A New Algorithm for Detecting Outliers in Linear Regression

Mehmet Hakan Satman¹

¹ Department of Econometrics, Istanbul University, Istanbul, Turkey

Correspondence: Mehmet Hakan Satman, Department of Econometrics, Istanbul University, Istanbul, Turkey. Tel: 90-212-440-0000 ext.11536. E-mail: mhsatman@istanbul.edu.tr

Received: June 12, 2013Accepted: July 5, 2013Online Published: July 15, 2013doi:10.5539/ijsp.v2n3p101URL: http://dx.doi.org/10.5539/ijsp.v2n3p101

Abstract

In this paper, we present a new algorithm for detecting multiple outliers in linear regression. The algorithm is based on a non-iterative robust covariance matrix and concentration steps used in LTS estimation. A robust covariance matrix is constructed to calculate Mahalanobis distances of independent variables which are then used as weights in weighted least squares estimation. A few concentration steps are then performed using the observations that have smallest residuals. We generate random data sets for $n = 10^3$, 10^4 , 10^5 and p = 5, 10 to show up the capabilities of the algorithm. In our Monte Carlo simulations, it is shown that our algorithm has very low masking and swamping ratios when the number of observations is up to 10^4 in the case of maximum contamination in X-Space. It is also shown that, the algorithm is successful in the case of Y-Space outliers when the contamination level, sample size and number of parameters are up to 30%, $n = 10^5$, and p = 10, respectively. Bias, variance and MSE statistics are calculated for different scenarios. The reported computation time of our implementation is quite short. It is concluded that the presented algorithm is suitable and applicable for detecting multiple outliers in regression analysis with its small masking and swamping ratios, accurate estimates of regression parameters except the intercept, and short computation time in large data sets and high level of contamination. A future work is required for reducing bias and variance of the intercept estimator in the model.

Keywords: outlier detection, linear regression, robust statistics

1. Introduction

Suppose the model is

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

where y is an n vector of dependent variable, X is an $n \times p$ matrix of independent variables, β is a p vector of unknown parameters, ϵ is an n vector of stochastic error term, p is the number of parameters, and n is the number of observations. Ordinary least squares (OLS) estimator $\hat{\beta}$ consistently estimates β with minimum variance among the other estimators when the assumptions are hold. Autocorrelated or heteroscedastic error term, including irrelevant variables in the regression equation decrease the efficiency of $\hat{\beta}$ whereas omitting a relevant variable and measurement errors in independent variables yield biased and inconsistent estimates. Since most of the statistical software packages include tests for classical assumptions of OLS, researchers have an ability of testing their hypotheses as fast as possible. However, the problem of outliers is generally considered as a symptom of heteroscedasticity in econometrics books or outlier detection routines are neglected in some software packages. In fact, outliers can be more dangerous than increasing the variability of conditional variances.

A single outlier can be detected by analysing OLS residuals or diagnostic measures derived from OLS estimates. Practitioners tend to analysis these statistics, however we don't know the number of outliers. When the data set contains more than one outlier, the OLS estimator $\hat{\beta}$ is usually affected and does not estimate the β correctly. As a result of this; residuals, fitted values and other properties of regression are also affected. Regression diagnostics should be calculated for subsets of observations rather than for each single observation. However, this operation is computationally inefficient.

As a result of technical difficulties, some authors developed more efficient algorithms for detecting regression outliers. Kianifard and Shallow (1989), Marasinghe (1985), Atkinson (1986), Hadi and Simonoff (1993), Peña and Yohai (1995), and Sebert et al. (1998) developed OLS based outlier detection algorithms which are not based

on calculating statistics for all subsets of potential outliers. The methods reported in Billor et al. (2000) and Billor et al. (2005) are modern revisions which can be categorized as robust methods. The success of these methods depends on the number of observations, the number of parameters, and the fraction and the direction of contamination (Wisnowski, 1999).

Robust regression is an other branch of outlier detection and has a vast of literature. Huber (1973) introduced the M-Estimator. Rousseeuw and Leroy (1987) investigated the properties of Least Median of Squares (LMS) and Least Trimmed Squares (LTS) estimators. LMS and LTS stay resistant when the number of outliers is up to 50% of data. However, robust regression estimators are generally based on optimizing non-smooth or discreate functions which consume too much computation time. There is an effort to speed up these methods. Salibian-Barrera and Yohai (2006) suggested a new algorithm for calculating S-Regression estimates. Rousseeuw and van Driessen (2006) suggested a new algorithm for calculating the LTS estimator for large data sets. Satman (2012) proposed a genetic algorithm (GA) based modification on the method given in Rousseeuw and van Driessen (2006) and showed that the GA based search obtains smaller objective values in reasonable CPU times. Torti et al. (2012) performed a simulation study to compare powers of fast robust regression estimators including forward seach (Atkinson, 2010). Shortly, as the level of technology increases, we can collect more data; as we collect more data, the need for the technology increases.

In this paper we devise a new algorithm for detecting regression outliers. In Section 2, we give a brief description of previous works and the problem of outlier detection. In Section 3, we present the devised algorithm. In Section 4, we perform a Monte Carlo simulation to unveil the success of our algorithm. In this simulations we show the MSE's (Mean Square Error) of our estimator as well as the masking and the swamping ratios. Finally, in Section 5, we conclude.

2. Preliminaries

In regression analysis, an observation is an outlier if the model does not fit this observation well. However, the model can not be estimated correctly by OLS when the data contains outliers. As a result of this, real outliers may be fitted well by the regression equation. This is same as Type I error, namely masking in outlier detection literature, rejecting outlyingness of observations when they are outliers in real (Lawrence, 1995). Similarly, clean observations may be misfitted by the regression equation, that is, we fail to reject outlyingness of observations when they are clean in real. This is the problem of swamping (Barnett & Lewis, 1978). Success of an outlier detection method is generally measured with its masking and swamping ratios.

The LMS estimator stays resistant when the level of contamination is up to 50% (Rousseeuw, 1984). The objective function of the LMS estimator is to minimize the median of squared residuals. Since, median is not a continuous function of squared residuals, gradient based optimization techniques are not applicable. Therefore, several algorithms were developed for the LMS in Winker et al. (2011), Nunkesser and Morell (2010) and Karr et al. (1995) among others.

Another robust regression estimator LTS has the same breakdown point property as the LMS, that is, it stays resistant when the data is contaminated up to 50%. The objective function of the LTS estimator can be written as

$$\min_{\hat{\beta}} \sum_{i=1}^{h} r_i^2 \tag{2}$$

where r_i^2 is the *i*th ordered squared residual and *h* is a custom integer which is approximately n/2.

Since the objective function given in (2) minimizes the sum of h smallest squared residuals, it can be re-written as

$$\min_{\hat{\beta}} \sum_{i=1}^{n} w_i e_i^2$$
subject to $\sum_{i=1}^{n} w_i = h$
(3)

where $w_i \in \{0, 1\}$ and e_i^2 is the *ith* squared residual. Note that the equation given in (3) is a constrained optimization problem with a non-linear objective function and binary variables. Several algorithms for optimizing the LTS objective function can be found in Atkinson and Cheng (1999), Agulló (2001), Giloni and Padberg (2002), Bai (2003), and Hofmann et al. (2010) among others.

Rousseeuw and Driessen (2006) developed the Fast LTS algorithm which is based on performing C-Steps (Concentration Steps) on randomly selected subsets. In this algorithm, the key point is to find out the right set of p observations and to perform C-Steps to enlarge the initial subset to h observations that minimizes the objective function. In our algorithm, we replace the random subset selection part of Fast LTS by a non-iterative procedure.

Outlier detection methods in multivariate data and linear regression are not independent subjects. In many algorithms, initially the X-Space is considered as multivariate data and an outlier detection procedure is performed. Then, regression outliers are detected using the observations that are labeled as clean in the first stage of algorithms. Hadi (1992) and Hadi (1994) developed and modified a method for detecting outliers in multivariate data by estimating the robust covariance matrix. Rousseeuw and Van Zomeren (1990) introduced the robust covariance estimator MVE (Minimum Volume Ellipsoid) and showed that plotting MVE based Mahalanobis distances versus LMS residuals gives a snapshot of outliers and their directions. MCD (Minimum Covariance Determinant) is an other robust covariance estimator which has a similar objective function with the LTS. In MCD, the h observations which have the minimum determinant of covariance matrix is searched (Rousseeuw & Van Driessen, 1999). Note that, all of the covariance estimators mentioned here requires many iterations and there is an effort to speed up these algorithms.

OLS estimators are directly related to the covariance structure of the data. Suppose the model is a special form of (1) with a single independent variable. Then the slope estimator is

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} \tag{4}$$

and it is very sensitive to outliers because both of the operators given in (4) are functions of sample sums. Huo et al. (2012) suggested a new measure of covariance, namely *comediance*. Comediance of two variates x and y is defined as

$$\hat{C}_{x,y} = med([x_i - med(x)] \times [y_i - med(y)])$$
(5)

where med(x) and med(y) are sample medians and i = 1, 2, ..., n. In their Monte Carlo simulations Huo et al. (2012) showed that the performance of the comediance is comparable with other robust covariance estimators including Campbell, Sign, and Rank. Note that this method requires computing three medians and does not include any iterative procedure (Note 1).

3. Proposed Method

In our algorithm, we combine the comediance measure introduced in Huo et al. (2012) and C-Steps suggested in Rousseeuw and Driessen (2006). Briefly, the algorithm constructs a robust covariance matrix which does not require too much computation time. Mahalanobis distances of independent variables are then calculated using this robust covariance matrix for dispersion and sample medians of variables for location. Then, h observations with minimum Mahalanobis distances are used to calculate OLS estimates. After all, C-Steps are performed using pobservations with smallest absolute residuals to enlarge the basic subset to h observations. The whole algorithm is given below.

3.1 Main Algorithm

Step 0. Let p is the number of regression parameters, p - 1 is the number of independent variables, n is the number of observations, w is an n vector of zeros, $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$, and $\lfloor k \rfloor$ is the integer part of k

Step 1. Construct the covariance matrix $\hat{\Sigma}$ of independent variables where (Note 2)

$$\hat{\sigma}_{ii} = med(|x_{ik} - med(x_i)|)$$

for i = 1, 2, ..., p - 1, k = 1, 2, ..., n, and

$$\hat{\sigma}_{ij} = med([x_{ik} - med(x_i)] \times [x_{jk} - med(x_j)])$$

 $j = 1, 2, ..., p - 1, i \neq j$, and $\hat{\sigma}_{ij}$ is the element of matrix $\hat{\Sigma}$ at row *i* and column *j*.

Step 2. Calculate the mahalanobis distances D where

$$\mathbf{D} = \sqrt{(x-\mu)'\hat{\boldsymbol{\Sigma}}^{-1}(x-\mu)}$$

and μ is the vector of medians of independent variables. Sort the values of **D** in ascending order. Let *k* is a vector of indices of first *h* smallest **D**_{*i*}'s. Set $w[k_1] = 1$, $w[k_2] = 1$, ..., and $w[k_h] = 1$.

Step 3. Perform a weighted least squares estimation for the model using the weights w and calculate absolute residuals.

Step 4. Perform C-Steps using p observations with first p ordered absolute residuals to enlarge the basic subset to h observations. Perform many C-Steps using the enlarged basic subset.

Step 5. Standardize the residuals obtained from the final C-Step using the formula $r_i = \frac{e_i - med(e)}{med(|e_i - med(e)|)}$. Report the observation *i* as an outlier if $|r_i| > 2.5$.

First iteration of C-Steps takes p observations and returns h observations with smallest absolute residuals. In remaining iterations, C-Steps are performed using h observations many times and the final h observations are returned. In those steps, it is expected that some outliers exit the basic subset and clean observations enter. Number of C-Steps is set to 10 by virtue of the graph shown in Figure 1 in our simulations.

4. Monte Carlo Simulations

We perform a simulation study to unveil the success of our algorithm. We report the MSE's of estimated regression parameters as well as the masking and the swamping ratios. Data are generated for $n = 10^3$, 10^4 and 10^5 . The number of parameters are set to p = 5 and p = 10. Any data set that generated with these *n*, *p* combinations can be considered as large (Note 3).

Data are generated using the linear model (1) where $\epsilon \sim \mathcal{N}(0, 1)$, $X_i \sim \mathcal{N}(0, 100)$, $\beta = [5, 5, \dots, 5]'$, $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2 , and $i = 1, 2, \dots, p$. The level of contamination is variable and set to $c_m = 20\%$, 30%, 40% and n - h for Y-Space outliers and $c_m = n - h$ for X-Space outliers, where $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ and $\lfloor k \rfloor$ is the integer part of k. Note that n - h is the maximum level of contamination, that is, a contamination level that is higher from n - h means that outliers turn into clean observations and via versa. Independent variables are contaminated by adding random variates that follow a $\mathcal{N}(100, 100)$. The dependent variable is contaminated using the formula $y_i := max(y_{1:h}) + C_R$ where $max(y_{1:h})$ is the *n*th order statistics of first h values of variable y and C_R is a random variable that follows a $\mathcal{N}(10, 100)$.

Tables 1-5 summarize the results. In Table 1, it is shown that the MSE's of estimators are directly related to the level of contamination when the data is contaminated only in Y-Space. Small bias and variance statistics indicate that the estimators are accurate when the contamination level is 20%. However, the accuracy of estimators tend to reduce as the level of contamination increases. Finally, in the case of maximum contamination, method loses its robustness and yields the highest MSE values.

| | cont | | %20 | | | %30 | | | 40% | | | n - h | |
|-----|-----------------|--------|---------|---------|--------|-----------|-----------|---------|-----------|-----------|---------|------------|------------|
| n | | bias | var | mse | bias | var | mse | bias | var | mse | bias | var | mse |
| | $\hat{\beta_0}$ | 0,675 | 119,016 | 119,471 | 19,197 | 5562,776 | 5931,294 | 79,894 | 25190,599 | 31573,614 | 202,975 | 49920,573 | 91119,425 |
| | $\hat{\beta_1}$ | -0,020 | 0,118 | 0,119 | -0,344 | 3,430 | 3,548 | -1,308 | 7,893 | 9,603 | -2,327 | 8,464 | 13,878 |
| | $\hat{\beta}_2$ | -0,010 | 0,062 | 0,063 | -0,392 | 2,791 | 2,944 | -1,136 | 6,675 | 7,964 | -2,129 | 9,791 | 14,322 |
| | β_3 | -0,003 | 0,016 | 0,016 | -0,427 | 3,618 | 3,801 | -1,199 | 7,717 | 9,153 | -2,349 | 10,008 | 15,524 |
| 103 | $\hat{\beta}_4$ | -0,014 | 0,127 | 0,127 | -0,375 | 2,710 | 2,851 | -1,139 | 7,480 | 8,776 | -2,325 | 11,380 | 16,784 |
| | βĵ | -0,018 | 0,112 | 0,112 | -0,407 | 3,889 | 4,054 | -1,331 | 8,039 | 9,810 | -2,264 | 10,648 | 15,772 |
| | $\hat{\beta_6}$ | -0,025 | 0,176 | 0,177 | -0,368 | 2,772 | 2,907 | -1,244 | 7,942 | 9,491 | -2,216 | 9,057 | 13,967 |
| | $\hat{\beta_7}$ | -0,016 | 0,096 | 0,096 | -0,378 | 3,110 | 3,253 | -1,261 | 8,398 | 9,988 | -2,247 | 11,818 | 16,868 |
| | $\hat{\beta_8}$ | -0,029 | 0,231 | 0,232 | -0,381 | 2,739 | 2,884 | -1,184 | 7,266 | 8,668 | -2,196 | 10,482 | 15,303 |
| | βĝ | -0,011 | 0,086 | 0,086 | -0,433 | 3,254 | 3,441 | -1,174 | 7,211 | 8,590 | -2,257 | 10,844 | 15,940 |
| | $\hat{\beta_0}$ | 0,544 | 66,519 | 66,815 | 45,825 | 16288,812 | 18388,700 | 124,577 | 46858,216 | 62377,524 | 274,151 | 77748,635 | 152907,395 |
| | β_1 | -0,015 | 0,060 | 0,060 | -0,957 | 7,249 | 8,164 | -1,730 | 9,984 | 12,975 | -2,357 | 8,116 | 13,672 |
| | β_2 | -0,014 | 0,110 | 0,110 | -0,882 | 6,604 | 7,382 | -1,703 | 9,832 | 12,732 | -2,390 | 8,873 | 14,586 |
| | β ₃ | -0,013 | 0,042 | 0,043 | -0,912 | 7,215 | 8,046 | -1,771 | 9,989 | 13,124 | -2,466 | 8,142 | 14,221 |
| 104 | $\hat{\beta}_4$ | -0,011 | 0,035 | 0,036 | -0,942 | 7,468 | 8,354 | -1,736 | 9,957 | 12,970 | -2,390 | 8,736 | 14,449 |
| | β̂5 | -0,016 | 0,061 | 0,061 | -0,896 | 6,770 | 7,572 | -1,785 | 10,017 | 13,203 | -2,487 | 7,448 | 13,632 |
| | $\hat{\beta_6}$ | -0,036 | 0,306 | 0,308 | -0,845 | 6,926 | 7,639 | -1,732 | 10,109 | 13,110 | -2,498 | 7,748 | 13,988 |
| | βĵ | -0,026 | 0,151 | 0,152 | -0,937 | 7,857 | 8,735 | -1,669 | 10,040 | 12,826 | -2,489 | 8,368 | 14,565 |
| | β_8 | -0,021 | 0,125 | 0,125 | -0,826 | 6,465 | 7,148 | -1,665 | 9,741 | 12,512 | -2,437 | 7,878 | 13,816 |
| | βĝ | -0,019 | 0,142 | 0,143 | -0,964 | 8,043 | 8,973 | -1,758 | 10,428 | 13,517 | -2,433 | 8,245 | 14,163 |
| | $\hat{\beta_0}$ | 0,093 | 4,175 | 4,184 | 40,846 | 16314,131 | 17982,498 | 119,308 | 53315,920 | 67550,397 | 342,996 | 101362,195 | 219008,474 |
| | β_1 | -0,005 | 0,015 | 0,015 | -0,832 | 7,499 | 8,191 | -1,564 | 10,164 | 12,609 | -2,790 | 8,485 | 16,270 |
| | β_2 | -0,006 | 0,020 | 0,020 | -0,814 | 6,946 | 7,608 | -1,563 | 9,919 | 12,362 | -2,575 | 9,562 | 16,192 |
| - | β_3 | -0,005 | 0,012 | 0,012 | -0,873 | 8,003 | 8,764 | -1,615 | 10,069 | 12,676 | -2,706 | 7,646 | 14,969 |
| 105 | $\hat{\beta}_4$ | -0,003 | 0,006 | 0,006 | -0,863 | 7,542 | 8,286 | -1,590 | 10,153 | 12,680 | -2,689 | 8,944 | 16,173 |
| | βĵ | -0,006 | 0,020 | 0,020 | -0,868 | 7,738 | 8,491 | -1,509 | 10,655 | 12,933 | -2,791 | 8,481 | 16,268 |
| | β_6 | -0,003 | 0,005 | 0,005 | -0,868 | 7,870 | 8,623 | -1,597 | 10,116 | 12,667 | -2,589 | 8,775 | 15,479 |
| | $\hat{\beta_7}$ | -0,002 | 0,004 | 0,004 | -0,813 | 7,449 | 8,110 | -1,521 | 9,842 | 12,156 | -2,678 | 7,353 | 14,526 |
| | $\hat{\beta_8}$ | -0,005 | 0,013 | 0,013 | -0,828 | 7,271 | 7,956 | -1,592 | 9,945 | 12,481 | -2,688 | 8,802 | 16,025 |
| | βĝ | -0,005 | 0,013 | 0,013 | -0,869 | 7,854 | 8,609 | -1,573 | 9,607 | 12,083 | -2,720 | 8,996 | 16,396 |

Table 1. Y-Outliers for p = 10

Table 2 shares a similar status including higher bias, variance and MSE statistics of the intercept estimator. Best results are obtained when $n = 10^5$ and $c_m = 20\%$.

| Ta | ble | 2. | Y- | Out | liers | for | р | = | 5 |
|----|-----|----|----|-----|-------|-----|---|---|---|
|----|-----|----|----|-----|-------|-----|---|---|---|

| | cont | | %20 | | | %30 | | | 40% | | | n - h | |
|-----------------|-----------------|--------|---------|---------|--------|-----------|-----------|---------|-----------|-----------|---------|-----------|-----------|
| n | | bias | var | mse | bias | var | mse | bias | var | mse | bias | var | mse |
| - | β_0 | 0,850 | 88,255 | 88,977 | 31,213 | 6654,392 | 7628,664 | 66,049 | 14906,204 | 19268,699 | 165,406 | 25268,142 | 52627,158 |
| | $\hat{\beta_1}$ | -0,070 | 0,627 | 0,632 | -0,982 | 6,696 | 7,660 | -1,441 | 7,769 | 9,847 | -2,673 | 6,519 | 13,667 |
| 10 ³ | $\hat{\beta_2}$ | -0,002 | 0,005 | 0,005 | -0,876 | 5,984 | 6,752 | -1,478 | 7,925 | 10,109 | -2,681 | 6,536 | 13,723 |
| | $\hat{\beta_3}$ | -0,058 | 0,446 | 0,449 | -0,847 | 6,037 | 6,754 | -1,467 | 8,201 | 10,355 | -2,678 | 6,627 | 13,797 |
| | $\hat{\beta_4}$ | -0,028 | 0,158 | 0,158 | -0,909 | 6,044 | 6,870 | -1,482 | 8,031 | 10,227 | -2,673 | 6,595 | 13,739 |
| - | β_0 | 3,191 | 380,347 | 390,529 | 38,331 | 9227,526 | 10696,798 | 94,507 | 22877,730 | 31809,338 | 176,187 | 36223,090 | 67264,823 |
| | $\hat{\beta_1}$ | -0,184 | 1,311 | 1,345 | -1,192 | 9,568 | 10,988 | -1,967 | 9,997 | 13,865 | -2,262 | 6,924 | 12,039 |
| 104 | $\hat{\beta}_2$ | -0,182 | 1,289 | 1,322 | -1,131 | 8,693 | 9,972 | -1,984 | 9,987 | 13,922 | -2,339 | 7,044 | 12,516 |
| | $\hat{\beta_3}$ | -0,184 | 1,290 | 1,324 | -1,118 | 8,157 | 9,407 | -1,984 | 10,138 | 14,075 | -2,306 | 7,040 | 12,358 |
| | $\hat{\beta_4}$ | -0,161 | 1,108 | 1,134 | -1,098 | 7,870 | 9,077 | -1,970 | 10,130 | 14,012 | -2,325 | 7,514 | 12,920 |
| | $\hat{\beta_0}$ | 0,052 | 0,532 | 0,535 | 36,156 | 10032,413 | 11339,679 | 103,871 | 29216,361 | 40005,465 | 226,611 | 47187,082 | 98539,506 |
| | $\hat{\beta_1}$ | -0,004 | 0,005 | 0,005 | -1,140 | 10,099 | 11,399 | -2,056 | 12,315 | 16,542 | -2,557 | 6,850 | 13,389 |
| 105 | β_2 | -0,004 | 0,005 | 0,005 | -1,090 | 9,301 | 10,488 | -2,025 | 11,886 | 15,986 | -2,636 | 6,601 | 13,548 |
| | βs | -0,003 | 0,003 | 0,003 | -1,108 | 9,707 | 10,935 | -2,030 | 11,601 | 15,722 | -2,674 | 7,051 | 14,202 |
| | $\hat{\beta_4}$ | -0,003 | 0,002 | 0,002 | -1,151 | 10,307 | 11,631 | -2,089 | 12,009 | 16,372 | -2,614 | 6,560 | 13,393 |

In Table 3, it is shown that the proposed method yields parameter estimates with relatively small biases and variances when independent variables are contaminated at maximum level. Algorithm performs well except for $n = 10^5$. Torti et al. (2012) stated that the primary interest in fitting regression models in applied statistics is to use the fitted model rather than solely to detect outliers. However, investigating the performance of an outlier detection method by examining masking and swamping ratios is not trivial. In Table 4, it is shown that, our method has very low masking and swamping ratios when $n = 10^3$ and $n = 10^4$. The masking ratio is relatively high for $n = 10^5$ and p = 5 but it is reduced when p = 10.

Table 3. X-Outliers, p = 5 and p = 10, $c_m = n - h$

| | n | | 10^{3} | | | 10^{4} | | | 10^{5} | |
|----|-----------------|--------|----------|-------|--------|----------|-------|--------|----------|--------|
| р | | bias | var | mse | bias | var | mse | bias | var | mse |
| | $\hat{eta_0}$ | 0,001 | 0,002 | 0,002 | -0,015 | 0,119 | 0,119 | -3,464 | 5,101 | 17,103 |
| | $\hat{eta_1}$ | -0,000 | 0,000 | 0,000 | -0,010 | 0,047 | 0,047 | -3,610 | 4,593 | 17,626 |
| 5 | $\hat{eta_2}$ | 0,000 | 0,000 | 0,000 | -0,010 | 0,056 | 0,056 | -3,609 | 4,590 | 17,616 |
| | $\hat{eta_3}$ | -0,000 | 0,000 | 0,000 | -0,010 | 0,046 | 0,046 | -3,611 | 4,595 | 17,634 |
| | $\hat{eta_4}$ | 0,000 | 0,000 | 0,000 | -0,009 | 0,042 | 0,042 | -3,609 | 4,589 | 17,615 |
| | $\hat{eta_0}$ | -0,123 | 2,072 | 2,087 | -0,023 | 0,275 | 0,275 | -0,859 | 3,809 | 4,546 |
| | $\hat{\beta_1}$ | -0,042 | 0,189 | 0,191 | -0,008 | 0,032 | 0,032 | -0,868 | 3,537 | 4,290 |
| | $\hat{eta_2}$ | -0,048 | 0,243 | 0,246 | -0,009 | 0,045 | 0,045 | -0,869 | 3,546 | 4,301 |
| | $\hat{eta_3}$ | -0,043 | 0,196 | 0,198 | -0,010 | 0,053 | 0,053 | -0,868 | 3,533 | 4,286 |
| 10 | $\hat{eta_4}$ | -0,051 | 0,267 | 0,269 | -0,010 | 0,053 | 0,053 | -0,866 | 3,522 | 4,273 |
| | $\hat{eta_5}$ | -0,047 | 0,227 | 0,229 | -0,009 | 0,042 | 0,042 | -0,869 | 3,543 | 4,297 |
| | $\hat{eta_6}$ | -0,047 | 0,228 | 0,230 | -0,010 | 0,055 | 0,055 | -0,869 | 3,541 | 4,295 |
| | $\hat{\beta_7}$ | -0,062 | 0,393 | 0,397 | -0,011 | 0,056 | 0,056 | -0,866 | 3,523 | 4,273 |
| | $\hat{eta_8}$ | -0,057 | 0,323 | 0,326 | -0,010 | 0,050 | 0,051 | -0,863 | 3,499 | 4,244 |
| | $\hat{eta_9}$ | -0,051 | 0,262 | 0,264 | -0,010 | 0,053 | 0,053 | -0,863 | 3,493 | 4,237 |

Table 4. Masking and swamping ratios for X-Space outliers under maximum level of contamination

| | р | =5 | p=10 | | | |
|----------|---------|----------|---------|----------|--|--|
| п | Masking | Swamping | Masking | Swamping | | |
| 10^{3} | 0,000 | 0,000 | 0,004 | 0,001 | | |
| 10^{4} | 0,001 | 0,000 | 0,001 | 0,000 | | |
| 10^{5} | 0,327 | 0,058 | 0,078 | 0,014 | | |

In Table 5, it is shown that the performance of our algorithm is convincing when the level of contamination is 20% and 30% in the case of Y-Space outliers. Masking and swamping ratios increase as the level of contamination increases, finally masking and swamping ratios reach their maximum at the maximum contamination level.

Choosing the right number of C-Steps is important. Generally, iterating more C-Steps yields more accurate estimates. However, C-Steps consumes time when the sample size is large. We set the number of C-Steps to 10 in our simulations. In Figure 1, LTS criterion versus number of choosen C-Steps is plotted. It is clear that, performing more than 10 C-Steps does not gain too much. When the computation time is not critical, more C-Steps can be performed.

In Figure 2, consumed CPU times of our implementation (Note 4) are presented. It is shown that, CPU times increase as *n* and *p* increase. However, it takes under a second when the number of observations is 3×10^4 . This property of our implementation points out that our method can also be used in online detection of outliers of real time data (Note 5).

Table 5. Masking and swamping ratios for Y-Space Outliers

| | с | 20% | | 30% | | 40 |)% | n-h | | |
|----|----------|-------|-------|-------|-------|-------|-------|-------|-------|--|
| р | п | М | S | М | S | М | S | М | S | |
| | 10^{3} | 0,002 | 0,023 | 0,038 | 0,018 | 0,110 | 0,039 | 0,245 | 0,407 | |
| 5 | 10^{4} | 0,004 | 0,022 | 0,042 | 0,020 | 0,113 | 0,033 | 0,235 | 0,426 | |
| | 10^{5} | 0,000 | 0,020 | 0,032 | 0,016 | 0,114 | 0,028 | 0,262 | 0,508 | |
| | 10^{3} | 0,001 | 0,024 | 0,019 | 0,011 | 0,085 | 0,032 | 0,235 | 0,321 | |
| 10 | 10^{4} | 0,000 | 0,020 | 0,035 | 0,016 | 0,102 | 0,028 | 0,251 | 0,442 | |
| | 10^{5} | 0,000 | 0,020 | 0,027 | 0,013 | 0,085 | 0,021 | 0,273 | 0,503 | |

M for masking, S for swamping.



Figure 1. Number of C-Steps and LTS criterion



Figure 2. CPU times consumed by our R implementation

5. Results and Discussion

It is important to be aware of outliers in regression analysis. When the data is contaminated, the parameter estimates, calculated residuals and fitted values are affected. Because of that, regression diagnostics are not useful and should be calculated for subsets of observations rather than each single observation. However, this operation consumes too much computation time. Despite the robust procedures are successful in detecting outliers even though the contamination level is up to 50%, they consume too much time and there is also an effort to speed up these procedures in the outlier detection literature.

In this paper, we proposed a new algorithm based on a non-iterative covariance matrix and C-Steps used in LTS estimation. Since this covariance matrix has not all desired statistical properties, it is useful at finding a clean basic subset which is then used in a robust fit. Standardized absolute residuals which are bigger than a predefined criterion can be labelled as outliers.

The proposed algorithm has low masking and swamping ratios in the case of X-Space outliers when the level of contamination, sample size and number of parameters are up to 50%, $n = 10^4$, and p = 10, respectively.

In the case of Y-Space outliers, performance of the algorithm reduces but it still stays resistant when the contamination level, sample size and number of parameters are up to 30%, $n = 10^5$, and p = 10, respectively.

Our algorithm is suitable and applicable for detecting multiple outliers in regression analysis when the data sets are large and the contamination level is high. Computational cost is low and it is applicable even in interpretable statistical software packages. Regression estimators have small biases, variances and MSE's except for the intercept parameter.

An effort for reducing the bias and variance of intercept estimator would not be trivial and it can be examined in future works.

References

- Agulló, J. (2001). New algorithms for computing the least trimmed squares regression estimator. *Computational Statistics & Data Analysis, 36*, 425-439. http://dx.doi.org/10.1016/S0167-9473(00)00056-6
- Atkinson, A. C. (1986). Masking Unmasked. Biometrika, 74-3, 533-541. http://dx.doi.org/10.1093/biomet/73.3.533
- Atkinson, A. C., & Cheng, T. C. (1999). Computing least trimmed squares regression with the forward search. *Statistics and Computing*, *9*, 251-263.
- Atkinson, A. C., Riani, M., & Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the Korean Statistical Society*, 39(2), 117-134. http://dx.doi.org/10.1016/j.jkss.2010.02.007
- Bai, E. W. (2003). A random least-trimmed-squares identification algorithm. *Automatica*, 39(9), 1651-1659. http://dx.doi.org/10.1016/S0005-1098(03)00193-6
- Barnett, V., & Lewis, T. (1978). Outliers in statistical data (286, 293). John Wiley & Sons.
- Billor, N., Chatterjee, S., & Hadi, A. S. (2005). A Re-weighted Least Squares Method for Robust Regression Estimation. *American Journal of Mathematical and Management Sciences*, 26-3/4, 229-252.
- Billor, N., Hadi, A. S., & Velleman, P. V. (2000). BACON: Blocked Adaptive Computationally Efficent Outlier Nominators. Computational Statistics & Data Analysis, 34, 279-298. http://dx.doi.org/10.1016/S0167-9473(99)00101-2
- Giloni, A., & Padberg, M. (2002). Least Trimmed Squares Regression, Least Median Squares Regression, and Mathematical Programming. *Mathematical and Computer Modelling*, 35, 1043-1060. http://dx.doi.org/10.1016/S0895-7177(02)00069-9
- Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B (Methodological)*, 761-771.
- Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, Series B (Methodological)*, 393-396.
- Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the Identification of Multiple Outliers in Linear Models. *Journal of the American Statistical Association*, 88-424, 1264-1272. http://dx.doi.org/10.1080/01621459.1993.10476407

- Hofmann, M., Gatu, C., & Kontoghiorghes, E. J. (2010). An exact least trimmed squares algorithm for a range of coverage values. *Journal of Computational and Graphical Statistics*, 19(1), 191-204.
- Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5), 799-821. http://dx.doi.org/10.1214/aos/1176342503
- Huo, L., Kim, T. H., & Kim, Y. (2012). Robust estimation of covariance and its application to portfolio optimization. *Finance Research Letters*, 9, 121-134. http://dx.doi.org/10.1016/j.frl.2012.06.001
- Karr, C. L., Weck, B., Massart, D. L., & Vankeerberghen, P. (1995). Least median squares curve fitting using a genetic algorithm. *Engineering Applications of Artificial Intelligence*, 8(2), 177-189. http://dx.doi.org/10.1016/0952-1976(94)00064-T
- Kianifard, F., & Shallow, W. H. (1989). Using Recursive Residuals, Calculated on Adaptively-Ordered Observations, to Identify Outliers in Linear Regression. *Biometrics*, 45-2, 571-585. http://dx.doi.org/10.2307/2531498
- Lawrence, A. J. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society, Series B (Methodological)*, 181-189.
- Marasinghe, M. G. (1985). A Multivariate Procedure for Detecting Several Outliers in Linear Regression. *Technometrics*, 27-4, 395-399.
- Nunkesser, R., & Morell, O. (2010). An evolutionary algorithm for robust regression. *Computational Statistics & Data Analysis*, 54(12), 3242-3248. http://dx.doi.org/10.1016/j.csda.2010.04.017
- Peña, D., & Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society, Series B (Methodological)*, 145-156.
- R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388), 871-880. http://dx.doi.org/10.1080/01621459.1984.10477105
- Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424), 1273-1283. http://dx.doi.org/10.1080/01621459.1993.10476408
- Rousseeuw, P. J., &Leroy A. M. (1987). *Robust Regression And Outlier Detection*. New York: John Wiley and Sons.
- Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212-223. http://dx.doi.org/10.1080/00401706.1999.10485670
- Rousseeuw, P. J., & van Driessen, K. (2006). Computing LTS Regression for Large Data Sets. *Data Mining and Knowledge Discovery*, *12*, 29-45. http://dx.doi.org/10.1007/s10618-005-0024-4
- Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633-639. http://dx.doi.org/10.1080/01621459.1990.10474920
- Salibian-Barrera, M., & Yohai, V. J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, *15*(2), 414-427. http://dx.doi.org/10.1198/106186006X113629
- Satman, M. H. (2012). A Genetic Algorithm Based Modification on the LTS Algorithm for Large Data Sets. *Communications in Statistics-Simulation and Computation, 41*(5), 644-652. http://dx.doi.org/10.1080/03610918.2011.598989
- Satman, M. H. (2013). galts: Genetic algorithms and C-steps based LTS (Least Trimmed Squares) estimation. *R* package version 1.3. Retrieved from http://cran.r-project.org/package=galts
- Sebert, D. M., Montgomery, D. C., & Rollier, D. A. (1998). A clustering algorithm for identifying multiple outliers in linear regression. *Computational statistics & data analysis*, 27(4), 461-484. http://dx.doi.org/10.1016/S0167-9473(98)00021-8
- Torti, F., Perrotta, D., Atkinson, A. C., & Riani, M. (2012). Benchmark testing of algorithms for very robust regression: FS, LMS and LTS. *Computational statistics & data analysis*, 56, 2501-2512. http://dx.doi.org/10.1016/j.csda.2012.02.003

- Winker, P., Lyra, M., & Sharpe, C. (2011). Least median of squares estimation by optimization heuristics with an application to the CAPM and a multi-factor model. *Computational Management Science*, 8(1), 103-123. http://dx.doi.org/10.1007/s10287-009-0103-x
- Wisnowski, J. W. (1999). Multiple Outliers in Linear Regression: Advances in Detection Methods, Robust Estimation and Variable Selection. (Phd Dissertation, Arizona State University).

Notes

Note 1. However, comediance is not affine equivariant for linear transformations of the X-space.

Note 2. Note that the $\hat{\Sigma}_{i,i}$ given in *Step 1* is the Median Absolute Deviation (MAD) statistic without the correction factor of 1.4826 (Rousseeuw & Croux, 1993). The MAD statistic has a lower efficiency without the correction factor when the data is normal. However, this version of MAD narrows the space spanned by the ellipsoid which is most likely free of outliers. In our simulation study, better results obtained using this definition of the MAD. Since MAD is a robust measure of scale, it is generally used instead of usual standard deviation. In our method, MAD is placed in the diagonal elements of $\hat{\Sigma}$ which stands for variances instead of the standard deviations.

Note 3. In statistics, the terms *large sample* or *large data set* are not well-defined. In robust statistics, a data set is generally considered as *large* when calculation of a required set of combinations is intractable or infeasible. Rousseeuw and van Driessen (2006) uses the term of *small data set* when calculation of all *p*-subsets is possible. In their Monte Carlo simulations, they generate data for n = 100, 500, 1000, 10000, 50000 and p = 2, 3, 5, 10. In our simulations, the selected *n*, *p* combinations constitute a subset of previous work and we include the sample size of $n = 10^5$ to generate larger data sets.

Note 4. We provide a function *medmad* in the R package *galts* (Satman, 2013) which is freely available at site http://cran.r-project.org/web/packages/galts/.

Note 5. The function *medmad* is written in R which is an interpreter (R Core Team, 2012). A C/C++ or Fortran implementation should result in shorter times.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).