The Korean Communications
in Statistics Vol. 7, No. 3, 2000
pp. 667-676

Identifying Multiple Leverage Points and Outliers in Multivariate Linear Models

Jong Young Yoo[1]

Abstract

This paper focuses on the problem of detecting multiple leverage points and outliers in multivariate linear models. It is well known that the identification of these points is affected by masking and swamping effects. To identify them, Rousseeuw(1985) used robust estimators of MVE(Minimum Volume Ellipsoids), which have the breakdown point of 50% approximately. And Rousseeuw and van Zomeren(1990) suggested the robust distance based on MVE, however, of which the computation is extremely difficult when the number of observations n is large. In this study, we propose a new algorithm to reduce the computational difficulty of MVE. The proposed method is powerful in identifying multiple leverage points and outliers and also effective in reducing the computational difficulty of MVE.

Keywords : leverage points, Mahalanobis distance, masking effect, MVE(minimum volume ellipsoids), outliers, robust distance, swamping effect

1. Introduction

We consider the problem of identifying and testing multiple leverage points and outliers in linear models. Linear models are commonly used to analyze data on many fields of study and these data often contain leverage points and outliers. In general, leverage points are located far away from the bulk of the explanatory variables and outliers do not follow the pattern of the majority of the data.

Classical methods of leverage points and outliers detection are powerful when some data set contain only a leverage point or an outlier. But it would be much more difficult to detect multiple leverage points and outliers in multivariate linear models. The identification of multiple leverage points and outliers can be perplexing because of masking and swamping effects. The masking effect occurs when a leverage point or an outlier obscures the existence

[1] Assistant Professor, Department of Computer Science and Statistics, Yongin University, Kyunggi-Do, 449-714, Korea.
E-mail : jyyoo@eve.yongin.ac.kr

of another, while the swamping effect occurs when non-leverage points or non-outliers are wrongly identified as the leverage points or outliers. It is difficult to control masking and swamping effects when we use non-robust estimators. To reduce the masking and swamping effects, several robust methods have been proposed in recent years.

We consider the standard linear model

$$y = X\beta + \varepsilon, \tag{1}$$

where $y=(y_1, y_2, \ldots, y_n)^T$ is an $n\times 1$ vector of values of the response variable, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $(p+1)\times 1$ vector of unknown parameters, $X=(x_1^T, x_2^T, \cdots, x_n^T)^T$ is an $n\times(p+1)$ matrix of explanatory variables with rank $p+1<n$, and $\varepsilon=(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ is an $n\times 1$ vector of independent normal random variables with mean $\mathbf{0}$ and unknown variance $\sigma^2 I_n$. We can estimate the unknown parameter β as the ordinary least squares(OLS) estimator $b = (X^TX)^{-1}X^Ty$. So the vector of OLS residuals can be written as

$$e = y - Xb = (I_n - P)y, \tag{2}$$

where $P = (p_{ij}) = X(X^TX)^{-1}X^T$. Then the residual sum of squares is $SSE = e^Te$ and the estimate of σ^2 is $e^Te/(n-p-1)$.

Consider the partition $x_i^T=(1, z_i^T)$ and let $D_i(M, V)$, $(i=1,2,\cdots,n)$ be the measure of the distance between the observation z_i and a location estimator M with a dispersion measure V. Then Mahalanobis distance, a classical method, can be expressed by the arithmetic mean of explanatory variables $\overline{Z}$ and the usual sample covariance matrix $S(Z)$.

$$MD_i = D_i(\overline{Z}, S(Z)) = \sqrt{(z_i - \overline{Z})^T S(Z)^{-1}(z_i - \overline{Z})} \quad i=1,2,\cdots,n\ . \tag{3}$$

With the significance level α, the cutoff value of MD_i would be $\sqrt{\chi^2_{p,1-\alpha/2}}$. So the values of MD_i that exceed the cutoff values may indicate that the corresponding observations are leverage points. However, leverage points do not necessarily have large values for MD_i, because of the masking effect and not all observations with large MD_i values are necessarily leverage points, because of the swamping effect. This is due to the fact that $\overline{Z}$ and $S(Z)$ are not robust. Therefore, it seems necessary to replace $\overline{Z}$ and $S(Z)$ in (3) by robust estimators respectively.

In the usual multiple linear regression model, we often use the diagonal elements of the hat matrix $P=X(X'X)^{-1}X'$ as diagnostic measure to identify leverage points. Hoaglin and Welsch(1978) pointed out the cutoff value of p_{ii} is $2(p+1)/n$. The method with diagonal element p_{ii} also has the problem of masking and swamping effects as well as the Mahalanobis distance. This can be explained by realizing that there exists a relation between p_{ii} and MD_i by

$$p_{ii}=\frac{(MD_i)^2}{n-1}+\frac{1}{n} \qquad i=1,2,\cdots,n \ . \tag{4}$$

Rousseeuw(1985) suggested the minimum volume ellipsoid(MVE) that covers at least half the observations to construct robust estimators. The MVE is defined as the pair (M, $V(Z)$), where M is a p vector and $V(Z)$ is a positive semidefinite $p\times p$ matrix such that the determinant $V(Z)$ is minimized subject to

$$\#\{i;(\ z_i-M)^T V(Z)^{-1}(\ z_i-M)\le\chi^2_{p,.50}\}\ge h, \quad i=1,2,\cdots,n \tag{5}$$

where h is an integer part of $(n+p+1)/2$. The advantage of the MVE estimator is that the breakdown point is as high as 50% approximately. However, the MVE estimates are computationally very expensive. For example, if we have an $n\times(p+1)$ data matrix X, then we need to compute the volumes of $n!/h!(n-h)!$ ellipsoids to select the MVE. So it is rarely possible to compute the MVE if n is large.

To reduce the computational difficulty of the MVE, Rousseeuw and Leroy(1987) suggested the resampling algorithm approximating computation of the MVE. The resampling algorithm draws several subsamples each of size $p+1$ and then for each subsample j, we compute

$$D_i(M_j,\ V(Z)_j)=\sqrt{(\ z_i-M_j)^T V(Z)_j^{-1}(\ x_i-M_j)}, \qquad i=1,2,\cdots,n \tag{6}$$

where M_j and $V(Z)_j$ denote the mean and covariance matrix for the j th subsample. Let m_j be the $100(h/n)$th percentile of the n values in equation (6). The volume of an ellipsoid based on M_j, $V(Z)_j$ and containing h observations is proportional to $\{m_j^p det(\ V(Z)_j)\}^{(1/2)}$. If we let k be the subsample for which $m_k^p det(\ V(Z)_k)$ is minimum, then the ellipsoid based on subsample k can be used as an approximation of the MVE containing h observations.

By using this resampling algorithm, Rousseeuw and van Zomeren(1990) suggested the robust distances

$$RD_i = D_i(\boldsymbol{M}_j, c_j V(Z)_j) = \sqrt{(\boldsymbol{z}_i - \boldsymbol{M}_j)^T (c_j V(Z)_j)^{-1} (\boldsymbol{z}_i - \boldsymbol{M}_j)}, \quad i = 1, 2, \cdots, n \tag{7}$$

to identify leverage points in the data set Z with the cutoff value $\sqrt{\chi^2_{p,1-\alpha/2}}$. A correction factor $c_j = \{1 + 15/(n-p)\}^2 m_j / \chi^2_{p,0.50}$ is used to achieve consistency at multivariate normal distributions.

RD_i is proved to be robust in the problems of masking and swamping effects, but has some difficulties in practice. Hadi(1992) pointed out the following three problems.

First, a decision has to be made on the number of subsamples.

Second, when Rousseeuw and van Zomeren(1990) calculate the RD_i of equation (7), they had the assumption of that X is a general position. (X is said to be in the general position when every subsample of size $p+1$ has rank p).

Third, even if all subsamples of size $p+1$ have rank p, it may happen that the covariance matrices for some subsamples have nearly zero determinants and hence the corresponding ellipsoids have nearly zero volumes.

Although Rousseeuw and van Zomeren's method using the resampling algorithm reduced the computational complexity from ${}_nC_h$ to ${}_nC_{p+1}$, the computation is still very expense. For example, if a consideration is given for 75 × 4 explanatory matrix, we need 1,215,450 computations of the MVE.

In this paper, we propose a procedure for reducing the numbers of calculation of the MVE by using the residual method which is suggested by Yoo and Kim(1996). The procedure is easy to compute and have good powers in identifying good and bad leverage points.

2. Proposed procedures

Let p_{kk} is the largest element of diagonal matrix of P. MAD means the median absolute deviation, e_k and s_k reveal the residual and standard deviation of the kth observation respectively.

The Suggested Algorithm

<Step 1> Find a clean subset of M, initially of size h which is an integer part of $(n+p+1)/2$.

① Calculate the regression coefficients $\boldsymbol{b}$ and residuals $\boldsymbol{e}$ by OLS.

② Select the largest diagonal element of p_{kk} of hat matrix P.

③ Calculate the MAD of residuals by using Yoo and Kim's model by simulating e_k in the interval of $(e_k - 3s_{k,}\ e_k + 3s_k)$.

④ Find a minimum MAD model and calculate the corresponding residual $\boldsymbol{e}_k'$.

⑤ Sort the $|\boldsymbol{e}_k' - median(\boldsymbol{e}_k')|$ in ascending order and select the first h data.

<Step 2> Find the subsample j of size $p+1$ which $m_j^p det(\ V(Z)_j)$ is minimum under the clean subset M.

<Step 3> Calculate the following SRD_i based on the selected subset of <Step 2>.

$$SRD_i = D_i(\boldsymbol{M}_j, c_j V(Z)_j) = \sqrt{(\boldsymbol{z}_i - \boldsymbol{M}_j)^T (c_j V(Z)_j)^{-1} (\boldsymbol{z}_i - \boldsymbol{M}_j)} \qquad i = i, 2, \cdots, n \tag{8}$$

where $c_j = c_{np} m_j / \chi^2_{p,0.50}$ and $c_{np} = \{1 + 15/(n-p)\}^2$.

<Step 4> Plot the standardized LS(Least Squares) residuals with the Mahalanobis distance MD_i's, the standardized LMS(Least Median of Squares) residuals with the robust distance RD_i's, the standardized LS residuals with the diagonal elements of hat matrix p_{ii} and the standardized LMS residuals with the suggested robust distance SRD_i's.

The theoretical background of <Step 1> can be found in Yoo and Kim(1996), <Step 2> and <Step 3> can be found Rousseeuw and van Zomeren(1990). By plotting <Step 4>, we can identify the good and bad leverage points and outliers.

3. Examples

In this section, we test the powers of the proposed procedure and compare this with other methods using stackloss data and artificial data. These data sets have been widely used to illustrate leverage points and outliers in linear regression.

3.1 Stackloss Data(Brownlee 1965, reused in Rousseeuw and Leroy(1987))

The stackloss data set consists of twenty one observations on three explanatory and one response variable. It is known that there are four leverage points(observations 1, 2, 3, 21) and one outlier point(observation 4). <Table 1> reveals the value of Mahalanobis distance MD_i, the robust distance RD_i, the diagonal elements of hat matrix $\boldsymbol{P}$ and the suggested robust distance SRD_i. With significant level of 0.05, the cutoff value of MD_i , RD_i and SRD_i are $\sqrt{\chi^2_{3,0.975}}$=3.06. The largest MD_i(observation 17) is only 2.70. The MD_i analyzes that there is no leverage point in this data. According to the p_{ii}, there is only one leverage point above the cutoff value 0.381. The robust distances RD_i, however, clearly point out the four leverage points(observation 1, 2, 3 and 21). To compute the RD_i, we should compute ${}_{21}C_4$=5,985 subsamples of size 4 observations. A search of all the subset found two subset ({7, 10, 14, 20} and {8, 10, 14, 20}) with the same minimum $m_j^p det(V(Z)_j)$. The robust RD_i is computed based on this two subset. When we apply the suggested algorithm to this data, we obtain the clean subset of M={10, 8, 20, 14, 16, 18, 19, 7, 5, 2, 13, 15}. Based on this clean subset, we compute ${}_{12}C_4$=715 subsamples of size 4 observations and find two same subset ({7, 10, 14, 20} and {8, 10, 14, 20}) as the robust distance method. Finally, we have the same result but the number of calculations are reduced from 5,985 to 715. In this data RD_i and SRD_i identify the leverage points correctly, but MD_i and p_{ii} fail to identify the leverage points.

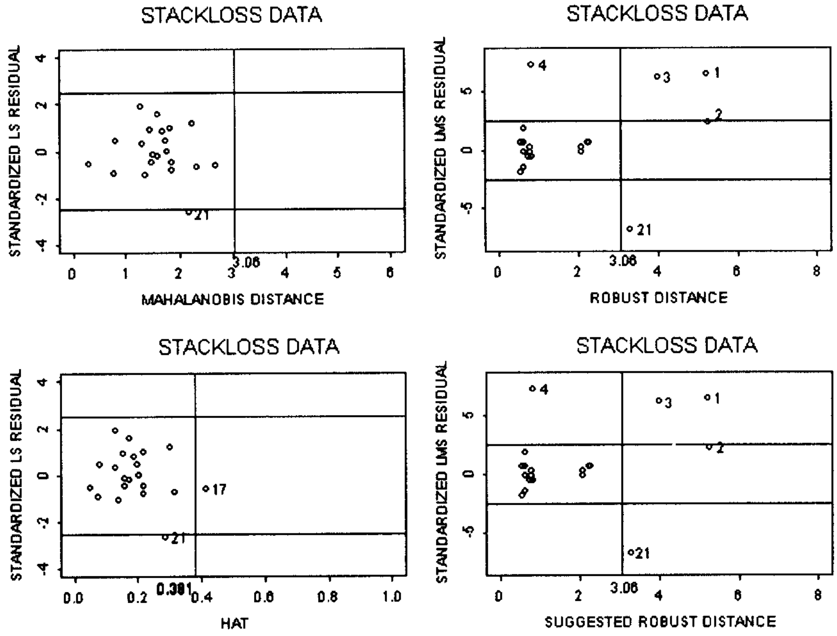
Table 1. MD_i, RD_i, p_{ii}, SRD_i for stackloss data

case	MD_i	RD_i	p_{ii}	SRD_i	case	MD_i	RD_i	p_{ii}	SRD_i
1	2.25	5.23	0.30	5.23	12	1.84	0.79	0.22	0.79
2	2.32	5.27	0.32	5.27	13	1.48	0.55	0.16	0.55
3	1.59	4.01	0.17	4.01	14	1.78	0.64	0.21	0.64
4	1.27	0.84	0.13	0.84	15	1.69	2.23	0.19	2.23
5	0.30	0.80	0.05	0.80	16	1.29	2.11	0.13	2.11
6	0.77	0.78	0.08	0.78	17	2.70	2.07	0.41	2.07
7	1.85	0.64	0.22	0.64	18	1.50	2.09	0.16	2.09
8	1.85	0.64	0.22	0.64	19	1.59	2.29	0.17	2.29
9	1.36	0.83	0.14	0.83	20	0.81	0.64	0.08	0.64
10	1.75	0.64	0.20	0.64	21	2.18	3.30	0.28	3.30
11	1.47	0.58	0.16	0.58					

<Figure 1> reveals four plots of the standardized LS residuals with the Mahalanobis

distance MD_i's, the standardized LMS residuals with the robust distance RD_i's, the standardized LS residuals with the diagonal elements of hat matrix p_{ii} and the standardized LMS residuals with the suggested robust distance SRD_i,'s respectively. According to the <Figure 1>, the MD_i and the p_{ii} have failed to find leverage points and outliers. But the RD_i and the SRD_i analyze exactly and point out that there are a good leverage point (observation 2), three bad leverage points(observation 1, 3, 21) and one outlier (observation 4).

Figure 1. Diagnostics plots of stackloss data



3.2 Artificial Data(Hawkins,Bradu,Kass(1984), reused in Rousseeuw and Leroy(1987))

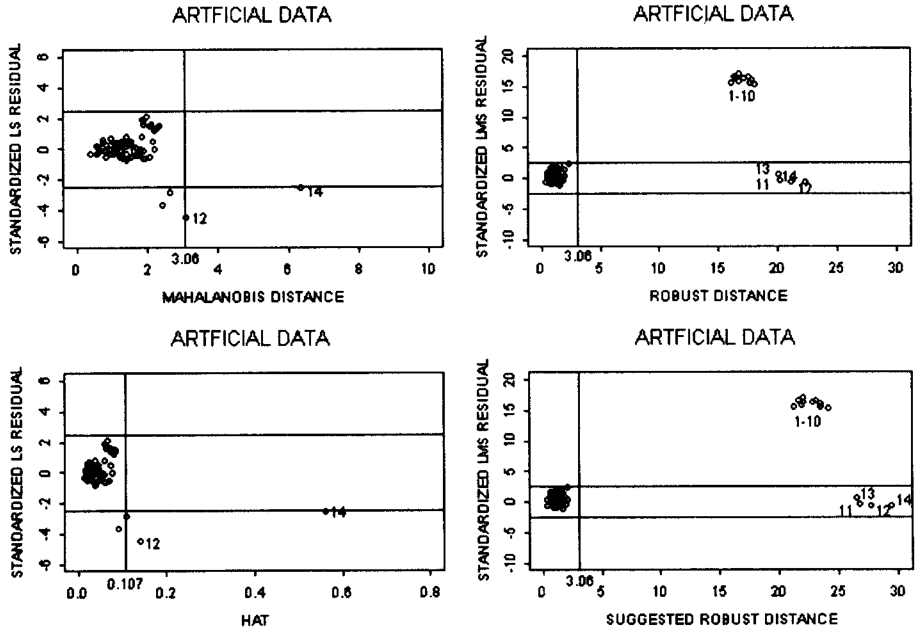
The artificial data set consists of 75 observations on three explanatory and one response variables. And it is known that it contains ten bad leverage points (1 ,2, 3, 4, 5, 6, 7, 8, 9, 10) and four good leverage points (11, 12, 13, 14). <Table 2> reveals MD_i, RD_i, p_{ii} and SRD_i of artificial data. In <table 2>, the MD_i and the p_{ii} say that two points(observation 12, 14) and three points(observation 12, 13, 14) are leverage points respectively. But the robust distances RD_i points out fourteen leverage points(observation 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14). To compute the RD_i, Rousseeuw and von Zomeren(1990) computed ${}_{75}C_4$ vales

<Table 2> MD_i, RD_i, p_{ii}, SRD_i for artificial data

case	MD_i	RD_i	p_{ii}	SRD_i	case	MD_i	RD_i	p_{ii}	SRD_i
1	1.92	16.20	0.06	21.32	39	1.27	1.34	0.03	1.21
2	1.86	16.62	0.06	22.07	40	1.11	0.55	0.03	0.80
3	2.32	17.65	0.09	23.21	41	1.70	1.48	0.05	1.24
4	2.23	18.18	0.08	24.17	42	1.77	1.74	0.06	1.59
5	2.10	17.82	0.07	23.58	43	1.87	1.18	0.06	1.61
6	2.15	16.80	0.08	22.01	44	1.42	1.82	0.04	1.52
7	2.01	16.82	0.07	22.13	45	1.08	1.25	0.03	1.31
8	1.92	16.44	0.06	21.69	46	1.34	1.70	0.04	1.38
9	2.22	17.71	0.08	23.53	47	1.97	1.65	0.07	1.75
10	2.33	17.21	0.09	22.87	48	1.42	1.37	0.04	1.16
11	2.45	20.23	0.09	26.87	49	1.57	1.27	0.05	1.02
12	3.11	21.14	0.14	27.82	50	0.42	0.83	0.02	1.12
13	2.66	20.16	0.11	26.56	51	1.30	1.19	0.04	1.26
14	6.38	22.38	0.56	29.47	52	2.08	1.61	0.07	1.57
15	1.82	1.54	0.06	1.26	53	2.21	2.41	0.08	2.13
16	2.15	1.88	0.08	1.81	54	1.41	1.26	0.04	1.41
17	1.38	1.03	0.04	1.20	55	1.23	0.66	0.03	0.99
18	0.85	0.73	0.02	0.45	56	1.33	1.21	0.04	1.24
19	1.15	0.59	0.03	1.08	57	0.83	0.93	0.02	0.96
20	1.59	1.49	0.05	1.31	58	1.40	1.31	0.04	1.23
21	1.09	0.87	0.03	0.65	59	0.59	0.96	0.02	1.16
22	1.55	0.90	0.05	1.50	60	1.89	1.89	0.06	1.98
23	1.09	0.94	0.03	0.93	61	1.67	1.31	0.05	2.07
24	0.97	0.83	0.03	0.99	62	0.76	1.22	0.02	1.23
25	0.80	1.26	0.02	1.24	63	1.29	1.17	0.04	1.34
26	1.17	0.86	0.03	1.52	64	0.97	1.14	0.03	1.19
27	1.45	1.35	0.04	1.19	65	1.15	1.40	0.03	1.08
28	0.87	1.00	0.02	0.65	66	1.30	0.78	0.04	0.96
29	0.58	0.72	0.02	0.68	67	0.63	0.37	0.02	0.42
30	1.57	1.97	0.05	1.76	68	1.55	1.64	0.05	1.27
31	1.84	1.43	0.06	1.35	69	1.07	1.17	0.03	1.34
32	1.31	0.95	0.04	1.22	70	1.00	1.04	0.03	1.35
33	0.98	0.73	0.03	0.98	71	0.64	0.64	0.03	0.69
34	1.18	1.42	0.03	1.18	72	1.05	0.52	0.03	0.77
35	1.24	1.26	0.03	1.14	73	1.47	1.14	0.04	1.24
36	0.85	0.86	0.02	1.16	74	1.65	0.96	0.05	1.26
37	1.83	1.26	0.06	1.88	75	1.90	1.99	0.06	1.66
38	0.75	0.92	0.02	1.10					

of $m_j^p det(V(Z)_j)$ and it is very hard to compute. When we apply this data to suggested algorithm, we obtain the clean subset of M={31, 32, 63, 71, 72, 20, 18, 35, 34, 40, 25, 56, 45, 58, 48, 19, 61, 55, 28, 30, 17, 75, 59, 66, 22, 46, 41, 37, 69, 65, 39, 42, 52, 33, 67, 74, 73, 29, 16, 50} and the calculation is reduced to ${}_{39}C_4$ times. Finally we find one subset {25, 56, 41, 73} with the minimum volume ellipsoids. The suggested robust SRD_i which is based on this subset points out fourteen leverage points(observation 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14). So, we have the same result but the calculations are reduced from ${}_{75}C_4$ to ${}_{39}C_4$. In this data, the RD_i and the SRD_i identify the leverage points correctly, but the MD_i and the h_{ii} fail to identify the leverage points or outliers. As shown in <Figure 2>. the RD_i and the SRD_i point out that there are four good leverage point(observation 11, 12, 13, 14) and ten bad leverage points (observation 1, 2, 3, 4, 5, 6, 7, 8, 9, 10).

Figure 2. Diagnostics plots of artificial data



4. Conclusion

Multiple leverage points and outliers deserve special attention as they often provide important clues about the model building and process under study. Thus regression diagnostics are very important to analyze several data correctly. In this paper, we attempted to prove effectiveness of the suggested algorithm in identifying and testing multiple leverage points and outliers through two examples. As a result, the suggested robust distance reduces the computational difficulty of Rousseeuw and van Zeremen's method and has good powers in identifying leverage points.

References

[1] Easton, G.S. (1994). A Simple Dynamic Graphical Diagnostics Method for Almost Any Model, *Journal of American Statistical Association*, 89, 201-207.

[2] Hadi, A.S. (1992). Identifying Multiple Outliers in Multivariate Data, *Journal of the Royal Statistical Society*, Ser.B, 54, 761-771.

[3] Hadi, A.S. and Simonoff, J.S. (1993). Procedures for the Identification of Multiple Outliers in Linear Models, *Journal of American Statistical Association*, 75, 1264-1272.

[4] Hoaglin, D.C. and Welsch, R.E. (1978) The hat matrix in regression and ANOVA. *American Statistician* 32 17-22.

[5] Rousseeuw, P.J. (1985) Multivariate estimation with high breakdown points. In *Mathematical Statistics and applications* (eds W. Grossman, G. Pflug, I. Vincze and W. Wertz), vol. B, pp.283-297. Dordrecht: Reidel.

[6] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, John Wiley & Sons.

[7] Rousseeuw, P.J. and van Zomeren, B.C.(1990) Unmasking multivariate outliers and leverage points(with comments). *Journal of the American Statistical Association,* 75, 633-651.

[8] Yoo, J.Y. and Kim. H.C.(1996) A Study of Hadi and Simonoff's Identifying Multiple Outliers Method. *The Korean Communications in Statistics,* Vol 3, 3, 11-23.

[9] Yoo. J.Y and Ahn. K.S. and Huh, M.Y.(1997) Dynamic Graphics Approach for Regression Diagnostics System(REDS), *The Korean Journal of Applied Statistics,* vol 10, 2, 241-251.