INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION

by

R. Dennis Cook

Technical Report #282

Department of Applied Statistics
University of Minnesota
Saint Paul, Minnesota 55108

February, 1977

Abstract

Characteristics of observations which cause them to be important in a least squares analysis of data arising from a non-designed experiment 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-test, studentized residuals, residual variances, convex hull of the independent variables and the estimated parameter vector. Outliers are discussed briefly and an example is presented.

1. INTRODUCTION

Consider the full rank linear regression model

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{Y}$ is an nx1 vector of observators, $\underset{\sim}{X}$ is an nxp full rank matrix of known constants, $\underset{\sim}{\beta}$ is a px1 vector of unknown parameters and $\underset{\sim}{\varepsilon}$ is an nx1 vector of randomly distributed errors such that $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ and $V(\underset{\sim}{\varepsilon}) = \underset{\sim}{I}\sigma^2$. Experimental designs for these models are usually constructed by first specifying the design space (e.g. some closed convex subset of R^p) and then choosing the design points within the space so that, in some optimal sense, the maximum amount of information will be obtained. For example, D-optimal designs produce confidence ellipsoids for $\underset{\sim}{\beta}$ with minimal volume. The analysis of designed experiments is usually well-known and straightforward. Also, the design space can be and usually is ignored during the analysis.

Unfortunately the idea of a design space is completely absent during most analyses of data arising from experiments in which the selection of the design points was relegated to Nature. In non-designed experiments the configuration of the design points in the factor space can have an important effect on the analysis. It has become increasingly apparent that failure to examine these configurations can result in severe misinterpretations and loss of information.

Behnken and Draper (1972) have noted that a wide variation in the variance of the residuals reflects a nonhomogeneous spacing of the design points. Box and Draper (1975) suggest that for a design to be insensitive to outliers the variances of the residuals should be constant. The comments

of Davies and Hutton (1975) also reflect the opinion that the residual variances should be examined.

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 $\underset{\sim}{\beta}$. The measure combines information from the studentized residuals and the residual variances, and shows that design points with relatively small residual variances will tend to be the more important. Huber (1975) mentioned that small residual variances typically correspond to "outlying" design points.

A close examination of the residual variances seems necessary in the analysis of any non-designed experiment. (In many designed experiments the residual variances are constant). However, the role that such an examination should play is somewhat vague. Recall that the covariance matrix of the residual vector, $\underset{\sim}{R}=(r_i)$, from a least squares analysis is given by

$$V(\underset{\sim}{R}) = (\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}')\sigma^2$$

where

$$\underset{\sim}{R} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}} = \underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}}$$

$$= (\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}')\underset{\sim}{Y}$$

Apart from a proportionately constant, the residual variances are determined by the design points (i.e. the rows of $\underset{\sim}{X}$). We find it convenient to refer to the smallest convex set containing all design points as the independent variable hull (IVH). Of course, the IVH and the design space are not in general the same.

Generally, in this note we discuss the role that the residual variances, residual correlations and the IVH play in the analysis of non-designed experiments. 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 observation are discussed in Section 3. Residual correlations and outliers are discussed in Sections 4 and 5. An example is presented.

2. INFLUENTIAL OBSERVATIONS

Cook (1977) proposed that the importance of the ith data point be measured by first computing the least squares estimate of β with and without the point and, second, measuring the distances, D_i, between the two estimates as a monotonic function of descriptive levels of significance. Specifically, let $\hat{\underset{\sim}{\beta}}$ denote the least squares estimate of $\underset{\sim}{\beta}$ based on the full data set and $\hat{\underset{\sim}{\beta}}_{(-i)}$ the analogous estimate without the ith point. Then

$$D_i \equiv \frac{(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(-i)})'\underset{\sim}{X}'\underset{\sim}{X}(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(-i)})}{ps^2} \qquad (1)$$

where $s^2 = \underset{\sim}{R}'\underset{\sim}{R}/(n-p)$. A large value of D_i indicates that the associated point has a strong influence on the least squares estimate of $\underset{\sim}{\beta}$. The magnitude of the distance between $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{\beta}}_{(-i)}$ is assessed by comparing D_i to the probability points of the central F-distribution with p and $n-p$ degrees of freedom. For example, suppose that D_i is equal to the 50% probability point, then the removal of the ith data point moves the least squares estimate to the edge of the usual 50% elliptical confidence region for $\underset{\sim}{\beta}$ centered at $\hat{\underset{\sim}{\beta}}$.

Let $v_{ii} = \underset{\sim}{x}_i'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}_i$ where $\underset{\sim}{x}_i'$ is the ith row of $\underset{\sim}{X}$. The quantities controlling the measure defined in equation (1) may be seen in an equivalent form that depends only on the full data set:

$$D_i = \frac{t_i^2}{p}\,\frac{V(\hat{y}_i)}{V(r_i)} \qquad (2)$$

where $t_i = r_i/s\sqrt{1 - v_{ii}}$ is the ith studentized residual, $V(\hat{y}_i) = \sigma^2 v_{ii}$ is the variance of the ith predicted value and $V(r_i) = \sigma^2(1 - v_{ii})$ is the variance of the ith residual. Clearly, D_i can be large if either t_i^2 or $V(\hat{y}_i)/(V(r_i))$ is large. These two components seem to measure the importance of two characteristics of each data point. The value of t_i^2 depends on the degree to which the ith observation conforms to the assumed model. It will be shown later that t_i^2 is a monotonic function of the likelihood ratio test statistic for the hypothesis that the ith observation is an outlier. Thus, a large value of D_i arising because of a large value of t_i^2 is an indication that the ith observation does not conform to the assumed model.

The ratio $V(\hat{Y}_i)/V(r_i)$ depends only on the design points and reflects characteristics of the ith point relative to the IVH. This ratio is a monotonically increasing function of v_{ii}. The particular characteristics which cause v_{ii} and hence $V(\hat{y}_i)/V(r_i)$ to be large may be seen by noting that for all $\underset{\sim}{x}$ in the IVH ($\underset{\sim}{x}$ need not be a design point)

$$\underset{\sim}{x}'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x} \leq \max_i v_{ii} . \qquad (3)$$

This follows because the levels of constant value of the quadratic form on the left of relationship (3) are ellipsoids and the ellipsoid passing through the design point corresponding to $\max v_{ii}$ must contain the IVH. Expression (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 $V(\hat{y}_i)/V(r_i)$ indicate "outlying" design points. Of course, the point with the largest prediction variance need not be the one whose Euclidean distance from the center of the design is the greatest since the values of $\underset{\sim}{x}'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}$ depend on the density of the points in the IVH. If $\underset{\sim}{x}_j$ is any design point with k replicates then

$$\underset{\sim}{x}_j'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}_j \leq \frac{1}{k} \quad . \tag{4}$$

This is easily justified by first noting that it is certainly true for $k=1$ and assuming that it is true for $k=k_0$. It can be shown to be true for $k=k_0 + 1$ by using the relationship

$$(\underset{\sim}{X}_+'\underset{\sim}{X}_+)^{-1} = (\underset{\sim}{X}'\underset{\sim}{X} + \underset{\sim}{x}_j\underset{\sim}{x}_j')^{-1}$$

$$= (\underset{\sim}{X}'\underset{\sim}{X})^{-1} - \frac{(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}_j\underset{\sim}{x}_j'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}}{1 + \underset{\sim}{x}_j'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}_j}$$

where $\underset{\sim}{X}$ and $\underset{\sim}{X}_+$ are the design matrices for k_0 and $k_0 + 1$ replicates of $\underset{\sim}{x}_j$.

Generally we may anticipate that the design point corresponding to $\max v_{ii}(\max V(\hat{y}_i)/V(r_i))$ will lie on the boundary of the IVH in a region where the density of the design points is relative low. In non-designed experiments it is not uncommon for $\max v_{ii}$ to be close to one.

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 $\underset{\sim}{x}_r$ denote a design point with k replicated observations, y_{r_j}, $j = 1,k$, it can be demonstrated by induction that

$$V(\underset{\sim}{x}_r' \hat{\underset{\sim}{\beta}} | y_{r_j},\ j = 1,k) = \sigma^2\, v_{rr}(1 - k\, v_{rr}) \ .$$

This conditional variance will be small when $\underset{\sim}{x}_r$ lies on the boundary of the IVH ($k\, v_{rr}$ is large) or when $\underset{\sim}{x}_r$ lies in the interior and k is large (v_{rr} is small). Generally, the remaining points will have little influence over the predictions around a design point with a small conditional variance.

Some additional insight about the dispersion of the points within the IVH may be obtained by inspecting the ordered values, $v_{(ii)}$. Since the ith design point lies on an ellipse whose value is v_{ii} any "large gaps" between the individual elements indicates a corresponding gap in the spacing of the design points. A large gap may be taken as an indication of a region in the design space with relatively inadequate coverage.

Equation (3) also provides a guide to the region in the factor space where a final model may be appropriate for the purpose of prediction. Prediction at any point $\underset{\sim}{x}$ for which (3) does not hold may be tantamount to extrapolation. In this case $\underset{\sim}{x}$ cannot be in the IVH. Using this rule it is possible to have extreme situations in which the design points are virtually the only points for which the model is appropriate for prediction.

The previous discussion relates the properties of D_i to the behavior of its individual components. It is also of interest to relate the behavior of D_i to its components simultaneously. If a point appears to be an outlier (t_i^2 is large) and at the same time lies in a high density region of the IVH ($V(\hat{y}_i)/V(r_i)$ is small) then it may be irrelevant whether the point is excluded or not at least as far as estimating $\underset{\sim}{\beta}$ is concerned. A small value of D_i indicates that $\hat{\underset{\sim}{\beta}}$ and $\hat{\beta}_{(-i)}$ are essentially the same. The same conclusion may hold if a point lies on the boundary of the IVH and fits the model extremely well.

3. CONSEQUENCES OF DELETING AN OBSERVATION

The measure discussed in the previous section provides a method for judging the importance of each observation through the implicit deletion of the observations. It does not, however, remove the necessity of actually deleting highly influential observations and examining the new solution. When this is done it must be remembered that the removal of any observation on the boundary of the IVH may affect a considerable change in the hull. Changing the IVH by removing an observation on its boundary may cause observations that were previously judged to be noninfluential to become influential. Such behavior indicates that the solution is not stable across the full IVH. This should be cause for concern since the final form of the solution depends on a few points whose presence or absence was originally left to chance. It may be desirable in such circumstances to post-design the experiment by deleting the observations and confining inference to a smaller region of the factor space. Of course, the influential observations should not be completely forgotten but should perhaps be the subject for future investigation. Our main contention is that the use of such observations without an independent verification of their authenticity or a well grounded firm belief in the model is not a sound practice.

3.1 Distance Measures

At present, sequential deletion of observations seems the most expedient method for detecting multiple influential points. In what follows we examine the effects of deleting a single influential point. Let $D_{j(-i)}$ denote the value of the distance measure for the jth point based on the data set from which the ith point has been removed, $i \neq j$.

The characteristics of $D_{j(-i)}$ seem best understood by expressing it in terms of the full data set.

Let

$$v_{ij} = \underset{\sim}{x}_i'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}_j$$

and

$$w_{k\ell} = \underset{\sim}{x}_k'(\underset{\sim}{X}_{(-i)}'\underset{\sim}{X}_{(-i)}')^{-1}\underset{\sim}{x}_\ell$$

where

$$\underset{\sim}{X}_{(-i)}'\underset{\sim}{X}_{(-i)} = \underset{\sim}{X}'\underset{\sim}{X} - \underset{\sim}{x}_i\underset{\sim}{x}_i'$$

and $\underset{\sim}{x}_r'$ is the rth row of $\underset{\sim}{X}$. Also, let $s^2_{(-i)}$ denote the usual estimate of σ^2 based the data set with the ith point removed.

The following relationships will be useful:

$$v_{ij} = w_{ij}/(1 + w_{ii}) \tag{5}$$

$$v_{jj} = w_{jj} - w_{ij}^2/(1 + w_{ii}) \tag{6}$$

$$\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(-i)} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}_i[y_i - \underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}}]/(1 - v_{ii}) \tag{7}$$

$$(n-p)s^2 = (n-1-p)s^2_{(-i)} + (y_i - \underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}})^2/(1 - v_{ii}) \ . \tag{8}$$

Expressions (5) and (6) are easily verified using the identity

$$(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = (\underset{\sim}{X}_{(-i)}'\underset{\sim}{X}_{(-i)})^{-1} - \frac{(\underset{\sim}{X}_{(-i)}'\underset{\sim}{X}_{(-i)})^{-1}\underset{\sim}{x}_i\underset{\sim}{x}_i'(\underset{\sim}{X}_{(-i)}'\underset{\sim}{X}_{(-i)})^{-1}}{1 + w_{ii}} \ .$$

Expression (7) was shown by Cook (1977) and expression (8) was shown by Beckman and Trussell (1974). The derivations are straightforward and will not be repeated here.

From equation (2),

$$D_{j(-i)} = t^2_{j(-i)} w_{jj}/p(1 - w_{jj}) \qquad (9)$$

where $t^2_{j(-i)}$ denotes the studentized residual for the jth point in the data set with the ith point removed. We now relate equation (9) to the full data set by dealing with $t^2_{j(-i)}$ and w_{jj} separately. Consider first w_{jj}: Using equations (5) and (6) it is easily verified that

$$v_{jj} = w_{jj} - \rho^2_{ij}(1 - v_{jj})$$

where ρ_{ij} denotes the correlation coefficient between the ith and jth residuals in the full data set. It follows that

$$w_{jj}/(1 - w_{jj}) = \frac{v_{jj}(1 - \rho^2_{ij}) + \rho^2_{ij}}{(1 - v_{jj})(1 - \rho^2_{ij})} . \qquad (10)$$

Clearly, this ratio will be large if either v_{jj} or ρ^2_{ij} 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. We see also that if the residual correlations are negligible then the variances of the predicted values will remain essentially unchanged when any point is deleted.

Next,

$$t^2_{j(-i)} = \frac{(y_j - \underset{\sim}{x}_j'\hat{\beta}_{(-i)})^2}{s^2_{(-i)}(1 - w_{jj})} .$$

Using the results in equations (5) - (7) a little algebra will verify that

$$t^2_{j(-i)} = s^2[t_j - \rho_{ij}t_i]^2/s^2_{(-i)}(1 - \rho^2_{ij})$$

where t_i and t_j are studentized residuals from the full data set. Using equation (8) this reduces to

$$t^2_{j(-i)} = \frac{(n - p - 1)(t_j - \rho_{ij}t_i)^2}{(n - p - t^2_i)(1 - \rho^2_{ij})} \quad . \tag{11}$$

Notice that if the residual correlations are negligible then the deletion of any point with a studentized residual larger than one will cause all remaining studentized residuals to increase.

Finally, combining (10) and (11) we have,

$$D_{j(-i)} = \frac{(n - p - 1)[t_j - \rho_{ij}t_i]^2 v_{jj}(1 - \rho^2_{ij}) + \rho^2_{ij}}{(n - p - t^2_i)(1 - \rho^2_{ij})^2(1 - v_{jj})p} \tag{12}$$

The rather complicated form of this expression makes definite predictions concerning its behavior difficult. Two general observations seem worthy of mention, however: If the residual correlations are all negligible then $D_{j(-i)} \geq D_j$ for all $j \neq i$ whenever $t^2_i \geq 1$. Also, a large residual correlation can cause $D_{j(-i)}$ to increase substantially.

It is clear from the previous discussion that residual correlations can play a substantial role in the isolation of influential observations. It seems natural to question those characteristics of two observations which cause their residuals to be highly correlated. This will be considered in Section 4.

3.2 Partial F-Tests

Partial F-tests for the hypothesis that the individual coefficients of $\underset{\sim}{\beta}$ are zero are commonly used to simplify the original model. When using this procedure it is not uncommon to find that whether a particular coefficient is retained in the model depends 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 $\hat{\underset{\sim}{\beta}}$ and define

$$T_k = \hat{\beta}_k / s\sqrt{d_k}$$

where d_k is the kth diagonal element of $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$. The partial F-statistic, F_k, for the hypothesis that β_k is zero can be expressed as

$$F_k = T_k^2 = \hat{\beta}_k^2 / s^2 d_k .$$

Further, let $\hat{\beta}_{k(-i)}$, $T_{k(-i)}$, $d_{k(-i)}$, and $F_{k(-i)}$ denote the analogous quantities based on the data set without the ith observation.

The factors controlling the behavior of the partial F-statistics can be seen by expressing $F_{k(-i)}$ in terms that depend only on the complete data set. We consider the three components comprising $F_{k(-i)}$ separately: Using equation (7),

$$\hat{\beta}_{k(-i)} = \hat{\beta}_k - c_{ki} r_i / (1 - v_{ii})$$

where

$$c_{ki} = \underset{\sim}{e}_k'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}_i$$

and $\underset{\sim}{e}_k$ is a pxl vector with a 1 in the kth position and zeros elsewhere. Also,

$$d_{k(-i)} = d_k + c_{ki}^2/(1 - v_{ii}) \ .$$

The desired expression for $s^2_{(-i)}$ derives immediately from equation (8). After substituting these three forms into

$$F_{k(-i)} = \hat{\beta}^2_{k(-i)}/s^2_{(-i)}d_{k(-i)}$$

a little algebra will verify that

$$F_{k(-i)} = \frac{(n-p-1)t_i^2}{(n-p-t_i^2)} \frac{[\frac{T_k}{t_i} - \gamma(\frac{v_{ii}}{1-v_{ii}})^{\frac{1}{2}}]^2}{(1+\gamma^2 v_{ii}/(1-v_{ii}))} \qquad (12a)$$

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

Recall that $v_{ii}/(1-v_{ii})$ appears in the expression of the distance measure, D_i, and will be relatively large for points on the boundary of the IVH. The term $(n-p-1)t_i^2/(n-p-t_i^2)$ is monotonic in t_i^2 and may be used to test whether the ith observation is an outlier. In fact, under the null hypothesis it is distributed as an F random variable with 1 and n-p-1 degrees of freedom. (This will be discussed in more detail in Section 5.)

It seems clear from inspection of equation (12a) that almost anything can happen to the partial F-statistics when an observation is removed. Two general observations seem particularly interesting, however: Suppose that the deleted observation appears to be an outlier (t_i^2 is large) and that $\gamma\,(v_{ii}/1-v_{ii})^{\frac{1}{2}}$ is negligible. The latter supposition

would hold when γ is small or the deleted observation lies in a dense region of the design space. (Empirical investigations indicate that typically γ is not negligible by itself.) In this case,

$$F_{k(-i)} \doteq F_k\left(\frac{n-p-1}{n-p-t_i^2}\right) > F_k \ .$$

Thus, deleting an observation with $t_i^2 > 1$ in a dense region of the design space will tend to increase all partial F-statistics.

Next, consider the deletion of a point that fits the model quite well $(t_i^2 \leq 1)$. In this case,

$$F_{k(-i)} \doteq \frac{[T_k - \gamma t_i (v_{ii}/1-v_{ii})^{\frac{1}{2}}]^2}{1 + \gamma^2 v_{ii}/1-v_{ii}} \ .$$

Using this approximation it is easily verified that

$$F_k - F_{k(-i)} \doteq \frac{[\gamma T_k v_{ii}/(1-v_{ii}) + t_i]^2}{1 + \gamma^2 v_{ii}/1-v_{ii}} - t_i^2 \ .$$

Thus, we can generally expect all partial F-statistics greater than one to decrease when a conforming point on the boundary of the design space is deleted.

4. RESIDUAL CORRELATIONS

The squared correlation coefficient, ρ_{ij}^2, between the ith and jth residuals can be expressed as,

$$\rho_{ij}^2 = v_{ij}^2/(1-v_{ii})(1-v_{jj}) \ . \tag{13}$$

To investigate the causes of a large value for ρ_{ij}^2, we shall hold the jth design point fixed and find how to choose the ith design point so that the correlation between the associated residuals is maximized. Specifically, 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. The required calculation is facilitated by writing ρ_{ij}^2 in terms of explicit quadratic forms in $\underset{\sim}{x}_i$. Using equations (5) and (6) it is easily verified that

$$\rho_{ij}^2 = \frac{w_{ij}^2}{(1+w_{ii})(1-w_{jj}) + w_{ij}^2} \ . \tag{14}$$

w_{ij}^2 and w_{ii} are quadratic forms in $\underset{\sim}{x}_i$ while w_{jj} is independent of $\underset{\sim}{x}_i$ and may be considered constant. If the model contains a constant term the first component of $\underset{\sim}{x}_i$ is constrained to be 1 and the supremum must be taken with respect to the last p-1 components of $\underset{\sim}{x}_i$. Let $\underset{\sim}{x}_r' = (1, \underset{\sim}{z}_r')$ and, assuming that the independent variables in the reduced data set are measured around their means,

$$(\underset{\sim}{X}_{(-i)}' \underset{\sim}{X}_{(-i)}')^{-1} = \begin{pmatrix} \frac{1}{n-1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{A} \end{pmatrix}.$$

The correlation may now be expressed in a more manageable form by substituting

$$w_{ij} = \frac{1}{n-1} + \underset{\sim}{z}_i' \underset{\sim}{A} \underset{\sim}{z}_j$$

into equation (14).

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 $G(c) = \{\underset{\sim}{z} | \underset{\sim}{z}' \underset{\sim}{A} \underset{\sim}{z} \leq c\}$. Of course, c may be chosen so that G contains the IVH corresponding to the reduced data set. From equation (14) it is not difficult to see that the supremum must be obtained on the boundary of $G(\min[c, n^2(w_{jj} - 1/(n-1))])$. It follows that

$$\sup_{\underset{\sim}{z}_i \in G} \rho_{ij}^2 = \frac{c' Q_{jj}^2}{(1-w_{jj})(1+\frac{1}{n-1}+c') + c' Q_{jj}^2} \qquad (15)$$

where

$$Q_{jj} = \frac{1}{(n-1)\sqrt{c'}} + (w_{jj} - \frac{1}{n-1})^{\frac{1}{2}}$$

and

$$c' = \min(c, n^2(w_{jj} - \frac{1}{n-1})) .$$

This value is attained at $\underset{\sim}{z}_i = \underset{\sim}{z}_j(c'/(w_{jj} - \frac{1}{n-1}))^{\frac{1}{2}}$. If the model does not contain a constant term the analogous expression is obtained by setting $c'=c$ and $1/n-1$ to zero.

This result shows that the higher correlations will arise between points on the boundary of the IVH and proportional points in the interior. Moreover, since equation (15) is monotonically increasing in w_{jj} we see that the highest correlations should occur between replicated design points on the boundary of the IVH. This observation confirms a conclusion reaching in Section 3; namely, if one of two highly correlated observations is deleted the remaining observation is likely to become extremely important. If one of two replicates of a design point on the boundary of the IVH is deleted the remaining point will stand alone and, thus, may become extremely important.

The previous discussion shows the general relationship between points whose residuals are the most highly correlated. It may be, however, that the highest possible correlation is quite small. If n is large and $c' = n^2(w_{jj} - 1/n-1)$ then a convenient approximation for equation (15) is

$$\sup_{\underset{\sim}{z}_i \in G} \rho_{ij}^2 \doteq w_{jj} .$$

Since equation (15) is monotonically increasing in $c \leq c'$,

$$\rho_{ij}^2 \leq w_{jj} \tag{16}$$

for all i. This in combination with equation (4) justifies the previous comment. Equation (16) may be used as a rough guide to the behavior of the correlations between any influential point and all other points.

Finally, it is worth noting that the correlations between residuals corresponding to replicated design points are $-v_{ii}/1-v_{ii}$.

5. OUTLIERS

Outliers play an important role in the concept of influential observations. The approaches used in the detection of outliers are manifold and no pretence of an exhaustive presentation is made. In this section we consider briefly the problem of detecting multiple outlying points as it relates to the previous discussion.

Consider the model,

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\theta} + \underset{\sim}{\varepsilon} .$$

All quantities are as previously defined except that $\underset{\sim}{\theta}$ is a vector consisting 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 unknown parameters in $\underset{\sim}{\theta}$ will now occupy the last ℓ positions. Generally, we are concerned with estimation of $\underset{\sim}{\theta}$ and the test of $H:\underset{\sim}{\theta} = 0$. The development is greatly facilitated by reparameterizing the model to obtain the following form:

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\alpha} + \underset{\sim}{TZ\theta}_{\ell} + \underset{\sim}{\varepsilon} \qquad (17)$$

where $\underset{\sim}{\alpha} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'\,(\underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\theta})$

$$\underset{\sim}{T} = (\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}')$$

$$Z = \begin{pmatrix} 0_{(n-\ell)x\ell} \\ I_{\ell x \ell} \end{pmatrix}$$

and $\underset{\sim}{\theta}_{\ell}$ is the $\ell x 1$ vector consisting of the last ℓ components of $\underset{\sim}{\theta}$.

Since $\underset{\sim}{X}$ and $\underset{\sim}{T}$ are orthogonal the least squares estimate of $\underset{\sim}{\theta}_{\ell}$ can be written as,

$$\hat{\underset{\sim}{\theta}}_{\ell} = (\underset{\sim}{Z}'\underset{\sim}{T}'\underset{\sim}{T}\underset{\sim}{Z})^{-1}\ \underset{\sim}{Z}'\underset{\sim}{T}'\underset{\sim}{Y}$$

$$= (\underset{\sim}{Z}'\underset{\sim}{T}\ )^{-1}\ \underset{\sim}{Z}'\underset{\sim}{T}\underset{\sim}{Y} \qquad (18)$$

It can be shown that $\underset{\sim}{Z}'\underset{\sim}{T}\underset{\sim}{Z}$ is positive definite as long as $[\underset{\sim}{X}'\vdots\underset{\sim}{Z}]$ is of full rank. Alternatively, $\hat{\underset{\sim}{\theta}}_{\ell}$ may be written as

$$\hat{\underset{\sim}{\theta}}_{\ell} = (\underset{\sim}{T}_{\ell})^{-1}\ \underset{\sim}{R}_{\ell}$$

where $\underset{\sim}{T}_{\ell}$ is the submatrix of $\underset{\sim}{T}$ consisting of the last ℓ rows and columns, and $\underset{\sim}{R}_{\ell}$ is a vector of the last ℓ components of $\underset{\sim}{R}_{\ell}$ residual vector from the model with $\underset{\sim}{\theta} = 0$.

The reduction in the sums of squares due to fitting $\hat{\underset{\sim}{\theta}}_{\ell}$ is

$$\underset{\sim}{R}'_{\ell}\ \underset{\sim}{T}_{\ell}^{-1}\underset{\sim}{R}_{\ell} \qquad (19)$$

and the usual normal theory F-statistic for the hypothesis $H:\underset{\sim}{\theta}_{\ell} = 0$ is

$$F_{\ell} = \frac{\underset{\sim}{R}_{\ell}\underset{\sim}{T}_{\ell}^{-1}\underset{\sim}{R}_{\ell}}{\underset{\sim}{R}'\underset{\sim}{R} - \underset{\sim}{R}'_{\ell}\underset{\sim}{T}_{\ell}^{-1}\underset{\sim}{R}_{\ell}} \cdot \frac{n-\ell-p}{\ell} \ .$$

The dependence on the studentized residuals and residual correlations can be seen by writing F_{ℓ} in the alternate form,

$$F_{\ell} = \frac{\underset{\sim}{t}'_{\ell} \underset{\sim}{\rho}^{-1} \underset{\sim}{t}_{\ell}}{n-p-\underset{\sim}{t}'_{\ell} \underset{\sim}{\rho}_{\ell}^{-1} \underset{\sim}{t}_{\ell}} \cdot \frac{n-\ell-p}{\ell} \qquad (20)$$

where $\underset{\sim}{t}_{\ell}$ and $\underset{\sim}{\rho}_{\ell}$ 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 clearly a montonically increasing function of the last studentized residual, t_n^2.

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 equation (19) must be computed for all $\binom{n}{\ell}$ possible combinations of points. The combination producing the maximum value is then chosen to be tested using the statistic

$$F_{max} = \max_{\binom{n}{\ell}} F_{\ell} \qquad (21)$$

which, of course, has the distribution of the maximum of $\binom{n}{\ell}$ correlated F-random variables. It is important to notice that the two most likely outlying points will not, in general, correspond to the two points with the largest studentized residuals. The residual correlations can produce two seemingly unlikely candidates as the most outlying values.

When $\ell \geq 2$ and $\underset{\sim}{\rho}_{\ell}$ is strongly diagonal the well-known approximation

$$\underset{\sim}{\rho}_{\ell}^{-1} \doteq 2I - \underset{\sim}{\rho}_{\ell}$$

can be used to display the effects of the residual correlations in determining the critical points: Substituting this approximation into equation (19) we obtain,

$$\underset{\sim}{R}_{\ell}' \underset{\sim}{T}_{\ell}^{-1} R_{\ell} \doteq \Sigma t_i^2 - \sum_{i \neq j} t_i t_j \rho_{ij}$$

where the summations are over those observations in the subset in question.

In short, if two or more outlying values are suspected the residual correlations should be inspected. The observations, if any, which are disturbing 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. At the very least, it seems wise to give these points special attention.

6. 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 4 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. For ease of reference the data set has been given in Table 1. The general purpose of this example is to illustrate selected results of the previous sections by considering 6 selected subsets of the data from Table 1. Tables 2, 3, and 4 give the values of D_i, v_{ii}, and t_i, respectively, for the six data sets. Table 5 gives the estimated coefficients, partial F-statistics, and mean square error for each data set. Note that the last data set used in each table consists of the observations that Daniel and Wood judged valid.

Consider first the complete data set. Inspection of the first column of Table 2 reveals that observation 21 is the most influential. Removal of this observation would move the least estimate of β to the edge of a 40% confidence ellipse for β based on $\hat{\beta}$. The reasons for this importance can be obtained from Tables 3 and 4. The four largest v_{ii} values are, $v_{1,1} = v_{2,2} = 0.409$, $v_{21,21} = 0.288$, $v_{19,19} = 0.212$. Observations 1 and 2 lie on the edge of the IVH

and are replicates. The spacing of these values indicates two gaps in the coverage of the design space. Observation 21 has the third largest value of v_{ii} and, thus, lies near the edge of the IVH. Moreover, from Table 4 we see that it has the largest studentized residual. Using Lund's (1975) tables of critical values we see that observation 21 may be declared an outlier at the 0.1 level but not at the 0.05 level of significance. Thus, there is some evidence to suggest that it is an outlier, although it is not overwhelming.

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, (i.e. the vector $\underset{\sim}{\theta}$ in equation (16) has at most one nonzero component). The two most likely candidates as outliers are found, using the quadratic form in equation (19), to be 4 and 21. These observations also have the two largest studentized residuals. The value of the statistic in equation (21) is $\max F_2 = 2.82$ which is about the 90% probability point of the F-distribution with 2 and 15 degrees of freedom. Since this probability point is an upper bound for the probability point of the distribution of $\max F_2$, we cannot justify the simultaneous deletion of observations 4 and 21 on significance levels alone.

In this example, we shall treat observation 21 as an outlier. Now, the conditional hypothesis $H:\theta_i = 0 | \theta_{21} \neq 0$ might be of interest. It seems intuitively obvious and is easy to show that the tests of this hypothesis is the same as the test of $H:\theta_i = 0$ based on the data set

with observation 21 removed. The second column in each table shows the results after the removal of observation 21. Note that now observation 4 appears to be an outlier although it is not the most influential, $\max D_i = D_2 = 0.593$. A comparison of the first two columns in Table 3 shows that all values of v_{ii} increased when observation 21 was deleted. Of course, this was predicted by equation (10). Also, the values in the second column of Table 3 may be taken as upper bounds on the residual correlations $\rho^2_{i,21}$ $(i = 1, 20)$ in the full data set.

Next, considering the data set with observations 4 and 21 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% confidence ellipse. The reasons for this importance are that observation 2 lies on the edge of the IVH and appears to be an outlier. (It may be rejected at the 0.05 level.) From Table 5 we see 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 the large value of t_4 and a small value of the correlation between $\hat{\beta}_3$ and $\underset{\sim}{x}_4\hat{\beta}$ $(\gamma = .957)$ 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 $\rho_{12} = -.421/(1 - .421) = -0.727$. Thus, when entertaining the removal of observation 2 we should anticipate that the characteristics of the IVH may change considerably and that observation 1 will become more influential. Inspection of the fourth column in each table shows this change. Notice that now the gaps in the spacing of the design points have widened considerably.

In the latest data set (2, 4, and 21 deleted) design points 1 and 3 are approximately proportional and thus have a very high residual correlation, $\rho_{13} = -0.988$. (If the third design point were replaced with a replicate of the first the residual correlation would increase to -0.993.) This high residual correlation suggests that we should anticipate extreme changes if observation 3 were deleted. The fifth column in each table shows the results of deleting observation 3.

For comparison, we have included in the last column of each table 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 the least squares estimates beyond the edge of a 99.95% confidence ellipse. This observation fits the model quite well and is important 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. It appears that for the final data set of Daniel and Wood the quadratic term is needed to model a single observation.

REFERENCES

[1] 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.

[2] Behnken, D.W. and Draper, N.R., (1972). Residuals and their variance patterns. Technometrics 14, 101-111.

[3] Box, G.E.P. and Draper, N.R., (1975). Robust design. Biometrika 62, 347-352.

[4] Cook, R.D., (1977). Detection of influential observations in linear regression. Technometrics 19, 15-18.

[5] Daniel, C. and Wood, F.S., (1971). Fitting Equations to Data. John Wiley & Sons, Inc., New York.

[6] Davies, R.B. and Hutton, B., (1975). The effects of errors in the independent variables in linear regression. Biometrika 62, 383-391.

[7] Huber, P.J., (1975). Robustness and designs. A Survey of Statistical Design and Linear Models. North-Holland, Amsterdam.

[8] Longley, J.W., (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. J. Amer. Statist. Assoc. 62, 819-841.

[9] Lund, R.E., (1975). Tables for an approximate test for outliers in linear models. Technometrics 17, 473-476.

TABLE 1

DATA ON THE OXIDATION OF AMONIA
TO NITRIC ACID

Observation Number	Air Flow x_1	(Air Flow)^2 x_1^2	Cooling Water Inlet Temperature x_2	Stack Loss y
1	80	6400	27	42
2	80	6400	27	37
3	75	5625	25	37
4	62	3844	24	28
5	62	3844	22	18
6	62	3844	23	18
7	62	3844	24	19
8	62	3844	24	20
9	58	3364	23	15
10	58	3364	18	14
11	58	3364	18	14
12	58	3364	17	13
13	58	3364	18	11
14	58	3364	19	12
15	50	2500	18	8
16	50	2500	18	7
17	50	2500	19	8
18	50	2500	19	8
19	50	2500	20	9
20	56	3136	20	15
21	70	4900	20	15

TABLE 2

Distance Measures, D_i, Based on Selected Subsets of Observations From Table 1.

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	*	*	*	*	*

TABLE 3

Values of v_{ii} Based on Selected Subsets of Observations From Table 1.

	Observations Deleted					
Observation	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	*	*	*	*	*

TABLE 4

Studentized Residuals, t_i, Based on
Selected Subsets of Observations from Table 1.

Observation	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	*	*	*	*	*

TABLE 5

Estimated Coefficients, $\hat{\beta}_k$, Partial F-Statistics, F_k, and Mean Square Error, MSE, Based on Selected Subsets of Observations from Table 1.

	Observations Deleted							
	None		(21)		(4,21)		(2,4,21)	
Term	$\hat{\beta}_k$	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
X_1	-0.46	0.21	0.07	0.01	-0.51	0.80	-1.10	5.95
X_1^2	0.009	1.27	0.006	0.97	-0.011	6.49	0.017	21.34
X_2	1.25	11.63	0.79	6.00	0.47	4.19	0.46	7.15
MSE	10.33		6.51		3.02		1.65	

	(1,2,4,21)		(1,3,4,21)	
Term	$\hat{\beta}_k$	F_k	$\hat{\beta}_k$	F_k
1	33.74	3.96	-15.41	1.5
X_1	-1.82	10.61	- 0.07	0.03
X_1^2	0.023	24.09	0.007	4.60
X_2	0.44	7.81	0.53	12.27
MSE	1.39		1.26	