DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN

Madison, Wisconsin 53706

TECHNICAL REPORT NO. 581

August 1979

INFLUENTIAL OBSERVATIONS AND OUTLIERS IN REGRESSION

by

N.R. Draper
Dept. of Statistics
University of Wis.
Madison, Wisconsin 53706

and

J.A. John
Dept. of Mathematics
The University
Southampton, England

Influential Observations and Outliers in Regression

N.R. Draper
Department of Statistics
University of Wisconsin
Madison, Wisconsin 53706

and

J.A. John
Department of Mathematics
The University
Southampton, England

Statistics offered by Cook (1977) and Andrews and Pregibon (1978) purport to reveal observations which are influential in the data set. Detailed examination of these statistics shows that two different types of influence are being measured, and this is illustrated with examples derived from a set of data given by Mickey, Dunn and Clark (1967). Recommendations are given for obtaining the best use of the statistics available.

Keywords: Influential observations; Outliers; Regression; Residuals.

1. INTRODUCTION

The fact that an observation is an outlier, that is, provides a large residual when the chosen model is fitted to the data does not necessarily mean that the observation is an influential one with respect to the fitted equation. When an outlier is omitted from the analysis, the fitted equation may change hardly at all. An example given by Andrews and Pregibon (1978), using data from Mickey, Dunn and Clark (1967) illustrates the point well. The observation with the largest residual (no. 19) is not at all influential. On the other hand, deletion of observation no. 18, which has a small residual, has a marked effect on the parameter estimates.

Cook (1977) introduced a statistic to indicate the influence of an observation with respect to a particular model. For a single observation, Cook also showed that the statistic contained information on whether the observation was also an outlier. Andrews and Pregibon (1978) proposed a statistic to identify one or more observations as either outliers, influential or both.

These statistics are examined further in this paper. It is shown that Cook's statistic together with two components of the Andrews and Pregibon statistic will provide considerable information not only on outlying and influential observations but also on the remoteness of observations in the factor space.

2. OUTLIER SUM OF SQUARES

Following the same notation as in John and Draper (1978), the basic regression model for n observations and p parameters is

$$E(\underset{\sim}{y}) = E\begin{pmatrix} \underset{\sim}{y}_1 \\ \underset{\sim}{y}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \end{pmatrix}\underset{\sim}{\beta}. \qquad (1)$$

The observations are divided into the K observations ($\underset{\sim}{y}_2$) which are being inspected as possible outliers or influential observations and the n-K observations ($\underset{\sim}{y}_1$) which are not. Natuarally, some rearrangement of rows may be needed to achieve the division in (1). The residuals from fitting this model are, using standard least squares analysis, given by

$$\underset{\sim}{r} = \begin{pmatrix} \underset{\sim}{r}_1 \\ \underset{\sim}{r}_2 \end{pmatrix} = (\underset{\sim}{I} - \underset{\sim}{R})\underset{\sim}{y} = \begin{pmatrix} \underset{\sim}{I} - \underset{\sim}{R}_{11} & -\underset{\sim}{R}_{12} \\ -\underset{\sim}{R}_{21} & \underset{\sim}{I} - \underset{\sim}{R}_{22} \end{pmatrix} \begin{pmatrix} \underset{\sim}{y}_1 \\ \underset{\sim}{y}_2 \end{pmatrix} \tag{2}$$

where

$$\underset{\sim}{R}_{ij} = \underset{\sim}{X}_i(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}_j' \tag{3}$$

is a submatrix of $\underset{\sim}{R} = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'$.

Deleting the suspect $\underset{\sim}{y}_2$ observations gives the model $E(\underset{\sim}{y}_1) = \underset{\sim}{X}_1\underset{\sim}{\beta}$. Alternatively, the model

$$E\begin{pmatrix} \underset{\sim}{y}_1 \\ \underset{\sim}{y}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{X}_1 & \underset{\sim}{0} \\ \underset{\sim}{X}_2 & \underset{\sim}{I} \end{pmatrix} \begin{pmatrix} \underset{\sim}{\beta} \\ \underset{\sim}{\gamma} \end{pmatrix} . \tag{4}$$

where $\underset{\sim}{\gamma}$ is a $K \times 1$ vector of additional parameters, can be used; see Draper (1961). The resulting estimators $\underset{\sim}{b}$ and $\underset{\sim}{c}$ of $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$ respectively are

$$\underset{\sim}{b} = (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{y} \tag{5}$$

and

$$\underset{\sim}{c} = (\underset{\sim}{I} - \underset{\sim}{R}_{22})^{-1}\underset{\sim}{r}_2 \quad . \tag{6}$$

Replacing $\underset{\sim}{y}_2$ by "missing value" estimates $\underset{\sim}{y}_2 - \underset{\sim}{c}$ in (1) and refitting the model (1) gives new residuals $\underset{\sim}{u}$ whose components are

$$\underset{\sim}{u}_1 = (\underset{\sim}{I} - \underset{\sim}{R}_{11} - \underset{\sim}{R}_{12}\,(\underset{\sim}{I}-\underset{\sim}{R}_{22})^{-1}\underset{\sim}{R}_{21})\underset{\sim}{y}_1$$

$$\underset{\sim}{u}_2 = \underset{\sim}{0} \qquad (7)$$

where the dimensions of $\underset{\sim}{u}_i$ are those of $\underset{\sim}{y}_i$ in (1). The above procedure of adjusting the vector $\underset{\sim}{u}_2$ necessitates that $\underset{\sim}{u}_2 = \underset{\sim}{0}$, whilst $\underset{\sim}{u}_1$ are also the residuals from fitting $E(\underset{\sim}{y}_1) = \underset{\sim}{X}_1\underset{\sim}{\beta}$. The u_i are the "revised residuals" of Gentleman and Wilk (1975, pp. 391, 394).

The extra sum of squares due to fitting $\underset{\sim}{\gamma}$ in model (4), as compared with model (1), is given by

$$Q_K = \underset{\sim}{r}_2'(\underset{\sim}{I} - \underset{\sim}{R}_{22})^{-1}\underset{\sim}{r}_2 \,. \qquad (8)$$

It measures the effect of outliers and can be used to form a test statistic as described by Gentleman and Wilk (1975) and John and Draper (1978).

John and Draper (1978) also show that Q_K is the sum of squares of K successive adjusted normalized uncorrelated residuals. This means that Q_K can be expressed as

$$Q_K = Q_{K-1} + u_K^2/V(u_K) \qquad (9)$$

where Q_{K-1} is the outlier sum of squares obtained from an analysis when K-1 of the observations are deleted, u_K is the residual corresponding to the Kth observation from this analysis and $V(u_K)\sigma^2$ is the variance of u_K.

3. ANDREWS-PREGIBON STATISTIC

The Andrews-Pregibon (1978) statistic, called hereafter the AP statistic, is based on matrices of independent variables with the y vector appended. For model (1), this matrix is

$$\underset{\sim}{X}_1^* = (\underset{\sim}{X} \vdots \underset{\sim}{y}) \tag{10}$$

and for model (4) it is

$$\underset{\sim}{X}_2^* = (\underset{\sim}{X} \vdots \underset{\sim}{D} \vdots \underset{\sim}{y}) \tag{11}$$

where

$$\underset{\sim}{D} = \begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{I} \end{pmatrix}.$$

The AP statistic is then defined as

$$R^{(K)}_{ij\ldots} = |\underset{\sim}{X}_2^{*\prime}\underset{\sim}{X}_2|/|\underset{\sim}{X}_1^{*\prime}\underset{\sim}{X}_1^*| \tag{12}$$

where ij... denote the K subscripts selected to form $\underset{\sim}{y}_2$. In the appendix it is established that

$$R^{(K)}_{ij\ldots} = (1 - Q_K/RSS).\ |\underset{\sim}{I} - \underset{\sim}{R}_{22}| \tag{13}$$

where RSS is the residual sum of squares obtained from fitting the full model (1), Q_K is given by (8) and $\underset{\sim}{R}_{22}$ is defined in (3).

Andrews and Pregibon (1978, p. 88) say concerning their statistic that the quantity $1 - \{R^{(K)}_{ij\ldots}\}^{1/2}$ "corresponds to the proportion of volume generated by $\underset{\sim}{X}^*_i$ attributable to the K observations (ij...). If this subset of observations lies 'far out' in the factor space, it will account for a large proportion of the volume of the space, lending some realistic interpretation to the term 'outliers'. Hence, small values of $R^{(K)}_{ij\ldots}$ are associated with deviant and/or influential observations. Regardless of which is actually the case, it is desirable to isolate subsets of the observations producing small $R^{(K)}_{ij\ldots}$ for further scouting".

The factorization in (13) indicates that two distinct quantities can be examined. The first factor will be small if Q_K is large and so identifies sets of outliers as in John and Draper (1978). The second term $|\underset{\sim}{I} - \underset{\sim}{R}_{22}|$ only involves the independent variables and, as will be shown later, provides a measure of the remoteness of the set of observations in the factor space.

4. COOK'S STATISTIC

Cook's (1977) statistic is basically, and in general,

$$C_{ij\ldots} = (\underset{\sim}{b} - \underset{\sim}{b}^*)'\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{b} - \underset{\sim}{b}^*)/ps^2 \qquad (14)$$

where $\underset{\sim}{b}$ is the least squares estimate (lse) of $\underset{\sim}{\beta}$ in (1), $\underset{\sim}{b}^*$ is the lse of

$\underset{\sim}{\beta}$ in (4), $s^2 = RSS/(n-p)$ and ij... denote, as before, the K subscripts selected to form $\underset{\sim}{y}_2$.

For K = 1, Cook has pointed out that the statistic can be written as

$$C_i = p^{-1} t_i^2 \cdot \{r_{ii}/(1-r_{ii})\} \tag{15}$$

where $t_i = r_i/\{s^2(1-r_{ii})\}^{1/2}$ is the ith standardized residual and r_{ij} is the (ij)th element of $\underset{\sim}{R}$. Thus t_i^2 is an outlier measure since it is a monotonic function of Q_1 in (8). The ratio $r_{ii}/(1-r_{ii})$ measures the influence of the ith data point in the sense that, to quote Cook (1977, p. 16), "A large value of this ratio indicates that the associated point has a heavy weight in the determination of $\hat{\underset{\sim}{\beta}}$ [i.e. $\underset{\sim}{b}$]. The two individual measures combine ... to produce a measure of the overall impact any single point has on the least squares solution". (For additional discussion, see Cook, 1979.)

Consider now a similar factorization for K > 1. Eliminating the additional parameters $\underset{\sim}{\gamma}$ from the normal equations for model (4) gives

$$\underset{\sim}{b}^* = \underset{\sim}{b} - (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim 2}{X}'\underset{\sim}{c} \tag{16}$$

where $\underset{\sim}{c}$ is given by (6). Hence

$$C_{ij\ldots} = \underset{\sim}{c}'R_{22}\underset{\sim}{c}/ps^2$$

Because

$$\underset{\sim}{c}'R_{22}\underset{\sim}{c} = \underset{\sim}{c}'\underset{\sim}{c} - Q_K,$$

it follows that Cook's statistic can be written as

$$C_{ij\ldots} = \frac{Q_K}{ps^2} \cdot \left(\frac{c'c}{Q_K} - 1\right) \tag{17}$$

The first component is an outlier measure as before. For K = 1, the ratio $\underset{\sim}{c}'\underset{\sim}{c}/Q_K$ can be written as $(1-r_{ii})^{-1}$ which is large for an influential observation since r_{ii} is then close to 1. For K > 1 this ratio can be expressed as

$$\frac{\underset{\sim}{c}'\underset{\sim}{c}}{Q_K} = \frac{\underset{\sim}{r}_2'(\underset{\sim}{I}-\underset{\sim}{R}_{22})^{-2}\underset{\sim}{r}_2}{\underset{\sim}{r}_2'(\underset{\sim}{I}-\underset{\sim}{R}_{22})^{-1}\underset{\sim}{r}_2} \tag{18}$$

It is difficult to see that any meaningful interpretation can be made of (18) in terms of the influence of a set of points. For K = 1, r_2 is a scalar but, in general, the residuals do not cancel out. Hence, although a useful factorization of Cook's statistic is possible for K = 1, the somewhat contrived factorization given in (17) seems to be of little value for K > 1. The overall statistic, however, remains a useful measure of the influence of a set of observations.

5. EXAMPLES

In the previous sections, three statistics have been considered, namely the outlier sum of squares Q_K, the Andrews-Pregibon statistic $R^{(K)}_{ij...}$ and the Cook statistic $C_{ij...}$ given by (8), (13) and (14) respectively. $R^{(K)}_{ij...}$ has also been factorized into two components.

These statistics will now be obtained from an analysis of a set of 21 observations (x,y) given by Mickey et. al (1967). These data have also been used by Andrews and Pregibon (1978). A straight line regression model was fitted to the full set of data and then to the 20 data points remaining when each observation was deleted in turn. For the computations, it was found to be simpler to use model (4) with additional dummy independent variables rather than to delete observations. The results of the analyses are given in Table 1, where the figures have been rounded to give a maximum of three figures

From Table 1, it is clear that observations 18 and 19 stand out. The values of Q_1 and the first component of the AP statistics indicate that observation 19 is an outlier whilst from the second component of AP and from C_i, observation 18 is shown to be influential. Andrews and Pregibon concluded that, since $R^{(1)}_{18} < R^{(1)}_{19}$, observation 18 is the more important point.

Although in this example there is close agreement between the second component in the AP statistic and C_i they are, in fact, different measures of influence. In general, the conclusions drawn from examing Cook's statistic and the AP statistic could well be different. To illustrate this

Table 1. Q_K, AP and Cook's statistics for single observations (original data).

Observation deleted	Q_1	$100(1-Q_1/RSS)$	$100(1-r_{ii})$	$100R_i^{(1)}$	$100C_i$
1	4	100	95	95	0
2	108	96	85	81	8
3	260	89	94	83	7
4	82	97	93	90	3
5	86	97	95	92	2
6	0	100	93	93	0
7	12	100	94	94	0
8	7	100	94	94	0
9	11	100	92	92	0
10	48	98	93	91	2
11	133	95	91	86	5
12	15	99	93	92	0
13	260	89	94	83	7
14	193	91	94	86	5
15	22	99	94	93	0
16	2	100	94	94	0
17	79	97	95	92	2
18	88	95	35	33	68
19	969	58	95	55	22
20	140	94	94	89	3
21	2	100	94	94	0

on the Mickey et. al. data, suppose the x value on observation 18 is changed from its original value of 42 to 62.43. This new point, called number 18*, has been chosen to lie on the least squares regression line fitted to the other 20 data points. Thus the addition of 18* will not change the fitted equation. The recalculated statistics, for a few deleted observations, are given in Table 2.

Table 2. Q_K, AP and Cook's statistics for single observations (amended data).

Observation deleted	Q_1	$100(1-Q_1/RSS)$	$100(1-r_{ii})$	$100R_i^{(1)}$	$1000C_i$
2	227	90	91	82	94
3	234	89	94	84	61
11	177	92	93	85	59
18*	0	100	16	16	0
19	861	61	95	58	188

The AP statistic picks out observation 18* and the factorization designates it as influential. However, it is clearly not influential in estimating the parameters as Cook's statistic confirms. The second component in the AP statistic, and hence the AP statistic itself, is small simply because the point is a long way from other points. Observation 19 is still very much an outlier, as shown by Q_1.

The same type of effect can be seen if a second observation is deleted. Table 3 gives the same statistics as before for the analyses of the original data with certain pairs of observations deleted; values for other pairs were calculated but are not shown here because they are of less interest.

Table 3. Q_K, AP and Cook's statistics for certain pairs of observations (original data).

Observations deleted	Q_2	$100(1-Q_2/RSS)$	$100\lvert \underset{\sim}{I}-\underset{\sim}{R}_{22}\rvert$	$100R_{ij}^{(2)}$	$100C_{ij}$
18, 2	442	81	20	16	637
18, 3	324	86	32	28	48
18, 11	277	88	30	27	152
18, 19	983	57	32	18	15
19, 2	1031	55	80	44	10
19, 3	1189	48	89	43	12
19, 11	1128	51	86	44	41

The AP statistic overall draws attention to the pair (2, 18) rather than (18, 19) by a small margin, while the factorization clarifies the position. For (2, 18) the second term is small indicating that the pair is influential, whilst for (18, 19) the first term is small indicating outliers. Of course, 19 on its own has already been shown to be an outlying observation so that any other observation paired with it must lead to a large Q_2, as (9) shows. Observations (3, 19) are the most outlying pair but most of the

Q_2 value is due to 19 alone. On the basis of Cook's statistic the pair (2, 18) is clearly the most influential in terms of affecting parameter estimates. The overall AP statistic does not, therefore, necessarily draw attention to points which are outliers or influential in terms of parameter estimation. As the second term in the factorization shows, it gives considerable weight to sets of points which can be regarded as remote in the factor space. This can again be illustrated by moving points 2 and 18 further out in the space so that they lie on the least squares line fitted from the other 19 points. Hence, Cook's statistic is zero when both points are deleted, but the AP statistic picks out these points as the most important pair of points. However, they are not outliers, nor are they influential in estimating parameters; they are simply remote in the predictor-response space (x,y).

6. RECOMMENDATIONS

Of the factors discussed here, which should be printed out and examined as part of a general regression routine? We recommend Q_K, Cook's statistic and the second factor $|\underset{\sim}{I} - \underset{\sim}{R}_{22}|$ of the AP statistic, for the following reasons.

1. Q_K in (8) provides a measure for outliers, large values being considered deviant. In certain circumstances, a test of significance for this can be made; see, for example, John and Draper (1978).

2. The form (14) of Cook's statistic ensures that it will be sensitive to changes in the fitted model, if observations are omitted. Thus Cook's

statistic will reveal which observations are influential in the sense that they affect the fitted equation's coefficients.

3. The second factor $|\underset{\sim}{I} - \underset{\sim}{R}_{22}|$ of the AP statistic is a spatial measure, and shows which observations are "influential" in the sense that they are isolated from the bulk of the data in the (p+1) dimensional space defined by the columns of $\underset{\sim}{X}$ and $\underset{\sim}{y}$. Note that such observations may or may not be influential in the sense described in the preceding paragraph.

These three measures contain all the basic information in the Cook and AP statistics. Note that, for K = 1, they correspond, apart from factors, to the t_i^2, D_i and c_i^2 of Cook (1977), and so achieve, for all K, the desirable features mentioned by Cook (1977, p. 349) for the K = 1 case in his reply to a letter by Obenchain (1977).

ACKNOWLEDGEMENTS

N.R. Draper was partially supported by the National Science Foundation under Grant MCS76-83899 and by the Wisconsin Alumni Research Foundation through the University of Wisconsin Graduate School.

REFERENCES

Andrews, D.F. and Pregibon, D. (1978). Finding the outliers that matter. J. Roy. Statist. Soc., B, 40, 85-93.

Cook, R.D. (1977). Detection of influential observations in linear regression. Technometrics, 19, 15-18. Additional correspondence, 348-350.

Cook, R.D. (1979). Influential observations in linear regression. J. Amer. Statist. Assoc., 74, 169-174.

Draper, N.R. (1961). Missing values in response surface designs. Technometrics, 3, 389-398.

Gentleman, J.F. and Wilk, M.B. (1975). Detecting outliers II. Supplementing the direct analysis of residuals. Biometrics, 31, 387-410.

John, J.A. and Draper, N.R. (1978). On testing for two outliers or one outlier in two-way tables. Technometrics, 20, 69-78.

Mickey, M.R., Dunn, O.J. and Clark, V. (1967). Note on the use of stepwise regression in detecting outliers. Computers and Biomedical Research, 1, 105-111.

Obenchain, R.L. (1977). Letter to the editor. Technometrics, 19, 348-349.

Rao, C.R. (1973). Linear Statistical Inference and its Applications. Second edition. New York: Wiley.

APPENDIX

From (11) it follows that

$$|\underset{\sim}{X}_2^{*\prime}\underset{\sim}{X}_2^{*}| = \begin{vmatrix} \underset{\sim}{X}'\underset{\sim}{X} & \underset{\sim}{X}_2' & \underset{\sim}{X}'\underset{\sim}{y} \\ \underset{\sim}{X}_2 & \underset{\sim}{I} & \underset{\sim}{y}_2 \\ \underset{\sim}{y}'\underset{\sim}{X} & \underset{\sim}{y}_2' & \underset{\sim}{y}'\underset{\sim}{y} \end{vmatrix}$$

$$= |\underset{\sim}{X}'\underset{\sim}{X}| \cdot \left| \begin{pmatrix} \underset{\sim}{I} & \underset{\sim}{y}_2 \\ \underset{\sim}{y}_2' & \underset{\sim}{y}'\underset{\sim}{y} \end{pmatrix} - \begin{pmatrix} \underset{\sim}{X}_2 \\ \underset{\sim}{y}'\underset{\sim}{X} \end{pmatrix} (\underset{\sim}{X}'\underset{\sim}{X})^{-1} (\underset{\sim}{X}_2', \underset{\sim}{X}'\underset{\sim}{y}) \right|$$

applying the well known expansion result for determinants (see, for example, Rao 1973, p. 32). Rearranging and repeating the expansion gives

$$|\underset{\sim}{X}_2^{*\prime}\underset{\sim}{X}_2^{*}| = |\underset{\sim}{X}'\underset{\sim}{X}| \cdot \begin{vmatrix} \underset{\sim}{I} - \underset{\sim}{R}_{22} & \underset{\sim}{r}_2 \\ \underset{\sim}{r}_2' & \text{RSS} \end{vmatrix}$$

$$= |\underset{\sim}{X}'\underset{\sim}{X}| \cdot |\underset{\sim}{I} - \underset{\sim}{R}_{22}| \cdot \{\text{RSS} - \underset{\sim}{r}_2'(\underset{\sim}{I} - \underset{\sim}{R}_{22})^{-1}\underset{\sim}{r}_2\}$$

$$= |\underset{\sim}{X}'\underset{\sim}{X}| \cdot \text{RSS} \cdot |\underset{\sim}{I} - \underset{\sim}{R}_{22}| \cdot (1 - Q_K/\text{RSS}).$$

A similar treatment provides

$$|\underset{\sim}{X}_1^{*\prime}\underset{\sim}{X}_1| = |\underset{\sim}{X}'\underset{\sim}{X}| \cdot \text{RSS}$$

so that the ratio is

$$R^{(K)}_{ij\ldots} = (1 - Q_K/\text{RSS}) \cdot |\underset{\sim}{I} - \underset{\sim}{R}_{22}|.$$