International Journal of Probability and Statistics 2016, 5(3): 65-72
DOI: 10.5923/j.ijps.20160503.01

Regression Estimation in the Presence of Outliers: A Comparative Study

Ahmed M. Gad[1,*], Maha E. Qura[2]

[1]Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt
[2]Department of Statistics, Mathematics & Insurance, Faculty of Commerce, Banha University, Egypt

Abstract In linear models, the ordinary least squares (OLS) estimators of parameters have always turned out to be the best linear unbiased estimators. However, if the data contain outliers, this may affect the least-squares estimates. So, an alternative approach; the so-called robust regression methods, is needed to obtain a better fit of the model or more precise estimates of parameters. In this article, various robust regression methods have been reviewed. The focus is on the presence of outliers in the *y*-direction (response direction). Comparison of the properties of these methods is done through a simulation study. The comparison's criteria were the efficiency and breakdown point. Also, the methods are applied to a real data set.

Keywords Linear regression, Outliers, High breakdown estimators, Efficiency estimators, Mean square errors

1. Introduction

Regression analysis is a statistical tool that is useful in all areas of sciences. The objective of linear regression analysis is to study how a dependent variable is linearly related to a set of explanatory variables. The linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i,$$
$$= \sum_{j=0}^{p} \beta_j x_{ij} + e_i, \quad i = 1,2,\ldots,n, \qquad (1)$$

where n is the number of observations, p is the number of explanatory variables ($n > p$), and $x_{i0} = 1$. The y_i denotes the i^{th} observed response, the x_{ij} represents the i^{th} observation of the explanatory variable x_j, β_j, $j = 0, 1, \ldots, p$, denote the regression coefficients, and e_i represents the random error (Birkes and Dodge, 1993). On the basis of the estimated parameters $\hat{\beta}_j$'s it is possible to fit the dependent variable as $\hat{y} = \sum_{j=0}^{p} \hat{\beta}_j x_{ij}$, and the estimates of the residuals $\hat{e}_i = y_i - \hat{y}_i$, for $1 \le i \le n$. In matrix notation the model can be written as $y = x\hat{\beta} + e_i$.

The objective of regression analysis is to find the estimates of the unknown parameters, $\hat{\beta}_j$'s. One method of obtaining the estimates of $\hat{\beta}_j$'s is the method of least squares, which minimizes the sum of squared distances of all the points from the actual observation to the regression surface (Fox, 1997). Therefore, the objective is to find those values of $\hat{\beta}$ that lead to the minimum value of $e'e = \sum_{i=1}^{n} e_i^2$, i.e.

$$\min_{\beta} \sum_{i=1}^{n} e_i^2 = \min_{\beta} \sum_{i=1}^{n} (y_i - x\hat{\beta})^2. \qquad (2)$$

The estimates $\hat{\beta}$ are unbiased if $E(\hat{\beta}) = \beta$. Among the class of linear unbiased estimators for β, the estimator $\hat{\beta}$ is the best in the sense that the variance of $\hat{\beta}$ is the minimum, $\mathrm{Cov}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$, where $\hat{\sigma}^2$ is the mean square error $(\mathrm{MSE}) = e'e/(n-p)$. For this reason the least squares estimates are the best linear unbiased estimators, referred to as BLUE.

In order to use regression correctly, the assumptions on which it is based need to be met. The assumptions of the OLS are that the errors are normally distributed, have equal variance at all levels of the independent variables (homoscedasticity), and are uncorrelated with both the independent variables and with each other. It is also assumed that the variables are measured without error. When the assumptions are met, model inferences such as confidence intervals and hypothesis testing are very powerful.

A common problem in regression analysis is the presence of outliers. As defined by Barnett and Lewis (1994), outliers are observations that appear inconsistent with the rest of data. The outliers may occur as a result of unusual but explainable events, such as faulty measurement, incorrect recording of data, failure of a measurement instrument, etc. Each different reason may require different treatment. Heavy-tailed distributions usually generate outliers, and these outliers may have a marked influence on parameter estimates (Chatterjee and Hadi, 1986). Rousseeuw and Leroy (1987) classified outliers as vertical outliers, good leverage points and bad leverage points. Vertical outliers are those observations have outlying values for the corresponding error term (on the *y*-direction) but are not outlying in the space of explanatory

* Corresponding author:
ahmed.gad@feps.edu.eg (Ahmed M. Gad)
Published online at http://journal.sapub.org/ijps

variables (the *x*-dimension). Good leverage points are those observations that are outlying in the space of explanatory variables but that are located close to the regression line. Bad leverage points are those observations that are both outlying in the space of explanatory variables and located far from the true regression line. There are outliers in both the directions (*x*, direction of the explanatory variables and *y*, direction of the response variable).

The aim of this paper is to compare the robust methods of estimation of regression parameters. This is specific to the outliers in *y*-direction. The comparison is conducted via simulation study and a real application. The paper is organized as follows. Section 2 is devoted to the evaluation and comparison criteria of regression estimators. In Section 3 we review the robust regression methods that achieve high breakdown point, high efficiency or both. Section 4 is devoted for the simulation study and Section 5 for the real application. In Section 6 we conclude the paper.

2. Evaluating Regression Estimators

The regression estimators can be evaluated and compared using different criteria. Although some criteria are more important than others, for a particular type of datasets, the ideal estimator would have positive characteristics of all criteria. These criteria include the following.

2.1. The Breakdown Point

The breakdown is the degree of robustness of an estimate in the presence of outliers (Hampel, 1974). The smallest possible breakdown point is *1/n* which tends to 0% when the sample size *n* becomes large. This is the case in the least squares estimators. If a robust estimator has a 50% breakdown point then 50% of the data could contain outliers and the coefficients would remain useable.

One aim of robust estimators is a high finite sample breakdown point ε_n^*. A formal finite sample definition of breakdown is given in Rousseeuw and Leroy (1987). Using a sample of *n* data points such that $Z = \{(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)\}$ and let T be a regression estimator. Applying T to the sample Z yields the regression estimator $T(Z) = \hat{\beta}$. Consider all possible corrupted samples Z' obtained by replacing the data points by the arbitrary values which allows for very bad outliers. The maximum bias that can be caused by this contamination is

$$\text{bias}(m; T, Z) = \sup_{Z'} \left\| T(Z') - T(Z) \right\|. \quad (3)$$

If the $\text{bias}(m; T, Z)$ is infinite, then the *m* outliers can have an arbitrarily large effect on T, which may be expressed by saving that estimator breaks down. Thus, the finite sample breakdown point ε^* of the estimator T_n at the sample of Z is defined as:

$$\varepsilon_n^*(T_n, Z) = \min\left\{\frac{m}{n}; \text{bias}(m; T, Z) \text{is infinite}\right\}. \quad (4)$$

2.2. The Mean Square Errors (MSE)

One of the performance criteria used to evaluate techniques (methods) of regression is the mean square errors which is given by

$$\text{MSE} = (\hat{\beta}_R - \beta)' (\hat{\beta}_R - \beta), \quad (5)$$

where $\hat{\beta}_R$ is a vector of robust parameter estimates and β is the vector of the true model coefficients.

2.3. The Bounded Influence

Bounded influence in the X-space is the estimator's resistance to being pulled towards the extreme observations in the X-space. Determining whether or not an estimator has a bounded influence is obtained by a study of the influence function. As defined by Seber (1977) influence function (IF) is a measure of the rate at which the estimator (*T*) responds to a small amount of contamination on *x*.

Hampel (1968, 1974) defined the influence function $IF_{T,F}(.)$ of the estimator *T*, at the underlying probability distribution, *F*, by

$$\text{IF}_{T,F}(x_0) = \lim_{\epsilon \to 0} \frac{T(\tilde{F}) - T(F)}{\epsilon} = \left[\frac{\partial T(\tilde{F})}{\partial \epsilon}\right]_{\epsilon=0}, \quad (6)$$

where $T(F)$ denotes the estimator of interest, expressed as a functional. The functional $T(\tilde{F})$ represent the estimator of interest under this altered c.d.f.. Thus, the influence function is actually a first derivative of an estimator, viewed as functional, and measures the influence of the point x_0 on the estimator T.

3. Robust Regression Methods

Many robust methods have been proposed in literature to achieve high breakdown point, high efficiency or both. Therefore, robust regression methods can be divided into three broad categories: high-breakdown point estimators, high efficiency estimators and multiple property estimators (related classes of efficient high-breakdown estimators). Each category contains a class of estimators derived under similar conditions and with comparable theoretical statistical properties.

3.1. High-Breakdown Point Estimators

High-breakdown point (HBP) regression estimators, also known as resistant estimators, can achieve up to a 50% breakdown point. They are useful for outliers detection and initial estimators. However, these estimators have low efficiency and unbounded influence. Due to this fact these estimators cannot be used as stand-alone estimators. The common estimators of this type are outlined in the following lines.

The repeated median (RM) estimator

Siegel and Benson (1982) propose the repeated median (RM) estimator with a 50% breakdown point. For any p observations $(x_{i1}, y_{i1}), \dots, (x_{ip}, y_{ip})$ the repeated median

estimator determines a unique parameter vector $\beta_j(i_1, \dots, i_p)$ which is defined as

$$\hat{\beta} = \text{med}_{i_1}(\dots(\text{med}_{i_p-1}(\text{med}_{i_p}\beta_j(i_1, \dots, i_p)))\dots). \quad (7)$$

The RM estimator is not affine equivariant. It was applied to a biological problem by Siegel and Benson (1982).

The least median squares (LMS) estimator

Rousseeuw (1984) propose the least median of squares (LMS) estimator. This approach can be thought of as being similar to the ordinary least squares, except that the median value is used instead of the mean value. The LMS estimator can be expressed as:

$$\min_\beta [med(e_i^2)], \quad (8)$$

where "*med*" denotes the median. The breakdown point of LMS is 50%. Therefore, it resists the effect of outliers even when they comprise nearly half of the data. The LMS lacks efficiency under normality although it is highly robust. The LMS can be robust with respect to outliers in both the *x* and the *y* directions but does not contain an influence function that is theoretically, bounded in the *x*-space. It satisfies all three equivariance properties; regression equivariance, scale equivariance, and affine equivariance (Rousseeuw and Leroy, 1987).

The least trimmed squares (LTS) estimator

Rousseeuw (1984) propose the least trimmed of squares (LTS) estimator as a high efficiency alternative to the LMS. The LTS is defined as

$$\min \sum_{i=1}^{h}(e^2)_{i:n}, \quad (9)$$

where $(e^2)_{1:n} \le (e^2)_{2:n} \le \dots \le (e^2)_{h:n} \le \dots \le (e^2)_{n:n}$ are the ordered squared residuals from the smallest to the largest. Rousseeuw and Leroy (1987) recommend using $h = n(1-\alpha)+1$, where α is the trimmed percentage.

Rousseeuw and van Driessen (2006) propose that the LMS estimator should be replaced by the least trimmed squares (LTS) estimator for large data sets. The best robustness properties are achieved when *h* is approximately *n/2*. In this case the breakdown point attains 50% (Rousseeuw and Leroy, 1987). The breakdown point of the LMS and the LTS are equal if $h = \frac{n+p+1}{2}$. The LTS has $n^{-1/2}$ convergence rate and it converges at a rate similar to the M-estimators (Rousseeuw, 1984). Zaman et al. (2001) mention that although the LTS method is good at finding out the outliers, it may sometimes eliminate too many observations and this may not give the true regression relation about the data. The LTS estimator is regression, scale, and affine equivariant (Rousseeuw and Leroy, 1987). The LTS-estimator statistical efficiency is better than the LMS, with higher asymptotic Gaussian efficiency of 7.1% (Rousseeuw and van Driessen, 2006).

The least Winsorized squares (LWS) estimator

Another alternative method to the LS estimation procedure is the Winsorized regression which is applied by altering the data values based upon the magnitude of the residuals (Yale and Forsythe, 1976). The aim of Winsorization is to diminish the effect of contamination on the estimators by reducing the effect of outliers in the sample. The estimator is given as:

$$\min \sum_{i=1}^{h}(e^2)_{i:n} + (n-h)(e^2)_{h:n}, \quad (10)$$

where h may depend on some fraction a. For simple linear regression, Winsorization is applied by ordering the values and modifying extreme *Y*-values, in an iterative fashion, by replacing the observed residual for an extreme *Y* -value with the next closest (and smaller) residual in the dataset, and, then, computing new *Y* -values using the formulation for an observed score as presented above. These new *Y*-values are used to compute new slope and intercept estimates for the regression line, and then a new set of residuals is obtained. The process of estimation, obtaining residuals, and data modification are continued for a specified number of iterations (Yale and Forsythe, 1976; Nevitt and Tam, 1998). The LWS is regression, scale, and affine equivariant similar to the LMS and the LTS (Rousseeuw and Leroy, 1987). Its breakdown is 50% when *h* is approximately *n/2* (Rousseeuw and Leroy, 1987).

The S-Estimator

Rousseeuw and Yohai (1984) develop a high-breakdown estimator that minimizes the dispersion of the residuals. The S –estimator $\hat{\beta}$ is defined by

$$\min s(e_1(\beta), \dots, e_n(\beta)), \quad (11)$$

with final scale estimate

$$\hat{\sigma} = s\big(e_1(\beta), \dots, e_n(\beta)\big), \quad (12)$$

where $e_i(\beta)$ is the i^{th} residual for candidate β. The dispersions $s(e_1(\beta), \dots, e_n(\beta))$ are defined as the solution of the equation:

$$\frac{1}{n-p}\sum_{i=1}^{n}\rho\left(\frac{y_i - X_i'\hat{\beta}}{s}\right) = K. \quad (13)$$

The constant K may be defined as $E_\Phi[\rho]$ or $\int \rho(x)d\Phi(x)$ to ensure that the S-estimator of the residual scale $\hat{\sigma}$ is consistent for σ_0 whenever it is assumed that the error distribution is normal with zero mean and σ_0^2 variance, where Φ is the standard normal distribution. The function ρ must satisfy the following conditions:

1. ρ is symmetric, continuously differentiable and $\rho(0) = 0$.
2. There exists the constant $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty]$.

The term S-estimator is used to describe the class of robust estimation because it is derived from a scale statistic in an implicit way. For ρ one often chooses the function

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & for \ \ |x| \le c; \\ \frac{c^2}{6} & for \ \ |x| > c, \end{cases} \quad (14)$$

where c is an appropriate tuning constant (Rousseeuw and Yohai, 1984). The derivative of this function is known as Tukey's biweight function:

$$\rho'(x) = \psi(x) = \begin{cases} x(1-(\frac{x}{c})^2)^2 & for\ |x| \le c; \\ 0 & for\ |x| > c. \end{cases} \quad (15)$$

In general, $\psi(x)$ will always be zero for $|x| > c$ because of condition 2; such $\psi(x)$ functions are usually called "re-descending" (Rousseeuw and Yohai, 1984). The breakdown point of the S-estimator can be 50%, assuming a condition is satisfied relating the constant K with the ρ function such that $\frac{K}{\rho(C)} = \frac{1}{2}$.

The S-estimator is regression, scale, and affine equivariant. It is also asymptotically normal with $n^{1/2}$ rate of convergence. Efficiency of the S-estimators can be increased at the expense of decreases in breakdown point (Rousseeuw and Leory, 1987). The S-estimator perform marginally better than the LMS and the LTS because the S-estimator can be used either as a high breakdown initial estimate with a high efficiency, or as a moderate breakdown (25%), moderate efficiency (75.9%) estimator (Rousseeuw and Leroy, 1987).

3.2. Efficiency Estimators

An efficient estimator provides parameter estimates close to those from an the OLS (the best linear unbiased estimator) which fits in an uncontaminated sample NID error terms. While the common efficient techniques are not high breakdown nor bounded-influence.

The L_1- norm or Least Absolute Deviations (LAD) Estimator

Boscovich introduced the method of least absolute deviations (LAD) in 1757, almost 50 years before the method of least squares was discovered by Legendre in France around 1805 (Birkes and Dodge, 1993). The least absolute deviations (LAD), also known as least absolute errors (LAE), the least absolute value (LAV) or the L_1 regression, is a mathematical optimization technique similar to the popular least squares technique in the attempts to find a function which closely approximates a set of data. (Cankaya, 2009).

In the LAD method, the coefficients are chosen so that the sum of the absolute deviations of the residuals is minimized as:

$$\min \sum_{i=1}^{n} |e_i|. \quad (16)$$

The LAD method is especially suitable when it is believed that the distribution of the errors has very heavy tails or is asymmetric or when the sample is very large. Because the LAD estimates have relatively low variance in the case of a heavy tailed error distribution, they have low bias in the case of an asymmetric error distribution and any reasonable estimate has low variance and control bias in the case of large samples (Birkes and Dodge, 1993). The LAD reduces the effects of outliers among the Y values but it does not protect against outliers among the X values (leverage points), which have a large effect on the model fit. Hence, it is a robust with respect to outliers in the y-direction and it is not robust with respect to the x-outliers (Wilcox, 2010). The breakdown point of the LAV is *1/n*. Thus the effect of one x-axis data outlier will change the regression line, causing the line to pass through the outlying data point.

The M-estimator (M)

Huber introduced in 1973 the idea of M-estimation as an alternative robust regression estimator to the least squares. The class of M-estimator models contains all models that are derived to be maximum likelihood models. The M-estimate is nearly as efficient as the OLS. This method is based on the idea of replacing the sum of squared residuals $\sum e^2$ used in the LS estimation by another function of residuals, $\sum \rho(e)$.

$$\min \sum_{i=1}^{n} \rho(e_i) = \min \sum_{i=1}^{n} \rho(y_i - x_i\hat{\beta}). \quad (17)$$

The function ρ gives the contribution of each residual to the objective function. A reasonable $\rho(.)$ should posses the following properties: $\rho(0) = 0$; $\rho(e_i) \ge 0$ (non-negativity); $\rho(e_i) = \rho(-e_i)$ (symmetric); $\rho(e_i) \ge \rho(e_j)$ for $|e_i| \ge |e_j|$ (monotonicity or non-decreasing function of $|x_i|$ and ρ is continuous (ρ is differentiable). Because the M- estimator is not scale invariant the minimization problem is modified by dividing the ρ function by a robust estimate of scale *s*, so the formula becomes

$$\min \sum_{i=1}^{n} \rho(\frac{e_i}{s}) = min \sum_{i=1}^{n} \rho(\frac{y_i - x_i\hat{\beta}}{s}), \quad (18)$$

where *s* is a robust estimate of scale. A popular choice for *s* is the median absolute deviation

$$s = \text{median}\ |e_i - \text{median}(e_i)|/0.6745.$$

The least squares estimator is a special case of the $\rho()$ function where $\rho(z) = \frac{1}{2}z^2$. The system of normal equations to solve this minimization problem is found by taking partial derivatives with respect to β and setting them equal to 0, yielding $\sum_{i=1}^{n} \psi\left(\frac{y_i - x_i'\hat{\beta}}{s}\right) x_i = 0$, where ψ is the derivative of ρ. In general, the ψ function is nonlinear and formula (18) must be solved by iterative methods. While several nonlinear optimization techniques could be employed, iteratively reweighted least squares (IRLS) is most widely used in practice and it is the only one considered for this research. The M-estimators can almost equivalently be described by a ρ function (posing a minimization problem) or by its derivative, an ψ function (yielding a set of implicit equation(s)), which is proportional to the influence function. Robust regression procedures can be classified by the behavior of their ψ function. The key to M- estimation is finding a good ψ function.

Although there are many specific proposals for the ψ-function, they can all be grouped into one of two classes: monotone and redescending. *A monotone ψ-function* (e.g. Huber's estimator) does not weight large outliers as much as least squares. The ψ -function of Huber estimator is constant-linearly increasing-constant. *A re-descending ψ-function* (e.g. Hample's and. Ramsay's) increases the weight assigned to an outlier until a specified distance and then decreases the weight to 0 as the outlying distance gets larger. Montgomery et al. (2012) introduce two types of a re-descending ψ -function: soft re-descender and hard re-descender. Alamgir et al. (2013) propose a new

re-descending M-estimator, called Alamgir re-descending M-estimator abbreviated as (ALARM). The ψ-function of ALARM estimator is defined as

$$\psi(e) = \begin{cases} \frac{16xe^{-2(e/b)}}{(1+e^{-(e/b)^4}} & \text{if } |e| \leq b; \\ 0 & \text{if } |e| > b, \end{cases} \quad (19)$$

where e denotes the error and b is a tuning constant. However, the M-estimators are not robust to x-axis outliers; therefore, their breakdown point is $1/n$ because of the effect of a single outlying observation. The M-estimators are statistically more efficient than L_1 at a central model with Gaussin errors (Rousseeuw and Yohai, 1984). Its asymptotic relative efficiency with respect to least squares is at least 0.81 (Birkes and Dodge, 1993). For regression analysis, some of the re-descending M-estimators can attain the maximum breakdown point. Moreover, some of them are the solutions of the problem of maximizing the efficiency under bounded influence function when the regression coefficient and the scale parameter are estimated simultaneously. Hence re-descending M-estimators satisfy several outlier robustness properties (Muthukrishnan and Radha, 2010).

3.3. Multiple Property Estimators (Related classes of efficient high breakdown estimators)

The discussion of robust estimators has clearly shown that no estimator has all the desirable properties. The multiple property estimators have been proposed to combine several properties into a single estimator.

The multi-stage (MM) estimator

Yohai (1987) introduces the multi-stage estimator (MM-estimator), which combines high breakdown with high efficiency. The MM-estimator is obtained using a three-stage procedure. In the first stage, an initial consistent estimate $\hat{\beta}_0$ with high breakdown point but possibly low normal efficiency is obtained. Yohai (1987) suggests using the S-estimator for this stage. In the second stage, a robust M-estimator of scale parameter $\hat{\sigma}$ of the residuals based on the initial value is obtained. In the third stage, an M-estimator $\hat{\beta}$ starting at $\hat{\beta}_0$ is obtained.

In practice, the LMS or S-estimate with Huber or bi-square functions is typically used as the initial estimate $\hat{\beta}_0$. Let $\rho_0(e) = \rho_1(e/k_0), \rho(e) = \rho_1(e/k_1)$, and assume that each of the ρ_i functions is bounded, i = 0 and 1. The scale estimate $\hat{\sigma}$ satisfies the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{y_i - x_i'\hat{\beta}}{\hat{\sigma}}\right) = 0.5. \quad (20)$$

If the ρ function is biweight, then $k_0 = 1.56$ ensures that the estimator has the asymptotic BP = 0.5. Although the MM-estimators have a high breakdown and are efficient, they do not necessarily have bounded influence, meaning that they may not perform especially well in the presence of high leverage points.

The τ- estimator

Yohai and Zamar (1988) introduce a new class of robust estimators; the τ -estimator. The τ -estimator has, simultaneously, the following properties: i) they are qualitatively robust, ii) their breakdown point is 0.5, and iii) they are highly efficient for regression models with normal errors.

In the τ-estimator, the coefficients are chosen so that a new estimator of the scale of the residuals is minimized as:

$$\min \tau_n(\beta), \quad (21)$$

where the τ-scale $\tau_n(\beta)$ is given by

$$\tau_n^2(\beta) = s_n^2(\beta)\frac{1}{nb_2}\sum_{i=1}^{n} \rho_2\left(\frac{y_i - x_i'\beta}{s_n(\beta)}\right), \quad (22)$$

with $s_n(\beta)$ is the M-estimator of scale that satisfies the solution

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{y_i - x_i'\beta}{s_n(\beta)}\right) = b_1. \quad (23)$$

The functions $\rho_j; j = 1,2$ are assumed to be symmetric, continuously differentiable bounded, strictly increasing on $[0.c_j]$, and constant on $[c_j, \infty)$, with $0 < c_j < +\infty, j = 1,2$.

The parameters b_1 and b_2 are tuned to obtain consistency for the scale at the normal error model:

$$b_j = E_\Phi[\rho_j(e)]; \quad j = 1,2, \quad (24)$$

where Φ is the standard normal distribution. Like the MM-estimators, the τ-estimators have the breakdown point of an S-estimator based on the loss function ρ_1, while its efficiency is determined by the function ρ_2 which is used in Eq. (18) (Yohai and Zamar, 1988). The τ-estimators possess theoretical advantages over the MM-estimators: they are associated with a robust and efficient scale estimate (Barrera et al., 2008).

The robust and efficiency weighted least squares (REWLSE) estimator

Gervini and Yohai (2002) introduce a new class of estimators that simultaneously attain the maximum breakdown point and full asymptotic efficiency under normal errors. They weight the least squares estimators with adaptively computed weights using the empirical distribution of the residuals of an initial robust estimator.

Consider a pair of initial robust estimators of parameters and scale $\hat{\beta}_0$ and $\hat{\sigma}_0$ respectively. If $\hat{\sigma}_0 > 0$, the standardized residuals are defined as:

$$r_i = \frac{y_i - X_i'\hat{\beta}_0}{\hat{\sigma}_0}. \quad (25)$$

A large value of $|r_i|$ would suggest that (x_i, y_i) is an outlier. In order to maintain the breakdown point value of the initial estimator and to have a high efficiency, they proposed the use of an adapted cut-off value, t_n, as $\min\{t: F_n^+(t) \geq 1 - d_n)$, where F_n^+ is the empirical cumulative distribution function of the standardized absolute residuals and d_n is the measure of the proportion of the outliers in the sample. The values of d_n are given as

$$d_n = \sup_{t \geq t_0}\{\Phi^+(t) - F_n^+(t)\}^+, \quad (26)$$

where Φ^+ denotes the normal cumulative distribution of the random errors, t_0=2.5 is the initial cut-off value and $\{\}^+$

denotes the positive part between $\Phi^{+}(t)$ and $F_n^{+}(t)$. The form of the weights, W, and the REWLS estimator, $\hat{\beta}_1$, are defined as

$$\hat{\beta}_{REWLSE}=\hat{\beta}_1=\begin{cases}(X^{\prime}WX)^{-1}X^{\prime}WY & \text{if } \hat{\sigma}_0>0;\\ \hat{\beta}_0, & \text{if } \hat{\sigma}_0>0.\end{cases} \quad (27)$$

The weight function, W, is chosen in order to have a hard-rejection to outliers

$$w_i=\begin{cases}1 & \text{if } |r_i|<t_{0;}\\ 0 & \text{if } |r_i|\geq t_0,\end{cases} \quad (28)$$

$W = \text{diag}(w_1, w_2, \ldots, w_n)$.

Note that with these weights, $\hat{\beta}_1$ estimator maintains the same breakdown point value of the initial estimator $\hat{\beta}_0$. Touati et al. (2010) modify the robust estimator of Gervini and Yohai (2002) and labeled it as the Robust and Efficient Weighted Least Squares Estimator (REWLSE), which simultaneously combines high statistical efficiency and high breakdown point by replacing the weight function by a new weight function.

4. Simulation Study

A simulation study is conducted to compare different methods of estimation. These methods are:

1. The ordinary least squares estimator (OLS);
2. The least median squares estimator (LMS);
3. The least trimmed squares estimator (LTS);
4. The S-estimator (S);
5. The least absolute value estimator (LAV);
6. The M-estimator; the Huber's M-estimator (M_{Huber}) with b=1.345;
7. The Tukey's M-estimator (M_{Tukey}) with k_1=4.685; and
8. The Hampel's M-estimators (M_{Hampel}) with a=1.7, b=3.4, and c=8.5;
9. The MM-estimators (MM) using bi-square weights and k_1=4.685, and
10. The robust and efficiency weighted least squares estimator (RWLSE).

The comparison criteria are the mean squares error (MSE), total mean squares error (TMSE), absolute bias (AB) and total absolute bias (TAB) of the estimates of the regression coefficients.

The data are generated according to the multiple linear regression model:

$$Y = 1 + X_1 + X_2 + X_3 + X_4 + e,$$

where $X_i \sim N(0,1), i = 1,2,3,4$ and the $X_i s$ are independent. The true value of the regression parameters are all equal one; $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. The data simulation is repeated 5000 times to obtain 5000 independent samples of Y and X of a given size *n*. The used sample sizes are *n* = 30 and *n* = 100. Comparisons of the properties of some robust methods are based on outliers in the y-direction (response direction). In order to cover the effects of various situations on the regression coefficients, seven scenarios of the density function of the errors (***e***) have been used. These scenarios are:

Scenario I: *e* ~ N (0, 1); the standard normal distribution.

Scenario II: *e* ~ t-distribution with df = 1; the t-distribution with degrees of freedom 1 (Cauchy distribution).

Scenario III: *e* ~ t-distribution with df = 5; the t-distribution with degrees of freedom 5.

Scenario IV: *e* ~ slash (0, 1); the slash distribution denoted by N (0, 1)/U (0, 1).

Scenario V: *e* ~ N (0, 1) with 20% outliers in *y*-direction generated from N (0, 10).

Scenario VI: *e* ~ N (0, 1) with 40% outliers in *y*-direction generated from N (0, 10).

Scenario VII: *e* ~ 0.80*N(0,1) + 0.20*N(0,10); contaminated mixture of normal.

For each scenario the mean square error (MSE), the total mean square error (TMSE), the absolute bias (AB) and the total absolute bias (TAB) have been obtained using the OLS-estimator, the LTS-estimator, the LMS-estimator, the S-estimator, the LAV-estimator, the M-estimators, the MM-estimator and the REWLSE. The results are not displayed for the sake of parsimony, however the following conclusions can be drawn.

Scenario I: the OLS estimate achieves the best performance. The decline in the performance of the other estimates compared to the performance of the OLS-estimate is the price paid by the methods due to the existence of outliers. The M-estimate, the MM-estimate, the REWLSE, and the LAV-estimate have better performance comparable to the estimates, as well as they provide a performance equal to the performance of the OLS-estimate. The performance of the high breakdown point estimates are poor compared to the OLS-estimate, the M-estimate, the LAV-estimate, the MM-estimate, or the REWLSE. As the sample size increases, the value of both the TMSE and the TAB decrease.

Scenario II: The OLS estimates give the worst result with regard to the TAB and the TMSE. The M_{Tukey} -estimate, the MM-estimate, the LAV-estimate, the REWLSE and the S-estimate tend to give lower TMSE and TAB values than other robust estimators. The LMS estimate and the LTS estimate have higher TMSE and TAB values than other robust estimators.

Scenario III: The OLS estimates give the highest TMSE and TAB values. The M_{Huber} estimates and the MM-estimates tend to give smaller TMSE and TAB values than the others. The LMS-estimates, the LTS estimates and the S-estimates have higher TMSE and TAB values. If the errors follow the t-distribution, the TMSE and TAB of each estimate decreases as the degrees of freedom (df) increases.

Scenario IV: The OLS estimates have the largest TMSE and TAB values. The M_{Tukey} estimate and the MM-estimates tend to give smaller TMSE and TAB than others. The LMS estimates and the LTS-estimates tend to give a higher TMSE and TAB than others.

Scenario V: It is noticed that when the data contain 20% outliers from the N (0, 10) in the *y*-direction, the

OLS-estimate has the largest TMSE and TAB. The REWLSE, the M_{Tukey} estimator and the MM-estimator, tend to give a lower TMSE and TAB than others.

Scenario VI: We can see that when the data contain 40% outliers from N (0, 10) in the *y*-direction, the OLS estimator decline in performance, while most of the other estimators are have good performances depending on the percentage and direction of contaminations. In general, the TMSE and TAB values decrease when the sample sizes increase while the TMSE and TAB values increase as the proportion of contaminations "outliers" increases.

Scenario VII: The OLS estimator has the largest TMSE and TAB values. The REWLSE, the M_{Tukey} estimator and the MM-estimator tend to give lower TMSE and TAB values than other robust estimators. The LMS-estimator and the LTS estimator achieve higher TMSE and TAB values than other robust estimators.

5. Application

A real data set (growth data) given by De Long and Summers (1991) is used. The aim is to evaluate the performance of various estimators. This data set measures the national growth of 61 countries from all over the world from the years 1960 to 1985. The data set contains many variables. They are the GDP growth per worker (GDP) which the response variable, the labor force growth (LFG), the relative GDP gap (GAP), the equipment investment (EQP), and non-equipment investment (NEQ). The main claim is that there is a strong and clear relationship between equipment investment and productivity growth. The regression equation they used:

$$GDP = \beta_0 + \beta_1\ LFG + \beta_2\ GAP + \beta_3 EQP + \beta_4\ NEQ + e, \quad (29)$$

where $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 are regression parameters.

Zaman et al. (2001) note that the growth study of De Long and Summers suffers from the problem of outliers in the y-direction (response direction). Zaman et al. (2001) use robust regression techniques and show that the 60th country in the data is an outlier.

We compare the OLS estimator, the LMS-estimator, the LTS-estimator, the S-estimator, the LAV-estimator, the Huber's M-estimator, the Tukey's M-estimator, the Hampel's M-estimator, the MM-estimator and the REWLSE by using bias and Mean square error values.

The results of the estimated regression parameters for the ten methods with all data points (complete data) are presented in Table (1). Also, results of the OLS method are displayed for the 60 data point (subset) ignoring the outlier; Zambia. The value of the bias and the MSE for the OLS estimate change from 0.1261 and 0.0302 (complete data) to 0 and 0.0139 (subset data). Thus, it is clear that the outlier strongly influences the OLS estimate. High-breakdown estimates (the LTS estimate, the LMS estimate, and the S-estimate) have low performance when the contamination is in the direction of the response variable only. The LAV estimate also has poor performance in this real data example.

Depending on the bias we can conclude that the Hampel's M-estimate and the REWLSE perform better than the others. When the MSE is considered, it indicates that the Hampel's M-estimate and the Huber's M-estimate perform better than the others.

Table (1). Estimates of regression coefficients, the bias and the MSE for growth data

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Bias	MSE
OLS-subset	-0.0222	0.0446	0.0245	0.2824	0.0849	0.0000	0.0139
OLS-complete	-0.0143	-0.0298	0.0203	0.2654	0.0624	0.1261	0.0302
LMS	-0.0249	0.1962	0.0170	0.2563	0.1296	0.2326	0.0681
LTS	-0.0286	0.2335	0.0198	0.2805	0.1221	0.2391	0.0748
S	-0.0263	0.1019	0.0216	0.2636	0.1239	0.1706	0.0403
LAV	-0.0248	0.1321	0.0228	0.3095	0.0886	0.1226	0.0316
M_{Huber}	-0.0217	0.0735	0.0236	0.2865	0.0808	0.0385	0.0154
M_{Tukey}	-0.0247	0.1040	0.0250	0.2968	0.0885	0.0804	0.0215
M_{Hampel}	-0.0199	0.0367	0.0230	0.2810	0.0766	0.0214	0.0143
MM	-0.0239	0.0883	0.0247	0.2921	0.0874	0.0578	0.0178
REWLSE	-0.0232	0.0732	0.0242	0.2891	0.0864	0.0380	0.0157

6. Conclusions

The mean square errors (MSE) and the bias are of interest in regression analysis in presence of outliers. The performances of different estimates are studied using a simulation study and a real data for outliers in *y*-direction. The estimates of the regression coefficients using nine methods are compared with the ordinary least-squares. Depending on the simulation study the Tukey's M-estimator give a lowest TAB and TMSE values than others, for all sample sizes and when the contamination is in the *y*-direction. For the real data the Hampel's M-estimators gives lower bias and MSE values than others when the contamination is in the *y*-direction.

The work can be extended in future to handle outliers in *x*-direction; good leverage points. Another possible future direction is to compare robust methods when outliers are in both y-direction and x-direction (bad leverage points).

ACKNOWLEDGEMENTS

The authors would like to thank referees and editors for their help and constructive comments that improve significantly the manuscript.

REFERENCES

[1] Alamgir, A. A., Khan, S. A., Khan, D. M. and Khalil, U. (2013) A new efficient redescending M-estimator: Alamgir redescending M-estimator, *Research Journal of Recent Sciences*, 2, 79-91.

[2] Anderson, C. (2001) *A comparison of five robust regression methods with ordinary least squares: relative efficiency, bias and test of the null hypothesis*, Ph. D. thesis, University of North Texas, USA.

[3] Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*, 3rd Edition, Wiley, New York, USA.

[4] Barrera, M. S., Willems, G. and Zamar, R. (2008) The fast τ estimator for regression, *Journal of Computational and Graphical Statistics*, 17, 1-24.

[5] Birkes, D. and Dodge, Y. D. (1993) *Alternative methods of regression*, Wiley, New York, USA.

[6] Cankaya, S. (2009) A comparative study of some estimation for parameters and effects of outliers in simple regression model for research on small ruminants, *Trop. Anim. Health Prod.*, 41, 35-41.

[7] Chatterjee, S. and Hadi, A.S. (1986) Influential observations, high leverage points, and outliers in linear regression, *Statistical Science*, 1, 379-393.

[8] De Long, J. B. and Summers, L. H., (1991) Equipment investment and economic growth, *Quarterly Journal of Economics*, 106, 445–501.

[9] Fox, J. (1997) *Applied Regression Analysis, Linear Model and Related Methods*, 3rd Edition, Sage Publication, USA.

[10] Gervini, D. and Yohai, V. J. (2002) A class of robust and fully efficient regression estimators, *The Annals of Statistics*, 30, 583-616.

[11] Hampel, F. R. (1968) *Contributions to the theory of robust estimation*, Ph.D. Thesis, University of California, Berkeley, USA.

[12] Hampel, F. R. (1974) The influence curve and its role in robust estimation, *Journal of the American Statistical Association,* 69, 383-393.

[13] Touati, F., Kahlouche, S. and Idres, M. (2010) Robust and efficient weighted least squares adjustment of relative gravity data *Gravity*, In *Geoid and Earth Observation*, edited by Mertikas, S.P., *International Association of Geodesy Symposia 135*, Springer-Verlag Berlin Heidelberg.

[14] Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012) *Introduction to linear regression analysis*, Wiley, New York, USA.

[15] Muthukrishnan, R. and Radha, M. (2010) M-estimators in regression models, *Journal of Mathematics Research*, 2, 23-27.

[16] Nevitt, J. and Tam, H. P. (1998) A comparison of robust and nonparametric estimators under the simple linear regression model, *Multiple Linear Regression Viewpoints*, 25, 54-69.

[17] Rousseeuw, P. J. (1984) Least median of squares regression, *Journal of the American Statistical Association*, 79, 871-880.

[18] Rousseeuw, P. J. and Yohai, V. (1984) Robust regression by means of S-estimators, In *Robust and Nonlinear Time Series Analysis*, 256-272, Springer Verilag, New York, USA.

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

[20] Rousseeuw, P. J., van Driessen, K. (2006) Computing LTS regression for large data sets, *Data Mining and Knowledge Discovery*, 12, 29-45.

[21] Seber, G. A. F. (1977) *Linear Regression Analysis*, Wiley, New York, USA.

[22] Siegel, A. F. and Benson, R. H. (1982) A robust comparison of biological shapes, *Biometrics*, 38, 341-350.

[23] Wilcox, R. R. (2010) *Fundamentals of Modern Statistical Methods*", 2nd. Edition, Springer, London.

[24] Yale, C. and Forsythe, A. B. (1976) Winsorized regression, *Technometrics*, *18*, 291- 300.

[25] Yohai, V. J. (1987) High breakdown point and high efficiency robust estimates for regression, *Annals of Statistics*, 15, 642–656.

[26] Yohai, V. J. and Zamar, R. H. (1988) High breakdown-point estimates of regression by means of the minimization of an efficient scale, *Journal of the American Statistical Association,* 83, 406-413.

[27] Zaman, A., Rousseeuw, P. J. and Orhan, M. (2001) Econometric applications of high-breakdown robust regression techniques, *Economics Letters*, 71, 1–8.