Chapter

Building the Regression Model II: Diagnostics

In this chapter we take up a number of refined diagnostics for checking the adequacy of a regression model. These include methods for detecting improper functional form for a predictor variable, outliers, influential observations, and multicollinearity. We conclude the chapter by illustrating the use of these diagnostic procedures in the surgical unit example. In the following chapter, we take up some remedial measures that are useful when the diagnostic procedures indicate model inadequacies.

10.1 Model Adequacy for a Predictor Variable—Added-Variable Plots

We discussed in Chapters 3 and 6 how a plot of residuals against a predictor variable in the regression model can be used to check whether a curvature effect for that variable is required in the model. We also described the plotting of residuals against predictor variables not yet in the regression model to determine whether it would be helpful to add one or more of these variables to the model.

A limitation of these residual plots is that they may not properly show the nature of the marginal effect of a predictor variable, given the other predictor variables in the model. *Added-variable plots*, also called *partial regression plots* and *adjusted variable plots*, are refined residual plots that provide graphic information about the marginal importance of a predictor variable X_k, given the other predictor variables already in the model. In addition, these plots can at times be useful for identifying the nature of the marginal relation for a predictor variable in the regression model.

Added-variable plots consider the marginal role of a predictor variable X_k, given that the other predictor variables under consideration are already in the model. In an added-variable plot, both the response variable Y and the predictor variable X_k under consideration are regressed against the other predictor variables in the regression model and the residuals are obtained for each. These residuals reflect the part of each variable that is not linearly associated with the other predictor variables already in the regression model. The plot of these residuals against each other (1) shows the marginal importance of this variable in reducing the residual variability and (2) may provide information about the nature of the marginal

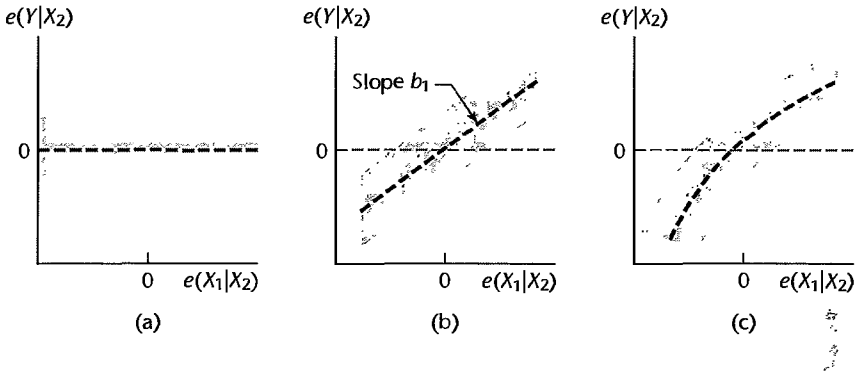


FIGURE 10.1 Prototype Added-Variable Plots.

regression relation for the predictor variable X_k under consideration for possible inclusion in the regression model.

To make these ideas more specific, we consider a first-order multiple regression model with two predictor variables X_1 and X_2. The extension to more than two predictor variables is direct. Suppose we are concerned about the nature of the regression effect for X_1, given that X_2 is already in the model. We regress Y on X_2 and obtain the fitted values and residuals:

$$\hat{Y}_i(X_2) = b_0 + b_2 X_{i2} \quad \textbf{(10.1a)}$$

$$e_i(Y|X_2) = Y_i - \hat{Y}_i(X_2) \quad \textbf{(10.1b)}$$

The notation here indicates explicitly the response and predictor variables in the fitted model. We also regress X_1 on X_2 and obtain:

$$\hat{X}_{i1}(X_2) = b_0^* + b_2^* X_{i2} \quad \textbf{(10.2a)}$$

$$e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2) \quad \textbf{(10.2b)}$$

The added-variable plot for predictor variable X_1 consists of a plot of the Y residuals $e(Y|X_2)$ against the X_1 residuals $e(X_1|X_2)$.

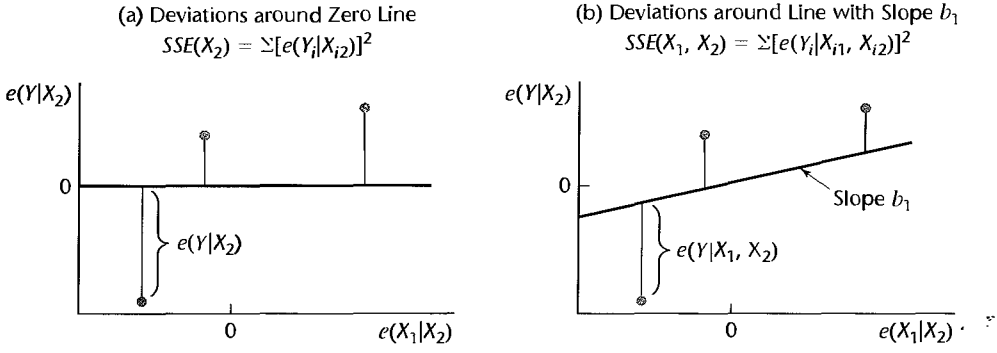
Figure 10.1 contains several prototype added-variable plots for our example, where X_2 is already in the regression model and X_1 is under consideration to be added. Figure 10.1a shows a horizontal band, indicating that X_1 contains no additional information useful for predicting Y beyond that contained in X_2, so that it is not helpful to add X_1 to the regression model here.

Figure 10.1b shows a linear band with a nonzero slope. This plot indicates that a linear term in X_1 may be a helpful addition to the regression model already containing X_2. It can be shown that the slope of the least squares line through the origin fitted to the plotted residuals is b_1, the regression coefficient of X_1 if this variable were added to the regression model already containing X_2.

Figure 10.1c shows a curvilinear band, indicating that the addition of X_1 to the regression model may be helpful and suggesting the possible nature of the curvature effect by the pattern shown.

Added-variable plots, in addition to providing information about the possible nature of the marginal relationship for a predictor variable, given the other predictor variables already in the regression model, also provide information about the strength of this relationship. To see how this additional information is provided, consider Figure 10.2. Figure 10.2a illustrates

FIGURE 10.2 Illustration of Deviations in an Added-Variable Plot.



an added-variable plot for X_1 when X_2 is already in the model, based on $n = 3$ cases. The vertical deviations of the plotted points around the horizontal line $e(Y|X_2) = 0$ shown in Figure 10.2a represent the Y residuals when X_2 alone is in the regression model. When these deviations are squared and summed, we obtain the error sum of squares $SSE(X_2)$. Figure 10.2b shows the same plotted points, but here the vertical deviations of these points are around the least squares line through the origin with slope b_1. These deviations are the residuals $e(Y|X_1, X_2)$ when both X_1 and X_2 are in the regression model. Hence, the sum of the squares of these deviations is the error sum of squares $SSE(X_1, X_2)$.

The difference between the two sums of squared deviations in Figures 10.2a and 10.2b according to (7.1a) is the extra sum of squares $SSR(X_1|X_2)$. Hence, the difference in the magnitudes of the two sets of deviations provides information about the marginal strength of the linear relation of X_1 to the response variable, given that X_2 is in the model. If the scatter of the points around the line through the origin with slope b_1 is much less than the scatter around the horizontal line, inclusion of the variable X_1 in the regression model will provide a substantial further reduction in the error sum of squares.

Added-variable plots are also useful for uncovering outlying data points that may have a strong influence in estimating the relationship of the predictor variable X_k to the response variable, given the other predictor variables already in the model.

Example 1

Table 10.1 shows a portion of the data on average annual income of managers during the past two years (X_1), a score measuring each manager's risk aversion (X_2), and the amount of life insurance carried (Y) for a sample of 18 managers in the 30–39 age group. Risk aversion was measured by a standard questionnaire administered to each manager; the higher the score, the greater the degree of risk aversion. Income and risk aversion are mildly correlated here, the coefficient of correlation being $r_{12} = .254$.

A fit of the first-order regression model yields:

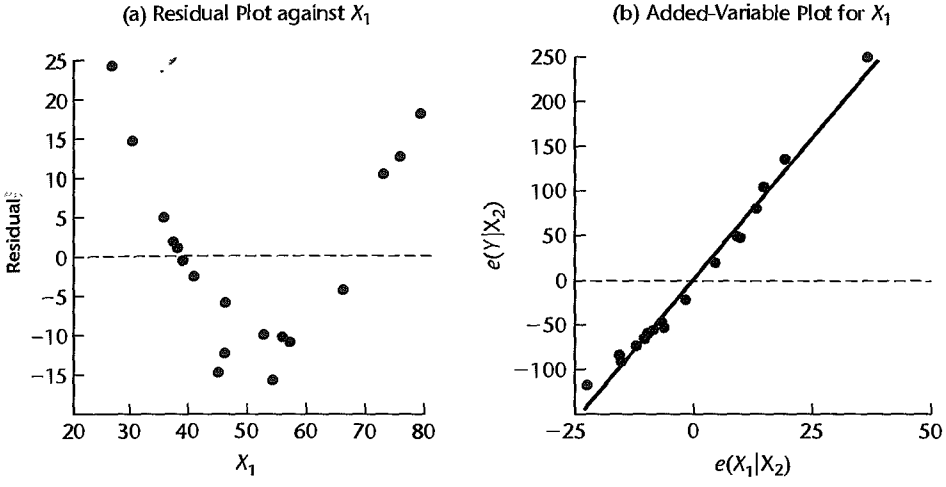
$$\hat{Y} = -205.72 + 6.2880X_1 + 4.738X_2 \qquad (10.3)$$

The residuals for this fitted model are plotted against X_1 in Figure 10.3a. This residual plot clearly suggests that a linear relation for X_1 is not appropriate in the model already containing X_2. To obtain more information about the nature of this relationship, we shall use an added-variable plot. We regress Y and X_1 each against X_2. When doing this, we

TABLE 10.1 Basic Data—Life Insurance Example.

Manager i	Average Annual Income (thousand dollars) X_{i1}	Risk Aversion Score X_{i2}	Amount of Life Insurance Carried (thousand dollars) Y_i
1	45.010	6	91
2	57.204	4	162
3	26.852	5	11
...	...	...	...
16	46.130	4	91
17	30.366	3	14
18	39.060	5	63

FIGURE 10.3 Residual Plot and Added-Variable Plot—Life Insurance Example.



obtain:

$$\hat{Y}(X_2) = 50.70 + 15.54X_2 \quad \textbf{(10.4a)}$$

$$\hat{X}_1(X_2) = 40.779 + 1.718X_2 \quad \textbf{(10.4b)}$$

The residuals from these two fitted models are plotted against each other in the added-variable plot in Figure 10.3b. This plot also contains the least squares line through the origin, which has slope $b_1 = 6.2880$. The added-variable plot suggests that the curvilinear relation between Y and X_1 when X_2 is already in the regression model is strongly positive, and that a slight concave upward shape may be present. The suggested concavity of the relationship is also evident from the vertical deviations around the line through the origin with slope b_1. These deviations are positive at the left, negative in the middle, and positive again at the right. Overall, the deviations from linearity appear to be modest in the range of the predictor variables.

Note also that the scatter of the points around the least squares line through the origin with slope $b_1 = 6.2880$ is much smaller than is the scatter around the horizontal line $e(Y|X_2) = 0$, indicating that adding X_1 to the regression model with a linear relation will substantially reduce the error sum of squares. In fact, the coefficient of partial determination for the linear effect of X_1 is $R^2_{Y1|2} = .984$. Incorporating a curvilinear effect for X_1 will lead to only a modest further reduction in the error sum of squares since the plotted points are already quite close to the linear relation through the origin with slope b_1.

Finally, the added-variable plot in Figure 10.3b shows one outlying case, in the upper right corner. The influence of this case needs to be investigated by procedures to be explained later in this chapter.

Example 2

For the body fat example in Table 7.1 (page 257), we consider here the regression of body fat (Y) only on triceps skinfold thickness (X_1) and thigh circumference (X_2). We omit the third predictor variable (X_3, midarm circumference) to focus the discussion of added-variable plots on its essentials. Recall that X_1 and X_2 are highly correlated ($r_{12} = .92$). The fitted regression function was obtained in Table 7.2c (page 258):

$$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$$

Figures 10.4a and 10.4c contain plots of the residuals against X_1 and X_2, respectively. These plots do not indicate any lack of fit for the linear terms in the regression model or the existence of unequal variances of the error terms.

Figures 10.4b and 10.4d contain the added-variable plots for X_1 and X_2, respectively, when the other predictor variable is already in the regression model. Both plots also show the line through the origin with slope equal to the regression coefficient for the predictor variable if it were added to the fitted model. These two plots provide some useful additional information. The scatter in Figure 10.4b follows the prototype in Figure 10.1a, suggesting that X_1 is of little additional help in the model when X_2 is already present. This information is not provided by the regular residual plot in Figure 10.4a. The fact that X_1 appears to be of little marginal help when X_2 is already in the regression model is in accord with earlier findings in Chapter 7. We saw there that the coefficient of partial determination is only $R^2_{Y1|2} = .031$ and that the t^* statistic for b_1 is only .73.

The added-variable plot for X_2 in Figure 10.4d follows the prototype in Figure 10.1b, showing a linear scatter with positive slope. We also see in Figure 10.4d that there is somewhat less variability around the line with slope b_2 than around the horizontal line $e(Y|X_1) = 0$. This suggests that: (1) variable X_2 may be helpful in the regression model even when X_1 is already in the model, and (2) a linear term in X_2 appears to be adequate because no curvilinear relation is suggested by the scatter of points. Thus, the added-variable plot for X_2 in Figure 10.4d complements the regular residual plot in Figure 10.4c by indicating the potential usefulness of thigh circumference (X_2) in the regression model when triceps skinfold thickness (X_1) is already in the model. This information is consistent with the t^* statistic for b_2 of 2.26 in Table 7.2c and the moderate coefficient of partial determination of $R^2_{Y2|1} = .232$. Finally, the added-variable plot in Figure 10.4d reveals the presence of one potentially influential case (case 3) in the lower left corner. The influence of this case will be investigated in greater detail in Section 10.4.

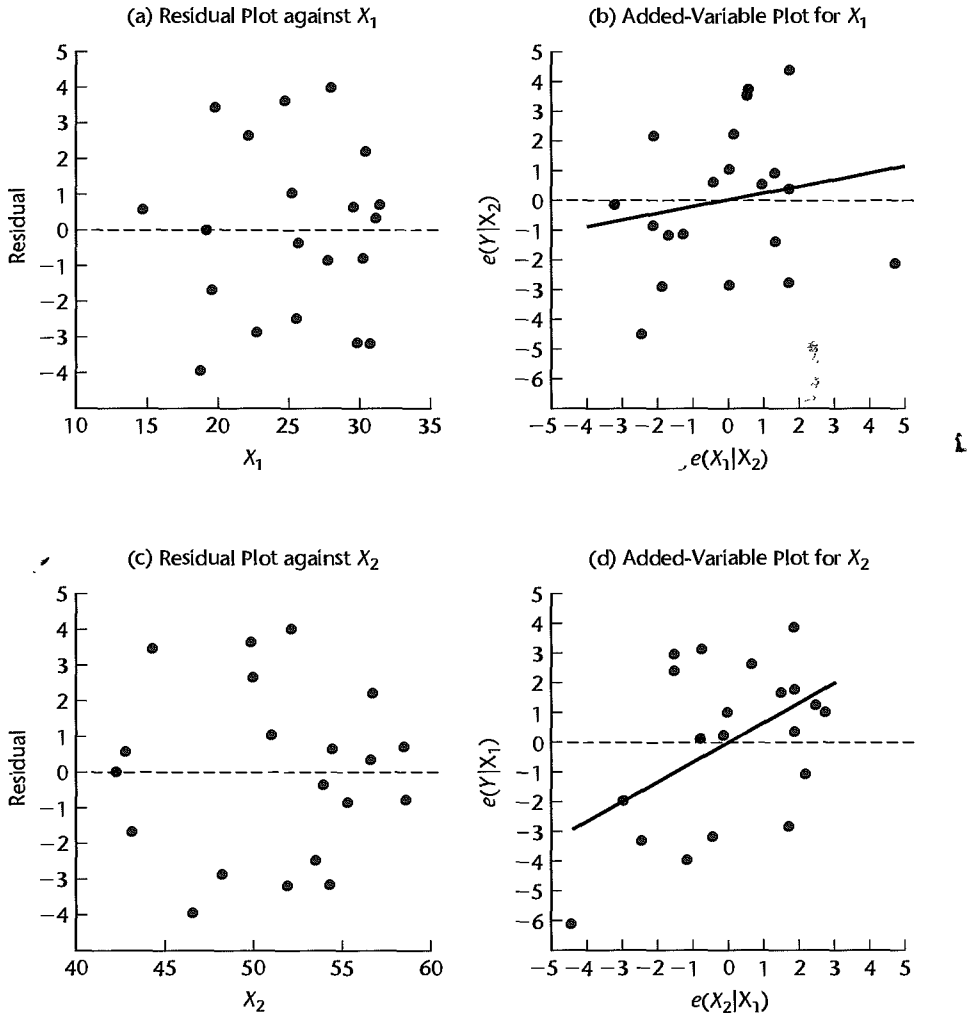


FIGURE 10.4 Residual Plots and Added-Variable Plots—Body Fat Example with Two Predictor Variables.

Comments

1. An added-variable plot only suggests the nature of the functional relation in which a predictor variable should be added to the regression model but does not provide an analytic expression of the relation. Furthermore, the relation shown is for X_k adjusted for the other predictor variables in the regression model, not for X_k directly. Hence, a variety of transformations or curvature effect terms may need to be investigated and additional residual plots utilized to identify the best transformation or curvature effect terms.

2. Added-variable plots need to be used with caution for identifying the nature of the marginal effect of a predictor variable. These plots may not show the proper form of the marginal effect of a predictor variable if the functional relations for some or all of the predictor variables already in the regression model are misspecified. For example, if X_2 and X_3 are related in a curvilinear fashion to the response variable but the regression model uses linear terms only, the added-variable plots for X_2

and X_3 may not show the proper relationships to the response variable, especially when the predictor variables are correlated. Since added-variable plots for the several predictor variables are all concerned with marginal effects only, they may therefore not be effective when the relations of the predictor variables to the response variable are complex. Also, added-variable plots may not detect interaction effects that are present. Finally, high multicollinearity among the predictor variables may cause the added-variable plots to show an improper functional relation for the marginal effect of a predictor variable.

3. When several added-variable plots are required for a set of predictor variables, it is not necessary to fit entirely new regression models each time. Computational procedures are available that economize on the calculations required; these are explained in specialized texts such as Reference 10.1.

4. Any fitted multiple regression function can be obtained from a sequence of fitted partial regressions. To illustrate this, consider again the life insurance example, where the fitted regression of Y on X_2 is given in (10.4a) and the fitted regression of X_1 on X_2 is given in (10.4b). If we now regress the residuals $e(Y|X_2) = Y - \hat{Y}(X_2)$ on the residuals $e(X_1|X_2) = X_1 - \hat{X}_1(X_2)$, using regression through the origin, we obtain (calculations not shown):

$$\widehat{e(Y|X_2)} = 6.2880[e(X_1|X_2)] \tag{10.5}$$

By simple substitution, using (10.4a) and (10.4b), we obtain:

$$[\hat{Y} - (50.70 + 15.54X_2)] = 6.2880[X_1 - (40.779 + 1.718X_2)]$$

or:

$$\hat{Y} = -205.72 + 6.2880X_1 + 4.737X_2 \tag{10.6}$$

where the solution for Y is the fitted value $\hat{Y}$ when X_1 and X_2 are included in the regression model. Note that the fitted regression function in (10.6) is the same as when the regression model was fitted to X_1 and X_2 directly in (10.3), except for a minor difference due to rounding effects.

5. A residual plot closely related to the added-variable plot is the *partial residual plot*. This plot also is used as an aid for identifying the nature of the relationship for a predictor variable X_k under consideration for addition to the regression model. The partial residual plot takes as the starting point the usual residuals $e_i = Y_i - \hat{Y}_i$ when the model including X_k is fitted, to which the regression effect for X_k is added. Specifically, the partial residuals for examining the effect of predictor variable X_k, denoted by $p_i(X_k)$, are defined as follows:

$$p_i(X_k) = e_i + b_k X_{ik} \tag{10.7}$$

Thus, for a partial residual, we add the effect of X_k, as reflected by the fitted model term $b_k X_{ik}$, back onto the residual. A plot of these partial residuals against X_k is referred to as a partial residual plot. The reader is referred to References 10.2 and 10.3 for more details on partial residual plots. ■

10.2 Identifying Outlying Y Observations—Studentized Deleted Residuals

Outlying Cases

Frequently in regression analysis applications, the data set contains some cases that are outlying or extreme; that is, the observations for these cases are well separated from the remainder of the data. These outlying cases may involve large residuals and often have dramatic effects on the fitted least squares regression function. It is therefore important to

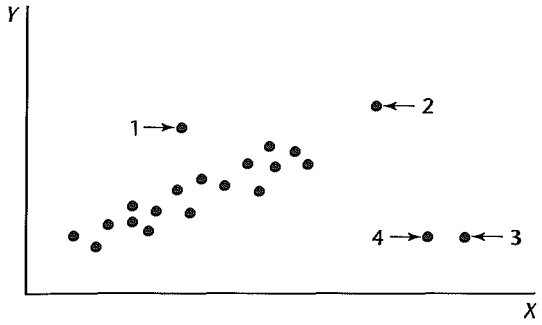


FIGURE 10.5
Scatter Plot for Regression with One Predictor Variable Illustrating Outlying Cases.

study the outlying cases carefully and decide whether they should be retained or eliminated, and if retained, whether their influence should be reduced in the fitting process and/or the regression model should be revised.

A case may be outlying or extreme with respect to its Y value, its X value(s), or both. Figure 10.5 illustrates this for the case of regression with a single predictor variable. In the scatter plot in Figure 10.5, case 1 is outlying with respect to its Y value, given X. Note that this point falls far outside the scatter, although its X value is near the middle of the range of observations on the predictor variable. Cases 2, 3, and 4 are outlying with respect to their X values since they have much larger X values than those for the other cases; cases 3 and 4 are also outlying with respect to their Y values, given X.

Not all outlying cases have a strong influence on the fitted regression function. Case 1 in Figure 10.5 may not be too influential because a number of other cases have similar X values that will keep the fitted regression function from being displaced too far by the outlying case. Likewise, case 2 may not be too influential because its Y value is consistent with the regression relation displayed by the nonextreme cases. Cases 3 and 4, on the other hand, are likely to be very influential in affecting the fit of the regression function. They are outlying with regard to their X values, and their Y values are not consistent with the regression relation for the other cases.

A basic step in any regression analysis is to determine if the regression model under consideration is heavily influenced by one or a few cases in the data set. For regression with one or two predictor variables, it is relatively simple to identify outlying cases with respect to their X or Y values by means of box plots, stem-and-leaf plots, scatter plots, and residual plots, and to study whether they are influential in affecting the fitted regression function. When more than two predictor variables are included in the regression model, however, the identification of outlying cases by simple graphic means becomes difficult because single-variable or two-variable examinations do not necessarily help find outliers relative to a multivariable regression model. Some univariate outliers may not be extreme in a multiple regression model, and, conversely, some multivariable outliers may not be detectable in single-variable or two-variable analyses.

We now discuss the use of some refined measures for identifying cases with outlying Y observations. In the following section we take up the identification of cases that are multivariable outliers with respect to their X values.

Residuals and Semistudentized Residuals

The detection of outlying or extreme Y observations based on an examination of the residuals has been considered in earlier chapters. We utilized there either the residual e_i:

$$e_i = Y_i - \hat{Y}_i \tag{10.8}$$

or the semistudentized residuals e_i^*:

$$e_i^* = \frac{e_i}{\sqrt{MSE}} \tag{10.9}$$

We introduce now two refinements to make the analysis of residuals more effective for identifying outlying Y observations. These refinements require the use of the hat matrix, which we encountered in Chapters 5 and 6.

Hat Matrix

The hat matrix was defined in (6.30a):

$$\underset{n\times n}{\mathbf{H}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{10.10}$$

We noted in (6.30) that the fitted values $\hat{Y}_i$ can be expressed as linear combinations of the observations Y_i through the hat matrix:

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \tag{10.11}$$

and similarly we noted in (6.31) that the residuals e_i can also be expressed as linear combinations of the observations Y_i by means of the hat matrix:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \tag{10.12}$$

Further, we noted in (6.32) that the variance-covariance matrix of the residuals involves the hat matrix:

$$\boldsymbol{\sigma}^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H}) \tag{10.13}$$

Therefore, the variance of residual e_i, denoted by $\sigma^2\{e_i\}$, is:

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}) \tag{10.14}$$

where h_{ii} is the ith element on the main diagonal of the hat matrix, and the covariance between residuals e_i and e_j $(i \neq j)$ is:

$$\sigma\{e_i, e_j\} = \sigma^2(0 - h_{ij}) = -h_{ij}\sigma^2 \qquad i \neq j \tag{10.15}$$

where h_{ij} is the element in the ith row and jth column of the hat matrix.

These variances and covariances are estimated by using MSE as the estimator of the error variance σ^2:

$$s^2\{e_i\} = MSE(1 - h_{ii}) \tag{10.16a}$$

$$s\{e_i, e_j\} = -h_{ij}(MSE) \qquad i \neq j \tag{10.16b}$$

We shall illustrate these different roles of the hat matrix by an example.

TABLE 10.2 Illustration of Hat Matrix.

(a) Data and Basic Results

i	(1) X_{i1}	(2) X_{i2}	(3) Y_i	(4) $\hat{Y}_i$	(5) e_i	(6) h_{ii}	(7) $s^2\{e_i\}$
1	14	25	301	282.2	18.8	.3877	352.0
2	19	32	327	332.3	−5.3	.9513	28.0
3	12	22	246	260.0	−14.0	.6614	194.6
4	11	15	187	186.5	.5	.9996	.2

(b) H

$$\begin{bmatrix} .3877 & .1727 & .4553 & -.0157 \\ .1727 & .9513 & -.1284 & .0044 \\ .4553 & -.1284 & .6614 & .0117 \\ -.0157 & .0044 & .0117 & .9996 \end{bmatrix}$$

(c) $s^2\{e\}$

$$\begin{bmatrix} 352.0 & -99.3 & -261.8 & 9.0 \\ -99.3 & 28.0 & 73.8 & -2.5 \\ -261.8 & 73.8 & 194.6 & -6.7 \\ 9.0 & -2.5 & -6.7 & .2 \end{bmatrix}$$

Example

A small data set based on $n = 4$ cases for examining the regression relation between a response variable Y and two predictor variables X_1 and X_2 is shown in Table 10.2a, columns 1–3. The fitted first-order model and the error mean square are:

$$\hat{Y} = 80.93 - 5.84X_1 + 11.32X_2 \qquad MSE = 574.9 \tag{10.17}$$

The fitted values and the residuals for the four cases are shown in columns 4 and 5 of Table 10.2a.

The hat matrix for these data is shown in Table 10.2b. It was obtained by means of (10.10) for the **X** matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 14 & 25 \\ 1 & 19 & 32 \\ 1 & 12 & 22 \\ 1 & 11 & 15 \end{bmatrix}$$

Note from (10.10) that the hat matrix is solely a function of the predictor variable(s). Also note from Table 10.2b that the hat matrix is symmetric. The diagonal elements h_{ii} of the hat matrix are repeated in column 6 of Table 10.2a.

We illustrate that the fitted values are linear combinations of the Y values by calculating $\hat{Y}_1$ by means of (10.11):

$$\begin{aligned} \hat{Y}_1 &= h_{11}Y_1 + h_{12}Y_2 + h_{13}Y_3 + h_{14}Y_4 \\ &= .3877(301) + .1727(327) + .4553(246) - .0157(187) \\ &= 282.2 \end{aligned}$$

This is the same result, except for possible rounding effects, as obtained from the fitted regression function (10.17):

$$\hat{Y}_1 = 80.93 - 5.84(14) + 11.32(25) = 282.2$$

The estimated variance-covariance matrix of the residuals, $\mathbf{s}^2\{\mathbf{e}\} = MSE(\mathbf{I} - \mathbf{H})$, is shown in Table 10.2c. It was obtained by using $MSE = 574.9$. The estimated variances of the residuals are shown in the main diagonal of the variance-covariance matrix in Table 10.2c and are repeated in column 7 of Table 10.2a. We illustrate their direct calculation for case 1 by using (10.16a):

$$s^2\{e_1\} = 574.9(1 - .3877) = 352.0$$

We see from Table 10.2a, column 7, that the residuals do not have constant variance. In fact, the variances differ greatly here because the data set is so small. As we shall note in Section 10.3, residuals for cases that are outlying with respect to the X variables have smaller variances.

Note also that the covariances in the matrix in Table 10.2c are not zero; hence, pairs of residuals are correlated, some positively and some negatively. We noted this correlation in Chapter 3, but also pointed out there that the correlations become very small for larger data sets.

Comment

The diagonal element h_{ii} of the hat matrix can be obtained directly from:

$$h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i \tag{10.18}$$

where:

$$\underset{p\times 1}{\mathbf{X}_i} = \begin{bmatrix} 1 \\ X_{i,1} \\ \vdots \\ X_{i,p-1} \end{bmatrix} \tag{10.18a}$$

Note that $\mathbf{X}_i$ corresponds to the $\mathbf{X}_h$ vector in (6.53) except that $\mathbf{X}_i$ pertains to the ith case, and that $\mathbf{X}_i'$ is simply the ith row of the $\mathbf{X}$ matrix, pertaining to the ith case. ■

Studentized Residuals

The first refinement in making residuals more effective for detecting outlying Y observations involves recognition of the fact that the residuals e_i may have substantially different variances $\sigma^2\{e_i\}$. It is therefore appropriate to consider the magnitude of each e_i relative to its estimated standard deviation to give recognition to differences in the sampling errors of the residuals. We see from (10.16a) that an estimator of the standard deviation of e_i is:

$$s\{e_i\} = \sqrt{MSE(1 - h_{ii})} \tag{10.19}$$

The ratio of e_i to $s\{e_i\}$ is called the *studentized residual* and will be denoted by r_i:

$$r_i = \frac{e_i}{s\{e_i\}} \tag{10.20}$$

While the residuals e_i will have substantially different sampling variations if their standard deviations differ markedly, the studentized residuals r_i have constant variance (when the model is appropriate). Studentized residuals often are called *internally studentized residuals*.

Deleted Residuals

The second refinement to make residuals more effective for detecting outlying Y observations is to measure the ith residual $e_i = Y_i - \hat{Y}_i$ when the fitted regression is based on all of the cases except the ith one. The reason for this refinement is that if Y_i is far outlying, the fitted least squares regression function based on all cases including the ith one may be influenced to come close to Y_i, yielding a fitted value $\hat{Y}_i$ near Y_i. In that event, the residual e_i will be small and will not disclose that Y_i is outlying. On the other hand, if the ith case is excluded before the regression function is fitted, the least squares fitted value $\hat{Y}_i$ is not influenced by the outlying Y_i observation, and the residual for the ith case will then tend to be larger and therefore more likely to disclose the outlying Y observation.

The procedure then is to delete the ith case, fit the regression function to the remaining $n - 1$ cases, and obtain the point estimate of the expected value when the X levels are those of the ith case, to be denoted by $\hat{Y}_{i(i)}$. The difference between the actual observed value Y_i and the estimated expected value $\hat{Y}_{i(i)}$ will be denoted by d_i:

$$d_i = Y_i - \hat{Y}_{i(i)} \tag{10.21}$$

The difference d_i is called the *deleted residual* for the ith case. We encountered this same difference in (9.16), where it was called the *PRESS* prediction error for the ith case.

An algebraically equivalent expression for d_i that does not require a recomputation of the fitted regression function omitting the ith case is:

$$d_i = \frac{e_i}{1 - h_{ii}} \tag{10.21a}$$

where e_i is the ordinary residual for the ith case and h_{ii} is the ith diagonal element in the hat matrix, as given in (10.18). Note that the larger is the value h_{ii}, the larger will be the deleted residual as compared to the ordinary residual.

Thus, deleted residuals will at times identify outlying Y observations when ordinary residuals would not identify these; at other times deleted residuals lead to the same identifications as ordinary residuals.

Note that a deleted residual also corresponds to the prediction error for a new observation in the numerator of (2.35). There, we are predicting a new $n + 1$ observation from the fitted regression function based on the earlier n cases. Modifying the earlier notation for the context of deleted residuals, where $n - 1$ cases are used for predicting the "new" nth case, we can restate the result in (6.63a) to obtain the estimated variance of d_i:

$$s^2\{d_i\} = MSE_{(i)}\left(1 + \mathbf{X}_i'\left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_i\right) \tag{10.22}$$

where $\mathbf{X}_i$ is the X observations vector (10.18a) for the ith case, $MSE_{(i)}$ is the mean square error when the ith case is omitted in fitting the regression function, and $\mathbf{X}_{(i)}$ is the $\mathbf{X}$ matrix with the ith case deleted. An algebraically equivalent expression for $s^2\{d_i\}$ is:

$$s^2\{d_i\} = \frac{MSE_{(i)}}{1 - h_{ii}} \tag{10.22a}$$

It follows from (6.63) that:

$$\frac{d_i}{s\{d_i\}} \sim t(n - p - 1) \tag{10.23}$$

Remember that $n - 1$ cases are used here in predicting the ith observation; hence, the degrees of freedom are $(n - 1) - p = n - p - 1$.

Studentized Deleted Residuals

Combining the above two refinements, we utilize for diagnosis of outlying or extreme Y observations the deleted residual d_i in (10.21) and studentize it by dividing it by its estimated standard deviation given by (10.22). The *studentized deleted residual*, denoted by t_i, therefore is:

$$t_i = \frac{d_i}{s\{d_i\}} \tag{10.24}$$

It follows from (10.21a) and (10.22a) that an algebraically equivalent expression for t_i is:

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}} \tag{10.24a}$$

The studentized deleted residual t_i in (10.24) is also called an *externally studentized residual*, in contrast to the internally studentized residual r_i in (10.20). We know from (10.23) that each studentized deleted residual t_i follows the t distribution with $n - p - 1$ degrees of freedom. The t_i, however, are not independent.

Fortunately, the studentized deleted residuals t_i in (10.24) can be calculated without having to fit new regression functions each time a different case is omitted. A simple relationship exists between MSE and $MSE_{(i)}$:

$$(n - p)MSE = (n - p - 1)MSE_{(i)} + \frac{e_i^2}{1 - h_{ii}} \tag{10.25}$$

Using this relationship in (10.24a) yields the following equivalent expression for t_i:

$$t_i = e_i\left[\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2}\right]^{1/2} \tag{10.26}$$

Thus, the studentized deleted residuals t_i can be calculated from the residuals e_i, the error sum of squares SSE, and the hat matrix values h_{ii}, all for the fitted regression based on the n cases.

Test for Outliers. We identify as outlying Y observations those cases whose studentized deleted residuals are large in absolute value. In addition, we can conduct a formal test by means of the Bonferroni test procedure of whether the case with the largest absolute studentized deleted residual is an outlier. Since we do not know in advance which case will have the largest absolute value $|t_i|$, we consider the family of tests to include n tests, one for each case. If the regression model is appropriate, so that no case is outlying because of a change in the model, then each studentized deleted residual will follow the t distribution with $n - p - 1$ degrees of freedom. The appropriate Bonferroni critical value therefore is $t(1 - \alpha/2n; n - p - 1)$. Note that the test is two-sided since we are not concerned with the direction of the residuals but only with their absolute values.

Example

For the body fat example with two predictor variables (X_1, X_2), we wish to examine whether there are outlying Y observations. Table 10.3 presents the residuals e_i in column 1,

TABLE 10.3 Residuals, Diagonal Elements of the Hat Matrix, and Studentized Deleted Residuals—Body Fat Example with Two Predictor Variables.

i	(1) e_i	(2) h_{ii}	(3) t_i
1	−1.683	.201	−.730
2	3.643	.059	1.534
3	−3.176	.372	−1.656
4	−3.158	.111	−1.348
5	.000	.248	.000
6	−.361	.129	−.148
7	.716	.156	.298
8	4.015	.096	1.760
9	2.655	.115	1.117
10	−2.475	.110	−1.034
11	.336	.120	.137
12	2.226	.109	.923
13	−3.947	.178	−1.825
14	3.447	.148	1.524
15	.571	.333	.267
16	.642	.095	.258
17	−.851	.106	.344
18	−.783	.197	.335
19	−2.857	.067	−1.176
20	1.040	.050	.409

the diagonal elements h_{ii} of the hat matrix in column 2, and the studentized deleted residuals t_i in column 3. We illustrate the calculation of the studentized deleted residual for the first case. The X values for this case, given in Table 7.1, are $X_{11} = 19.5$ and $X_{12} = 43.1$. Using the fitted regression function from Table 7.2c, we obtain:

$$\hat{Y}_1 = -19.174 + .2224(19.5) + .6594(43.1) = 13.583$$

Since $Y_1 = 11.9$, the residual for this case is $e_1 = 11.9 - 13.583 = -1.683$. We also know from Table 7.2c that $SSE = 109.95$ and from Table 10.3 that $h_{11} = .201$. Hence, by (10.26), we find:

$$t_1 = -1.683\left[\frac{20 - 3 - 1}{109.95(1 - .201) - (-1.683)^2}\right]^{1/2} = -.730$$

Note from Table 10.3, column 3, that cases 3, 8, and 13 have the largest absolute studentized deleted residuals. Incidentally, consideration of the residuals e_i (shown in Table 10.3, column 1) here would have identified cases 2, 8, and 13 as the most outlying ones, but not case 3.

We would like to test whether case 13, which has the largest absolute studentized deleted residual, is an outlier resulting from a change in the model. We shall use the Bonferroni simultaneous test procedure with a family significance level of $\alpha = .10$. We therefore require:

$$t(1 - \alpha/2n; n - p - 1) = t(.9975; 16) = 3.252$$

Since $|t_{13}| = 1.825 \leq 3.252$, we conclude that case 13 is not an outlier. Still, we might wish to investigate whether case 13 and perhaps a few other outlying cases are influential in determining the fitted regression function because the Bonferroni procedure provides a very conservative test for the presence of an outlier.

10.3 Identifying Outlying X Observations—Hat Matrix Leverage Values

Use of Hat Matrix for Identifying Outlying X Observations

The hat matrix, as we saw, plays an important role in determining the magnitude of a studentized deleted residual and therefore in identifying outlying Y observations. The hat matrix also is helpful in directly identifying outlying X observations. In particular, the diagonal elements of the hat matrix are a useful indicator in a multivariable setting of whether or not a case is outlying with respect to its X values.

The diagonal elements h_{ii} of the hat matrix have some useful properties. In particular, their values are always between 0 and 1 and their sum is p:

$$0 \leq h_{ii} \leq 1 \qquad \sum_{i=1}^{n} h_{ii} = p \tag{10.27}$$

where p is the number of regression parameters in the regression function including the intercept term. In addition, it can be shown that h_{ii} is a measure of the distance between the X values for the ith case and the means of the X values for all n cases. Thus, a large value h_{ii} indicates that the ith case is distant from the center of all X observations. The diagonal element h_{ii} in this context is called the *leverage* (in terms of the X values) of the ith case.

Figure 10.6 illustrates the role of the leverage values h_{ii} as distance measures for our earlier example in Table 10.2. Figure 10.6 shows a scatter plot of X_2 against X_1 for the four cases, and the center of the four cases located at $(\bar{X}_1, \bar{X}_2)$. This center is called the *centroid*. Here, the centroid is $(\bar{X}_1 = 14.0, \bar{X}_2 = 23.5)$. In addition, Figure 10.6 shows the leverage value for each case. Note that cases 1 and 3, which are closest to the centroid, have the smallest leverage values, while cases 2 and 4, which are farthest from the center, have the largest leverage values. Note also that the four leverage values sum to $p = 3$.

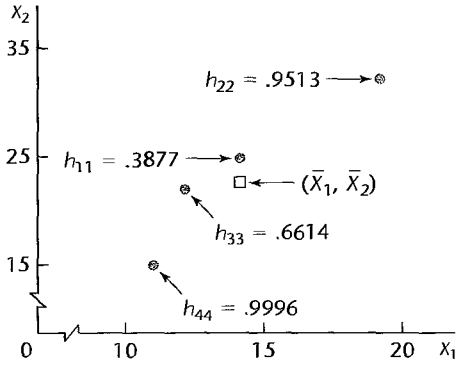


FIGURE 10.6 Illustration of Leverage Values as Distance Measures—Table 10.2 Example.

If the ith case is outlying in terms of its X observations and therefore has a large leverage value h_{ii}, it exercises substantial leverage in determining the fitted value $\hat{Y}_i$. This is so for the following reasons:

1. The fitted value $\hat{Y}_i$ is a linear combination of the observed Y values, as shown by (10.11), and h_{ii} is the weight of observation Y_i in determining this fitted value. Thus, the larger is h_{ii}, the more important is Y_i in determining $\hat{Y}_i$. Remember that h_{ii} is a function only of the X values, so h_{ii} measures the role of the X values in determining how important Y_i is in affecting the fitted value $\hat{Y}_i$.

2. The larger is h_{ii}, the smaller is the variance of the residual e_i, as we noted earlier from (10.14). Hence, the larger is h_{ii}, the closer the fitted value $\hat{Y}_i$ will tend to be to the observed value Y_i. In the extreme case where $h_{ii} = 1$, the variance $\hat{\sigma}^2\{e_i\}$ equals 0, so the fitted value $\hat{Y}_i$ is then forced to equal the observed value Y_i.

A leverage value h_{ii} is usually considered to be large if it is more than twice as large as the mean leverage value, denoted by $\bar{h}$, which according to (10.27) is:

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n} \tag{10.28}$$

Hence, leverage values greater than $2p/n$ are considered by this rule to indicate outlying cases with regard to their X values. Another suggested guideline is that h_{ii} values exceeding .5 indicate very high leverage, whereas those between .2 and .5 indicate moderate leverage. Additional evidence of an outlying case is the existence of a gap between the leverage values for most of the cases and the unusually large leverage value(s).

The rules just mentioned for identifying cases that are outlying with respect to their X values are intended for data sets that are reasonably large, relative to the number of parameters in the regression function. They are not applicable, for instance, to the simple example in Table 10.2 where there are $n = 4$ cases and $p = 3$ parameters in the regression function. Here, the mean leverage value is $3/4 = .75$, and one cannot obtain a leverage value twice as large as the mean value since leverage values cannot exceed 1.0.

Example

We continue with the body fat example of Table 7.1. We again use only the two predictor variables triceps skinfold thickness (X_1) and thigh circumference (X_2) so that the results using the hat matrix can be compared to simple graphic plots. Figure 10.7 contains a scatter

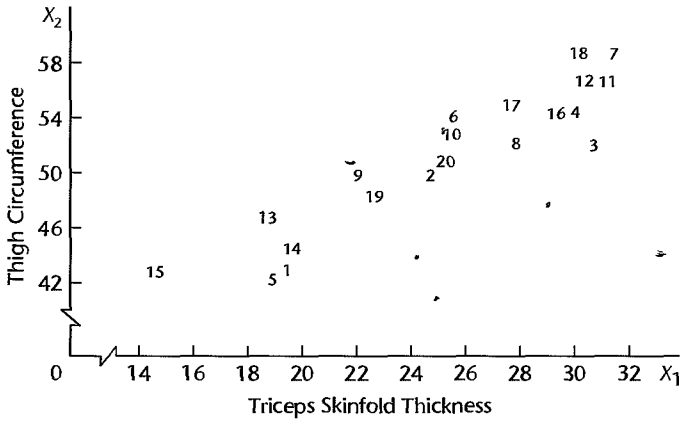


FIGURE 10.7 Scatter Plot of Thigh Circumference against Triceps Skinfold Thickness—Body Fat Example with Two Predictor Variables.

plot of X_2 against X_1, where the data points are identified by their case number. We note from Figure 10.7 that cases 15 and 3 appear to be outlying ones with respect to the pattern of the X values. Case 15 is outlying for X_1 and at the low end of the range for X_2, whereas case 3 is outlying in terms of the pattern of multicollinearity, though it is not outlying for either of the predictor variables separately. Cases 1 and 5 also appear to be somewhat extreme.

Table 10.3, column 2, contains the leverage values h_{ii} for the body fat example. Note that the two largest leverage values are $h_{33} = .372$ and $h_{15,15} = .333$. Both exceed the criterion of twice the mean leverage value, $2p/n = 2(3)/20 = .30$, and both are separated by a substantial gap from the next largest leverage values, $h_{55} = .248$ and $h_{11} = .201$. Having identified cases 3 and 15 as outlying in terms of their X values, we shall need to ascertain how influential these cases are in the fitting of the regression function.

Use of Hat Matrix to Identify Hidden Extrapolation

We have seen that the hat matrix is useful in the model-building stage for identifying cases that are outlying with respect to their X values and, that, therefore, may be influential in affecting the fitted model. The hat matrix is also useful after the model has been selected and fitted for determining whether an inference for a mean response or a new observation involves a substantial extrapolation beyond the range of the data. When there are only two predictor variables, it is easy to see from a scatter plot of X_2 against X_1 whether an inference for a particular (X_1, X_2) set of values is outlying beyond the range of the data, such as from Figure 10.7. This simple graphic analysis is no longer available with larger numbers of predictor variables, where extrapolations may be hidden.

To spot hidden extrapolations, we can utilize the direct leverage calculation in (10.18) for the new set of X values for which inferences are to be made:

$$h_{\text{new,new}} = \mathbf{X}'_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{\text{new}} \tag{10.29}$$

where $\mathbf{X}_{\text{new}}$ is the vector containing the X values for which an inference about a mean response or a new observation is to be made, and the $\mathbf{X}$ matrix is the one based on the data set used for fitting the regression model. If $h_{\text{new,new}}$ is well within the range of leverage values h_{ii} for the cases in the data set, no extrapolation is involved. On the other hand, if $h_{\text{new,new}}$ is much larger than the leverage values for the cases in the data set, an extrapolation is indicated.

10.4 Identifying Influential Cases—*DFFITS*, Cook's Distance, and *DFBETAS* Measures

After identifying cases that are outlying with respect to their Y values and/or their X values, the next step is to ascertain whether or not these outlying cases are influential. We shall consider a case to be *influential* if its exclusion causes major changes in the fitted regression function. As noted in Figure 10.5, not all outlying cases need be influential. For example, case 1 in Figure 10.5 may not affect the fitted regression function to any substantial extent.

We take up three measures of influence that are widely used in practice, each based on the omission of a single case to measure its influence.

Influence on Single Fitted Value—*DFFITS*

A useful measure of the influence that case i has on the fitted value $\hat{Y}_i$ is given by:

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} \tag{10.30}$$

The letters *DF* stand for the difference between the fitted value $\hat{Y}_i$ for the ith case when all n cases are used in fitting the regression function and the predicted value $\hat{Y}_{i(i)}$ for the ith case obtained when the ith case is omitted in fitting the regression function. The denominator of (10.30) is the estimated standard deviation of $\hat{Y}_i$, but it uses the error mean square when the ith case is omitted in fitting the regression function for estimating the error variance σ^2. The denominator provides a standardization so that the value $(DFFITS)_i$ for the ith case represents the number of estimated standard deviations of $\hat{Y}_i$ that the fitted value $\hat{Y}_i$ increases or decreases with the inclusion of the ith case in fitting the regression model.

It can be shown that the *DFFITS* values can be computed by using only the results from fitting the entire data set, as follows:

$$(DFFITS)_i = e_i\left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right]^{1/2}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} = t_i\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \tag{10.30a}$$

Note from the last expression that the *DFFITS* value for the ith case is a studentized deleted residual, as given in (10.26), increased or decreased by a factor that is a function of the leverage value for this case. If case i is an X outlier and has a high leverage value, this factor will be greater than 1 and $(DFFITS)_i$ will tend to be large absolutely.

As a guideline for identifying influential cases, we suggest considering a case influential if the absolute value of *DFFITS* exceeds 1 for small to medium data sets and $2\sqrt{p/n}$ for large data sets.

Example

Table 10.4, column 1, lists the *DFFITS* values for the body fat example with two predictor variables. To illustrate the calculations, consider the *DFFITS* value for case 3, which was identified as outlying with respect to its X values. From Table 10.3, we know that the studentized deleted residual for this case is $t_3 = -1.656$ and the leverage value is $h_{33} = .372$. Hence, using (10.30a) we obtain:

$$(DFFITS)_3 = -1.656\left(\frac{.372}{1-.372}\right)^{1/2} = -1.27$$

The only *DFFITS* value in Table 10.4 that exceeds our guideline for a medium-size data set is for case 3, where $|(DFFITS)_3| = 1.273$. This value is somewhat larger than our guideline of 1. However, the value is close enough to 1 that the case may not be influential enough to require remedial action.

Comment

The estimated variance of $\hat{Y}_i$ used in the denominator of (10.30) is developed from the relation $\hat{\mathbf{Y}} = \mathbf{HY}$ in (10.11). Using (5.46), we obtain:

$$\sigma^2\{\hat{\mathbf{Y}}\} = \mathbf{H}\sigma^2\{\mathbf{Y}\}\mathbf{H}' = \mathbf{H}(\sigma^2\mathbf{I})\mathbf{H}'$$

TABLE 10.4 *DFFITS*, Cook's Distances, and *DFBETAS*—Body Fat Example with Two Predictor Variables.

	(1)	(2)	(3)	(4)	(5)
				DFBETAS	
i	$(DFFITS)_i$	D_i	b_0	b_1	b_2
1	−.366	.046	−.305	−.132	.232
2	.384	.046	.173	.115	−.143
3	−1.273	.490	−.847	−1.183	1.067
4	−.476	.072	−.102	−.294	.196
5	.000	.000	.000	.000	.000
6	−.057	.001	.040	.040	−.044
7	.128	.006	−.078	−.016	.054
8	.575	.098	.261	.391	−.333
9	.402	.053	−.151	−.295	.247
10	−.364	.044	.238	.245	−.269
11	.051	.001	−.009	.017	−.003
12	.323	.035	−.131	.023	.070
13	−.851	.212	.119	.592	−.390
14	.636	.125	.452	.113	−.298
15	.189	.013	−.003	−.125	.069
16	.084	.002	.009	.043	−.025
17	−.118	.005	.080	.055	−.076
18	−.166	.010	.132	.075	−.116
19	−.315	.032	−.130	−.004	.064
20	.094	.003	.010	.002	−.003

Since $\mathbf{H}$ is a symmetric matrix, so $\mathbf{H}' = \mathbf{H}$, and it is also idempotent, so $\mathbf{HH} = \mathbf{H}$, we obtain:

$$\sigma^2\{\hat{\mathbf{Y}}\} = \sigma^2\mathbf{H} \tag{10.31}$$

Hence, the variance of $\hat{Y}_i$ is:

$$\sigma^2\{\hat{Y}_i\} = \sigma^2 h_{ii} \tag{10.32}$$

where h_{ii} is the ith diagonal element of the hat matrix. The error term variance σ^2 is estimated in (10.30) by the error mean square $MSE_{(i)}$ obtained when the ith case is omitted in fitting the regression model. ■

Influence on All Fitted Values—Cook's Distance

In contrast to the *DFFITS* measure in (10.30), which considers the influence of the ith case on the fitted value $\hat{Y}_i$ for this case, Cook's distance measure considers the influence of the ith case on all n fitted values. Cook's distance measure, denoted by D_i, is an aggregate influence measure, showing the effect of the ith case on all n fitted values:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE} \tag{10.33}$$

Note that the numerator involves similar differences as in the *DFFITS* measure, but here each of the n fitted values $\hat{Y}_j$ is compared with the corresponding fitted value $\hat{Y}_{j(i)}$ when the ith case is deleted in fitting the regression model. These differences are then squared and summed, so that the aggregate influence of the ith case is measured without regard to the signs of the effects. Finally, the denominator serves as a standardizing measure. In matrix

terms, Cook's distance measure can be expressed as follows:

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{pMSE} \qquad \textbf{(10.33a)}$$

Here, $\hat{\mathbf{Y}}$ as usual is the vector of the fitted values when all n cases are used for the regression fit and $\hat{\mathbf{Y}}_{(i)}$ is the vector of the fitted values when the ith case is deleted.

For interpreting Cook's distance measure, it has been found useful to relate D_i to the $F(p, n-p)$ distribution and ascertain the corresponding percentile value. If the percentile value is less than about 10 or 20 percent, the ith case has little apparent influence on the fitted values. If, on the other hand, the percentile value is near 50 percent or more, the fitted values obtained with and without the ith case should be considered to differ substantially, implying that the ith case has a major influence on the fit of the regression function.

Fortunately, Cook's distance measure D_i can be calculated without fitting a new regression function each time a different case is deleted. An algebraically equivalent expression is:

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \qquad \textbf{(10.33b)}$$

Note from (10.33b) that D_i depends on two factors: (1) the size of the residual e_i and (2) the leverage value h_{ii}. The larger either e_i or h_{ii} is, the larger D_i is. Thus, the ith case can be influential: (1) by having a large residual e_i and only a moderate leverage value h_{ii}, or (2) by having a large leverage value h_{ii} with only a moderately sized residual e_i, or (3) by having both a large residual e_i and a large leverage value h_{ii}.

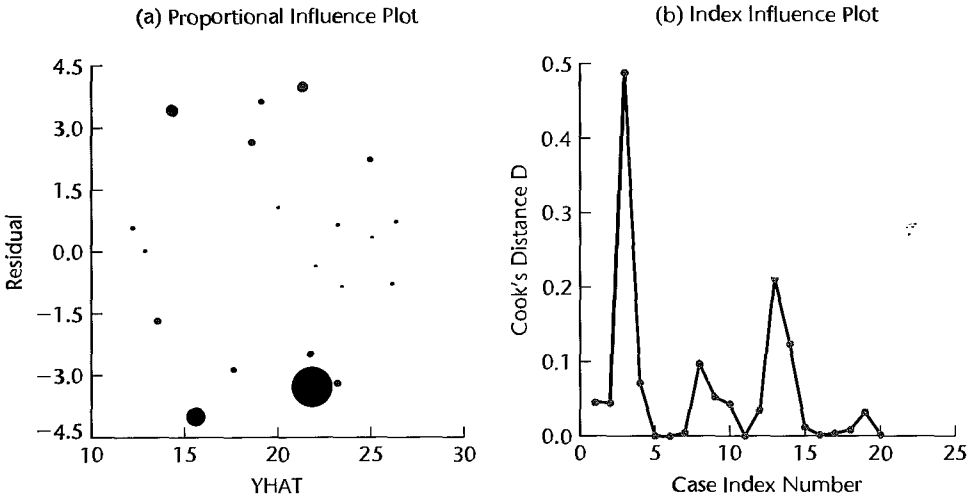
Example

For the body fat example with two predictor variables, Table 10.4, column 2, presents the D_i values. To illustrate the calculations, we consider again case 3, which is outlying with regard to its X values. We know from Table 10.3 that $e_3 = -3.176$ and $h_{33} = .372$. Further, $MSE = 6.47$ according to Table 7.2c and $p = 3$ for the model with two predictor variables. Hence, we obtain:

$$D_3 = \frac{(-3.176)^2}{3(6.47)}\left[\frac{.372}{(1-.372)^2}\right] = .490$$

We note from Table 10.4, column 2 that case 3 clearly has the largest D_i value, with the next largest distance measure $D_{13} = .212$ being substantially smaller. Figure 10.8 presents the information provided by Cook's distance measure about the influence of each case in two different plots. Shown in Figure 10.8a is a proportional influence plot of the residuals e_i against the corresponding fitted values $\hat{Y}_i$, the size of the plotted points being proportional to Cook's distance measure D_i. Figure 10.8b presents the information about the Cook's distance measures in the form of an index influence plot, where Cook's distance measure D_i is plotted against the corresponding case index i. Both plots in Figure 10.8 clearly show that one case stands out as most influential (case 3) and that all the other cases are much less influential. The proportional influence plot in Figure 10.8a shows that the residual for the most influential case is large negative, but does not identify the case. The index influence plot in Figure 10.8b, on the other hand, identifies the most influential case as case 3 but does not provide any information about the magnitude of the residual for this case.

FIGURE 10.8 Proportional Influence Plot (Points Porportional in Size to Cook's Distance Measure) and Index Influence Plot—Body Fat Example with Two Predictor Variables.



To assess the magnitude of the influence of case 3 ($D_3 = .490$), we refer to the corresponding F distribution, namely, $F(p, n - p) = F(3, 17)$. We find that .490 is the 30.6th percentile of this distribution. Hence, it appears that case 3 does influence the regression fit, but the extent of the influence may not be large enough to call for consideration of remedial measures.

Influence on the Regression Coefficients—*DFBETAS*

A measure of the influence of the ith case on each regression coefficient b_k ($k = 0, 1, \ldots, p - 1$) is the difference between the estimated regression coefficient b_k based on all n cases and the regression coefficient obtained when the ith case is omitted, to be denoted by $b_{k(i)}$. When this difference is divided by an estimate of the standard deviation of b_k, we obtain the measure *DFBETAS*:

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}} \qquad k = 0, 1, \ldots, p - 1 \tag{10.34}$$

where c_{kk} is the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Recall from (6.46) that the variance-covariance matrix of the regression coefficients is given by $\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Hence the variance of b_k is:

$$\sigma^2\{b_k\} = \sigma^2 c_{kk} \tag{10.35}$$

The error term variance σ^2 here is estimated by $MSE_{(i)}$, the error mean square obtained when the ith case is deleted in fitting the regression model.

The *DFBETAS* value by its sign indicates whether inclusion of a case leads to an increase or a decrease in the estimated regression coefficient, and its absolute magnitude shows the size of the difference relative to the estimated standard deviation of the regression coefficient. A large absolute value of $(DFBETAS)_{k(i)}$ is indicative of a large impact of the

ith case on the kth regression coefficient. As a guideline for identifying influential cases, we recommend considering a case influential if the absolute value of *DFBETAS* exceeds 1 for small to medium data sets and $2/\sqrt{n}$ for large data sets.

Example

For the body fat example with two predictor variables, Table 10.4 lists the *DFBETAS* values in columns 3, 4, and 5. Note that case 3, which is outlying with respect to its X values, is the only case that exceeds our guideline of 1 for medium-size data sets for both b_1 and b_2. Thus, case 3 is again tagged as potentially influential. Again, however, the *DFBETAS* values do not exceed 1 by very much so that case 3 may not be so influential as to require remedial action.

Comment

Cook's distance measure of the aggregate influence of a case on the n fitted values, which was defined in (10.33), is algebraically equivalent to a measure of the aggregate influence of a case on the p regression coefficients. In fact, Cook's distance measure was originally derived from the concept of a confidence region for all p regression coefficients β_k ($k = 0, 1, \ldots, p-1$) simultaneously. It can be shown that the boundary of this joint confidence region for the normal error multiple regression model (6.19) is given by:

$$\frac{(\mathbf{b}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})}{pMSE} = F(1-\alpha; p, n-p) \tag{10.36}$$

Cook's distance measure D_i uses the same structure for measuring the combined impact of the ith case on the differences in the estimated regression coefficients:

$$D_i = \frac{\left(\mathbf{b}-\mathbf{b}_{(i)}\right)'\mathbf{X}'\mathbf{X}\left(\mathbf{b}-\mathbf{b}_{(i)}\right)}{pMSE} \tag{10.37}$$

where $\mathbf{b}_{(i)}$ is the vector of the estimated regression coefficients obtained when the ith case is omitted and $\mathbf{b}$, as usual, is the vector when all n cases are used. The expressions for Cook's distance measure in (10.33a) and (10.37) are algebraically identical. ■

Influence on Inferences

To round out the determination of influential cases, it is usually a good idea to examine in a direct fashion the inferences from the fitted regression model that would be made with and without the case(s) of concern. If the inferences are not essentially changed, there is little need to think of remedial actions for the cases diagnosed as influential. On the other hand, serious changes in the inferences drawn from the fitted model when a case is omitted will require consideration of remedial measures.

Example

In the body fat example with two predictor variables, cases 3 and 15 were identified as outlying X observations and cases 8 and 13 as outlying Y observations. All three influence measures (*DFFITS*, Cook's distance, and *DFBETAS*) identified only case 3 as influential, and, indeed, suggested that its influence may be of marginal importance so that remedial measures might not be required.

The analyst in the body fat example was primarily interested in the fit of the regression model because the model was intended to be used for making predictions within the range of the observations on the predictor variables in the data set. Hence, the analyst considered

the fitted regression functions with and without case 3:

$$\text{With case 3: } \hat{Y} = -19.174 + .2224X_1 + .6594X_2$$

$$\text{Without case 3: } \hat{Y} = -12.428 + .5641X_1 + .3635X_2$$

Because of the high multicollinearity between X_1 and X_2, the analyst was not surprised by the shifts in the magnitudes of b_1 and b_2 when case 3 is omitted. Remember that the estimated standard deviations of the coefficients, given in Table 7.2c, are very large and that a single case can change the estimated coefficients substantially when the predictor variables are highly correlated.

To examine the effect of case 3 on inferences to be made from the fitted regression function in the range of the X observations in a direct fashion, the analyst calculated for each of the 20 cases the relative difference between the fitted value $\hat{Y}_i$ based on all 20 cases and the fitted value $\hat{Y}_{i(3)}$ obtained when case 3 is omitted. The measure of interest was the average absolute percent difference:

$$\frac{\sum_{i=1}^{n} \left| \frac{\hat{Y}_{i(3)} - \hat{Y}_i}{\hat{Y}_i} \right| 100}{n}$$

This mean difference is 3.1 percent; further, 17 of the 20 differences are less than 5 percent (calculations not shown). On the basis of this direct evidence about the effect of case 3 on the inferences to be made, the analyst was satisfied that case 3 does not exercise undue influence so that no remedial action is required for handling this case.

Some Final Comments

Analysis of outlying and influential cases is a necessary component of good regression analysis. However, it is neither automatic nor foolproof and requires good judgment by the analyst. The methods described often work well, but at times are ineffective. For example, if two influential outlying cases are nearly coincident, as depicted in Figure 10.5 by cases 3 and 4, an analysis that deletes one case at a time and estimates the change in fit will result in virtually no change for these two outlying cases. The reason is that the retained outlying case will mask the effect of the deleted outlying case. Extensions of the single-case diagnostic procedures described here have been developed that involve deleting two or more cases at a time. However, the computational requirements for these extensions are much more demanding than for the single-case diagnostics. Reference 10.4 describes some of these extensions.

Remedial measures for outlying cases that are determined to be highly influential by the diagnostic procedures will be discussed in the next chapter.

10.5 Multicollinearity Diagnostics—Variance Inflation Factor

When we discussed multicollinearity in Chapter 7, we noted some key problems that typically arise when the predictor variables being considered for the regression model are highly correlated among themselves:

1. Adding or deleting a predictor variable changes the regression coefficients.

2. The extra sum of squares associated with a predictor variable varies, depending upon which other predictor variables are already included in the model.

3. The estimated standard deviations of the regression coefficients become large when the predictor variables in the regression model are highly correlated with each other.

4. The estimated regression coefficients individually may not be statistically significant even though a definite statistical relation exists between the response variable and the set of predictor variables.

These problems can also arise without substantial multicollinearity being present, but only under unusual circumstances not likely to be found in practice.

We first consider some informal diagnostics for multicollinearity and then a highly useful formal diagnostic, the variance inflation factor.

Informal Diagnostics

Indications of the presence of serious multicollinearity are given by the following informal diagnostics:

1. Large changes in the estimated regression coefficients when a predictor variable is added or deleted, or when an observation is altered or deleted.
2. Nonsignificant results in individual tests on the regression coefficients for important predictor variables.
3. Estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical considerations or prior experience.
4. Large coefficients of simple correlation between pairs of predictor variables in the correlation matrix $\mathbf{r}_{XX}$.
5. Wide confidence intervals for the regression coefficients representing important predictor variables.

Example

We consider again the body fat example of Table 7.1, this time with all three predictor variables—triceps skinfold thickness (X_1), thigh circumference (X_2), and midarm circumference (X_3). We noted in Chapter 7 that the predictor variables triceps skinfold thickness and thigh circumference are highly correlated with each other. We also noted large changes in the estimated regression coefficients and their estimated standard deviations when a variable was added, nonsignificant results in individual tests on anticipated important variables, and an estimated negative coefficient when a positive coefficient was expected. These are all informal indications that suggest serious multicollinearity among the predictor variables.

Comment

The informal methods just described have important limitations. They do not provide quantitative measurements of the impact of multicollinearity and they may not identify the nature of the multicollinearity. For instance, if predictor variables X_1, X_2, and X_3 have low pairwise correlations, then the examination of simple correlation coefficients may not disclose the existence of relations among groups of predictor variables, such as a high correlation between X_1 and a linear combination of X_2 and X_3.

Another limitation of the informal diagnostic methods is that sometimes the observed behavior may occur without multicollinearity being present. ■

Variance Inflation Factor

A formal method of detecting the presence of multicollinearity that is widely accepted is use of variance inflation factors. These factors measure how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

To understand the significance of variance inflation factors, we begin with the precision of least squares estimated regression coefficients, which is measured by their variances. We know from (6.46) that the variance-covariance matrix of the estimated regression coefficients is:

$$\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{10.38}$$

For purposes of measuring the impact of multicollinearity, it is useful to work with the standardized regression model (7.45), which is obtained by transforming the variables by means of the correlation transformation (7.44). When the standardized regression model is fitted, the estimated regression coefficients b_k^* are standardized coefficients that are related to the estimated regression coefficients for the untransformed variables according to (7.53). The variance-covariance matrix of the estimated standardized regression coefficients is obtained from (10.38) by using the result in (7.50), which states that the $\mathbf{X}'\mathbf{X}$ matrix for the transformed variables is the correlation matrix of the X variables $\mathbf{r}_{XX}$. Hence, we obtain:

$$\sigma^2\{\mathbf{b}^*\} = (\sigma^*)^2\mathbf{r}_{XX}^{-1} \tag{10.39}$$

where $\mathbf{r}_{XX}$ is the matrix of the pairwise simple correlation coefficients among the X variables, as defined in (7.47), and $(\sigma^*)^2$ is the error term variance for the transformed model.

Note from (10.39) that the variance of b_k^* ($k = 1, \ldots, p-1$) is equal to the following, letting $(VIF)_k$ denote the kth diagonal element of the matrix $\mathbf{r}_{XX}^{-1}$:

$$\sigma^2\{b_k^*\} = (\sigma^*)^2(VIF)_k \tag{10.40}$$

The diagonal element $(VIF)_k$ is called the *variance inflation factor* (VIF) for b_k^*. It can be shown that this variance inflation factor is equal to:

$$(VIF)_k = \left(1 - R_k^2\right)^{-1} \qquad k = 1, 2, \ldots, p-1 \tag{10.41}$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the $p-2$ other X variables in the model. Hence, we have:

$$\sigma^2\{b_k^*\} = \frac{(\sigma^*)^2}{1 - R_k^2} \tag{10.42}$$

We presented in (7.65) the special results for $\sigma^2\{b_k^*\}$ when $p - 1 = 2$, for which $R_k^2 = r_{12}^2$, the coefficient of simple determination between X_1 and X_2.

The variance inflation factor $(VIF)_k$ is equal to 1 when $R_k^2 = 0$, i.e., when X_k is not linearly related to the other X variables. When $R_k^2 \neq 0$, then $(VIF)_k$ is greater than 1, indicating an inflated variance for b_k^* as a result of the intercorrelations among the X variables. When X_k has a perfect linear association with the other X variables in the model so that $R_k^2 = 1$, then $(VIF)_k$ and $\sigma^2\{b_k^*\}$ are unbounded.

Diagnostic Uses. The largest *VIF* value among all X variables is often used as an indicator of the severity of multicollinearity. A maximum *VIF* value in excess of 10 is frequently taken as an indication that multicollinearity may be unduly influencing the least squares estimates.

The mean of the *VIF* values also provides information about the severity of the multicollinearity in terms of how far the estimated standardized regression coefficients b_k^* are from the true values β_k^*. It can be shown that the expected value of the sum of these squared errors $(b_k^* - \beta_k^*)^2$ is given by:

$$E\left\{\sum_{k=1}^{p-1}(b_k^* - \beta_k^*)^2\right\} = (\sigma^*)^2 \sum_{k=1}^{p-1}(VIF)_k \qquad \textbf{(10.43)}$$

Thus, large *VIF* values result, on the average, in larger differences between the estimated and true standardized regression coefficients.

When no X variable is linearly related to the others in the regression model, $R_k^2 \equiv 0$; hence, $(VIF)_k \equiv 1$, their sum is $p - 1$, and the expected value of the sum of the squared errors is:

$$E\left\{\sum_{k=1}^{p-1}(b_k^* - \beta_k^*)^2\right\} = (\sigma^*)^2(p-1) \qquad \text{when } (VIF)_k \equiv 1 \qquad \textbf{(10.43a)}$$

A ratio of the results in (10.43) and (10.43a) provides useful information about the effect of multicollinearity on the sum of the squared errors:

$$\frac{(\sigma^*)^2 \sum(VIF)_k}{(\sigma^*)^2(p-1)} = \frac{\sum(VIF)_k}{p-1}$$

Note that this ratio is simply the mean of the *VIF* values, to be denoted by $(\overline{VIF})$:

$$(\overline{VIF}) = \frac{\sum_{k=1}^{p-1}(VIF)_k}{p-1} \qquad \textbf{(10.44)}$$

Mean *VIF* values considerably larger than 1 are indicative of serious multicollinearity problems.

Example

Table 10.5 contains the estimated standardized regression coefficients and the *VIF* values for the body fat example with three predictor variables (calculations not shown). The maximum of the *VIF* values is 708.84 and their mean value is $(\overline{VIF}) = 459.26$. Thus, the expected sum of the squared errors in the least squares standardized regression coefficients is nearly 460 times as large as it would be if the X variables were uncorrelated. In addition, all three *VIF* values greatly exceed 10, which again indicates that serious multicollinearity problems exist.

TABLE 10.5 Variance Inflation Factors—Body Fat Example with Three Predictor Variables.

Variable	b_k^*	$(VIF)_k$
X_1	4.2637	708.84
X_2	−2.9287	564.34
X_3	−1.5614	104.61

Maximum $(VIF)_k = 708.84$ $(\overline{VIF}) = 459.26$

It is interesting to note that $(VIF)_3 = 105$ despite the fact that both r_{13}^2 and r_{23}^2 (see Figure 7.3b) are not large. Here is an instance where X_3 is strongly related to X_1 and X_2 together ($R_3^2 = .990$), even though the pairwise coefficients of simple determination are not large. Examination of the pairwise correlations does not disclose this multicollinearity.

Comments

1. Some computer regression programs use the reciprocal of the variance inflation factor to detect instances where an X variable should not be allowed into the fitted regression model because of excessively high interdependence between this variable and the other X variables in the model. Tolerance limits for $1/(VIF)_k = 1 - R_k^2$ frequently used are .01, .001, or .0001, below which the variable is not entered into the model.

2. A limitation of variance inflation factors for detecting multicollinearities is that they cannot distinguish between several simultaneous multicollinearities.

3. A number of other formal methods for detecting multicollinearity have been proposed. These are more complex than variance inflation factors and are discussed in specialized texts such as References 10.5 and 10.6. ■

10.6 Surgical Unit Example—Continued

In Chapter 9 we developed a regression model for the surgical unit example (data in Table 9.1). Recall that validation studies in Section 9.6 led to the selection of model (9.21), the model containing variables X_1, X_2, X_3, and X_8. We will now utilize this regression model to demonstrate a more in-depth study of curvature, interaction effects, multicollinearity, and influential cases using residuals and other diagnostics.

To examine interaction effects further, a regression model containing first-order terms in X_1, X_2, X_3, and X_8 was fitted and added-variable plots for the six two-factor interaction terms, X_1X_2, X_1X_3, X_1X_8, X_2X_3, X_2X_8, and X_3X_8, were examined. These plots (not shown) did not suggest that any strong two-variable interactions are present and need to be included in the model. The absence of any strong interactions was also noted by fitting a regression model containing X_1, X_2, X_3, and X_8 in first-order terms and all two-variable interaction terms. The P-value of the formal F test statistic (7.19) for dropping all of the interaction terms from the model containing both the first-order effects and the interaction effects is .35, indicating that interaction effects are not present.

Figure 10.9 contains some of the additional diagnostic plots that were generated to check on the adequacy of the first-order model:

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_8 X_{i8} + \varepsilon_i \tag{10.45}$$

where $Y_i' = \ln Y_i$. The following points are worth noting:

1. The residual plot against the fitted values in Figure 10.9a shows no evidence of serious departures from the model.

2. One of the three candidate models (9.23) subjected to validation studies in Section 9.6 contained X_5 (patient age) as a predictor. The regression coefficient for age (b_5) was negative in model (9.23), but when the same model was fit to the validation data, the sign of b_5 became positive. We will now use a residual plot and an added-variable plot to study graphically

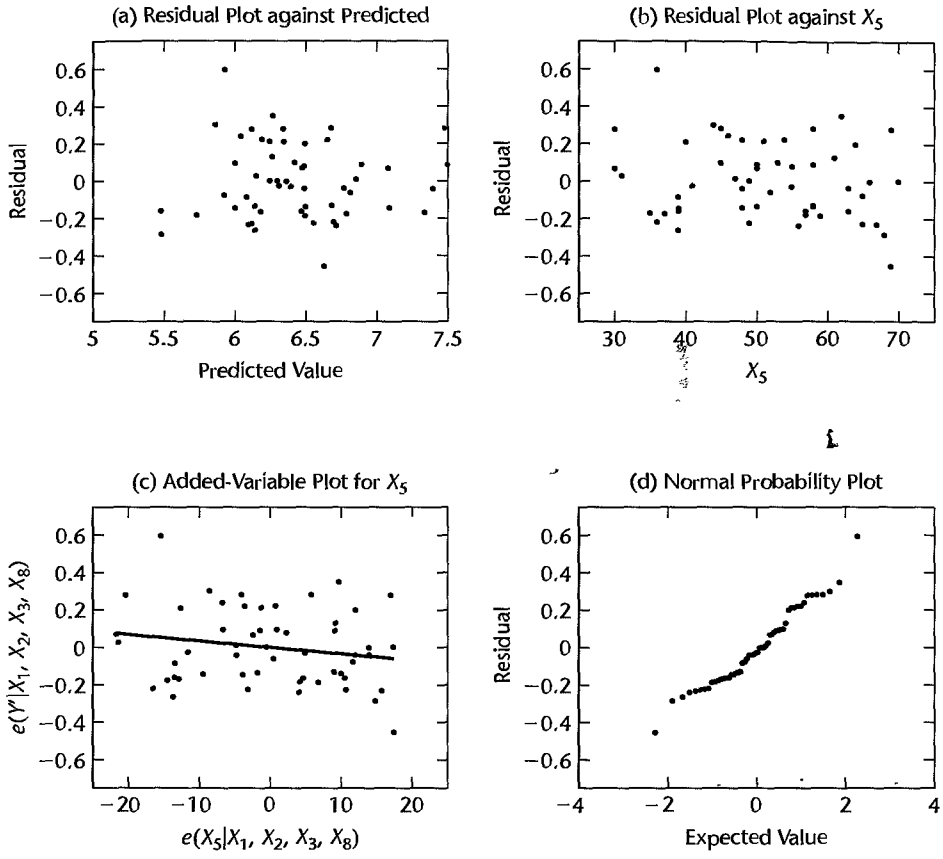


FIGURE 10.9 Residual and Added-Variable Plots for Surgical Unit Example—Regression Model (10.45).

the strength of the marginal relationship between X_5 and the response, when X_1, X_2, X_3, and X_8 are already in the model. Figure 10.9b shows the plot of the residuals for the model containing X_1, X_2, X_3, and X_8 against X_5, the predictor variable not in the model. This plot shows no need to include patient age (X_5) in the model to predict logarithm of survival time. A better view of this marginal relationship is provided by the added-variable plot in Figure 10.9c. The slope coefficient b_5 can be seen again to be slightly negative as depicted by the solid line in the added-variable plot. Overall, however, the marginal relationship between X_5 and Y' is weak. The P-value of the formal t test (9.18) for dropping X_5 from the model containing X_1, X_2, X_3, X_5 and X_8 is 0.194. In addition, the plot shows that the negative slope is driven largely by one or two outliers—one in the upper left region of the plot, and one in the lower right region. In this way the added-variable plot provides additional support for dropping X_5.

3. The normal probability plot of the residuals in Figure 10.9d shows little departure from linearity. The coefficient of correlation between the ordered residuals and their expected values under normality is .982, which is larger than the critical value for significance level .05 in Table B.6.

Multicollinearity was studied by calculating the variance inflation factors:

Variable	$(VIF)_k$
X_1	1.10
X_2	1.02
X_3	1.05
X_8	1.09

As may be seen from these results, multicollinearity among the four predictor variables is not a problem.

Figure 10.10 contains index plots of four key regression diagnostics, namely the deleted studentized residuals t_i in Figure 10.10a. the leverage values h_{ii} in Figure 10.10b, Cook's distances D_i in Figure 10.10c, and $DFFITS_i$ values in Figure 10.10d. These plots suggest further study of cases 17, 28, and 38. Table 10.6 lists numerical diagnostic values for these cases. The measures presented in columns 1–5 are the residuals e_i in (10.8), the studentized deleted residuals t_i in (10.24), the leverage values h_{ii} in (10.18), the Cook's distance measures D_i in (10.33), and the $(DFFITS)_i$ values in (10.30). The following are noteworthy points about the diagnostics in Table 10.6:

1. Case 17 was identified as outlying with regard to its Y value according to its studentized deleted residual, outlying by more than three standard deviations. We test formally whether case 17 is outlying by means of the Bonferroni test procedure. For a family significance level of $\alpha = .05$ and sample size $n = 54$, we require $t(1-\alpha/2n; n-p-1) = t(.99954; 49) = 3.528$. Since $|t_{17}| = 3.3696 \leq 3.528$, the formal outlier test indicates that case 22 is not an outlier. Still, t_{17} is very close to the critical value, and although this case does not appear to be outlying to any substantial extent, we may wish to investigate the influence of case 17 to remove any doubts.

2. With $2p/n = 2(5)/54 = .185$ as a guide for identifying outlying X observations, cases 23, 28, 32, 38, 42, and 52 were identified as outlying according to their leverage values. Incidentally, the univariate dot plots identify only cases 28 and 38 as outlying. Here we see the value of multivariable outlier identification.

3. To determine the influence of cases 17, 23, 28, 32, 38, 42, 32, and 52, we consider their Cook's distance and *DFFITS* values. According to each of these measures, case 17 is the most influential, with Cook's distance $D_{17} = .3306$ and $(DFFITS)_{17} = 1.4151$. Referring to the F distribution with 5 and 49 degrees of freedom, we note that the Cook's value corresponds to the 11th percentile. It thus appears that the influence of case 38 is not large enough to warrant remedial measures, and consequently the other outlying cases also do not appear to be overly influential.

A direct check of the influence of case 17 on the inferences of interest was also conducted. Here, the inferences of primary interest are in the fit of the regression model because the model is intended to be used for making predictions in the range of the X observations. Hence, each fitted value $\hat{Y}_i$ based on all 54 observations was compared with the fitted value $\hat{Y}_{i(17)}$ when case 17 is deleted in fitting the regression model. The average of the absolute percent differences:

$$\left| \frac{\hat{Y}_{i(17)} - \hat{Y}_i}{\hat{Y}_i} \right| 100$$

FIGURE 10.10 Diagnostic Plots for Surgical Unit Example—Regression Model (10.45).

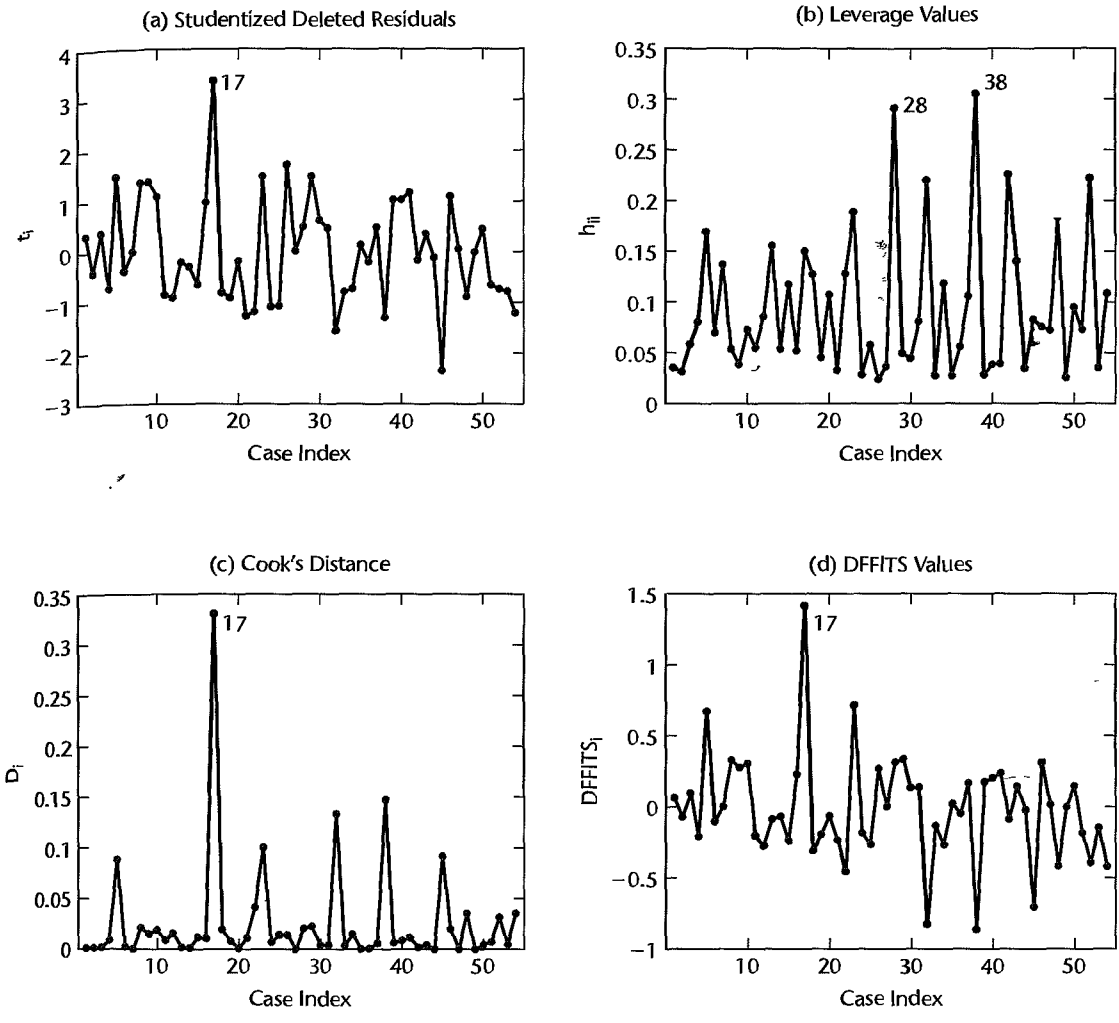


TABLE 10.6 Various Diagnostics for Outlying Cases—Surgical Unit Example, Regression Model (10.45).

	(1)	(2)	(3)	(4)	(5)
Case Number i	e_i	t_i	h_{ii}	D_i	$(DFFITS)_i$
17	0.5952	3.3696	0.1499	0.3306	1.4151
23	0.2788	1.4854	0.1885	0.1001	0.7160
28	0.0876	0.4896	0.2914	0.0200	0.3140
32	−0.2861	−1.5585	0.2202	0.1333	−0.8283
38	−0.2271	−1.3016	0.3059	0.1472	−0.8641
42	−0.0303	−0.1620	0.2262	0.0016	−0.0876
52	−0.1375	−0.7358	0.2221	0.0312	−0.3931

is only .42 percent, and the largest absolute percent difference (which is for case 17) is only 1.77 percent. Thus, case 17 does not have such a disproportionate influence on the fitted values that remedial action would be required.

4. In summary, the diagnostic analyses identified a number of potential problems, but none of these was considered to be serious enough to require further remedial action.

Cited References

10.1. Atkinson, A. C. *Plots, Transformations, and Regression*. Oxford: Clarendon Press, 1987.

10.2. Mansfield, E. R., and M. D. Conerly. "Diagnostic Value of Residual and Partial Residual Plots," *The American Statistician* 41 (1987), pp. 107–16.

10.3. Cook, R. D. "Exploring Partial Residual Plots," *Technometrics* 35 (1993), pp. 351–62.

10.4. Rousseeuw, P. J., and A. M. Leroy. *Robust Regression and Outlier Detection*. New York: John Wiley & Sons, 1987.

10.5. Belsley, D. A.; E. Kuh; and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons, 1980.

10.6. Belsley, D. A. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: John Wiley & Sons, 1991.

Problems

10.1. A student asked: "Why is it necessary to perform diagnostic checks of the fit when R^2 is large?" Comment.

10.2. A researcher stated: "One good thing about added-variable plots is that they are extremely useful for identifying model adequacy even when the predictor variables are not properly specified in the regression model." Comment.

10.3. A student suggested: "If extremely influential outlying cases are detected in a data set, simply discard these cases from the data set." Comment.

10.4. Describe several informal methods that can be helpful in identifying multicollinearity among the X variables in a multiple regression model.

10.5. Refer to **Brand preference** Problem 6.5b.

a. Prepare an added-variable plot for each of the predictor variables.

b. Do your plots in part (a) suggest that the regression relationships in the fitted regression function in Problem 6.5b are inappropriate for any of the predictor variables? Explain.

c. Obtain the fitted regression function in Problem 6.5b by separately regressing both Y and X_2 on X_1, and then regressing the residuals in an appropriate fashion.

10.6. Refer to **Grocery retailer** Problem 6.9.

a. Fit regression model (6.1) to the data using X_1 and X_2 only.

b. Prepare an added-variable plot for each of the predictor variables X_1 and X_2.

c. Do your plots in part (a) suggest that the regression relationships in the fitted regression function in part (a) are inappropriate for any of the predictor variables? Explain.

d. Obtain the fitted regression function in part (a) by separately regressing both Y and X_2 on X_1, and then regressing the residuals in an appropriate fashion.

10.7. Refer to **Patient satisfaction** Problem 6.15c.

a. Prepare an added-variable plot for each of the predictor variables.

b. Do your plots in part (a) suggest that the regression relationships in the fitted regression function in Problem 6.15c are inappropriate for any of the predictor variables? Explain.

10.8. Refer to **Commercial properties** Problem 6.18c.

a. Prepare an added-variable plot for each of the predictor variables.

b. Do your plots in part (a) suggest that the regression relationships in the fitted regression function in Problem 6.18c are inappropriate for any of the predictor variables? Explain.

10.9. Refer to **Brand preference** Problem 6.5.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .10$. State the decision rule and conclusion.

b. Obtain the diagonal elements of the hat matrix, and provide an explanation for the pattern in these elements.

c. Are any of the observations outlying with regard to their X values according to the rule of thumb stated in the chapter?

d. Management wishes to estimate the mean degree of brand liking for moisture content $X_1 = 10$ and sweetness $X_2 = 3$. Construct a scatter plot of X_2 against X_1 and determine visually whether this prediction involves an extrapolation beyond the range of the data. Also, use (10.29) to determine whether an extrapolation is involved. Do your conclusions from the two methods agree?

e. The largest absolute studentized deleted residual is for case 14. Obtain the *DFFITS*, *DFBETAS*, and Cook's distance values for this case to assess the influence of this case. What do you conclude?

f. Calculate the average absolute percent difference in the fitted values with and without case 14. What does this measure indicate about the influence of case 14?

g. Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

*10.10. Refer to **Grocery retailer** Problems 6.9 and 6.10.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .05$. State the decision rule and conclusion.

b. Obtain the diagonal element of the hat matrix. Identify any outlying X observations using the rule of thumb presented in the chapter.

c. Management wishes to predict the total labor hours required to handle the next shipment containing $X_1 = 300{,}000$ cases whose indirect costs of the total hours is $X_2 = 7.2$ and $X_3 = 0$ (no holiday in week). Construct a scatter plot of X_2 against X_1 and determine visually whether this prediction involves an extrapolation beyond the range of the data. Also, use (10.29) to determine whether an extrapolation is involved. Do your conclusions from the two methods agree?

d. Cases 16, 22, 43, and 48 appear to be outlying X observations, and cases 10, 32, 38, and 40 appear to be outlying Y observations. Obtain the *DFFITS*, *DFBETAS*, and Cook's distance values for each of these cases to assess their influence. What do you conclude?

e. Calculate the average absolute percent difference in the fitted values with and without each of these cases. What does this measure indicate about the influence of each of the cases?

f. Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

*10.11. Refer to **Patient satisfaction** Problem 6.15.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .10$. State the decision rule and conclusion.

b. Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

c. Hospital management wishes to estimate mean patient satisfaction for patients who are $X_1 = 30$ years old, whose index of illness severity is $X_2 = 58$, and whose index of anxiety level is $X_3 = 2.0$. Use (10.29) to determine whether this estimate will involve a hidden extrapolation.

d. The three largest absolute studentized deleted residuals are for cases 11, 17, and 27. Obtain the *DFFITS*, *DFBETAS*, and Cook's distance values for this case to assess its influence. What do you conclude?

e. Calculate the average absolute percent difference in the fitted values with and without each of these cases. What does this measure indicate about the influence of each of these cases?

f. Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

10.12. Refer to **Commercial Properties** Problem 6.18.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .01$. State the decision rule and conclusion.

b. Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

c. The researcher wishes to estimate the rental rates of a property whose age is 10 years, whose operating expenses and taxes are 12.00, whose occupancy rate is 0.05, and whose square footage is 350,000. Use (10.29) to determine whether this estimate will involve a hidden extrapolation.

d. Cases 61, 8, 3, and 53 appear to be outlying X observations, and cases 6 and 62 appear to be outlying Y observations. Obtain the *DFFITS*, *DFBETAS*, and Cook's distance values for each case to assess its influence. What do you conclude?

e. Calculate the average absolute percent difference in the fitted values with and without each of the cases. What does this measure indicate about the influence of each case?

f. Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

10.13. **Cosmetics sales.** An assistant in the district sales office of a national cosmetics firm obtained data, shown below, on advertising expenditures and sales last year in the district's 44 territories. X_1 denotes expenditures for point-of-sale displays in beauty salons and department stores (in thousand dollars), and X_2 and X_3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Y denotes sales (in thousand cases). The assistant was instructed to estimate the increase in expected sales when X_1 is increased by 1 thousand dollars and X_2 and X_3 are held constant, and was told to use an ordinary multiple regression model with linear terms for the predictor variables and with independent normal error terms.

i:	1	2	3	...	42	43	44
X_{i1}:	5.6	4.1	3.7	...	3.6	3.9	5.5
X_{i2}:	5.6	4.8	3.5	...	3.7	3.6	5.0
X_{i3}:	3.8	4.8	3.6	...	4.4	2.9	5.5
Y_i:	12.85	11.55	12.78	...	10.47	11.03	12.31

a. State the regression model to be employed and fit it to the data.

b. Test whether there is a regression relation between sales and the three predictor variables; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

c. Test for each of the regression coefficients β_k ($k = 1, 2, 3$) individually whether or not $\beta_k = 0$; use $\alpha = .05$ each time. Do the conclusions of these tests correspond to that obtained in part (b)?

d. Obtain the correlation matrix of the X variables.

e. What do the results in parts (b), (c), and (d) suggest about the suitability of the data for the research objective?

10.14. Refer to **Cosmetics sales** Problem 10.13.

a. Obtain the three variance inflation factors. What do these suggest about the effects of multicollinearity here?

b. The assistant eventually decided to drop variables X_2 and X_3 from the model "to clear up the picture." Fit the assistant's revised model. Is the assistant now in a better position to achieve the research objective?

c. Why would an experiment here be more effective in providing suitable data to meet the research objective? How would you design such an experiment? What regression model would you employ?

10.15. Refer to **Brand preference** Problem 6.5a.

a. What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables?

b. Find the two variance inflation factors. Why are they both equal to 1?

*10.16. Refer to **Grocery retailer** Problem 6.9c.

a. What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables?

b. Find the three variance inflation factors. Do they indicate that a serious multicollinearity problem exists here?

*10.17. Refer to **Patient satisfaction** Problem 6.15b.

a. What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables?

b. Obtain the three variance inflation factors. What do these results suggest about the effects of multicollinearity here? Are these results more revealing than those in part (a)?

10.18. Refer to **Commercial properties** Problem 6.18b.

a. What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables?

b. Obtain the four variance inflation factors. Do they indicate that a serious multicollinearity problem exists here?

10.19. Refer to **Job proficiency** Problems 9.10 and 9.11. The subset model containing only first-order terms in X_1 and X_3 is to be evaluated in detail.

a. Obtain the residuals and plot them separately against $\hat{Y}$, each of the four predictor variables, and the cross-product term X_1X_3. On the basis of these plots, should any modifications in the regression model be investigated?

b. Prepare separate added-variable plots against $e(X_1|X_3)$ and $e(X_3|X_1)$. Do these plots suggest that any modifications in the model form are warranted?

c. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumptions, using Table B.6 and $\alpha = .01$. What do you conclude?

d. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .05$. State the decision rule and conclusion.

e. Obtain the diagonal elements of the hat matrix. Using the rule of thumb in the text, identify any outlying X observations. Are your findings consistent with those in Problem 9.10a? Should they be? Comment.

f. Cases 7 and 18 appear to be moderately outlying with respect to their X values, and case 16 is reasonably far outlying with respect to its Y value. Obtain *DFFITS*, *DFBETAS*, and Cook's distance values for these cases to assess their influence. What do you conclude?

g. Obtain the variance inflation factors. What do they indicate?

10.20. Refer to **Lung pressure** Problems 9.13 and 9.14. The subset regression model containing first-order terms for X_1 and X_2 and the cross-product term X_1X_2 is to be evaluated in detail.

a. Obtain the residuals and plot them separately against $\hat{Y}$ and each of the three predictor variables. On the basis of these plots, should any further modifications of the regression model be attempted?

b. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here?

c. Obtain the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

d. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .05$. State the decision rule and conclusion.

e. Obtain the diagonal elements of the hat matrix. Using the rule of thumb in the text, identify any outlying X observations. Are your findings consistent with those in Problem 9.13a? Should they be? Discuss.

f. Cases 3, 8, and 15 are moderately far outlying with respect to their X values, and case 7 is relatively far outlying with respect to its Y value. Obtain *DFFITS*, *DFBETAS*, and Cook's distance values for these cases to assess their influence. What do you conclude?

*10.21. Refer to **Kidney function** Problem 9.15 and the regression model fitted in part (c).

a. Obtain the variance inflation factors. Are there indications that serious multicollinearity problems exist here? Explain.

b. Obtain the residuals and plot them separately against $\hat{Y}$ and each of the predictor variables. Also prepare a normal probability plot of the residuals.

c. Prepare separate added-variable plots against $e(X_1|X_2, X_3)$, $e(X_2|X_1, X_3)$, and $e(X_3|X_1, X_2)$.

d. Do the plots in parts (b) and (c) suggest that the regression model should be modified?

*10.22. Refer to **Kidney function** Problems 9.15 and 10.21. Theoretical arguments suggest use of the following regression function:

$$E\{\ln Y\} = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln(140 - X_2) + \beta_3 \ln X_3$$

a. Fit the regression function based on theoretical considerations.

b. Obtain the residuals and plot them separately against $\hat{Y}$ and each predictor variable in the fitted model. Also prepare a normal probability plot of the residuals. Have the difficulties noted in Problem 10.21 now largely been eliminated?

c. Obtain the variance inflation factors. Are there indications that serious multicollinearity problems exist here? Explain.

d. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .10$. State the decision rule and conclusion.

e. Obtain the diagonal elements of the hat matrix. Using the rule of thumb in the text, identify any outlying X observations.

f. Cases 28 and 29 are relatively far outlying with respect to their Y values. Obtain *DFFITS*, *DFBETAS*, and Cook's distance values for these cases to assess their influence. What do you conclude?

Exercises

10.23. Show that (10.37) is algebraically equivalent to (10.33a).

10.24. If $n = p$ and the $\mathbf{X}$ matrix is invertible, use (5.34) and (5.37) to show that the hat matrix $\mathbf{H}$ is given by the $p \times p$ identity matrix. In this case, what are h_{ii} and $\hat{Y}_i$?

10.25. Show that (10.26) follows from (10.24a) and (10.25).

10.26. Prove (9.11), using (10.27) and Exercise 5.31.

Projects

10.27. Refer to the **SENIC** data set in Appendix C.1 and Project 9.25. The regression model containing age, routine chest X-ray ratio, and average daily census in first-order terms is to be evaluated in detail based on the model-building data set.

a. Obtain the residuals and plot them separately against $\hat{Y}$, each of the predictor variables in the model, and each of the related cross-product terms. On the basis of these plots, should any modifications of the model be made?

b. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption, using Table B.6 and $\alpha = .05$. What do you conclude?

c. Obtain the scatter plot matrix, the correlation matrix of the X variables, and the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

d. Obtain the studentized deleted residuals and prepare a dot plot of these residuals. Are any outliers present? Use the Bonferroni outlier test procedure with $\alpha = .01$. State the decision rule and conclusion.

e. Obtain the diagonal elements of the hat matrix. Using the rule of thumb in the text, identify any outlying X observations.

f. Cases 62, 75, 106, and 112 are moderately outlying with respect to their X values, and case 87 is reasonably far outlying with respect to its Y value. Obtain *DFFITS*, *DFBETAS*, and Cook's distance values for these cases to assess their influence. What do you conclude?

10.28. Refer to the **CDI** data set in Appendix C.2 and Project 9.26. The regression model containing variables 6, 8, 9, 13, 14, and 15 in first-order terms is to be evaluated in detail based on the model-building data set.

a. Obtain the residuals and plot them separately against $\hat{Y}$, each predictor variable in the model, and the related cross-product term. On the basis of these plots, should any modifications in the model be made?

b. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption, using Table B.6 and $\alpha = .01$. What do you conclude?

c. Obtain the scatter plot matrix, the correlation matrix of the X variables, and the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

d. Obtain the studentized deleted residuals and prepare a dot plot of these residuals. Are any outliers present? Use the Bonferroni outlier test procedure with $\alpha = .05$. State the decision rule and conclusion.

e. Obtain the diagonal elements of the hat matrix. Using the rule of thumb in the text, identify any outlying X observations.

f. Cases 2, 8, 48, 128, 206, and 404 are outlying with respect to their X values, and cases 2 and 6 are reasonably far outlying with respect to their Y values. Obtain *DFFITS*, *DFBETAS*, and Cook's distance values for these cases to assess their influence. What do you conclude?

Case Studies

10.29. Refer to the **Website developer** data set in Appendix C.6 and Case Study 9.29. For the best subset model developed in Case Study 9.29, perform appropriate diagnostic checks to evaluate outliers and assess their influence. Do any serious multicollinearity problems exist here?

10.30. Refer to the **Prostate cancer** data set in Appendix C.5 and Case Study 9.30. For the best subset model developed in Case Study 9.30, perform appropriate diagnostic checks to evaluate outliers and assess their influence. Do any serious multicollinearity problems exist here?

10.31. Refer to the **Real estate** data set in Appendix C.7 and Case Study 9.31. For the best subset model developed in Case Study 9.31, perform appropriate diagnostic checks to evaluate outliers and assess their influence. Do any serious multicollinearity problems exist here?