Influential Observations in Linear Regression

R. DENNIS COOK*

Characteristics of observations which cause them to be influential in a least squares analysis are investigated and related to residual variances, residual correlations, and the convex hull of the observed values of the independent variables. It is shown how deleting an observation can substantially alter an analysis by changing the partial F-tests, the studentized residuals, the residual variances, the convex hull of the independent variables, and the estimated parameter vector. Outliers are discussed briefly, and an example is presented.

KEY WORDS: Deleting observations; Outliers; Partial F-tests; Residual correlations; Studentized residuals.

1. INTRODUCTION

In the least squares analysis of data based on a full-rank linear regression model, an observation may be judged influential if important features of the analysis are altered substantially when the observation is deleted. Past investigations have indicated that the influence of an observation is partially manifested through the associated residual and residual variance.

Consider the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \ , \tag{1.1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times p$ full-rank matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\mathbf{e}$ is a vector of independent random variables each with zero mean and variance σ^2. Let $\mathbf{V} = (v_{ij}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and recall that the covariance matrices of the residuals, $\mathbf{R} = (r_i)$, and predicted values, $\hat{\mathbf{Y}} = (\hat{y}_i)$, from a least squares analysis are given by $(\mathbf{I} - \mathbf{V})\sigma^2$ and $\mathbf{V}\sigma^2$, respectively.

Behnken and Draper (1972) note that wide variation in the variance of the residuals reflects nonhomogeneous spacing of the design points (i.e., the rows of $\mathbf{X}$). Huber (1975) mentions that large values of v_{ii} typically correspond to "outlying" design points. Hoaglin and Welsch (1978) describe how the elements of $\mathbf{V}$ may be used to detect high-leverage design points.

Others have linked the v_{ii} to our ability to detect outliers. Huber (1975) suggests that it will be difficult to spot outlying observations if $\max v_{ii}$ is not considerably smaller than 1. Davies and Hutton (1975) state that if $v_{ii} > 0.2$, then "it is possible for a moderate error in the corresponding observation to affect the estimates significantly and yet go undetected when the residuals are checked." Box and Draper (1975) suggest that for a design to be insensitive to outliers, the v_{ii} should be constant.

A close examination of the residual variances or, equivalently, the variances of the predicted values, seems necessary in the analysis of any experiment. Apart from a proportionality constant, these variances are determined by the design points. We will refer to the smallest convex set containing all design points as the independent variable hull (IVH).

Cook (1977) developed a measure based on confidence ellipsoids for judging the contribution of each data point to the determination of the least squares estimate of $\boldsymbol{\beta}$. Other diagnostic measures have been suggested by Welsch and Kuh (1977) and Andrews and Pregibon (1977).

In this note, we discuss the roles that the residual variances, the residual correlations, and the IVH play in an analysis. In Section 2 the measure proposed by Cook (1977) is reviewed, some comments on its use are given, and relationships between the residual variances and the IVH are discussed. Consequences of deleting an influential observation are discussed in Section 3. Outliers are discussed in Section 4 and an example is presented.

2. INFLUENTIAL OBSERVATIONS

Cook (1977) proposed that the influence of the ith data point be judged by using the distance measure,

$$D_i \equiv [(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})]/(ps^2) \ . \tag{2.1}$$

Here, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ denote the estimates of $\boldsymbol{\beta}$ with and without the ith data point, respectively, and $s^2 = \mathbf{R}'\mathbf{R}/(n - p)$. A large value of D_i indicates that the associated ith point has a strong influence on the estimate of $\boldsymbol{\beta}$. The magnitude of the distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ may be assessed by comparing D_i to the probability points of the central F-distribution with p and $n - p$ degrees of freedom. This is equivalent to finding the level of the confidence ellipsoid centered at $\hat{\boldsymbol{\beta}}$ that passes through $\hat{\boldsymbol{\beta}}_{(i)}$, and entails nothing more than a monotonic transformation of D_i to a familiar scale.

The quantities controlling D_i can be seen in an equivalent form that depends only on the full data set:

$$D_i = t_i^2 w_i / p \ , \tag{2.2}$$

where $t_i = r_i / s(1 - v_{ii})^{\frac{1}{2}}$ is the ith studentized residual and $w_i = v_{ii}/(1 - v_{ii})$ is the ratio of the variance of the ith predicted value to that of the ith residual. Clearly,

[*] R. Dennis Cook is Associate Professor and Director of the Statistical Center, School of Statistics, University of Minnesota, St. Paul, MN 55108. Research for this article was supported in part by Grant 1 R01 GM25587-01 from the National Institute of General Medical Sciences, U.S. Department of Health, Education and Welfare. The author would like to thank Professors C. Bingham and S. Weisberg and the referees for their valuable comments.

© Journal of the American Statistical Association
March 1979, Volume 74, Number 365
Theory and Methods Section

D_i can be large if either t_i^2 or w_i is large. These two components measure the importance of two characteristics of each data point. The ith studentized residual, t_i^2, is a monotonic function of the likelihood-ratio test statistic for the hypothesis that the ith observation is not an outlier (Gentleman and Wilk 1975).

The ratio w_i depends only on the design points and reflects characteristics of the location of the ith point within the IVH. The characteristics which cause v_{ii} and hence w_i to be large may be seen by noting that for all $\mathbf{x}$ in the IVH ($\mathbf{x}$ need not be a design point),

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} \leq \max_i v_{ii} \ . \tag{2.3}$$

This follows because the surfaces of constant value of the quadratic form on the left of (2.3) are ellipsoids and the ellipsoid passing through the design point corresponding to $\max v_{ii}$ must contain the IVH. Expression (2.3) shows that the point with the largest variance of a predicted value must lie on the boundary of the IVH. Thus, large values of w_i indicate "outlying" design points. Of course, the point with the largest prediction variance need not be the one whose Euclidean distance from the centroid of the design is the greatest, since the v_{ii} depend also on the density of the points in the IVH: If $\mathbf{x}_j$ is any design point with k replicates, then

$$\mathbf{x}_j'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j \leq 1/k \ . \tag{2.4}$$

In general, we may anticipate that the design point corresponding to $\max v_{ii}$ will lie on the boundary of the IVH in a region where the density of the design points is relatively low.

The influence of a design point can also be seen by considering the variance of a predicted value conditional on the corresponding observed value. Letting $\mathbf{x}_r$ denote a design point with k replicated observations, y_{rj}, $j = 1, \ldots, k$, it can be demonstrated by induction that

$$V(\mathbf{x}_r'\hat{\boldsymbol{\beta}} \,|\, y_{rj}) = \sigma^2 v_{rr}(1 - k v_{rr}) \ . \tag{2.5}$$

This conditional variance will be small when $\mathbf{x}_r$ lies on the boundary of the IVH (kv_{rr} is large), or when $\mathbf{x}_r$ lies in the interior and k is large (v_{rr} is small).

The distance measure, D_i, can be easily extended to accommodate the situation in which q linearly independent combinations of the elements of $\boldsymbol{\beta}$ are of interest. Let $\hat{\boldsymbol{\phi}} = \mathbf{C}\hat{\boldsymbol{\beta}}$, where $\mathbf{C}$ is a $q \times p$ rank q matrix. The distance, $D_i(\boldsymbol{\phi})$, between $\hat{\boldsymbol{\phi}} = \mathbf{C}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\phi}}_{(i)} = \mathbf{C}\hat{\boldsymbol{\beta}}_{(i)}$ is defined to be:

$$D_i(\boldsymbol{\phi}) \equiv \frac{(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{(i)})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{(i)})}{qs^2} .$$

Note that

$$D_i(\boldsymbol{\phi}) \leq (p/q) D_i \tag{2.6}$$

for all i. When $\boldsymbol{\phi}$ consists of a subset, $\boldsymbol{\beta}_2$, of $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, we find that

$$D_i(\boldsymbol{\beta}_2) = \frac{t_i^2}{q} \frac{v_{ii} - v_{ii}^1}{1 - v_{ii}} , \tag{2.7}$$

where $v_{ii}^1\sigma^2$ is the variance of the ith predicted value from the regression on the first $p - q$ variables. Thus, the influence of an observation on a selected subset of $\hat{\boldsymbol{\beta}}$ may be easily determined from the results of two separate regressions.

3. CONSEQUENCES OF DELETING AN OBSERVATION

In this section we examine the behavior of the studentized residuals, the residual variances, and partial F-tests when an influential observation is deleted. Also, the importance and the causes of large residual correlations are discussed.

3.1 Studentized Residuals and Residual Variances

Let

$$v_{kl(i)} = \mathbf{x}_k'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_l \ ,$$

where $\mathbf{X}_{(i)}'\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'$ and $\mathbf{x}_i$ is the ith row of $\mathbf{X}$. Also, let $t_{j(i)}^2$ denote the jth studentized residual based on the data set with the ith point removed. The behavior of $t_{j(i)}^2$ and $v_{kl(i)}$ is best understood by expressing them in terms of the full data set. The following relationships will be useful:

$$v_{kl} = v_{kl(i)} - v_{ki(i)}v_{li(i)}/(1 + v_{ii(i)}) \ ; \tag{3.1}$$

$$v_{kl(i)} = v_{kl} + v_{ki}v_{li}/(1 - v_{ii}) \ ; \tag{3.2}$$

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i r_i/(1 - v_{ii}) \ ; \tag{3.3}$$

$$(n - p)s^2 = (n - p - 1)s_{(i)}^2 + r_i^2/(1 - v_{ii}) \ ; \tag{3.4}$$

where $s_{(i)}^2$ denotes the usual estimate of σ^2 based on the data set 'without the ith point. Expressions (3.1) and (3.2) are easily verified using the general identity:

$$[\mathbf{B} + \mathbf{u}\mathbf{z}']^{-1} = \mathbf{B}^{-1} - (\mathbf{B}^{-1}\mathbf{u}\mathbf{z}'\mathbf{B}^{-1})/(1 + \mathbf{u}'\mathbf{B}^{-1}\mathbf{z}) \ , \tag{3.5}$$

where $\mathbf{u}$ and $\mathbf{z}$ are column vectors and $\mathbf{B}$ is nonsingular. Expression (3.3) was shown by Miller (1974) and Cook (1977), and expression (3.4) is given in Beckman and Trussell (1974).

Letting ρ_{ij} denote the correlation between the ith and jth residuals in the full data set, it follows from (3.2) that

$$v_{jj(i)}/(1 - v_{jj(i)}) = \frac{v_{jj}(1 - \rho_{ij}^2) + \rho_{ij}^2}{(1 - v_{jj})(1 - \rho_{ij}^2)} \quad \text{for} \quad j \neq i \ . \tag{3.6}$$

This ratio will be large if either v_{jj} or ρ_{ij}^2 is large. A large value of v_{jj} would have been detected in the analysis of the full data set. Thus, if the variance of the jth predicted value increases substantially when the ith observation is deleted, it must be due to a large correlation between the ith and jth residuals in the full data set. Moreover, if the residual correlations are negligible, then the variances of the predicted values will remain essentially unchanged when any point is deleted.

Next, using (3.2)–(3.4), we find that

$$t_{j(i)}^2 = \frac{(n - p - 1)(t_j - \rho_{ij}t_i)^2}{(n - p - t_i^2)(1 - \rho_{ij}^2)} . \tag{3.7}$$

Notice that if the residual correlations are negligible, then the deletion of any point with a studentized residual larger than one will increase all remaining studentized residuals. Also a large value of ρ_{ij}^2 can cause the jth studentized residual to increase substantially when the ith point is deleted.

The distance measure in (2.2) for the data set without the ith observation can be expressed in terms of the full data set as the product of the right-hand sides of equations (3.6) and (3.7).

3.2 Residual Correlations

The previous discussion shows that large residual correlations can play a substantial role in locating influential observations. To investigate the causes of a large ρ_{ij}^2, we shall hold the jth design point fixed and find how to choose the ith design point so that ρ_{ij}^2 is maximized. Specifically, for first-order models, we consider $\sup\rho_{ij}^2$, where the supremum is to be taken over some convex subset of the factor space that consists of all permissible values for the ith design point. If the model contains a constant term, the first component of $\mathbf{x}_i$ is constrained to be 1, and the supremum must be taken with respect to the last $p - 1$ components of $\mathbf{x}_i$. Let $\mathbf{x}_r' = (1, \mathbf{z}_r')$ and, assuming that the independent variables in the reduced data set are measured around their means, let

$$(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1} = \begin{bmatrix} 1/(n-1) & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix}.$$

The required calculation is facilitated by using

$$v_{ij(i)} = 1/(n-1) + \mathbf{z}_i'\mathbf{A}\mathbf{z}_j$$

and equation (3.1) to express ρ_{ij}^2 in terms of explicit quadratic forms in $\mathbf{z}_i$.

The largest possible value for ρ_{ij}^2 will obviously depend on the subset of the factor space over which the supremum is taken. Lacking definite guidance, we choose a subset that seems reasonable and is, at least, expedient: For an arbitrary positive constant c, the supremum will be taken over $G(c) = \{\mathbf{z} \mid \mathbf{z}'\mathbf{A}\mathbf{z} \leq c\}$.

Letting $m = \min[c,\ n^2(v_{jj(i)} - 1/(n-1))]$, we find that the supremum must be obtained on the boundary of $G(m)$ and that

$$\sup_{\mathbf{z}_i \in G} \rho_{ij}^2 = \frac{mQ_{jj}^2}{(1 - v_{jj(i)})\{1 + [1/(n-1)] + m\} + mQ_{jj}^2}, \qquad (3.8)$$

where

$$Q_{jj} = \frac{1}{(n-1)\sqrt{m}} + \left(v_{jj(i)} - \frac{1}{n-1}\right)^{\frac{1}{2}}.$$

The supremum is attained at

$$z_i = z_j\{m/[v_{jj(i)} - 1/(n-1)]\}^{\frac{1}{2}}.$$

If the model does not contain a constant term, the analogous expression is obtained by replacing m with c and $1/(n-1)$ with zero. Also, if $m = n^2(v_{jj(i)} - 1/(n-1))$, then (3.8) reduces to

$$\sup_{\mathbf{z}_i \in G} \rho_{ij}^2 = v_{jj(i)} \frac{n}{n-1} - \frac{1}{n-1}.$$

These results show that higher correlations will arise between points on the boundary of the IVH and proportional points in the interior. Moreover, since the right side of (3.8) is monotonically increasing in $v_{jj(i)}$, we see that the highest correlations should occur between replicated design points on the boundary of the IVH. This observation confirms a conclusion reached in Section 3.1; namely, that if one of the two highly correlated observations is deleted, then the remaining observation is likely to become extremely influential. If one of two replicates at a design point on the boundary of the IVH is deleted, the remaining point will stand alone and, thus, may become extremely influential.

Finally, the correlations between residuals corresponding to a replicated design point are $-w_i = -v_{ii}/(1 - v_{ii})$.

3.3 Partial F-Tests

Partial F-tests for the hypothesis that the individual coefficients of $\boldsymbol{\beta}$ are zero are commonly used to simplify the original model. It is not uncommon to find that retention of a particular coefficient depends strongly on the presence of a single observation. This behavior seems particularly prevalent when the model contains polynomial terms.

Let $\hat{\beta}_k$ denote the kth component of $\hat{\boldsymbol{\beta}}$ and $T_k = \hat{\beta}_k/s(d_k)^{\frac{1}{2}}$, where d_k is the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. The partial F-statistic for the hypothesis that β_k is zero can be expressed as $F_k = T_k^2$.

Using (3.3)–(3.5), the partial F-statistic, $F_{k(i)}$, for the data set with the ith observation removed can be expressed as:

$$F_{k(i)} = \frac{(n-p-1)[T_k - \gamma t_i(w_i)^{\frac{1}{2}}]^2}{(n-p-t_i^2)(1+\gamma^2 w_i)}, \qquad (3.9)$$

where γ denotes the correlation between $\hat{\beta}_k$ and $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$.

Two general observations on the relationship between $F_{k(i)}$ and F_k are noteworthy: If $t_i^2 > 1$ and $\gamma(w_i)^{\frac{1}{2}}$ is negligible, then $F_{k(i)} > F_k$. Thus, deleting a point with $t_i^2 > 1$ in a dense region of the IVH will tend to increase all partial F-statistics. Next, if $t_i^2 \leq 1$, $F_k > 1$, and $\gamma(w_i)^{\frac{1}{2}}$ is sufficiently large, then $F_{k(i)} < F_k$. Thus, we can generally expect all $F_k > 1$ to decrease when a conforming point ($t_i^2 \leq 1$) on the boundary of the IVH is deleted.

4. OUTLIERS

In this section we briefly consider the problem of detecting multiple outliers in light of the foregoing discussion.

Consider the modified form of model (1.1), $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\theta} + \mathbf{e}$, where $\boldsymbol{\theta}$ consists of $n - \ell$ zeros and $\ell < n - p$ unknown parameters. Without loss of generality, we may assume that we wish to test the last ℓ observations as outliers. The vector of unknown parameters, $\boldsymbol{\theta}_\ell$, will now occupy the last ℓ positions of $\boldsymbol{\theta}$.

Gentleman and Wilk (1975) find the least squares estimate of $\boldsymbol{\theta}_\ell$ and show that the reduction in the sums of squares due to fitting the estimate is

$$\mathbf{R}_\ell'\mathbf{T}_\ell^{-1}\mathbf{R}_\ell \ , \qquad (4.1)$$

where $\mathbf{T}_\ell$ is the submatrix of $\mathbf{I} - \mathbf{V}$ consisting of the last ℓ rows and columns, and $\mathbf{R}_\ell$ is the vector of the last ℓ components of $\mathbf{R}$. The dependence of the usual normal theory F-statistic $F^{(\ell)}$ for the hypothesis $H: \boldsymbol{\theta}_\ell = 0$ on the studentized residuals and residual correlations can be seen by writing

$$F^{(\ell)} = \frac{\mathbf{t}_\ell'\boldsymbol{\varrho}_\ell^{-1}\mathbf{t}_\ell}{n - p - \mathbf{t}_\ell'\boldsymbol{\varrho}_\ell^{-1}\mathbf{t}_\ell} \cdot \frac{n - \ell - p}{\ell} \ , \qquad (4.2)$$

where $\mathbf{t}_\ell$ and $\boldsymbol{\varrho}_\ell$ are the matrices of studentized residuals and residual correlations for the last ℓ observations. For $\ell = 1$, we have $F^{(1)} = (n - p - 1)t_n^2/(n - p - t_n^2)$, which is a monotonically increasing function of the last studentized residual t_n^2.

Recall from equation (2.2) that as v_{ii} increases, the potential influence of the ith observations also increases. Thus, it seems important that we should have a good chance of detecting outliers at design points with a large v_{ii}. However, the noncentrality parameter, λ_ℓ, associated with the test of $H: \boldsymbol{\theta}_\ell = 0$ is $\lambda_\ell = \boldsymbol{\theta}_\ell'\mathbf{T}_\ell\boldsymbol{\theta}_\ell/2\sigma^2$, which reduces to $\lambda_1 = \theta_n^2(1 - v_{nn})/2\sigma^2$ for $\ell = 1$. This confirms the comments of Huber (1975) and Davies and Hutton (1975), and shows that, in general, outliers at points on the boundary of the IVH will be the most difficult to detect. Also, since the v_{ii} are increasing functions of p, outliers become more difficult to detect as the model is enlarged.

The above presentation is conditional on the a priori specification of the observations to be tested. It is perhaps more common to ask for the ℓ most likely outlying points. When this is the case, the quadratic form in (4.1) must be computed for all $C_{n,\ell}$ possible combinations of points. The combination producing the maximum value is then chosen to be tested using the statistic max$F^{(\ell)}$ which, of course, has the distribution of the maximum of $C_{n,\ell}$ correlated F-random variables. It is important to notice that the most likely pair of outlying points will not, in general, consist of the two points with the largest studentized residuals. The residual correlations can combine two seemingly unlikely candidates as the most likely outlier pair.

In short, if two outlying values are suspected, the residual correlations should be inspected. The observations, if any, which are most influencing the analysis may not be the ones with the larger studentized residuals. Points on the boundary of the IVH will generally be associated with higher residual correlations. It seems wise to give these points special attention.

If more than two outliers or influential observations are suspected, an examination of the residual correlations, while helpful, may not be sufficient. Bingham (1977) and Welsch and Peters (1978) suggest methods for examining more than two at a time.

5. EXAMPLE

Daniel and Wood (1971) considered a set of data on the oxidation of ammonia to nitric acid. The original data set is from Brownlee (1965), and consists of 21 observations with three possible explanatory variables. After a reasonably extensive analysis, Daniel and Wood decided that four observations (1, 3, 4, and 21) were outliers and that one of the explanatory variables was not needed. Their final model contained a linear and quadratic term for one explanatory variable, a linear term for the other, and was based on 17 "valid" observations.

In this example, we adopt the final model of Daniel and Wood but include all data points. The purpose of this example is to illustrate selected results of the previous sections by considering six selected subsets of the data. Tables 1, 2, and 3 give the values of D_i, v_{ii}, and t_i, respectively, for the six data sets. Table 4 gives the estimated coefficients, partial F-statistics, and mean squared error for each data set.

1. Distance Measures, D_i

Observation	Observations deleted					
	None	(21)	(4,21)	(2,4,21)	(1,2,4,21)	(1,3,4,21)
1	0.162	0.107	0.210	2.042	*	*
2	0.193	0.593	1.331	*	*	12.175
3	0.125	0.123	0.333	0.403	21.160	*
4	0.304	0.539	*	*	*	*
5	0.003	0.012	0.003	0.006	0.006	0.001
6	0.021	0.037	0.019	0.031	0.033	0.021
7	0.042	0.050	0.007	0.010	0.008	0.003
8	0.014	0.014	0.010	0.023	0.034	0.056
9	0.043	0.040	0.010	0.008	0.002	0.028
10	0.028	0.008	0.013	0.030	0.049	0.051
11	0.028	0.008	0.013	0.030	0.049	0.051
12	0.062	0.009	0.002	0.009	0.019	0.028
13	0.001	0.044	0.107	0.177	0.190	0.213
14	0.001	0.019	0.034	0.054	0.055	0.068
15	0.002	0.007	0.005	0.005	0.002	0.006
16	0.002	0.001	0.007	0.019	0.035	0.027
17	0.004	0.000	0.000	0.001	0.003	0.003
18	0.004	0.000	0.000	0.001	0.003	0.003
19	0.008	0.001	0.008	0.011	0.007	0.006
20	0.008	0.010	0.037	0.078	0.117	0.088
21	0.699	*	*	*	*	*

Consider first the complete data set. Inspection of the first column of Table 1 reveals that observation 21 is the most influential. Removal of this observation would move $\hat{\boldsymbol{\beta}}$ to the edge of a 40 percent confidence ellipse for $\boldsymbol{\beta}$ based on $\hat{\boldsymbol{\beta}}$. The reasons for this importance can be obtained from Tables 2 and 3. The three largest v_{ii} values are $v_{1,1} = v_{2,2} = 0.409$, and $v_{21,21} = 0.288$. Observations 1 and 2 lie on the edge of the IVH and are replicates. Observation 21 has the third largest value of v_{ii} and, thus, lies near the edge of the IVH. Moreover, from Table 3 we see that it has the largest studentized residual. Using Lund's (1975) tables of critical values, we see that

2. Values of v_{ii}

Obser-vation	Observations deleted: None	(21)	(4,21)	(2,4,21)	(1,2,4,21)	(1,3,4,21)
1	0.409	0.421	0.421	0.727	*	*
2	0.409	0.421	0.421	*	*	0.993
3	0.176	0.199	0.201	0.308	0.983	*
4	0.191	0.192	*	*	*	*
5	0.103	0.108	0.125	0.125	0.125	0.131
6	0.134	0.134	0.164	0.164	0.165	0.169
7	0.191	0.192	0.238	0.239	0.240	0.242
8	0.191	0.192	0.238	0.239	0.240	0.242
9	0.163	0.170	0.206	0.208	0.218	0.208
10	0.139	0.175	0.175	0.176	0.179	0.179
11	0.139	0.175	0.175	0.176	0.179	0.179
12	0.212	0.272	0.275	0.276	0.279	0.280
13	0.139	0.175	0.175	0.176	0.179	0.179
14	0.092	0.110	0.111	0.112	0.116	0.113
15	0.188	0.189	0.191	0.191	0.195	0.193
16	0.188	0.189	0.191	0.191	0.195	0.193
17	0.187	0.195	0.195	0.195	0.198	0.197
18	0.187	0.195	0.195	0.195	0.198	0.197
19	0.212	0.232	0.234	0.234	0.236	0.237
20	0.064	0.064	0.069	0.070	0.076	0.070
21	0.288	*	*	*	*	*

3. Studentized Residuals, t_i

Obser-vation	Observations deleted: None	(21)	(4,21)	(2,4,21)	(1,2,4,21)	(1,3,4,21)
1	0.97	0.77	1.08	−1.75	*	*
2	−1.06	−1.81	−2.71	*	*	0.57
3	1.54	1.40	2.30	1.91	1.21	*
4	2.27	3.01	*	*	*	*
5	−0.31	−0.63	−0.31	−0.40	−0.40	−0.12
6	−0.73	−0.97	−0.62	−0.80	−0.82	−0.64
7	−0.84	−0.92	−0.30	−0.35	−0.31	−0.18
8	−0.50	−0.48	0.36	−0.54	0.66	0.84
9	−0.94	−0.89	−0.39	−0.36	−0.18	−0.66
10	0.84	0.38	0.49	0.75	0.94	0.96
11	0.84	0.38	0.49	0.75	0.94	0.96
12	0.96	0.31	0.16	0.30	0.45	0.53
13	−0.17	−0.91	−1.42	−1.82	−1.87	−1.98
14	−0.25	−0.79	−1.04	−1.30	−1.29	−1.41
15	0.17	0.34	0.29	0.29	0.19	0 32
16	−0.17	−0.10	−0.35	−0.57	−0.76	−0.67
17	−0.26	−0.01	−0.01	−0.10	−0.23	−0.21
18	−0.26	−0.01	−0.01	−0.10	−0.23	−0.21
19	−0.35	0.09	0.33	0.37	0.31	0.27
20	0.68	0.75	1.42	2.04	2.38	2.16
21	−2.63	*	*	*	*	*

observation 21 may be declared an outlier at the 0.1 level but not at the 0.05 level of significance.

We are now faced with the decision to declare observation 21 an outlier or accept the data set as it stands. It must be remembered that when inspecting the studentized residuals, we are implicitly assuming the existence of at most one outlier. The most likely candidates for a pair of outliers are found, using the quadratic form in expression (4.1), to be 4 and 21. These observations also have the two largest studentized residuals. Also, $\max F^{(2)} = 21.62$ which, using a Bonferroni inequality, is significant at the 0.01 level. The three most likely candidates are 2, 4, and 21 with $\max F^{(3)} = 30.78$. Notice that these three points do not have the three largest studentized residuals. We consider next the data sets obtained by sequentially deleting points 21, 4, and 2.

The second column in each table shows the results after the removal of observation 21. Now, the conditional hypothesis $H\colon \theta_i = 0$ given $\theta_{21} \neq 0$ might be of interest. The test of this hypothesis is the same as the test of $H\colon \theta_i = 0$, based on the data set with observation 21 removed. Note that now observation 4 appears to be an outlier, although it is not the most influential, $\max D_i = D_2 = 0.593$. Table 3 shows that all values of v_{ii} increased when observation 21 was deleted. Of course, this was predicted by equation (3.6).

Next, when observations 4 and 21 are deleted, we find ourselves in a situation similar to that of the full data set. Observation 2 is the most influential, $D_2 = 1.33$. Removal of this observation will move the latest estimate of β to the edge of a 70 percent confidence ellipse. The reasons for this importance are that observation 2 lies on the edge of the IVH and appears to be an outlier. Table 4 shows that now the coefficient of X_1^2, $\hat\beta_3$, is highly significant. The increase in the partial F-statistic for $\hat\beta_3$ is due to a large value of t_4 and a small value of the correlation between $\hat\beta_3$ and $\mathbf{x}_4'\hat\beta$ ($\gamma = 0.057$) in the data set with only observation 21 deleted.

Recall that observation 2 is very influential and is one of two replicates (1 and 2) on the edge of the IVH. The residual correlation between observations 1 and 2 is $\rho_{12} = -0.421/(1 - 0.421) = -0.727$. Thus, when entertaining the removal of observation 2, we should anticipate that the characteristics of the IVH may change con-

4. Estimated Coefficients, $\hat\beta_k$, Partial F-Statistics, F_k, and Mean Squared Error, MSE

	Observations deleted											
	None		(21)		(4,21)		(2,4,21)		(1,2,4,21)		(1,3,4,21)	
Term	$\hat\beta_k$	F_k	$\hat\beta_k$	F_k	$\hat\beta_k$	F_k	$\hat\beta_k$	F_k	$\hat\beta_k$	F_k	$\hat\beta_k$	F_k
1	−14.30	0.19	−25.90	1.00	−3.74	0.04	13.26	0.85	33.74	3.96	−15.41	1.5
X_1	−0.46	0.21	0.07	0.01	−0.51	0.80	−1.10	5.95	−1.82	10.61	−0.07	0.03
X_1^2	0.009	1.27	0.006	0.97	−0.011	6.49	0.017	21.34	0.023	24.09	0.007	4.60
X_2	1.25	11.63	0.79	6.00	0.47	4.19	0.46	7.15	0.44	7.81	0.53	12.27
MSE	10.33		6.51		3.02		1.65		1.39		1.26	

siderably and that observation 1 will become more influential. Inspection of the fourth column in each table shows this change. In the corresponding data set (2, 4, and 21 deleted), design points 1 and 3 are approximately proportional and thus have a very high residual correlation, $\rho_{13} = -0.988$. This high ρ_{13} suggests that we should anticipate extreme changes when observation 1 is deleted. The fifth column in each table shows the results of deleting observation 1.

For comparison, the last column of each table gives the results based on the subset of the data that Daniel and Wood judged valid. Notice that observation 2 is extremely influential. The removal of this observation would move $\hat{\boldsymbol{\beta}}$ beyond the edge of a 99.95 percent confidence ellipse. This observation fits the model quite well and is influential because it stands alone on the edge of the IVH. From this, we can anticipate that if it were removed, all partial F-statistics would decrease. In fact, when observation 2 is removed, the first three partial F-statistics are all less than one, while F_4 decreases but remains fairly large. In addition, $D_2(\boldsymbol{\beta}_2)$ can be used to see how observation 2 influences subsets of $\hat{\boldsymbol{\beta}}$. Using equation (2.7), we find that $D_2(\beta_4) = 0.235$, while $D_2(\boldsymbol{\beta}_2) > 6.0$ for all other subsets, $\boldsymbol{\beta}_2$. Thus, observation 2 has little influence on $\hat{\beta}_4$, but has a strong influence on the estimates of all other coefficients. Moreover, inequality (2.6) can be used to verify that observation 2 is the only one which could have a strong influence on any subset. It appears that for the final data set of Daniel and Wood, the quadratic term is needed to model a single observation.

[*Received March 1977. Revised July 1978.*]

REFERENCES

Andrews, David F., and Pregibon, Daryl (1977), "Finding the Outliers that Matter," Technical Report No. 1, Institute of Applied Statistics, University of Toronto.

Beckman, R.J., and Trussell, H.J. (1974), "The Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Multiple Regression," *Journal of the American Statistical Association*, 69, 199–201.

Behnken, Donald W., and Draper, Norman R. (1972), "Residuals and Their Variance Patterns," *Technometrics*, 14, 101–111.

Bingham, Christopher (1977), "Some Identities Useful in the Analysis of Residuals from Linear Regression," Technical Report No. 300, School of Statistics, University of Minnesota.

Box, George E.P., and Draper, Norman R. (1975), "Robust Designs," *Biometrika*, 62, 347–352.

Brownlee, K.A. (1965), *Statistical Theory and Methodology* in Science and Engineering, 2nd ed., New York: John Wiley & Sons.

Cook, R. Dennis (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

Daniel, Cuthbert, and Wood, Fred S. (1971), *Fitting Equations to Data*, New York: John Wiley & Sons.

Davies, R.B., and Hutton, B. (1975), "The Effect of Errors in the Independent Variables in Linear Regression," *Biometrika*, 62, 383–391.

Gentleman, J.F., and Wilk, M.B. (1975), "Detecting Outliers. II. Supplementing the Direct Analysis of Residuals," *Biometrics*, 31, 387–410.

Hoaglin, David C., and Welsch, Roy E. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17–22.

Huber, Peter J. (1975), "Robustness and Designs," in *A Survey of Statistical Design and Linear Models*, ed. Jagdish N. Srivastava, Amsterdam: North-Holland Publishing Co.

Lund, Richard E. (1975), "Tables for an Approximate Test for Outliers in Linear Models," *Technometrics*, 17, 473–476.

Miller, Rupert G., Jr. (1974), "An Unbalanced Jackknife," *The Annals of Statistics*, 2, 880–891.

Welsch, Roy E., and Kuh, Edwin (1977), "Linear Regression Diagnostics," Working Paper No. 173, National Bureau of Economic Research.

———, and Peters, Stephan C. (1978), "Finding Influential Subsets of Data in Regression Models," in *Proceedings of the Eleventh Interface Symposium on Computer Science and Statistics*, eds. R. Gallant and T. Gerry, 240–244.