Detection of Influential Observation in Linear Regression

R. Dennis Cook

Department of Applied Statistics
University of Minnesota
St. Paul, Minnesota 55108

A new measure based on confidence ellipsoids is developed for judging the contribution of each data point to the determination of the least squares estimate of the parameter vector in full rank linear regression models. It is shown that the measure combines information from the studentized residuals and the variances of the residuals and predicted values. Two examples are presented.

KEY WORDS

Influential observations
Confidence ellipsoids
Variances of residuals
Outliers

1. INTRODUCTION

It is perhaps a universally held opinion that the overall summary statistics (e.g., R^2, $\hat{\beta}$) arising from data analyses based on full rank linear regression models can present a distorted and misleading picture. This has led to the recommendation and use of a number of procedures that can isolate peculiarities in the data; plots of the residuals (R_i) and examination of standardized residuals are probably the two most widely used. The studentized residuals, t_i, (i.e. the residual divided by its standard error) have been recommended (see, e.g., [2], [4], [6]) as more appropriate than the standardized residuals (i.e., the residual divided by the square root of the mean square for error) for detecting outliers. Also, approximate critical values for the maximum absolute studentized residual are available [8].

Behnken and Draper [2] have illustrated that the estimated variances of the predicted values (or, equivalently, the estimated variances of the residuals, $\hat{V}(R_i)$) contain relevant information beyond that furnished by residual plots or studentized residuals. Specifically, they state "A wide variation in the [variance of the residuals] reflects a peculiarity of the $\mathbf{X}$ matrix, namely a nonhomogeneous spacing of the observations and will thus often direct attention to data deficiencies." The opinion that these variances contain additional information was also put forth by Huber [6] and Davies and Hutton [4]. Box and Draper [3] developed a robust design criterion based on the sums of squares of the variances of the predicted values.

If a potentially critical observation has been detected using one or more of the above measures then the examination of the effects of deleting the observation seems a natural next step. However, the problem of determining which point(s) to delete can be very perplexing, especially in large data sets, because each point now has two associated measures (t_i, $\hat{V}(R_i)$) which must be judged simultaneously. For example, assuming the mean square for error to be 1.0, which point from the set $(t_i, \hat{V}(R_i))$ = (1, .1), (1.732, .25), (3, .5), (5.196, .75) is most likely to be critical?

It is the purpose of this note to suggest an easily interpretable measure that combines information from both t_i and $V(R_i)$, and that will naturally isolate "critical" values.

2. DEVELOPMENT

Consider the model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times p$ full rank matrix of known constants, β is an $n \times p$ vector of unknown parameters and $\mathbf{e}$ is an $n \times 1$ vector of randomly distributed errors such that $E(\mathbf{e}) = \mathbf{O}$ and $V(\mathbf{e}) = \mathbf{I}\sigma^2$. Recall that the least squares estimate of β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The corresponding residual vector is

$$\mathbf{R} = (R_i) = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$$

The covariance matrices of $\hat{\mathbf{Y}}$ and $\mathbf{R}$ are, respectively,

$$V(\hat{\mathbf{Y}}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 \qquad (1)$$

Received December 1975; revised September 1976

and

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

Finally, the normal theory $(1-\alpha) \times 100\%$ confidence ellipsoid for the unknown vector, β, is given by the set of all vectors β^*, say, that satisfy

$$\frac{(\beta^* - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta^* - \hat{\beta})}{ps^2} \leq F(p, n-p, 1-\alpha) \quad (3)$$

where $s^2 = \mathbf{R}'\mathbf{R}/(n-p)$ and $F(p, n-p, 1-\alpha)$ is the $1-\alpha$ probability point of the central F-distribution with p and $n-p$ degrees of freedom.

To determine the degree of influence the ith data point has on the estimate, $\hat{\beta}$, a natural first step would be to compute the least squares estimate of β with the point deleted. Accordingly, let $\hat{\beta}_{(-i)}$ denote the least squares estimate of β with the ith point deleted. An easily interpretable measure of the distance of $\hat{\beta}_{(-i)}$ from $\hat{\beta}$ is (3). Thus, the suggested measure of the critical nature of each data point is now defined to be

$$D_i \equiv \frac{(\hat{\beta}_{(-i)} - \hat{\beta})'\mathbf{X}'\mathbf{X}(\hat{\beta}_{(-i)} - \hat{\beta})}{ps^2} \qquad i = 1, 2, \cdots n \quad (4)$$

This provides a measure of the distance between $\hat{\beta}_{(-i)}$ and $\hat{\beta}$ in terms of descriptive levels of significance. Suppose, for example, that $D_i \approx F(p, n-p, .5)$, then the removal of the ith data point moves the least squares estimate to the edge of the 50% confidence region for β based on $\hat{\beta}$. Such a situation may be cause for concern. For an uncomplicated analysis one would like each $\hat{\beta}_{(-i)}$ to stay well within a 10%, say, confidence region.

On the surface it might seem that any desirability this measure has would be overshadowed by the computations necessary for the determination of $n+1$ regressions. However, it is easily shown that (see [1])

$$\hat{\beta} - \hat{\beta}_{(-i)} = (\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i[Y_i - \mathbf{x}_i'\hat{\beta}] \quad (5)$$

where $\mathbf{X}_{(-i)}$ is obtained by removing the ith row, $\mathbf{x}_i'$, from $\mathbf{X}$ and Y_i is the ith observation. Also, letting $v_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ and assuming $v_i < 1$, as well as be the case if $\mathbf{X}_{(-i)}$ has full rank p,

$$(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}/(1-v_i),$$

from which it follows that

$$\frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1-v_i} = (\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i \quad (6)$$

Substitution of (6) into (5) yields

$$\hat{\beta} - \hat{\beta}_{(-i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1-v_i}[Y_i - \mathbf{x}_i'\hat{\beta}]$$

It follows immediately that

$$D_i = \left[\frac{Y_i - \mathbf{x}_i'\hat{\beta}}{s\sqrt{1-v_i}}\right]^2 \frac{v_i}{p(1-v_i)}. \quad (7)$$

Note that D_i depends on three relevant quantities all relating to the full data set: The number of parameters, p, the ith studentized residual,

$$t_i = \left[\frac{Y_i - \mathbf{x}_i'\hat{\beta}}{s\sqrt{1-v_i}}\right]$$

and the ratio of the variance of the ith predicted value, $V(\hat{Y}_i) = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\sigma^2 = v_i\sigma^2$ (see, equation 1), to the variance of the ith residual, $V(R_i) = \sigma^2(1-v_i)$ (see, equation 2). Thus D_i can be written as simply

$$D_i = \frac{t_i^2}{p}\frac{V(\hat{Y}_i)}{V(R_i)} \quad (8)$$

Clearly, t_i^2 is a measure of the degree to which the ith observation can be considered as an outlier from the assumed model. In addition, it is easily demonstrated that if the possible presence of a single outlier is modeled by adding a parameter vector $\theta' = (0, 0, \ldots, 0, \theta, 0, \ldots 0)$ (both θ and its position within θ' are unknown) to the model, then max t_i^2 is a monotonic function of the normal theory likelihood ratio test of the hypothesis that $\theta = 0$.

The ratios $V(\hat{Y}_i)/V(R_i)$ measure the relative sensitivity of the estimate, $\hat{\beta}$, to potential outlying values at each data point. They are also monotonic functions of the v_i's which are the quantities Box and Draper [3] used in the development of their robust design (i.e., insensitive to outliers) criterion. A large value of the ratio indicates that the associated point has heavy weight in the determination of $\hat{\beta}$. The two individual measures combine in (8) to produce a measure of the overall impact any single point has on the least squares solution.

A little care must be exercised when using D_i since $\hat{\beta}_{(-i)}$ is essentially undefined in extreme cases when $V(R_i) = 0$ (see the lemma in [1]); i.e., when $\mathbf{X}_{(-i)}$ has rank less than p.

Returning to the example given in the Introduction we see that each point has an equal overall impact on the regression since in each case $pD_i = 9.0$. To continue the example, suppose $p = 3$ and $n - p = 24$, then $D_i = 3.0$ and the removal of any of the four points would move the estimate of β to the edge of the 95% confidence region about $\hat{\beta}$. However, inspection of the individual components shows that the reasons for the extreme movement are different. The two points (3, .5) and (5.196, .75) could be rejected as containing outliers in the dependent variable while the remaining two could not. Inspection of $\mathbf{X}$ would be necessary to determine why the remaining points are important. It may be, for example, that they

correspond to outlying values in the independent variables.

In any analysis additional information may be gained by examining t_i and $V(\hat{Y}_i)/V(R_i)$ separately. A three column output of t_i, $V(\hat{Y}_i)/V(R_i)$ and D_i would seem to be a highly desirable option in any multiple regression program.

The following two examples should serve as additional demonstrations of the use of D_i. No attempt at a "complete" analysis is made.

3. EXAMPLES

Example 1—Longley Data

Longley [7] presented a data set relating six economic variables to total derived employment for the years 1947 to 1962. Table 1 lists the residuals standardized by s, t_i, $V(\hat{Y}_i)/V(R_i)$, D_i and the year. Notice first that there are considerable differences between R_i/s and t_i. Second, the point with the largest D_i value corresponds to 1951. Removal of this point will move the least squares estimate to the edge of a 35% confidence region around $\hat{\beta}$. The second largest D_i value is at 1962 and its removal will move the estimate of β to approximately the edge of a 15% confidence region. Clearly, 1951 and 1962 have the greatest impact on the determination of $\hat{\beta}$. The point with the largest studentized residual is 1956; however, the effect of this point on $\hat{\beta}$ is not important relative to the effects of 1951 and 1962. The identification of the points with max $|t_i|$ and max $V(\hat{Y}_i)/V(R_i)$ (or max v_i) would not have isolated 1951. (It is interesting to note that 1951 was the first full year of the Korean conflict).

Example 2—Hald Data

The data for this example were previously published by Hald (see [2] and p.165 of [5]). There are 4 regressors and 13 observation points. Table 2 lists R_i/s, t_i, $V(\hat{Y}_i)/V(R_i)$ and D_i. In contrast to the previous example the data here seem well behaved. Observation number 8 has the largest D_i value but its removal moves the least squares estimate to the edge of only the 10% confidence region for β.

4. EXTENSIONS

It is easily seen that D_i is invariant under changes of scale. If the scale of each variable is thought to be an important consideration it may be more desirable to compute the squared length of $(\hat{\beta}_{(-i)} - \hat{\beta})$. It is easily shown that

$$\frac{(\hat{\beta}_{(-i)} - \hat{\beta})'(\hat{\beta}_{(-i)} - \hat{\beta})}{ps^2} = \frac{t_i^2}{p} \frac{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-2}\mathbf{x}_i}{1 - v_i}$$

The proposed measure was developed under the implicit presumption that β is the parameter of interest. This may not always be the case. If interest is in q, say, linearly independent combinations of the elements of β, then it would be more reasonable to measure the influence each data point has on the determination of the least squares estimates of these combinations. Let $\mathbf{A}$ denote a qxp rank q matrix and let $\psi = \mathbf{A}\beta$ denote the combinations of interest. A generalized measure of the importance of the ith point is now defined to be

$$D_i(\mathbf{A}) = \frac{(\hat{\psi}_{(-i)} - \hat{\psi})'\mathbf{B}^{-1}(\hat{\psi}_{(-i)} - \hat{\psi})}{qs^2}$$

TABLE 1—*Longley Data*

Year	R_i/s	$\lvert t_i \rvert$	$V(\hat{Y}_i)/V(R_i)$	D_i
1947	0.88	1.15	0.74	0.14
48	-0.31	0.48	1.30	0.04
49	0.15	0.19	0.57	*
50	-1.34	1.70	0.59	0.24
51	1.02	1.64	1.60	0.61
52	-0.82	1.03	0.59	0.09
53	-0.54	0.75	0.97	0.08
54	-0.04	0.06	1.02	*
55	0.05	0.07	0.84	*
56	1.48	1.83	0.49	0.23
57	-0.06	0.07	0.56	*
58	-0.13	0.18	0.93	*
59	-0.51	0.64	0.60	0.04
60	-0.28	0.32	0.30	*
61	1.12	1.42	0.59	0.17
62	-0.68	1.21	2.21	0.47

*: smaller than 5×10^{-3}

TABLE 2—*Hald Data*

Observation	R_i/s	$\lvert t_i \rvert$	$V(\hat{Y}_i)/V(R_i)$	D_i
1	0.002	0.003	1.22	*
2	0.62	0.76	0.50	0.06
3	-0.68	1.05	1.36	0.30
4	-0.71	0.84	0.42	0.06
5	0.10	0.13	0.56	*
6	1.61	1.71	0.14	0.08
7	-0.59	0.74	0.58	0.06
8	-1.24	1.69	0.69	0.31
9	0.56	0.67	0.42	0.04
10	0.12	0.21	2.34	0.02
11	0.81	1.07	0.74	0.17
12	0.40	0.46	0.36	0.02
13	-0.94	1.12	0.44	0.11

*: smaller than 2×10^{-3}

where

$$\mathbf{B} = \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' \quad \text{and} \quad \hat{\psi}_{(-i)} = \mathbf{A}\hat{\beta}_{(-i)}.$$

Since

$$\hat{\psi} - \hat{\psi}_{(-i)} = \mathbf{A}\,(\hat{\beta} - \hat{\beta}_{(-i)})$$

it follows that

$$D_i(\mathbf{A}) = \frac{t_i^2}{q}\,\frac{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\mathbf{B}^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1-v_i}. \qquad (9)$$

To obtain the descriptive levels of significance the values of this generalized measure should, of course, be compared to the probability points of the central F-distribution with q and $n-p$ degrees of freedom.

The case when $q = 1$, i.e., when $\mathbf{A}$ is chosen to be a $1xp$ vector $\mathbf{z}'$, say, warrants special emphasis. From (9) it is easily seen that

$$D_i(\mathbf{z}') = pD_i\rho^2(\mathbf{x}_i'\hat{\beta}, \mathbf{z}'\hat{\beta}) \qquad (10)$$

where $D_i = D_i(\mathbf{I})$ and $\rho(\cdot\,,\,\cdot)$ denotes the correlation coefficient. If $\mathbf{z}'$ is a vector of values of the independent variables then $D_i(\mathbf{z}')$ measures the distance between the predicted mean value of y at $\mathbf{z}$ using the ith data point ($\mathbf{z}'\hat{\beta}$) and the predicted value at $\mathbf{z}$ without the ith point ($\mathbf{z}'\hat{\beta}_{(-i)}$). Note also that when $\mathbf{z}'$ is chosen to be a vector of the form $(0, \ldots, 1, 0, \ldots, 0)$, $D_i(\mathbf{z}')$ measures the distance between the corresponding components of $\hat{\beta}$ and $\hat{\beta}_{(-i)}$.

The maximum value of $D_i(\mathbf{z}')$ for a fixed i is obtained by choosing $\mathbf{z}' = \mathbf{x}_i'$,

$$D_i(\mathbf{z}') \le D_i(\mathbf{x}_i') = pD_i$$

for all $\mathbf{z}$. Thus, when prediction of mean values or the individual components of β are of interest it may not be necessary to use (10) directly: If $D_i(\mathbf{x}_i')$ shows a negligible difference between $\mathbf{x}_i\hat{\beta}$ and $\mathbf{x}_i'\hat{\beta}_{(-i)}$ then the difference between $\mathbf{z}'\hat{\beta}$ and $\mathbf{z}'\hat{\beta}_{(-i)}$ must also be negligible for all $\mathbf{z}$.

5. ACKNOWLEDGEMENT

The author would like to thank Professor C. Bingham for his suggestions and criticisms.

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*, 102–111.

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

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

[5] Draper, N. R. and Smith, H., (1966). *Applied Regression Analysis.* Wiley, New York.

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

[7] 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.

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