Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Outlier detection by means of robust regression estimators for use in engineering science*

Serif HEKIMOGLU[†‡1], R. Cuneyt ERENOGLU[†1], Jan KALINA[2]

([1]*Department of Geodesy and Photogrammetry Engineering, Yildiz Technical University, Istanbul 34349, Turkey*)

([2]*Department of Probability and Mathematical Statistics, Charles University, Praha 18675, Czech Republic*)

[†]E-mail: {hekim, ceren}@yildiz.edu.tr

Received Feb. 29, 2008; Revision accepted Nov. 6, 2008; Crosschecked Dec. 29, 2008

Abstract: This study compares the ability of different robust regression estimators to detect and classify outliers. Well-known estimators with high breakdown points were compared using simulated data. Mean success rates (MSR) were computed and used as comparison criteria. The results showed that the least median of squares (LMS) and least trimmed squares (LTS) were the most successful methods for data that included leverage points, masking and swamping effects or critical and concentrated outliers. We recommend using LMS and LTS as diagnostic tools to classify outliers, because they remain robust even when applied to models that are heavily contaminated or that have a complicated structure of outliers.

Key words: Linear regression, Outlier, Mean success rate (MSR), Leverage point, Least median of squares (LMS), Least trimmed squares (LTS)

doi:10.1631/jzus.A0820140 **Document code:** A **CLC number:** O21

INTRODUCTION

In engineering science, fitting a model to noisy data is a common task since all real data are contaminated. The most common form of regression analysis is the least squares (LS) method, which achieves optimum results when data include only normally distributed random errors. Unfortunately, this method is extremely sensitive to outliers and breaks down when the data include a leverage point (Huber, 1981; Hampel *et al.*, 1986; Rousseeuw and Leroy, 1987; Shevlyakov and Vilchevski, 2001) or a gross error in the *y*-direction (Hekimoglu, 2005). Non-robust estimators fail completely in estimating the regression parameters for contaminated data, while robust methods can successfully detect bad observations.

Several robust estimators have been developed in recent decades (Huber, 1981; Hampel *et al.*, 1986; Rousseeuw and Leroy, 1987). Non-parametric estimators insensitive to outliers have also been developed including the Theil-Sen estimator (Sen, 1968). The most commonly used methods with a high breakdown point include the repeated median (Siegel, 1982), the least median of squares (LMS) (Rousseeuw, 1984), and the least trimmed squares (LTS) (Rousseeuw and Leroy, 1987). In all these methods the parameters are obtained by fitting sub-samples. When many outliers occur, these methods may not fit the data correctly. Thus, the effectiveness of any method may be defined by its ability to detect outliers.

In this study, we investigated the local reliability of the robust regression methods against the outliers, leverage points and/or gross errors in the *y*-direction. An important question is how to compare the effectiveness of these methods when the outliers are small. To measure the local reliability of a robust method the idea of the mean success rate (MSR) was introduced (Hekimoglu and Koch, 1999). MSR was also applied to outlier detection in linear regression models (Hekimoglu and Erenoglu, 2005) and in geomatics engineering models (Hekimoglu and Erenoglu, 2007).

‡ Corresponding author

* Project (No. 28-05-03-03) supported by the Yildiz Technical University Research Fund, Turkey

There are some general problems with outlier detection using the robust methods. The success of the identification of the outliers is greatly affected by the presence of any leverage points, masking effects, swamping effects, critical outliers, or gross errors in the *y*-direction. Moreover, their success also depends on how an outlier is defined. In this study, we compared the ability of robust methods to detect outliers in linear regression models using MSR. We also determined how many outliers could be detected reliably.

The rest of this paper is organized as follows. We first introduce the mathematical model and then provide information about the concepts of breakdown points and outliers. The definition of the MSR and its application to outlier detection in linear regression models are discussed. The results of Monte Carlo simulations and real data experiments are presented and discussed. Finally, the conclusions are summarized.

MODEL

We consider the classical linear model given by

$$y_i = b_0 + b_1 x_{i1} + ... + b_p x_{ip} + e_i,\ i = 1,2,...,n, \quad (1)$$

where n is the size of the sample (number of cases), y_i is the response variable, x_{ip} are the regressors, b_0, b_1, ..., b_p are regression parameters and p+1 is the number of regression parameters. In classical theory, the random error e_i is normally distributed with mean zero and variance σ^2 and is independent of the regressors. The residual r_i of the ith case is the difference between what is actually observed and what is estimated:

$$r_i = y_i - \hat{y}_i, \quad (2)$$

where $\hat{y}_i$ is the estimated value of y_i using any estimator. The residual vector $\boldsymbol{r}$ of the least squares estimate is also given with the hat matrix $\boldsymbol{H}=\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}$ by

$$\boldsymbol{r}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}, \quad (3)$$

where $\boldsymbol{A}$ is the $n\times(p+1)$ design matrix, $\boldsymbol{I}$ is the $n\times n$ identity matrix, and $\boldsymbol{y}$ is the $n\times 1$ data vector.

BREAKDOWN POINTS

The concept of a breakdown point was first introduced by Hampel (1968; 1971) and later used as a functional analytical procedure by Huber (1981). A simplified version for finite samples was presented by Donoho (1982) and Donoho and Huber (1983). The breakdown point is the smallest amount of contamination that may cause an estimator to take an arbitrarily large aberrant value. The version of the breakdown point for finite samples provides an opportunity to compare different statistical estimators. This approach has been used in regression analysis and in scale and location models (Hampel, 1975; Stahel, 1981; Donoho, 1982; Rousseeuw, 1984; 1985; Lopuhaa and Rousseeuw, 1991; Davies, 1993; Gather and Hilker, 1997; Davies and Gather, 2005). The maximal breakdown point of a regression equivariant estimator is given by $\{[(n-p-1)/2]+1\}/n$ (Rousseeuw and Leroy, 1987).

CONCEPT OF OUTLIERS

In statistical literature, outlier analysis always plays an important role (Barnett and Lewis, 1994). We denote by e_i random errors with normal distribution and zero expectation, α the significance level, σ the standard deviation of the normal distribution, and $z_{1-\alpha/2}$ the corresponding critical value as follows:

$$P\left(\left|\frac{e_i-\mu}{\sigma}\right| > z_{1-\alpha/2}\right) = \alpha. \quad (4)$$

Hekimoglu (1997) considered observations that fulfill $|r_i|>3\sigma$ to be outliers.

In this study, we distinguish between random and non-random (structural) outliers. Random outliers are those that occur accidentally during the measuring process. Their signs and magnitudes change randomly according to a uniform distribution. Non-random outliers are defined here as those caused by the same unknown disturbing source in the measuring process. All of them have the same sign although their magnitudes can change randomly, according to any distribution.

Masking and swamping effects

It is well known that masking and swamping effects are common problems in outlier detection procedures (Hadi and Simonoff, 1993; Hekimoglu, 2005). Let the observations include two bad outliers $\bar{y}_i$ and $\bar{y}_k$. The contaminated residuals $\bar{r}_i$ can be written as follows:

$$\bar{r}_i = -(1-h_{ii})\bar{y}_i + h_{ik}\bar{y}_k + \sum_{j=1}^{n} h_{ij} y_j, \quad j \neq i,\ k \neq i,\ j \neq k,\ i = 1,2,...,n, \tag{5}$$

where h_{ii} is the ith diagonal element of the hat matrix. If the second term on the right-hand side of Eq.(5) has the opposite sign to the first term, then the two terms may cancel each other. Hence a bad observation becomes a good observation. This is called the masking effect. Let the observation y_i be a good one. If the contributions of the second and third terms are added, the observation y_i might become a bad observation. This is the swamping effect.

Leverage points

If the x-value for a particular observation lies far away from the x-values of the majority of the observations, this point is called a leverage point (Rousseeuw and Leroy, 1987) or an outlier in the x-direction. Its partial redundancy number, defined by $1-h_{ii}$, has the smallest value among all the observations.

Critical outliers

We first assume a simple regression (Table 1). Let the outliers in a linear regression model lie close together with respect to their x-values and let their partial redundancy numbers $1-h_{ii}$ be smaller than those of the other observations. In these circumstances we call them critical outliers. This kind of outlier is considered as a special case of the non-random outliers. The two possible outliers are considered as being $(\bar{y}_1, \bar{y}_2)$ or $(\bar{y}_9, \bar{y}_{10})$ in a regression model and the three possible outliers as $(\bar{y}_1, \bar{y}_2, \bar{y}_3)$ or $(\bar{y}_8, \bar{y}_9, \bar{y}_{10})$.

Equileverage design (ELD)

It is well known that the breakdown point of Huber's method is zero. To eliminate bad effects of the leverage points, the concept of equiredundancy design was first proposed by Staudte and Sheather (1990). If, and only if, each observation has the same geometrical and stochastic effects on the other observations, the diagonal elements h_{ii} of the hat matrix $\boldsymbol{H}$ become equal. To achieve this, we multiplied the weights by $(1-h_{ii})/h_{ii}$. The balanced weights, which provide the equiredundancy, can then be used as the pseudo weights of observations for Huber's method.

Table 1 The x_i, y_i values of a simple regression and the partial redundancy numbers

Point	x_i	y_i	$1-h_{ii}$
1	1	2	0.66
2	2	3	0.75
3	3	4	0.82
4	4	5	0.87
5	5	6	0.90
6	6	7	0.90
7	7	8	0.87
8	8	9	0.82
9	9	10	0.75
10	10	11	0.66

OUTLIER DETECTION

The concept of the MSR introduced by Hekimoglu and Koch (1999) was crucial in our simulation study. To explain the idea, let us take a good sample $\boldsymbol{o}$ coming from a linear model with a known variance of errors. A contaminated sample $\bar{\boldsymbol{o}}$ may be obtained by replacing any m of the original measurements from the sample $\boldsymbol{o}$ with arbitrary values. An estimator denoted as T is applied to the contaminated sample $\bar{\boldsymbol{o}}$.

The estimator T is considered successful if the absolute value of the residual of each contaminated observation is greater than $z_{1-\alpha/2}\sigma$. An outlier interval $int(\sigma)$ for $\boldsymbol{o}$ will be any interval of the form

$$int(\sigma)_{kl} = l\sigma - k\sigma = (l-k)\sigma,\ l > k > z_{1-\alpha/2}. \tag{6}$$

Let a certain contaminated sample $\bar{\boldsymbol{o}}$ contain m outliers of any magnitude in the given outlier interval $int(\sigma)$. The partial reliability of an estimator T that is applied to this contaminated sample $\bar{\boldsymbol{o}}$ is measured by the success rate (SR), namely as the number of successful identifications divided by the number of

the contaminated samples (experiments). So the SR is given by identifying all of the m outliers in the contaminated sample $\bar{\boldsymbol{o}}$ depending on the given interval $int(\sigma)$ of the given sample $\boldsymbol{o}$.

Many corrupted samples $\bar{\boldsymbol{o}}$ can be generated from each sample $\boldsymbol{o}$. Therefore, the definition of reliability must be generalized to a ratio of the sum of individual SRs of particular samples to the number of samples. This is called the MSR of an estimator for each number of outliers (Hekimoglu and Koch, 1999). Thus, the SR is obtained from many experiments using only one sample. However, the MSR is the mean value of SRs computed from many samples.

For instance, let there be two bad observations in a sample. First, we apply an estimator to the corrupted sample. If the estimator can detect both bad observations exactly, it is accepted that the estimator is successful. Otherwise, the estimator is considered unsuccessful. The definition of MSR can be extended easily to a dataset without outliers. Then the MSRs are equal to the probability that an absolute value of a residual exceeds the critical value $z_{1-\alpha/2}\sigma$.

MONTE CARLO SIMULATION

Simulations were carried out for different datasets and different situations to study the performance of different robust regression estimators in outlier detection. We programmed the algorithms for different robust estimators used during this study in the language of technical computing—MATLAB. MATLAB toolboxes were also used for the exact solution of LMS and LTS (Stromberg, 1993). Huber's M-estimator and LMS were implemented using functions of rlm and rqs, respectively, in the R MASS package. The LTS was implemented in function ltsReg in the R robustbase package. These methods are also available in the SAS procedure ROBUSTREG (Chen, 2002). We began by simulating data with a known variance of the random errors and this known value of the variance was also used in outlier detection.

Let a simple straight line be defined as

$$y_i = b_0 + b_1 x_i,\ i = 1,2,...,n_1,\ \text{with } b_0=1,\ b_1=1, \quad (7)$$

and a multiple linear regression model be defined as

$$z_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + b_3 x_{3j} + b_4 x_{4j},\ j = 1,2,...,n_2,$$
$$\text{with } b_0=2,\ b_1=-1,\ b_2=0.5,\ b_3=1.2,\ b_4=1.5. \quad (8)$$

The regressors were generated from a uniform distribution on [0, 1].

Observations without outliers

The random errors of e_{1i} (i=1,2,...,n_1) and e_{2j} (j=1,2,...,n_2) were generated from the normal distribution $\boldsymbol{e}\sim N(\mu=0,\ \sigma^2=0.02^2)$ by a random number generator, a subroutine of MATLAB. To obtain 'good' observations y_i' and z_j', the random errors e_i (i.e., e_{1i} or e_{2j}) were added to the y_i- or z_j-values as follows:

$$y_i' = y_i + e_{1i}, \quad (9)$$
$$z_j' = z_j + e_{2j}. \quad (10)$$

One hundred sets for y_i' and z_j' were generated by creating a subset of the random error vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$.

Bad observations

To simulate a 'bad' observation such as $\bar{y}_i$ or $\bar{z}_j$, the random error of a 'good' observation was replaced by an outlier δ_y or δ_z. Thus, the magnitude δ_y or δ_z of an outlier was added to the y_i- or z_j-value (e.g., $\bar{y}_i = y_i + \delta_{y_i}$ or $\bar{z}_j = z_j + \delta_{z_j}$). Outliers were generated by the uniform distribution for a given interval in the outlier region as in (Hekimoglu and Koch, 1999). We distinguished two kinds of outliers: random and non-random.

1. Random outliers

The magnitude δ_y of one random outlier was generated by the uniform distribution for a given interval $int(\sigma)$ in the outlier region as follows:

$$int(\sigma) = 3\sigma < \left|\delta_{y_k}\right| < 6\sigma,\quad \delta_{y_k} = \text{sign}(t_1)\delta_{\bar{y}_k}, \quad (11)$$

$$\text{sign}(t_1) = \begin{cases} +, & 0.5 < t_1 \le 1, \\ -, & 0 < t_1 \le 0.5, \end{cases} \quad (12)$$

$$\delta_{\bar{y}_k} = 3\sigma + t_2\Delta,\ k = n_1 t_3,\ \Delta = 6\sigma - 3\sigma = 3\sigma,\quad 0 < t_2 \le 1,\ 0 < t_3 \le 1, \quad (13)$$

where t_1, t_2 and t_3 are distributed uniformly and Δ is

the length of outlier interval *int*(σ). Note that k must be an integer because it is an observation number. The interval *int*(σ) for small outliers lies between 3σ and 6σ, and for large outliers between 6σ and 12σ. For the last case, Eq.(13) is changed to

$$\delta_{\bar{y}_k} = 6\sigma + t_2\Delta,\ k = n_1 t_3,\ \Delta = 12\sigma - 6\sigma = 6\sigma,\quad 0 < t_2 \le 1,\ 0 < t_3 \le 1. \tag{14}$$

The magnitudes δ_{y_i} and δ_{y_k} of two or more random outliers were generated by the uniform distribution for a given interval *int*(σ) in the outlier region (Hekimoglu and Koch, 1999). In the multiple linear regression, the magnitude δ_z and the magnitudes of δ_{z_i} and δ_{z_k} were generated for one and two random outliers, respectively, as done in simple regression.

This algorithm was computed 100 times for each sample. First, 100 different y_i' (or z_j') samples were created by a subset of the random error vectors $\boldsymbol{e}_1$ (or $\boldsymbol{e}_2$). Then 100 different contaminated samples of $\bar{y}_i$ (or $\bar{z}_j$) were simulated for each sample of y_i' (or z_j') by randomly changing the number of observations, the magnitudes of outliers, and the signs of the outliers. Thus, we computed 100 different SRs. The mean value of these is called the MSR and represents the reliability of the robust estimator.

2. Non-random outliers

The magnitude δ_{y_i} (or δ_{z_j}) of a non-random outlier was also generated by the uniform distribution for a given interval in the outlier region. Two or more outliers were generated as before. When the signs of multiple outliers were the same, i.e., all plus or all minus, we described the outliers as non-random.

The random errors were simulated with a standard deviation of 0.02 in each situation. In this study, the Theil-Sen estimator, the LMS and LTS methods, the repeated median method and Huber's method were each applied as estimators for 0, 1, 2, 3, 4 and 5 outliers (either random or non-random) that lay between 3σ and 6σ and then between 6σ and 12σ. First, the MSRs for all the methods for identifying outliers were computed using both simple and multiple linear models. Then, leverage points were added to these models. The LTS estimator was computed with the trimming constant h ensuring the maximal breakdown point, namely

$$h = [n/2] + [(p+2)/2], \tag{15}$$

where '[ ]' denotes the integer function.

Simple regression model

The simple regression model with one independent variable was used, as shown in Eq.(7). The data are shown in Fig.1 (n_1=10).

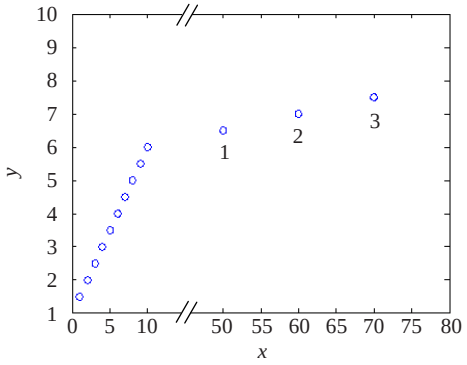

Fig.1 Plot of x and y values for simple regression
1, 2 and 3 are leverage points

The following methods were applied to the samples to detect the outliers for various cases:

(1) LMS (Rousseeuw, 1984; Stromberg, 1993).

(2) LTS (Rousseeuw and Leroy, 1987).

(3) Theil-Sen estimator (Theil, 1950).

(4) Repeated median (RM) method (Siegel, 1982).

(5) Huber's M-estimation method (Huber, 1981).

For the Theil-Sen estimator, the slope of the line of fit is taken to be the median of the set of C_N^2 slopes that result from passing a line through each pair of distinct points in the dataset. The line's intercept can be estimated similarly. It can be shown that the breakdown point of the Theil-Sen estimator is about 29.3%. The repeated median method fits to a set of N distinct points in the plane $\{p_1, p_2, ..., p_n\}$ and is defined as follows. For each point $p_i=(x_i, y_i)$, let θ_i denote the median of the $N-1$ slopes of the lines passing through p_i and each of the other points of the set. The repeated median slope, θ^*, is defined as the median of the multi-set $\{\theta_i\}$ (the RM-intercept is defined analogously, in terms of line intercepts). The repeated median estimator has a breakdown point of 50% (Kamgar-Parsi and Netanyahu, 1989).

For Huber's method, the tuning constant c is taken as 1.5. The MSRs of all the estimators were computed for random and non-random outliers separately (Tables 2 and 3, respectively). The LMS and LTS methods were shown to flag outliers even when the sample had none. This is obviously a significant disadvantage of these methods. The other methods identified fewer false outliers. Furthermore, the MSRs for random outliers were greater than those for non-random ones. The MSRs decreased as the number of outliers increased and increased when the magnitude of the outliers increased.

When there was no leverage point in the data, the Theil-Sen estimator had the highest overall MSR among all the estimators. The LMS and LTS methods required more computing time than the other methods. When the data included five outliers, i.e., five good observations and five bad observations, all the methods failed to detect the outliers.

Next, we considered a contaminated sample containing leverage points. For one leverage point, the 11th observation was considered with (y_{11}=12, x_{11}=50) and for three leverage points of the 11th, 12th and 13th (y_{11}=12, x_{11}=50, y_{12}=13, x_{12}=100, y_{13}=14, x_{13}=150). The MSRs for one and three leverage points are shown in Tables 4 and 5, respectively. The results showed when leverage points existed in the data, the LMS and LTS methods had the largest overall MSR for detecting outliers. The second most successful method was the repeated median method and the third was the Theil-Sen estimator. Huber's method broke down when the data included two or more leverage points or when the x-value of the first leverage point reached 500 (i.e., x_{11}=500). When the number of leverage points in the data increased, the number of detectable outliers in the y-direction decreased. If the sample included leverage points, the MSRs of the robust methods decreased.

Multiple regression model

The multiple regression model given in Eq.(8) was generated with four independent variables and n_2=13. This model was contaminated with different numbers (1, 2, 3, 4 or 5) of outliers. The MSRs for each estimator were computed for random outliers (Table 6). Clearly, the MSRs had decreased significantly compared with those for a simple regression analysis (Table 2). The LMS and LTS methods identified extra outliers when none existed. If the number of outliers was increased, the MSRs decreased. Moreover, if the magnitude of the outlier increased, the MSRs also increased. If there was no leverage point and no outlier in the data, Huber's method did not identify an outlier, unlike the other methods tested.

Table 2 MSRs for simple regression in the case of random outliers and no leverage point

Method	*MSR* (%)										
	$\delta>3\sigma$	$\delta=3\sigma$~6σ					$\delta=6\sigma$~12σ				
	n_o=0	n_o=1	2	3	4	5	n_o=1	2	3	4	5
Theil-Sen	4	80	69	54	60	0	95	92	85	96	0
Least median of squares	16	73	67	60	52	0	84	85	88	97	0
Least trimmed squares	14	79	71	65	53	0	91	92	94	99	0
Repeated median	6	79	70	55	41	3	93	91	86	77	6
Huber's	0	81	74	56	41	2	97	96	90	77	5

δ: magnitude of the outlier; n_o: number of outliers

Table 3 MSRs for simple regression in the case of non-random outliers and no leverage point

Method	*MSR* (%)							
	$\delta=3\sigma$~6σ				$\delta=6\sigma$~12σ			
	n_o=2	3	4	5	n_o=2	3	4	5
Theil-Sen	71	61	48	0	87	90	97	0
Least median of squares	69	60	47	0	85	87	95	0
Least trimmed squares	74	65	53	0	89	91	96	0
Repeated median	63	35	15	0	88	72	52	0
Huber's	57	23	3	0	90	57	13	0

δ: magnitude of the outlier; n_o: number of outliers

Table 4 MSRs for simple regression in the case of random outliers and one leverage point

Method	*MSR* (%)										
	$\delta>3\sigma$	$\delta=3\sigma\sim6\sigma$					$\delta=6\sigma\sim12\sigma$				
	$n_o=0$	$n_o=1$	2	3	4	5	$n_o=1$	2	3	4	5
Theil-Sen	5	73	59	33	41	0	88	79	53	54	0
Least median of squares	16	73	60	47	40	0	84	82	71	52	0
Least trimmed squares	14	78	64	52	44	0	89	86	76	57	0
Repeated median	5	77	67	40	30	0	92	89	68	51	0
Huber's	3	80	78	44	21	0	93	89	65	40	0

δ: magnitude of the outlier; n_o: number of outliers

Table 5 MSRs for simple regression in the case of random outliers and three leverage points

Method	*MSR* (%)										
	$\delta>3\sigma$	$\delta=3\sigma\sim6\sigma$					$\delta=6\sigma\sim12\sigma$				
	$n_o=0$	$n_o=1$	2	3	4	5	$n_o=1$	2	3	4	5
Theil-Sen	28	10	5	1	0	0	11	6	1	0	0
Least median of squares	12	76	43	30	0	0	85	88	46	0	0
Least trimmed squares	13	81	47	35	0	0	89	93	52	0	0
Repeated median	21	54	42	11	0	0	66	55	17	0	0
Huber's	0	0	0	0	0	0	0	0	0	0	0

δ: magnitude of the outlier; n_o: number of outliers

Table 6 MSRs for multiple regression in the case of random outliers and no leverage point

Method	*MSR* (%)										
	$\delta>3\sigma$	$\delta=3\sigma\sim6\sigma$					$\delta=6\sigma\sim12\sigma$				
	$n_o=0$	$n_o=1$	2	3	4	5	$n_o=1$	2	3	4	5
Theil-Sen	8	64	55	57	43	0	76	74	68	35	0
Least median of squares	26	45	37	27	14	0	56	61	66	65	0
Least trimmed squares	28	49	40	31	15	0	60	64	67	66	0
Repeated median	6	63	56	34	23	0	74	73	69	62	0
Huber's	0	61	40	19	10	0	91	78	52	33	2

δ: magnitude of the outlier; n_o: number of outliers

Furthermore, one and three leverage points were added to the multiple regressions. The MSRs are shown in Tables 7 and 8, respectively. From these results it follows that the LMS and LTS methods were the estimators most resistant to leverage points. If the data included no leverage points, all the methods were unsuccessful in detecting the five outliers. However, it may not mean that the estimators break down.

The MSRs for the Theil-Sen estimator and repeated median methods were higher than those of the LMS and Huber's methods for data without leverage points. When there were five outliers in the data, all of the methods failed to detect outliers in the *y*-direction. The LMS and LTS methods had the highest MSRs overall when leverage points existed in the data. The repeated median method had the second highest MSRs overall. It is also noted that the number of detectable outliers decreased for the Theil-Sen estimator when the number of leverage points increased. Consequently, these simulation results verify the known theoretical properties of the robust estimators. For example, Huber's estimator has a low breakdown point, the LMS and LTS are robust methods with high breakdown points, and the Theil-Sen estimator has a lower breakdown point than LMS or LTS.

Experiment with real data

We used the pilot-plant dataset from (Daniel and Wood, 1971). The response variable y_i corresponds to the acid content determined by titration, and the

Table 7 MSRs for multiple regression in the case of random outliers and one leverage point

Method	*MSR* (%)										
	$\delta>3\sigma$	$\delta=3\sigma$~6σ					$\delta=6\sigma$~12σ				
	$n_o=0$	$n_o=1$	2	3	4	5	$n_o=1$	2	3	4	5
Theil-Sen	5	58	48	26	21	0	70	63	43	25	0
Least median of squares	27	52	43	28	19	0	68	52	34	23	0
Least trimmed squares	28	56	46	30	21	0	69	55	36	25	0
Repeated median	9	61	54	40	0	0	73	72	54	0	0
Huber's	0	0	0	0	0	0	0	0	0	0	0

δ: magnitude of the outlier; n_o: number of outliers

Table 8 MSRs for multiple regression in the case of random outliers and three leverage point

Method	*MSR* (%)										
	$\delta>3\sigma$	$\delta=3\sigma$~6σ					$\delta=6\sigma$~12σ				
	$n_o=0$	$n_o=1$	2	3	4	5	$n_o=1$	2	3	4	5
Theil-Sen	31	9	4	0	0	0	9	4	0	0	0
Least median of squares	22	54	38	0	0	0	66	62	0	0	0
Least trimmed squares	24	57	41	0	0	0	68	61	0	0	0
Repeated median	14	44	33	0	0	0	53	44	0	0	0
Huber's	0	0	0	0	0	0	0	0	0	0	0

δ: magnitude of the outlier; n_o: number of outliers

regressors x_i are the organic acid content determined by extraction and weighing. The data are plotted in Fig.2 and include 20 points.

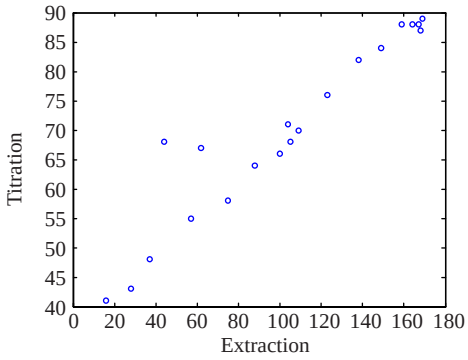

Fig.2 Plot of extraction vs titration values

The simple regression model with one independent variable was fitted to the data by using the robust estimators. For Huber's method, the estimated variance computed from the LSE is used instead of the a priori variance, σ^2. The residuals from these estimators are shown in Table 9. The results showed that the residuals of the 4th and 6th data points were significantly greater than the others although the residuals from each estimator differed. Thus, the corresponding points are outliers and all the methods successfully identified them. We also computed the median of absolute deviations of the residuals (Rousseeuw and Leroy, 1987), shown in the last row of Table 9. By taking the median absolute deviation (MAD) value of residuals from Huber's method, i.e., $\hat{\sigma}$ =1.8017, as the estimated value of the standard deviation σ, we see that the Theil-Sen and repeated median methods also flagged the 10th observation as an outlier ($r_{10}>3\hat{\sigma}$) while the other methods did not.

However, we cannot only measure the ability of the methods to identify outliers in the case study. Residuals help us to also find out which observation point includes the gross error.

DISCUSSION

Thus far, we have assumed a known variance of random errors. Otherwise, the MAD estimator can be used instead of the value of the (unknown) variance (Hampel *et al*., 1986). It can be computed as follows:

$$MAD = 1.4826 \times median(|r_i|),\ \ i = 1,2,...,n, \quad (16)$$

where $|r_i|$ is the absolute residual. Thus, we can discuss the effect of the variance factor on outlier detection.

Table 9 Residuals of the methods for the real data experiment

Experiment No.	Residual				
	Theil-Sen	LMS	LTS	Repeated median	Huber's method
1	−0.1928	1.6170	1.0810	0.1894	−0.8872
2	1.7671	0.0851	−0.4400	2.2902	0.7053
3	−0.6758	2.6809	2.1597	−0.1024	−1.8688
4	−8.7962	12.0851	11.5967	−7.8000	−11.0914
5	1.7609	1.6809	1.1964	2.8074	−0.6655
6	−14.9907	18.8298	18.3555	−13.8133	−17.7583
7	2.9893	1.0638	0.5950	4.2371	0.0380
8	3.1698	−1.0426	−1.5606	3.7835	1.8719
9	3.9291	0.7660	0.3135	5.3883	0.4267
10	5.3920	−1.0638	−1.5256	6.7305	2.2046
11	−1.8641	2.8298	2.2821	−1.6329	−2.1648
12	2.6127	−0.6383	−1.1603	3.1761	1.4460
13	−1.8039	2.1277	1.5635	−1.7841	−1.5535
14	3.9553	−1.0638	−1.5624	4.8207	2.0013
15	1.7069	0.7872	0.2786	2.4414	0.0940
16	−0.3610	0.5319	−0.0361	−0.3915	0.0206
17	0.0819	−0.0638	−0.6358	0.0011	0.5947
18	0.5047	−0.4255	−0.9959	0.4440	0.9651
19	−0.6897	1.3191	0.7628	−0.5692	−0.7017
20	1.5047	−1.7447	−2.3158	1.4440	1.9651
Rousseeuw's *MAD*	2.6061	1.9715	1.9764	2.4674	1.8017

MAD: median absolute deviation

Table 10 MSRs for the Theil-Sen estimator with and without MAD for simple regression

Method	*MSR* (%)						
	$\delta>3\sigma$	$\delta=3\sigma$~6σ			$\delta=6\sigma$~12σ		
	$n_o=0$	$n_o=1$	2	3	$n_o=1$	2	3
With MAD	4	59	45	26	84	86	74
Without MAD	4	80	69	54	95	92	85

δ: magnitude of the outlier; n_o: number of outliers

As an example, we applied this procedure to the Theil-Sen estimator for the simple regression Eq.(7). The MAD was obtained using residuals according to Eq.(16), and the computed MSRs using Theil-Sen estimation are shown in the first row of Table 10. The MSRs decreased compared with those in the second row of Table 10, where the variance was known. Note that the second row of Table 10 is the same as the first row of Table 2.

It can be seen that Huber's estimator broke down even if the data included only one leverage point. When the ELD is considered, the MSRs of Huber's method to detect the outlier increase and it does not break down (Hekimoglu, 2005). Thus, we first multiplied the weights by $(1-h_{ii})/h_{ii}$, and the balanced weights were obtained. In addition, Huber's method with the balanced weights was applied to the simple regression where the sample included two leverage points. The MSR of Huber's method with ELD was significantly greater than that of Huber's method, but still smaller than the MSR of the LMS (Table 11).

Why does a robust estimator fail to identify the outlier(s) correctly? The identification of multiple outliers is complicated owing to masking and swamping effects. In addition, the performance of the estimators strongly depends on the following variables or properties: the partial redundancy numbers of the observations, the magnitude of the outliers, the number of outliers, the type of the outliers (random or non-random), random errors, the position of outliers

Table 11 MSRs of Huber's method with ELD for simple regression in the case of random outliers and two leverage points

Method	*MSR* (%)						
	$\delta>3\sigma$	$\delta=3\sigma$~6σ			$\delta=6\sigma$~12σ		
	$n_o=0$	$n_o=1$	2	3	$n_o=1$	2	3
Without ELD	0	0	0	0	0	0	0
With ELD	9	58	48	26	70	63	43

δ: magnitude of the outlier; n_o: number of outliers. ELD: equileverage design

in the data (e.g., if two outliers are close to each other or not) and the tuning constant *c* (Hekimoglu, 2005).

Let us discuss the situation when a robust method cannot identify outliers. Let the sample be given as $\{(y_1, x_1), (y_2, x_2), \ldots, (y_{10}, x_{10})\}$ for simple regression. Let $\bar{y}_3$ be contaminated by an outlier. If a robust method is applied to the sample, the following results may arise: $\bar{y}_3$ is identified as a bad observation; $\bar{y}_3$ and another observation are identified as bad ones; instead of $\bar{y}_3$, another observation is identified as a bad one; no outlier is identified.

Why does this occur? Robust methods with a high breakdown point are not efficient against small outliers in the *y*-direction (Hekimoglu, 2005). They detect outliers even though the sample includes no outlier due to estimating the parameters. For detecting outliers, more efficient results are obtained for M-estimators.

We have performed another simulation study to investigate the effect of the critical value that defines an outlier. Let the observation samples be contaminated by an outlier whose magnitude changes between 3σ and 6σ and let the critical value be $k\sigma$. For this test, the constant *k* was increased step by step from 2.50 to 3.00 and different samples were simulated. The performance of different methods is shown in Table 12, where the most successful results are shown in bold. They are 2.90σ for the LMS method, 2.80σ for the repeated median method and 2.65σ for Huber's method. There is no single optimal critical value for all the methods.

We have also considered the masking effect. Let the simple regression again be given as detailed in Table 1. Consider two bad observations located close to each other, such as $(\bar{y}_1,\bar{y}_2)$, $(\bar{y}_2,\bar{y}_3)$ and $(\bar{y}_9,\bar{y}_{10})$, separately. The partial redundancy numbers of these two bad observations are smaller than those of the others. The magnitude of one of them is greater than that of the other. The MSRs of all the robust methods for non-random outliers are shown in Table 13. They are seen to have decreased considerably compared with the MSRs given in Tables 2 and 3 for random and non-random outliers, respectively. However, the MSRs of the LMS are greater than before. This is also valid for the multiple regression.

In the next part of the study, we simulated the swamping effect on the simple regression given in Table 1. Thus, we designed that $(\bar{y}_1,\bar{y}_3)$, $(\bar{y}_2,\bar{y}_4)$, $(\bar{y}_7,\bar{y}_9)$ and $(\bar{y}_8,\bar{y}_{10})$ are bad observations for separate cases in simple regression. The MSRs of all the robust methods for non-random outliers are shown in Table 14. They have decreased considerably compared with the values given in Table 3. However, the MSRs of the LMS are larger than those of the other methods. These results are also valid for multiple regression models.

While investigating the effect of critical outliers, we considered the cases where $(\bar{y}_1,\bar{y}_2)$, $(\bar{y}_9,\bar{y}_{10})$, $(\bar{y}_1,\bar{y}_2,\bar{y}_3)$, $(\bar{y}_8,\bar{y}_9,\bar{y}_{10})$, $(\bar{y}_1,\bar{y}_2,\bar{y}_3,\bar{y}_4)$ or $(\bar{y}_7,\bar{y}_8,\bar{y}_9,\bar{y}_{10})$ are the sets of bad observations for the simple regression previously given in Table 1. The MSRs of all the robust methods for non-random outliers are shown in Table 15. The values decreased dramatically when compared to those given in Table 3. Furthermore, Huber's method was more successful than LMS in the case of random outliers. Note that the results for random outliers are not given due to lack of space. LMS was more successful than the other methods in the case of non-random critical outliers. It was also effective for multiple regression.

We have also considered the effect of concentrated outliers. These lie far away from the bulk of the data and act in a similar way to leverage points. However, they lie close to each other (Fig.3). The concentrated outliers were simulated for both simple and multiple regressions. The MSRs of all the robust methods for non-random outliers are shown in Table 16. They behaved in the same way as in the study of leverage points.

Table 12 MSRs of the robust methods for different values of *k*

Method	*MSR* (%)										
	k=3.00	2.95	2.90	2.85	2.80	2.75	2.70	2.65	2.60	2.55	2.50
Least median of squares	73.3	73.3	**73.4**	72.4	71.6	70.4	69.1	67.6	65.6	63.9	60.8
Huber's	81.3	82.3	82.8	83.8	84.1	84.4	84.3	**84.8**	83.8	83.7	83.4
Repeated median	78.6	79.1	79.7	79.8	**80.2**	79.8	79.8	79.1	78.8	77.0	76.0

Bold numbers represent the most successful results

Table 13 MSRs for non-random outliers in the case of masking effect

Bad observation	*MSR* (%)		
	Theil-Sen	LMS	Huber's
$(\bar{y}_1,\bar{y}_2)$	38	41	26
$(\bar{y}_2,\bar{y}_3)$	39	42	28
$(\bar{y}_9,\bar{y}_{10})$	37	41	26

Table 14 MSRs for non-random outliers in the case of swamping effect

Bad observation	*MSR* (%)		
	Theil-Sen	LMS	Huber's
$(\bar{y}_1,\bar{y}_3)$	29	31	17
$(\bar{y}_2,\bar{y}_4)$	31	32	18
$(\bar{y}_7,\bar{y}_9)$	30	32	18
$(\bar{y}_8,\bar{y}_{10})$	29	31	17

Table 15 MSRs for non-random in the case of critical outliers

Bad observation	*MSR* (%)		
	Theil-Sen	LMS	Huber's
$(\bar{y}_1,\bar{y}_2)$	58	60	46
$(\bar{y}_9,\bar{y}_{10})$	57	61	45
$(\bar{y}_1,\bar{y}_2,\bar{y}_3)$	37	43	32
$(\bar{y}_8,\bar{y}_9,\bar{y}_{10})$	38	42	31
$(\bar{y}_1,\bar{y}_2,\bar{y}_3,\bar{y}_4)$	26	39	22
$(\bar{y}_7,\bar{y}_8,\bar{y}_9,\bar{y}_{10})$	25	39	24

Table 16 MSRs for non-random outliers in the case of concentrated outliers

Bad observation	*MSR* (%)		
	Theil-Sen	LMS	Huber's
$\bar{y}_{11}$	72	73	80
$(\bar{y}_{11},\bar{y}_{12})$	51	76	0
$(\bar{y}_{11},\bar{y}_{12},\bar{y}_{13})$	10	76	0
$(\bar{y}_{11},\bar{y}_{12},\bar{y}_{13},\bar{y}_{14})$	0	75	0

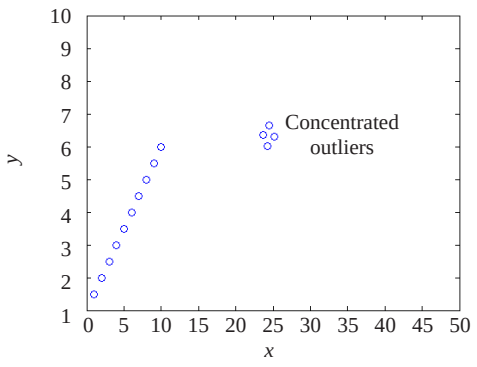

Fig.3 The concentrated outliers for simple regression

Gross outliers in the *y*-direction are errors with a very large magnitude with respect to the magnitude of other outliers, for example 1000σ or more. We considered linear regression models with gross errors at both ends of the interval for the observations, where the partial redundancies were smaller than for the other observations. If the sample had even a gross error in the *y*-direction, the results from LSE and M-estimator were disturbed. However, gross errors did not affect the LMS or LTS results. In other words, LMS and LTS identified gross errors successfully.

CONCLUSION

In this study, four robust methods with high breakdown points and Huber's method were compared using the MSR in different outlier scenarios in the *x*- and *y*-directions. A simulation compared their ability to detect and classify outliers in linear regression models. The results showed that the mean success rates decreased as the number of regressors in the regression model increased. Moreover, as the magnitude of outliers increased, the MSRs also increased. If the actual number of outliers increased, the MSRs decreased. The MSRs of random outliers were greater than those for non-random outliers. Huber's method

was more successful than the other methods when the data did not include any leverage points. Among the robust estimators with high breakdown points, the LMS and LTS methods were the most successful against the leverage points and gross errors in the *y*-direction. In addition, the LMS and LTS methods were the most successful when masking and/or swamping effects occurred in the data and when the outliers were concentrated. The median-based methods failed in some cases when the data did not include any outliers, because they have the tendency to classify non-outlying observations as outliers. The success rates of the methods also changed according to the actual definition of an outlier. However, the MSR may be interpreted as a local estimated value of the finite sample breakdown point that can be used to compare the reliability of the robust methods for outlier detection. We recommend using LMS and LTS as diagnostic tools to classify outliers, because they retain their robustness even for models that are heavily contaminated or that have a complicated structure of outliers.

Finally, in future work, the simulation study could be extended to cover the case of unknown variance for the robust methods used.

References

Barnett, V., Lewis, T., 1994. Outliers in Statistical Data (3rd Ed.). John Wiley and Sons, New York.

Chen, C., 2002. Robust Regression and Outlier Detection with the ROBUSTREG Procedure. SUGI Paper No.265-27. SAS Institute, Cary, NC.

Daniel, C., Wood, F.S., 1971. Fitting Equations to Data. Wiley, New York.

Davies, P.L., 1993. Aspects of robust linear regression. *Ann. Stat.*, **21**(4):1843-1899. [doi:10.1214/aos/1176349401]

Davies, P.L., Gather, U., 2005. Breakdown and groups with discussion and rejoinder. *Ann. Stat.*, **33**(3):977-1035. [doi:10.1214/009053604000001138]

Donoho, D.L., 1982. Breakdown Properties of Multivariate Location Estimators. PhD Qualifying Paper, Harvard University, Boston.

Donoho, D.L., Huber, P.J., 1983. The Notion of Breakdown Point. *In*: Bickel, P.J., Doksum, K., Hodges, J.L.J. (Eds.), A Festschrift for Erich L. Lehmann. Wadsworth, Belmont, p.157-184.

Gather, U., Hilker, T., 1997. A note on Tyler's modification of the MAD for the Stahel-Donoho estimator. *Ann. Stat.*, **25**(5):2024-2026. [doi:10.1214/aos/1069362384]

Hadi, A.S., Simonoff, J.S., 1993. Procedures for the identification of multiple outliers in linear models. *J. Am. Stat. Assoc*., **88**(424):1264-1272. [doi:10.2307/2291266]

Hampel, F.R., 1968. Contributions to the Theory of Robust Estimation. PhD Thesis, University of California, Berkeley.

Hampel, F.R., 1971. A general qualitative definition of robustness. *Ann. Math. Stat.*, **42**(6):1887-1896. [doi:10.1214/aoms/1177693054]

Hampel, F.R., 1975. Beyond location parameters: robust concepts and methods (with discussion). *Bull. Inst. Int. Stat.*, **46**:375-391.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.R., Shatel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Hekimoglu, S., 1997. Finite sample breakdown points of outlier detection procedures. *ASCE J. Surv. Eng.*, **123**(1):15-31. [doi:10.1061/(ASCE)0733-9453(1997)123:1(15)]

Hekimoglu, S., 2005. Do robust methods identify outliers more reliably than conventional test for outlier? *Zeitschrift für Vermessungwesen*, **3**:174-180.

Hekimoglu, S., Koch, K.R., 1999. How Can Reliability of the Robust Methods Be Measured? *In*: Altan, M.O., Gründig, L. (Eds.), Third Turkish-German Joint Geodetic Days, **1**:179-196.

Hekimoglu, S., Erenoglu, R.C., 2005. Estimation of Parameters for Linear Regression Using Median Estimator. Int. Conf. on Robust Statistics, University of Jyvaskyla, Finland, p.26.

Hekimoglu, S., Erenoglu, R.C., 2007. Effect of heteroscedasticity and heterogeneousness on outlier detection for geodetic networks. *J. Geod.*, **81**(2):137-148. [doi:10.1007/s00190-006-0095-z]

Huber, P.J., 1981. Robust Statistics. John Wiley and Sons, New York.

Kamgar-Parsi, B., Netanyahu, N.S., 1989. A nonparametric method for fitting a straight line to a noisy image. *IEEE Trans. Pattern Anal. Mach. Intell.*, **11**(9):998-1001. [doi:10.1109/34.35504]

Lopuhaa, H.P., Rousseeuw, P.J., 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Stat.*, **19**(1):229-248. [doi:10.1214/aos/1176347978]

Rousseeuw, P.J., 1984. Least median of squares regression. *J. Am. Stat. Assoc.*, **79**(388):871-880. [doi:10.2307/2288718]

Rousseeuw, P.J., 1985. Multivariate Estimation with High Breakdown Point. *In*: Grossman, W., Pflug, G., Vincze, I., Werz, W. (Eds.), Mathematical Statistics and Applications. Reidel, Dordrecht, p.283-297.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. John Wiley and Sons, New York.

Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc.*, **63**(324):1379-1389. [doi:10.2307/2285891]

Shevlyakov, G.L., Vilchevski, N.O., 2001. Robustness in Data Analysis: Criteria and Methods. VSP International Science Publishers, Utrecht.

Siegel, A.F., 1982. Robust regression using repeated medians. *Biometrika*, **69**(1):242-244. [doi:10.1093/biomet/69.1.

242]

Stahel, W.A., 1981. Breakdown of Covariance Estimators. Research Rep. 31, Fachgruppe für Statistik, ETH, Zurich.

Staudte, R.G., Sheather, S.J., 1990. Robust Estimation and Testing. Wiley, New York.

Stromberg, A.J., 1993. Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM J. Sci. Comput.*, **14**(6):1289-1299. [doi:10.1137/0914076]

Theil, H., 1950. A rank-invariant method of linear and polynomial regression analysis. *Nederlandse Akademie Wetenchappen Series A*, **53**:386-392.