Statistical Science
1986, Vol. 1, No. 3, 379–416

Influential Observations, High Leverage Points, and Outliers in Linear Regression

Samprit Chatterjee and Ali S. Hadi

Abstract. A bewilderingly large number of statistical quantities have been proposed to study outliers and influence of individual observations in regression analysis. In this article we describe the inter-relationships which exist among the proposed measures. An examination of these relationships leads us to conclude that only three of these measures along with some graphical displays can provide an analyst a complete picture of outliers (major discrepant points) and points which excessively influence the fitted regression equation. Illustrative examples based on real data are presented.

Key words and phrases: Influence, leverage, outliers, regression diagnostics, residuals.

1. NOTATION

We consider a multiple linear regression model:

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is an $N \times 1$ vector of values of the response (dependent) variable, X is an $N \times p$ full-column rank matrix of known predictors (carriers, factors, regressors, explanatory variables) possibly including one constant predictor, β is a $p \times 1$ vector of unknown coefficients (parameters) to be estimated, and ε is an $N \times 1$ vector of independent random variables each with zero mean and unknown variance σ^2.

Following standard notation such as that in Velleman and Welsch (1981), we use y_i and x_i to denote the ith row of Y and X, respectively, and X_j to denote the jth column of X. By the ith observation we mean $(x_i : y_i)$, that is, the ith row in the matrix $(X : Y)$. We also use the subscript notation "(i)" or "$[j]$" to indicate the deletion of the ith observation or the jth variable, respectively. Thus, for example $X_{(i)}$ is the matrix X with the ith row deleted, $X_{[j]}$ is the matrix X with the jth column deleted, and $\hat{\beta}_{(i)}$ is the estimated parameter vector when the ith observation is deleted. We reserve the symbols $\hat{Y}$, e, and SSE to denote the vector of fitted values, the vector of residuals, and the residual sum of squares when Y is regressed on X, respectively, and the symbols R_j and W_j to denote the vectors of residuals when Y and X_j are regressed on $X_{[j]}$, respectively. Finally, we use M^{-1}, M^T, M^{-T} to denote the inverse, transpose, and inverse of the transpose of a matrix M, respectively.

Samprit Chatterjee is Professor of Statistics, New York University, 100 Trinity Place, New York, New York 10006. Ali S. Hadi is Assistant Professor of Statistics, Cornell University, NYSSILR, Ithaca, New York 14850.

2. INTRODUCTION

In fitting the multiple linear regression model (1) by the method of least squares, we have:

$$\hat{\beta} = (X^TX)^{-1}X^TY, \tag{2}$$

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X^TX)^{-1}, \tag{3}$$

$$\hat{Y} = X\hat{\beta} = PY, \tag{4}$$

where

$$P = X(X^TX)^{-1}X^T, \tag{5}$$

$$\mathrm{Var}(\hat{Y}) = \sigma^2P, \tag{6}$$

$$e = Y - \hat{Y} = (I - P)Y, \tag{7}$$

$$\mathrm{Var}(e) = \sigma^2(I - P), \tag{8}$$

and

$$\hat{\sigma}^2 = \frac{e^Te}{N - p}, \tag{9}$$

the residual mean square estimate of $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

It is well known that these (and other) quantities can be substantially influenced by one observation or a few observations; that is, not all the observations have an equal importance in least squares regression and, hence, in the conclusions that result from an analysis. It is, therefore, important for an analyst to be able to identify such observations and assess their effects on various aspects of the analysis. To this end,

several methods have been proposed in the statistical literature.

Before reviewing these methods, we first define what is meant by influence. A definition, which seems most appropriate, is given by Belsley, Kuh, and Welsch (1980):

> An influential observation is one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates ... than is the case for most of the other observations.

This definition, although of a subjective nature, implies that one should, at least, be able to order observations in a sensible way according to some measure of their influence.

An observation, however, may not have the same impact on all regression outputs. The question "Influence on what?" is, therefore, an important one. An observation may have influence on $\hat{\beta}$, a linear combination of $\hat{\beta}$, the estimated variance of $\hat{\beta}$, the fitted values, and/or the goodness-of-fit statistics. The primary goal of the analysis should determine which influence to consider. For example, if estimation of β is of primary concern, then measuring the influence of observations on $\hat{\beta}$ is appropriate; or if prediction is the primary goal, then measuring influence on the fitted values may be more appropriate than measuring influence on $\hat{\beta}$.

Influence measures are numerous. We review the most common ones and show the inter-relationships that exist among them. These measures can be classified into five groups:

1. Measures based on residuals,
2. Measures based on the prediction matrix,
3. Measures based on the volume of confidence ellipsoids,
4. Measures based on influence functions, and
5. Measures based on partial influence.

3. ANALYSIS OF RESIDUALS

One of the early methods of detecting model failures is examining the least squares residuals

$$e_i = y_i - x_i\hat{\beta} \tag{10}$$

where x_i is the ith row of X, or preferably, examining a scaled version of e_i, that is,

$$e_i(\sigma) = \frac{e_i}{\sigma\sqrt{1 - p_i}} \tag{11}$$

where p_i is the ith diagonal element of P (cf. (5)). Two special cases of (11) are:

$$t_i = e_i(\hat{\sigma}) = \frac{e_i}{\hat{\sigma}\sqrt{1 - p_i}} \tag{12}$$

where $\hat{\sigma}$ is as defined in (9), and

$$t_i^* \equiv e_i(\hat{\sigma}_{(i)}) = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - p_i}} \tag{13}$$

where

$$\hat{\sigma}^2_{(i)} = \frac{Y^T_{(i)}(I - P_{(i)})Y_{(i)}}{(N - p - 1)} = \frac{(N - p)\hat{\sigma}^2}{(N - p - 1)} - \frac{e_i^2}{(N - p - 1)(1 - p_i)} \tag{14}$$

is the residual mean square when the ith observation is omitted. Identity (14) was given by Beckman and Trussell (1974). We avoid using terminology here, because it is both confusing and conflicting. For example, (13) is called the "cross-validatory" or "jackknife" residuals by Atkinson (1981a), "RSTUDENT" by Belsley, Kuh, and Welsch (1980), and "studentized" residuals by Velleman and Welsch (1981).

Several authors, e.g., Velleman and Welsch (1981), Atkinson (1981a), and Belsley, Kuh, and Welsch (1980), prefer t_i^* over t_i for the following reasons:

1. t_i^* is the t-statistic for testing the significance of the coefficient of the ith unit vector u_i in the mean-shift outlier model $Y = X\beta + u_i\delta + \varepsilon$ (see, e.g., Belsley, Kuh, and Welsch (1980)) and under Gaussian assumptions, it follows a t-distribution with $(N - p - 1)$ degrees of freedom (df) (Beckman and Trussell (1974)) for which tables are available, whereas, $t_i^2/(N - p)$ follows a beta distribution (Ellenberg, 1973).

2. A little algebra will verify that (see Atkinson, 1981a):

$$t_i^* = t_i\sqrt{\{(N - p - 1)/(N - p - t_i^2\}}, \tag{15}$$

from which we see that t_i^* is a monotonic transformation of t_i and that $t_i^{*2} \to \infty$ as $t_i^2 \to (N - p)$. Therefore, t_i^* reflects large deviations more dramatically than does t_i.

3. The estimate $\hat{\sigma}^2_{(i)}$ is robust to problems of gross errors in the ith observation.

We now define what is meant by outliers in multiple linear regression. An outlier in the response-factor space is a point $(x_i{:}y_i)$ with large t_i or t_i^*. Outliers are usually detected by plotting t_i or t_i^* versus other variables such as Y, each X_j, and in serial order (see, e.g., Chatterjee and Price (1977), Seber (1977), Daniel and Wood (1980), and Draper and Smith (1981)).

An outlier need not be influential. As an example, consider the data given by Mickey, Dunn, and Clark (1967) and plotted in Figure 1. If a straight line regression model is fitted to the data, we see clearly that the observation marked by an "o" is an outlier. The fitted line, however, will hardly change if this

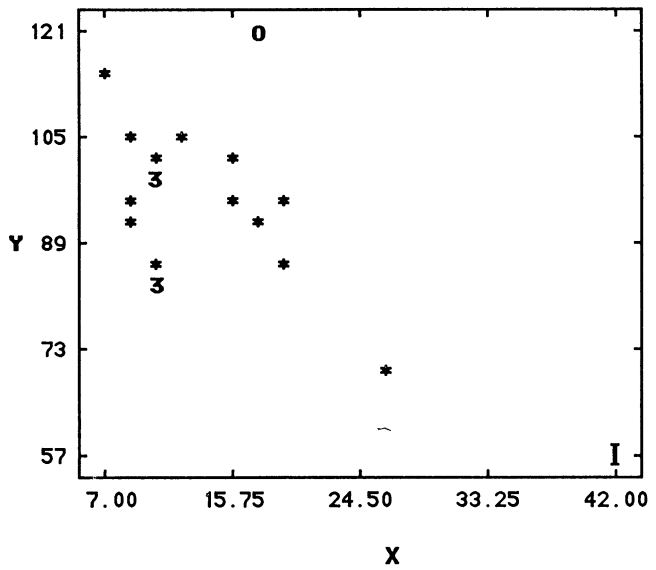

FIG. 1. *Plot of Mickey, Dunn, and Clark* (1967) *data. Y denotes a child's score in an aptitude test and X denotes the age of the child (in months) at first word.*

data point is deleted. This observation has little influence on $\hat{\beta}$. If an observation has little influence on the results, there is little point in agonizing over how deviant it appears. This example illustrates the existence of an outlier that does not matter (Andrews and Pregibon, 1978).

On the other hand, influential observations need not be outliers in the sense of having large residuals. The observation marked by "I" in Figure 1 illustrates this situation. Another (rare but good) example of this situation is found in Draper and Smith (1981). Consider fitting a straight line to a set of data consisting of five observations, four at $x = a$ and one at $x = b$. It is easy to show that the residual at $x = b$ is zero whatever the value of the corresponding y. This observation, however, is extremely influential due to the fact that one parameter estimate is completely determined by this observation, as can be seen in Figure 2, where if the y value at $x = b$ changed from the point marked * to the point marked o, a completely different line is obtained. This discussion points up the fact that examination of residuals alone will not detect aberrant or unusual observations such as the one indicated by I in Figure 1 and the one at $x = b$ in Figure 2. Graphical methods based on residuals alone will fail to detect these unusual points. Observations with these characteristics (small residuals and highly influential on the fit) often occur in real data (an example is given in Section 10). To study this problem we need the additional concept of "leverage," which we discuss in the next section.

4. THE PREDICTION MATRIX

The matrix P defined in (5) plays an important role, as can be seen in (4)–(8), in determining $\hat{Y}$, e, and their covariance matrices. The ith diagonal element of P,

$$p_i = x_i(X^TX)^{-1}x_i^T, \tag{16}$$

can be thought of as the amount of leverage of the response value y_i on the corresponding value $\hat{y}_i$. P is sometimes called the Hat matrix because it maps Y into $\hat{Y}$, i.e., $\hat{Y} = PY$. It is also a projection matrix because it generates the perpendicular projection of Y (an N-dimensional vector) into a p-dimensional subspace. We call it the prediction matrix, because applying it to Y produces the predicted values. Detailed discussion of the properties and importance of P in data analysis can be found in Hoaglin and Welsch (1978) and Cook and Weisberg (1982).

Hoaglin and Welsch (1978) recommended examination of p_i for high leverage design points and of t_i^* for outliers and suggested using $2p/N$ as a calibration point for p_i. For other calibration points, see Velleman and Welsch (1981).

We define a high leverage point in the factor space to be a point x_i with large p_i. Points which are isolated in the factor space (i.e., far removed from the main body of points in the X space) will have high leverage. They can be thought of as outliers in the factor space.

As with outliers, high leverage points need not be influential, and influential observations are not necessarily high leverage points. Two such examples can be found in Coleman (1977). If we augment the matrix X by the vector Y, that is

$$X^* = (X{:}Y), \tag{17}$$

the corresponding prediction matrix P^* of X^* is related to P by

$$P^* = P + \frac{(I - P)YY^T(I - P)}{Y^T(I - P)Y}. \tag{18}$$

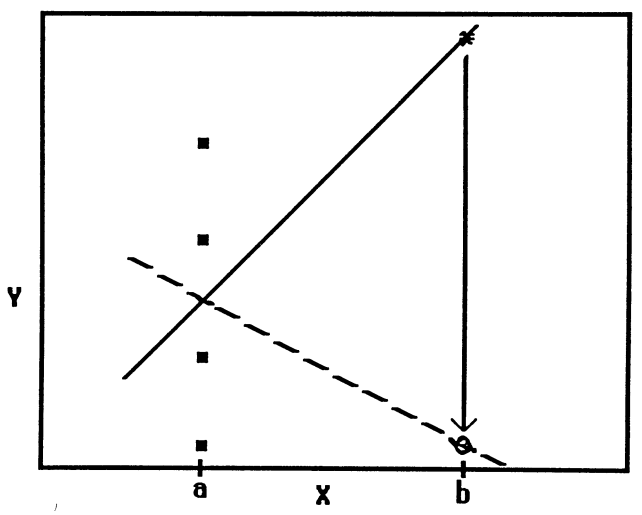

FIG. 2. *The point at $x = b$ is an extremely influential one, yet it has a zero residual regardless of the corresponding value of y.*

This can be proved by decomposing P^* into a sum of two orthogonal projections (see, e.g., Cook and Weisberg, 1982). Because X^* contains information about X and Y, one might be tempted to use

$$p_i^* = [x_i : y_i] \begin{bmatrix} X^TX & X^TY \\ Y^TX & Y^TY \end{bmatrix}^{-1} \begin{bmatrix} x_i^T \\ y_i \end{bmatrix} \tag{19}$$

as an influence measure. Using (18) we can write (19) as

$$p_i^* = p_i + \frac{e_i^2}{e^Te}. \tag{20}$$

From (20), however, we see that p_i^* will be large whenever p_i or e_i^2 is large. Hence, p_i^* cannot distinguish between two different situations: a high leverage point in the factor space and an outlier in the response-factor space. It is useful, therefore, to distinguish between sources of influence. An observation may influence (some or all) regression results because: 1) it is an outlying response value, 2) it is a high leverage point in the factor space, or 3) it is a combination of both.

This classification is helpful in a search for a remedial action. For example, an observation of type (1) may indicate inadequacies of distributional assumptions. An observation, of type (2), could be the most important one in the data set since it may provide the only information in a region where the ability to take observations is limited (Cook and Weisberg, 1980). If this data point is correct, then the only good remedial action, we believe, is to collect more data.

5. VOLUME OF CONFIDENCE ELLIPSOIDS

A measure of the influence of the ith observation on the estimated regression coefficients can be based on the change in volume of confidence ellipsoids with and without the ith observation. We review here four such measures.

5.1 Andrews-Pregibon Statistic

Andrews and Pregibon (1978) suggested using the ratio

$$\mathrm{AP}_i = \frac{|X_{(i)}^{*T}X_{(i)}^*|}{|X^{*T}X^*|} \tag{21}$$

to assess the relative change in $|X^{*T}X^*|$ when the ith observation is omitted. Small values of AP_i call for special attention. Note that $|X^{*T}X^*| = e^Te\,|X^TX|$. Also, AP_i is related to p_i^* defined in (19) and (20) by:

$$\mathrm{AP}_i = 1 - p_i - \frac{e_i^2}{e^Te} = 1 - p_i^*. \tag{22}$$

Hence, what applies to p_i^* also applies to AP_i; that is, AP_i does not distinguish between a high leverage point in the factor space and an outlier in the response-factor space.

5.2 The Likelihood Distance

Let $\mathrm{L}(\hat\beta)$ and $\mathrm{L}(\hat\beta_{(i)})$ be the log likelihood evaluated at $\hat\beta$ and $\hat\beta_{(i)}$, respectively. A measure of the influence of the ith observation on $\hat\beta$ can be based on the distance between $\mathrm{L}(\hat\beta)$ and $\mathrm{L}(\hat\beta_{(i)})$. Cook and Weisberg (1982) define the likelihood distance as

$$\begin{aligned} \mathrm{LD}_i &= 2[\mathrm{L}(\hat\beta) - \mathrm{L}(\hat\beta_{(i)})] \\ &= N \log\left[\left(\frac{N}{N-1}\right)\frac{N-p-1}{t_i^{*2}+N-p-1}\right] \\ &\quad + \frac{t_i^{*2}(N-1)}{(1-p_i)(N-p-1)} - 1. \end{aligned} \tag{23}$$

The likelihood distance is related to the asymptotic confidence region $\{\beta\colon 2[\mathrm{L}(\hat\beta) - \mathrm{L}(\beta)] \le \chi^2_{\alpha,p+1}\}$, where $\chi^2_{\alpha,p+1}$ is the upper α point of the χ^2 distribution with $(p+1)$ degrees of freedom. Consequently, LD_i is compared to χ^2_{p+1}. This measure of influence is based on the probability model used, whereas other measures of influence are strictly numerical.

5.3 Covariance Ratio

As suggested by Belsley, Kuh, and Welsch (1980), the influence of the ith observation on $\mathrm{Var}(\hat\beta)$ can be measured by comparing the ratio of $\det\{\mathrm{Var}(\hat\beta_{(i)})\}$ to $\det\{\mathrm{Var}(\hat\beta)\}$; that is,

$$\begin{aligned} \mathrm{CVR}_i &= \frac{\det\{\hat\sigma_{(i)}^2(X_{(i)}^TX_{(i)})^{-1}\}}{\det\{\hat\sigma^2(X^TX)^{-1}\}} \\ &= (\hat\sigma_{(i)}^2/\hat\sigma^2)^p/(1-p_i) \\ &= \left(\frac{N-p-t_i^2}{N-p-1}\right)^p \Big/ (1-p_i). \end{aligned} \tag{24}$$

A rough calibration point for (24) is $|\mathrm{CVR}_i - 1| > 3p/N$. (Belsley, Kuh, and Welsch (1980) call (24) COVRATIO. We have abbreviated the mnemonic further for simplicity.)

5.4 Cook-Weisberg Statistic

Cook and Weisberg (1980) propose the logarithm of the ratio of the volume of the $(1-\alpha)100\%$ confidence ellipsoids with and without the ith observation as a measure of influence. This measure reduces to

$$\begin{aligned} \mathrm{CW}_i &= \frac{1}{2}\log(1-p_i) \\ &\quad + \frac{p}{2}\log\left\{\frac{(N-p-1)F_{(\alpha;p,N-p)}}{(N-p-t_i^2)F_{(\alpha;p,N-p-1)}}\right\} \\ &= -\frac{1}{2}\log(\mathrm{CVR}_i) \\ &\quad + \frac{p}{2}\log\{F_{(\alpha;p,N-p)}/F_{(\alpha;p,N-p-1)}\} \end{aligned} \tag{25}$$

where $F_{(\alpha;\cdot,\cdot)}$ is the upper α-point of the F-distribution with the appropriate degrees of freedom. Cook and Weisberg (1980) say this about CW_i:

> "If this quantity is large and positive, then deletion of the ith case [observation] will result in a substantial decrease in volume ... [and if it is] large and negative, the case will result in a substantial increase in volume ..."

Apart from a constant (the ratio of F values), (25) is equivalent to (24). Inspection of (25) indicates that CW_i will be large and negative where t_i^2 is small and p_i is large, and large and positive where t_i^2 is large and p_i is small. But, if both t_i^2 and p_i are large (or small), then CW_i and CVR_i tend to be small. These two factors may offset each other and, therefore, reduce the capability of CVR_i and CW_i of detecting influential observations. We have observed from analysis of many data sets, however, that CW_i and CVR_i successfully pick out influential observations. This is perhaps because points with large p_i tend to pull the fitted equation toward them and consequently have small t_i^2.

6. INFLUENCE FUNCTIONS

An alternative class of measures of the influence of the ith observation is based on the idea of the influence function introduced by Hampel (1968, 1974),

$$\mathrm{IF}_i(x_i; y_i; F; T) = \lim_{\varepsilon \to 0} \frac{T[(1-\varepsilon)F + \varepsilon\delta_{x_i,y_i}] - T[F]}{\varepsilon}, \tag{26}$$

where $T(\cdot)$ is a (sufficiently regular) vector-valued statistic based on a random sample from the cdf F and $\delta_{x_i,y_i} = 1$ at (x_i, y_i) and 0 otherwise. IF_i measures the influence on T of adding one observation (x_i, y_i) to a very large sample. For a finite sample, several approximations to (26) are possible; three of the most promising ones are the empirical influence curve, the sample influence curve, and the sensitivity curve.

Let $\hat{F}$ be the empirical distribution function based on the full sample and $\hat{F}_{(i)}$ be the empirical distribution function when the ith observation is omitted. The empirical influence curve (EIC) is found by substituting $\hat{F}_{(i)}$ for F and $\hat{\beta}_{(i)}$ for $T(\hat{F}_{(i)})$ in (26) and obtaining

$$\begin{aligned}\mathrm{EIC}_i &= (N-1)(X_{(i)}^T X_{(i)})^{-1} x_i^T (y_i - x_i\hat{\beta}_{(i)}) \\ &= (N-1)(X^TX)^{-1}x_i^T \frac{e_i}{(1-p_i)^2},\end{aligned} \tag{27}$$

where

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)} \tag{28}$$

is the estimate of β when the ith observation is omitted. The sample influence curve (SIC) is found by omitting the limit in (26) and taking $F = \hat{F}$, $T(\hat{F}) = \hat{\beta}$, $\varepsilon = -1/(N-1)$, and obtaining

$$\begin{aligned}\mathrm{SIC}_i &= (N-1)(X^TX)^{-1}x_i^T(y_i - x_i\hat{\beta}_{(i)}) \\ &= (N-1)(X^TX)^{-1}x_i^T \frac{e_i}{(1-p_i)}.\end{aligned} \tag{29}$$

The sensitivity curve (SC) is obtained by setting $F = \hat{F}_{(i)}$, $T(\hat{F}_{(i)}) = \hat{\beta}_{(i)}$, and $\varepsilon = 1/N$. This yields

$$\mathrm{SC}_i = N(X^TX)^{-1}x_i^T \frac{e_i}{1-p_i}. \tag{30}$$

Clearly SIC_i and SC_i are equivalent. Miller (1974) showed that

$$\hat{\beta} - \hat{\beta}_{(i)} = (X^TX)^{-1}x_i^T \frac{e_i}{1-p_i}. \tag{31}$$

In comparing SIC_i and SC_i to (31), we see that SIC_i and SC_i are easier to interpret; they are proportional to the distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$. However, EIC_i is more sensitive to p_i.

Since IF_i is a vector, it must be normalized so that observations can be ordered in a meaningful way. The class of norms which are location/scale invariant is given by

$$D_i(M; c) = \frac{(\mathrm{IF}_i)^T M (\mathrm{IF}_i)}{c} \tag{32}$$

for any appropriate choice of M and c. A large value of $D_i(M; c)$ indicates that the ith observation has strong influence on the estimated coefficients relative to M and c. We examine $D_i(M; c)$ for four commonly suggested choices of M and c.

6.1 Cook's Distance

If we use the sample influence curve to approximate the influence function and substitute $M = X^TX$ and $c = (N-1)^2 p\hat{\sigma}^2$ in (32), we obtain

$$\begin{aligned}\mathrm{C}_i &= D_i(X^TX; (N-1)^2 p\hat{\sigma}^2) \\ &= \frac{t_i^2}{p}\frac{p_i}{1-p_i}.\end{aligned} \tag{33}$$

This measure is called Cook's distance and was proposed by Cook (1977a). Although C_i should not be used as a test of significance (see Obenchain, 1977), Cook (1977a) suggested that each C_i be compared with the quantiles of the central F distribution with p and $(N-p)$ degrees of freedom.

C_i can also be written as (see Bingham, 1977):

$$\mathrm{C}_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^T(\hat{Y} - \hat{Y}_{(i)})}{p\hat{\sigma}^2}$$

where $\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$ is the vector of predicted values when $Y_{(i)}$ is regressed on $X_{(i)}$. Thus, C_i can be interpreted as the scaled Euclidean distance between the two vectors of fitted values when the fitting is done by including or excluding the ith observation.

6.2 Welsch-Kuh Distance

The impact of the ith observation on the ith predicted value can be measured by scaling the change in prediction at x_i when the ith observation is omitted, that is,

$$\frac{|\hat{y}_i - \hat{y}_{i(i)}|}{\sigma\sqrt{p_i}} = \frac{|x_i(\hat{\beta} - \hat{\beta}_{(i)})|}{\sigma\sqrt{p_i}}. \tag{34}$$

Welsch and Kuh (1977), Welsch and Peters (1978), and Belsley, Kuh, and Welsch (1980) suggest using $\hat{\sigma}^2_{(i)}$ as an estimate of σ^2 and called (34) DFFITS$_i$. For simplicity, we will refer to (34) by WK$_i$. Thus

$$\text{WK}_i = \frac{|x_i(\hat{\beta} - \hat{\beta}_{(i)})|}{\hat{\sigma}_{(i)}\sqrt{p_i}} = |t_i^*|\sqrt{\{p_i/(1-p_i)\}}. \tag{35}$$

If (29) is used to approximate (26), then WK$_i$ = $\sqrt{D_i(X^TX;\ (N-1)\hat{\sigma}^2_{(i)})}$, and if (30) is used to approximate (26), then WK$_i$ = $\sqrt{D_i(X^TX;\ N\hat{\sigma}^2_{(i)})}$. Belsley, Kuh, and Welsch (1980) suggested using $2\sqrt{(p/N)}$ as a calibration point for WK$_i$. Arguing that (34) is a t-like statistic, Velleman and Welsch (1981) recommended that "values greater than 1 or 2 seem reasonable to nominate points for special attention."

6.3. Welsch's Distance

Using (27) to approximate (26) and setting $M = X^T_{(i)}X_{(i)}$ and $c = (N-1)\hat{\sigma}^2_{(i)}$, (32) becomes

$$\begin{aligned} \text{W}_i^2 &= D_i(X^T_{(i)}X_{(i)};\ (N-1)\hat{\sigma}^2_{(i)}) \\ &= (N-1)t_i^{*2}\frac{p_i}{(1-p_i)^2}. \end{aligned} \tag{36}$$

Welsch (1982) suggested using W$_i$ as a diagnostic tool and, for $n > 15$, using $3\sqrt{p}$ as a calibration point for W$_i$. Equations (35) and (36) indicate that

$$\text{W}_i = \text{WK}_i\sqrt{\frac{N-1}{1-p_i}}. \tag{37}$$

Hence, W$_i$ is more sensitive than WK$_i$ to p_i. However, the fact that WK$_i$ is easier to interpret led some authors to prefer WK$_i$ over W$_i$.

6.4 Modified Cook's Distance

A modified version of C$_i$ (cf. (33)) has also been proposed. The measure suggested is

$$\begin{aligned} \text{C}_i^* &= \sqrt{D_i(X^TX;\ \frac{p(N-1)^2}{N-p}\hat{\sigma}^2_{(i)})} \\ &= |t_i^*|\sqrt{\frac{N-p}{p}\frac{p_i}{1-p_i}} \\ &= \text{WK}_i\sqrt{\{(N-p)/p\}} \end{aligned} \tag{38}$$

which, aside from a constant factor, is the same as WK$_i$. C$_i^*$ was originally suggested by Welsch and Kuh (1977) and subsequently by Atkinson (1981a), who contended that this modification:

* gives more emphasis to extreme values,
* makes C$_i^*$ more suitable for graphical displays (a half normal plot was suggested), and
* makes the plots of C$_i^*$ and $|t_i^*|$ identical for the balanced case, where $p_i = p/N$, for all i.

Atkinson (1982), added: "signed values of the C$_i^*$ can be plotted in the same way as residuals, for example, against explanatory variables in the model." C$_i^*$ can also be plotted in serial order.

TABLE 1
Influence measures based on the influence function for a small dataset

Row	Y	X	e_i	p_i	C$_i$	C$_i^*$	W$_i$	WK$_i$
1	2.5	1	−0.01	0.58	0.00	−0.06	−0.14	−0.04
2	3.5	3	0.09	0.24	0.01	0.17	0.30	0.12
3	4	4	0.14	0.18	0.01	0.21	0.36	0.14
4	4.5	5	0.19	0.18	0.03	0.28	0.49	0.20
5	6	8	0.34	0.58	1.04	2.24	5.48	1.58
6	4	6	−0.76	0.24	0.64	∞	∞	∞

The basic difference between C$_i$ and C$_i^*$, W$_i$ and WK$_i$ is in the choice of the scale estimate. An advantage of using $\hat{\sigma}^2$, as an estimate of Var(ε_i), is that comparison of the distances from observation to observation is meaningful because they refer to a fixed metric. For further discussion, see Cook and Weisberg (1982). Using $\hat{\sigma}^2$ as a scale estimate, however, is sometimes noninformative. To illustrate we give here a numerical version of an example given by Dempster and Gasko-Green (1981) and cited by Velleman and Welsch (1981). Here all of the observations but one lie on the line $y = 2 + 0.5x$ (see Table 1). C$_i$ can indicate that some observations on the line (e.g., point 5) are more influential than the one observation not on the line (e.g., point 6), whereas C$_i^*$, W$_i$, and WK$_i$ are infinite for that point.

7. PARTIAL INFLUENCE

The influence measures discussed thus far assume that all regression coefficients are of equal interest. An influence measure which involves all regression coefficients can be noninformative and misleading (see comments by Pregibon in the discussion following Atkinson, 1982). An observation can be an outlier and/or influential only in one dimension or a few dimensions. (For example, observation 17 in the example given in Section 10 is influential only on X_4 and X_3, but not on X_1, X_2, or X_5.) Further, an observation with a moderate influence on all regression coefficients may be judged more influential than one with a large influence on one coefficient and negligible influence on all others.

Information about a single regression coefficient is, therefore, of interest. In this section, we present measures for assessing the influence that an observation has on a single regression coefficient and examine several diagnostic plots which have been suggested for studying this effect.

7.1 Influence of an Observation on a Single Coefficient

A statistic for the impact of the ith observation on a subset of β can be found in Cook and Weisberg (1980). A special case is:

$$D_{ij} = \frac{t_i^2(p_i - p_{i[j]})}{1 - p_i} \tag{39}$$

which measures the influence of the ith observation on the jth coefficient. Using a version of (20), it is straightforward to show that

$$D_{ij} = \frac{t_i^2}{1 - p_i}\frac{w_{ij}^2}{W_j^T W_j} \tag{40}$$

where w_{ij} is the ith element of

$$W_j = (I - P_{[j]})X_j, \tag{41}$$

the vector of residuals when X_j is regressed on $X_{[j]}$.

Belsley, Kuh, and Welsch (1980) suggested using

$$D_{ij}^* = \frac{\hat{\beta}_j - \hat{\beta}_j(i)}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} = \frac{t_i^* c_{ij}}{\sqrt{\{(1 - p_i)C_j^T C_j\}}} \tag{42}$$

where C_j is the jth column of

$$C = X(X^TX)^{-1}, \tag{43}$$

the Moore-Penrose inverse of X, and t_i^* is as defined in (13). Belsley, Kuh, and Welsch (1980) call (42) DFBETAS$_{ij}$. Here, we use D_{ij}^* for simplicity. In the Appendix, we show that

$$D_{ij}^* = \frac{t_i^* w_{ij}}{\sqrt{\{(1 - p_i)W_j^T W_j\}}} \tag{44}$$

which, apart from the difference in scale estimate, is the same as $\sqrt{D_{ij}}$ (cf. (40)). Belsley, Kuh, and Welsch (1980) suggest nominating points with values of $|D_{ij}^*|$ exceeding $2/\sqrt{N}$ for special attention.

7.2 Partial Leverage

Analogous to (20), we write

$$p_i = p_{i[j]} + \delta_{ij}^2, \tag{45}$$

where $\delta_{ij}^2 = w_{ij}^2/W_j^T W_j$ represents the contribution of the jth variable to the leverage of the ith observation, or, in other words, the change in the ith diagonal element of the prediction matrix when X_j is added to (or omitted from) the regression model. The vector $\delta_j^2 = (\delta_{1j}^2, \cdots, \sigma_{Nj}^2)^T$ is the normalized vector of squared residuals obtained from the regression of X_j on all other columns of X.

Because δ_{ij}^2 is the leverage of the ith observation in the added variable plot for X_j (the regression of R_j on W_j), data points with large δ_{ij}^2 can exert an undue influence on the selection (omission) of the jth variable in most automatic variable selection methods (Velleman and Welsch, 1981).

If all observations have equal partial leverage (i.e., δ_{ij}^2), then δ_{ij}^2 equals $1/N$. Therefore, analogous to the choice of the calibration point for p_i (see Hoaglin and Welsch, 1978), a reasonable rule of thumb is that δ_{ij}^2 be regarded as large if it exceeds $2/N$. Also, signed values of δ_{ij} may be plotted versus X_j, or alternatively, δ_{ij} may be plotted in serial order.

7.3 Added Variable Plots

Suppose we wish to fit the model

$$Y = X_{[j]}\beta + X_j\theta_j + \varepsilon \tag{46}$$

where β is now of dimension $(p - 1) \times 1$. By multiplying (46) by $(I - P_{[j]})$ and noting that $(I - P_{[j]})X_{[j]} = 0$, we obtain

$$(I - P_{[j]})Y = (I - P_{[j]})X_j\theta_j + (I - P_{[j]})\varepsilon$$

or

$$R_j = W_j\theta_j + \varepsilon^*, \tag{47}$$

where R_j and W_j, as implicitly defined in (47), are the residuals vectors when Y and X_j are regressed on $X_{[j]}$, respectively. Taking the expectation of (47), we obtain $E(R_j) = W_j\theta_j$, which suggests a plot of

$$R_j \quad \text{versus} \quad W_j \tag{48}$$

This plot was introduced by Mosteller and Tukey (1977) and has several attractive features. It should appear as a straight line through the origin with slope $\hat{\theta}_j$. In fact, the residuals from the multiple regression model (46) and the residuals from the simple regression model (47) are identical. The scatter of the points will visually indicate which of the data points are most influential in determining the magnitude of $\hat{\theta}_j$. Belsley, Kuh, and Welsch (1980) have called this plot a partial regression leverage plot, but we prefer the name added variable plot suggested by Cook and Weisberg (1982). For properties and details, see, e.g., Belsley, Kuh, and Welsch (1980) and Cook and Weisberg (1982).

7.4 Partial Residuals Plots

Using the well known identity (see Bingham, 1977)

$$\hat{\theta}_j = \frac{X_j^T(I - P_{[j]})Y}{X_j^T(I - P_{[j]})X_j} = \frac{W_j^T Y}{W_j^T W_j} \tag{49}$$

and a version of (18), we can write R_j as

$$R_j = \left\{I - P + \frac{(I - P_{[j]})X_j X_j^T(I - P_{[j]})}{X_j^T(I - P_{[j]})X_j}\right\}Y \tag{50}$$

$$= e + (I - P_{[j]})X_j\hat{\theta}_j. \tag{51}$$

The added variable plot in (48) can, then, be written as

$$(52) \quad e + (I - P_{[j]})X_j\hat{\theta}_j \quad \text{versus} \quad (I - P_{[j]})X_j.$$

A special case of (52) is obtained by replacing $P_{[j]}$ by 0, yielding the plot of

$$(53) \quad e + X_j\hat{\theta}_j \quad \text{versus} \quad X_j.$$

This plot, which was introduced by Ezekiel (1924) and rediscovered by Larsen and McCleary (1972), has been discussed by many others (e.g., Wood (1973), Daniel and Wood (1980), Henderson and Velleman (1981), and Atkinson (1982) and his discussants). Larsen and McCleary (1972) called (53) a partial residual plot and Daniel and Wood (1980) called it a component plus residual plot. Since the horizontal scale on the partial residual plots is X_j, the plot often (but not always) indicates nonlinearity, thereby suggesting the need for transformation if necessary. It is not easy, however, to determine which of the data points have a major role in determining $\hat{\theta}_j$. The partial residuals plots seem to be better for the analysis of transformation, while added variable plots help more in high leverage and influential data. Of course, the two plots are identical when the columns of X are orthogonal. The main contribution of these plots is that they tell us about the influence that an observation will exercise on the fit if a particular variable which is currently not in the equation is brought into the equation.

7.5 Augmented Partial-Residual Plots

As has been mentioned above, the partial residual plots can fail to indicate the need for transformation. Mallows (1985) has proposed a modification of the partial residual plot by augmenting the linear component with a nonlinear component. This modification appears to be more sensitive to nonlinearity. Mallows calls this an augmented partial residual plot. One can calculate the augmented partial residual plot by first fitting the model

$$Y = X_{[j]}\beta + X_j\theta_j + Z_j\Upsilon_j + \varepsilon$$

where $Z_j = f(X_j)$ is some nonlinear function of X_j, and then plotting

$$e = X_j\hat{\theta}_j + f(X_j)\hat{\Upsilon}_j \quad \text{versus} \quad X_j.$$

Mallows recommends taking the nonlinear component as quadratic. Some early results of the augmented partial residual plot have been encouraging but more work needs to be done before we can conclude that this plot is superior to the partial residual plot for detecting nonlinearity in the regressors.

8. JOINT INFLUENCE OF MULTIPLE OBSERVATIONS

The methods that we have described can be used to detect individual observations which are influential.

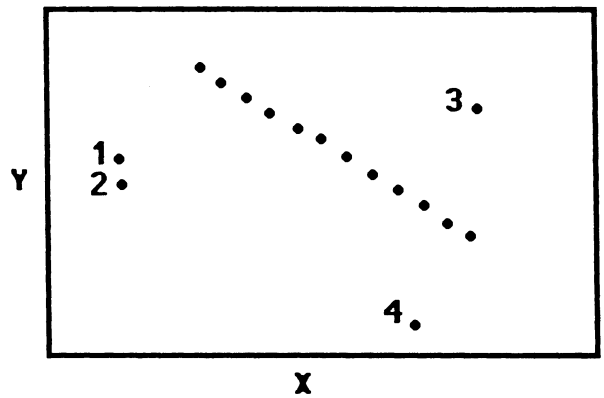

FIG. 3. *An illustration of joint influence. Points* 1 *and* 2 *are jointly (but not individually) influential, whereas points* 3 *and* 4 *are individually (but not jointly) influential.*

There are situations however when an observation is not influential singly but taken in a group with other observations may be highly influential. An illustration is given in Figure 3. Points 1 and 2 singly are not influential, but jointly they have a large effect on the fit. This situation is sometimes called the "masking" effect, because the influence of one obsrvation is masked by the presence of another neighboring observation. On the other hand, points 3 and 4 behave differently. Individually they are influential, but jointly (i.e., when both omitted) they are not. The methods for detecting the influence of a single observation can be generalized for detecting subsets of observations which are jointly influential. Not much progress has been made in that direction, perhaps due to the computational burden associated with the multiple observations procedures. The number of subsets involved here is very large. Also, in addition to the residuals and the ith diagonal elements of the prediction matrix, the multiple observation procedures generally require the computation of the off diagonal elements of the prediction matrix.

A procedure, which is independent of any particular measure of influence, employing cluster analysis to detect subsets of influential observations has been proposed by Gray and Ling (1984). A modification to the Gray and Ling procedures has been suggested by Hadi (1985). More efficient computational procedures for detecting subsets of influential observations are still needed however.

9. SUMMARY OF VARIOUS INFLUENCE MEASURES

A summary of the influence measures that we have discussed together with their calibration points is shown in Table 2. As we noted before, each influence measure is designed to detect a specific phenomenon in the data. They are all closely related, as they are functions of the basic building blocks in model construction (e.g., the residuals e, the residual mean

TABLE 2
Summary of influence measures

Measures based on	Formula	Calibration point	Equation
Residuals	$t_i = e_i/\hat{\sigma}\sqrt{1 - p_i}$	$\simeq N(0, 1)$	(12)
	$t_i^* = t_i \sqrt{\dfrac{N - p - 1}{N - p - t_i^2}}$	$\simeq t(N - p - 1)$	(15)
Prediction matrix	$p_i = x_i(X^TX)^{-1}x_i^T$	$2p/N$	(16)
Volume of confidence ellipsoids	$p_i^* = 1 - \mathrm{AP}_i = p_i + e_i^2/e^Te$	$2(p + 1)/N$	(20)
	$\mathrm{LD}_i = N \log\left\{\left(\dfrac{N}{N-1}\right)\dfrac{N - p - 1}{t_i^* + N - p - 1}\right\} + \dfrac{t_i^*(N-1)}{(1 - p_i)(N - p - 1)} - 1$	χ_p^2	(23)
	$\mathrm{CVR}_i = \left(\dfrac{N - p - t_i^2}{N - p - 1}\right)^p \Big/ (1 - p_i)$	$\lvert \mathrm{CVR}_i - 1 \rvert > 3p/N$	(24)
	$\mathrm{CW}_i = \text{const.} - \tfrac{1}{2}\log(\mathrm{CVR}_i)$		(25)
Influence function	$\mathrm{C}_i = p_i t_i^2/p(1 - p_i)$	$F(p, N - p)$	(33)
	$\mathrm{WK}_i = \lvert t_i^* \rvert \sqrt{\dfrac{p_i}{1 - p_i}}$	$2\sqrt{(p/N)}$	(35)
	$\mathrm{W}_i = \mathrm{WK}_i \sqrt{\dfrac{N-1}{1 - p_i}}$	$3\sqrt{p}$	(37)
	$\mathrm{C}_i^* = \mathrm{WK}_i\sqrt{\{(N - p)/p\}}$	$2\sqrt{\{(N - p)/N\}}$	(38)
Partial influence	$D_{ij} = \dfrac{t_i^2 w_{ij}^2}{\mathrm{W}_j^T\mathrm{W}_j(1 - p_i)}$		(40)
	$D_{ij}^* = \dfrac{t_i^* w_{ij}}{\sqrt{\{\mathrm{W}_j^T\mathrm{W}_j(1 - p_i)\}}}$	$2/\sqrt{N}$	(44)
	$\delta_{ij}^2 = \dfrac{w_{ij}^2}{\mathrm{W}_j^T\mathrm{W}_j}$	$2/N$	(45)

square $\hat{\sigma}^2$, the ith element of the prediction matrix p_i). In any particular application, the analyst does not have to look at all of these measures since there is a great deal of redundancy in them. Their relative merits and importance have not been established. From our experience with several data sets, examining WK_i, CW_i, and D_{ij} or, alternatively, C_i^*, CVR_i, and D_{ij}^* seem sufficient for detecting influential observations. The three quantities in each set measure different aspects of influence and give a comprehensive picture.

10. EXAMPLE

10.1 Data Description and Global Analysis

As an illustrative example, we use the result of a laboratory experiment performed by Moore (1975). This example has also been used by Weisberg (1981) to illustrate the contribution of the individual observations to the Mallows C_p statistic. The data were collected on a single sample, kept in suspension in water for 220 days. The data as presented by Weisberg (1981) are reproduced in Table 3. The measured variables are: Y_1 = log(oxygen demand in dairy wastes), mg/min; X_1 = biological oxygen demand, mg/liter; X_2 = total Kjeldahl nitrogen, mg/liter; X_3 = total solids, mg/liter; X_4 = total volative solids (a component of X_3), mg/liter; and X_5 = chemical oxygen demand, mg/liter.

A linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \tag{54}$$

is fitted to the data and the results of the fit are shown

TABLE 3
Moore's data (1975)

Row	X_1	X_2	X_3	X_4	X_5	Y
1	1125	232	7160	85.9	8905	1.5563
2	920	268	8804	86.5	7388	0.8976
3	835	271	8108	85.2	5348	0.7482
4	1000	237	6370	83.8	8056	0.7160
5	1150	192	6441	82.1	6960	0.3130
6	990	202	5154	79.2	5690	0.3617
7	840	184	5896	81.2	6932	0.1139
8	650	200	5336	80.6	5400	0.1139
9	640	180	5041	78.4	3177	−0.2218
10	583	165	5012	79.3	4461	−0.1549
11	570	151	4825	78.7	3901	0.0000
12	570	171	4391	78.0	5002	0.0000
13	510	243	4320	72.3	4665	−0.0969
14	555	147	3709	74.9	4642	−0.2218
15	460	286	3969	74.4	4840	−0.3979
16	275	198	3558	72.5	4479	−0.1549
17	510	196	4361	57.7	4200	−0.2218
18	165	210	3301	71.8	3410	−0.3979
19	244	327	2964	72.5	3360	−0.5229
20	79	334	2777	71.9	2599	−0.0458

TABLE 4
Moore's data: regression summary

Variable	$\hat{\beta}_j$	S.E.$(\hat{\beta}_j)$	t_j	R_j^2	$W_j^T W_j$	$\hat{\beta}_j^2 W_j^T W_j$
Const.	−2.1561					
X_1	−0.0000	0.0005	−0.017	0.86	2.551E5	2.072E-5
X_2	0.0013	0.0013	1.041	0.23	4.293E4	7.434E-2
X_3	0.0001	0.0001	1.662	0.78	1.159E7	1.893E-1
X_4	0.0079	0.0140	0.564	0.59	3.497E2	2.182E-2
X_5	0.0001	0.0001	1.921	0.77	1.260E7	2.529E-1

SSE = 0.9595 $R^2 = 0.81$ $\hat{\sigma}^2 = 0.0685$
SST = 5.0679 $F = 11.99$ $N = 20$
Durbin-Watson = 2.13

in Table 4. Entries in column 5 of Table 4 are the multiple correlation coefficients squared when X_j is regressed on $X_{[J]}$. Entries in column 6 are the corresponding sum of squares of residuals. Examination of the regression results leads us to the following conclusions:

1. The plot of t_i versus $\hat{y}_i$ (not shown) does not indicate systematic failure of model (54). Observation number 1, however, has the largest residual, $t_1 = 2.64$.

2. The fit is significant with ($p < 0.01$) as indicated by the F value = 11.99 and $R^2 = 0.81$,

3. None of the t values is significant. This is, perhaps, due to the high correlation among the explanatory variables which can be seen from column 5 of Table 4 where R_j^2 is large for all $j \neq 2$. This may suggest that a linear model based on a subset of the

TABLE 6
Moore's data with observations arranged within each measure in decreasing order of influence and clusters indicated in parentheses

Measures based on	Influence measures	Influential observations	Reference equation
Residuals	t_i	(1, 20)	(12)
	t_i^*	(1, 20)	(15)
Prediction matrix	p_i	17	(16)
Volume of confidence ellipsoids	p_i^*	17	(20)
	LD_i	(17, 1), 20	(23)
	CVR_i	1, 17, (20, 15, 7)	(24)
	CW_i	1, 17, (20, 15, 7)	(25)
Influence function	C_i	17, (1, 20)	(33)
	WK_i	(17, 1, 20)	(35)
	W_i	17, (1, 20)	(37)
	C_i^*	(17, 1, 20)	(38)

TABLE 5
Influence measures for Moore's data

Row	t_i	t_i^*	p_i	p_i^*	LD_i	CVR_i	CW_i	WK_i	W_i	C_i^*	C_i
1	2.64*	3.58*	0.34	0.67	14.60*	0.04*	1.57*	2.55*	13.68*	3.90*	0.59*
2	−0.79	−0.78	0.50	0.52	0.89	2.39	−0.50	−0.78	−4.81	−1.19	0.10
3	0.47	0.46	0.49	0.49	0.30	2.75	−0.57	0.44	2.70	0.68	0.03
4	−0.21	−0.20	0.25	0.25	0.04	2.04	−0.42	−0.12	−0.59	−0.18	0.00
5	−1.04	−1.04	0.28	0.34	0.64	1.34	−0.21	−0.66	−3.39	−1.00	0.07
6	0.82	0.81	0.37	0.40	0.57	1.84	−0.37	0.62	3.43	0.95	0.07
7	−1.42	−1.47	0.15	0.27	0.69	0.73	0.10	−0.63	−2.97	−0.96	0.06
8	−0.28	−0.27	0.09	0.09	0.03	1.65	−0.31	−0.08	−0.38	−0.13	0.00
9	−0.05	−0.05	0.36	0.36	0.03	2.45	−0.51	−0.04	−0.20	−0.05	0.00
10	−0.46	−0.44	0.16	0.17	0.07	1.69	−0.33	−0.19	−0.91	−0.29	0.01
11	0.74	0.73	0.22	0.26	0.23	1.58	−0.29	0.39	1.95	0.60	0.03
12	0.21	0.20	0.14	0.14	0.03	1.77	−0.35	0.08	0.37	0.12	0.00
13	−0.16	−0.15	0.09	0.10	0.03	1.70	−0.33	−0.05	−0.23	−0.08	0.00
14	0.10	0.09	0.20	0.20	0.03	1.94	−0.39	0.05	0.23	0.07	0.00
15	−1.66	−1.78	0.17	0.33	1.25	0.51	0.27	−0.81	−3.87	−1.24	0.09
16	0.36	0.35	0.26	0.27	0.08	2.00	−0.41	0.21	1.06	0.32	0.01
17	0.97	0.97	0.92*	0.92*	15.55*	12.51*	−1.33*	3.26*	49.72*	4.98*	1.78*
18	0.05	0.05	0.23	0.23	0.03	2.03	−0.42	0.03	0.13	0.04	0.00
19	−1.06	−1.07	0.36	0.42	0.97	1.48	−0.26	−0.81	−4.43	−1.24	0.11
20	1.89	2.11*	0.41	0.56	5.07*	0.45	0.33	1.74*	9.86*	2.66*	0.41*

explanatory variables may do as well. Weisberg (1981) concluded that for the purpose of variable selection, no reason is apparent for rejecting the assumption of unbiasedness of model (54) for the region covered by the observed data.

4. Even though all observations were taken on the same sample over time the model has survived the Durbin-Watson test with ($p < 0.01$).

10.2 Influence Analysis

We now examine the several influence measures which we have described earlier. These are shown in Table 5. For economy of space, the plot for only three of these measures in serial order is given in Figures 4–6. A summary of these plots is given in Table 6 where observations that appear to be most influential

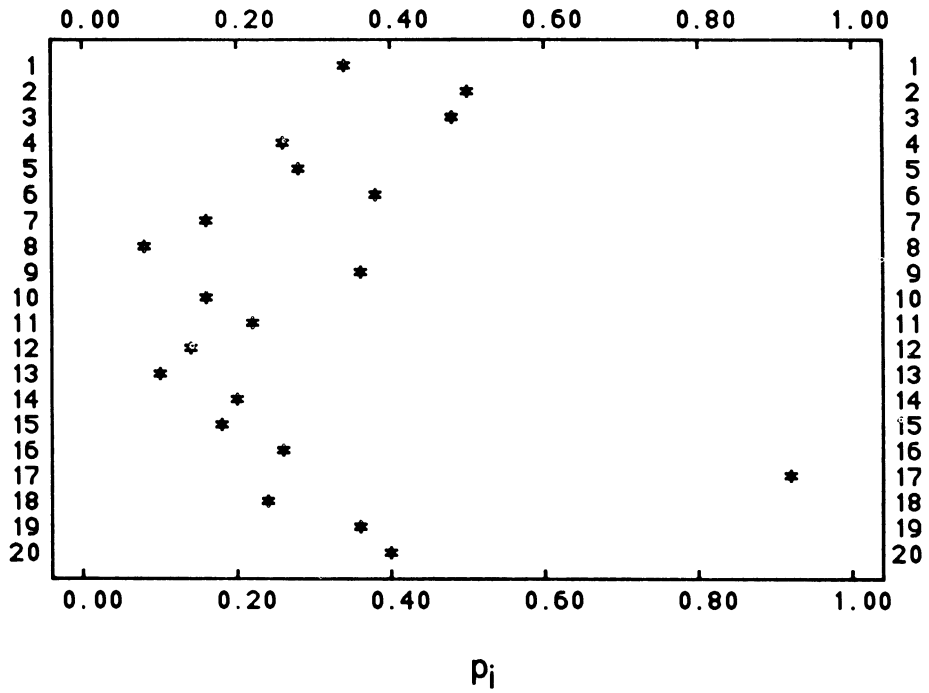

FIG. 4. *Moore's data: plot of p_i in serial order.*

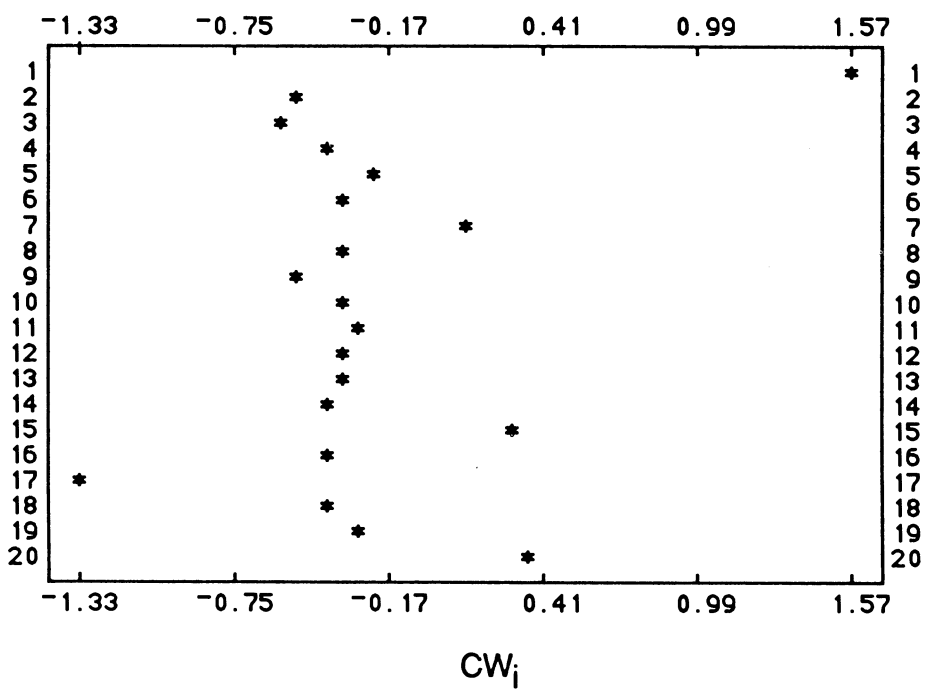

FIG. 5. *Moore's data: plot of CW_i in serial order.*

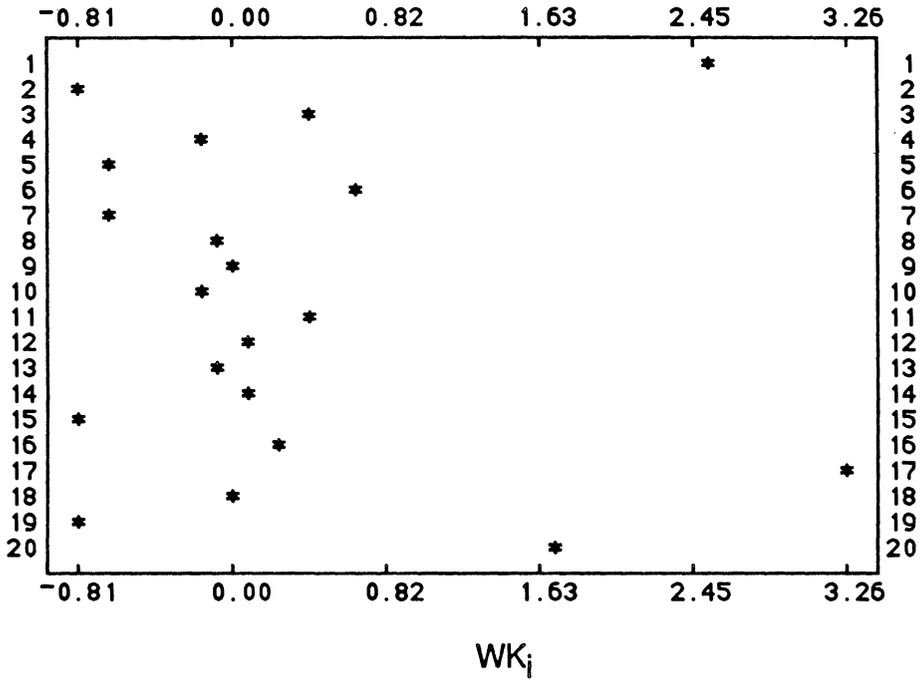

FIG. 6. *Moore's data: plot of* WK_i *in serial order.*

TABLE 7
The D_{ij} *in* (40) *from Moore's data*

No.	X_1	X_2	X_3	X_4	X_5
1	0.010	0.052	0.128	0.008	1.277*
2	0.134	0.001	0.354	0.004	0.018
3	0.000	0.012	0.090	0.000	0.049
4	0.000	0.001	0.002	0.000	0.005
5	0.251*	0.016	0.059	0.002	0.037
6	0.325*	0.044	0.136	0.000	0.094
7	0.043	0.075	0.001	0.003	0.148
8	0.001	0.001	0.000	0.002	0.000
9	0.000	0.000	0.000	0.000	0.001
10	0.002	0.017	0.001	0.006	0.002
11	0.000	0.058	0.003	0.018	0.028
12	0.001	0.003	0.000	0.001	0.000
13	0.000	0.000	0.000	0.001	0.000
14	0.000	0.001	0.000	0.000	0.000
15	0.008	0.250	0.098	0.000	0.017
16	0.026	0.012	0.000	0.000	0.019
17	0.077	0.042	1.541*	9.753*	0.003
18	0.000	0.000	0.000	0.000	0.000
19	0.036	0.377	0.112	0.011	0.018
20	0.000	1.003*	0.100	0.083	0.070

TABLE 8
Moore's data showing the D_{ij}^* *in* (44)

No.	X_1	X_2	X_3	X_4	X_5
1	−0.134	0.310	−0.486	0.125	1.535*
2	0.360	−0.030	−0.586	0.064	−0.133
3	−0.005	0.105	0.291	0.013	−0.215
4	−0.004	−0.025	0.040	−0.012	−0.066
5	−0.502*	−0.129	0.244	0.047	0.194
6	0.563*	0.206	−0.364	0.015	−0.304
7	0.215	0.285	0.026	−0.057	−0.400
8	0.031	0.035	−0.005	−0.040	−0.014
9	−0.017	0.003	−0.002	−0.007	0.031
10	0.047	0.126	−0.033	−0.073	0.038
11	−0.018	−0.237	0.053	0.133	−0.164
12	−0.023	−0.049	−0.018	0.030	0.021
13	−0.006	−0.017	0.003	0.025	−0.003
14	−0.001	−0.028	−0.018	0.007	0.007
15	−0.093	−0.538	0.336	0.007	−0.141
16	−0.156	−0.106	0.021	0.001	0.133
17	0.278	−0.205	1.239*	−3.117*	0.057
18	−0.018	−0.011	0.005	0.002	0.009
19	−0.191	−0.617	0.336	−0.107	0.135
20	0.000	1.118*	−0.352	0.322	−0.295

are arranged within each measure in descending order of influence. The analyses based on the influence measures lead to the following conclusions:

1. As had been expected (see (38)), the plots for WK_i and C_i^* are identical.

2. The measures based on the influence functions (WK_i, W_i, C_i^*, and C_i) pinpoint observations number 17, 1, and 20 as different from the others. Because W_i puts more emphasis on p_i, the influence of observations 1 and 20 is not clear in the plot of W_i; this is due to the relatively small values of p_1 and p_{20}. The influence of these two observations is clearer in the plot of WK_i (and C_i^*) compared to that of C_i.

3. While measures based on the influence function appear to be in agreement, those based on the volume of confidence ellipsoids do not. Observation number 17 is declared to be the most influential one by p_i^* and LD_i, while CVR_i and CW_i declare observation 1 to be

the most influential one. On the other hand, CW_i and CVR_i pinpoint five points (in three clusters) as different from all others. By declaring too many observations to be influential, one might think that CW_i and CVR_i are conservative measures. On the contrary, each of these observations is influential on at least one dimension (variable). This can be seen upon inspection of D_{ij} or D_{ij}^*; see (40) and (42), respectively.

We first examine the effects of deleting the ith observation on the jth coefficient as measured by D_{ij}, the results are shown in Table 7. We see immediately that no observation is uniformly most influential on all coefficients. For example, the most influential observation on $\hat{\beta}_1$ is observation number 6, on $\hat{\beta}_2$ is observation number 20, on $\hat{\beta}_3$ and $\hat{\beta}_4$ is observation number 17 (which was declared to be the most influential by all measures except t_i, t_i^*, and CW_i), and on $\hat{\beta}_5$ is observation number 1.

The effects of deleting the ith observation on the jth coefficient as measured by D_{ij}^* are shown in Table 8. Inspection of D_{ij}^* leads to the same conclusions as those obtained by examining D_{ij}. Therefore, a measure which involves all coefficients may be noninformative. We document this point further by looking at, for example, p_i and C_i when the jth variable is deleted. From Table 9, we see that observation number 17 is most influential only when X_4 is included in the model. This indicates that observation number 17 has a large influence basically in one dimension.

We have seen that observations number 1, 7, 15, 17, and 20 are influential either individually or in groups. The impact of deleting these observations on the t values and on the best subset based on the minimum RMS criterion (see Seber, 1977) is shown in Table 10. Examination of Table 10 indicates that:

1. Variables X_3 and X_5 are included in the best subset in all cases except when observation number 17 is omitted; this causes X_3 to be replaced by X_4. Therefore, X_3 and X_5 are the most influential variables,

2. The model with minimum RMS is X_3 and X_5 with observation number 1, 7, and 20 deleted. Note the change in R^2, F values, and the minimum RMS. (Notice that in comparing the F values, one should keep in mind that the degrees of freedom are different in each case.)

TABLE 9
Moore's data showing influential observations according to p_i and C_i when variable X_j is omitted

Variable deleted	Influential observations	
	p_i	C_i
None	17	17
X_1	17	17
X_2	17	17
X_3	17	17
X_4	2, 3	1, 20
X_5	17	17

11. CONCLUSION

We have discussed and reviewed the various measures which have been presented for studying outliers, high leverage points, and influential observations in the context of linear regression. The existence of the interrelationship between these measures enables us to reduce the vast number of measures to a few well chosen ones. The measures suggested concentrate on different aspects of the problem. Some of the quantities proposed concentrate on the lack of fit, others on leverage in the space of explanatory variables, and still others on the interaction between the two. We showed that three measures are sufficient to display the major characteristics of a data set with reference to its leverage, influence, and lack of fit.

TABLE 10
Moore's data showing effects of omitting selected observations on various regression outputs

Observation deleted	t_1[a]	t_2	t_3	t_4	t_5	Best subset[b]			
						$R^2 \times 100$	F	RMS × 1000	Variable included
None	−0.02	1.04	1.66	0.56	1.92	81	22	61	2, 3, 5
1	0.11	1.10	2.72	0.64	0.99	83	24	33	2, 3, 5
7	−0.23	0.79	1.70	0.64	2.32	83	25	57	2, 3, 5
15	0.07	1.59	1.42	0.60	2.20	83	25	53	2, 3, 5
17	−0.28	1.22	0.26	1.10	1.86	82	22	60	2, 4, 5
20	−0.02	0.04	2.18	0.30	2.41	86	48	45	3, 5
1, 20	0.12	−0.06	3.67	0.37	1.63	89	63	20	3, 5
7, 20	−0.32	−0.41	2.43	0.38	3.15	89	62	36	3, 5
1, 7, 20	−0.17	−0.49	4.01	0.46	2.38	92	79	17	3, 5
7, 15, 20	−0.25	0.17	2.11	0.46	−3.33	90	65	33	3, 5

[a] t_j is the t statistic for testing the significance of the jth variable.
[b] Based on the minimum RMS criterion.

APPENDIX: PROOF OF (44)

For a proof of (44), we use the triangular decomposition of positive definite (pd) matrices, i.e., if S is a pd matrix, then there exists a unique unit lower triangle matrix L and a unique diagonal matrix D with positive diagonal elements, such that:

$$(55)\quad \begin{cases} LSL^T = D \\ \text{or equivalently} \\ S = L^{-1}DL^{-T} \\ \text{or equivalently} \\ S^{-1} = L^TD^{-1}L. \end{cases}$$

See, e.g., Stewart (1973) or Maindonald (1984), for proof. Substituting the triangular decomposition (55) of

$$(X^TX) = \begin{bmatrix} X_{[j]}^TX_{[j]} & X_{[j]}^TX_j \\ X_j^TX_{[j]} & X_j^TX_j \end{bmatrix}$$

into a partitioned form of (31), where L and D are partitioned conformably into

$$(56)\qquad L = \begin{bmatrix} \Lambda & 0 \\ \lambda^T & 1 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} \Delta & 0 \\ 0^T & \delta \end{bmatrix},$$

we obtain

$$\begin{bmatrix} \hat\beta \\ \hat\theta_j \end{bmatrix} - \begin{bmatrix} \hat\beta_{(i)} \\ \hat\theta_{j(i)} \end{bmatrix} = \frac{e_i}{1-p_i}\begin{bmatrix} \Lambda^T & \lambda \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \Delta & 0 \\ 0 & \delta \end{bmatrix}^{-1}\begin{bmatrix} \lambda & 0 \\ \Lambda^T & 1 \end{bmatrix}\begin{bmatrix} x_{i[j]}^T \\ x_{ij} \end{bmatrix}$$

$$= \frac{e_i}{1-p_i}\begin{bmatrix} w_{ij}\delta^{-1} & \lambda + \lambda^T\Delta^{-1} & \Lambda x_{i[j]}^T \\ & w_{ij}\delta^{-1} & \end{bmatrix}.$$

It follows that

$$\hat\theta_j - \hat\theta_{j(i)} = \frac{e_i}{1-p_i}\frac{w_{ij}}{\delta} = \frac{e_i}{1-p_i}\frac{w_{ij}}{W_j^TW_j}$$

because the jth diagonal element of D is the residual sum of squares obtained when X_j is regressed on the preceding variables. After scaling, the result follows.

ACKNOWLEDGMENTS

We would like to thank the Executive Editor and anonymous reviewers for their helpful comments on an earlier version of this article.

REFERENCES

ANDREWS, D. F. and PREGIBON, D. (1978). Finding the outliers that matter. *J. Roy. Statist. Soc. Ser. B* **40** 85–93.

ATKINSON, A. C. (1981a). Two graphical displays for outlying and influential observations in regression. *Biometrika* **68** 13–20.

ATKINSON, A. C. (1982). Regression diagnostics, transformations, and constructed variables (with discussion). *J. Roy. Statist. Soc. Ser. B* **44** 1–36.

BECKMAN, R. J. and TRUSSELL, H. J. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *J. Amer. Statist. Assoc.* **69** 199–201.

BELSLEY, D. A., KUH, E. and WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

BINGHAM, C. (1977). Some identities useful in the analysis of residuals from linear regression. Technical Report 300, School of Statistics, Univ. Minnesota.

CHATTERJEE, S. and PRICE, B. (1977). *Regression Analysis by Example*. Wiley, New York.

COLEMAN, D. E. (1977). Finding leverage groups. Working Paper 195, National Bureau of Economic Research, Inc.

COOK, R. D. (1977a). Detection of influential observations in linear regression. *Technometrics* **19** 15–18.

COOK, R. D. and WEISBERG, S. (1980). Characterization of an empirical influence function for detecting influential cases in regression. *Technometrics* **22** 495–508.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

DANIEL, C. and WOOD, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2nd ed. Wiley, New York.

DEMPSTER, A. P. and GASKO-GREEN, M. (1981). New tools for residual analysis. *Ann. Statist.* **9** 945–959.

DRAPER, N. R. and SMITH, H. (1981). *Applied Regression Analysis*, 2nd ed. Wiley, New York.

ELLENBERG, J. H. (1973). The joint distribution of the studentized least squares residuals from a general linear regression. *J. Amer. Statist. Assoc.* **68** 941–943.

EZEKIEL, M. (1924). A method for handling curvilinear correlation for any number of variables. *J. Amer. Statist. Assoc.* **19** 431–453.

GRAY, J. B. and LING, R. F. (1984). K-clustering as a detection tool for influential subsets in regression (with discussion). *Technometrics* **26** 305–330.

HADI, A. S. (1985). K-clustering and the detection of influential subsets (letter to the editor with response). *Technometrics* **27** 323–325.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, Univ. California, Berkeley.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.

HENDERSON, H. V. and VELLEMAN, P. F. (1981). Building multiple regression models interactively. *Biometrics* **37** 391–411.

HOAGLIN, D. C. and WELSCH, R. E. (1978). The hat matrix in regression and ANOVA. *Amer. Statist.* **32** 17–22.

LARSEN, W. A. and MCCLEARY, S. A. (1972). The use of partial residual plots in regression analysis. *Technometrics* **14** 781–790.

MAINDONALD, J. H. (1984). *Statistical Computation*. Wiley, New York.

MALLOWS, C. L. (1985). Augmented partial residuals. Unpublished manuscript.

MICKEY, M. R., DUNN, O. J. and CLARK, V. (1967). Note on the use of stepwise regression in detecting outliers. *Comput. Biomed. Res.* **1** 105–109.

MILLER, R. G. (1974). An unbalanced jackknife. *Ann. Statist.* **2** 880–891.

MOORE, J. (1975). Total biochemical oxygen demand of dairy manures. Ph.D. thesis, Univ. Minnesota, Dept. Agricultural Engineering.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, Mass.

OBENCHAIN, R. L. (1977). Letter to the editor. *Technometrics* **19** 348–351.

SEBER, G. A. F. (1977). *Linear Regression Analysis*. Wiley, New York.

STEWART, G. W. (1973). *Introduction to Matrix Computations*. Academic, New York.

VELLEMAN, P. F. and WELSCH, R. E. (1981). Efficient computing of regression diagnostics. *Amer. Statist.* **35** 234–242.

WEISBERG, S. (1981). A statistic for allocating C_p to individual cases. *Technometrics* **23** 27–31.

WELSCH, R. E. (1982). Influence functions and regression diagnostics. In *Modern Data Analysis* (R. L. Launer and A. F. Siegel, eds.). Academic, New York.

WELSCH, R. E. and KUH, E. (1977). Linear regression diagnostics. Technical Report 923-77, Sloan School of Management, Massachusetts Institute of Technology.

WELSCH, R. E. and PETERS, S. C. (1978). Finding influential subsets of data in regression models. *Proc. Eleventh Interface Symp. Comput. Sci. Statist.* 240–244.

WOOD, F. S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics* **15** 677–695.

Comment

R. Dennis Cook

Chatterjee and Hadi present a disturbing account of the disorientation that can result from attempting to sort through the variety of methods that are available for studying influence, leverage, and outliers in linear regression. Their admonition that the goals of an analysis must be used to guide our choice of methodology is entirely appropriate. The question "Influence on what?" is indeed important, particularly when it is asked of a specific method. I find that answers to this question can form a useful guidebook to influence methodology and can thereby remove much of the perceived confusion. With this key question in mind, Chatterjee and Hadi describe several useful distinctions between the various methods, but some confusion evidently remains, as exemplified by the all-but-one-point-on-a-line problem. For further clarity, it is necessary to take a closer look at the appropriate uses of various influence diagnostics. Beginning with a general introduction, the following discussion is intended to emphasize critical distinctions between selected methods and to further illustrate the importance of Chatterjee and Hadi's question. Unless indicated otherwise, notation is the same as that used by Chatterjee and Hadi.

1. INTRODUCTION

Statistical models are extremely useful devices for extracting and understanding the essential features of a set of data. Models, however, are nearly always approximate descriptions of more complicated processes and therefore are nearly always wrong. Because of this inexactness, considerations of model adequacy are extremely important. The recent paper by Freedman and Navidi (1986) in combination with the discussants' remarks provides a forceful lesson on modeling. Depending on the situation, a universally compelling demonstration of the adequacy of a model may not be possible. But what we can always do is strive for the reassurance that what we have done is sensible in light of the available information, that the data do not contradict the model or vice versa, and that reasonable alternative formulations will not lead to drastically different conclusions. How much reassurance we may need depends on the particular problem. In well studied situations where we have considerable prior information and experience, a little reassurance may be sufficient, while in fresh problems we may require much more. But some reassurance is always necessary.

Many methods are available for gaining necessary reassurance. For example, we may empirically validate a model through continued observation of the process under study or use robust methods to mitigate the impact of questionable aspects of the model. In addition, diagnostic methods should be used to look for contradictory or other relevant information in the observed data. The absence of such information will not prove that the model is accurate, but it can provide the reassurance that the model is not contradicted by available information or unduly influenced by isolated characteristics of the data.

Chatterjee and Hadi describe their experiences with a particular class of diagnostic methods that are intended to aid in assessing the role that individual observations play in determining a fitted model. A fitted model can be viewed as a smoothed representation that captures global and essential features of the data, but this view is not always appropriate. Key features of a fitted model can be dominated by a single observation and conclusions in such situations tend to depend critically on the model. It seems generally recognized that a concern for influential observations should be part of any analysis, and in recent years there has been a proliferation of methods for their detection.

2. t_i AND t_i^*

Chatterjee and Hadi discuss several reasons for preferring t_i^* over t_i, but their discussion seems to lack

R. Dennis Cook is Professor and Chair, Department of Applied Statistics, University of Minnesota, St. Paul, Minnesota 55108.