9 Multiple Regression: Model Validation and Diagnostics

In Sections 7.8.2 and 7.9 we discussed some consequences of misspecification of the model. In this chapter we consider various approaches to checking the model and the attendant assumptions for adequacy and validity. Some properties of the residuals [see (7.11)] and the hat matrix are developed in Sections 9.1 and 9.2. We discuss outliers, the influence of individual observations, and leverage in Sections 9.3 and 9.4.

For additional reading, see Snee (1977), Cook (1977), Belsley et al. (1980), Draper and Smith (1981, Chapter 6), Cook and Weisberg (1982), Beckman and Cook (1983), Weisberg (1985, Chapters 5, 6), Chatterjee and Hadi (1988), Myers (1990, Chapters 5–8), Sen and Srivastava (1990, Chapter 8), Montgomery and Peck (1992, pp. 67–113, 159–192), Jørgensen (1993, Chapter 5), Graybill and Iyer (1994, Chapter 5), Hocking (1996, Chapter 9), Christensen (1996, Chapter 13), Ryan (1997, Chapters 2, 5), Fox (1997, Chapters 11–13) and Kutner et al. (2005, Chapter 10).

9.1 RESIDUALS

The usual model is given by (7.4) as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with assumptions $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, where $\mathbf{y}$ is $n \times 1$, $\mathbf{X}$ is $n \times (k+1)$ of rank $k + 1 < n$, and β is $(k+1) \times 1$. The error vector $\boldsymbol{\varepsilon}$ is unobservable unless $\boldsymbol{\beta}$ is known. To estimate $\boldsymbol{\varepsilon}$ for a given sample, we use the residual vector

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \hat{\mathbf{y}} \tag{9.1}$$

as defined in (7.11). The n residuals in (9.1), $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n$, are used in various plots and procedures for checking on the validity or adequacy of the model.

We first consider some properties of the residual vector $\hat{\varepsilon}$. Using the least-squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in (7.6), the vector of predicted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ can be

Linear Models in Statistics, *Second Edition*, by Alvin C. Rencher and G. Bruce Schaalje written as

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}, \qquad (9.2)\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (see Section 8.2). The $n \times n$ matrix $\mathbf{H}$ is called the *hat matrix* because it transforms $\mathbf{y}$ to $\hat{\mathbf{y}}$. We also refer to $\mathbf{H}$ as a *projection matrix* for essentially the same reason; geometrically it projects $\mathbf{y}$ (perpendicularly) onto $\hat{\mathbf{y}}$ (see Fig. 7.4). The hat matrix $\mathbf{H}$ is symmetric and idempotent (see Problem 5.32a).

Multiplying $\mathbf{X}$ by $\mathbf{H}$, we obtain

$$\mathbf{H}\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X}. \qquad (9.3)$$

Writing $\mathbf{X}$ in terms of its columns and using (2.28), we can write (9.3) as

$$\mathbf{H}\mathbf{X} = \mathbf{H}(\mathbf{j}, \mathbf{x}_1, \ldots \mathbf{x}_k) = (\mathbf{H}\mathbf{j}, \mathbf{H}\mathbf{x}_1, \ldots, \mathbf{H}\mathbf{x}_k),$$

so that

$$\mathbf{j} = \mathbf{H}\mathbf{j}, \quad \mathbf{x}_i = \mathbf{H}\mathbf{x}_i, \quad i = 1, 2, \ldots, k. \qquad (9.4)$$

Using (9.2), the residual vector $\hat{\boldsymbol{\varepsilon}}$ (9.1) can be expressed in terms of $\mathbf{H}$:

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}. \qquad (9.5)\end{aligned}$$

We can rewrite (9.5) to express the residual vector $\hat{\boldsymbol{\varepsilon}}$ in terms of $\boldsymbol{\varepsilon}$:

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta}) + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \\ &= (\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \qquad [\text{by (9.3)}] \\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}. \qquad (9.6)\end{aligned}$$

In terms of the elements h_{ij} of $\mathbf{H}$, we have $\hat{\varepsilon}_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j$, $i = 1, 2, \ldots, n$. Thus, if the h_{ij}'s are small (in absolute value), $\hat{\boldsymbol{\varepsilon}}$ is close to $\boldsymbol{\varepsilon}$.

The following are some of the properties of $\hat{\boldsymbol{\varepsilon}}$ (see Problem 9.1). For the first four, we assume that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$:

$$E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0} \qquad (9.7)$$

$$\text{cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \sigma^2(\mathbf{I} - \mathbf{H}) \tag{9.8}$$

$$\text{cov}(\hat{\boldsymbol{\varepsilon}}, \mathbf{y}) = \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \sigma^2(\mathbf{I} - \mathbf{H}) \tag{9.9}$$

$$\text{cov}(\hat{\boldsymbol{\varepsilon}}, \hat{\mathbf{y}}) = \mathbf{O} \tag{9.10}$$

$$\bar{\hat{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i/n = \hat{\boldsymbol{\varepsilon}}'\mathbf{j}/n = 0 \tag{9.11}$$

$$\hat{\boldsymbol{\varepsilon}}'\mathbf{y} = \text{SSE} = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} \tag{9.12}$$

$$\hat{\boldsymbol{\varepsilon}}'\hat{\mathbf{y}} = 0 \tag{9.13}$$

$$\hat{\boldsymbol{\varepsilon}}'\mathbf{X} = \mathbf{0}' \tag{9.14}$$

In (9.7), the residual vector $\hat{\boldsymbol{\varepsilon}}$ has the same mean as the error term $\boldsymbol{\varepsilon}$, but in (9.8) $\text{cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$ differs from the assumption $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Thus the residuals $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n$ are not independent. However, in many cases, especially if n is large, the h_{ij}'s tend to be small (for $i \neq j$), and the dependence shown in $\sigma^2(\mathbf{I} - \mathbf{H})$ does not unduly affect plots and other techniques for model validation. Each $\hat{\varepsilon}_i$ is seen to be correlated with each y_j in (9.9), but in (9.10) the $\hat{\varepsilon}_i$'s are uncorrelated with the $\hat{y}_j$'s.

Some sample properties of the residuals are given in (9.11)–(9.14). The sample mean of the residuals is zero, as shown in (9.11). By (9.12), it can be seen that $\hat{\varepsilon}$ and y are correlated in the sample since $\hat{\boldsymbol{\varepsilon}}'\mathbf{y}$ is the numerator of

$$r_{\hat{\varepsilon}y} = \frac{\hat{\boldsymbol{\varepsilon}}'(\mathbf{y} - \bar{y}\mathbf{j})}{\sqrt{(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})(\mathbf{y} - \bar{y}\mathbf{j})'(\mathbf{y} - \bar{y}\mathbf{j})}} = \frac{\hat{\boldsymbol{\varepsilon}}'\mathbf{y}}{\sqrt{(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})(\mathbf{y} - \bar{y}\mathbf{j})'(\mathbf{y} - \bar{y}\mathbf{j})}}.$$

However, $\hat{\boldsymbol{\varepsilon}}$ and $\hat{\mathbf{y}}$ are orthogonal by (9.13), and therefore

$$r_{\hat{\varepsilon}\hat{y}} = 0. \tag{9.15}$$

Similarly, by (9.14), $\hat{\boldsymbol{\varepsilon}}$ is orthogonal to each column of $\mathbf{X}$ and

$$r_{\hat{\varepsilon}x_i} = 0, \quad i = 1, 2, \ldots, k. \tag{9.16}$$

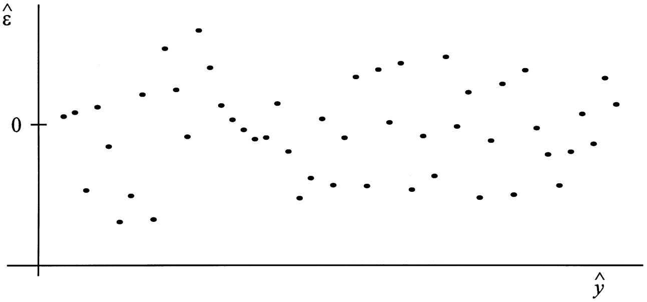

Figure 9.1 Ideal residual plot when model is correct.

If the model and attendant assumptions are correct, then by (9.15), a plot of the residuals versus predicted values, $(\hat{\varepsilon}_1, \hat{y}_1)$, $(\hat{\varepsilon}_2, \hat{y}_2), \ldots, (\hat{\varepsilon}_n, \hat{y}_n)$, should show no systematic pattern. Likewise, by (9.16), the k plots of the residuals versus each of $x_1, x_2, \ldots, x_k$ should show only random variation. These plots are therefore useful for checking the model. A typical plot of this type is shown in Figure 9.1. It may also be useful to plot the residuals on normal probability paper and to plot residuals in time sequence (Christensen 1996, Section 13.2).

If the model is incorrect, various plots involving residuals may show departures from the fitted model such as outliers, curvature, or nonconstant variance. The plots may also suggest remedial measures to improve the fit of the model. For example, the residuals could be plotted versus any of the x_i's, and a simple curved pattern might suggest the addition of x_i^2 to the model. We will consider various approaches for detecting outliers in Section 9.3 and for finding influential observations in Section 9.4. Before doing so, we discuss some properties of the hat matrix in Section 9.2.

9.2 THE HAT MATRIX

It was noted following (9.2) that the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is symmetric and idempotent. We now present some additional properties of this matrix. These properties will be useful in the discussion of outliers and influential observations in Sections 9.3 and 9.4.

For the centered model

$$\mathbf{y} = \alpha\mathbf{j} + \mathbf{X}_c\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{9.17}$$

in (7.32), $\hat{\mathbf{y}}$ becomes

$$\hat{\mathbf{y}} = \hat{\alpha}\mathbf{j} + \mathbf{X}_c\hat{\boldsymbol{\beta}}_1, \tag{9.18}$$

and the hat matrix is $\mathbf{H}_c = \mathbf{X}_c(\mathbf{X}'_c\mathbf{X}_c)^{-1}\mathbf{X}'_c$, where

$$\mathbf{X}_c = \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{X}_1 = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix}.$$

By (7.36) and (7.37), we can write (9.18) as

$$\begin{aligned} \hat{\mathbf{y}} &= \bar{y}\mathbf{j} + \mathbf{X}_c(\mathbf{X}'_c\mathbf{X}_c)^{-1}\mathbf{X}'_c\mathbf{y} = \left(\frac{1}{n}\mathbf{j}'\mathbf{y}\right)\mathbf{j} + \mathbf{H}_c\mathbf{y} \\ &= \left(\frac{1}{n}\mathbf{J} + \mathbf{H}_c\right)\mathbf{y}. \end{aligned} \tag{9.19}$$

Comparing (9.19) and (9.2), we have

$$\mathbf{H} = \frac{1}{n}\mathbf{J} + \mathbf{H}_c = \frac{1}{n}\mathbf{J} + \mathbf{X}_c(\mathbf{X}'_c\mathbf{X}_c)^{-1}\mathbf{X}'_c. \tag{9.20}$$

We now examine some properties of the elements h_{ij} of $\mathbf{H}$.

Theorem 9.2. If $\mathbf{X}$ is $n \times (k+1)$ of rank $k + 1 < n$, and if the first column of $\mathbf{X}$ is $\mathbf{j}$, then the elements h_{ij} of $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ have the following properties:

(i) $(1/n) \leq h_{ii} \leq 1$ for $i = 1, 2, \ldots, n$.

(ii) $-.5 \leq h_{ij} \leq .5$ for all $j \neq i$.

(iii) $h_{ii} = (1/n) + (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'(\mathbf{X}'_c\mathbf{X}_c)^{-1}(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)$, where $\mathbf{x}'_{1i} = (x_{i1}, x_{i2}, \ldots, x_{ik})$, $\bar{\mathbf{x}}'_1 = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$, and $(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'$ is the ith row of the centered matrix $\mathbf{X}_c$.

(iv) $\text{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = k + 1$.

PROOF

(i) The lower bound follows from (9.20), since $\mathbf{X}'_c\mathbf{X}_c$ is positive definite. Since $\mathbf{H}$ is symmetric and idempotent, we use the relationship $\mathbf{H} = \mathbf{H}^2$ to find an upper bound on h_{ii}. Let $\mathbf{h}'_i$ be the ith row of $\mathbf{H}$. Then

$$\begin{aligned} h_{ii} &= \mathbf{h}'_i\mathbf{h}_i = (h_{i1}, h_{i2}, \ldots, h_{in}) \begin{pmatrix} h_{i1} \\ h_{i2} \\ \vdots \\ h_{in} \end{pmatrix} = \sum_{j=1}^{n} h_{ij}^2 \\ &= h_{ii}^2 + \sum_{j \neq i} h_{ij}^2. \end{aligned} \tag{9.21}$$

Dividing both sides of (9.21) by h_{ii} [which is positive since $h_{ii} \geq (1/n)$], we obtain

$$1 = h_{ii} + \frac{\sum_{j \neq i} h_{ij}^2}{h_{ii}}, \tag{9.22}$$

which implies $h_{ii} \leq 1$.

(ii) (Chatterjee and Hadi 1988, p. 18.) We can write (9.21) in the form

$$h_{ii} = h_{ii}^2 + h_{ij}^2 + \sum_{r \neq i,j} h_{ir}^2$$

or

$$h_{ii} - h_{ii}^2 = h_{ij}^2 + \sum_{r \neq i,j} h_{ir}^2.$$

Thus, $h_{ij}^2 \leq h_{ii} - h_{ii}^2$, and since the maximum value of $h_{ii} - h_{ii}^2$ is $\frac{1}{4}$, we have $h_{ij}^2 \leq \frac{1}{4}$ for $j \neq i$.

(iii) This follows from (9.20); see Problem 9.2b.

(iv) See Problem 9.2c. □

By Theorem 9.2(iv), we see that as n increases, the values of h_{ii} will tend to decrease.

The function $(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'(\mathbf{X}_c'\mathbf{X}_c)^{-1}(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)$ in Theorem 9.2(iii) is a standardized distance. The standardized distance (Mahalanobis distance) defined in (3.27) is for a population covariance matrix. The matrix $\mathbf{X}_c'\mathbf{X}_c$ is proportional to a sample covariance matrix [see (7.44)]. Thus, $(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'(\mathbf{X}_c'\mathbf{X}_c)^{-1}(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)$ is an estimated standardized distance and provides a good measure of the relative distance of each $\mathbf{x}_{1i}$ from the center of the points as represented by $\bar{\mathbf{x}}_1$.

9.3 OUTLIERS

In some cases, the model appears to be correct for most of the data, but one residual is much larger (in absolute value) than the others. Such an outlier may be due to an error in recording or may be from another population or may simply be an unusual observation from the assumed distribution. For example, if the errors ε_i are distributed as $N(0, \sigma^2)$, a value of ε_i greater than 3σ or less than -3σ would occur with frequency .0027.

If no explanation for an apparent outlier can be found, the dataset could be analyzed both with and without the outlying observation. If the results differ sufficiently to affect the conclusions, then both analyses could be maintained until additional data become available. Another alternative is to discard the outlier, even though no explanation has been found. A third possibility is to use *robust* methods that accommodate the outlying observation (Huber 1973, Andrews 1974, Hampel 1974, Welsch 1975, Devlin et al. 1975, Mosteller and Turkey 1977, Birch 1980, Krasker and Welsch 1982).

One approach to checking for outliers is to plot the residuals $\hat{\varepsilon}_i$ versus $\hat{y}_i$ or versus i, the observation number. In our examination of residuals, we need to keep in mind that by (9.8), the variance of the residuals is not constant:

$$\text{var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}). \tag{9.23}$$

By Theorem 9.2(i), $h_{ii} \leq 1$; hence, $\text{var}(\hat{\varepsilon}_i)$ will be small if h_{ii} is near 1. By Theorem 9.2(iii), h_{ii} will be large if $\mathbf{x}_{1i}$ is far from $\bar{\mathbf{x}}_1$, where $\mathbf{x}_{1i} = (x_{i1}, x_{i2}, \ldots, x_{ik})'$ and $\bar{\mathbf{x}}_1 = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)'$. By (9.23), such observations will tend to have small residuals, which seems unfortunate because the model is less likely to hold far from $\bar{\mathbf{x}}_1$. A small residual at a point where $\mathbf{x}_{1i}$ is far from $\bar{\mathbf{x}}_1$ may result because the fitted model will tend to pass close to a point isolated from the bulk of the points, with a resulting poorer fit to the bulk of the data. This may mask an inadequacy of the true model in the region of $\mathbf{x}_{1i}$.

An additional verification that large values of h_{ii} are accompanied by small residuals is provided by the following inequality (see Problem 9.4):

$$\frac{1}{n} \leq h_{ii} + \frac{\hat{\varepsilon}_i^2}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}} \leq 1. \tag{9.24}$$

For the reasons implicit in (9.23) and (9.24), it is desirable to scale the residuals so that they have the same variance. There are two common (and related) methods of scaling.

For the first method of scaling, we use $\text{var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$ in (9.23) to obtain the standardized residuals $\hat{\varepsilon}_i/\sigma\sqrt{1 - h_{ii}}$, which have mean 0 and variance 1. Replacing σ by s yields the *studentized residual*

$$r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1 - h_{ii}}}, \tag{9.25}$$

where $s^2 = \text{SSE}/(n - k - 1)$ is as defined in (7.24). The use of r_i in place of $\hat{\varepsilon}_i$ eliminates the location effect (due to h_{ii}) on the size of residuals, as discussed following (9.23).

A second method of scaling the residuals uses an estimate of σ that excludes the ith observation

$$t_i = \frac{\hat{\varepsilon}_i}{s_{(i)}\sqrt{1 - h_{ii}}}, \tag{9.26}$$

where $s_{(i)}$ is the standard error computed with the $n - 1$ observations remaining after omitting $(y_i, \mathbf{x}_i') = (y_{i1}, x_{i1}, \ldots, x_{ik})$, in which y_i is the ith element of $\mathbf{y}$ and $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. If the ith observation is an outlier, it will more likely show up as such with the standardization in (9.26), which is called the *externally studentized residual* or the *studentized deleted residual* or *R student*.

Another option is to examine the *deleted residuals*. The ith deleted residual, $\varepsilon_{(i)}$, is computed with $\hat{\boldsymbol{\beta}}_{(i)}$ on the basis of $n - 1$ observations with $(y_i, \mathbf{x}_i')$ deleted:

$$\hat{\varepsilon}_{(i)} = y_i - \hat{y}_{(i)} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}. \tag{9.27}$$

By definition

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}'\mathbf{y}_{(i)}, \tag{9.28}$$

where $\mathbf{X}_{(i)}$ is the $(n-1)\times\ (k+1)$ matrix obtained by deleting $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ik})$, the ith row of $\mathbf{X}$, and $\mathbf{y}_{(i)}$ is the corresponding $(n-1) \times 1$ $\mathbf{y}$ vector after deleting y_i. The deleted vector $\hat{\boldsymbol{\beta}}_{(i)}$ can also be found without actually deleting $(y_i, \mathbf{x}_i')$ since

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{\hat{\varepsilon}_i}{1 - h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \tag{9.29}$$

(see Problem 9.5).

The deleted residual $\hat{\varepsilon}_{(i)} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$ in (9.27) can be expressed in terms of $\hat{\varepsilon}_i$ and h_{ii} as

$$\hat{\varepsilon}_{(i)} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}} \tag{9.30}$$

(see Problem 9.6). Thus the n deleted residuals can be obtained without computing n regressions. The scaled residual t_i in (9.26) can be expressed in terms of $\hat{\varepsilon}_{(i)}$ in (9.30) as

$$t_i = \frac{\hat{\varepsilon}_{(i)}}{\sqrt{\widehat{\text{var}}(\varepsilon_{(i)})}} \tag{9.31}$$

(see Problem 9.7).

The deleted sample variance $s_{(i)}^2$ used in (9.26) is defined as $s_{(i)}^2 = \text{SSE}_{(i)}/(n - k - 2)$, where $\text{SSE}_{(i)} = \mathbf{y}_{(i)}'\mathbf{y}_{(i)} - \hat{\boldsymbol{\beta}}_{(i)}'\mathbf{X}_{(i)}'\mathbf{y}_{(i)}$. This can be found without excluding the ith observation as

$$s_{(i)}^2 = \frac{\text{SSE}_{(i)}}{n - k - 2} = \frac{\text{SSE} - \hat{\varepsilon}_i^2/(1 - h_{ii})}{n - k - 2} \tag{9.32}$$

(see Problem 9.8).

Another option for outlier detection is to plot the ordinary residuals $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ against the deleted residuals $\hat{\varepsilon}_{(i)}$ in (9.27) or (9.30). If the fit does not change substantially when the ith observation is deleted in computation of $\hat{\boldsymbol{\beta}}$, the plotted points should approximately follow a straight line with a slope of 1. Any points that are relatively far from this line are potential outliers.

If an outlier is from a distribution with a different mean, the model can be expressed as $E(y_i) = \mathbf{x}_i'\boldsymbol{\beta} + \theta$, where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. This is called the *mean-shift outlier model.* The distribution of t_i in (9.26) or (9.31) is $t(n - k - 1)$, and t_i can therefore be used in a test of the hypothesis $H_0 : \theta = 0$. Since n tests will be made, a Bonferroni adjustment to the critical values can be used, or we can simply focus on the largest t_i values.

The n deleted residuals in (9.30) can be used for model validation or selection by defining the *prediction sum of squares* (PRESS):

$$\text{PRESS} = \sum_{i=1}^{n} \hat{\varepsilon}_{(i)}^2 = \sum_{i=1}^{n} \left(\frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right)^2. \tag{9.33}$$

Thus, a residual $\hat{\varepsilon}_i$ that corresponds to a large value of h_{ii} contributes more to PRESS. For a given dataset, PRESS may be a better measure than SSE of how well the model will predict future observations. To use PRESS to compare alternative models when the objective is prediction, preference would be shown to models with small values of PRESS.

9.4 INFLUENTIAL OBSERVATIONS AND LEVERAGE

In Section 9.3, we emphasized a search for outliers that did not fit the model. In this section, we consider the effect that deletion of an observation $(y_i, \mathbf{x}_i')$ has on the estimates $\hat{\boldsymbol{\beta}}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}$. An observation that makes a major difference on these estimates is called an *influential observation.* A point $(y_i, \mathbf{x}_i')$ is potentially influential if it is an outlier in the y direction or if it is unusually far removed from the center of the x's.

We illustrate influential observations for the case of one x in Figure 9.2. Points 1 and 3 are extreme in the x direction; points 2 and 3 would likely appear as outliers in the y direction. Even though point 1 is extreme in x, it will not unduly influence the slope or intercept. Point 3 will have a dramatic influence on the slope and intercept since the regression line would pass near point 3. Point 2 is also influential, but much less so than point 3.

Thus, influential points are likely to be found in areas where little or no other data were collected. Such points may be fitted very well, sometimes to the detriment of the fit to the other data.

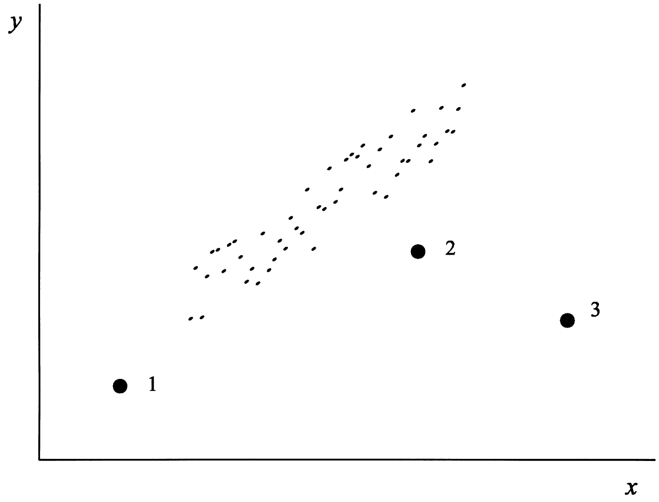

Figure 9.2 Simple linear regression showing three outliers.

To investigate the influence of each observation, we begin with $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ in (9.2), the elements of which are

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_i. \tag{9.34}$$

By (9.22), if h_{ii} is large (close to 1), then the h'_{ij}s, $j \neq i$, are all small, and y_i contributes much more than the other y's to $\hat{y}_i$. Hence, h_{ii} is called the *leverage* of y_i. Points with high leverage have high potential for influencing regression results. In general, if an observation $(y_i, \mathbf{x}'_i)$ has a value of h_{ii} near 1, then the estimated regression equation will be close to y_i; that is, $\hat{y}_i - y_i$ will be small.

By Theorem 9.2(iv), the average value of the h_{ii}'s is $(k+1)/n$. Hoaglin and Welsch (1978) suggest that a point with $h_{ii} > 2(k+1)/n$ is a high leverage point. Alternatively, we can simply examine any observation whose value of h_{ii} is unusually large relative to the other values of h_{ii}.

In terms of fitting the model to the bulk of the data, high leverage points can be either good or bad, as illustrated by points 1 and 3 in Figure 9.2. Point 1 may reduce the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$. On the other hand, point 3 will drastically alter the fitted model. If point 3 is not the result of a recording error, then the researcher must choose between two competing fitted models. Typically, the model that fits the bulk of the data might be preferred until additional points can be observed in other areas.

To formalize the influence of a point $(y_i, \mathbf{x}'_i)$, we consider the effect of its deletion on $\boldsymbol{\beta}$ and $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. The estimate of $\boldsymbol{\beta}$ obtained by deleting the ith observation $(y_i, \mathbf{x}'_i)$ is defined in (9.28) as $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}$. We can compare $\hat{\boldsymbol{\beta}}_{(i)}$ to $\hat{\boldsymbol{\beta}}$ by means of *Cook's distance*, defined as

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(k+1)s^2}. \tag{9.35}$$

This can be rewritten as

$$\begin{aligned} D_i &= \frac{(\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)} - \mathbf{X}\hat{\boldsymbol{\beta}})}{(k+1)s^2} \\ &= \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{(k+1)s^2}, \end{aligned} \tag{9.36}$$

in which D_i is proportional to the ordinary Euclidean distance between $\hat{\mathbf{y}}_{(i)}$ and $\hat{\mathbf{y}}$. Thus if D_i is large, the observation $(y_i, \mathbf{x}_i')$ has substantial influence on both $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{y}}$. A more computationally convenient form of D_i is given by

$$D_i = \frac{r_i^2}{k+1}\left(\frac{h_{ii}}{1-h_{ii}}\right) \tag{9.37}$$

TABLE 9.1 Residuals and Influence Measures for the Chemical Data with Dependent Variable y_1

Observation	y_i	$\hat{y}_i$	$\hat{\varepsilon}_i$	h_{ii}	r_i	t_i	D_i
1	41.5	42.19	−0.688	0.430	−0.394	−0.383	0.029
2	33.8	31.00	2.798	0.310	1.457	1.520	0.239
3	27.7	27.74	−0.042	0.155	−0.020	−0.019	0.000
4	21.7	21.03	0.670	0.139	0.313	0.303	0.004
5	19.9	19.40	0.495	0.129	0.230	0.222	0.002
6	15.0	12.69	2.307	0.140	1.076	1.082	0.047
7	12.2	12.28	−0.082	0.228	−0.040	−0.039	0.000
8	4.3	5.57	−1.270	0.186	−0.609	−0.596	0.021
9	19.3	20.22	−0.917	0.053	−0.408	−0.396	0.002
10	6.4	4.76	1.642	0.233	0.811	0.801	0.050
11	37.6	35.68	1.923	0.240	0.954	0.951	0.072
12	18.0	13.09	4.906	0.164	2.320	2.800	0.264
13	26.3	27.34	−1.040	0.146	−0.487	−0.474	0.010
14	9.9	13.51	−3.605	0.245	−1.795	−1.956	0.261
15	25.0	26.93	−1.929	0.250	−0.964	−0.961	0.077
16	14.1	15.44	−1.342	0.258	−0.674	−0.661	0.039
17	15.2	15.44	−0.242	0.258	−0.121	−0.117	0.001
18	15.9	19.54	−3.642	0.217	−1.780	−1.937	0.220
19	19.6	19.54	0.058	0.217	0.028	0.027	0.000

(see Problem 9.9). Muller and Mok (1997) discuss the distribution of D_i and provide a table of critical values.

Example 9.4. We illustrate several diagnostic tools for the chemical reaction data of Table 7.4 using y_1. In Table 9.1, we give $\hat{\varepsilon}_i$, h_{ii}, and some functions of these from Sections 9.3 and 9.4.

The guideline for h_{ii} in Section 9.4 is $2(k+1)/n = 2(4)/19 = .421$. The only value of h_{ii} that exceeds .421 is the first, $h_{11} = .430$. Thus the first observation has potential for influencing the model fit, but this influence does not appear in $t_1 = -.383$ and $D_1 = .029$. Other relatively large values of h_{ii} are seen for observations 2, 11, 14, 15, 16, and 17. Of these only observation 14 has a very large (absolute) value of t_i. Observation 12 has large values of $\hat{\varepsilon}_i$, r_i, t_i and D_i and is a potentially influential outlier.

The value of PRESS as defined in (9.33) is PRESS = 130.76, which can be compared to SSE = 80.17. □

PROBLEMS

9.1 Verify the following properties of the residual vector $\hat{\boldsymbol{\varepsilon}}$ as given in (9.7)–(9.14):

(**a**) $E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$

(**b**) $\text{cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$

(**c**) $\text{cov}(\hat{\boldsymbol{\varepsilon}}, \mathbf{y}) = \sigma^2(\mathbf{I} - \mathbf{H})$

(**d**) $\text{cov}(\hat{\boldsymbol{\varepsilon}}, \hat{\mathbf{y}}) = \mathbf{O}$

(**e**) $\bar{\hat{\boldsymbol{\varepsilon}}} = \sum_{i=1}^{n} \hat{\varepsilon}_i / n = 0$

(**f**) $\hat{\boldsymbol{\varepsilon}}'\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$

(**g**) $\hat{\boldsymbol{\varepsilon}}'\hat{\mathbf{y}} = 0$

(**h**) $\hat{\boldsymbol{\varepsilon}}'\mathbf{X} = \mathbf{0}'$

9.2 (**a**) In the proof of Theorem 9.2(ii), verify that the maximum value of $h_{ii} - h_{ii}^2$ is $\frac{1}{4}$.

(**b**) Prove Theorem 9.2(iii).

(**c**) Prove Theorem 9.2(iv).

9.3 Show that an alternative expression for h_{ii} in Theorem 9.2(iii) is the following:

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1) \sum_{r=1}^{k} \frac{1}{\lambda_r} \cos^2 \theta_{ir},$$

where θ_{ir} is the angle between $\mathbf{x}_{1i} - \bar{\mathbf{x}}_1$ and $\mathbf{a}_r$, the rth eigenvector of $\mathbf{X}_c'\mathbf{X}_c$ (Cook and Weisberg 1982, p. 13). Thus h_{ii} is large if $(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)$ is large or if θ_{ir} is small for some r.

9.4 Show that $\frac{1}{n} \leq h_{ii} + \hat{\varepsilon}_i^2/\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \leq 1$ as in (9.24). The following steps are suggested:

(**a**) Let $\mathbf{H}^*$ be the hat matrix corresponding to the augmented matrix $(\mathbf{X}, \mathbf{y})$. Then

$$\begin{aligned}\mathbf{H}^* &= (\mathbf{X},\ \mathbf{y})[(\mathbf{X},\mathbf{y})'(\mathbf{X},\mathbf{y})]^{-1}(\mathbf{X},\mathbf{y})' \\ &= (\mathbf{X},\mathbf{y})\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{y} \\ \mathbf{y}'\mathbf{X} & \mathbf{y}'\mathbf{y}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{X}' \\ \mathbf{y}'\end{pmatrix}.\end{aligned}$$

Use the inverse of a partitioned matrix in (2.50) with $\mathbf{A}_{11} = \mathbf{X}'\mathbf{X}$, $\mathbf{a}_{12} = \mathbf{X}'\mathbf{y}$, and $a_{22} = \mathbf{y}'\mathbf{y}$ to obtain

$$\begin{aligned}\mathbf{H}^* &= \mathbf{H} + \frac{1}{b}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{y}\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &\quad - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\mathbf{y}' + \mathbf{y}\mathbf{y}'] \\ &= \mathbf{H} + \frac{1}{b}[\mathbf{H}\mathbf{y}\mathbf{y}'\mathbf{H} - \mathbf{y}\mathbf{y}'\mathbf{H} - \mathbf{H}\mathbf{y}\mathbf{y}' + \mathbf{y}\mathbf{y}'],\end{aligned}$$

where $b = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

(**b**) Show that the above expression factors into

$$\mathbf{H}^* = \mathbf{H} + \frac{(\mathbf{I}-\mathbf{H})\mathbf{y}\mathbf{y}'(\mathbf{I}-\mathbf{H})}{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}} = \mathbf{H} + \frac{\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}'}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}},$$

which gives $h_{ii}^* = h_{ii} + \hat{\varepsilon}_i^2/\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$.

(**c**) The proof is easily completed by noting that $\mathbf{H}^*$ is a hat matrix and therefore $(1/n) \leq h_{ii}^* \leq 1$ by Theorem 9.2(i).

9.5 Show that $\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \hat{\varepsilon}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i/(1-h_{ii})$ as in (9.29). The following steps are suggested:

(**a**) Show that $\mathbf{X}'\mathbf{X} = \mathbf{X}_{(i)}'\mathbf{X}_{(i)} + \mathbf{x}_i\mathbf{x}_i'$ and that $\mathbf{X}'\mathbf{y} = \mathbf{X}_{(i)}'\mathbf{y}_{(i)} + \mathbf{x}_i y_i$.

(**b**) Show that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{(i)}'\mathbf{y}_{(i)} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i$.

(**c**) Using the following adaptation of (2.53)

$$(\mathbf{B} - \mathbf{c}\mathbf{c}')^{-1} = \mathbf{B}^{-1} + \frac{\mathbf{B}^{-1}\mathbf{c}\mathbf{c}'\mathbf{B}^{-1}}{1 - \mathbf{c}'\mathbf{B}^{-1}\mathbf{c}}.$$

show that

$$\hat{\boldsymbol{\beta}}_{(i)} = \left[(\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}\right]\mathbf{X}_{(i)}'\mathbf{y}_{(i)}.$$

(**d**) Using the result of parts (b) and (c), show that

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{\hat{\varepsilon}_i}{1 - h_{ii}} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i.$$

9.6 Show that $\hat{\varepsilon}_{(i)} = \hat{\varepsilon}_i/(1 - h_{ii})$ as in (9.30).

9.7 Show that $t_i = \hat{\varepsilon}_{(i)}/\sqrt{\widehat{\text{var}}(\hat{\varepsilon}_{(i)})}$ in (9.31) is the same as $t_i = \hat{\varepsilon}_i/s_{(i)}\sqrt{1 - h_{ii}}$ in (9.26). The following steps are suggested:

(**a**) Using $\hat{\varepsilon}_{(i)} = \hat{\varepsilon}_i/(1 - h_{ii})$ in (9.30), show that $\text{var}(\hat{\varepsilon}_{(i)}) = \sigma^2/(1 - h_{ii})$.

(**b**) If $\text{var}(\hat{\varepsilon}_{(i)})$ in part (a) is estimated by $\widehat{\text{var}}(\hat{\varepsilon}_{(i)}) = s^2_{(i)}/(1 - h_{ii})$, show that $\hat{\varepsilon}_{(i)}/\sqrt{\widehat{\text{var}}(\varepsilon_{(i)})} = \hat{\varepsilon}_i/s_{(i)}\sqrt{1 - h_{ii}}$.

9.8 Show that $\text{SSE}_{(i)} = \mathbf{y}'_{(i)}\mathbf{y}_{(i)} - \mathbf{y}'_{(i)}\mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}$ can be written in the form

$$\text{SSE}_{(i)} = \text{SSE} - \hat{\varepsilon}_i^2/(1 - h_{ii})$$

as in (9.32). One way to do this is as follows:

(**a**) Show that $\mathbf{y}'_{(i)}\mathbf{y}_{(i)} = \mathbf{y}'\mathbf{y} - y_i^2$.

(**b**) Using Problem 9.5a,d, we have

$$\mathbf{y}'_{(i)}\mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{y}'\mathbf{X} - y_i\mathbf{x}'_i)\left[\hat{\boldsymbol{\beta}} - \frac{\hat{\varepsilon}_i}{1 - h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\right].$$

Show that this can be written as

$$\mathbf{y}'_{(i)}\mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)} = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} - y_i^2 + \frac{\hat{\varepsilon}_i^2}{1 - h_{ii}}.$$

(**c**) Show that

$$\text{SSE}_{(i)} = \text{SSE} - \hat{\varepsilon}_i^2/(1 - h_{ii}).$$

9.9 Show that $D_i = r_i^2 h_{ii}/(k + 1)(1 - h_{ii})$ in (9.37) is the same as D_i in (9.35). This may be done by substituting (9.29) into (9.35).

9.10 For the gas vapor data in Table 7.3, compute the diagnostic measures $\hat{y}_i$, $\hat{\varepsilon}_i$, h_{ii}, r_i, t_i, and D_i. Display these in a table similar to Table 9.1. Are there outliers or potentially influential observations? Calculate PRESS and compare to SSE.

9.11 For the land rent data in Table 7.5, compute the diagnostic measures $\hat{y}_i$, $\hat{\varepsilon}_i$, h_{ii}, r_i, t_i, and D_i. Display these in a table similar to Table 9.1. Are there outliers or potentially influential observations? Calculate PRESS and compare to SSE.

9.12 For the chemical reaction data of Table 7.4 with dependent variable y_2, compute the diagnostic measures $\hat{y}_i$, $\hat{\varepsilon}_i$, h_{ii}, r_i, t_i, and D_i. Display these in a table similar to Table 9.1. Are there outliers or potentially influential observations? Calculate PRESS and compare to SSE.