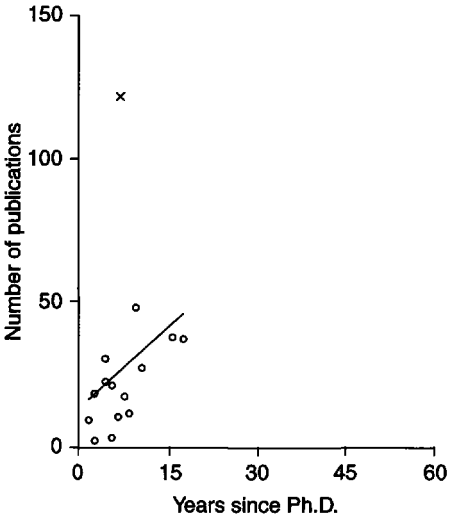
Figure 10.3.7(A)

In Fig. 10.3.7(A), the new point has been added at the mean of X and mean of Y (case 6, $X = 7.79, Y = 20.93$). The new regression equation is identical to the original regression equation based on the 14 cases. For the new case (case 6) in the regression equation based on 15 cases, $h_{ii} = 0.0667$, which is equal to the minimum leverage value of $1/n$; t_i (the externally studentized residual, the measure of discrepancy) $= 0$; and Cook's D_i (the measure of overall influence) $= 0$.

Figure 10.3.7(B)

Figure 10.3.7(B) adds the new point at an extreme value of X and the corresponding value of Y that falls exactly on the original regression line. The point is added at $X = 60, Y = 121.07$.

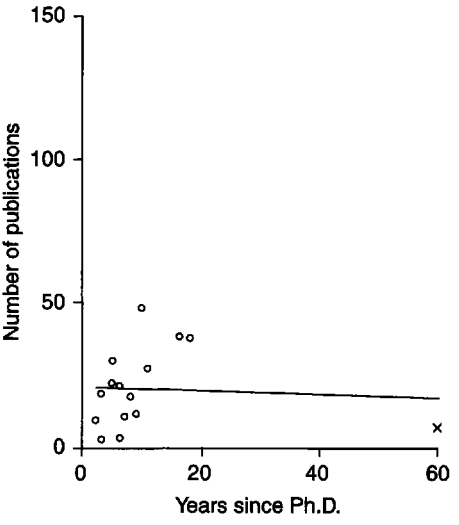
(C) At the mean of X, extreme value of Y.



$\hat{Y} = 1.92X + 12.66.$
$R^2 = .09.$
For case 6 (added point):

h_{ii}	$= 0.067.$
t_i	$= 8.96.$
$DFFITS_i$	$= 2.39.$
$DFBETAS_{ij}$ for B_0	$= 1.18.$
$DFBETAS_{ij}$ for B_1	$= 0.00.$

(D) Extreme value of X, extreme value of Y (not on original regression line).



$\hat{Y} = -0.06X + 20.61.$
$R^2 = .004.$
For case 6 (added point):

h_{ii}	$= 0.90.$
t_i	$= -3.29.$
$DFFITS_i$	$= -10.13.$
$DFBETAS_{ij}$ for B_0	$= 4.05.$
$DFBETAS_{ij}$ for B_1	$= -9.75.$

Note: Years since Ph.D. is shown on the abscissa. Number of publications is shown on the ordinate. The added case (case 6) in each Part is denoted by ×. The best fitting linear regression line is superimposed in each plot. The data presented in Fig. 10.3.7(D) were previously displayed in Fig. 10.2.1(B). The results of the regression analysis and diagnostic statistics are also presented for each panel.

FIGURE 10.3.7 Continued.

To calculate the value of Y, we substituted $X = 60$ into the original regression equation: $\hat{Y} = 1.92(60) + 6.00 = 121.07$. The new regression equation is $\hat{Y} = 1.92X + 6.00, R^2 = .88, M_X = 11.27, M_Y = 27.61$. Note that the new regression equation is identical to the original. The new R^2 has doubled—from .43 based on the original 14 cases to .88 in the new data set. This result indicates how selecting cases with extreme values on X can potentially increase the R^2 and the power of significance tests of the regression coefficients.[8] Extreme cases located on the regression line stabilize the regression line and decrease the *SE*s of both the slope and intercept. For the new case (case 6) in the regression equation based on 15 cases, $h_{ii} = .90, t_i = 0$, and $DFFITS_i = 0$. Only the measure of leverage reflects this outlying case.

Figure 10.3.7(C)

Figure 10.3.7(C) illustrates what happens when the new point is added at an extreme value of Y, but the value of X is at the mean ($M_X = 7.79$) of the original cases. For purposes of

[8] It may be useful to refer back to Section 2.11, where the discussion of the impact of the range of values on r is discussed.

comparison, we add the point at the same Y value ($Y = 121.07$) as in Fig. 10.3.7(B). As shown in Fig. 10.3.7(C), the slope of the new B_1 is identical to that of the original data set, but B_0 has increased from 6.00 to 12.66. The addition of a point far from the original regression line leads to a large decrease in R^2—from .43 for the original sample of 14 cases to .09 with the case added. For case 6 in this new regression equation based on 15 cases, $h_{ii} = .0667$, $t_i = 8.96$, and $DFFITS_i = 2.39$. Leverage is at the minimum possible value, but the externally studentized residual is very large, and the measure of global influence, $DFFITS_i$, has a large value. To understand the effect of case 6 on the global measure of influence, it is useful to consider the measures of specific influence, the values of $DFBETAS_{ij}$ for the slope and intercept. For case 6, the value of $DFBETAS_{ij}$ for B_0 is high, 1.18. In contrast, the value of $DFBETAS_{ij}$ for B_1 is 0. These values of $DFBETAS_{ij}$ indicate that all of the change was in the intercept; the slope has not changed.

Figure 10.3.7(D)

Finally, Fig. 10.3.7(D) reprises the example used throughout this section—an outlier is added that is extreme on both X and Y ($X = 60; Y = 6$). As we saw earlier, the regression equation changes dramatically from $\hat{Y} = 1.92X + 6.00$ to $\hat{Y} = -0.06X + 20.61$ with the addition of the outlier, case 6. Case 6 has high leverage ($h_{ii} = 0.90$), high discrepancy ($t_i = -3.29$), and high measures of both global ($DFFITS_i = -10.13$) and specific influence ($DFBETAS_{ij} = 4.05$ for intercept; $DFBETAS_{ij} = -9.75$ for slope).

10.3.5 Summary and Suggestions

In summary, each of the diagnostic statistics provides different information about the effect of an outlier on the regression equation. Leverage (h_{ii}) informs us about how far the point is from the centroid of the predictor space, and discrepancy (t_i = externally studentized residual) informs us about how far the point is from the regression line with case i deleted. The two measures of global influence, $DFFITS_i$ and Cook's D_i, provide interchangeable information about the overall influence of the single case on the regression equation, whereas $DFBETAS_{ij}$ informs us about how the single case i affects each regression coefficient. Each of these sources of information is useful in studying the effects of outliers. Table 10.3.1 summarizes the diagnostic statistics and provides rule of thumb cutoff values.

We present here some suggestions for looking at diagnostic statistics, deferring our consideration of possible remedial actions until the next section.

1. When the data are initially received, it is very useful to examine histograms with a large number of bins (or boxplots) of each variable to look for univariate outliers. Plots of leverage values can help identify multivariate outliers in initial data screening. For example, calculating leverage values for each participant based on the responses to each item of a 20-item scale can help identify any participants with unusual response patterns that may be problematic. Recall that leverage is based only on the IVs.[9] These statistics can be useful as a final step in the initial data checking and cleaning.

2. For any data analysis that may be reported, it is useful to examine diagnostic statistics. The extent of scrutiny of these statistics depends on the nature of the study. If the study is one of a series of replications, inspection of graphical displays for any obvious outliers is normally sufficient. Consistency of the findings across replications provides assurance that the presence

[9] Standard regression programs require that a regression equation be specified to calculate leverage. We recommend that a regression analysis be specified that includes all IVs of potential interest and an arbitrary numeric variable that is *complete* for all cases (e.g., case number) as the DV. Leverage is not affected by the DV that is chosen.

TABLE 10.3.1
Summary of Regression Diagnostics for Individual Cases

Diagnostic index	Measures	Proposed cutoff	Expected cases identified
Leverage (h_{ii})	Extremity on IVs	$2(k+1)/n$ for large n $3(k+1)/n$ for small n	5%
Centered leverage (h_{ii}^*)	Extremity on IVs	$2k/n$ for large n $3k/n$ for small n	5%
Externally studentized residuals (t_i)	Discrepancy of Y_i from regression line excluding the case	± 3.0 or ± 4.0 for large n ± 2.0 for small n	0.3%, 0.01% 5%
DFFITS	Influence: change in predicted Y if case omitted from estimate	$\pm 2\sqrt{\frac{k+1}{n}}$ for large n ± 1.0 for small n	5% —
Cook's D	Influence measured as aggregate change in set of B_is if case omitted from estimate	1.0 or F distribution value for $\alpha = .50$	—
DFBETAS	Influence measured as change in a specific B_i if case omitted from estimate	$\pm 2/\sqrt{n}$ for large n ± 1.0 for small n	5% —

Note: Some proposed minimum cutoffs for diagnostic statistics. Only a few of the most extreme cases that exceed minimum cutoffs merit examination.

of an outlier is not responsible for the results. On the other hand, if the data set is unique and unlikely to be replicated (e.g., a study of 40 individuals with a rare medical disorder), very careful scrutiny of the data is in order.

3. In large samples, visual inspection of index plots becomes difficult. Analysts may use boxplots or histograms with a large number of bins to identify outlying values, and then identify these cases in the data set. Alternatively, analysts may save the values of the diagnostic statistics, order them from lowest to highest, and plot only the highest values (e.g., top 50) on an index plot. Ideally, the most extreme values relative to the remainder should be apparent.

4. If the researchers' interest is in overall prediction or they have not made any prediction about specific regression coefficients, we encourage examination of influence statistics focused on $DFFITS_i$ or Cook's D_i. If the researchers have made a priori predictions about specific regression coefficients, then we encourage examination of the associated $DFBETAS_{ij}$ regardless of whether the measure of global influence is extreme. If the researchers find an unpredicted new result, we also encourage examination of the associated $DFBETAS_{ij}$ (and ideally replication in a new sample[10]) prior to reporting the result. This helps assure that "exciting new results" are indeed potentially exciting and not merely produced by an outlier. Note that measures of influence are associated with a specific regression equation; the values of these diagnostic statistics will change if the regression equation is modified.

[10]Maxwell (2000) provides a striking demonstration of the importance of replication in the interpretation of unpredicted findings in multiple regression.

5. In regression equations including power polynomial (e.g., X^2) or interaction (e.g., XZ) terms, outlying points can have profound effects on measures of both global and specific influence even in moderate and large samples. For example, Pillow, West, and Reich (1991) found that a single extreme outlier in a sample of over 300 cases produced an inexplicable three-way interaction and that the originally predicted results were obtained when this case was deleted. Even though a case may be only moderately extreme on X and on Z separately, the product of these values may yield an extreme point that has a substantial influence on the results of the regression analysis. Such outliers can create spurious effects or mask a priori predicted effects. Very careful screening for outliers is encouraged in such regression equations.

6. Cases that do not have high values of influence but that are extreme in terms of the externally standardized residual do not greatly alter the estimates of the regression coefficients (except for B_0). Nonetheless, they do affect the standard errors and hence the power of the statistical tests. Measures of discrepancy are associated with a specific regression equation; these diagnostic statistics will change if the regression equation is modified.

7. The values of the diagnostic statistics change whenever a case is removed. If a serious outlier is detected and removed, the effects of its removal should be studied. The diagnostic statistics for the new data set should be recomputed before any additional cases are considered for removal. If the removal of outlying cases continues to produce new outlying cases after a few repetitions of this process, other strategies of addressing the outlier (e.g., transformation) should be sought.

Other sources (Bollen & Jackman, 1990; Chatterjee & Hadi, 1988) present a fuller discussion of cutoff values for the diagnostic statistics and illustrations of the use of diagnostic statistics with real data sets.

10.4 SOURCES OF OUTLIERS AND POSSIBLE REMEDIAL ACTIONS

When outliers are discovered in data, the researcher needs to decide what remedial actions, if any, should be undertaken. This decision needs to be based on careful detective work using clues provided by the regression diagnostic statistics to try to understand the source of the outliers and their influence on the results of the regression analysis. However, good detective work also depends on a clear understanding of the substantive problem that is being studied, the methods through which the data were collected, the nature of the sample, and the population to which generalization is sought. In some cases, a clear understanding of the source of the outliers may not emerge even with the best detective work. In such cases, the choice of the optimal remedial action will be associated with considerable uncertainty. More than one remedial action may be investigated, or researchers may choose to employ the remedial action that is most commonly used in their substantive area.

10.4.1 Sources of Outliers

Outliers can arise from many sources. To help readers in thinking about this problem, we have grouped sources of outliers into two general classes: contaminated observations and rare cases.

Contaminated Observations

Outliers may occur because the data have been contaminated in some way. Here we present several of the possible sources of contamination in the behavioral sciences together with examples. Contaminated observations can and should be minimized by careful research procedure

and data preparation. Nonetheless, contaminated observations will occur even for the most careful researchers.

1. *Error of execution of the research procedure.* An interviewer may misread some of the questions; an experimenter may deliver the wrong or an incomplete treatment.

2. *Inaccurate measurement of the dependent measure.* Equipment may fail so that measurement of the DV (e.g., response time) is not accurately recorded.

3. *Errors in recording or keying the data.* An interviewer may write down the participant's response incorrectly, or the data may not be keyed into the computer properly.

4. *Errors in calculation of the measures.* The researcher may incorrectly count up the number of responses or make a mistake in the calculation of a measure (e.g., percentage of correct responses).

5. *Nonattentive participants.* In certain cases, participants may be fatigued, ill, or drunk, and be unable to respond in their typical manner to the experimental materials.

Each of the diagnostic statistics (leverage, discrepancy, and influence) can potentially aid in detecting contaminated data. Whenever researchers detect outliers, they should first attempt to rule out the possibility that the outliers represent contaminated data. Data and calculations should be checked for accuracy; research notes should be checked for procedural anomalies that may explain the outlier.[11] If it can be verified that the outliers represent contaminated data, these data points should *not* be included in the data analysis. The researcher should replace the contaminated data with the correct data for the case if possible or delete the contaminated case from the data set. In other situations, it may be possible to make a second observation on the case and to replace the outlying value in the data set. To illustrate, in our example of faculty publications, it may be possible to check personnel records or to reinterview the faculty member in question (case 6) to determine the correct value for time since Ph.D. The corrected value should always be used in the data set.

Rare Cases

For other cases, the outlying observations may be correct (or alternatively, not show any detectable evidence of contamination). The outlying case may represent a valid, but extremely rare observation in the population. For example, imagine researchers are conducting a study of the relationship between year in college (freshman = 1; senior = 4) and sexual attitudes. Suppose that the sample contains a single 12-year-old freshman male. Although very rare, such individuals do exist in the U.S. college population. The results of the regression analysis could *potentially* be affected by this one outlying case if his sexual attitudes differed greatly from those of his classmates.

When outliers having high influence are detected that are not contaminants, how they should be treated in the data analysis can often be a serious issue that can be difficult to resolve. In general, there is a tension between two scientific goals. On one side, eliminating or minimizing the effect of the rare case can often lead to a regression equation that provides a more accurate description of the X-Y relationship in the population of interest. On the other side, the outlier may represent a subtle signal of an important phenomenon or the inadequacy of the specific regression model that was tested. Research attempting to understand rare cases has sometimes

[11] In complex experiments or interviews, many observations (including apparently valid observations) may be associated with anomalies in the procedure. The researcher should examine outliers as well as a random sample of other observations to be sure the two classes of observations can be distinguished on the basis of procedural problems.

led to important lines of scientific research. The discovery of the Antarctic ozone hole and the discovery of penicillin both depended on understanding the source of rare cases.

In some research situations, the decision may be relatively straightforward. The rare case may be thought of as a different kind of contaminant—the participant is from a different population than the one of interest. In such cases, researchers may conclude that they wish to restrict their generalization to a specific population and exclude the rare case(s) from the analysis. Returning to our example on college sexual attitudes, the researcher may decide to exclude all students from the analysis who began college at less than a minimum age (e.g., 16 years old). The result is that the regression equation will better characterize a population of normatively aged college students. The high potential for a small number of young (prepuberty) students to influence the results of the regression analysis if they have different sexual attitudes is eliminated. Now, however, generalization of the results is limited to students who were at least 16 years old when they began college.

Unfortunately, in practice the source of any rare cases will be difficult (or impossible) to determine. Rare cases can arise from many sources.

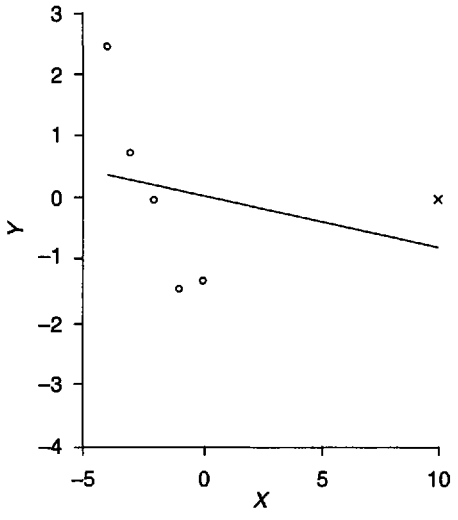
1. There may be an *undetected* (and perhaps unknowable) contaminant in the data. For example, an experimenter may fail to observe and record a procedural error in an experiment.
2. One or more of the predictor variables may have an unusual distribution that produces extreme values. For example, even though intelligence is usually thought of as having a normal distribution, there are a substantially larger than expected number of cases in the population with very low intelligence because of specific constitutional insults or anomalies. In some samples such cases may cause potential distortions of the relationship between variables.
3. The dependent variable may have properties that lead to a potential for occasional large residuals. For example, the distribution of the number of absences from work during a month in a group of typically very healthy workers may include rare high outliers associated with workers who experienced major illnesses such as a heart attack.
4. The regression model being studied may be incorrect. This issue is illustrated in Fig. 10.4.1 through a series of scatterplots based on a small artificial data set developed by Huber (1981). Note that in Fig. 10.4.1(A) and (B) there is one point marked by × which is an outlier. Figure 10.4.1(A) depicts the poor fit of a linear regression to the full set of 6 data points. Figure 10.4.1(B) illustrates that a regression equation including a quadratic term may fit the data extremely well and the case marked by an × is no longer an outlier.


CH10EX02

Figure 10.4.1 nicely illustrates the dilemma created by outliers. In reality, researchers often do not know for sure that the point marked × has not in fact been contaminated by unknown influences. Nor will they always have prior theory or empirical work that would lead them to expect this form of curvilinear relationship between X and Y. Figure 10.4.1(C) illustrates what happens if we delete the outlier: The linear regression fits well, indicating a strong negative relationship between X and Y, $B_1 = -0.98, t(3) = -5.57, p = .01$. The proposal of a curvilinear relationship is based only on a single outlying data point; this is generally a very risky practice, as the odds of being able to replicate this unexpected curvilinear relationship in another sample are low. Indeed, $DFBETAS_{ij}$ for case 6 for the nonlinear (quadratic) term in the regression equation is -14.74, indicating the extraordinary degree to which the nonlinear function is dependent on the inclusion of this one case. At the same time, a decision to exclude the outlier in Fig. 10.4.1(C) may mean the researcher has discarded the basis for an important potential insight. Collecting a larger sample in which there are a sizeable number of cases at the higher values of X would enable the researcher to resolve this dilemma.

Consider yet again our earlier example, in which very young students were excluded from the analysis to provide a better characterization of the relationship between year in college and sexual attitudes in "typical" college students. The distinctly different college experience

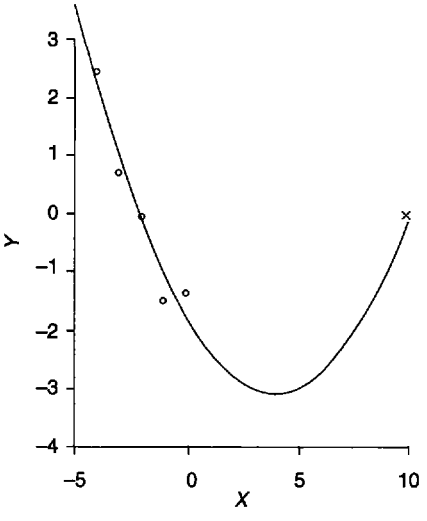
(A) Fit of linear model.



$\hat{Y} = -0.08X + 0.07.$

$R^2 = .08.$

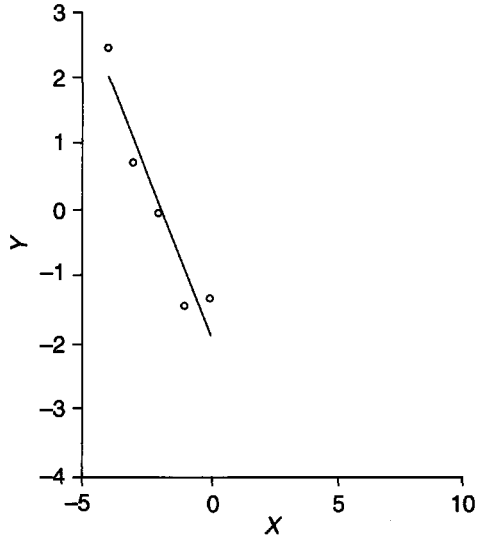
(B) Fit of quadratic model.



$\hat{Y} = -0.66X + 0.08X^2 - 1.74.$

$R^2 = .95.$

(C) Fit of linear model with outlier deleted.



$\hat{Y} = -0.98X - 1.87.$

$R^2 = .91.$

Note: In Panels (A) and (B), the outlier is denoted by ×.

FIGURE 10.4.1 Scatterplot of Huber's (1981) example.

of very young students, especially with regard to their social and sexual attitudes and relationships, may be a very interesting focus of study in its own right. The outlying cases may be providing an important signal about differences between young and normatively aged college students. Other researchers might note the information about such outliers in published reports, perceive this possibility, and design a study in which a large number of very young students were included in the sample. More precise estimates of the effects of number of years of college attendance on sexual attitudes for both the older, normative population and this younger population could be obtained.

Given this tension between accurately characterizing relationships for the nonoutlying participants versus missing important information provided by the rare case(s), it is important to provide information about outliers in published research reports. Kruskal (1960) has emphatically argued that no matter the reason, apparent outliers should be reported.[12]

> I suggest that it is of great importance to preach the doctrine that apparent outliers should *always* be reported, even when one feels that their causes are known or rejects them for whatever good rule or reason. The immediate pressures of practical statistical analysis are almost uniformly in the direction of suppressing announcements that do not fit the pattern; we must maintain a strong sea-wall against these pressures (p. 158, italics in original).

10.4.2 Remedial Actions

As we have discussed, addressing data that are known to be contaminated is easy. The contaminated data point(s) are simply corrected, deleted, or replaced as is appropriate. In contrast, three different general approaches may be taken with outliers that represent rare cases. First, the data may be analyzed with the outliers deleted. Second, the regression model may be revised, adding terms that may account for outliers, or the data may be transformed so that the outliers are no longer present. Or third, alternative robust regression methods (to be described later) may be used that attempt to minimize the influence of outliers on the results of the regression equation.

Deletion of Outliers

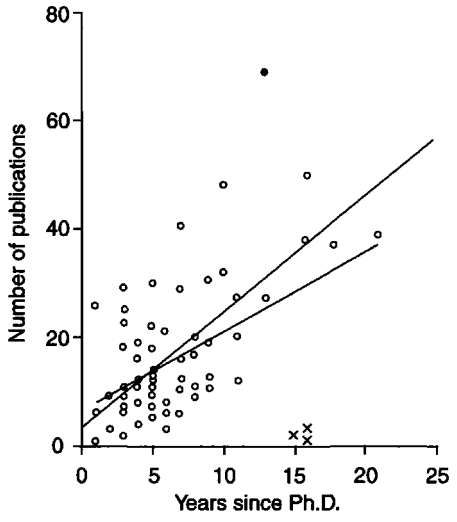
The classic method of addressing outliers is to delete them and to reanalyze the remaining data (e.g., Chatterjee & Wiseman, 1983; Stevens, 1984). This is the simplest method and in many (but certainly not all) cases, it will provide estimates of the regression coefficients that are very similar to those produced by more complex robust regression procedures. Researchers will typically base their conclusions on the regression equation with outlying cases deleted. However, the nature of the outliers and the results of the original regression analysis with all cases included should normally be reported, at least in footnotes.

There are several potential problems with this approach. First, as noted in Section 10.3.5, analysis of the diagnostic statistics for the new data set (now with the original outliers deleted) may yield still other cases with extreme values on the diagnostic statistics. Second, the specific cases chosen for deletion will often depend on the subjective judgment of a particular researcher. Other researchers analyzing the same data set may come to different conclusions. Third, although diagnostic statistics do a good job of detecting single outliers, the presence of several outliers in close proximity (known as a "clump") can sometimes mask the problem.

Figure 10.4.2 illustrates this third problem with a scatterplot of years since Ph.D. versus publications. The 62 cases we considered in Chapter 3 are plotted as circles (one case to be discussed later is plotted as a solid circle). We add to this data set a clump of three outliers


CH10EX3

[12]Journal space limitations preclude the provision of extensive information about outliers. Succinct information about the number and type of outliers and their effects on the regression analysis can typically be reported.



Note: The 62 data points denoted by circles (o) are the original 62 cases presented in Table 3.5.1. Three outliers denoted by × have been added to form a clump at the lower right of the plot. One data point with high discrepancy from the original data set has been darkened (•). The longer line that touches the *y* axis is the regression line for the original 62 cases. The shorter line is the regression line for all 65 cases including the clump of 3 outliers.

FIGURE 10.4.2 Illustration of a clump of outliers: Scatterplot of years since Ph.D. vs number of publications.

having 15–16 years since the Ph.D. and few publications; these cases are denoted with an ×. The first regression equation for the 62 original cases is $\hat{Y} = 2.13X + 3.72$, $R^2 = .42$. The second regression equation for the 65 cases (including the clump of 3 outliers) is $\hat{Y} = 1.46X + 6.88$, $R^2 = .23$. These regression lines are illustrated in Fig. 10.4.2.

The values of the diagnostic statistics corresponding to these three cases (case numbers 63, 64, 65 in data file) for the second regression equation would *not* be considered to be unusually extreme by most guidelines. Leverage values for the three cases are moderate ($h_{ii} = .06, .07, .07$, respectively, compared to a rule of thumb *minimum* cutoff value of $h_{ii} = .09$ for small samples (see Table 10.3.1). $DFFITS_i = -.58, -.71, -.66$; $DFBETAS_{ij} = .27, .36, .33$ for the intercept and $DFBETAS_{ij} = -.50, -.63, -.59$ for the slope, respectively. These values are not large relative to rule of thumb cutoffs of ± 1.0 that have been proposed for small samples. The values of the externally studentized residuals for the clump of outliers, $t_i = -2.29, -2.54, -2.34$, respectively, might draw some attention, but would likely be overshadowed by another data point from the original sample ($X = 13, Y = 69$; marked by •) that has a much higher externally studentized residual, $t_i = 3.89$. What makes the three outliers stand out is that they form a visually distinct clump with similar values on X and Y. In situations with multiple predictors, clumps of outliers can not always be easily detected by visual inspection of scatterplots.

Because the standard diagnostic statistics perform poorly in finding clumps of outliers, other methods need to be used. Promising methods of detecting clumps of outliers in complex data sets have recently been developed (e.g., Hadi, 1994; Hadi & Simonoff, 1993), but to

date they have not yet become available in common statistical packages. Alternatively, robust regression approaches (to be discussed later) minimize the weight given to outlying cases in the calculation of regression coefficients. These approaches will produce improved estimates of the regression coefficients and the standard errors, even when there are clumps of outliers.

Respecification and Transformation

There is a second and often overlooked consideration when outliers are discovered. Outliers may result from misspecification of the regression model rather than any problems with the data. If the appropriate regression model is specified, the originally outlying cases may be well fit by the new model. The data set presented in Fig. 10.4.1 illustrated this idea: The original equation illustrated in Fig. 10.4.1(A) does not fit the data well and produces an extreme outlier with high influence. However, the nonlinear model illustrated in Fig. 10.4.1 fits all of the data very well—there is a high R^2 and all of the residuals are small. Consequently, it is very important to use the approaches presented in Section 4.4 to look for evidence of model misspecification before outliers are deleted. Models that specify curvilinear effects (see Chapter 6) or interactions between IVs (see Chapter 7) can sometimes address the problem of outliers.

Alternatively, a linear regression model may be used, but the individual variables may be transformed so that the data are more appropriate for a linear regression equation. Recall from Chapter 6 that transformation uses a mathematical expression to change the value of the IV, DV, or both for each case. For example, an outcome variable with a few cases with high values (i.e., a long upper tail) is sometimes more appropriately analyzed when the value of Y for each case replaced by a new value equal to the logarithm of the original value, i.e., $Y^*_{\text{new}} = \log(Y_{\text{original}})$. Chapter 6 presented a thorough discussion of transformations.

Robust Approaches

Robust approaches refer to a family of techniques that use alternatives to the ordinary least squares (OLS) method to estimate the regression coefficients. Robust approaches can be thought of as a kind of insurance policy (Anscombe, 1960). Ideally, robust approaches should perform better than OLS when there are outliers or the residuals have a non-normal distribution with many extreme residuals in the tails. And when the data are well behaved, robust approaches should perform *almost* as well as OLS. We should only pay a small cost in terms of lower statistical power when there are no outliers and the assumptions of OLS regression are fully met.

In Section 10.3 we showed that in OLS estimation the values of $B_0, B_1, \ldots, B_k$ are chosen so as to minimize the sum of the squared residuals. When an observed Y_i is far from the regression line of the other cases, it can strongly affect the values of the Bs that are chosen. As we have shown, cases that have both high leverage and high discrepancy (high influence) can greatly alter the values of the regression coefficients that are chosen. Under these circumstances, robust alternatives to OLS may be considered. Examples of four approaches to robust estimation are briefly presented next.

One alternative estimator is *least absolute deviation* (LAD; also called L^1). This method chooses values of the regression coefficients $B_0, \ldots, B_k$ so as to minimize the value of $\Sigma|Y - \hat{Y}|$, where $|Y - \hat{Y}|$ refers to the magnitude of the difference, ignoring its sign (absolute value). Because the difference between Y_i and $\hat{Y}_i$ is not squared, this estimator may provide better results than OLS when there are cases with high discrepancy (externally studentized residuals). However, the LAD estimator is potentially very sensitive to cases that are also extreme on X(high leverage). A single outlying case with high influence can have a greater impact on the regression results when LAD is used rather than OLS.

A second approach is the *least trimmed squares* (LTS) estimator. In LTS the squared residuals for each of the n cases $e_1^2, e_2^2, \ldots, e_n^2$ are ordered from lowest to highest. The analyst chooses a proportion of the cases (e.g., proportion = .25) with the highest value of e^2 to be "trimmed", that is, removed from the analysis producing the regression estimates. LTS chooses values of the regression coefficients $B_0, \ldots, B_k$ so as to minimize the value of $\Sigma_{i=1}^{n'}(Y - \hat{Y})^2$ where n is the number of squared residuals remaining after the largest residuals are trimmed. Rousseeuw, Van Aelst, and Hubert (1999) note that LTS generally performs well, but that it can on rare occasions mislead by providing highly inaccurate estimates when there is a clump of outliers.

A third approach is known as *M-estimation* (Huber, 1981). This approach uses a variant of weighted least squares regression (see Section 4.5.4) in which the function to be minimized is $\Sigma w_i e_i^2$, where w_i is the weight that is given to the ith case. In M-estimation the weight for each case is chosen by how far the residual is from the regression line. Huber suggested that residuals that fall on or near the regression line be given full weight so that $w_i = 1$ and that residuals that fall beyond a threshold be given weights that decrease as $|Y - \hat{Y}|$ becomes larger. Like the LAD estimator, M-estimation will provide better results than OLS when cases with high discrepancy, but it also shares the major liability of LAD—it can produce poor results relative to OLS when cases are also extreme on X, that is, that have high leverage as well as discrepancy (i.e., high influence).

Bounded influence estimators (also called generalized M-estimators or GM estimators) follow the same general logic as M-estimation, except that the weights are chosen based on consideration of both leverage and discrepancy. For example, Welsch (1980) proposed that cases having high values of the $DFFITS_i$ statistic be given less weight. The bounded influence estimators give very good performance in many situations but can provide poor estimates in some cases when there are clumps of outliers or when outliers on Y have low leverage.

In summary, the four estimators considered—LAD, LTS, M-estimates, and bounded influence estimates—represent examples of four of the general approaches to robust estimation that have been proposed. Each of these robust statistics can potentially produce greatly improved estimates relative to OLS when certain patterns of outliers are present in the data; however, rare conditions do exist under which each of the robust estimators can be badly misleading and produce very poor estimates relative to OLS. In addition, OLS will always produce accurate estimates of regression coefficients with the smallest possible standard errors when its assumptions are met.

There are currently few published applications of robust statistics in the behavioral sciences. Perhaps the primary reason for the lack of use to date is that many of the common statistical packages have been slow to incorporate robust regression procedures. Currently, many statistical packages do not include robust estimators, include them in another specialized module (e.g., SAS NLIN), or include only the earlier developed procedures such as LAD and M-estimation. Some of the procedures (e.g., LTS) are computer intensive and may require considerable computer time when applied to large data sets. Alternative procedures described in Staudte and Sheather (1990) should be used for significance testing and construction of confidence intervals for regression coefficients. And robust techniques must be used cautiously because they can hide problems associated with the use of a misspecifed regression model (Cook, Hawkins, & Weisberg, 1992).

Despite some limitations, robust approaches are a very valuable addition to our available tools for multiple regression when there are outliers present in the data. Researchers may usefully compare the results obtained using (a) OLS regression and (b) two robust approaches that are believed to have different strengths and weakness (e.g., LTS; M-estimation). When the different approaches lead to similar conclusions, we gain increased confidence in our results. When the results do not agree, information from these analyses, diagnostic statistics, and

careful examination of scatterplot matrices can often be very helpful in understanding the source of the differences.

We have provided only a brief introduction to the complex topic of robust regression. Readers wishing to use robust regression techniques in their own research should consult recent chapters and texts (e.g., Draper & Smith, 1998, Chapter 25; Ryan, 1997, Chapter 11; and Wilcox, 1997, Chapter 8, for introductions; Rousseeuw, 1998; Rousseeuw & Leroy, 1987; Staudte & Sheather, 1990 for more advanced treatments).

10.5 MULTICOLLINEARITY

We now shift our attention from problems that may arise from specific cases in the data set to problems that may arise from specific IVs. In multiple regression, we assume that each IV can *potentially* add to the prediction of the dependent variable Y. However, as one of the independent variables, X_i, becomes increasingly correlated with the set of other IVs in the regression equation, X_i will have less and less unique information that it can potentially contribute to the prediction of Y. This causes a variety of problems when the multiple correlation of X_i with the set of other predictors, $R_{X_i \cdot X_1 X_2 \ldots (X_i) \ldots X_k}$, becomes very high. The individual regression coefficients can change appreciably in magnitude and even in sign, making such coefficients difficult to interpret. As the predictors become increasingly correlated, the estimate of the individual regression coefficients also becomes more and more unreliable, a problem that is reflected in large standard errors. In the limiting case of *exact collinearity*, in which X_i is perfectly correlated with the other predictor variables (that is, when X_i can be perfectly predicted from the remaining IVs), the individual regression coefficients cannot even be properly computed. Short of exact collinearity, small changes in the data such as adding or deleting a few observations can lead to large changes in the results of the regression analysis. This set of problems that result from high correlations between some of the IVs is known as *multicollinearity*. Multicollinearity depends only on the set of IVs—regardless of the Y that is chosen, the degree of multicollinearity will be the same.

10.5.1 Exact Collinearity

Exact collinearity occurs when one IV has a correlation or multiple correlation of 1.0 with the other IVs. Exact collinearity indicates that a mistake was made in setting up the regression analysis. Consider the regression equation $\hat{Y} = B_1X_1 + B_2X_2 + B_0$. If X_1 and X_2 are the same or if one is a linear transformation of the other as when X_1 and X_2 represent the same variable expressed in different units (e.g., $X_1 =$ person's weight in pounds; $X_2 =$ person's weight in kilograms), neither variable conveys any *unique* information to the prediction of Y. Exact collinearity can also occur for more subtle reasons. For example, consider the regression equation $\hat{Y} = B_1X_1 + B_2X_2 + B_3D + B_0$, in which X_1 is the person's score at time 1, X_2 is the person's score at time 2, and D is the difference between the time 1 and time 2 scores, $D = X_1 - X_2$. Note that D contains no unique information that is not contained in X_1 and X_2 so that the multiple correlation of D with X_1 and X_2, $R_{D.X_1X_2}$, will be 1.0. As another example, suppose a researcher asks students to explain their performance on their first statistics exam by dividing 100 points among four potential explanations. These explanations are $X_1 =$ ability in statistics, $X_2 =$ difficulty of the test, $X_3 =$ amount of effort studying the material, and $X_4 =$ luck. The researcher wishes to predict Y, the student's performance on the final exam. But, note that for each student $X_1 + X_2 + X_3 + X_4$ must equal a constant, here 100 points. This means that if for a particular student we know that the value of $X_1 = 20, X_2 = 10, X_3 = 20$, then we know that the value of X_4 must be 50, i.e., $100 - X_1 - X_2 - X_3$. Once again, X_4 adds no unique

information that is not contained in X_1, X_2, and X_3, and the multiple correlation between any one of the IVs and the other three must be 1.0. These more subtle forms of exact collinearity typically occur when the sum of the predictors must equal a constant value or when composite scores and the original scores from which they are derived are included in the same regression equation.

When exact collinearity occurs, there is no mathematically unique solution for the regression coefficients. Major statistical packages perform an initial check to determine if one (or more) of the IVs is highly redundant with the other IVs in the regression equation.[13] When they detect this problem, some regression programs will not run. Other programs will attempt to "fix" the problem by arbitrarily dropping one (or more) IVs with exact collinearity from the regression model, perhaps even an IV in which the researcher has particular interest.

10.5.2 Multicollinearity: A Numerical Illustration

Multicollinearity occurs when highly related IVs are included in the same regression model. In cross-sectional research, serious multicollinearity most commonly occurs when multiple measures of the same or similar constructs (e.g., depression, anxiety) are used as the IVs in a regression equation. In longitudinal research, serious multicollinearity most commonly occurs when similar measures collected at several previous time points are used to predict the participants' score at a later time point (e.g., all four test scores in a statistics class are used as IVs in a regression equation predicting the final exam score). As with exact collinearity, highly related IVs can occur in more subtle ways as well. Some measures which purport to measure different constructs are based on overlapping sets of similar items so that they will be highly related. For example, some MMPI-based scales used in clinical psychology are based on partially overlapping sets of items.

As was introduced in Section 3.8, multicollinearity may lead to unstable regression coefficients that are associated with large standard errors. Multicollinearity can also lead to complexities in interpreting regression coefficients. We illustrate these ideas with two examples. In our first example we examine data with two IVs to explore what happens to the regression estimates as r_{12} takes on increasingly large values. In this example $r_{Y1} = .30$ and $r_{Y2} = .40$ are kept constant. These values represent a moderate and a moderate to large effect size, respectively, according to the normative values presented in Chapter 2. The variances are $sd_Y^2 = 5, sd_{X1}^2 = 3, sd_{X2}^2 = 4$; x_1 and x_2 are centered, M_Y is set equal to 20, and $n = 100$.


CH10EX4

Table 10.5.1 presents the results of several regression analyses based on this data set. First, consider Table 10.5.1A, which presents the results when $r_{12} = 0$. The intercept $B_0 = 20$ (the mean value of Y), $B_1 = .39, SE_{B_1} = .11, B_2 = .45, SE_{B_2} = .10$ (to two decimals). We can construct the 95% confidence interval for each B_i, $CI = B_i \pm t\,SE_{B_i}$. Thus, the 95% confidence interval for $B_1 = .16$ to .61 and the 95% confidence interval for $B_2 = .25$ to .64. Examining null hypothesis significance tests, for B_1, $t = 3.41$, $df = 97, p < .001$ and for $B_2, t = 4.55$, and again, $p < .001$. These results indicate that both x_1 and x_2 make independent contributions to the prediction of Y.

Now, let us consider what happens as we increase the value of r_{12}. Comparing the values of B_1 across Parts A–E of Table 10.5.1, we see that B_1 decreases in value and ultimately becomes *negative* at the highest values of r_{12} (e.g., $B_1 = -1.03$ for $r_{12} = .949$). In contrast, B_2 initially decreases in value, but reaches a minimum following which it rapidly increases[14] at very high levels of r_{12}. The standard errors of B_1 and B_2 initially increase slowly in magnitude as r_{12}

[13] Statistical packages typically compare one of the indices of multicollinearity (to be presented later) with a very extreme cutoff value (e.g., tolerance = .0001). This procedure detects cases in which exact collinearity is obscured by computer rounding errors.

[14] The rapid increase of B_2 at high levels of r_{12} is an example of statistical suppression discussed in Section 3.4.

TABLE 10.5.1
Effects of Multicollinearity:
Two-Independent-Variable Example

A. $r_{12} = 0.00; r_{Y1} = .30; r_{Y2} = .40; R^2 = .250.$

Variable	B	SE	pr^2	Tolerance	VIF
Intercept	20.000	0.196			
x_1	0.387	0.114	0.107	1.000	1.000
x_2	0.447	0.098	0.176	1.000	1.000

B. $r_{12} = 0.10; r_{Y1} = .30; r_{Y2} = .40; R^2 = .228.$

Variable	B	SE	pr^2	Tolerance	VIF
Intercept	20.000	0.198			
x_1	0.339	0.116	0.081	0.990	1.010
x_2	0.418	0.100	0.152	0.990	1.010

C. $r_{12} = 0.50; r_{Y1} = .30; r_{Y2} = .40; R^2 = .173.$

Variable	B	SE	pr^2	Tolerance	VIF
Intercept	20.000	0.205			
x_1	0.172	0.138	0.016	0.750	1.333
x_2	0.373	0.119	0.092	0.750	1.333

D. $r_{12} = 0.90; r_{Y1} = .30; r_{Y2} = .40; R^2 = .179.$

Variable	B	SE	pr^2	Tolerance	VIF
Intercept	20.000	0.205			
x_1	−0.407	0.272	0.023	0.190	5.263
x_2	0.765	0.236	0.098	0.190	5.263

E. $r_{12} = 0.949; r_{Y1} = .30; r_{Y2} = .40; R^2 = .224.$

Variable	B	SE	pr^2	Tolerance	VIF
Intercept	20.000	0.199			
x_1	−1.034	0.366	0.076	0.099	10.060
x_2	1.297	0.317	0.147	0.099	10.060

Note: $sd_Y^2 = 5.00; sd_1^2 = 3.00; sd_2^2 = 4.00; M_Y = 20.$

increases in value, but then rise rapidly as r_{12} becomes close to 1. Indeed, when $r_{12} = .949$, the standard errors of B_1 and B_2 are more than 3 times larger than when $r_{12} = 0$. Necessarily, this increase in both SE_{B_1} and SE_{B_2} leads to corresponding increases in the associated *CI*s: −.95 to +.13 for B_1 and +.30 to 1.23 for B_2. Thus, Table 10.5.1 illustrates both the increased difficulty that can arise in interpreting the regression coefficients and the increase in *SE*s that occurs as two predictors become highly correlated.

Table 10.5.2A provides a second illustration of the effects of multicollinearity, this time with four centered IVs, x_1, x_2, x_3, and x_4. This example compares the results of two regression analyses. In the first analysis, presented in Table 10.5.2A, the IVs are uncorrelated. In the second analysis, presented in Table 10.5.2B, x_1, x_2 and x_3 are highly intercorrelated ($r_{12} = r_{13} = r_{23} = .933$). Note in the second analysis that x_4 is uncorrelated with x_1, x_2, or x_3. As in our first example presented in Table 10.5.1, the correlations of the IVs with Y were kept at

TABLE 10.5.2
Effects of Multicollinearity: Four-Independent-Variable Example

A. $r_{12} = r_{13} = r_{23} = 0.00$; $r_{14} = 0$; $R^2 = .495$.

Variable	B	SE	pr^2	Tolerance	VIF
Intercept	20.000	0.162			
x_1	0.387	0.094	0.151	1.000	1.000
x_2	0.391	0.082	0.195	1.000	1.000
x_3	0.400	0.073	0.240	1.000	1.000
x_4	0.391	0.082	0.195	1.000	1.000

B. $r_{12} = r_{13} = r_{23} = 0.933$; $r_{14} = 0$; $R^2 = .325$.

Variable	B	SE	pr^2	Tolerance	VIF
Intercept	20.000	0.187			
x_1	−0.806	0.345	0.054	0.099	10.067
x_2	0.137	0.299	0.002	0.099	10.067
x_3	0.868	0.267	0.100	0.099	10.067
x_4	0.391	0.094	0.154	1.000	1.000

Note: $r_{Y1} = .30$; $r_{Y2} = .40$; $r_{Y3} = .30$; $r_{Y4} = .40$; $sd_Y^2 = 5.00$; $sd_1^2 = 3.00$; $sd_2^2 = 4.00$; $sd_3^2 = 3.00$; $sd_4^2 = 4.00$; $M_Y = 20$.

constant values within the range .30 to .40 and the variances of all variables were also kept constant, within the range 3 to 5.

Table 10.5.2A displays the results of the first analysis when the IVs are uncorrelated. Under these circumstances, the B_1 to B_4 regression coefficients range from .39 to .40 and the *SE*s range from .07 to .09 (to two decimals). In contrast, Table 10.5.2B shows the results when x_1 to x_3 are highly intercorrelated, but x_4 is uncorrelated with x_1, x_2, or x_3. The same pattern that we observed in the first example emerges for x_1 to x_3. The regression coefficients now range from −.0.81 to +0.86 and their *SE*s now range from 0.27 to 0.35, over a three fold increase in *SE*s relative to the uncorrelated case. In contrast, note that the value of $B_4 = 0.39$, is *exactly* the same value obtained in Table 10.5.2A when all IVs were uncorrelated. The *SE* of B_4 does increase from .082 to .094, but this increase is very modest relative to the increase in the *SE*s of the regression coefficients.

Both examples illustrate the increased complexity in interpreting the meaning of the *B*s and the increase in *SE* when there are high correlations among the IVs. However, the *B*s of IVs that are unrelated to the other predictors (here, x_4) are not affected and their *SE*s are only minimally affected, even when there is a high degree of multicollinearity among the other predictors in the regression equation.

10.5.3 Measures of the Degree of Multicollinearity

$r^2_{X_i X_j}$

The squared correlation between each of the pairs of predictor variables provides an index of bivariate multicollinearity. As its value increases toward 1.0, the magnitude of potential problems associated with multicollinearity increases correspondingly. With two IVs, this index

is sufficient. As the number of IVs in the regression model increases, this index becomes increasingly likely to miss substantial multicollinearity.

The Variance Inflation Factor

The variance inflation factor (*VIF*) provides an index of the amount that the variance of each regression coefficient is increased relative to a situation in which all of the predictor variables are uncorrelated. To understand the *VIF*, recall from Section 3.6 that the standard error of B_i is

(3.6.1) $$SE_{B_i} = \frac{sd_Y}{sd_{X_i}} \sqrt{\frac{1 - R^2_{Y.12\ldots k}}{n - k - 1}} \sqrt{\frac{1}{1 - R^2_{i.12\ldots(i)\ldots k}}}.$$

where $R^2_{i.12\ldots(i)\ldots k}$ is the squared multiple correlation between X_i and the other predictor variables in the regression equation. Squaring this equation, we get the variance of B_i, $V(B_i)$:

(10.5.1) $$V(B_i) = \frac{sd^2_Y}{sd^2_{X_i}} \left(\frac{1 - R^2_{Y.12\ldots k}}{n - k - 1}\right) \left(\frac{1}{R^2_{i.12\ldots(i)\ldots k}}\right).$$

The *VIF* is simply the third term in Eq. (10.5.1), so

(10.5.2) $$VIF \left(\frac{1}{1 - R^2_{i.12\ldots(i)\ldots k}}\right).$$

A *VIF* is calculated for each term in the regression equation, excluding the intercept. A commonly used rule of thumb is that any *VIF* of 10 or more provides evidence of serious multicollinearity involving the corresponding IV. Given the relationships between Eq. (3.6.1) and (10.5.1), $\sqrt{VIF}$ will represent the amount that the *SE* of B_i will increase relative to the situation in which all of the predictor variables are uncorrelated. Thus, a *VIF* of 10 means that there is a $\sqrt{10} = 3.16$ or slightly more than a threefold increase in SE_{B_i} relative to the situation of no correlation between any of the IVs. Table 10.5.1E and Table 10.5.2B include values of intercorrelations among IVs that produce *VIF*s of about 10. Note that the *SE*s show slightly more than a threefold increase relative to their values when $r_{X_iX_j} = 0$. As illustrated in Table 10.5.1, in the two-predictor case, the *VIF*s for X_1 and X_2 will always be equal. In the three or more predictor case, the VIFs will, in general, not be equal. The *VIF*s for X_1 to X_3 in Table 10.5.2 are equal only because the correlations between predictors are precisely equal, $r_{12} = r_{13} = r_{23}$.

We remind readers that extremely high intercorrelations between predictors were necessary to produce *VIF*s = 10 in Tables 10.5.1 and 10.5.2. We believe that this common rule of thumb guideline is too high (lenient) for most behavioral science applications. We discuss issues in measuring multicollinearity later in this section.

Tolerance

Some statistical packages present the tolerance in addition to or instead of the *VIF*. The tolerance is the reciprocal of the *VIF*,

(10.5.3) $$\text{tolerance} = \frac{1}{VIF} = 1 - R^2_{X_i.X_1X_2\ldots(i)\ldots p}.$$

and therefore tells us how much of the variance in X_i is independent of other IVs. This relationship can be verified in Tables 10.5.1 and 10.5.2. A commonly used rule of thumb is that

tolerance values of .10 or less indicate that there may be serious problems of multicollinearity in the regression equation (and, of course, are equivalent to a *VIF* of 10). Some statistical packages use very low values of tolerance (e.g., .0001) as a means of detecting exact collinearity.

Condition Number

The correlation matrix of the IVs may be decomposed into a set of orthogonal dimensions. Orthogonal dimensions are completely nonoverlapping and share no variance in common. Major statistical programs will perform this decomposition, which is known as principal components analysis. When this analysis is performed on the correlation matrix of the k IVs, a set of k *eigenvalues* or characteristic roots of the matrix is produced. The proportion of the variance in the IVs accounted for by each orthogonal dimension i is λ/k, where λ is the eigenvalue. The eigenvalues are ordered from largest to smallest so that each orthogonal dimension in turn accounts for a smaller proportion of the variance of the IVs. With two independent variables, if the two IVs are uncorrelated, each eigenvalue will equal 1.0, so that each *independent* dimension will account for $\lambda_i/2$, or .50 of the variance in the set of IVs. As the IVs become increasingly correlated, more and more of the variance in the IVs is associated with the first dimension, so that the value of the first eigenvalue will become larger and the value of the second eigenvalue will become correspondingly smaller. When r_{12} reaches its maximum, $r_{12} = 1.0, \lambda_1 = 2$ and $\lambda_2 = 0$, so that the first dimension will account for 1.0 of the variance in the IVs, whereas the second dimension will account for no additional variance in the IVs.

The condition number[15] κ (kappa) is defined as the square root of the ratio of the largest eigenvalue (λ_{max}) to the smallest eigenvalue (λ_{min}).

(10.5.4)
$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}.$$

Traditionally, a rule of thumb has been suggested that values of κ (kappa) that are 30 or larger indicate highly severe problems of multicollinearity. However, no strong statistical rationale exists for this choice of 30 as a threshold value above which serious problems of multicollinearity are indicated.

Some Issues in Measuring Multicollinearity

In discussing the measures of multicollinearity, we noted rule of thumb cutoff values that have been offered above which multicollinearity appear to be problematic in most behavioral science applications. Note that the problems associated with multicollinearity differ in degree; they do not differ in kind. Thus, there is no good statistical rationale for the choice of any of the traditional rule of thumb threshold values for separating acceptable from unacceptable levels of multicollinearity[16]. A review of Table 10.5.1 indicates that the magnitude and direction of the regression coefficients may change appreciably at values of the *VIF* that are substantially less than the typically suggested threshold value of 10. As presented in Section 3.4.1, when B increases or changes its sign when one or more other IVs are added to the equation, we have

---

[15]The condition number may be computed from matrices involving uncentered scores on the IVs, centered scores on the IVs, or the correlation matrix of the IVs. This choice can lead to differences in the eigenvalues that are obtained. Consequently the value of the condition number may depend on the specific matrix which the statistical package uses to compute eigenvalues. Typically, the eigenvalues are based on the $(\mathbf{X}_c'\mathbf{X}_c)$ matrix, where $\mathbf{X}_c$ is the $(n \times k)$ matrix of the centered predictor values. Belsley (1984, 1991) and R. D. Cook (1984) discuss this issue and indicate cases under which computations based on each matrix may be preferred.

[16]For example, some authors have proposed values of 6 or 7 as a threshold value for the *VIF* or 15 or 20 as a threshold value for the condition index.

a situation of statistical suppression. Although such findings may sometimes be theoretically anticipated, they are often indicative of a serious problem of multicollinearity. Thus, the values of the multicollinearity indices at which the interpretation of regression coefficients may become problematic will often be considerably smaller than traditional rule of thumb guidelines such as $VIF = 10$.

Multicollinearity indices provide useful information but do not substitute for more basic checks on the data. First, researchers should carefully examine the scatterplot matrix of the predictor variables and the leverages for each case, looking for outlying observations that may affect the relationship between each pair of IVs. As we saw earlier in this chapter, outliers can greatly increase or decrease the magnitude of the relationship between variables, leading to values of multicollinearity indices that may be too high or too low. Second, researchers should compare the results of simple univariate regression analyses in which the outcome is regressed separately on each predictor variable with the results of the full multiple regression analysis in which the outcome is regressed on all of the predictor variables of interest. Probably the easiest way to accomplish this is by comparison of r_{YX} for each IV with its corresponding standardized β in the regression equation. Large, unexpected changes in direction and magnitude of these coefficients suggest a substantial influence of multicollinearity.

Relatively high values of the standard multicollinearity indices may occur in some of the more complex regression analyses we considered in earlier chapters of this book. In these analyses a single substantive IV was represented by more than one term in the regression equation. In Chapter 6 we used several polynomial terms (e.g., X, X^2) to represent a variable's nonlinear relationship with Y. In Chapter 7 we used terms that are products of IVs (e.g., XZ) to represent interactions. In Chapter 8 we introduced coding schemes (e.g., dummy codes C_1, C_2, C_3) that used multiple terms to represent qualitative variables such as religious affiliation or experimental treatment groups. In these circumstances, high values on standard measures of multicollinearity are not necessarily problematic—the degree of multicollinearity depends on the particular scaling of the IVs. For example, we showed in Chapter 7 that in a regression model with an interaction term the correlation between X and XZ can be reduced by centering each of the IVs. Fox and Monette (1992) present a general index of multicollinearity that is not affected by the scaling of the IVs that is appropriate in such applications.

VIF, tolerance, and the condition number implemented in most statistical packages do not take multicollinearity involving the intercept into account. This characteristic is fully appropriate in most applications in the behavioral sciences. However, in some areas of economics and the physical and biological sciences, IVs such as interest rates and body size are measured on ratio level measurement scales (see Chapter 1). With a ratio level of measurement, the value of the intercept estimated when each of the IVs takes on a true value of 0 can be a parameter that is of considerable theoretical interest. In such cases, alternative versions of the *VIF*, tolerance, or condition number discussed by Belsley (1991) should be calculated.[17]

10.6 REMEDIES FOR MULTICOLLINEARITY

When a researcher is interested solely in the prediction of Y or in the value of R^2, multicollinearity has little effect and no remedial action is needed. However, in research testing a substantive theory in which the researcher is interested in the value of each B_i, high values

[17]Belsley (1991, Chapter 5) has also developed a useful extension of the condition number that more precisely pinpoints the sources of multicollinearity. This approach to detecting multicollinearity is of particular value in time series analysis and complex econometric models that include large numbers of IVs.

of multicollinearity present a potentially serious problem. Four general approaches to solving problems of multicollinearity have been proposed.

10.6.1 Model Respecification

In some cases, it may be possible to revise the regression model so that the degree of multicollinearity is reduced. This remedy is particularly applicable when the analyst has included several highly correlated variables that can be thought of as measuring the same underlying construct. For example, suppose a researcher were interested in the effects of socioeconomic status (SES) and IQ on undergraduate GPA (Y). Suppose the researcher has collected several measures of SES—mother's income (X_1), father's income (X_2), mother's education (X_3), and father's education (X_4), father's occupational status (X_5), and mother's occupational status (X_6)—as well as IQ (X_7)—and has included all seven predictors in a regression equation,

$$\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6 + B_7X_7 + B_0.$$

The IVs X_1 to X_6 are all measures of SES and are likely to be moderately to very highly correlated, leading to high levels of multicollinearity. In such cases, it is often useful to combine the variables measuring the underlying construct, here SES, into a single index. The simplest way to do this is to convert each of the measures to z scores. The z scores are then averaged to produce an overall index of SES for each person in the sample. In cases where theory or prior empirical work point to differential importance of each of the variables assessing the construct, more complex weighting schemes can be used to form the overall index. In either case, a different regression model is now estimated,

$$\hat{Y} = B_1z_{\text{SES}} + B_2X_7 + B_0$$

Note that in this equation, B_1 represents the unique contribution of the index of SES over and above IQ to the prediction of GPA, whereas in the previous equation B_1 represented the unique contribution of mother's income over and above the five other measures of socioeconomic status and IQ to the prediction of GPA. Thus, we have respecified the model so that it answers a different question than the one posed by the original analysis, but a question that in many cases may more adequately represent the researcher's question of interest.

An alternative approach to model respecification is to drop one (or more) IVs from the regression equation. Multicollinearity measures provide information about sources of multicollinearity, but they do not tell the researcher which IVs should be retained in the regression equation. When either theory or prior empirical work exists, it should be used as a strong guide. For example, if a variable has been thrown into the regression equation "to see what happens," it is a prime candidate for deletion. In other situations, choosing a variable or variables to delete among several variables that are contributing to high multicollinearity may be largely an arbitrary decision. Deletion of IVs on the basis of correlation with other IVs always carries a risk. If the IV in question is truly relevant to the theory, the estimates of all other IVs will be biased in its absence.

This caution about dropping variables from the regression equation takes on particular importance in complex regression equations in which multiple terms are used to represent a curvilinear effect (Chapter 6), an interaction (Chapters 7 and 9), or a categorical IV (Chapter 8). Deletion of lower order terms that are included in higher order terms leads to poorly structured regression models with Bs that are not readily interpretable (Peixoto, 1987). For example, in Chapter 7 we presented the interpretation of regression equation specifying a linear XZ interaction, $\hat{Y} = B_1X + B_2Z + B_3XZ + B_0$. If the X term were now dropped from the equation,

$\hat{Y} = B_2Z + B_3XZ + B_0$, the B_3 coefficient for the XZ term no longer represents purely the interaction between X and Z, but confounds the interaction with the first order effect of X. Lower order terms should *not* be dropped (see Chapter 7 and Aiken & West, 1991, Chapter 3).

10.6.2 Collection of Additional Data

The collection of additional data reduces some but not all of the problems associated with multicollinearity. Larger sample sizes will always improve the precision of the estimate of B With small samples, the degree of multicollinearity will typically be overestimated if there are a large number of IVs in the regression model. Large samples will reduce this problem. However, the pattern of correlations among the IVs would not be expected to change as sample size increases. Thus, the use of large samples alone cannot eliminate difficulties that arise in *interpreting* regression coefficients when IVs are highly multicollinear.

An alternative approach is to try to reduce the correlations among the IVs. In some cases, it may be possible to manipulate one or more of the predictors in an experimental setting. For example, suppose a researcher is studying stress (X_1) and coping skills (X_2) as predictors of well being. X_1 and X_2 are very highly correlated in the population. To minimize this correlation, study participants could be randomly assigned in a laboratory experiment to a highly or mildly stressful experience so that $r_{X_1X_2}$ would on average be expected to be 0. Alternatively, if the scores of potential participants on the IVs (but not the DV) were known in the population (e.g., a large school district), the researcher could devise a sampling plan so that the correlation between X_1 and X_2 would be substantially reduced in magnitude (see McClelland & Judd, 1993; Pitts & West, 2001). Such procedures can permit a greater understanding of the independent effects of each of the predictor variables as reflected in the unstandardized regression coefficients (Bs). At the same time, they change the estimates of standardized effect sizes for each predictor variable and R^2 so that these statistics no longer estimate the values in the original population. Pitts and West (2001) discuss these procedures, their strengths, their limitations, and appropriate methods of estimating population effect sizes.

10.6.3 Ridge Regression

As was the case with outliers, alternative estimation techniques exist that can provide "improved" estimates of each regression coefficient and its associated standard error when multicollinearity is present. Ridge regression is an alternative estimation method that may be used when there is an *extremely* high degree of multicollinearity in a data set (Darlington, 1978). In ridge regression a constant is added to the variance of each IV. This procedure leads to a biased estimate of each regression coefficient B_i—the estimate is slightly attenuated (too close to 0) so that it is no longer on average equal to the value of β_i^* in the population. However, the estimate of SE_B may be substantially reduced. When multicollinearity is extremely high, it may be advantageous to trade off a small increase in bias for a substantial increase in the precision of the estimate of the regression coefficient. The regression coefficients will be far less sensitive to small changes in the data set such as adding or deleting a case.

The details of implementing ridge regression are presented in Ryan (1997, Chapter 12). Draper and Smith (1998, Chapter 17) provide a discussion of the strengths and weaknesses of ridge regression and note that ridge regression estimates are not always superior to OLS estimates. Unlike OLS estimates, ridge regression estimates of the regression coefficients are biased; consequently, alternative methods presented in Neter, Kutner, Nachtsheim, and Wasserman (1996, Chapter 10) must be used to construct confidence intervals and conduct significance tests. Although the SAS, SPSS, and SYSTAT regression modules do not presently

include ridge regression, software is available in other statistical packages (see Ryan, 1997, for an overview). Box 10.6.1 presents an illustration of ridge regression.

BOX 10.6.1
Illustration of Ridge Regression

To illustrate how ridge regression works, consider a case with two predictors X_1 and X_2. Multicollinearity between two predictor variables can be assessed by r_{12}^2, which can be calculated using Eq. (2.3.5). We have squared and rewritten original Eq. (2.3.5) below for ease of presentation:

(2.3.5) $$r_{12}^2 = \frac{\left[\sum x_1 x_2/(n-1)\right]^2}{sd_1^2 sd_2^2}$$

The numerator $\left[\Sigma x_1 x_2/(n-1)\right]^2$, is the squared covariance between x_1 and x_2, and the denominator terms are the variances of the two IVs. Suppose the covariance of x_1 and x_2 = 141, sd_1^2 = 100, and sd_2^2 = 200. Substituting these values into Eq. (2.3.5) gives r_{12}^2 = .994. From Eq. (10.5.2), the *VIF* is $1/(1 - .994) = 168.1$, an extremely high value. Now what happens to r_{12}^2 and the *VIF* if we add a constant value of 10 to each variance? r_{12}^2 now equals $(141)^2/((110)(210)) = .861$ so that the *VIF* will be $1/(1 - .861) = 7.19$. Thus, the addition of a relatively small constant to the variance of each predictor decreases the correlation between the IVs and the value of the *VIF*, and hence greatly reduces the standard errors of the tests of the regression coefficients. The central problem in ridge regression is to choose the value of the constant that will provide the maximum benefit in terms of improvement of the precision of the estimate at the minimum cost in terms of bias in the estimates of the regression coefficient. Several methods exist for making this choice; they are discussed in Draper and Smith (1998) and Ryan (1997).

10.6.4 Principal Components Regression

In Section 10.5.3, we briefly noted that a set of independent dimensions can be created that are combinations of the predictor variables. In principal components regression, we regress the dependent variable on these independent dimensions rather than on the original set of predictor variables. To create these independent dimensions, a procedure known as principal components analysis is used (see e.g., Harris, 2001; Tabachnick & Fidell, 2001). These dimensions are termed *components*. Principal components analysis produces a set of k eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_k)$, each of which has a corresponding eigenvector. The elements of the eigenvectors are weights, with a different set of weights being produced for each component. These weights allow the researcher to transform the participant's original score on the set of IVs into a score on each component, C_i. The raw scores for each IV are first standardized and then the weights are applied. Thus, with four IVs the participant's score on component i would be

(10.6.1) $$C_i = w_{i1}z_1 + w_{i2}z_2 + w_{i3}z_3 + w_{i4}z_4,$$

where z_1 to z_4 are the participant's z-scores that correspond to scores on the original IVs, X_1 to X_4, and w_{i1} to w_{i4} are the weights for component i determined by the principal components analysis.

Each principal component represents an orthogonal dimension. The components are ordered from largest to smallest in terms of the variance of the original IVs that is reproduced in the component. Thus, each principal component, in turn, accounts for a smaller and smaller proportion of the variance in the IVs. The last component (or last few components) will often account for little variance in the IVs; small components are sources of multicollinearity in the original data. Thus, we have created a new set of orthogonal variables $C_1, C_2, \ldots, C_k$ that collectively represent all of the information that is contained in the k original IVs, but that reorganize the information into orthogonal sources. Indeed, if we regressed Y on C_1 to C_k, we will obtain the identical R^2 as we would if we regressed Y on X_1 to X_k.

To illustrate principal components regression, suppose we had a regression equation with 4 IVs and that the four components accounted for 61%, 28%, 10.5%, and 0.5% of the variance in the IVs. The unique feature of this procedure is that we may be able to discard the last orthogonal component (which accounts for very little variance) with little loss of information. We then regress Y on the three remaining orthogonal component scores,

$$\hat{Y} = \tilde{B}_1 C_1 + \tilde{B}_2 C_2 + \tilde{B}_3 C_3 + \tilde{B}_0, \tag{10.6.2}$$

where $\tilde{B}_i$ represents the regression coefficient for the ith component. Since the components are orthogonal, the discarding of small components has no impact on statistical inference about the $\tilde{B}_i$ terms. We simply use the standard significance testing and confidence interval procedures discussed in Chapter 3.

Unfortunately, however, these $\tilde{B}_i$ are only rarely interpretable. The component scores are linear combinations of the original IVs (see Eq. 10.6.1) and will not typically have a clear meaning. Consequently, most sources recommend that researchers transform the regression coefficients for the components back to the regression coefficients for the original scores, B_i. The procedure of discarding small components means that this transformation will not reproduce the results of the regression analysis on the original IVs. On the positive side, dropping components that account for small proportions of variance eliminates major sources of multicollinearity. The result is that the back transformed regression coefficients, B_i, for the original IVs will be biased, but will be more robust to small changes in the data set than are original OLS estimates. Once again, constructing confidence intervals and performing significance tests on the B_is becomes more complex because these estimates are biased and do not follow a t distribution (see Chatterjee & Price, 1991).

Chatterjee and Price (1991) present a good introduction and Jackson (1991), Jolliffe (1986) and Hadi and Ling (1998) present a thorough discussion of the strengths and weaknesses of regression on principal components. In some data sets the small rather than the large components may account for most of the predictable variance in Y so that the small components cannot be discarded without substantially affecting the results of the regression analysis. Regression on principal components is limited to regression equations that are linear in the variables (e.g., no interactions or polynomial terms; see Cronbach, 1987). Software to conduct principal components analysis is available in the SAS PROC Factor, SPSS Factor, and SYSTAT.

10.6.5 Summary of Multicollinearity Considerations

In most areas of the behavioral sciences, severe multicollinearity according to traditional statistical standards does not occur. Cases only rarely occur in which conventional statistical rules of thumb such as a $VIF = 10$ or a condition index $= 30$ are exceeded. Instead, multicollinearity occurs to a lesser degree, but enough to produce regression coefficients that may be difficult to interpret. In these cases of "moderate" multicollinearity, alternative estimation procedures such as ridge regression or principal components regression cannot be counted on to produce

better estimates than OLS regression. Instead, the researcher's focus should be on attempting to understand the nature and the source of the multicollinearity problem. In some cases, the regression model may be modified by combining IVs that are measures of the same underlying construct or by dropping variables from the equation. The primary risk here is that important IVs may be inadvertently dropped from the model.[18] In other cases, it may be possible to experimentally manipulate one of the IVs or to systematically sample participants so that the IVs are less correlated. Careful selection of the IVs with regard to their relationship to theoretical constructs will also often help reduce the problem of multicollinearity.

10.7 SUMMARY

We begin this chapter with a consideration of outliers, atypical data points that can affect the results in multiple regression analysis. After an illustration of the potential effects of outliers particularly in small samples (Section 10.2), we consider measures of extremity of a single case on the IVs (leverage, h_{ii}), on the DV (discrepancy, t_i), and on the overall results of the regression equation (global influence, $DFFITS_i$ and Cook's D_i) as well as the individual regression coefficients (specific influence, $DFBETAS_{ij}$). A small number of cases may be identified as outliers if they are extreme relative to the other cases in the data set and exceed minimum rule of thumb cutoff values (Section 10.3). Outliers may be produced by contaminated observations or rare cases. Contaminated observations may be corrected or the case may be removed from the data set; in contrast, the proper procedures to take with rare cases are more difficult, involving both statistical and substantive considerations. Potential remedial actions include deleting the case(s) from the data set (possibly changing to population to which the results may be generalized), respecification of the regression equation, transformation of the variables to account for the case(s), and robust regression approaches that downweight the influence of the outliers in the regression analysis (Section 10.4).

We then consider the problem of multicollinearity in regression analysis which occurs when the IVs become highly correlated. Several measures of the degree of multicollinearity including $r^2_{X_iX_j}$, the variance inflation factor, tolerance, and the condition number are presented. Standard cutoff values for these measures are presented, but they appear to be far too high for many behavioral science applications (Section 10.5). The advantages and disadvantages of several remedies for multicollinearity including model respecification, collection of additional data, ridge regression, and principal components regression are presented. Careful design of studies, selection of conceptually relevant measures, and specification of regression models can often help avoid problems of multicollinearity (Section 10.6).

[18] See our presentation of sensitivity analysis in Chapter 5.