The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 291-304

A Hybrid Algorithm for Identifying Multiple Outlers in Linear Regression[1)]

Bu-yong Kim[2)], Hee-young Kim[3)]

Abstract

This article is concerned with an effective algorithm for the identification of multiple outliers in linear regression. It proposes a hybrid algorithm which employs the least median of squares estimator, instead of the least squares estimator, to construct an initial clean subset in the stepwise forward search scheme. The performance of the proposed algorithm is evaluated and compared with the existing competitor via an extensive Monte Carlo simulation. The algorithm appears to be superior to the competitor for the most of scenarios explored in the simulation study. Particularly it copes with the masking problem quite well. In addition, the orthogonal decomposition and its updating techniques are considered to improve the computational efficiency and numerical stability of the algorithm.

1. Introduction

It is advisable to assess the presence or absence of any outliers, and then to identify them correctly prior to the regression analysis. A number of procedures have been developed for the identification of regression outliers. Some of the direct procedures are mainly based on the residuals from the least squares (LS) fit. For instance, Tietjen *et al.* (1973) suggest using the maximum of the absolute residuals as a diagnostic measure, and provide critical values based on the simulation. But outliers may cause a poor LS fit and hence distort the residuals since the LS estimation accommodates the outlying observations at the expense of other points. Other procedures are based on the principle of deleting one observation at a time. Those procedures, however, work well only when the data set contains single outlier because they are affected by the outliers which are to be identified. Marasinghe (1985) employs the sequential application of deleting-one scheme to choose candidate outliers. But it requires to determine the number of outliers in advance and suffers from the masking and swamping effects. Kianifard and Swallow (1990) propose a forward search procedure which is based on a deleting-one method and recursive residuals. However it turns out that the procedure is also

1) This Research was supported by the Sookmyung Women's University Special Research Grants
2) Department of Statistics, Sookmyung Women's University, Yongsan-gu, Seoul, Korea
3) Strategy and Innovation Dept., Kookmin Credit Card Co. Ltd, Jongro-gu, Seoul, Korea

susceptible to the masking effect.

The principle of single-deletion has been extended to the procedures for multiple outliers to cope with the masking and swamping problems. The most common procedures for the multiple-deletion construct all possible subsets of observations and test whether the observations in a subset are outlying significantly relative to the remaining observations. However, the drawback of them is that a great deal of computation is required since they have to check all possible subsets.

On the other hand, the indirect procedures based on the robust estimators have been suggested for the identification of outliers (see e.g. Rousseeuw and Leroy (1987)). Rousseeuw and Zomeren (1990) suggest a diagnostic measure that is based on the Mahalanobis-type robust distance with the minimum covariance determinant (MCD) estimators of the mean and the covariance matrix. They propose a display in which the standardized residuals from a high breakdown point fit, such as the least median of squares (LMS) and the least trimmed sum of squares (LTS), are plotted versus the robust distances. The plot makes it possible to classify the data into regular observations, vertical outliers, and bad leverage points. However, the cutoff values are chosen by somewhat arbitrary way since the distribution of LMS residuals is not known for small samples. Also, this approach tends to swamp too many of observations because of its high robustness.

Hadi and Simonoff (1993) suggest an effective direct algorithm H-S which employs a forward search scheme. The algorithm starts with constructing a basic clean subset with the observations corresponding to the smallest absolute value of the adjusted residuals from the LS fit. Then the basic subset is iteratively increased, on the basis of the absolute scaled residuals, to the initial clean subset of specified size. After an initial subset is formed, this algorithm updates the subset iteratively until the final clean subset is obtained. This uses the internally studentized residual and the scaled prediction error as diagnostics for reforming the subset, and tests the outlyingness of the observations relative to the clean subset. However, H-S appears to fail in the presence of multiple outliers which are clustered in an outlying cloud far from the main bulk of the data.

This article is mostly concerned with an effective identification algorithm which is improved in terms of the computational efficiency and numerical stability. As an attempt to modify H-S algorithm, we suggest a hybrid algorithm which constructs the initial clean subset based on the LMS estimator, and then increases the clean subset size one at a time by either adding one observation or exchanging several observations. In Section 3, the updating techniques of the orthogonal decomposition are suggested to reduce the computational burden that may be increased rapidly with the data size. The performance of the proposed algorithm is evaluated and compared with H-S algorithm in Section 4.

2. Proposed Algorithm

Consider the standard linear regression model $\boldsymbol{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is an n vector of responses, X is an $n\times p$ matrix of regressor variables with intercept, $\boldsymbol{\beta}$ is a p vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an n vector of random errors.

Construction of initial clean subset : Since the algorithm H-S is totally based on the LS estimator, it is not guaranteed to cope with the masking problem when a data set has many severe outliers. In particular, the basic or initial clean subset may not be constructed appropriately since the LS fit is strongly affected by the outliers. It is found, in several data sets with multiple outliers, that H-S can not detect the outliers correctly. Consequently, the central question is whether more clean initial subset can be obtained even if the masking and swamping effects are present in the data set. Thus, a new algorithm is instead proposed for constructing the initial clean subset.

The proposed algorithm is a variant of H-S algorithm, but it employs the LMS estimator which has very high breakdown point. (Statistical properties and algorithms of the LMS estimator can be found in Rousseeuw (1984), Basset (1991), and Kim (1996).) The algorithm employs a single-phase approach to find the initial clean subset, whereas H-S algorithm uses a two-phase approach. Let C_0 denote the set of indices of the observations in an initial clean subset that can be presumed to be free of outliers. While the H-S algorithm starts with a basic subset of size $p+1$ and increases the subset by one observation until it constructs the initial clean subset of size $[(n+p-1)/2]$, the proposed algorithm constructs the initial subset C_0 of size $n-[n/2]+p-1$ corresponding to the smallest absolute standardized residuals from the LMS fit. (The initial subset size is chosen on the basis of the exact fit property of the LMS estimator.) This approach makes it possible to form the appropriate clean subset even though the masking and swamping effects are serious.

Test for outlyingness : After an initial clean subset is constructed, this algorithm performs a stepwise forward search, which is suggested by Hadi and Simonoff (1993), to separate the data into a subset of clean observations and a subset of potential outliers. That is, it updates the subset C_0 iteratively until the final clean subset is obtained. The internally studentized residuals and the scaled prediction errors based on the clean subset are used as criteria for the inclusion and deletion of observations. Let X_C and $\boldsymbol{y}_C$ be the current subset of observations indexed by C. And provided that X_C is of full column rank, let $\widehat{\boldsymbol{\beta}_C}$ and $\widehat{\sigma_C^2}$ be the LS estimator of $\boldsymbol{\beta}$ based on the observations in the subset X_C and $\boldsymbol{y}_C$ and the corresponding residual mean squares, respectively. (If the subset X_C is not of full column rank, we may increase the subset by adding observations according to their ordered

diagnostics until it becomes a matrix of full rank.) Tests are performed whether the potential outliers are significantly outlying relative to the clean subset as follows. For the observations in the clean subset C, the internally studentized residuals $|y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_C| / [\hat{\sigma}_C \{ 1 - \boldsymbol{x}_i'(X_C'X_C)^{-1} \boldsymbol{x}_i \}^{1/2}]$ are compared with the critical value. On the other hand, for the observations in the potential outlier subset $\overline{C}$, the scaled prediction errors $|y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_C| / [\hat{\sigma}_C \{ 1 + \boldsymbol{x}_i'(X_C'X_C)^{-1} \boldsymbol{x}_i \}^{1/2}]$ are compared with the same critical value. It is known that those statistics follow t distribution with $c-p$ degrees of freedom, where c denotes the size of current subset C. Therefore, utilizing the Bonferroni approach we can set the percentile $t_{a/2(c+1)}(c-p)$ as the critical value for the $(c+1)$th largest diagnostic. If the diagnostic is not significant, then one observation corresponding to the $(c+1)$th largest diagnostic is chosen to be included in the new clean subset. At the final iterate, the observations which are included in the complementary subset $\overline{C}$ are regarded as the outliers. Of course, if the final subset $\overline{C}$ is a null set, then we declare that the data set has no outliers. The detail steps of the proposed algorithm K-K are illustrated in the following.

Algorithm K-K : <*Step 1*> Find the $n-[n/2]+p-1$ smallest absolute standardized residuals from the LMS fit. Then construct an initial clean subset C_0 with the corresponding indices of the observations. Set $C = C_0$ and $c = n-[n/2]+p-1$.

<*Step 2*> Compute the following diagnostic measures from the LS fit,

$$\xi_i = \begin{cases} |e_{C_i}| / \{ \hat{\sigma}_C (1-h_{C_i})^{1/2} \} & \text{for } i \in C \\ |e_{C_i}| / \{ \hat{\sigma}_C (1+h_{C_i})^{1/2} \} & \text{for } i \in \overline{C} \end{cases},$$

where $e_{C_i} = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_C$ and $h_{C_i} = \boldsymbol{x}_i'(X_C'X_C)^{-1} \boldsymbol{x}_i$.

<*Step 3*> Arrange the observations in ascending order according to ξ_i. Let $\xi^*_{(c+1)}$ denote the $(c+1)$th largest value of the ξ_i. If $\xi^*_{(c+1)} \geq t_{a/2(c+1)}(c-p)$, then declare all observations corresponding to ξ_j $(j \geq c+1)$ as outliers and stop. Otherwise, go to step 2 with a new subset C including the index corresponding to $\xi^*_{(c+1)}$. Stop if $c+1 = n$, and declare that no outliers are in the data set.

3. Computational Aspects

The orthogonal decomposition approach can be employed to deal with the numerical

instability problem which may occur at each iterate of the algorithm K-K as well as H-S. The full column matrix X_C can be decomposed as $X_C = Q\,[\,T' \vdots O'\,]'$, where Q is a $c \times c$ orthogonal matrix, T is a $p \times p$ upper triangular matrix, and O is a $(c-p) \times p$ zero matrix. Let $Q'\,\boldsymbol{y}_C = [\,\boldsymbol{a}_1' \vdots \boldsymbol{a}_2'\,]'$, where $\boldsymbol{a}_1$ is a p vector and $\boldsymbol{a}_2$ is a $c-p$ vector, and let $W = T^{-1}$, $\boldsymbol{w}_i = W'\,\boldsymbol{x}_i$ for simplicity. Then the statistics at the second step of the algorithm can be computed as follows,

$$\widehat{\beta_C} = W\boldsymbol{a}_1, \quad \widehat{\sigma_C} = \{\ \boldsymbol{a}_2'\,\boldsymbol{a}_2/(c-p)\ \}^{1/2},$$

$$e_{C_i} = \begin{cases} Q\,[\,\boldsymbol{0}' \vdots \boldsymbol{a}_2'\,]', & h_{C_i} = \boldsymbol{w}_i'\,\boldsymbol{w}_i \quad \text{for } i \in C \\ y_i - \boldsymbol{x}_i' W \boldsymbol{a}_1, & h_{C_i} = \boldsymbol{x}_i WW'\,\boldsymbol{x}_i \quad \text{for } i \in \overline{C} \end{cases},$$

where $\boldsymbol{0}$ is a p vector of zeros.

However, a great deal of computation is required since the matrix Q and T have to be computed at each iterate of the algorithm. Moreover, the computational cost may escalate with the size of data set. Therefore, it is necessary to make some improvement in the computational efficiency of the algorithm. Kim and Kim (1997) have found that a few observations are changed at the forward search scheme, moreover at many iterates only one observation is added to the previous clean subset. Thus, in order to improve its computational efficiency, we may employ the updating techniques of the orthogonal decomposition in two cases of the deletion and addition of observations.

Addition of observations : Let X_+ denote new matrix in which a row is augmented to the matrix X_C, and define the orthogonal matrix Q_+ as follows

$$X_+ = \begin{bmatrix} X_C \\ \cdots\cdots \\ \boldsymbol{x}'_{c+1} \end{bmatrix}, \quad Q_+' = \begin{bmatrix} Q' \vdots \boldsymbol{0} \\ \cdots\cdots\cdots \\ \boldsymbol{0}' \vdots 1 \end{bmatrix}, \quad Q_+'X_+ = \begin{bmatrix} R \\ \cdots\cdots \\ \boldsymbol{x}'_{c+1} \end{bmatrix},$$

where $R = [\,T' \vdots O'\,]'$, and $\boldsymbol{0}$ is a c vector of zeros. Now by applying Givens rotations to the elements in the last row of Q_+X_+, we can obtain upper trapezoidal matrix can be updated as follows

$$G_{c+1,p} \cdots G_{c+1,1}\,Q_+'X_+ = \begin{bmatrix} \widetilde{R} \\ \cdots\cdots \\ \boldsymbol{0}' \end{bmatrix},$$

where $G_{i,j}$ denotes the Givens rotation matrix operating on rows i and j. Therefore, the orthogonal matrix and upper trapezoidal matrix can be updated as follows

$$\widetilde{Q}' = G_{c+1,p} \cdots G_{c+1,1}\,Q_+',$$

$$\widetilde{R} = G_{c+1,p} \cdots G_{c+1,1} \begin{bmatrix} R \\ \cdots\cdots \\ \boldsymbol{x}'_{c+1} \end{bmatrix}.$$

Additionally, this updating approach can be easily extended to the case of addition of multiple observations. Suppose s rows are added and they are augmented to the matrix R accordingly. To make the matrix upper trapezoidal, Givens rotations have to be operated on the rows. Then by left-multiplying

$$Q_{++}' = \begin{bmatrix} Q'_{c\times c} & \vdots & O_{c\times s} \\ \cdots\cdots & & \cdots\cdots \\ O_{s\times c} & \vdots & I_{s\times s} \end{bmatrix}$$

by the rotation matrices, we can obtain updated matrices as follows.

$$\widetilde{Q}' = (G_{c+s,p} \cdots G_{c+s,1}) \cdots (G_{c+1,p} \cdots G_{c+1,1})\, Q_{++}',$$

$$\widetilde{R} = (G_{c+s,p} \cdots G_{c+s,1}) \cdots (G_{c+1,p} \cdots G_{c+1,1}) \begin{bmatrix} R \\ \cdots\cdots \\ \boldsymbol{x}'_{c+1} \\ \vdots \\ \boldsymbol{x}'_{c+s} \end{bmatrix}.$$

Deletion of observations : Let X_- denote the submatrix reformed after the last row $\boldsymbol{x}_c'$ is deleted from the matrix X_C. Then the matrices Q and R are partitioned

$$\begin{bmatrix} Q_1 & \vdots & \boldsymbol{b} \\ \cdots\cdots & & \cdots\cdots \\ \boldsymbol{a}' & \vdots & h \\ \cdots\cdots & & \cdots\cdots \\ Q_2 & \vdots & \boldsymbol{d} \end{bmatrix} \begin{bmatrix} X_- \\ \cdots\cdots \\ \boldsymbol{x}_c' \end{bmatrix} = \begin{bmatrix} T \\ \cdots\cdots \\ \mathbf{0}' \\ \cdots\cdots \\ O \end{bmatrix},$$

where Q_1 is a $p\times(c-1)$ matrix, Q_2 is a $(c-p-1)\times(c-1)$ matrix, $\boldsymbol{b}$ is a p vector, $\boldsymbol{d}$ is a $c-p-1$ vector, $\boldsymbol{a}$ is a $c-1$ vector, and h is a scalar. By a Householder transformation, $\boldsymbol{d}$ is transformed to zero vector and hence $\boldsymbol{a}'$, h and Q_2 are modified accordingly, but Q_1, $\boldsymbol{b}$, and the right-side are not changed as follows,

$$\begin{bmatrix} Q_1 & \vdots & \boldsymbol{b} \\ \cdots\cdots & & \cdots\cdots \\ \widehat{\boldsymbol{a}}' & \vdots & \hat{h} \\ \cdots\cdots & & \cdots\cdots \\ \widehat{Q}_2 & \vdots & \mathbf{0} \end{bmatrix} \begin{bmatrix} X_- \\ \cdots\cdots \\ \boldsymbol{x}_c' \end{bmatrix} = \begin{bmatrix} T \\ \cdots\cdots \\ \mathbf{0}' \\ \cdots\cdots \\ O \end{bmatrix}.$$

Next by applying Givens rotations, $\boldsymbol{b}$ is transformed into zero vector and hence Q_1, $\widehat{\boldsymbol{a}}'$, $\hat{h}$, T and $\mathbf{0}'$ are modified accordingly, but $\widehat{Q}_2$ and O are unchanged as

$$\begin{bmatrix} Q_1^* & \vdots & 0 \\ \cdots & \cdots & \cdots \\ \boldsymbol{a}^{*\prime} & \vdots & h^* \\ \cdots & \cdots & \cdots \\ \widehat{Q}_2 & \vdots & 0 \end{bmatrix} \begin{bmatrix} X_- \\ \cdots \\ \boldsymbol{x}_c' \end{bmatrix} = \begin{bmatrix} T^* \\ \cdots \\ \boldsymbol{w}' \\ \cdots \\ O \end{bmatrix}.$$

Since the left-side matrix is also orthogonal, the rows and columns are clearly of unit Euclidean length. Therefore $h^* = \pm 1$ and $\boldsymbol{a}^{*\prime} = \mathbf{0}'$, that is,

$$\begin{bmatrix} Q_1^* & \vdots & 0 \\ \cdots & \cdots & \cdots \\ \mathbf{0}' & \vdots & \pm 1 \\ \cdots & \cdots & \cdots \\ \widehat{Q}_2 & \vdots & 0 \end{bmatrix} \begin{bmatrix} X_- \\ \cdots \\ \boldsymbol{x}_c' \end{bmatrix} = \begin{bmatrix} T^* \\ \cdots \\ \pm \boldsymbol{x}_c' \\ \cdots \\ O \end{bmatrix}.$$

Now removing the $(p+1)$th column and $(c+1)$th row from the orthogonal matrix of left-side, and also removing the $(p+1)$th column from the right-side matrix, we obtain the updated matrices as $\widetilde{Q}' = \begin{bmatrix} Q_1^* \\ \cdots \\ \widehat{Q}_2 \end{bmatrix}$, $\widetilde{R} = \begin{bmatrix} T^* \\ \cdots \\ O \end{bmatrix}$,

where $Q_1^* = G_{p,c} \cdots G_{1,c} \cdot Q_1$, $T^* = G_{p,c} \cdots G_{1,c} \cdot T$, $\widehat{Q}_2 = H_{p+1(c)} Q_2$, and $H_{p+1(c)}$ denotes the Householder transformation operated on the c-th column with pivoting $(p+1)$th position.

In addition, when multiple rows are deleted from the matrix X_C, the updating of orthogonal transformation can be performed by operating the above steps to the rows repeatedly.

4. Comparisons of Performance

Hadi and Simonoff (1993) have conducted Monte Carlo experiments to compare the power of the procedures, and shown that their algorithm with method M1 is the most effective one for the identification of outliers. Recently, Wisnowski *et al.* (2001) validate that H-S algorithm is effective when the residual outlying distance is large. In this article we also perform Monte Carlo simulation studies to figure out the effectiveness of the proposed algorithm. The performance of the algorithm is evaluated and compared with H-S on both its ability to identify the outliers and the probability of swamping. Three measures of performance are considered: $p_1 = P$(all outliers are exactly identified), $p_2 = P$(at least one of the outliers is identified), $p_3 = P$(at least one inlying observation is wrongly identified as an outlier). The probability of masking is $1 - p_2$, whereas the probability of swamping is p_3 that is also called the false alarm rate. Thus excellent performance of an algorithm is indicated by large value of p_1 and p_2, and by small value of p_3.

We take the total number of observations $n=25, 35, 45$ and the number of regressor variables $p=2, 3$ in the data set. For each value of n and p, 1000 data sets are generated from the specified distributions. The simulation considers various data configurations such as the density and geometry of outliers, number of observations, and dimension of regressor variables. In order to generate the data sets with various outliers, z($z=1, 3, 7$ for $n=25$; $z=1, 5, 9$ for $n=35$; $z=5, 9, 11$ for $n=45$) observations are contaminated by artificial values according to the predetermined rules. In the tables and figures, the acronyms HV, LV, HL, LL, HLV, and LLV represent, respectively, the scenarios in which the data set has high vertical outliers, low vertical outliers, high leverage points, low leverage points, high leverage points with high vertical outliers, and low leverage points with high vertical outliers. The test for outlyingness is performed at 5% significance level. However, for comparison purposes, the critical values of the algorithms are slightly adjusted to ensure that the expected false alarm probability is close to 0.05 under the null hypothesis of no outliers.

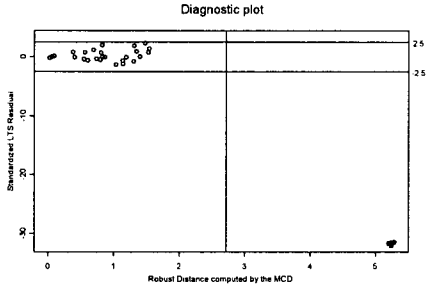
Code development and Monte Carlo simulations are conducted in SAS/IML, and the figures are obtained by S-Plus. The simulation results presented in Table 3 - Table 5 are summaries based on 1000 runs for each scenario.

Simple model case : For each data set, the regressors and errors are generated, respectively, from the $U(0,20)$ and $N(0,1)$ distributions with the randomly selected seed, 8940(n=25), 8940(n=35), and 123(n=45). Then the responses are obtained according to the linear regression model with the coefficients arbitrarily selected to be 0.0 for the intercept and 1.0 for the regressor variable. In order to plant the vertical outliers and bad leverage points, z inlying observations are replaced by the shifted values according to the rules in Table 1. To confirm that the artificial outliers are planted appropriately, the diagnostic plots (Fig. 1 - Fig. 4) suggested by Rousseeuw and Zomeren (1990) are investigated for a data set with 7 outliers out of 35 observations. The pilot studies illustrate that the rules confirm well to the four outlier scenarios.

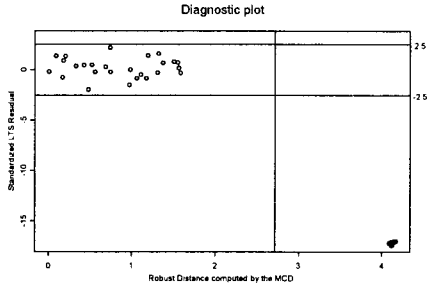
<Table 1> Rules for planting outliers : simple model case

Scenario	x-shift ($i=1,\ldots,z$)	y-shift ($i=1,\ldots,z$)
HL	$x_i = 40-0.06(i-1)$	$y_i = 10-0.01z(i+1)$
LL	$x_i = 35-0.06(i-1)$	$y_i = 16-0.01z(i+1)$
HV	$x_i = 3-0.06(i-1)$	$y_i = x_i+15$
LV	$x_i = 3-0.06(i-1)$	$y_i = x_i+7$

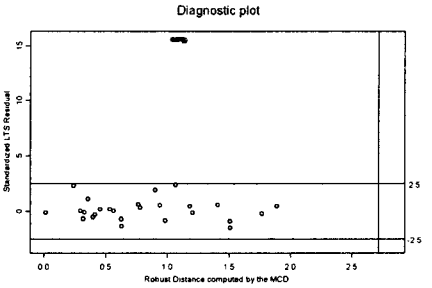
Multiple model case : The data sets for the multiple model case are generated in a similar manner. The number of regressor variables p is set to 3 with $n=25,\ 35,\ 45$. The values of two regressors and errors are generated from the $U(0,20)$, $N(0,3)$ and $N(0,1)$ distributions, respectively. Seeds randomly chosen for X_1 are 91590(n=25), 98755(n=35), 8940(n=45), and those for X_2 are 33333(n=25), 1231(n=35), 5643(n=45). The regression coefficients are arbitrarily specified as $\beta_0=0.0,\ \beta_1=1.0,\ \beta_2=1.0$ to generate the responses. Also to contaminate the data



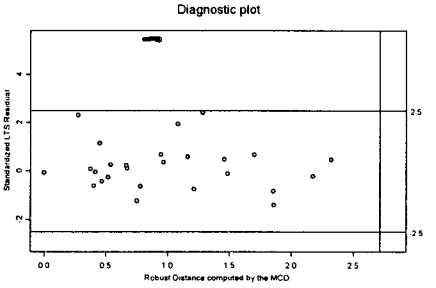
<Fig. 1> Diagnostic plot for HL



<Fig. 2> Diagnostic plot for LL



<Fig. 3> Diagnostic plot for HV

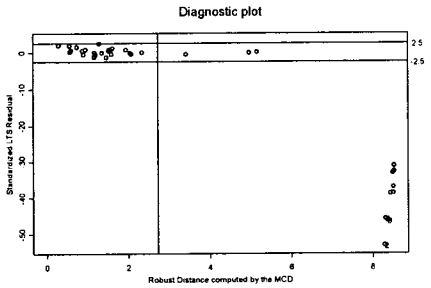


<Fig. 4> Diagnostic plot for LV

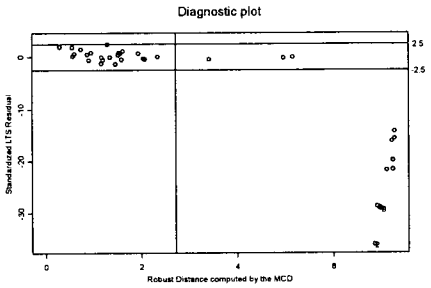
sets with vertical outliers and bad leverage points, z observations are shifted away from the group of inlying observations according to the rules in Table 2. It can be found from the diagnostic plots (Fig. 5 - Fig. 8) for a data set of each scenario that the outliers are planted in the appropriate way.

<Table 2> Rules for planting outliers : multiple model case

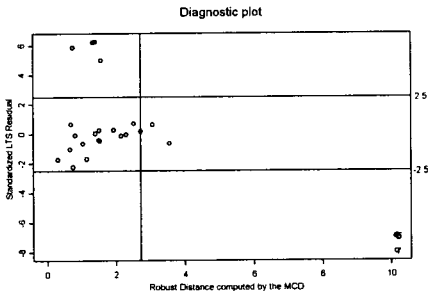
	HL	LL	HLV	LLV
x-shift	$x_{1i}=25-0.05(i-1)$ $x_{2i}=15+0.03(i-1)$ $(i=1,\ldots,z)$	$x_{1i}=40-0.05(i-1)$ $x_{2i}=20+0.03(i-1)$ $(i=1,\ldots,z)$	L/ $x_{1i}=40+0.03(i-1)$ $(i=1,\ldots,[z/2]+1)$ V/ $x_{1i}=7+0.03(i-1)$ $(i=[z/2]+2,\ldots,z)$	L/ $x_{1i}=32+0.03(i-1)$ $(i=1,\ldots,[z/2]+1)$ V/ $x_{1i}=7+0.03(i-1)$ $(i=[z/2]+2,\ldots,z)$
y-shift	-	-	L/ $y_i=6-0.05(i+1)$ $(i=1,\ldots,[z/2]+1)$ V/ $y_i=x_i+25$ $(i=[z/2]+2,\ldots,z)$	L/ $y_i=6-0.05(i+1)$ $(i=1,\ldots,[z/2]+1)$ V/ $y_i=x_i+15$ $(i=[z/2]+2,\ldots,z)$



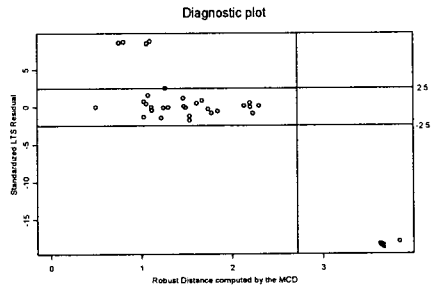
<Fig. 5> Diagnostic plot for HL



<Fig. 6> Diagnostic plot for LL



<Fig. 7> Diagnostic plot for HLV



<Fig. 8> Diagnostic plot for LLV

Summary of simulation results : The results of Monte Carlo simulations are summarized in Table 3 - Table 5. The estimates of p_1, p_2, and p_3 are denoted, respectively, by $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$, which are the averages of proportions over 3 outlier configurations. (The complete and detail tables are available from the authors.) The performance of the proposed algorithm appears to be very similar in both the simple model case and the multiple model case. The results indicate that the algorithm K-K has excellent capability and is superior to the H-S algorithm for the most of the scenarios considered in the experiments. In particular, the superiority of K-K increases as the percentage of outliers increases. Also the results demonstrate that K-K is quite more powerful than H-S in case of high leverage and/or high vertical outlier scenarios, and is less affected by the masking effects. In addition, the algorithm outperforms as the number of observations increases. On the other hand, it turns out that the proposed algorithm is less effective than H-S for the swamping problem since it is partly based on the high breakdown estimator. (Wisnowski *et al.* (2001) also note that H-S has the unusually low false alarm probability.) However, the degree of ineffectiveness of K-K is relatively low in many scenarios. In summary, the simulation studies indicate that the proposed algorithm performs better than H-S in general.

<Table 3> Estimated measures of performance ($n=25$)

Simple model				Multiple model			
Scenario	Measure	H-S	K-K	Scenario	Measure	H-S	K-K
HL	$\hat{p}_1$	0.727	0.903	HL	$\hat{p}_1$	0.648	0.915
	$\hat{p}_2$	0.782	0.973		$\hat{p}_2$	0.718	1.000
	$\hat{p}_3$	0.111	0.097		$\hat{p}_3$	0.062	0.085
LL	$\hat{p}_1$	0.716	0.872	LL	$\hat{p}_1$	0.680	0.915
	$\hat{p}_2$	0.772	0.941		$\hat{p}_2$	0.786	1.000
	$\hat{p}_3$	0.104	0.123		$\hat{p}_3$	0.065	0.085
HV	$\hat{p}_1$	0.886	0.922	HLV	$\hat{p}_1$	0.606	0.872
	$\hat{p}_2$	0.951	0.993		$\hat{p}_2$	0.788	0.985
	$\hat{p}_3$	0.088	0.078		$\hat{p}_3$	0.068	0.093
LV	$\hat{p}_1$	0.875	0.827	LLV	$\hat{p}_1$	0.586	0.715
	$\hat{p}_2$	0.933	0.894		$\hat{p}_2$	0.672	0.825
	$\hat{p}_3$	0.078	0.132		$\hat{p}_3$	0.078	0.118

<Table 4> Estimated measures of performance ($n=35$)

Simple model				Multiple model			
Scenario	Measure	H-S	K-K	Scenario	Measure	H-S	K-K
HL	$\hat{p}_1$	0.690	0.944	HL	$\hat{p}_1$	0.543	0.948
	$\hat{p}_2$	0.730	0.995		$\hat{p}_2$	0.575	1.000
	$\hat{p}_3$	0.071	0.056		$\hat{p}_3$	0.043	0.052
LL	$\hat{p}_1$	0.683	0.934	LL	$\hat{p}_1$	0.467	0.948
	$\hat{p}_2$	0.721	0.985		$\hat{p}_2$	0.606	1.000
	$\hat{p}_3$	0.067	0.064		$\hat{p}_3$	0.035	0.052
HV	$\hat{p}_1$	0.907	0.949	HLV	$\hat{p}_1$	0.736	0.935
	$\hat{p}_2$	0.955	0.999		$\hat{p}_2$	0.874	1.000
	$\hat{p}_3$	0.058	0.051		$\hat{p}_3$	0.044	0.065
LV	$\hat{p}_1$	0.923	0.906	LLV	$\hat{p}_1$	0.754	0.919
	$\hat{p}_2$	0.973	0.957		$\hat{p}_2$	0.816	0.991
	$\hat{p}_3$	0.057	0.073		$\hat{p}_3$	0.049	0.063

<Table 5> Estimated measures of performance ($n=45$)

Simple model				Multiple model			
Scenario	Measure	H-S	K-K	Scenario	Measure	H-S	K-K
HL	$\hat{p}_1$	0.531	0.950	HL	$\hat{p}_1$	0.467	0.943
	$\hat{p}_2$	0.556	0.999		$\hat{p}_2$	0.522	1.000
	$\hat{p}_3$	0.043	0.049		$\hat{p}_3$	0.038	0.057
LL	$\hat{p}_1$	0.545	0.947	LL	$\hat{p}_1$	0.585	0.943
	$\hat{p}_2$	0.571	0.996		$\hat{p}_2$	0.716	1.000
	$\hat{p}_3$	0.044	0.052		$\hat{p}_3$	0.041	0.057

HV	$\widehat{p}_1$	0.905	0.951	HLV	$\widehat{p}_1$	0.609	0.938
	$\widehat{p}_2$	0.954	1.000		$\widehat{p}_2$	0.814	0.998
	$\widehat{p}_3$	0.054	0.049		$\widehat{p}_3$	0.039	0.055
LV	$\widehat{p}_1$	0.934	0.927	LLV	$\widehat{p}_1$	0.710	0.909
	$\widehat{p}_2$	0.984	0.942		$\widehat{p}_2$	0.766	0.983
	$\widehat{p}_3$	0.053	0.057		$\widehat{p}_3$	0.044	0.057

5. Concluding Remarks

This paper proposes a hybrid algorithm which combines the indirect approach and the direct approach. An attempt is made to improve the algorithm suggested by Hadi and Simonoff (1993). The proposed algorithm is different from H-S in the method of constructing the initial clean subset. The former finds the clean subset on the basis of the least median of squares estimator which is highly robust, whereas the latter does it using the least squares estimator which is strongly affected by the outliers. Therefore, the proposed algorithm can construct more clean initial subset and hence identify the outliers more correctly.

The performance of the proposed algorithm is investigated in a wide variety of outlier scenarios and regression conditions. The algorithm K-K seems to be superior to H-S particularly when there exist a lot of high leverage points in the data set. In general, the simulation studies reveal that the proposed algorithm is more effective in identifying multiple outliers and is relatively resistant to the masking problem. Furthermore, K-K does not require presetting the number of outliers, does not require Monte Carlo simulation to determine cutoff values, and achieves the computational efficiency and numerical stability.

Although the proposed algorithm has excellent identifying capability, it has some difficulty with the swamping effect as expected. Despite this drawback, we may conclude that the algorithm K-K outperforms the algorithm H-S in the sense that swamping is a less serious problem than masking.

References

[1] Basset, Jr. G. W. (1991). Equivariant, Monotonic, 50% Breakdown Estimators, *The American Statistician*, 45, 135-137.

[2] Hadi, A. S. and Simonoff, J. S. (1993), Procedures for the Identification of Multiple Outliers in Linear Models, *Journal of the American Statistical Association*, 88, 1264-1272.

[3] Kianifard, F. and Swallow, W. H. (1990), A Monte Carlo Comparison of Five Procedures

for Identifying Outliers in Linear Regression, *Commun. Statist.-Theory Meth.*, 19, 1913-1938.

[4] Kim, B. Y. (1996), L_∞-estimation based Algorithm for the Least Median of Squares Estimator, *The Korean Communications in Statistics*, 3, 299-307.

[5] Kim, B. Y. and Kim S. B. (1997), Improvements in Computational Efficiency and Accuracy of an Algorithm for the Identification of Regression Outliers, *Journal of Natural Sciences*, 8, 135-142.

[6] Marasinghe, M. G. (1985), A Multistage Procedure for Detecting Several Outliers in Linear Regression, *Technometrics*, 27, 395-399.

[7] Rousseeuw, P. J. (1984), Least Median of Squares Regression, *Journal of the American Statistical Association*, 79, 871-880.

[8] Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, Wiley-Interscience, New York.

[9] Rousseeuw, P. J. and Zomeren, B. C. (1990), Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633-639.

[10] Tietjen, G. L., Moore, R. H., and Beckman, R. J. (1973), Testing for a Single Outlier in Simple Linear Regression, *Technometrics*, 15, 717-721.

[11] Wisnowski, J. W., Montgomery, D. C. and Simpson, J. R. (2001). A Comparative Analysis of Multiple Outlier Detection Procedures in the Linear Regression Model, *Computational Statistics and Data Analysis*, 36, 351-382.

[2001년 12월 접수, 2002년 4월 채택]