CHAPTER 8

More on Checking Fitted Models

We have already discussed the basic residuals plots in Chapter 2, and the Durbin–Watson test for serial correlation in Chapter 7. In many regression problems, these checks are perfectly adequate. In this chapter we discuss further techniques that can be of value. Readers should form their own opinions about which methods are likely to be useful in their own regression applications.

8.1. THE HAT MATRIX H AND THE VARIOUS TYPES OF RESIDUALS

Up to this point we have featured only the "ordinary residuals," that is, the $e_i = Y_i - \hat{Y}_i$ values. These are perfectly adequate for residuals checks on most regressions. It is also possible to look at other types of residual quantities. Some workers would look at these other types routinely; some only when they wanted to submit the regression fit to more study. These other types of residuals have been created to overcome problems that are occasionally concealed by the ordinary $e_i = Y_i - \hat{Y}_i$ residuals. We first remind the reader of the notation: the model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the normal equations are $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$ with solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, if $\mathbf{X}'\mathbf{X}$ is nonsingular. The fitted values are $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$, say, where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{8.1.1}$$

This matrix $\mathbf{H}$ occurs repeatedly in regression work. It is usually called the "hat matrix," as mentioned below, above (8.1.11).

The elements of $\mathbf{H}$ will be denoted by h_{ij}, namely,

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix} \tag{8.1.2}$$

$\mathbf{H}$ is symmetric, that is, $\mathbf{H}' = \mathbf{H}$, so that $h_{ij} = h_{ji}$. $\mathbf{H}$ is also idempotent, that is,

$$\begin{aligned} \mathbf{H}^2 = \mathbf{H}\mathbf{H} &= \{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{H}. \end{aligned} \tag{8.1.3}$$

(In fact, $\mathbf{H}^p = \mathbf{H}$, where p is any power.)

The residuals can be expressed as

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}. \tag{8.1.4}$$

(The matrix $\mathbf{I} - \mathbf{H}$ is also symmetric and idempotent, by the way. We show this below.)

Variance–Covariance Matrix of e

Since $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, and because $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$, it follows that

$$\mathbf{e} - E(\mathbf{e}) = (\mathbf{I} - \mathbf{H})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon} \tag{8.1.5}$$

and the variance–covariance matrix of $\mathbf{e}$ is defined as

$$\mathbf{V}(\mathbf{e}) = E\{[\mathbf{e} - E(\mathbf{e})][\mathbf{e} - E(\mathbf{e})]'\} = (\mathbf{I} - \mathbf{H})E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')(\mathbf{I} - \mathbf{H})'. \tag{8.1.6}$$

Now $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \mathbf{V}(\boldsymbol{\epsilon}) = \mathbf{I}\sigma^2$ if $E(\boldsymbol{\epsilon}) = \mathbf{0}$, as we usually assume, and when unweighted least squares are used. Furthermore, $(\mathbf{I} - \mathbf{H})' = (\mathbf{I} - \mathbf{H}') = \mathbf{I} - [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I} - \mathbf{H}$. Thus the matrix $\mathbf{I} - \mathbf{H}$ is symmetric, and

$$\begin{aligned} \mathbf{V}(\mathbf{e}) &= (\mathbf{I} - \mathbf{H})\mathbf{I}\sigma^2(\mathbf{I} - \mathbf{H})' \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\sigma^2 \\ &= (\mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{HH})\sigma^2 \\ &= (\mathbf{I} - \mathbf{H})\sigma^2 \end{aligned} \tag{8.1.7}$$

since $\mathbf{HH} = \mathbf{H}^2 = \mathbf{H}$. Thus $V(e_i)$ is given by the ith diagonal element $1 - h_{ii}$, and $\text{cov}(e_i, e_j)$ is given by the (i, j)th element $-h_{ij}$ of the matrix $(\mathbf{I} - \mathbf{H})\sigma^2$. The correlation between e_i and e_j is given by

$$\rho_{ij} = \frac{\text{cov}(e_i, e_j)}{\{V(e_i) \cdot V(e_j)\}^{1/2}} = \frac{-h_{ij}}{\{(1 - h_{ii})(1 - h_{jj})\}^{1/2}}. \tag{8.1.8}$$

The values of these correlations thus depend entirely on the elements of the matrix $\mathbf{X}$, since σ^2 cancels. [In situations where we "design our experiment," that is, choose our $\mathbf{X}$ matrix, we thus have the opportunity to affect these correlations. We cannot get all zero correlations, of course, because the n residuals carry only $(n - p)$ degrees of freedom and are linked by the normal equations.]

Other Facts About H

1. $$\begin{aligned} \text{SS (all parameters)} = \text{SS}(\mathbf{b}) &= \mathbf{b}'\mathbf{X}'\mathbf{Y} \\ &= \hat{\mathbf{Y}}'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{H}'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{HY} \\ &= \mathbf{Y}'\mathbf{H}^2\mathbf{Y} \\ &= \hat{\mathbf{Y}}'\hat{\mathbf{Y}}. \end{aligned} \tag{8.1.9}$$

2. The average $V(\hat{Y}_i)$ over all the data points is

$$\sum_{i=1}^{n} V(\hat{Y}_i)/n = \text{trace } (\mathbf{H}\sigma^2)/n = p\sigma^2/n, \tag{8.1.10}$$

where p is the number of parameters. (See Exercise R in "Exercises for Chapters 5 and 6.")

3. $\mathbf{H1} = \mathbf{1}$ when the model contains a β_0 term. This means that every row of $\mathbf{H}$ adds up to 1. So does every column, since $\mathbf{H}$ is symmetric. Note that $\mathbf{1}' = \mathbf{1}'\mathbf{H}' = \mathbf{1}'\mathbf{H}$ and $\mathbf{1}'\mathbf{H}\mathbf{1} = n$.

4. Because $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$, $\mathbf{H}$ is often called the "hat matrix," that is, the matrix that converts Y's into $\hat{Y}$'s. The diagonal elements are often called the leverages, since examination of

$$\hat{Y}_i = h_{ii}Y_i + \sum_{j \neq i} h_{ij}Y_j \tag{8.1.11}$$

indicates via h_{ii} how heavily Y_i contributes to $\hat{Y}_i$. The messages from the leverages are not clear-cut. Cook and Weisberg (1982, p. 15) say that "for any $h_{ii} > 0$, $\hat{Y}_i$ will be dominated by $h_{ii}Y_i$ if Y_i is sufficiently different from the other elements of $\mathbf{Y}$ (that is, an outlier)." Hadi (1992, p. 5) says that "because high-leverage points are outliers in the X-space, some authors define leverage in terms of outlyingness in the X-space. However, leverage and outlyingness in the X-space are two different concepts .... High-leverage points are outliers in the X-space but the converse is not necessarily true." We shall deemphasize use of the leverages.

Internally Studentized Residuals[1]

It is clear from (8.1.7) that $V(e_i) = (1 - h_{ii})\sigma^2$, and these may well vary considerably. Usually σ^2 would be estimated by s^2, the residual mean square, that is, by

$$s^2 = \mathbf{e}'\mathbf{e}/(n - p) = \Sigma e_i^2/(n - p). \tag{8.1.12}$$

We can thus *studentize* the residuals by defining

$$s_i = \frac{e_i}{s(1 - h_{ii})^{1/2}}, \tag{8.1.13}$$

namely, by dividing each residual by its standard error $\{\hat{V}(e_i)\}^{1/2} = \{(1 - h_{ii})s^2\}^{1/2}$. These studentized residuals are said to be *internally studentized* because the s has, within it, e_i itself. So e_i is both on top and (concealed) underneath.

Extra Sum of Squares Attributable to e_i

We recall that, since $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$,

$$\begin{aligned} e_i &= -h_{i1}Y_1 - h_{i2}Y_2 - \cdots + (1 - h_{ii})Y_i \cdots - h_{in}Y_n \\ &= \mathbf{c}'\mathbf{Y}, \end{aligned} \tag{8.1.14}$$

say, where $\mathbf{c}' = (-h_{i1}, -h_{i2}, \ldots, (1 - h_{ii}), \ldots - h_{in})$. Now

[1]Some versions of MINITAB call these standardized residuals. They are obtained by allocating a column for them in the regression command.

$$\begin{aligned}\mathbf{c}'\mathbf{c} &= \sum_{j=1}^{n} h_{ij}^2 + (1 - 2h_{ii}) \\ &= (1 - h_{ii}).\end{aligned} \tag{8.1.15}$$

This follows because the summation is the ith row of $\mathbf{H}$ multiplied by the ith column of $\mathbf{H}$, which gives h_{ii} by the idempotency of $\mathbf{H}$. So, by (6.3.3) applied to (8.1.14), the 1 df extra SS for e_i is

$$\mathrm{SS}(e_i) = e_i^2/(1 - h_{ii}). \tag{8.1.16}$$

Thus

$$s^2(i) = \frac{(n - p)s^2 - e_i^2/(1 - h_{ii})}{n - p - 1} \tag{8.1.17}$$

provides an estimate of σ^2 after deletion of the contribution of e_i.

Externally Studentized Residuals[2]

Analogously to (8.1.13), we can define

$$t_i = \frac{e_i}{s(i)(1 - h_{ii})^{1/2}} \tag{8.1.18}$$

as the *externally studentized* residuals. An advantage of this is that, if e_i is large, it is thrown into emphasis even more by the fact that $s(i)$ has excluded it. The t_i follow a $t(n - p - 1)$ distribution under the usual normality of errors assumption.

Example. Consider again the data of Table 2.1 used to illustrate pure error. We show the calculation details for the influential observation Y_{23}. The ordinary residual is 2.36; $h_{23,23} = 0.158$, $1 - h_{23,23} = 0.842$; $s^2 = 0.7275$,

$$s^2(23) = \{21(0.7275) - (2.36)^2/0.842\}/20 = 0.4336,$$

much smaller than s^2. The internally studentized residual is

$$s_{23} = 2.36/\{0.7275(0.842)\}^{1/2} = 3.01.$$

The externally studentized residual is

$$t_{23} = 2.36/\{0.4336(0.842)\}^{1/2} = 3.90.$$

We see how external studentization has placed the t-residual nearly four standard errors away from zero rather than about three, and has thus given it extra prominence. (Of course, it was already large enough to notice at three standard errors, in this example!)

Calculations like those above are carried out automatically by many programs and the reader does not actually have to go through the details.

The amended residuals can be used in place of the ordinary residuals in any of the plots mentioned in Chapter 2. Comparison of e_i, s_i, and t_i plots is sometimes useful.

[2]Some versions of MINITAB call these the t-residuals. Their derivation is also described differently in MINITAB source material, where they are described as $\{Y_i - \hat{Y}(i)\}/\{\mathrm{MSE}_i + V(\hat{Y}(i))\}^{1/2}$, where $\hat{Y}(i)$ is predicted from a regression on data omitting Y_i and $\mathrm{MSE}_i = s^2(i)$. They are, nevertheless, identical to the t_i we describe. A subcommand "tresids C19" gives them in column 19, say.

Other Comments

Gray and Woodall (1994) mention several issues.

1. The internally studentized residuals s_i are such that the marginal distribution of $s_i^2/(n-p)$ is $\beta(\frac{1}{2}, (n-p-1)/2)$, which implies that $|s_i|$ cannot exceed $(n-p)^{1/2}$. This bound is reached only when deleting the ith set of data results in a perfect fit to the data that remain. When the residual degrees of freedom $(n-p)$ are small, no $|s_i|$ residual can be very large as a consequence.
2. Max $|s_i|$ can be tested, if desired; references are given.
3. The externally studentized residuals t_i are not bounded and t_i has a marginal t-distribution with $(n-p-1)$ df.

LaMotte (1994) covers some of the same ground from the viewpoint of when ratio statistics are actually t-variables and when not.

Berk and Booth (1995) indicate that several types of diagnostic plots can be useful to detect curvature omitted in a first-order model, but all of the diagnostics can be misleading in certain circumstances.

8.2. ADDED VARIABLE PLOT AND PARTIAL RESIDUALS

Added Variable Plot

We suggested in Section 2.6 that residuals could also be plotted against variables that are new candidates for entry. For example, if $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$ is a column of values of a variable observed with the rest of the data, we could plot the e_i versus the z_i. We can also first regress $\mathbf{z}$ on the other X's in the model and get its residual vector

$$\mathbf{e}_z = \mathbf{z} - \hat{\mathbf{z}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})\mathbf{z}.$$

This will remove from $\mathbf{z}$ any effect due to the columns of $\mathbf{X}$. We can then plot $\mathbf{e}$ versus $\mathbf{e}_z$. Such a plot is called an *added variable plot*. Note that the plot is centered around the origin since both variables are residuals from a model with β_0 in it. Assessing the slope of a straight line through the origin in this plot is essentially an assessment of the value of adding z to the group of regression predictors. So this could also be done directly using an extra sum of squares F-test.

Partial Residuals

Partial residuals are residuals that have not been adjusted for a particular X variable. Suppose we focus on X_i and rewrite $\mathbf{X} = (\mathbf{X}_1, \mathbf{x}_i)$ and

$$\hat{\mathbf{Y}} = \mathbf{X}_1\mathbf{b}_1 + x_i b_i, \tag{8.2.2}$$

where correspondingly

$$\mathbf{b} = (\mathbf{b}_1', b_i)' = \begin{pmatrix} \mathbf{b}_1 \\ b_i \end{pmatrix}.$$

Then the set of partial residuals for X_i would be

$$\mathbf{e}_i^* = \mathbf{Y} - \mathbf{X}_1\mathbf{b}_1 \tag{8.2.3}$$
$$= \mathbf{e} + x_i b_i$$

where $\mathbf{e}$ is the usual vector of residuals $\mathbf{Y} - \hat{\mathbf{Y}}$. A plot of e_i^* versus X_i has slope b_i and the vector $\mathbf{e}$ would provide the residuals if a straight line fit were made.

(This has also been called a "component plus residual plot." A variation adds back a quadratic term in X_i to the residuals as well as a first-order term.)

8.3. DETECTION OF INFLUENTIAL OBSERVATIONS: COOK'S STATISTICS

First we consider the (rather extreme) example where we fit a straight line to a set of data consisting of five observations, four at $X = a$, and one at $X = b$. If $V(Y_i) = \sigma^2$ we can show that, at $X = a$, $V(e_i) = 0.75\sigma^2$, $i = 1, 2, 3, 4$, while, at $X = b$, $V(e_5) = 0$. At first sight, a zero variance seems very desirable but it in fact arises because the fitted straight line is determined as the join of the average level of Y at $X = a$ and the single observed level of Y at $X = b$. The residual at $X = b$ is zero whatever the value of the corresponding Y and, in fact, the parameter estimates depend heavily on this particular observation. A large error in this observation is not detectable in the model-fitting process, and examination of the residuals would not reveal it either, if it existed. The observation at $X = b$ is an extremely *influential* one, whether it is correct or not.

The fact that an observation provides a large outlier is not, of course, good, but it does not necessarily mean that the observation is influential in fitting the chosen model. For example, in Figure 8.1, where the data of Table 8.1 are plotted, we see that the observation marked 19 will certainly be an outlier for most simple models fitted through the data. Its possible influence is moderated by the fact that there are

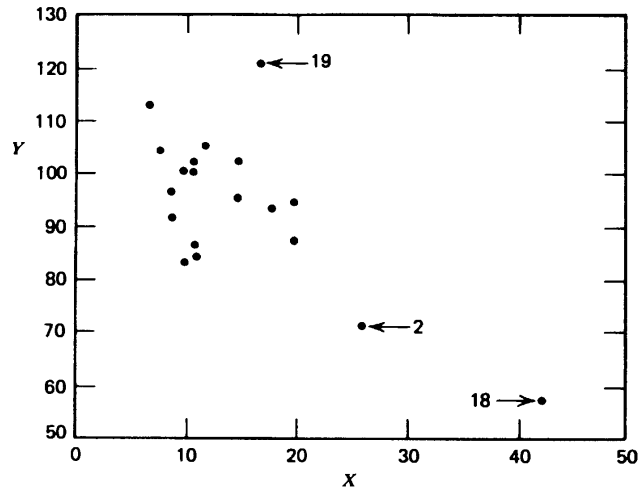

Figure 8.1. A regression with an observation (19) that may not be influential and one (18) that may well be. X represents the age of a child at first word (in months) and Y represents the child's score on an aptitude test. Reproduced by permission from Andrews and Pregibon (1978). The original data were recorded by Dr. L. M. Linde of UCLA and were given by Mickey, Dunn, and Clark (1967). See Table 8.1 for the data.

TABLE 8.1. Age at First Word (X) and Gesell Adaptive Score (Y)

Case	X	Y
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

Source: Data from Mickey, Dunn, and Clark (1967) but recorded by L. M. Linde of UCLA.

other observations at neighboring X-values. Observation 18, on the other hand, could well be an influential one. Being alone in its territory, it may have a major influence on the position of the fitted model there. It may or may not have a large residual, depending on the model fitted and the rest of the data.

In any data set where the estimation of one or more parameters depends heavily on a very small number of the observations, problems of interpretation can arise. One way to anticipate such problems is to check whether the deletion of observations greatly affects the fit of the model and the subsequent conclusions. If it does, the conclusions are shaky and more data are probably needed.

Cook (1977) proposed that the influence of the ith data point be measured by the squared scaled distance

$$D_i = (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}(i))'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}(i))/(ps^2). \tag{8.3.1}$$

Here, $\hat{\mathbf{Y}} = \mathbf{Xb}$ is the usual vector of predicted values, while $\hat{\mathbf{Y}}(i) = \mathbf{Xb}(i)$ is the vector of predicted values from a least squares fit when the ith data point is deleted, where $\mathbf{b}(i)$ is the corresponding least squares estimator.

When $\mathbf{v}$ is a vector, $\mathbf{v}'\mathbf{v}$ is the squared length of that vector, so D_i is the squared distance between (the ends of) the vectors $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Y}}(i)$, divided by ps^2. If omission of the ith observation makes little difference to the fitted values, D_i will be small. Large D_i indicate observations whose deletion greatly affects the fitted values. Because $\hat{\mathbf{Y}} - \hat{\mathbf{Y}}(i) = \mathbf{X}\{\mathbf{b} - \mathbf{b}(i)\}$, we can also write

$$D_i = \{\mathbf{b} - \mathbf{b}(i)\}'\mathbf{X}'\mathbf{X}\{\mathbf{b} - \mathbf{b}(i)\}/(ps^2). \tag{8.3.2}$$

A third representation of D_i is the form

$$D_i = \left\{\frac{e_i}{s(1-h_{ii})^{1/2}}\right\}^2 \left\{\frac{h_{ii}}{1-h_{ii}}\right\} \frac{1}{p}, \tag{8.3.3}$$

where e_i is the ith residual when the full data set is used, s^2 is the estimate of the variance $V(Y_i) = \sigma^2$ provided by the residual mean square when the full data set is used, and h_{ii} is the ith diagonal entry of the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We see that the first factor in Eq. (8.3.3) is the ith internally studentized residual, that is, the residual divided by its standard error (see Section 8.1), while the second term is the ratio (variance of the ith predicted value)/(variance of the ith residual). Note that $0 \leq h_{ii} \leq 1$. D_i can be large if either the first or second factor is large, and these factors measure two separate characteristics of each data point.

For our example above, with four observations at $X = a$, one at $X = b$,

$$D_i = \left\{\frac{e_i^2}{0.75s^2}\right\}\left(\frac{1}{3}\right)\left(\frac{1}{2}\right), \qquad i = 1, 2, 3, 4, \tag{8.3.4}$$

and

$$D_5 = \text{indeterminate},$$

where, for $i = 1, 2, 3, 4$, each $e_i = Y_i - \overline{Y}_a$, where $\overline{Y}_a = (Y_1 + Y_2 + Y_3 + Y_4)/4$ and $e_5 = 0$. The fifth observation, at $X = b$, is thus "flagged" as being a peculiar one and examination of the circumstances reveals its overwhelming influence.

For the Mickey, Dunn, and Clark data of Table 8.1, the values of the Cook's statistics for omission of observations 1, 2, . . . , 21 are 0.01 times, respectively,

0, 8, 7, 3, 2, 0, 0, 0, 0, 2, 5, 0, 7, 5, 0, 0, 2, 68, 22, 3, 0.

Observation 18 is very influential, and 19 is somewhat influential. There is no formal test for this. We simply compare the sizes of the "big ones" with the base level indicated by the bulk of the numbers. This base level is in the single digits here and the statistics 68 (for observation 18) and 22 (for observation 19) are clearly higher. (Some writings suggest that the Cook's statistics follow an F-distribution, but this is incorrect.) (What happens to the fit when 18 is omitted is shown in Figure 8.2.)

Evaluation of the residuals for these data show that observation 19 appears to be an outlier. Because X is the age (in months) of a child at first word, it is clearly not sensible to plan to collect data to fill in the gap between the lower X-values where most of the data occur, and the X-values of observations 2 and 18, which may not be reliable anyway. If we omit 2 and 18 as atypical X-values and omit 19 as an outlier, the message sent by the data is much reduced. That could well be the appropriate course of action here.

Evaluation of D_i as a routine technique is recommended. It is widely available as a standard option on many regression systems. In MINITAB, the subcommand "cookd C20;" will place the Cook's statistics in column 20, for example.

Higher-Order Cook's Statistics

Cook's statistics can also be evaluated for omitted pairs, omitted triplets, and so on, in an obvious extension of (8.3.1) and (8.3.2). Much more calculation is required in these cases and this is not routine. Here are a few example calculations of Cook's

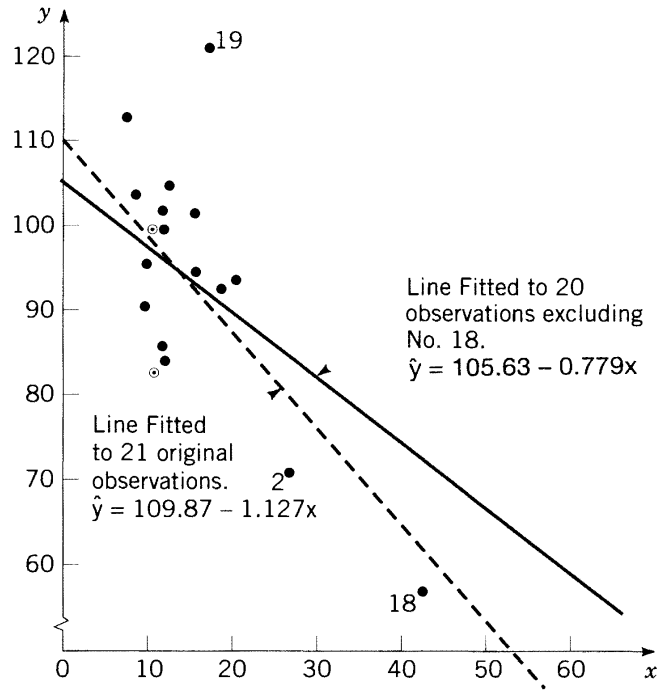

Figure 8.2. Plot of the Mickey, Dunn, and Clark (1967) data together with fitted lines that include and exclude data point 18.

statistics for some omitted pairs of observations from the data of Table 8.1. (These numbers are in their original scaling and have *not* been multiplied by 100 to clear decimals.)

Omitted	Cook's Statistic
18, 2	6.37
18, 3	0.48
18, 11	1.52
18, 19	0.15
19, 2	0.10
19, 3	0.12
19, 11	0.41

Clearly (18, 2) is the most influential pair. The statistic's low value for the pair (18, 19) comes as somewhat of a surprise at first sight. Each exerts a clockwise pull on the line but when both are omitted together, the line does not move very much.

Omitting more than one observation at a time can be useful in cases where two or more points conceal each or one another in a single deletion check. For example, if there were a point (let us call it 22) close to point 18 in the Table 8.1 data, omission of 18 might not give a large Cook's statistic value because 22 would "hold the line close" to its original position. Similarly, omission of 22 alone might have the same result. Omission of both 18 and 22 would, however, reveal them as an influential pair, if that were the case.

Another Worked Example

For the data used to illustrate the lack of fit test in Section 2.1, the Cook's statistics are 0.01 times these numbers:

$$4, 0, 7, 4, 0, 8, 2, 4, 3, 0, 0, 1, 0, 4, 6, 1, 0, 5, 0, 0, 1, 2, 85.$$

The last observation shows up as very influential. We have already mentioned its strangeness in Section 2.1, where it showed up clearly in the plot of Figure 2.2. Evaluation of Cook's statistics usually enables us to pinpoint such observations even when we have not been able to look at a plot, for example, in a case with multiple predictors.

Plots

Some suggestions for plotting the values of Cook's statistics are given by Hines and Hines (1995).

8.4. OTHER STATISTICS MEASURING INFLUENCE

The DFFITS Statistics

These statistics are cousins of Cook's influence measures. They are defined by Belsley, Kuh, and Welsch (1980) as

$$\text{DFFITS}_i = [\{\mathbf{b} - \mathbf{b}(i)\}'\mathbf{X}'\mathbf{X}\{\mathbf{b} - \mathbf{b}(i)\}/s^2(i)]^{1/2} \tag{8.4.1}$$

$$= [D_i p s^2/s^2(i)]^{1/2}, \tag{8.4.2}$$

where D_i is Cook's statistic (8.3.1). [For $s^2(i)$, see Section 8.1. It is an estimate of σ^2 obtained after deletion of the contribution of the ith residual.]

Atkinson's Modified Cook's Statistics

By replacing s^2 by $s^2(i)$, scaling by a factor $(n - p)/p$ instead of $1/p$, and taking the square root, we obtain another set of relatives of Cook's statistics:

$$A_i = [\{\mathbf{b} - \mathbf{b}(i)\}'\mathbf{X}'\mathbf{X}\{\mathbf{b} - \mathbf{b}(i)\}(n - p)/\{ps^2(i)\}]^{1/2} \tag{8.4.3}$$

$$= [D_i(n - p)s^2/s^2(i)]^{1/2} \tag{8.4.4}$$

$$= \text{DFFITS}_i\{(n - p)/n\}^{1/2}. \tag{8.4.5}$$

See Atkinson (1985).

Cook's, DFFITS, and Atkinson's statistics tell parallel stories for most sets of data. We suggest you go with your preference. While we favor Atkinson's for technical reasons, we usually go with Cook's, which is more widely available.

8.5. REFERENCE BOOKS FOR ANALYSIS OF RESIDUALS

How much effort should one put into examining residuals? It depends on the circumstances. If your regression is one of a series, and tomorrow you move on to other experiments and to other data, perhaps not too much effort beyond the standard plots

we have described is needed. On the other hand, if you spent a year in an African jungle and emerged with 61 precious observations, you may wish to spend another year examining all their aspects. In this case, you may need to consult some of the excellent books listed below for sophisticated analyses we have not told you about. The same references appear in Section 2.8.

Atkinson (1985); Barnett and Lewis (1994); Belsley (1991); Belsley, Kuh, and Welsch (1980); Chatterjee and Hadi (1988); Cook and Weisberg (1982); Hawkins (1980); Rousseeuw and Leroy (1987).

EXERCISES FOR CHAPTER 8

A. Find the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the three types of residuals (ordinary, s-residuals, and t-residuals) for any of the regression examples or exercises in this book. Check the characteristics and properties of $\mathbf{H}$ given in Section 8.1. Compare standard plots of the three types of residuals, for example, versus order, overall, or versus $\hat{Y}$. For most (but not all) regressions, it makes little difference which type of residual is used in these plots. When the plots differ materially, the reasons why should be explored. Check for outliers and influential points.

B. If you were asked to fit a straight line to the data

$$(X, Y) = (1, 3), (2, 2.5), (2, 1.2), (3, 1), \text{ and } (6, 4.5),$$

what would you say about it?

C. Find the Cook's statistics for any of the regression examples or exercises in this book. Are any of the observations influential? What can be done if they are?

D. (*Source:* "The hat matrix in regression and anova," by D. C. Hoaglin and R. E. Welsch. *The American Statistician,* **32,** 1978, 17–22. See also p. 146.) We recall that $\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{HY}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (sometimes called the *hat matrix*) is symmetric and idempotent. Show that each of the diagonal terms h_{uu} (sometimes called the *leverage* of Y_u on $\hat{Y}_u$) for $u = 1, 2, \ldots, n$ lies between 0 and 1, that the n h_{uu} add to p, the number of parameters (note that $\mathbf{X}$ is $n \times p$), and that, if Y_u is replaced by $Y_u^* = Y_u + 1$ in the regression calculation, then $\hat{Y}_u^* = \hat{Y}_u + h_{uu}$.