응용통계연구 제 6권 1호
1993년 3월 pp.67 - 78

A Study of A New Statistic for Detection of Outliers and/or Influential Observations in Regression Diagnostics[1)]

Eun Mee Kang[2)]

Abstract

A new diagnostic statistic for detecting outliers and influential observations in linear models is suggested and studied in this paper. The proposed statistic is a weighted sum of two measures; one is for detecting outliers and the other is for detecting influential observations. The merit of this statistic is that it is possible to distinguish outliers from influential observations. We have done some Monte-Carlo Simulation to find the probability distribution of this statistic.

1. Introduction for Diagnostic Measures

Recently a great number of research papers have been published on the area of outliers and influential observations for diagnostic purposes, and there still remain many unsolved problems.

It is known that observations of, in the opinion of the investigator, standing apart from the bulk of the data have been called "outliers". It is also known that observations are judged as "influential" if important features of the analysis are substantially altered when the observations are deleted. Note that Chatterjee and Hadi(1986) emphasize that the meaning of "influential" should be clarified. Here, "influential" means "influential on the estimate of $\underline{\beta}$".

---

1) This paper was supported in part by NON-DIRECTED RESEARCH FUND, KOREA RESEARCH FOUNDATION, 1990

2) Department of Statistics, Sungshin Women's University, Seoul, 136-742, Korea

A great deal of measures have been proposed to detect outliers and influential observations for regression models.

Suppose the linear regression model can be written as

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \tag{1.1}$$

where $\underline{y}$ is a $n \times 1$ vector of observations , X is a $n \times p$ full rank matrix of known constants, $\underline{\beta}$ is a $p \times 1$ vector of regression coefficients and $\underline{\varepsilon}$ is a $n \times 1$ vector of randomly distributed errors such that $E(\underline{\varepsilon}) = \underline{0}$ and $V(\underline{\varepsilon}) = I\sigma^2$. In fitting the model (1.1) by least squares , we ususally obtain the fitted or predicted value from $\hat{\underline{y}} = X\hat{\underline{\beta}}$ where $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$. From this, it is simple to see that

$$\hat{\underline{y}} = H\underline{y} \tag{1.2}$$

$$\underline{e} = \underline{y} - \hat{\underline{y}} = (I - H)\underline{y} \tag{1.3}$$

where $H = X(X'X)^{-1}X'$ is the hat matrix and $\underline{e}$ is the residual vector. Note that $\hat{\underline{y}}$ is the perpendicular projection of $\underline{y}$ into the subspace generated by columns of X. Since , H is symmetric and idempotent, the average of diagonal elements h_{ii} of the hat matrix is p/n. Thus we determine a high leverage point by looking at the diagonal elements of H and paying particular attention to any design point for which $h_{ii} > 2p/n$. We may say that if h_{ii} is large, the data point may be considered as influential.

For a measure of an outlier, many authors have suggested the standardized residuals r_i

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}} \tag{1.4}$$

where $s = \sqrt{\underline{e}'\underline{e}/(n-p)}$. However, s^2 tends to overestimate σ^2 when there exists an outlier. For such case, $s_{(i)}^2$ is a better choice as an estimate of σ^2, where

$s_{(i)}^2$ is the residual mean squares of $n-1$ observations after discarding the i-th possible outlier case. Then we obtain the studentized residual r_i^*,

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \tag{1.5}$$

which is t-distributed with $(n-p-1)$ degrees of freedom.

A useful measure of influence which is called the Cook's statistic

$$D_I = (\hat{\beta} - \hat{\beta}_{(I)})' X' X(\hat{\beta} - \hat{\beta}_{(I)})/ps^2 \tag{1.6}$$

is obtained by Cook (1977, 1979), where $\hat{\beta}_{(I)}$ is the least squares estimate of β obtained by deleting k rows and k observations indexed by I from X and y, respectively. If there is a single observation deleted, D_I is written as D_i. Cook suggests that if observed D_i is equal to or greater than $F(p, n-1 ; \alpha)$ where α is less than 0.5, then y_i may be significant as an influential observation.

Andrews and Pregibon (1978) suggest a statistic using the ratio

$$R_I = \frac{(n-p-k)s_{(I)}^2 |X_{(I)}' X_{(I)}|}{(n-p)s^2 |X' X|} \tag{1.7}$$

for identifying subsets of k influential cases where $X_{(I)}$ is obtained by deleting k rows indexed by I from X. Small values of R_I are associated with deviants and/or influential observations.

For multiple outlier case Gentleman and Wilk (1975b) suggest Q_k

$$Q_k = RSS_C - RSS_m \tag{1.8}$$

where k indicates k outliers, $RSSc$ is the residual sum of squares when the complete set of original data is used to fit the specified model, and $RSSm$ is the residual sum of squares when the extreme observations are regarded as missing.

We write the basic model

$$E(\underline{y}) = E\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \underline{\beta} \tag{1.9}$$

where X_1 is an $(n-k)\times p$ matrix which contains no outliers, and X_2 is a $k\times p$ matrix which contains k outliers. From Little(1985), Q_k can be expressed as

$$\begin{aligned} Q_k &= Q_{k_1} + Q_{k_2} \\ &= \underline{e}_2' \underline{e}_2 + \underline{e}_1' X_1 (X_1' X_1)^{-1} X_1' \underline{e}_1 \end{aligned} \tag{1.10}$$

where $Q_{k_1} = \underline{e}_2' \underline{e}_2$, $Q_{k_2} = \underline{e}_1' X_1 (X_1' X_1)^{-1} X_1' \underline{e}_1$, and

$$\begin{aligned} \underline{e} &= \begin{bmatrix} \underline{e}_1 \\ \underline{e}_2 \end{bmatrix} = (I-H)\underline{y} \\ &= \begin{bmatrix} I-H_{11} & -H_{12} \\ -H_{21} & I-H_{22} \end{bmatrix} \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} \end{aligned} \tag{1.11}$$

Here, $H_{ij} = X_i(X'X)^{-1}X_j'$. By LIttle(1985), it is not difficult to show that

$$Q_{k_2} = (\tilde{\underline{\beta}} - \hat{\underline{\beta}})' X_1' X_1 (\tilde{\underline{\beta}} - \hat{\underline{\beta}}) \tag{1.12}$$

where $\tilde{\underline{\beta}} = (X_1' X_1)^{-1} X_1' \underline{y}_1$. Note that Q_{k_2} with proper scaling factor fits in the general scheme of normed influence measure discussed by Cook and Weisberg (1982). Hence, if Q_{k_1} has large proportion in the largest Q_k, the k observations can be candidates for outliers, and if Q_{k_2} has large proportion, the k observations can be influential observations.

2. Proposition of a New Statistic

For regression models, the standardized residual r_i and the studentized residual r_i^* only serve to detect outliers, while such measures as h_{ii}, the Cook's statistic D_I are only used for detecting influential observations. Such measures as the Andrews-Pregibon's R_I and the Gentleman and Wilk's Q_k may detect the observations which are outliers and/or influential observations. However, in practice we want to know which observations are outliers and which observations

are influential. Statistics which can distinguish outliers from influential observations have not been suggested.

As mentioned in the previous section Q_k is decomposed into Q_{k_1} and Q_{k_2}, where Q_{k_1} mainly detects outliers and Q_{k_2} mainly detects influential observations. However, the magnitude of Q_{k_1} and Q_{k_2} heavily depends on the unit of observations. Hence, to make Q_{k_1} and Q_{k_2} scale invariant, under changes of scale and non-singular linear transformations of the rows of X, we need to divide them by some scaling factor. We propose the following statistic which is a weighted sum of Q_{k_1} and Q_{k_2} divided by some scaling factor

$$WQ_k = wQ_{k_1}/(s.f.) + (1-w)Q_{k_2}/(s.f.) \tag{2.1}$$

where w is the weight factor, i.e. $0 \le w \le 1$ and s.f. is a scaling factor.

Now we choose an appropriate scaling factor for detecting outliers and influential observations. The appropriate scaling factor we want to propose is $ks^2_{(I)}$. It is clear that $s^2/ks^2_{(I)}$ is scale free and $ks^2_{(I)}$ belongs to the most frequently used types of scaling factor in the normed influential measures of Cook and Weisberg(1982).

Note that when w = 0, WQ_k becomes $Q_{k_2}/ks^2_{(I)}$ which is similar to the Cook's D_I in the equation (1.6). When w = 1, WQ_k becomes $Q_{k_1}/ks^2_{(I)}$ which is the sum of squares of the largest k residuals divided by $ks^2_{(i)}$. Hence, $Q_{k_1}/ks^2_{(I)}$ can detect k outliers. When w = 0.5, WQ_k becomes $Q_k/2ks^2_{(I)}$ which behaves like the statistic Q_k(it detects the same points as Q_k). However, the most important point of this statistic is that as the weight changes from 0 to 1, it can show the influential observations at first and then gradually changes to outliers.

Next, it is of interest to compare the diagnostic measures with the proposed statistic. The value of Q_k consists of two parts, the outlier part and the influential part. However, Little(1985) shows that the value of Q_k seems to be dominated by the outlier part, Q_k is categorized as a measure of detecting outliers in Table 1.

According to Table 1 each of all the diagnostic statistics is some function of r_i, r_i^* and h_{ii}. And WQ_k can be represented in a similar form. When w = 0, WQ_k is $h_{ii}r_i^*$, when w = 0.5, WQ_k becomes $r_i^{*2}/2$ and when w = 1, WQ_k = $(1-h_{ii})r_i^{*2}$. Using the formula (A2.1) in appendix of Cook and Weisberg(1982), these relations are easily obtained.

Measures for detection of an outlier	Measures for detection of an influential observation	Proposed statistics
1. $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$	1. h_{ii} = leverage	1. when w=0, $WQ_k = h_{ii}r_i^{*2}$
2. $r_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$	2. $D_i = \frac{h_{ii}}{p(1-h_{ii})^2} r_i^2$	2. When w=0.5, $WQ_k = r_i^{*2}/2$
3. $Q_k = s^2 r_i^2 = s_{(i)}^2 r_i^{*2}$	3. $R_i = (1-h_{ii})(1-\frac{r_i^2}{n-p})$	3. When w=1.0 $WQ_k = (1-h_{ii})r_i^{*2}$

Table 1. Comparisons of measures for detecting an outlier and/or influential observation

Table 2 shows the relationships among D_I, R_I, Q_k and WQ_k, when the number of outliers or influential observations are greater than 1. The second equation of R_I is obtained from Draper and John(1981) where RSS_C is defined in the equation (1.8). As we already noted, when w=0 the WQ_k looks similar to D_I where the numerator of the first equation is the equation(1.12), around w=0.5 it behaves like Q_k and when w=1.0, the WQ_k is the sum of the scaled residuals.

Measures for detection of an outlier	Measures for detection of an influential observation	Proposed statistics
1. $Q_k = \underline{e_2}'(I-H_{22})^{-1}\underline{e_2}$ $= \underline{e_2}'\underline{e_2} + \underline{e_1}'X_1(X_1'X_1)^{-1}X_1'\underline{e_1}$	1. $D_I = \frac{(\hat{\beta}-\hat{\beta}_{(I)})'X'X(\hat{\beta}-\hat{\beta}_{(I)})}{ps^2}$	1. when w=0, $WQ_k = \frac{(\hat{\beta}-\hat{\beta}_{(I)})X'X(\hat{\beta}-\hat{\beta}_{(I)})}{ks_{(I)}^2}$ $= \frac{e_1\cdot(X_{(I)}'X_{(I)})^{-1}X_{(I)}'e_1}{ks_{(I)}^2}$
	2. $R_I = \frac{(n-p-k)s_{(I)}^2\lvert X_{(I)}'X_{(I)}\rvert}{(n-p)s^2\lvert X'X\rvert}$ $= \left(1-\frac{Q_k}{RSS}\right)\lvert I-R_{22}\rvert$	2. When w=0.5, $WQ_k = \frac{Q_k}{2ks_{(I)}^2}$
		3. When w=1.0 $WQ_k = \frac{e_2'e_2}{ks_{(I)}^2}$

Table 2. Comparisons of measures for more than 1 outlier and/or influential observation

3. Example

Here we are interested in the maximum value of WQ_k, depending on k which is the assumed number of outliers. for convinience, we may write this as max WQ_k The statistic max WQ_k will now be obtained from the analysis of a set of 21 observations (x,y) which is similar to the data set given by Mickey, Dunn and Clark (1967). Among the 21 observations, the only difference is (42,51) which is changed into (42,35) in this paper. The reason of this change is that we want to see some diversified change in detected points.

Case	x	y	Case	x	y
1	15	95	11	7	113
2	26	71	12	9	96
3	10	83	13	10	83
4	9	91	14	11	84
5	15	102	15	11	102
6	20	87	16	10	100
7	18	93	17	12	105
8	11	100	18	42	35
9	8	104	19	17	121
10	20	94	20	11	86
			21	10	100

Table 3. Age at First Work(x) and Gesell Adaptive Score(y)

The observations appear in Table 3 and plotted in Figure 1. A straight line regression model is fitted to the full set of data and then to the 20 data points, remained when each observation is deleted in turn. Our test statistic max WQ_k is obtained where the scale factor (s.f.) is $ks^2_{(n)}$, and k is the assumed number of outliers.

Table 4 shows the weights ranged from 0 to 1, the value of max WQ_k and the deleted observation number. When the weights are small($0 \le w \le 0.2$), the number 18 is deleted. However, when the weights become larger ($w \ge 0.2$), the deleted number changes from 18 to 19. The reason for this is that the residual for observation 19 is too large than any others.

For the removal of two cases, the results are summarized in Table 5. The results show that the deleted observations are varying with respect to w. It

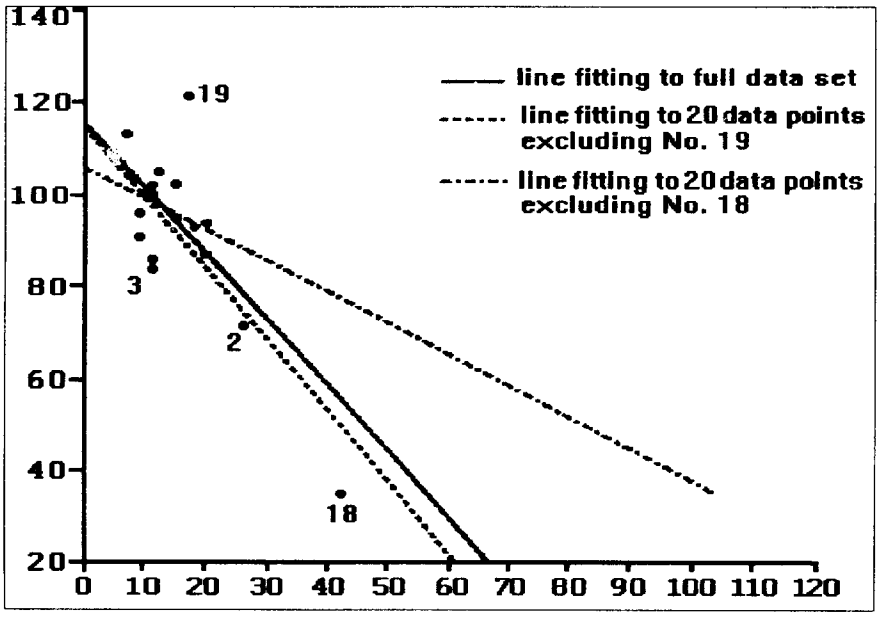

Figure 1. Plot of Example Data

seems that the (2,18) observations are most influential and (3,19) are outliers.

given weights	max WQ_k	deleted observation
0.0	2.609	18
0.1	2.480	18
0.2	2.942	19
0.3	4.750	19
0.4	5.208	19
0.5	6.341	19
0.6	7.474	19
0.7	8.607	19
0.8	9.740	19
0.9	10.873	19
1.0	12.005	19

Table 4. Detected Observation and the value of max WQ_k when One Point Detection for Example data

given weights	max WQ_k	deleted observation
0.0	3.108	2,18
0.1	2.881	2,18
0.2	2.653	2,19
0.3	3.012	18,19
0.4	3.694	18,19
0.5	4.376	18,19
0.6	5.161	3,19
0.7	5.986	3,19
0.8	6.810	3,19
0.9	7.635	3,19
1.0	8.459	3,19

Table 5. Detected Observation and the value of max WQ_k when Two Point Detection for Example data

Table 6 shows the detected points when other statistics are used in this example. In one point detection case, the detected observation is the number 19

when using the outlier detecting statistic Q_k, r_i and r_i^*, while the number 18 is detected by using the statistics of influential observations such as D_I, R_I and the leverage h_{ii}. From these results, the number 19 is the most outlying case and the number 18 is the most influential and remote point which has been already shown in Table 4.

For two points detection R_I and D_I detect the numbers 2 and 18 where Q_k detects the numbers 18 and 19. In Table 5 the results include these points and in addition, when $w > 0.5$, the points 3 and 19 are detected. Hence, we note that the proposed statistic max WQ_k extensively shows the influential points and the outlying points as w varies from 0 to 1.

Statistics	Detected Points
Q_k	19 (12.68) 18, 19 (8.75)
R_I	18 (0.27) 2, 18 (0.07)
D_I	18 (3.27) 2, 18 (11.29)
r_i	19 (2.79)
r_i^*	19 (3.55)
h_{ii}	18 (0.65)

Table 6. Detected Observations for Exemple in Other Statistics(the values of the test statistics for detected points are given within the parentheses)

5. Concluding remarks

In practical regression situations, the outliers and influential observations are the same, but very often they are different i.e., some points are influential outliers, some points are influential but are not outliers and vice virsa. The proposed statistic WQ_k is a composite statistic which can detect outliers and influential observations separately, if they are different. When the detected points have both influential and outlier properties, then the proposed statistic tells us that they have both. Hence, we believe that it is very useful to use the proposed statistic in practice. Of course, it is more efficient to use this statistic when the outliers and influential observations are different.

Note that there are no serious probrems in computational cost in calculating WQ_k. The number of multiplications for all possible Q_k is proportional

to $(k^2+1)n^k/k!$ from the equation $Q_k = \underline{e}_2'(I-H_{22})^{-1}\underline{e}_2$, and the number of multiplications for all possible WQ_k is proportional to 7(number of weight points) $(k^2+1)n^k/k!$. Since the size of k is much smaller than the size of n, the cost doesn't make much trouble.

We have done some Monte Carlo simulation for simple regression cases to find the probability distribution of max WQ_k and the critical values for testing possible outliers/influential observations. Due to the limitation of pages in this paper, the detailed results are not given.

The simulation results show that the distribution is skewed to the right when the weight w is small and it moves steadily to the right hand side when w becomes larger and it tends to be symmetric. The results also show that the critical values are highly robust to design patterns in simple regression. We hope to report some details of the simulation work in a separate paper later.

References

[1] Andrews, D. F., and Pregibon, D. (1978), "Finding the outliners That Matter," *Journal of the Royal Statistical Society,* Ser. B, **40**, 87-93.

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

[3] Chatterjee, S. C. and Hadi, A. S. (1986), "Influential Observations, High Leverage Points, and Outliers in Linear Regression," *Statistical Science,* **1**, 379-416.

[4] Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics,* **19**, 15-18.

[5] -----(1979), "Influential Observations in Linear Regression", *Journal of the American Statistical Associations,* **74**, 169-174.

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

[7] Gentleman, J. F., and Wilk, M. B. (1975a), "Detecting Outliers in a Two-Way Table: I. Statistical Behavior of Residuals," *Technometrics,* **17**, 1-14.

[8] ----- (1975b),"Detecting Outliers II. Supplementing the Direct Analysis of Residuals," *Biometrics,* **31**, 387-410.

[9] Little, J. K. (1985), "Influence and Quadratic Form in the Andrews-Pregibon Statistic," *Technometrics*, **27**, 13-15.

[10] Mickey, M. R., Dunn, O. J., and Clark, V. (1976), "Note on the Use of Stepwise Regression in Detecting Outliers," *Computers and Biomedical Research*, 1, 105-111.

회귀진단에서 이상치와 영향관측치를 동시에 발견하는 새로운 통계량에 관한 연구[1)]

강 은 미[2)]

요 약

회귀진단에서 이상치와 영향을 많이 주는 측정치를 발견하는 새로운 통계량을 제안하였다. 이 제안된 통계량은 이상치를 찾는 측도와 영향추정치를 찾는 측도의 가중합으로 해석될 수 있으며, 가중치를 변화시킴으로써 이상치와 영향추정치들을 일목요연하게 찾아낼 수 있다는 장점이 있다. 씨뮬레이션을 이용하여 제안된 통계량의 분포형태를 살펴 보았다.

[1)] 본 연구는 1990년도 학술진흥재단 자유공모과제 연구비에 의해서 수행되었음.
[2)] (136-742) 서울특별시 성북구 동선동 성신여자대학교 자연과학대학 통계학과 부교수