연세경제연구 제 Ⅲ 권, 제 2호
Yonsei Economic Studies, Vol. Ⅲ, No. 2 (September 1996)

Improved Procedure for Least Median of Squares Estimation and Application to Identification of Multiple Outliers

Bu-yong Kim

This article addresses properties of the least median of squares estimator. Computational aspects of the LMS estimation are reviewed, and the degree of approximation and optimality of the algorithms are investigated. An algorithm which yields more optimal solution than the traditional algorithm is proposed on the basis of $L\infty$-estimation. To improve the computational efficiency of the algorithm, updating scheme is employed in computing the projected gradient at some iterations of the linear scaling transformation approach. Also we suggest an effective procedure for outlier detection which uses the LMS estimator to construct an initial clean subset in the sequential procedure. The proposed procedure appears to overcome the masking and swamping problems quite well.

Ⅰ. Introduction

Consider the problem of estimating the parameters of a multiple linear regression model

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

where y denotes an n -vector of response variable, X an $n \times p$ matrix of regressor variables with rank $p<n$, β a p -vector of regression parameters, and ϵ an n -vector of random errors.

Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea. The author acknowledges that this paper was supported in part by Sookmyung Women's University, 1996, and thanks Dr. Peter Rousseeuw for making the PROGRESS source program available. He is also grateful to Dr. Ali S. Hadi for providing a copy of the REGOUT program.

Data sets for regression analysis occasionally contain one or more outliers, and it motivates the use of robust estimators. Various robust estimators have been proposed for regression analysis. For instance, the minimum L_1-norm estimator has long been considered as an acceptable robust estimator alternative to least squares estimator in the presence of vertical outliers. Statistical properties and empirical tests of the L_1-estimator can be found in Rosenberg and Carlson [11], Bassett and Koenker [2], and Dielman and Pfaf-fenberger [3]. It is known that the finite sample breakdown point, a measure of the global insensitivity of the estimator to multiple outliers in the data, of L_1-estimator is equal to $1/n$ because of the possibility of leverage points. Recently, Kim [7] has suggested a definition of vertical breakdown point of the L_1-estimator, which is $(n-p)/2n$. Also M-estimator introduced by Huber [6] is well known to be statistically efficient and highly robust to vertical outliers. However the finite sample breakdown point of M-estimator is still no better than $1/n$. To cope with this vulnerability to leverage points, the GM-estimators such as Mallows-type and Schweppe-type estimator have been suggested, which bound the influence of outliers by means of some weight function. However, Mallows-type estimator downweights leverage points regardless of the magnitude of the corresponding residual, and hence decreases the efficiency. In contrast, Schweppe-type estimator downweights the influence of leverage points only if the corresponding residual is large. Those GM-estimators are less sensitive to leverage points, but it turns out that the breakdown point of the estimator is at most $1/(p+1)$. Several high breakdown point estimators have been proposed to deal with those shortcomings. Siegel [18] introduces the repeated median estimator of which the breakdown point is equal to 0.5 . However, the estimator is not equivariant for linear transformations. Also Rousseeuw [12] proposes the least trimmed squares estimator which has the breakdown point of 0.5. This estimator is equivariant and has a good asymptotic efficiency, but requires extensive computation.

On the other hand, Rousseeuw [13] has proposed a high breakdown point estimator that is called the least median of squares estimator given by

$$\underset{\hat{\boldsymbol{\beta}}}{\text{minimize}} \quad \underset{i}{\text{median}} \quad e_i^2, \tag{2}$$

where $e_i = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$. Of course, the squared residual (e_i^2) may be replaced by the absolute residual ($|e_i|$) for the sake of computation. It has been proved that a solution to the problem (2) exists under an assumption.

Ⅱ. Properties of LMS Estimator

The LMS estimator has some desirable theoretical properties and hence is widely used for robust regression. This estimator has high robustness with respect to leverage points as well as vertical outliers. Rousseeuw [13] shows substantial advantage of the LMS estimator against four traditional robust competitors(Huber's M-estimator, Mallows' GM-estimator, Schweppe's GM-estimator, and repeated median estimator). In addition, it has been shown that if $p>1$ and observations are in general position, then the finite sample breakdown point of the LMS estimator is $([n/2]-p+2)/n$, where $[\cdot]$ represents the largest integer function. Taking the limit for $n\rightarrow\infty$ with p fixed, we find that the breakdown point of LMS estimator is as high as 0.5, the best that can be expected.

Another desirable property of the LMS estimator is that it is regression equivariant, scale equivariant, and affine equivariant (These properties are extensively reviewed by Bassett [1]). Furthermore, this estimator has the exact fit property. That is, when at least $n - [n/2] +1$ of the observations satisfy the relation $y_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ exactly and are in general position, then the LMS estimate equals to $\hat{\boldsymbol{\beta}}$ whatever the other observations are.

On the other hand, the main disadvantage of LMS estimator is that it is less efficient than other robust estimators. To deal with the inefficiency of the LMS estimator, Yohai [21] suggests an approach which uses the LMS estimate and high breakdown point scale estimate of σ as the starting value in the iterative algorithm of M-estimation. Another drawback of the LMS estimator is that it requires a great deal of computation.

Ⅲ. Algorithms for LMS Estimation

In this section we shall concentrate on the computational aspects of the LMS estimation and the behavior and performance of the LMS algorithms. A brief overview is given with respect to the algorithm suggested by Rousseeuw [13] and the program PROGRESS given by Rousseeuw and Leroy [15]. The L_∞-estimation based algorithm suggested by Kim [8] is briefly described and some modifications are made to improve the computational efficiency. Empirical comparisons of algorithms are performed in terms of the degree of approximation and optimality, and it is shown that the proposed algorithm clearly yields

more optimal solution than PROGRESS.

1. Resampling Algorithm

It is probably impossible to write down a closed form formula for the LMS estimation. Therefore, several algorithms (for instance, Steele and Steiger [20], Souvaine and Steele [19]) have been suggested in the recent past. However, since those algorithms suffer from the computational complexity, resampling approximation algorithm suggested by Rousseeuw [13] and a variant given by Marazzi [9] are widely used in the LMS estimation. The basic resampling algorithm starts with selecting a subsamples of size p having different observations. For each subsample, the solution of a system of p linear equations is obtained, which is called the trial estimate (On the contrary, Marazzi's algorithm obtains the trial estimate by the least squares method applied to a subsample of size $q(>p)$). And the predicted residuals and the corresponding LMS objective function with respect to whole observations in data set are calculated. Then the trial estimate with minimum value of objective function is finally taken as the LMS estimate. In this algorithm, the h-th order statistic is treated as a median, where $h = [n/2] + [(p+1)/2]$ with which the maximum possible breakdown point is attained.

In fact, this algorithm does not, in general, produce exact solutions, but the approximate ones. Moreover, the program PROGRESS adapts the approximate method in which one may choose the faster version or the extensive search version since a lot of computation is required when all possible subsamples are considered. PROGRESS actually employs the predetermined values of m for different combinations of n and p. As a result it may yield very approximate estimates (Details of computational results are described in Section III.3).

2. Proposed Algorithm

An algorithm LINFLMS has been proposed by Kim [8] in an attempt to obtain more optimal solutions than the traditional algorithms for the LMS estimation. In fact, its computational inefficiency problem still remains. In this section some modifications are made with the aim of improving the computational efficiency of the algorithm LINFLMS.

For completeness, the L_∞-estimation procedure for linear regression model (1) is outlined. Suppose that a subsample of size h is taken from the data set. The L_∞-estimation problem is defined as follows

$$\underset{\hat{\beta}}{\text{minimize}} \quad \| e \|_{\infty}, \tag{3}$$

where $\| \cdot \|_{\infty}$ denotes the L_{∞}-norm. This problem can be reformulated as the linear programming problem

$$\text{minimize} \quad \lambda$$
$$\text{subject to} \quad \begin{bmatrix} X & \ell \\ -X & \ell \end{bmatrix} \begin{bmatrix} \widehat{\beta_{\infty}} \\ \lambda \end{bmatrix} \geq \begin{bmatrix} y \\ -y \end{bmatrix} \tag{4}$$

where λ denotes the maximum absolute residual that is to be minimized, and $\ell = (1, \cdots, 1)' \in R^h$. To overcome the computational inefficiency, the dual problem corresponding to the formulation (4) is usually solved,

$$\text{maximize} \quad \{c'r : Ar = b, r \geq 0\}, \tag{5}$$

$$c = \begin{bmatrix} y \\ -y \end{bmatrix}, r = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, A = \begin{bmatrix} X' & -X' \\ \ell' & \ell' \end{bmatrix}, b = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

where $r \in R^{2h}$, and ξ_1 and ξ_2 represent dual variables.

Kim [8] suggests an algorithm which is based on a linear scaling transformation scheme, since it requires a great deal of computation to solve the problem by the simplex-type algorithms particularly when the data set is large. At each iteration, given a feasible solution gamma $r_{\langle k \rangle}$, the algorithm employs the scaling linear transformation such that

$$r = \Gamma_{\langle k \rangle} \tau, \text{ where } \Gamma_{\langle k \rangle} = \text{diag } \{r_{\langle k \rangle 1}, \cdots, r_{\langle k \rangle 2h}\}.$$

This transformation reformulates the problem (5) in terms of τ coordinates as follows,

$$\text{maximize } \{(\Gamma_{\langle k \rangle} c)' \tau : A\Gamma_{\langle k \rangle} \tau = b, \tau \geq 0 \}. \tag{6}$$

The steps of the L_{∞}-algorithm are briefly described as follows.

Initialization : Set the number of iteration $k = 1$, and let $r_{\langle 0 \rangle}$ be the initial feasible solution.

Step 1 : Given $r_{\langle k \rangle}$, $\Gamma_{\langle k \rangle} = \text{diag } \{r_{\langle k \rangle 1}, \cdots, r_{\langle k \rangle 2h}\}$. Compute the projected gradient $p_{\langle k \rangle}$, and the direction of motion $d_{\langle k \rangle}$

$$\boldsymbol{p}_{\langle k\rangle} = [I-(A\Gamma_{\langle k\rangle})'(A\Gamma^2_{\langle k\rangle}A')^{-1}A\Gamma_{\langle k\rangle}]\Gamma_{\langle k\rangle}\boldsymbol{c},$$
$$\boldsymbol{d}_{\langle k\rangle} = \Gamma_{\langle k\rangle}\boldsymbol{p}_{\langle k\rangle}$$

Step 2 : If $\|\boldsymbol{p}_{\langle k\rangle}\|_\infty < \omega$, for small enough $\omega = 10^{-8}$, then go to Step 5.

Step 3 : Compute the step length

$$\eta_{\langle k\rangle} = \delta\eta_{\max}, \text{ where } 1/\eta_{\max} = \underset{i=1,\cdots,2h}{\text{maximum}}\{-d_{\langle k\rangle i}/r_{\langle k\rangle i}\} > 0,\ \delta = 0.97.$$

Step 4 : Set the new iterate

$$\boldsymbol{r}_{\langle k+1\rangle} = \boldsymbol{r}_{\langle k\rangle} + \eta_{\langle k\rangle}\boldsymbol{d}_{\langle k\rangle}$$

increment k by one and return to Step 1.

Step 5 : Compute the primal solution

$$\boldsymbol{w} = (A\Gamma^2_{\langle k\rangle}A')^{-1}A_{\langle k\rangle}\Gamma^2\boldsymbol{c}.$$

Then pick the first p entries of the vector $\boldsymbol{w}$ as the L_∞-estimates, $\widehat{\boldsymbol{\beta}_\infty}$, and stop.

In practice, the initial feasible solution can be obtained by adding one artificial variable r_{2h+1} to the A matrix ; arch $\hat{A} = [A : \boldsymbol{b}-A\boldsymbol{1}]$, $\boldsymbol{1} = (1, \cdots, 1)' \in R^{2h}$, and assigning big-M to the artificial variable. In Step 1, $\boldsymbol{p}_{\langle k\rangle}$ is the prpojection of the gradient of the objective function onto the null space of the equality constraints. The step length eta $\eta_{\langle k\rangle}$ in Step 3 should be chosen to ensure the feasibility of new point. And a step is taken along $\boldsymbol{d}_{\langle k\rangle}$, yielding the new iterate in Step 4. If the termination criterion is met in Step 2, then the algorithm is completed (The convergence of the algorithm has been proved by Sherali *et al.* [17]). Then the solution $\widehat{\boldsymbol{\beta}_\infty}$, of the primal problem (4) can be obtained in Step 5 since $\Gamma_{\langle k\rangle}\boldsymbol{c}$ is in the orthogonal complement of the null space of $A\Gamma_{\langle k\rangle}$, and hence there exists a vector $\boldsymbol{w}$ such that $\boldsymbol{w}'A\Gamma_{\langle k\rangle} = (\Gamma_{\langle k\rangle}\boldsymbol{c})'$. That is, the current L_∞-estimate $\widehat{\boldsymbol{\beta}_\infty}$ consists of the first p entries of the vector $\boldsymbol{w}$.

The algorithm LINF is very computer intensive. Therefore, to improve the computational efficiency of the algorithm, updating technique is employed in computing the matrix$(A\Gamma^2_{\langle k\rangle}A')^{-1}$ since the most computer intensive part of the algorithm is clearly in computing the projected gradient $\boldsymbol{p}_{\langle k\rangle}$ at each iteration. Let $Z_{(k)} = A\Gamma^2_{\langle k\rangle}A'$ and $\Delta = \Gamma^2_{\langle k+1\rangle} - \Gamma^2_{\langle k\rangle}$, and R denote the index set of the nonnull rows in Δ. Then at the $(k+1)$th

iteration,

$$(A\Gamma^2_{\langle k+1\rangle}A')^{-1} = (Z_{\langle k\rangle} + A\Delta A')^{-1}$$
$$= (I + Z^{-1}_{\langle k\rangle}A_R\,\Delta_{RR}\,A_R')^{-1}\,Z^{-1}_{\langle k\rangle}$$
$$= Z^{-1}_{\langle k\rangle} - Z^{-1}_{\langle k\rangle}A_R\,\Delta_{RR}(I + A_R'Z^{-1}_{\langle k\rangle}A_R\,\Delta_{RR})^{-1}A_R'Z^{-1}_{\langle k\rangle}$$

where I denotes properly dimensioned identity matrices. Therefore the inverse matrix $(A\Gamma^2_{\langle k+1\rangle}A')^{-1}$ can be easily updated from $(A\Gamma^2_{\langle k\rangle}A')^{-1}$ at each iteration. The dimension of the matrix $A_R'Z^{-1}_{\langle k\rangle}A_R\,\Delta_{RR}$ is smaller than that of $A\Gamma^2_{\langle k+1\rangle}A'$ when the number of nonnull rows is smaller than $p+1$. This situation happens particularly when the algorithm approaches to the optimum, that is, many elements of $\hat{p}_{\langle k\rangle}$ are equal to zero. As a consequence, the improvement in the computational efficiency is expected to be obtained if this updating scheme is employed at the iterations for which the number of nonnull elements of $\hat{p}_{\langle k\rangle}$ is smaller in comparison to $p+1$ and the number of regressors is considerably large. Now, the main algorithm for the LMS-estimation can be described as follows.

Initialization : Set the iteration number t=1, and a very large number $Q^* = \infty$. Construct all possible s subsamples of size $p+2$ of the n observations (Or, specify the appropriate replication number $m(<s)$ to lessen the computation required).

Step 1 : For one of the subsamples (or, when the replication number m is specified, select a subsample randomly), compute the L_∞-estimate $\widehat{\beta}_\infty$ by the algorithm LINF, and predicted residuals $e = y - X\widehat{\beta}_\infty$.

Step 2 : Find the objective function value, that is, the h-th absolute residual $Q = \tilde{e}_{(h)}$, where $h = [n/2] + [(p+1)/2]$.

Step 3 : If $Q<Q^*$, set $Q^* = Q$ and $\hat{\beta} = \widehat{\beta}_\infty$.

Step 4 : If t=s (or, if $t=m$ when m is specified), then return $\hat{\beta}$ as the LMS estimate. Otherwise, set $t = t+1$ and go to Step 1.

3. Empirical Comparisons of Algorithms

The proposed algorithm is implemented on PC with the source program written in FORTRAN and supported by IMSL subroutines. In order to assess the behavior and performance of the algorithms, first of all, computational studies are conducted to measure to what extent the estimates computed by resampling the limited number (m) of subsamples differ from the estimates obtained by resampling all possible s subsamples. Computational results from the algorithms PROGRESS and LINFLMS1 are compared on the basis of 17 data sets (viz., stackloss data, Hadi-Simonoff simulated data, wood specific gravity data, Hertzsprung-Russell diagram data, salinity data, telephone call data, monthly payments data, pension funds data, phosphorus content data, delivery time data, air quality data, education expenditure data, pilot-plant data, inflation in China data, Coleman data, aircraft data, cloud point data) which are commonly introduced in the literatures on robust regression (The details on the data sets are given by Rousseeuw and Leroy [15]). The computational results are summarized in Appendix.

Throughout the remainder of this article, PROG(T) indicates Rousseeuw-Leroy's algorithm PROGRESS with all possible subsets version, PROG(E) with the extensive search version, and PROG(Q) with the faster version, respectively. And LINFL(T), LINFL(E), and LINFL(Q) are the counterparts of the proposed algorithm LINFLMS1, respectively. As would be expected, several notable facts are found from the computational results. One of them is that the objective function values from PROG(E), PROG(Q), LINFL(E), and LINFL(Q) are not quite close to the values from PROG(T) and LINFL(T). It implies that the estimates computed by the extensive version and faster version are very approximate ones.

On the other hand, the optimality of the proposed algorithm LINFLMS1 is compared empirically with PROGRESS with regard to the objective function value. On the whole, the computational results indicate that the solutions from PROGRESS are far from optimum, and LINFLMS1 leads to improvements in optimality although LINFL(E) and LINFL(Q) provide approximate solutions with a median absolute residual larger than PROG(E) and PROG(Q) in some cases (However, appropriate choice of the values m for those approximation versions would improve the optimality of LINFL(E) and LINFL(Q)). Apparently, the smallest values of the objective function are produced by LINFL(T), that

is, algorithm LINFLMS1 yields more optimal solution than PROGRESS.

Ⅳ. Application of LMS Estimator to Outlier Identification

It is necessary to ascertain the presence or absence of any outliers in regression data sets, and then identify them prior to proper regression analysis. In this article we are mainly concerned with regression outliers, that is, cases for which observations deviate from the linear relation followed by the majority of the data. Therefore, we define the regression outliers as either the vertical outliers or the bad leverage points in regression data.

A vast criteria have been proposed in recent years for the identification of single outlier or multiple outliers. Several diagnostics are based on the residuals obtained by the least squares fit. But outliers may cause a poor fit for the majority of the data since the least squares estimation accommodates outlying observations. Other diagnostics are based on the principle of deleting one observation at a time. When there is only one outlier in the regression data, well-known diagnostics employing the principle of deleting one observation at a time work quite well. However, this method may fail to detect multiple outliers. Of course, the principle of deleting one observation at a time has been extended to the diagnostics for multiple outliers. However, it is known that multiple deletion methods require a large number of computation, and sometimes suffer from either masking or swamping or both problems.

1. Outlier Identification Based on Robust Estimator

To cope with those problems mentioned above, robust estimator based diagnostic measures are constructed for the identification of multiple outliers. Since robust regression tries to accommodate the majority of the data with dampening the impact of outlying observations, it consequently yields large residuals for the outliers. The residuals from a robust regression may be inspected for anomalies in the same way as the residuals from a least squares estimation. This approach can identify multiple outliers which usually can not be detected by means of the least squares residuals.

Now we briefly review several procedures for outlier identification based on the robust

estimator. McKean *et al.* [10] show that robust residuals from M and GM estimation can be standardized in the same manner as their least squares counterparts. So we may plot $e_i/\hat{\sigma}^*$(where $\hat{\sigma}^*$is a robust estimates of σ) versus the fitted value of y_i or the index i. The observation i is classified as an outlier if $|e_i/\hat{\sigma}^*|$ is large. As a robust estimate of σ, we usually use the median absolute deviation (MAD) denoted by $\hat{\sigma}^* = 1.4826\ median_i\ |e_i - median_j\ (e_j)|$. Recently, Rousseeuw and Croux [4] suggest an alternative to MAD, $\hat{\sigma}^* = 1.1926\ median_i\ (median_j\ |e_i - e_j|)$, which is more efficient and not slanted towards symmetric distributions.

Another simple approach is to look at the standardized residuals from the LMS fit. Its basic principle is to fit the data by the LMS estimation method, after which outliers may be identified as those observations that are far from the fitted line. Also we may plot the standardized residuals $e_i/\hat{\sigma}^*$ (where $\hat{\sigma}^* = 1.4826\sqrt{median_i\ e_i^2}$) versus the fitted value of y_i or the index i. This plot may be used to identify the outliers that are far from the linear pattern of majority. Rousseeuw and Leroy [15] recommend ± 2.5 as the cutoff values. However this cutoff value is chosen by somewhat arbitrary way since the distribution of LMS residuals is unknown for small samples.

On the other hand, Rousseeuw and Zomeren [16] define a diagnostic measure that is based on the Mahalanobis-type robust distance $RD_i = \sqrt{\{x_i' - T(X)\}\ C(X)^{-1}\{x_i - T(X)\}}$ with high breakdown point estimators of location and covariance. They propose a display in which the standardized residuals ($e_i/\hat{\sigma}^*$) from the LMS fit are plotted versus the robust distances (RD_i). They use ± 2.5 and $\sqrt{\{x^2_{0.975}\ (k - 1)}$ as the cutoff values for the standardized residuals and the robust distances, respectively. We may classify the data into regular observations with small RD_i and small $|e_i/\hat{\sigma}^*|$, vertical outliers with small RD_i and large $|e_i/\hat{\sigma}^*|$, and bad leverage points with large RD_i and large $|e_i/\hat{\sigma}^*|$, However, those robust approaches tends to declare larger number of observations as outliers because of their high robustness.

2. Proposed Identification Procedure

Hadi and Simonoff [4] has proposed a sequential procedure for the identification of outliers, which is superior to various existing procedures. It attempts to separate the data into a subset of clean data points and a subset of potential outliers, and then tests whether the potential outliers are outlying relative to the clean subset. It starts with finding an initial

subset of size $[(n+p-1)/2]$, and updates the subset by a stepwise method until all non-outlier observations are included in the clean subset. They suggest two methods for constructing the initial subset, but here attention is focused on one method they recommend. The method is to fit the regression model to the full data, compute and order the n observations by a diagnostic measure $e_i/\sqrt{1-h_{ii}}$, where h_{ii} is the i-th diagonal entry of hat matrix, and then construct an initial basic subset including the first $(p+1)$th observations. Let B of size b be a set of indices of the observations in the initial basic subset, and $\boldsymbol{X}_B$ and $\boldsymbol{y}_B$ be the subsets of data indexed by B. And let $\widehat{\boldsymbol{\beta}_B}$ denote the least squares estimates obtained from the subset $(\boldsymbol{X}_B, \boldsymbol{y}_B)$. Now fit the linear regression model to the initial basic subset, and compute the internally studen-tized residuals ($|e_{B_i}|/\sqrt{1-h_{Bi}}$) if $i \in B$, or the scaled prediction errors ($|e_{B_i}|/\sqrt{1+h_{Bi}}$) if $i \notin B$, where $e_{B_i} = y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}_B}$ and $h_{B_i} = \boldsymbol{x}_i' (\boldsymbol{X}_B' \boldsymbol{X}_B)^{-1} \boldsymbol{x}_i$. And then arrange the observations in ascending order accordingly. If $b < [(n+p-1)/2]$, then reform a new basic subset by taking the first $b+1$ ordered points, and repeat until b is equal to $[(n+p-1)/2]$. If $b=[(n+p-1)/2]$, the initial subset is completely formed. This procedure iteratively update the initial subset until the final clean subset is obtained. If the final clean subset is of size n, the data set does not have any outliers. The observations which are not included in the clean subset is declared as the outliers.

The above procedure is computationally simple, and can form the appropriate initial clean subset when the masking and swamping effects are not serious. However, since this method still uses the least squares method, it is not guaranteed to overcome the masking and swamping problems when the data set has a large number of outliers and hence the least squares fit is strongly affected by the outliers. Also it requires a large number of computation particularly when the data set is large. The central question is whether an initial subset which is more clean can be obtained by the robust estimation technique. Our goal is to proposing a new procedure for constructing more clean initial subset, even if the masking and/or swamping effects are present in the data, by employing the LMS estimator which has very high breakdown point.

The proposed method employs a single-phase approach to find the initial clean subset, whereas Hadi-Simonoff method involves two phases. At the initial step of the algorithm we construct a clean subset of size $n-[n/2]+p-1$ corresponding to the smallest absolute standardized residuals from the LMS fit (The subset size is determined on the basis of the exact fit property of the LMS estimator). By adapting this approach we can construct a

more clean subset, and reduce the amount of computation.

After forming the initial clean subset based on the LMS estimate, we may expand the subset using a stepwise approach. The internally studentized residual and the scaled prediction error based on the clean subset are used as criteria for including and deleting of observations. Let C of size $c(\leq n)$ and bar $\bar{C}$ of size $n-c$ be the sets of indices of the observations in the clean subset and in the potential outlier subset, respectively, and $\boldsymbol{X}_c$ and $\boldsymbol{y}_c$ be subsets of data indexed by C. And let $\widehat{\boldsymbol{\beta}_c}$ and $\widehat{\sigma_c^2}$ denote the least squares estimates obtained from the subset and the corresponding residual mean square, respectively. To test whether a observation already in the subset is significantly far from the remaining observations, the internally studentized residual $(y_i-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}_c})/\widehat{\sigma_c}\sqrt{1-\boldsymbol{x}'(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{x}_i}$ can be compared with some critical value. On the other hand, the scaled prediction error $(y_i-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}_c})/\widehat{\sigma_c}\sqrt{1+\boldsymbol{x}_i'(\mathbf{X}_c'\mathbf{X}_c)^{-1}\boldsymbol{x}_i}$ can be compared with the same cutoff value in order to choose a observation that is to be included in the clean subset. Hadi and Simonoff [4] state that these two statistics follow a t distribution with $c-p$ degrees of freedom for each set C and bar $\bar{C}$. Utilizing the Bonferroni-type approach we can set $t_{\alpha/2(c+1)}(c-p)$ as the cutoff value. The steps in the proposed identification procedure are described in detail below.

Step 1 : Compute LMS estimate and absolute standardized residuals $|e_i/\hat{\sigma}^*|$.

Then construct an index set C corresponding to the observations having the $(n-[n/2]+p-1)$th smallest absolute standardized residuals.

Set $c = n-[n/2]+p-1$.

Step 2 : Compute the following diagnostic measures,

$$d_i = \begin{cases} e_{c_i}/\widehat{\sigma_c}\sqrt{1-h_{c_i}} & \text{if } i \in C \\ e_{c_i}/\widehat{\sigma_c}\sqrt{1+h_{c_i}} & \text{if } i \in \bar{C} \end{cases}$$

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

Step 3 : Arrange the observations in ascending order according to $|d_i|$.

Let $d^*_{(c+1)}$ denote the $(c+1)$th largest value of the $|d_i|$.

If $d^*_{(c+1)} \geq t_{\alpha/2(c+1)}(c-p)$, then label all observations corresponding to $|d_i| \geq t_{\alpha/2(c+1)}(c-p)$ as outliers and stop.

Step 4 : If $c+1<n$, then go to Step 2 with a new set C including the index corresponding

to $d^*_{(c+1)}$. Otherwise, declare no outliers in the data and stop.

3. Applications and Comparisons

The performance of the proposed procedure IDOUT is compared with other well known procedures such as Rousseeuw and Leroy [15], Rousseeuw and Zomeren [16], and Hadi and Simonoff [4], using several real and artificial data sets. For brevity we use the abbreviated title RL, RZ, and HS for the above-mentioned procedures, respectively.

(1) Telephone call data : It contains the number of international phone calls from Belgium in the years 1950-1973. It is clear that this real data set has six extreme vertical outliers with indices 15, 16, 17, 18, 19, and 20, and two moderate vertical outliers with indices 14 and 21. Application of the least squares fit yields masking in the observations 14 and 21, and swamping in the observations 22, 23, and 24. The proposed method IDOUT as well as RL, RZ, and HS methods identifies the outliers correctly.

(2) Hadi-Simonoff data : It is an artificial data set generated by Hadi and Simonoff [4] to illustrate the methods of outlier detection. First three observations are true outliers, and point 17 is under swamping. All four methods, RL, RZ, HS, and IDOUT appear to identify the outliers correctly.

(3) Stackloss data : This real data set describes the operation of a plant for the oxidation of ammonia to nitric acid and consist of 21 observations with three regressors. Many authors have found out that observations 1, 3, 4, and 21 are severe outliers, and observation 2 is a minor outlier. Methods RL and RZ identify all the outliers, whereas HS and IDOUT detect only the severe outliers.

(4) Hertzsprung-Russell stars data : It is formed from the Hertzsprung-Russell diagram of the star cluster CYG OB1, which contains measurements concerning 47 stars in the direction of Cygnus. The regressor is the logarithm of the effective temperature at the star surface as estimated by spectroscopy and the response variable is the logarithm of its light intensity. It has bad leverage points with indices 11, 20, 30, and 34. Application of the LMS estimation to this data set yields hat $\hat{y} = -12.298 + 3.898x$, which fits the majority of data very well, whereas the least squares estimation yields hat $\hat{y} = 6.78 - 0.409x$, which is a poor fit and brings about masking and swamping problems. Two different results are due to the fact that the least squares estimates are heavily affected by the leverage points. Since the HS procedure forms the initial clean subset using the least

squares method, it fails to identify the outliers. On the other hand, the proposed procedure IDOUT which is based on the LMS estimates identifies four outliers correctly.

(5) Hawkins-Bradu-Kass data : This artificial data set consists of 75 observations with 3 regressors. The first 10 observations are bad leverage points which also have the masking effects, and the next 4 points are good leverage points. Hawkins *et al.* [5] show that M-estimator method fails to identify the outliers because the first 10 observations are masked outliers. The proposed method IDOUT as well as RL and HS methods clearly identifies the 10 outliers. Also RZ method classifies 10 observations into bad leverage points, and the next 4 points into good leverage points. On the other hand, if we select the first 45 observations only in order to induce the swamping effect in the data set, then HS method wrongly declares the good leverage points 11-14 as outliers, whereas IDOUT identifies 10 outliers correctly.

(6) Number of fires data : It contains the number of reported fire claims from 1976 to 1980 in Belgium. We notice that the number for 1976 is extraordinarily large, i.e., an outlier. We find that HS method can not identify the outlier, whereas the proposed method IDOUT as well as RL and RZ methods can identify them. This result implies that more clean initial subset is formed by the proposed procedure based on the high breakdown point estimator.

Ⅴ. Concluding Remarks

An attempt has been made to improve the LMS algorithm. Although only the limited number of data sets are investigated in this short contribution, the results indicate that the proposed algorithm yields more optimal estimates in the LMS estimation. In addition, some modifications are made to improve the algorithm with respect to computational efficiency.

The proposed procedure for outlier detection is different from the procedure of Hadi-Simonoff in the construction of 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. As a result the proposed procedure can construct more clean subset and therefore identify outliers more correctly. The examples demonstrate that the proposed procedure appears to

be more effective in identifying multiple outliers and deal with the masking and swamping problems. However, it is worth noting that comparisons of the procedures are performed only on the limited number of examples. Further investigation on the effectiveness of the procedures is the subject of ongoing research.

Appendix

Comparisons on the degree of approximation and optimality

Stackloss data ($p = 4, n = 21$)

$\widehat{\beta_0}$	−36.37500	−34.50000	−34.55000	−35.41490	−33.59547	−31.06765
$\widehat{\beta_1}$	7.29167E-1	7.14286E-1	7.75000E-1	7.50000E-1	7.67588E-1	8.70588E-1
$\widehat{\beta_2}$	4.16667E-1	3.57143E-1	2.00000E-1	4.04255E-1	3.66835E-1	3.47059E-1
$\widehat{\beta_3}$	−4.86825E-17	5.49177E-17	1.37469E-16	−2.12765E-2	−4.52262E-2	−1.41176E-1
med.$\lvert e_i \rvert$	0.583336	0.642857	0.799999	0.531916	0.63819070	0.891180

Hadi−Simonoff simulated data ($p = 3, n = 25$)

$\widehat{\beta_0}$	−5.26127E-1	−5.90322E-1	−6.12456E-1	−1.27579E-1	−4.68631E-1	−2.72725E-2
$\widehat{\beta_1}$	1.05964	1.06629	1.03038	9.88149E-1	1.06293	9.67042E-1
$\widehat{\beta_2}$	1.06546	1.06969	1.10878	1.07642	1.04986	1.08336
med.$\lvert e_i \rvert$	0.429407	0.460623	0.490291	0.395018	0.40739820	0.476889

Wood specific gravity data ($p = 6, n = 20$)

$\widehat{\beta_0}$	2.88082E-1	4.34739E-1	4.34739E-1	4.43711E-1	3.99533E-1	3.99533E-1
$\widehat{\beta_1}$	2.38439E-1	2.68699E-1	2.68699E-1	2.25482E-1	1.82231E-1	1.82231E-1
$\widehat{\beta_2}$	−6.99169E-2	−2.38057E-1	−2.38057E-1	4.07652E-2	1.51189E-1	1.51189E-1
$\widehat{\beta_3}$	−5.69512E-1	−5.35723E-1	−5.35723E-1	−6.79559E-1	−5.64190E-1	−5.64190E-1
$\widehat{\beta_4}$	−3.83893E-1	−2.93732E-1	−2.93732E-1	−4.08702E-1	−3.55694E-1	−3.55694E-1
$\widehat{\beta_5}$	6.82049E-1	4.50959E-1	4.50959E-1	5.33891E-1	5.43279E-1	5.43279E-1
med.$\lvert e_i \rvert$	0.005739	0.007330	0.007330	0.004791	0.004998	0.004998

Hertzsprung-Russell diagram data ($p = 2$, $n = 47$)

$\widehat{\beta_0}$	12.76002	−12.76002	−6.70002	−12.75999	−10.62252	−11.11239
$\widehat{\beta_1}$	4.00000	4.00000	2.66667	3.99999	3.50000	3.60869
med.$\lvert e_i \rvert$	0.280001	0.280001	0.296668	0.260001	0.267503	0.273696

Salinity data ($p = 4$, $n = 28$)

$\widehat{\beta_0}$	36.69960	36.69960	36.69960	37.36716	36.28203	36.28203
$\widehat{\beta_1}$	3.56153E-1	3.56153E-1	3.56153E-1	3.61832E-1	3.59903E-1	3.59903E-1
$\widehat{\beta_2}$	−7.31262E-2	−7.31262E-2	−7.31262E-2	−8.62551E-2	−7.57605E-2	−7.57605E-2
$\widehat{\beta_3}$	−1.29808	−1.29808	−1.29808	−1.32665	−1.28228	−1.28228
med.$\lvert e_i \rvert$	0.374394	0.374394	0.374394	0.314614	0.379380	0.379380

Telephone call data ($p = 2$, $n = 24$)

$\widehat{\beta_0}$	−5.61000	−5.61000	−5.61000	−5.61749	−5.76031	−5.76031
$\widehat{\beta_1}$	1.15385E-1	1.15385E-1	1.15385E-1	1.15500E-1	1.18125E-1	1.18125E-1
med.$\lvert e_i \rvert$	0.089231	0.089231	0.089231	0.086000	0.092813	0.092813

Monthly payments data ($p = 2$, $n = 12$)

$\widehat{\beta_0}$	4.586250	4.936430
$\widehat{\beta_1}$	−3.37500E-2	−1.01429E-1
med.$\lvert e_i \rvert$	0.503750	0.409286

Pension funds data ($p = 2$, $n = 18$)

$\widehat{\beta_0}$	83.183290	46.069010
$\widehat{\beta_1}$	8.169551	8.935798
med.$\lvert e_i \rvert$	168.164000	157.742400

Phosphorus content data ($p = 3$, $n = 18$)

$\widehat{\beta_0}$	44.044070	59.355500
$\widehat{\beta_1}$	1.018435	1.200058
$\widehat{\beta_2}$	4.79991E-1	1.66958E-1
med.$\lvert e_i \rvert$	6.375675	4.752113

Delivery time data ($p = 3$, $n = 25$)

$\widehat{\beta_0}$	4.016476	3.740078
$\widehat{\beta_1}$	1.428218	1.464273
$\widehat{\beta_2}$	1.37797E-2	1.40675E-2
med.$\lvert e_i \rvert$	0.964509	0.885840

Air quality data ($p = 4$, $n = 31$)

	PROG(T)	LINFL(T)
$\widehat{\beta_0}$	−100.934600	−80.017940
$\widehat{\beta_1}$	1.03365E-3	3.35698E-4
$\widehat{\beta_2}$	−1.10824E-1	9.18469E-1
$\widehat{\beta_3}$	1.821568	1.347511
med. $\lvert e_i \rvert$	8.034264	7.716761

Education expenditures data ($p = 4$, $n = 50$)

	PROG(T)	LINFL(T)
$\widehat{\beta_0}$	−158.762200	−210.765000
$\widehat{\beta_1}$	1.03969E-1	4.01286E-2
$\widehat{\beta_2}$	2.22909E-2	4.28468E-2
$\widehat{\beta_3}$	7.47647E-1	7.41842E-1
med. $\lvert e_i \rvert$	18.378810	16.635130

Pilot-plant data ($p = 2$,$n = 20$)

	PROG(T)	LINFL(T)
$\widehat{\beta_0}$	36.518940	35.637790
$\widehat{\beta_1}$	3.10606E-1	3.14961E-1
med. $\lvert e_i \rvert$	0.787880	0.708664

Inflation in China data ($p = 2$, $n = 9$)

	PROG(T)	LINFL(T)
$\widehat{\beta_0}$	−2.468001	−3.422502
$\widehat{\beta_1}$	1.02000E-1	1.25000E-1
med. $\lvert e_i \rvert$	0.092000	0.07249930

Coleman data ($p = 6$, $n = 20$)

	PROG(T)	LINFL(T)
$\widehat{\beta_0}$	23.641860	21.843950
$\widehat{\beta_1}$	5.38716E-1	−3.33593E-1
$\widehat{\beta_2}$	5.55068E-2	5.54329E-2
$\widehat{\beta_3}$	6.26059E-1	6.23313E-1
$\widehat{\beta_4}$	7.55318E-1	9.09080E-1
$\widehat{\beta_5}$	−2.107389	−2.061756
med. $\lvert e_i \rvert$	0.473412	0.292645

Aircraft data ($p = 5$, $n = 23$)

	PROG(T)	LINFL(T)
$\widehat{\beta_0}$	15.167250	10.666760
$\widehat{\beta_1}$	−5.288352	−2.588279
$\widehat{\beta_2}$	1.771048	1.505752
$\widehat{\beta_3}$	1.89026E-3	9.26011E-4
$\widehat{\beta_4}$	−1.08864E-3	−4.83286E-4
med. $\lvert e_i \rvert$	3.112727	2.155865

Cloud point data ($p = 2$, $n = 19$)

	PROG(T)	LINFL(T)
$\widehat{\beta_0}$	24.533330	25.187500
$\widehat{\beta_1}$	8.66667E-1	7.74999E-1
med. $\lvert e_i \rvert$	0.233334	0.212499

◈ *REFERENCES* ◈

1. Basset, Jr. G. W., "Equivalent, Monotonic, 50% Breakdown Estimators," *The American Statistician*, 45, 1991, pp. 135~137.
2. Bassett, G. and Koenker, R., "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association*, 73, 1978, pp. 618~622.
3. Dielman, T. and Pfaffenberger, R., "Least Absolute Value Estimation in Linear Regression: A Review," *TIMS/Studies in the Management Sciences*, 19, 1982, pp. 31~52.
4. Hadi, A. S. and Simonoff, J. S., "Procedures for the Identification of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, 88, 1993, pp. 1264~1272.
5. Hawkins, D. M., Bradu, D., and Kass, G. V., "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," *Technometrics*, 26, 1984, pp. 197~208.
6. Huber, P. J., "Robust Regression : Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 1973, pp. 799~821.
7. Kim, B. Y., "On the Robustness of L_1-estimator in Linear Regression Models," *The Korean Communications in Statistics*, 2, 1995, pp. 277~287.
8. _____, "L_∞-estimation Based Algorithm for the Least Median of Squares Estimator," *The Korean Communications in Statistics*, 3, 1996, pp. 299~307.
9. Marazzi, A., "Algorithms and Programs for Robust Linear Regression," in *Directions in Robust Statistics and Diagnostics : Part I*, (eds.) W. Stahel and S. Weisberg, New York: Springer-Verlag, 1991, pp. 183~199.
10. McKean, J. W., Sheather, S. J., and Hettmansperger, T. P., "The Use and Interpretation of Residuals Based on Robust Estimation," *Journal of the American Statistical Association*, 88, 1993, pp. 1254~1263.
11. Rosenberg, B. and Carlson, D., "A Simple Approximation of the Sampling Distribution of Least Absolute Residuals Regression Estimates," *Communications in Statistics-Simulation and Computation*, B. 6, 1977, pp. 421~437.
12. Rousseeuw, P. J., "Regression Techniques with High Breakdown Point," *IMS Bull.*, 12, 1983, p. 155.
13. _____, "Least Median of Squares Regression," *Journal of the American Statistical*

Association, 79, 1984, pp. 871~880.

14. _____ , and Croux, C., "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1993, pp. 1273~1283.
15. _____ , and Leroy, A. M., *Robust Regression and Outlier Detection*, 1987, Wiley-Interscience, New York.
16. _____ , and Zomeren, B. C., "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 1990, pp. 633~639.
17. Sherali, H. D., Skarpness, B. O., and Kim, B. Y., "An Assumption-Free Convergence Analysis for a Perturbation of the Scaling Algorithm for Linear Programs, with Application to the L_1 Estimation Problem," *Naval Research Logistics*, 35, 1988, pp. 473~492.
18. Siegel, A. F., "Robust Regression Using Repeated Medians," *Biometrika*, 69, 1982, pp. 242~244.
19. Souvaine, D. L. and Steele, J. M., "Time-and Space-Efficient Algorithms for Least Median of Squares Regression," *Journal of the American Statistical Association*, 82, 1987, pp. 794~801.
20. Steele, J. M. and Steiger, W. L., "Algorithms and Complexity for Least Median of Squares Regression," *Discrete Applied Mathematics*, 14, 1986, pp. 93~100.
21. Yohai, V. J., "High Breakdown Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics*, 15, 1987, pp. 642~656.