Journal of the KSQC
Vol. 17, No. 1 1989

A Study on Influential Subset Detection in Regression

Kim, Seung Gu*
Yum, Joon Keun**

ABSTRACT

This study is suggested to find a methodology of influential subset detecting by using diagonal elements of the hat matrix, h_{ij}, and off-diagonal elements of the modified hat matrix, h_{ij}^*. And this study shows that elements of two matrix is very useful to detect a masking subset in various situations by analytic and emperical explanations. Also a desirable procedure to detect the significant subset and the calibration points for h_{ij} and h_{ij}^* are recommended.

1. Introduction

A interests of the influential observations in regressions began with R. D. Cook (1977) and D. C. Hoaglin and R. E. Welsch (1978). A many papers were suggested by being revealed the charaterizations and structures of this observation's after the middle of 1970's.

An i or j observations of (A) in (Fig. 1. 1) are an influential observation since when one of two is deleted a fitted model was seriously changed. But in case of (B) in (Fig. 1. 1), deletion of one of two hardly changes the fitted model but deletion of both simultaneously affects the fitted model. Sometimes, this is called a "masking effects", in case of (A), both is not affetive, so we call these observations an individual influential observations. And an i, j

* Doctoral degree in department of statistics, Dong Guk Univ.
** Department of statistics, Dong Guk Univ.

observations of (B) is called a jointly influential observations. A measures of the sensitivity that an observation affects the model have been suggested many times, and is almost systemized in an individul observations case.

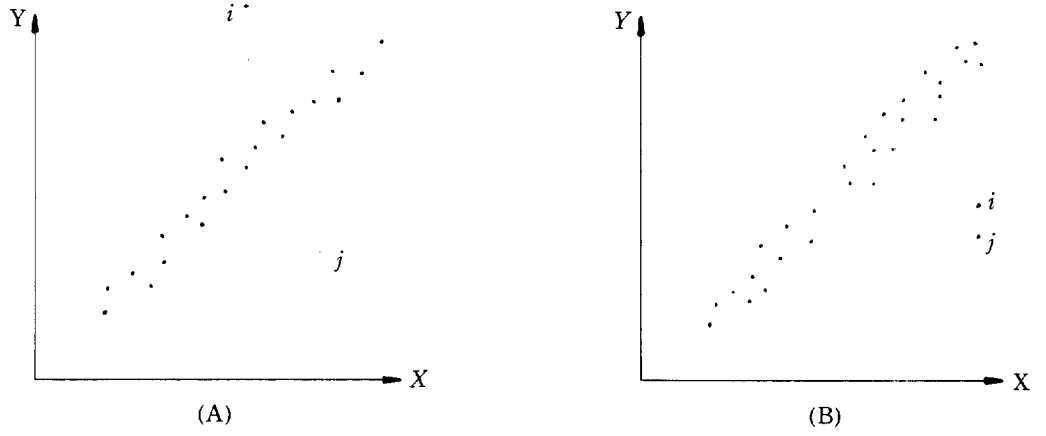

Fig. 1.1

S. Chatterjee and A. S. Hadi (1986) discussed the relationship between these measures. The methodology of detecting subset, however, was nearly suggested, except Gray and Ling(1984), Hadi(1985), and Cook(1980).

This paper is suggested to study on effective mothod for identifying the significant influential subset by using the off diagonal elements of hat matrix H and the modified hat matrix H^*.

2. Basic Concept and Assumptions.

Suppose the linear regression model,

$$Y = X\beta + \varepsilon \quad , \tag{2.1}$$

where Y is an $(n \times 1)$ vectors of observations, X is an $(n \times p)$ full rank matrix, β is a $(p \times 1)$ vector of coefficients and ε is a $(n \times 1)$ vector of random error such that $E(\varepsilon) = 0$, $V(\varepsilon) = I\sigma^2$.

Then we have the LSE $\hat{\beta} = (X'X)^{-1}X'Y$ and

$$\begin{aligned} Y = X\hat{\beta} &= X(X'X)^{-1}X'Y \\ &= HY, \text{ where } H = X(X'X)^{-1}X' \end{aligned} \tag{2.2}$$

$$V(\hat{Y}) = \sigma^2 H \tag{2.3}$$

$$V(e) = \sigma^2 (I - H), \text{ where } e = Y - \hat{Y} \tag{2.4}$$

A matrix H is called a "Hat Matrix" which maps Y into $\hat{Y}$.

Let $h_{ij} = i^{th}$ diag(H), $i = 1, 2, \cdots, n$ (2.5)

The elements h_{ij} of H can be interpreted as the amount of effect exerted by Y_j an $\hat{Y}_i$. From (2.4)

$$0 \leq h_{ii} \leq 1 \tag{2.6}$$

Also, H is a symmetric idempotent matrix and hence

$$tr(H) = \text{rank}(H) = p, \tag{2.7}$$

so $$\sum_i h_{ii} = p \tag{2.8}$$

and $$\sum_j h_{ij}^2 = h_{ii} \tag{2.9}$$

For residuals e_i, e_j, $V(e_i) = \sigma^2 \sqrt{1 - h_{ii}}$ and $\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$ and the correlation coefficients of residuals is

$$e_{ij} = \frac{-h_{ij}}{\sqrt{(1-h_{ii}) \cdot (1-h_{jj})}} \tag{2.10}$$

3. Influential Subset.

we want to find a subsets that influence the linear model, but it should be noted that it's not desired that subset we try to find include one or more individual influential points which may have relatively large t_i and h_{ii} just as (A) in (Fig. 3.1). Since the effects of subset included influential observations is largely caused by that points, it is not regarded as masking effected subset.

Accordingly, subset excluded individual influential observations must be detected as the influential subset just as an observations in the dotted circle of (B) in (Fig. 3.1).

The features of this subset's is that correlation coefficients of residuals, defined as (2.10), between elements in the subset, $|e_{ij}|$, are large. And other feature is that h_{ii} may equal to h_{ij}, for $i \neq j$.

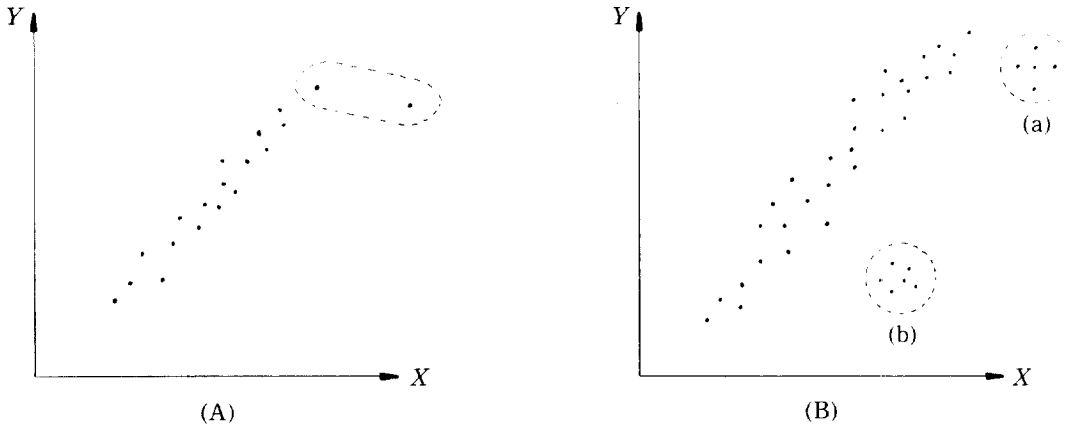

Fig. 3.1

R. D. Cook and S. Weisberg (1972, page 141−142) discussed that two points which lie on same (or opposite) positions from the centroid in x−space have large $|e_{ij}|$ and they have $|e_{ij}|=0$ when they are on perpendicular position in x− space each other.

According to this important discussion, we pay attention to the subset of observations that may have large $|e_{ij}|$.

4. Hat Matrix H and Modified Hat Matrix H^*

(1) Aspects of hat matrix H

The i^{th} diagonal elements of H, h_{ii}, is called a leverage such that lever Y to $\hat{Y}$. The leverage is the potential effects since it has only effects of $x's$.

The off diagonal elements, h_{ij}, have a relationship to $e_{ij}=-h_{ij}\sqrt{(1-h_{ii})(1-h_{jj})}$, as we have seen in (2.10). So that, we can detect a significant subsets by large h_{ij}. But $|h_{ij}|$ may be small even though observations have stong relation in x− space if h_{ii} and h_{jj} are large. On the other hands, $|h_{ii}|$ may be overestimated when h_{ii} and h_{jj} are small. Especially when subset consisted of high leveraged observations, just as (a) of (B) in (Fig. 3.1), is detected, h_{ij} poorly detects a significant subset.

(2) Aspects of modified hat matrix H^*

A matrix, saying a modified hat matrix, is defined as

$$H^*=X^*(X^{*\prime}X^*)^{-1}X^{*\prime}, \text{ where } X^*=[X\ ;\ Y] \tag{4.1}$$

$$=H+\frac{ee'}{(n-p)\hat{\sigma}^2} \tag{4.2}$$

Hence, $(i,j)^{th}$ off diagonal elements of H^* is

$$h_{ij}^* = h_{ij} + e_i e_j / [(n-p)\hat{\sigma}^2]$$

$$= h_{ij} + (1-h_{ii})^{\frac{1}{2}}(1-h_{jj})^{\frac{1}{2}} r_i r_j / (n-p) \qquad (4.3)$$

, where $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ is the internal studentized residual.

It is well known that $E(r_i)=0$, $V(r_i)=1$ and $E(r_i\, r_j)=e_{ij}$.

Then we have

$$E(h_{ij}^*) = h_{ij} + (1-h_{ii})^{\frac{1}{2}}(1-h_{jj})^{\frac{1}{2}} E(r_i r_j)/(n-p)$$

$$= h_{ij} + (1-h_{ii})^{\frac{1}{2}}(1-h_{jj})^{\frac{1}{2}} \cdot e_{ij}$$

$$= h_{ij}[1-1/(n-p)]$$

$$= -e_{ij}(1-h_{ii})^{\frac{1}{2}}(1-h_{jj})^{\frac{1}{2}}\left(1-\frac{1}{n-p}\right) \qquad (4.4)$$

From (4.3), we are indecated that h_{ij}^* is a combination of the leverage part and the scaled residual part of two elements. Moreover, formula (4.4) indicates that it is expected that h_{ij}^* also have a relationship with e_{ij}.

Hence we expected that a large h_{ij}^* can identify the significant subset which consists of low-leveraged outliers just as (b) of (B) in (Fig. 3.1).

However, it is possible that subset consisted of high leverage points might not be detected distinctly by h_{ij}^* since it's values are seen to be small relative to that of outliers, not-high leverage subset.

5. Calibration points of h_{ij} and h_{ij}^*

Hoagin and Welsch (1978) recommended examination of h_{ii} for high leverage observations suggesting $2p/n$ as a calibration point for h_{ii}.

Since the modified hat matrix H^* is a natural expention of H, the desirable calibration point of h_{ii}^* is $2(p+1)/n$.

And from formula (2.9)

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 \leq n(\max_j h_{ij})^2 \qquad (5.1)$$

If $h_{ii} \geq 2p/n$, that is, i^{th} observation is an high leveraged point, then

$$\max_j |h_{ij}| \geq \sqrt{2p/n} \qquad (5.2)$$

Accordingly, we suggested the calibration point for observation's pair with highly correlated residuals is $\sqrt{2p/n}$ for h_{ij} and $\sqrt{2(p+1)n}$ for h_{ij}^*.

6. Numerical Examples.

In order to know that the method suggested detects the significant subset which is made a different type, we use 22 imaginary data to simple regression.

The first data include two jointly subset which observations are not high leverage but may be an outliers.

Observations subset (20, 21) and (17, 18, 19) is seen not to be very influential subset as seeing (A) in (Fig. 6.1). Indeed, even if one of two subset are deleted, the fitted model is slightly changed.

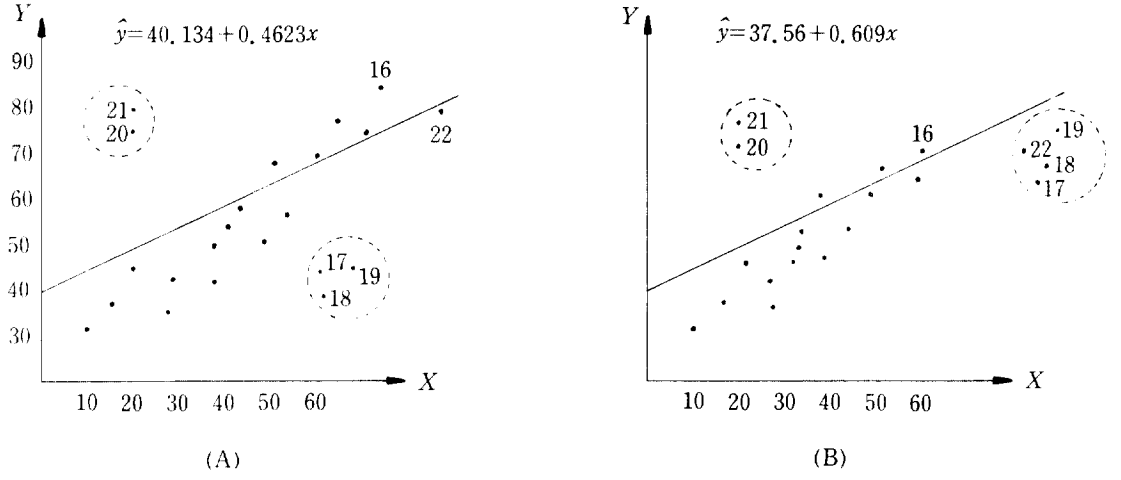

Fig. 6.1

But it is necessary to indentify the subset of observations which appears on similar positions in regression space repeatedly.

Table (6.1) is 22 imaginary data and shows their features.

Subset (20, 21) and (17, 18, 19) is to be detected, and since observation 22 is a distinctive influential observation, it should not be included in these subset. Table (6.2) is the first data's elements of H and H^*.

Table 6.1 22 imaginary observations and absolute residuals and leverage h_{ii}. A parenthesis is the second data modifying 17, 18, 19 and it's components.

No.	X	Y	$\mid ei \mid$	h_{ii}
1	11	33	12.22(11.26)	0.171(0.140)
2	16	38	9.53(9.30)	0.130(0.112)
3	25	38	13.69(14.78)	0.076(0.074)
4	20	45	4.83(4.74)	0.102(0.093)
5	26	48	4.15(5.39)	0.075(0.070)
6	32	47	7.93(10.04)	0.053(0.055)
7	32	53	1.93(4.04)	0.053(0.055)
8	28	56	2.92(1.39)	0.064(0.064)
9	35	60	3.68(1.13)	0.048(0.050)
10	38	55	2.70(5.69)	0.046(0.047)
11	43	60	0.01(3.74)	0.049(0.046)
12	40	68	9.37(6.09)	0.046(0.046)
13	48	70	7.67(3.22)	0.060(0.049)
14	55	74	8.44(2.96)	0.090(0.063)
15	51	77	13.29(8.39)	0.071(0.054)
16	60	83	15.27(8.91)	0.122(0.078)
17	54(76)	45(81)	20.10(2.82)	0.085(0.061)
18	56(77)	43(82)	20.00(2.44)	0.096(0.167)
19	58(80)	47(84)	19.95(2.61)	0.108(0.189)
20	22	74	23.70(23.05)	0.091(0.084)
21	22	77	26.70(26.05)	0.091(0.084)
22	76	84	8.73(4.70)	0.278(0.221)

h_{ij}* is well detects two significant subsets with calibration point $\sqrt{2(p+1)/n} \doteqdot 0.1113$ but h_{ij} doesn't with calibration point $\sqrt{2p/n} \doteqdot 0.091$.

Next, let points 17, 18 and 19 move near to point 22 to make subset which consists of high leverage point but not outiers as seeing (B) in (Fig. 6. 1).

So that, the second data has two subset, that is, (20, 21) is a subset of outliers but not high influential observations, whereas(17, 18, 19, 22) has the contrary features.

Also h_{ij}* is well detecting method with rule $\sqrt{2(p+1)/n}$ but h_{ij} also poorly detects the subset with low leverage observations, such that (20, 21), on the contrary, overestimates the subset with high leverage observations, so that point 16 is tend to be included in subset (17, 18, 19, 22).

7. Conclusions.

Analytic and emperical discussion which have been done in this paper give some results, that

Table 6.2 The first data ; 15 elements of H^* and H, the rests are omitted since there are not significant larger value than calibration points $\sqrt{2(p+1)}/n \doteqdot 0.1113$ and $\sqrt{2p}/n \doteqdot 0.091$ for h_{ij}^* and h_{ij} respectively. Mark '*' notes the significant larger value than calibration points after individual influential observation 22 was deleted.

	No.	15	16	17	18	19	20	21	22
Hat matrix H	15	.067							
	16	.084	.112						
	17	.073	.093*	.080					
	18	.076	.099*	.084	.089				
	19	.080	.106*	.089	.094*	.099			
	20	.015	−.008	.008	.002	−.003	.088		
	21	.015	−.088	.008	.002	−.003	.088	.088	
	22	.127	.188	.147	.161	.174	−.069	−.069	.350
Modified Hat matrix H^*	15	.117							
	16	.042	.181						
	17	.008	.022	.190					
	18	.002	.017	.210*	.233				
	19	.017	.037	.199*	.221*	.211			
	20	.093	.079	−.120	−.144	−.130	.236		
	21	.103	.091	−.136	−.161	−.146	.255*	.275	
	22	.153	.213	.096	.102	.121	−.004	.003	.298

is, on detecting an significant subset, which has two styles.

One is the subset of which elements are observations with high leverage.

The other is the subset of which elements are observations with large residuals.

The former is over−detected as an influential subset, and the latter is bejond−detected by h_{ij}.

h_{ij}^*, whereas, is a good tools to detect the significant subset of two situations with calibration point $\sqrt{2(p+1)}/n$.

According to the results in this papers, we suggests the procedures to detect the significant or influential subset.

1. Calculating the modified hat matrix $H^* = (h_{ij}^*)$
2. Excluding individual influential observations after detecting it by several measures.
3. Identifying pairs of observations of which $|h_{ij}^*|$ is larger than calibration point $\sqrt{2(p+1)}/n$. (Note that negative sign of h_{ij}^* means the opposite positions of two observation space.)
4. Identifying observations corresponding to pairs selected, and calculating it's leverage h_{ii}.
5. Classifying the subset of observation by the simililar magnitude of h_{ii}.

Table 6.3 The second data ; 15 elements of H^* and H, the rests are omitted since there are not significant larger value than calibration points $\sqrt{2(p+1)}/n \doteqdot 0.1113$ and $\sqrt{2p/n} \doteqdot 0.091$ for h_{ij}^* and h_{ij} respectively. Mark '*' notes the significant larger value than calibration points.

	No.	15	16	17	18	19	20	21	22
Hat matrix H	15	.054							
	16	.062	.078						
	17	.076	.107*	.161					
	18	.077	.108*	.164*	167				
	19	.080	.114*	.174*	178*	.189			
	20	.028	.010	−.021	−.023	−.029	.084		
	21	.028	.010	−.021	−.023	−.029	.084	.084	
	22	.084	.121*	.187*	.192*	.204*	−.037	−.037	.221
Modified Hat matrix H^*	15	.086							
	16	.096	.115						
	17	.066	.094	.177					
	18	.068	.101	.180*	.184				
	19	.071	.107	.191*	.195*	.207			
	20	.115	.100	−.068	−.067	−.071	.345		
	21	.127	.112	−.073	−.071	−.075	.379*	.417	
	22	.077	.109	.172*	175*	.186*	−.034	−.035	.171

REFERENCES

1. Chatterjee, S and Hadi, A. S(1986), "*Influential observations, high leverage points, and outliers in Linear Regression*", Statistical Science, Vol. 1, No. 3, 379−416.
2. Cook, R. D(1977), "*Detection of Influential observations in Linear Regression*", technometrics, Vol. 19, 15−18.
3. Cook, R. D(1980), "*Charaterizations of an EIF for detecting Influential Cases in Regression*", technometrics, Vol. 22, No. 4, 495−508.
4. Cook, R. D and Weisberg, S(1982), "*Residuals and Influence in Regression*", Chapman and Hall.
5. Gray, J. B and Ling, R. F(1984), "*K−clustering as a detection tool for influential subset in regression*", technometrics, Vol. 26, No. 4, 305−318.
6. Hadi. A.S(1985), "*Letters to the Editor*," technometrics, Vol. 27, No. 3, 323-324.
7. Hoaglin, D. C and Welsch, R. E(1978), "*The hat matrix in Regression and ANOVA*", The American Statistitian, Vol. 32, No. 1, 17−22.