The Korean Communications
in Statistics Vol. 6, No. 1, 1999
pp. 25-31

A Reference Value for Cook's Measure

Jae June Lee[1]

Abstract

A single outlier can influence on the least squares estimators and can invalidate analysis based on these estimators. The Cook's statistic has been introduced to measure influence of individual data point on parameter estimation and the quantile of the F distribution is recommended as a reference value. But, in practice, subjective judgement is applied in the choice of appropriate quantile. A simple reference value is introduced in this paper, which is developed by approximating conditional quantities of Cook's measure. The performance of the proposed criterion is evaluated through analysis of real data set.

1. Introduction

We consider the multiple linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{y}$ is an n-vector of response, $\boldsymbol{X}$ is an n by $p'(=p+1)$ full rank matrix of p' independent variables possibly including one constant predictor, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$ is a p'-vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is an n-vector of errors with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $Var[\boldsymbol{\varepsilon}] = \sigma^2 I$.

It is well known that a single outlier can influence on the least squares estimators and can invalidate analysis based on these estimators. Several outlier diagnostics have been introduced in the past [Belsly et al. (1980), Cook and Weisberg (1982), and Chatterjee and Hadi (1986)]. Cook's measure (Cook (1977a)) has been developed to measure influence of individual data points on parameter estimation in the least squares regression problem and is commonly used to determine suspecious cases. A case is declared to be influential if the statistic is large compared with some reference value. Though the cutoff should not be used as a test of significance (Obenchain (1977)), Cook (1977a) suggested that this measure be compared with the quantiles of the central F distribution with p' and $n-p'$ degrees of freedom. For example, the 0.50 quantile of the $F(p', n-p')$ distribution [(Cook (1977a,b), Weisberg (1985), and Leger and Altman (1993)] is taken as a reference value because it can be interpreted that deletion of the i-th case would move the estimate of $\boldsymbol{\beta}$ to the edge of a 50% confidence

[1] Associate Professor, Department of Statistics, Inha University, Inchon 402-751, Korea

ellipsoid relative to $\hat{\boldsymbol{\beta}}$. Rather than using fixed quantile, one may investigate a few cases that appear to be most influential (i.e., having largest values). Sometimes, this approach leads to cutoffs as low as lower 0.10 quantile as shown in the literature (Chatterjee and Hadi (1986)). Weisberg (1985, pp 120) suggests the value 1.0, which is the limiting value of the F statistic as n and p become large. In practice, subjective judgement, in some degree, has been applied to determine the reference value.

In this paper, we propose a simple cutoff criterion which is based on the approximate conditional expected value and standard deviation of Cook's statistics. The performance of the proposed criterion is evaluated through analysis of real data sets.

2. A Simple Cutoff Criterion for Cook's Distance

In this paper, matrices and column vectors are denoted by boldface uppercase and lowercase letters, respectively. Also, the subscipt notation (i) is used to indicate the deletion of the i -th observation, and the special character ^ above any quantity is used to mean an estimator based on the method of least squares. For example, $\hat{\boldsymbol{\beta}}_{(i)}$ is the least squares estimator of $\boldsymbol{\beta}$ when the i-th case is deleted.

2.1 Cook's Measure

Assuming that the ($p' + 1$)-random vector ($\boldsymbol{x}, y$) have a joint cdf F with

$$E_F\left[\begin{pmatrix}\boldsymbol{x}\\ y\end{pmatrix}(\boldsymbol{x}', y)\right] = \begin{pmatrix}\Sigma(F) & \gamma(F)\\ \gamma'(F) & \tau(F)\end{pmatrix}$$

and $\boldsymbol{\Sigma}(F)$ be nonsingular. The functional corresponding to the least squares estimator of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = T(F) = \boldsymbol{\Sigma}^{-1}(F)\,\boldsymbol{\gamma}(F)$$

Then, the influence function of $\boldsymbol{\beta}$, $\boldsymbol{IF}_{\boldsymbol{\beta}}$, is given by [Hinkley (1977), Hample et al. (1986), and Cook et al. (1982)]

$$\boldsymbol{IF}_{\boldsymbol{\beta}}(y, \boldsymbol{x}; F) = \boldsymbol{\Sigma}^{-1}(F)\,\boldsymbol{x}(y - \boldsymbol{x}'\ T(F)) \tag{2}$$

An influence measure based on $\boldsymbol{IF}_{\boldsymbol{\beta}}$ can be constructed by normalizing the vector $\boldsymbol{IF}_{\boldsymbol{\beta}}$ to form a norm which is location/scale invariant, and written as

$$D_i(\boldsymbol{M}, c) = \frac{(\boldsymbol{IF}_i)'\boldsymbol{M}(\boldsymbol{IF}_i)}{c} \tag{3}$$

for any appropriate choice of $\boldsymbol{M}$ and c. The Cook's distance (Cook (1977a)) is obtained by substituting $\boldsymbol{M} = \boldsymbol{X}'\boldsymbol{X}$, $c = (n-1)^2 p's^2$, and $\boldsymbol{IF}_i$ as an sample influence function of $\boldsymbol{\beta}$ in the equation (3) as,

$$C_i = \frac{r_i^2}{p'} \frac{h_{ii}}{1-h_{ii}} = \frac{h_{ii}}{(1-h_{ii})^2} \frac{e_i^2}{s^2 p'} \tag{4}$$

where s^2 is an unbiased estimator of σ^2, $p' = p+1$, $h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$ is the i-th diagonal element of hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, and

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}. \tag{5}$$

The quantity C_i is also written as the equation (6),

$$\begin{aligned} C_i &= (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'(X'X)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})/p's^2 \\ &= (\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)})'(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)})/p's^2 \end{aligned} \tag{6}$$

and thus can be interpreted as the scaled Euclidean distance between the two vectors of fitted values when the fitting is done by including or excluding the i-th observation. The observation is declared to be influential if the distance is large compared with some reference value.

2.2 A Reference Value

Cook (1977a) suggests that the value of C_i can be compared to a quantile of $F(p', n-p')$ distribution because the equation of C_i in (6) is similar to a confidence ellipsoid for $\boldsymbol{\beta}$ based on $\hat{\boldsymbol{\beta}}$ which is given by the set of all $\boldsymbol{\beta}$ such that

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\boldsymbol{X}'X)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/p's^2 \le F(1-\alpha; p', n-p') \quad .$$

Though, usually, the 50-percentile of the $F(p', n-p')$ distribution is recommended as a cutoff value [Cook (1977a), Weisberg (1981)], subjective judgement, in some degree, has been used in determining the reference value (i.e. appropriate percentile of the $F(p', n-p')$). It is obvious that C_i does not follow an F distribution (Cook (1977a)), and therefore the criterion should be used just as a rough guide to detect influential cases.

We consider another simple reference value which is based on the approximate conditional expected value and standard deviation of C_i. Assuming normality of ε, the $r_i^2/(n-p')$ follow a beta distribution with parameters 1/2 and $(n-p'-1)/2$ (Ellenberg (1973)). By conditioning h_{ii}'s, the conditional mean and variance of C_i, from equation (4), become

$$E(C_i|h_{ii}) = \frac{h_{ii}}{1-h_{ii}} \frac{n-p'}{p'} \frac{a}{b} = \frac{1}{p'} \frac{h_{ii}}{1-h_{ii}} \tag{7}$$

$$Var(C_i|h_{ii}) = (\frac{h_{ii}}{1-h_{ii}})^2 (\frac{n-p'}{p'})^2 \frac{c}{d},$$

where $a=1/2, b=1/2+(n-p'-1)/2$, $c=E(C_i|h_{ii})[1-E(C_i|h_{ii})]$ and $d=a+b+1$.

For a data set with no aberrant observation (i.e., under H_0 in point of testing hypothesis), we can expect h_{ii} as p'/n, the average of h_{ii}, when the sample size is n. Therefore, by substituting p'/n for h_{ii} in conditional expressions (7) and by approximating $n-p'+2 \approx n-p'-1$, we roughly get

$$E(C_i|h_{ii}=p'/n) \approx \frac{1}{n-p'} \tag{8}$$

$$Var(C_i|h_{ii}=p'/n) \approx \frac{2}{(n-p')^2} \quad .$$

Based on the conditional expressions in equations (8), a simple reference value for Cook's measure is proposed as follows:

identify the i-th observation as an influential case, if for some k

$$C_i \geq \frac{1}{n-p'} + k \times \frac{\sqrt{2}}{n-p'} \tag{9}$$

For the value of k in (9), some value between 2 and 3 is recommended from our experience. The performance of the proposed criterion will be studied in analysis of real data sets.

3. A Numerical Example

As an illustrative example, the data set of a laboratory experiment performed by Moore in 1975 is analyzed. The data set has been used by Weisberg (1981) to illustrate the contribution of the individual observations to the Mallow's C_p statistics. Chatterjee and Hadi (1986) also used these data to compare performance of various influence measures. The measured variables are one dependent variable, y, and 5 independent variables, $x_1, x_2, \cdots, x_5$, and the data set consists of 20 $(= n)$ cases. A model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5 + \varepsilon$ is fitted to the data, and the resulting diagnostic statistics, including Cook's measure, are given in Table (1).

Table 1 : Diagnostic Statistics from Moore's Data

case	r_i	h_{ii}	C_i	case	r_i	h_{ii}	C_i
1	2.64	0.337	0.589	11	0.75	0.225	0.027
2	-0.79	0.502	0.104	12	0.21	0.135	0.001
3	0.47	0.485	0.035	13	-0.16	0.095	0.000
4	-0.21	0.251	0.002	14	0.10	0.198	0.000
5	-1.04	0.284	0.072	15	-1.66	0.171	0.094
6	0.82	0.371	0.066	16	0.36	0.262	0.008
7	-1.42	0.153	0.060	17	0.98	0.918	1.779
8	-0.28	0.087	0.001	18	0.05	0.234	0.000
9	-0.05	0.364	0.000	19	-1.06	0.364	0.108
10	-0.46	0.159	0.007	20	1.89	0.406	0.406

By referring to to these summaries in Table (1), we note following facts:

i) case 17 has a high h_{ii} value, and cases 2, 3, 20 have relatively high h_{ii} values

ii) case 1 has large r_i value, and cases 7, 15, 20 have relatively large r_i values

iii) the 0.50, lower 0.10, and upper 0.05 quantiles of $F(6, 14)$ distributions are 0.9357, 0.3470, and 2.8480, respectively. Therefore, influential obervations detected by corresponding quantile reference values are, denoting the upper quantile probability as α, are shown as Table (2). Chatterjee and Hadi (1986) identified the cases 1, 17, 20 as influential cases, based on the reference value 0.3470 (α = 0.90).

Table 2 : Moore Data: influential cases

criterion (α)	cutoff value	influential cases
0.05	2.8480	none
0.10	2.2427	none
0.50	0.9537	17
0.90	0.3470	1,17,20
0.95	0.2527	1,17,20

iv) Based on the reference formula proposed in equation (9), the reference values corresponding to k = 2.0, 2.5, and 3.0 are 0.2734, 0.3239, and 0.3740, respectively. Therefore, the reference value with any k in interval $(2, 3)$ leads to cases 1, 17, and 20 being detected as influential observations. We note that the reference value with $k = 2.5$ (i.e. 0.3239) is quite close to 0.3470 which is the value with lower 10-percentile ($\alpha = 0.90$) of $F(6, 14)$ distribution.

4. Concluding Remarks

Cook (1977a) suggested quantiles of the central F distribution with p' and $n-p'$ degrees of freedom as reference values. Especially, the 0.50 quantile of the $F(p', n-p')$ distribution is recommended, because it can be interpreted that deletion of the i-th case would move the estimate of β to the edge of a 50% confidence ellipsoid relative to $\hat{\beta}$. But, sometimes the quantiles as low as lower 0.10 quantile are used in the analysis of real data sets, as shown in the literature (Chatterjee and Hadi(1986)). In practice, based on subjective judgement, a certain quantile is chosen as a reference value, which is appropriate for the data set to be analysed.

The proposed reference value introduced in this paper is developed by using approximate conditional quantities of Cook's measure. The reference value can be determined by relatively less subjectively, and the proposed criterion works reasonably well, as shown in the analysis of Moore's data.

Reference

[1] Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics*, New York: John Wiley.

[2] Chatterjee, S. and Hadi, A.S. (1986), Influential observations, high leverage points, and outliers in linear regression(with discussion), *Statistical Science,* 1, 379-416.

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

[4] Cook, R. D. (1977b), Letter to the editor, *Technometrics,* Vol. 19, 348.

[5] Cook, R. D. (1979), Influential observations in linear regression, *Journal of the American Statistical Association,* Vol. 74, 169-194.

[6] Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression,* New York : Chapman and Hall.

[7] Ellenberg, J. H. (1973), The joint distribution of the standardized least squares residuals from a general linear regression, *Journal of the American Statististical Association,* Vol. 68, 941-943.

[8] Hample, F. R. (1974), The influence curve and its role in robust estiamtion, *Journal of the American Statistical Association,* Vol. 69, 383-393.

[9] Hample, F. R., Ronchentti, E. M., Rousseeuw, P. J., and Stahel, W.A. (1986), *Robust Statistics,* Wiley, New York.

[10] Hinkley, D. V. (1977), Jackknifing in unbalanced situation, *Technometrics,* 19, 285-92.

[11] Leger, C. and Altman, N. (1993), Assesing influence in variable selection problems, *Journal of the American Statistical Association,* Vol. 88, 547-556.

[12] Obenchain, R. L. (1977), Letter to the editor, *Technometrics,* Vol. 19, 348-351.

[13] Weisberg, S. (1985), *Applied Linear Regression (2nd ed.),* New York : John Wiley.

[14] Weisberg, S. (1981), A statistic for allocating C_p to individual cases, *Technometrics,* Vol. 23, 27-31.