Journal of the Korean Statistical Society (1998), 27: 3, pp 369-383

An Efficient Mallows-Type One-Step GM-Estimator in Linear Models†

Moon Sup Song[1], Changsoon Park[2] and Ho Soo Nam[3]

ABSTRACT

This paper deals with a robust regression estimator. We propose an efficient one-step GM-estimator, which has a bounded influence function and a high breakdown point. The main idea of this paper is to use the Mallows-type weights which depend on both the predictor variables and the residuals from a high breakdown initial estimator. The proposed weighting scheme severely downweights the bad leverage points and slightly downweights the good leverage points. Under some regularity conditions, we compute the finite-sample breakdown point and prove the asymptotic normality. Some simulation results and a numerical example are also presented.

Keywords: LTS estimator; Influence Function; Breakdown Point; Mallows-type GM-estimator; MVE estimator; Leverage Point

1. INTRODUCTION

Regression analysis is a very extensively used technique in statistical data analysis. In this paper we consider the multiple linear regression model which is given as follows:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \ , \quad i = 1, 2, \cdots, n \ , \tag{1.1}$$

where $\{(\mathbf{x}_i^T, y_i) : i = 1, 2, \cdots, n\}$ is a sequence of independent and identically distributed (iid) random variables with distribution function $F(\mathbf{x}, y)$, $\mathbf{x}_i$ is a $p \times 1$ random vector, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. While, ϵ_i's are iid, independent of $\mathbf{x}_i$ and symmetric about 0 with scale parameter σ.

When the errors are assumed to be normal with mean 0 and variance σ^2, the classical least squares (LS) estimator is regarded as the most suitable one.

†The present studies were supported by the Basic Science Research Institute Program, Ministry of Education, Korea, 1997, Project BSRI-97-1415.

[1]Department of Statistics, Seoul National University, Seoul, 151-742, Korea.

[2]Department of Applied Statistics, Chung-Ang University, Seoul, 156-756, Korea.

[3]Department of Industrial Engineering, Dongseo University, Pusan, 617-716, Korea.

However, in many practical situations the errors are not of this form. Worse than all, if there is an extremely bad leverage point then the LS estimator is readily broken down.

To overcome the non-robustness of the LS estimator, various alternative approaches such as the Huber's M-estimator (Huber (1973)) and trimmed least squares estimator (Ruppert and Carroll (1980)) have been tried. These estimators are robust to the potential regression outliers. However, these estimators cannot restrict the influence of leverage points which are outliers in x-direction, since these were designed only to protect the effects of outliers in y-direction.

The influence function defined by Hampel (1974) describes the effect of an infinitesimal contamination on a statistic. This is very useful as a measure of gross-error sensitivity, local-shift sensitivity or rejection point. However, it is a local concept. As a complement of the influence function, the breakdown point introduced by Hampel (1971) can be considered. The breakdown point represents the global reliability of an estimator or test statistic, i.e., it describes the limiting fraction of bad outliers a statistic can cope with. Many researchers in robust regression make efforts to robustify the estimators or test statistics to achieve the two main goals: bounded influence and high breakdown point.

The generalized M (GM)-estimator which is frequently called as bounded influence estimator has been discussed by Maronna and Yohai (1981), Krasker and Welsch (1982), etc. In general, the GM-estimator of $\boldsymbol{\beta}$ in the regression model (1.1) is defined implicitly by the solution of the equation

$$\sum_{i=1}^{n} \eta\left(\mathbf{x}_i, \frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\mathbf{x}_i = \mathbf{0}, \tag{1.2}$$

where η is a real-valued function. There have been several proposals for choosing η, and most of them can be written in the form

$$\eta(\mathbf{x}, r) = w(\mathbf{x})\psi(\mathbf{x}, r),$$

where $\psi(\mathbf{x}, r)$ is the ψ-function in M-estimation. The GM-estimator pursues essentially the bounded influences in both $\mathbf{x}$ and y directions. The GM-estimator uses weights which depend on the carriers to bound the influence of position in the factor space. However, these earlier GM-estimators have breakdown points of at most $1/p$, which decreases to 0 as p (number of parameters) increases.

Another goal to pursue in robust regression is to attain a high breakdown point. The least median of squares (LMS) estimator and the least trimmed squares (LTS) estimator of Rousseeuw (1984), and the S-estimator of Rousseeuw

and Yohai (1984) are the outcomes of high breakdown regression. These estimators have maximal breakdown point of 50%. Inspite of their virtue of high breakdown points, there are some defects to these estimators. One of them is the computational difficulty to obtain the estimators in practice. To overcome the difficulties, some fast and efficient algorithms to calculate the approximate solution have been proposed by many researchers in recent years. Another defect is the low efficiency of these estimators. This poor efficiency is caused by the structure of their object functions to obtain high breakdown points.

Simpson, Ruppert and Carroll (1992) and Coakley and Hettmansperger (1993) proposed bounded-influence and high-breakdown regression estimators which are one-step GM-estimators based on high breakdown initial estimators such as LMS or LTS estimator. The estimator proposed by Simpson *et al.* (1992) is the Mallows-type one-step GM-estimator with η-function of the form

$$\eta(\mathbf{x}, r/\sigma) = w(\mathbf{x})\psi(r/\sigma),$$

where $r = y - \mathbf{x}^T\boldsymbol{\beta}$. While, to improve the efficiency of the Mallows-type GM-estimator Coakley and Hettmansperger (1993) presented the Schweppe-type one-step GM-estimator with η-function of the form

$$\eta(\mathbf{x}, r/\sigma) = w(\mathbf{x})\psi(r/\sigma w(\mathbf{x})).$$

It is regarded that these two estimators combine the bounded influence approaches with the high breakdown procedures.

Song, Park and Nam (1996) proposed a bounded influence and high breakdown regression estimator, which is based on the weight function depending on both the residuals and the design points. That is, the η-function is of the form

$$\eta\left(\mathbf{x}, \frac{r}{\sigma}\right) = w\left(\mathbf{x}, \frac{r}{\sigma}\right)\psi\left(\frac{r}{\sigma w\left(\mathbf{x}, r/\sigma\right)}\right), \tag{1.3}$$

where

$$w\left(\mathbf{x}, \frac{r}{\sigma}\right) = \min\left(1, \frac{a \cdot \sigma v(\mathbf{x})}{r}\text{sign}(r)\right). \tag{1.4}$$

Here, a is a tuning constant and $v(\mathbf{x})$ is a measure of leverageness defined as follows (see Simpson *et al.*, 1992):

$$v(\mathbf{x}) = \min\left[1, \left\{\frac{b}{(\mathbf{z} - \mathbf{m}_\mathbf{z})^T\mathbf{C}_\mathbf{z}^{-1}(\mathbf{z} - \mathbf{m}_\mathbf{z})}\right\}^{1/2}\right], \tag{1.5}$$

where $\mathbf{z}$ is a $(p-1) \times 1$ vector of predictor variables such that $\mathbf{x}^T = (1, \mathbf{z}^T)$, and $\mathbf{m}_\mathbf{z}$ and $\mathbf{C}_\mathbf{z}$ are the minimum volume ellipsoid (MVE) estimates of location and covariance of $\{\mathbf{z}\}$, respectively, and b is the $(1-\gamma)$ quantile of the chi-squared distribution with $p-1$ degrees of freedom ($\gamma = 0.05$ or 0.025). The equation (1.3) is a form of the Schweppe-type, which is designed to overcome the defect that the Mallows-type downweights the leverage points regardless of the contribution to the model of these points.

Recently, Song and Kim (1997) proposed a one-step pairwise GM-estimator, which minimizes the weighted sum of absolute pairwise differences of residuals.

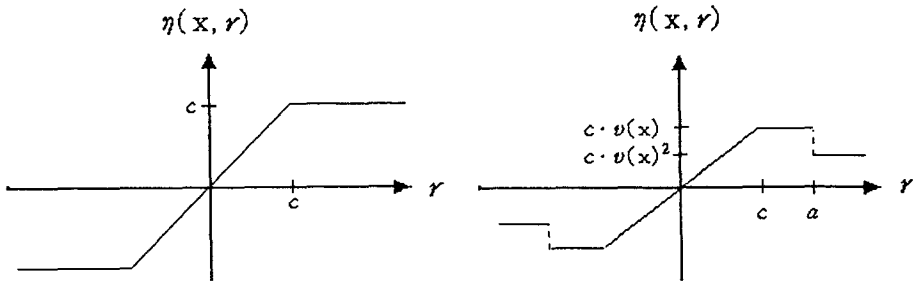
The simulation study in Song et al. (1996) shows that the one-step GM-estimator based on (1.3) performs better than those of Simpson et al. (1992) and Coakley and Hettmansperger (1993). However the weight function (1.4) downweights the nonleverage points, i.e. the points with $v(\mathbf{x}) = 1$, if the residuals are moderately large. Note that the y-direction outliers are to be downweighted by the ψ-function. Thus the points with moderately large residuals are to be seriously downweighted. These facts make the estimator based on (1.4) inefficient. In this paper, we thus propose a Mallows-type one-step GM-estimator with a weight function which downweights only the leverage points. Under some regularity conditions, the proposed one-step GM-estimator has a bounded influence function and a high breakdown point. To compare the performance of the proposed estimator with other estimators, we perform some Monte Carlo simulations in various situations. As a numerical example, the well-known stackloss data are analyzed.

2. THE PROPOSED ESTIMATOR AND ITS PROPERTIES

The main point of the proposed estimator is on the weight function, which is given by

$$w\left(\mathbf{x}, \frac{r}{\sigma}\right) = \begin{cases} v(\mathbf{x}) & \text{if } |r/\sigma| \le d, \\ v(\mathbf{x})^2 & \text{if } |r/\sigma| > d. \end{cases} \tag{2.1}$$

where $v(\mathbf{x})$ is the same measure of leverage in (1.5) and d is a tuning constant. Note that the measure of leverageness, $v(\mathbf{x})$, is less than 1 only for the leverage points. Thus the weight function (2.1) downweights only the leverage points. The leverage points with small residuals are to be slightly downweighted, while those with large residuals are severely downweighted.



(a) $v(x) = 1$ (b) $0 < v(x) < 1$

Figure 2.1: Shape of the η-function with Huber's ψ (tuning constant c $(< d)$).

Now the proposed estimator is given by the solution of the equation

$$\sum_{i=1}^{n} \eta(\mathbf{x}_i, r_i(\beta/\sigma)\mathbf{x}_i = \mathbf{0}\ , \tag{2.2}$$

where

$$\eta\left(\mathbf{x}, \frac{r}{\sigma}\right) = w\left(\mathbf{x}, \frac{r}{\sigma}\right)\psi\left(\frac{r}{\sigma}\right). \tag{2.3}$$

If the Huber's ψ-function, which is given by

$$\psi\left(\frac{r}{\sigma}\right) = \begin{cases} r/\sigma & \text{if } |r/\sigma| \leq c \\ c \cdot \text{sign}(r) & \text{if } |r/\sigma| > c, \end{cases}$$

is applied to (2.3), then the shape of the η-function is given in Figure 2.1. By taking a first order Taylor-series expansion of the left side of (2.2) about $\widehat{\boldsymbol{\beta}}_0$, the proposed one-step GM-estimator based on the scoring method is defined by

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_0 + \widehat{\sigma}_0 H_0^{-1} g_0, \tag{2.4}$$

where

$$g_0 = \sum_{i=1}^{n} \eta(\mathbf{x}_i, r_i(\widehat{\boldsymbol{\beta}}_0)/\widehat{\sigma}_0)\mathbf{x}_i$$

and

$$H_0 = X^T \widehat{A} X\ , \quad \widehat{A} = \text{diag}\left(\frac{1}{n}\sum_{j=1}^{n} \eta'(\mathbf{x}_i, \frac{r_j(\widehat{\boldsymbol{\beta}}_0)}{\widehat{\sigma}_0})\right).$$

Here, X is the $n \times p$ matrix having rows $\mathbf{x}_i^T$, $\eta'(\mathbf{x}, r) = \partial\eta(\mathbf{x}, r)/\partial r$, $\widehat{\sigma}_0$ is a high breakdown scale estimator such as the median absolute deviation(MAD), and $\widehat{\boldsymbol{\beta}}_0$ is an initial estimator of $\boldsymbol{\beta}$ such as the LMS or LTS estimators. We use the LTS estimator throughout this paper.

Now consider the properties of the proposed estimator (2.4) under the Assumptions (C1) through (C9) of Song et al. (1996).

Theorem 2.1. *(1) Under the assumptions (C1) and (C2), the proposed estimator has a bounded influence function.*
(2) Under the assumptions (C2) and (C4), the proposed one-step estimator has a breakdown point of $([n/2] - p + 1)/n$.
(3) Under the assumptions (C2) and (C5) through (C9),

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2\Sigma),$$

where $\Sigma = D^{-1}ED^{-1}$ *with*

$$D = \int \eta'\left(\mathbf{x}, \frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\mathbf{x}\mathbf{x}^T dF(\mathbf{x}, y)$$

and

$$E = \int \eta^2\left(\mathbf{x}, \frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\mathbf{x}\mathbf{x}^T dF(\mathbf{x}, y).$$

Proof: The proof is similar to that of Song et al. (1996). For the proof of the bounded influence, note that the ψ function is bounded and $||w(\mathbf{x}, r)\mathbf{x}||$ is bounded as a function of $\mathbf{x}$ and r, bacause $||v(\mathbf{x})\mathbf{x}||$ and $||v(\mathbf{x})^2\mathbf{x}||$ are finite when $||\mathbf{x}||$ or $|r|$ goes to infinity. □

For inferences about $\boldsymbol{\beta}$, the asymptotic variance of $\sqrt{n}\ \widehat{\boldsymbol{\beta}}$, n $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 D^{-1}ED^{-1}$, can be estimated as follows:

$$n\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}_0^2\ \widehat{D}^{-1}\widehat{E}\,\widehat{D}^{-1}. \tag{2.7}$$

Here, $\widehat{D}$ and $\widehat{E}$ are computed as follows (Song et al., 1996):

$$\widehat{D} = \frac{1}{n}(X^T\widehat{A}X)\ , \quad \widehat{E} = \frac{1}{n}(X^T\widehat{V}X),$$

where $\widehat{V}$ is the diagonal matrix with diagonal elements $\sum_{j=1}^n \eta^2(\mathbf{x}_i, r_j(\widehat{\boldsymbol{\beta}}_0)/\widehat{\sigma}_0)/n$, $i = 1, \cdots, n$.

3. SOME MONTE CARLO RESULTS

In this section we want to compare the proposed estimator with the well-known estimators such as the LS, LTS, Huber-M, Mallows-type GM, and Song et al.(1996)'s GM estimators. To perform a Monte Carlo study we consider the following four cases:

Case 1) No leverage points and no outliers.

Case 2) No leverage points but with some outliers.

Case 3) Some bad leverage points.

Case 4) Some bad leverage points and some good leverage points.

The simulation model is

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \cdots, n. \tag{4.1}$$

The regression parameters are set as $\alpha = \beta_1 = \beta_2 = 1$ for simplicity. The number of observations is $n = 50$, and the number of replications is 500. The explanatory variables and error terms are generated as follows.

For Case 1, the explanatory variables x_{ij}'s are independently generated from the normal distribution $N(0,1)$, for $i = 1, \cdots, 50$ and $j = 1, 2$. The error term ϵ_i's are independently generated from the standard normal $N(0,1)$. These explanatory variables and errors are newly generated in each replication.

For Case 2, we generate x_{ij}'s by the same method as in Case 1. But, the ϵ_i's are generated from the contaminated normal. The distribution function of ε-contaminated normal(CN(ε, σ)) is given by

$$F(\epsilon) = (1 - \varepsilon)\Phi(\epsilon) + \varepsilon\Phi(\epsilon/\sigma),$$

and $\varepsilon = 0.2$ and $\sigma = 5$ are used in our Monte Carlo study. We expect that some mild or extreme outliers are generated from this distribution comparing to $N(0,1)$ of Case 1.

For Case 3, first we generate x_{ij}'s from $N(0, 3^2)$ and ϵ_i's from $N(0,1)$. Then y_i's are obtained according to (4.1). To make some bad leverage points, we randomly select 10% points of the data, and replace these $(\mathbf{x}_i, y_i)$ by $(\mathbf{x}_i + 15, y_i)$.

In Case 4, we need some "bad" and some "good" leverage points. We first generate x_{ij}'s from $N(0,1)$ and replace 10% randomly selected $\mathbf{x}_i$ by $\mathbf{x}_i + 3$. Next,

we generate ϵ_i's from $N(0,1)$ and obtain y_i's according to (4.1). Thus, we expect about 5 good leverage points. To make some bad leverage points, we replace 10% points $(\mathbf{x}_i, y_i)$ randomly selected from the remaining observations by $(\mathbf{x}_i + 6, y_i)$. The latter 10% points must be bad leverage points.

We compared 6 estimators in our simulation study, and the results are summarized in Table 3.1. The performance of the estimators are compared in two ways, the bias and the MSE. Thus the empirical mean and the empirical MSE are listed in Table 3.1. In the table we use the following abbreviations:

LS : Least squares estimator

LTS : Least trimmed squares estimator

HME : Huber's one-step M-estimator

MGM : Mallows-type one-step GM-estimator

SPN : Song, Park and Nam(1996)'s one-step GM-estimator

NEW : Newly proposed one-step GM-estimator

We use the Huber's ψ-function with the tuning constant $c = 1.5$, for the M- and GM-estimators. The one-step M- and GM-estimators are obtained by the scoring method. In the proposed estimator, the tuning constant $d = 2.0$ is applied to compute weights. The constant $d = 2.0$ is chosen as a reasonable value through a simulation study in various situations. For the measure of leverageness $v(\mathbf{x}_i)$, we use the constant $b = \chi^2_{2,0.975}$ for three types of GM-estimators.

All computations in this Monte Carlo simulation were carried out on PC/ Pentium-II by using S-PLUS (Version 3.2). The LTS and MVE estimates were obtained by S-PLUS functions *ltsreg* and *cov.mve* respectively. Also the S-PLUS function *rnorm* and *runif* were used to generate normal and contaminated normal random variates.

According to the simulation results in Table 3.1, the LS has best performances in MSE in Case 1, as expected. However, there are only a little differences in bias except the LTS. In Case 2, the M-estimator (HME) and 3 GM-estimators (MGM, SPN and NEW) show similar performances. While, the LS and LTS are significantly inferior in terms of MSE.

In Case 3 and Case 4, MGM and SPN are better than the Huber's M-estimate. The LS is significantly biased and has larger MSE than the others. The proposed estimator(NEW) dominates the others in the viewpoint of MSE in Case 3 and

Table 3.1: Empirical MEAN and MSE

Estimators	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
Case 1 : No leverage points & No outliers			
	Empirical Mean		
LS	1.00797	0.99589	1.00630
LTS	1.02949	0.98996	1.02612
HME	1.00530	0.99432	1.00585
MGM	1.00454	0.99401	1.00547
SPN	1.00523	0.99444	1.00570
NEW	1.00431	0.99387	1.00518
	Empirical MSE		
LS	0.02011	0.02101	0.02368
LTS	0.10627	0.12955	0.11254
HME	0.02227	0.02529	0.02683
MGM	0.02248	0.02559	0.02746
SPN	0.02230	0.02550	0.02691
NEW	0.02242	0.02554	0.02754

Estimators	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
Case 2 : No leverage points & Some outliers			
	Empirical Mean		
LS	1.01095	0.97171	0.99853
LTS	1.02228	0.99956	0.97873
HME	1.00840	0.99109	0.99144
MGM	1.00885	0.99234	0.99153
SPN	1.00851	0.99257	0.99075
NEW	1.00913	0.99315	0.99101
	Empirical MSE		
LS	0.11857	0.14122	0.13245
LTS	0.10906	0.12220	0.12118
HME	0.04104	0.05049	0.04617
MGM	0.04107	0.05093	0.04527
SPN	0.04035	0.04950	0.04550
NEW	0.04114	0.05073	0.04480

Case 3 : 10% Bad Leverage Points			
	Empirical Mean		
LS	0.52592	0.19895	0.20353
LTS	1.02929	0.99787	1.00247
HME	0.97755	0.94349	0.94701
MGM	0.98716	0.95443	0.95808
SPN	1.00883	0.99493	0.99853
NEW	0.99718	0.97032	0.97408
	Empirical MSE		
LS	0.49607	0.67592	0.66471
LTS	0.10816	0.01214	0.01413
HME	0.02897	0.01026	0.00967
MGM	0.02774	0.00771	0.00710
SPN	0.02763	0.00658	0.00646
NEW	0.02654	0.00504	0.00465

Case 4 : 10% Bad & 10% Good Leverage Points			
	Empirical Mean		
LS	1.10501	0.28202	0.31877
LTS	1.00187	0.98782	0.98602
HME	1.01509	0.86511	0.87179
MGM	1.00159	0.89575	0.90231
SPN	0.99937	0.96360	0.96771
NEW	0.99878	0.94579	0.94184
	Empirical MSE		
LS	0.06997	0.59018	0.53777
LTS	0.12226	0.08231	0.09408
HME	0.02918	0.05460	0.05681
MGM	0.02826	0.04060	0.04284
SPN	0.02821	0.03945	0.04164
NEW	0.02700	0.03167	0.03372

The minimum MSE in each column is underlined.

Case 4. And, we can conclude that in general the proposed estimator has better performances in MSE than other competitors under heavy-tailed error distributions or in the presence of good or bad leverage points.

4. A NUMERICAL EXAMPLE

In this section we apply the robust estimators to the well-known stackloss data set presented by Brownlee (1965). The data describe the operation of a plant for the oxidation of ammonia to nitric acid, which was observed for 21 days. The data set is tabulated in Table 4.1 The stackloss(y) is regressed on the rate of operation(x_1), the cooling water inlet temperature(x_2), and the acid consentration(x_3). That is, the model is regarded as

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \ , \quad i = 1, \cdots, 21.$$

Table 4.1: Stackloss Data

Index (i)	*Rate* (x_1)	*Temperature* (x_2)	*Acid Concentration* (x_3)	*Stackloss* (y)
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

These real data have been examined by many statisticians. Most of the robust diagnostic results show that the four points (1, 3, 4, 21) are outliers (see Dodge, 1996). To analyze this data set, we consider the 6 estimators: LS, LTS, HME, MGM, SPN, and NEW, which were discussed in Section 3. The estimated values of parameters are tabulated in Table 4.2. The results of HME, MGM, SPN and NEW are similar. However the LS and LTS are somewhat different from the other four M- or GM-estimates. We now analyze the data in detail by using the proposed estimator.

Table 4.2: Estimates for the Stackloss Data

Method	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
LS	-39.9196	0.71564	1.29528	-0.15212
LTS	-35.8542	0.74559	0.37120	-0.00544
HME	-36.9235	0.72561	0.76187	-0.07077
MGM	-36.8406	0.72225	0.74149	-0.06469
SPN	-36.3481	0.71631	0.68559	-0.05375
NEW	-36.7290	0.71675	0.72402	-0.05822

Some diagnostic measures for the controversial points are computed and summarized in Table 4.3. Usually the Mahalanobis distance (MD) is defined by

$$MD = \sqrt{(\mathbf{z} - \widehat{\boldsymbol{\mu}})^T \widehat{\Sigma}^{-1} (\mathbf{z} - \widehat{\boldsymbol{\mu}})},$$

where $\mathbf{z}$ is a random vector of predictor variables given by $\mathbf{z} = (x_1, x_2, x_3)^T$, and $\widehat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$ are estimators of mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, respectively. In the table MD_i is the Mahalanobis distance based on the sample mean vector and sample covariance matrix, and RD_i is the robust Mahalanobis distance based on the MVE estimators. Note that in Table 4.3 the MVE estimates used in computing RD_i's are as follows:

$$\text{Center } \mathbf{m_z} = (56.706,\ 20.235,\ 85.529)^T$$

and

$$\text{Covariance Matrix } \mathbf{C_z} = \begin{pmatrix} 22.090 & 7.128 & 15.156 \\ 7.128 & 5.945 & 5.052 \\ 15.156 & 5.052 & 30.484 \end{pmatrix}$$

Another diagnostic measure for the leverageness is the diagonals of the hat matrix which are denoted by h_{ii}. The h_{ii} has the following relation with MD_i:

$$h_{ii} = \frac{MD_i^2}{n-1} + \frac{1}{n}.$$

According to Hoaglin and Welsch (1978), if $h_{ii} \geq 2p/n$ then the corresponding point can be considered as a leverage point.

Table 4.3: Diagnostic Measures for the Controversial Points

Index(i)	h_{ii}	MD_i	RD_i	$v(\mathbf{x}_i)$	$r_i(\widehat{\boldsymbol{\beta}}_0)$	$w(\mathbf{x}_i, r_i(\widehat{\boldsymbol{\beta}}_0)/\widehat{\sigma}_0)$
1	0.3016	2.2536	5.6985*	0.5365	8.7112	0.2879
2	0.3174	2.3247	5.8108*	0.5262	3.7671	0.2769
3	0.1746	1.5937	4.3264*	0.7067	8.1639	0.4994
4	0.1285	1.2719	1.6376	1.0000	9.2309	1.0000
21	0.2845	2.1768	3.7698*	0.8111	-8.3106	0.6578

* Distances exceeding the cutoff value $\sqrt{\chi^2_{3,.975}} = 3.06$ are marked by *.

The four points 1, 2, 3 and 21 are identified as leverage points in view of robust Mahalanobis distance RD_i or $v(\mathbf{x}_i)$. But the regression diagnostic based on the diagonals h_{ii} of the hat matrix fails to identify these leverage points. No points exceed the cutoff value $2p/n = 0.381$. To investigate the outliers in y-direction, the residuals $r_i(\widehat{\boldsymbol{\beta}}_0)$ based on the initial estimate $\widehat{\boldsymbol{\beta}}_0$ are computed. The weights $w(\mathbf{x}_i, r_i(\widehat{\boldsymbol{\beta}}_0)/\widehat{\sigma}_0)$ used in the proposed estimator are also included in the table to identify 'good' and 'bad' leverage points. By inspecting the residuals, there are 4 extreme outliers (1, 3, 4, 21) and a mild outlier (2).

Rousseeuw and Zomeren (1990) proposed a robust diagnostic method based on the LMS estimator and the robust Mahalanobis distance. Figure 4.1 shows the robust diagnostic result based on the proposed estimator. In the figure the horizontal axis is the robust Mahalanobis distance RD_i, and the vertical axis is the standardized residual $r_i(\widehat{\boldsymbol{\beta}})/\widehat{\sigma}$, where the residual $r_i(\widehat{\boldsymbol{\beta}})$ is obtained from the proposed fit. If the absolute value of a standardized residual is larger than 2.5, the corresponding point can be regarded as a regression outlier. While, if a robust Mahalanobis distance exceeds the cutoff value $b = \sqrt{\chi^2_{3,.975}} = 3.06$, then the corresponding point is considered as a leverage point. In a view of these

points there are 4 leverage points and one regression outlier. Among them, the fourth observation (4) is a regression outlier, the second observation (2) may be classified as a good leverage point, and the observations (1, 3, 21) as bad leverage points. Rousseeuw and van Zomeren (1990) classified the second observation (2) as a bad leverage point, while the proposed method classifies it as a good leverage point.

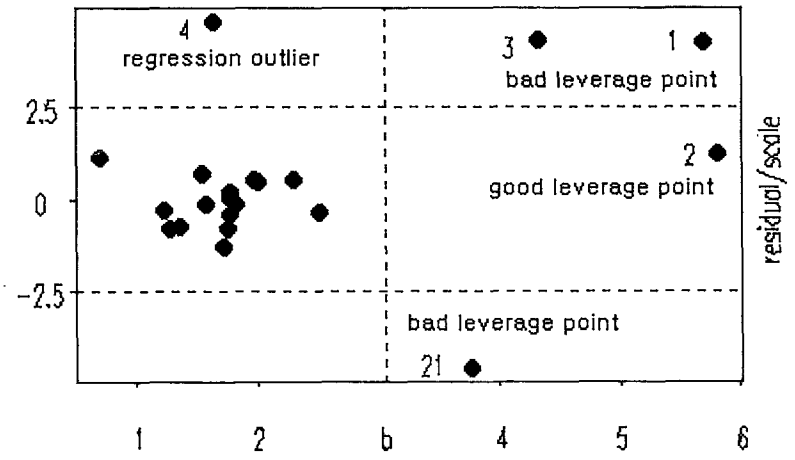


Figure 4.1 : Diagnostic Results for the Stackloss Data.

ACKNOWLEDGEMENTS

The authors would like to thank a referee for his/her careful reading of the manuscript and for helpful comments.

REFERENCES

Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering* (2nd ed.), New York: John Wiley.

Coakley, C.W. and Hettmansperger, T.P. (1993). "A Bounded Influence, High Breakdown, Efficient Regression Estimator," *Journal of the American Statistical Association*, 88, 872-880.

Dodge, Y. (1996). "The Guinea Pig of Multiple Regression," In *Robust Statistics, Data Analysis, and Computer Intensive Methods*, edited by H. Rieder, Lecture Notes in Statistics No.109, 91-117. New York: Springer-Verlag.

Hampel, F.R. (1971). "A General Qualitative Definition of Robustness," *Annals of Mathmetical Statistics*, 42, 1887-1896.

Hampel, F.R. (1974). "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383-393.

Hoaglin, D.C. and Welsch, R.E. (1978). "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17-22.

Huber, P.J. (1973). "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *The Annals of Statistics,* 1, 799-821.

Krasker, W.S. and Welsch, R.E. (1982). "Efficient Bounded-Influence Regression Estimation," *Journal of the American Statistical Association*, 77, 595-604.

Marazzi, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics,* Wadsworth & Brooks, California.

Maronna, R.A. and Yohai, V.J. (1981). "Asymptotic Behavior of General M-estimates for Regression and Scale with Random Carriers," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 58, 7-20.

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

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

Rousseeuw, P.J. and Wagner, J. (1994). "Robust Regression with a Distributed Intercept Using Least Median of Squares," *Computational Statistics & Data Analysis*, 17, 65-75.

Rousseeuw, P.J. and Yohai, V. (1984). "Robust Regression by Means of S-Estimators," In *Robust and Nonlinear Time Series*, edited by J. Franke, W. Hardle, and R.D. Martin, Lectures Notes in Statistics No. 26, 256-272. New York: Springer-Verlag.

Ruppert, D. and Carroll, R.J. (1980). "Trimmed Least Squares Estimation in Linear Model," *Journal of the American Statistical Association*, 75, 828-838.

Simpson, D.G., Ruppert, D. and Carroll, R.J. (1992). "On One-Step GM Estimates and Stability of Inferences in Linear Regression," *Journal of the American Statistical Association*, 87, 439-450.

Song, M.S. and Kim, J.H (1997). "A Study on a One-Step Pairwise GM-Estimator in Linear Model," *Journal of the Korean Statistical Society*, 26, 1-22.

Song, M.S., Park, C.S. and Nam, H.S. (1996). "A High Breakdown and Efficient GM-Estimator in Linear Models," *Journal of the Korean Statistical Society*, 25, 471-487.