British Journal of Mathematics & Computer Science
13(4): 1-11, 2016, Article no.BJMCS.22876
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Ordinary Least Squares and Robust Estimators in Linear Regression: Impacts of Outliers, Error and Response Contaminations

David Adedia[1*], Atinuke Adebanji[2], Moremi Labeodan[3] and Shola Adeyemi[4]

[1] *Department of Basic Sciences, University of Health and Allied Sciences, P.M.B. 31, Ho, Ghana.*
[2] *Department of Mathematics, Kwame Nkrumah University of Science and Technology, P.M.B. KNUST, Kumasi, Ghana.*
[3] *School of Business and Economics, Monash University (South African Campus), P/B X60, Roodepoort, South Africa.*
[4] *Sirtex Medical Europe GmbH, Bonn, Germany.*

Article Information

DOI: 10.9734/BJMCS/2016/22876
Editor(s):
(1) Jacek Dziok, Institute of Mathematics, University of Rzeszow, Poland.
Reviewers:
(1) Cliff Richard Kikawa, Tshwane University of Technology, South Africa.
(2) Weihua Zhao, Nantong University, China.
(3) Diana Blkova, University of Economics, Czech Republic.
Complete Peer review History: http://sciencedomain.org/review-history/12717

Review Article

Received: 02 November 2015
Accepted: 28 November 2015
Published: 17 December 2015

Abstract

The Ordinary Least Squares Estimator (OLSE) is the best method for linear regression if the classical assumptions are satisfied for estimating weights. When these assumptions are violated, the robust methods give more reliable estimates while the OLSE is strongly affected adversely. In order to assess the sensitivity of some estimators using more than five criteria, a secondary dataset on Anthropometric measurements from Komfo Anokye Teaching Hospital, Kumasi-Ghana, is used. In this study, we compare the performance of the Huber Maximum Likelihood Estimator (HMLE), Least Trimmed Squares Estimator (LTSE), S Estimator (SE) and Modified Maximum Likelihood Estimator (MMLE) relative to the OLSE when the dataset has normal errors; 10, 20 and 30 percent outliers; 20% error contamination and lognormal contamination in the response

**Corresponding author: E-mail: dadedia@uhas.edu.gh;*

variable. In the assessment, we use coefficients and their standard errors, relative efficiencies, Root Mean Square Errors, and the coefficients of determination of the estimators. We also use the power of the test to assess the effects of the aberrations on the post hoc power analysis of the estimators. The results show the SE and MMLE outperform the HMLE and LTSE while the OLSE breaks down completely. The LTSE performs well when the trimming is done to eliminate only the outliers. Also, SE and MMLE resist the effect of all aberrations in the data and also have good post hoc power analysis.

Keywords: Ordinary least squares estimator; robust estimators; power of the test; outliers; errors.

2010 Mathematics Subject Classification: 62F35, 62G35.

1 Introduction

The Ordinary Least Squares estimator is developed by [1], for estimating regression parameters. According to [2], this method performs well when the assumptions the method impose on the dataset are satisfied, but failure of these assumptions renders the OLSE incapable of providing stable results. The OLSE is mainly susceptible to the effects of vertical outliers, errors, leverages, and contaminations in the data. Good leverages are those leverages that are outlying in the dimension of predictors but are close to the regression line. Bad leverages do affect intercept and the slopes of the OLSE, and they are outliers in the space of predictors and are far from the fitted line. Vertical outliers are those observations that have outlying values for the corresponding error term, and their presence affects the ordinary least squares estimation. A dataset is said to be contaminated if data points from other distribution are present in the data of the distribution under study. A contaminant can be an outlier or an inlier. OLSE is sensitive to these aberrations leading to unreliable estimates including inflated standard errors. As a result, robust methods were introduced to take care of the limitations of the OLSE. Some pioneers of Robust regression methods include; ([3]-[9]) and many others. In this study, we assess the sensitivity of OLSE and some robust regression estimators to vertical outliers, errors and contaminations in the data. One of the focuses of this study is to compare the performance of the estimators across the six criteria: coefficients and their standard errors, relative efficiencies, Root Mean Square Errors, coefficients of determination and the power of the test.

2 Multiple Linear Regression

The multiple linear regression model can be written in matrices notation as

$$y = X\beta + e \tag{2.1}$$

where y is an $n \times 1$ vector of observed response values, X is the $n \times p$ matrix of the predictor variables, β is the $p \times 1$ vector contains the unknown parameters and has to be estimated, and e is the $n \times 1$ vector of random error terms. Therefore to fit this model to the data, we have to use a regression estimator to estimate the unknown parameters in β, to have $\hat{\beta}$, where $\hat{\beta}^T = (\hat{\beta}_1........\hat{\beta}_P)$. The expected value of y_i, that is the fitted value, $E(y_i)$ is given by

$$\hat{y_i} = X_i^T \hat{\beta}. \tag{2.2}$$

As a result the residuals can be computed using $e_i = y_i - \hat{y_i}$, where $i = 1, 2, ..., n$ and n is the sample size. According to [10], if the assumptions of the error terms are met, that is the $e_i \sim N(0, \sigma^2)$, then the least squares regression estimator is the maximum likelihood estimator for β. In practice, the assumption of normality often holds approximately in that it describes majority of the observations but some observations are found to have a different or no-pattern at all. The presence of such atypical observations can have a large distorting influence on the classical optimal statistical method.

2.1 The ordinary least squares estimator

The least squares estimator aims to minimize the sum of the square residuals as $\sum e^2 = (Y - X\beta)^T(Y - X\beta)$. Therefore, using the OLSE to estimate the regression parameters in (2.1), we have

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{2.3}$$

We can compute the least squares estimates directly from any dataset when $X^T X$ is nonsingular. However, if the assumptions of the OLSE are not met, the OLSE cannot be used to estimate the regression parameters because the estimates are highly distorted. As a result, some robust methods were introduced by ([11]-[14]) to address this inadequacy.

2.2 Robust regression estimators

Tiku and Akkaya [15] defined robust estimators as estimators which perform well when errors are normally distributed and do not breakdown when errors deviate from normality. These estimators are defined by properties of efficiency with high breakdown points.

2.3 Huber maximum likelihood estimator

The class of M-estimator models contains all models that are derived to be maximum likelihood models. The most common method of robust regression is M-estimation by [11]. The M-estimator minimizes a function ρ of the errors given as

$$\sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = \sum_{i=1}^{n} \rho\left(\frac{y_i - x^T\beta}{s}\right), \tag{2.4}$$

where "s" is an estimate of scale from a linear combination of the residuals. The function ρ gives the contribution of each residual to the objective function. Alma [2] stipulates that a reasonable ρ should have the following properties, $\rho(e) \ \geq 0$, $\rho(0) = 0$, $\rho(e) = \rho(-e)$, $\rho(e_i) \geq \rho(e_i^{'})$ for $|e_i| \geq \left|e_i^{'}\right|$, ρ is continuous [16], and the objective function of the least squares estimation is given by $\rho(e_i) = e_i^2$. So we minimize equations (2.4) with respect to each of the p parameters in (2.4).

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{e_i}{s}\right) = \sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - x_i^T\beta}{s}\right) = 0 \tag{2.5}$$

where $j = 1, 2, ..., p$ and $i = 1, 2, ..., n$; and $\psi(u) = \frac{\partial\rho}{\partial u}$ is the score function. We define a weight function as $w(u) = \frac{\psi(u)}{u}$, where $u = \frac{y_i - x_i^T}{s}$, which results in $w_i = w\left(\frac{e_i}{s}\right)$, for $i = 1, 2, ..., n$ with $w_i = 1$ if $e_i = 0$. Substituting this into (2.5) results in

$$\sum_{i=1}^{n} x_{ij} w_i \left(\frac{y_i - x_i^T\beta}{s}\right) = 0 \tag{2.6}$$

Since $s \neq 0$, we define the weight matrix $W = diag(w_i : i = 1, 2, ..., n)$. Solving (2.6) above for β the subject results in the equation

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y \tag{2.7}$$

2.4 Least trimmed squares estimator

Rousseeuw's 1984 LTSE, [9], is given by

$$\hat{\beta} = min \sum_{i=1}^{h} (e_i^2) \tag{2.8}$$

where ($e_1^2 \leq e_2^2 \leq, ..., \leq e_n^2$ are the ordered squared residuals) and LTSE is computed by minimizing the h ordered squared residuals, where $h = \left(\left[\frac{n}{2}\right] + 1\right)$; when n and h are the sample size and the trimming constant, respectively [17]. Using the h trimmed dataset ensures that estimates have a high breakdown point of 50% but a low efficiency of 7.13%, [18].

Rousseeuw and Leroy [13], suggested a trimming constant of $h = [n[1 - \alpha] + 1]$ where α is the trimmed percentage. The largest squared residuals are deleted and the least squares method is applied on the trimmed dataset. Moreover, this helps to discard outliers and their influence on the regression estimators. LTSE can be very efficient based on the size of the trimmed dataset (h) and the level of outliers in the data. However, in situations where there are more outliers and only some are trimmed this method can also perform as poorly as the ordinary least squares method of estimation. Also, if more observations are deleted where there are only few outliers, good data points will be discarded from the dataset. Least-trimmed-squares has a break-down point of 50%, hence makes the LTSE a high break-down method of estimation. This implies that half of the data has to be influential points before estimates of the least trimmed estimator be affected when the method of the ordinary least squares is applied. LTSE essentially proceeds with OLSE after the deletion of the most extreme positive or negative residuals. LTSE on the other hand, can misrepresent the trend in the data if it is characterized by clusters of extreme cases or if the data set is relatively small. The breakdown value is $\frac{n-h}{n}$ for the LTSE estimate.

2.5 S-estimator

The S-estimation introduced by [12] minimizes the dispersion of the residuals. The high breakdown S-estimator possesses a desirable property, that is, it is affine, scale and regression equivariant [18]. S-estimator minimizes the dispersion of the scaled residuals, that is, S Estimator is the $\hat{\beta}$ that makes $s(r(\beta_1), ...r(\beta_n)$ being minimal. The robust S-estimation minimizes a robust M-estimate of the residual scale

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{e_i}{s} \right) = k. \tag{2.9}$$

Differentiating (2.9) we obtain the estimating equations for S-estimator

$$\frac{1}{n} \sum_{i=1}^{n} x_i \psi \left(\frac{e_i}{s} \right) = 0 \tag{2.10}$$

where ψ is replaced with an appropriate weight function. Biweight or Huber function is usually used as with most M-estimation procedures. Although S-estimates have a breakdown point of Break Down Point $(BDP) = 0.5$, it comes at the cost of a very low relative efficiency [19]. The choice of the tuning constant is $a = 1.548$ and $k = 0.1995$ for 50% breakdown and about 29% asymptotic efficiency. To increase the efficiency of the S-estimator, if $a = 5.182$, the Gaussian efficiency rises to 96.6% but the breakdown point drops to 10%. Tradeoffs breakdown and efficiency are based on the selection of tuning constant a, and k. The final scale estimate, s, is the standard deviation of the residuals from the fit that minimized the dispersion of the residuals. [12] stated that if we set the tuning constant $a = 1.547$,this makes the S-estimator to have 50% BDP.

2.6 Modified maximum likelihood estimator

The MMLE is a special type of M-estimation developed by [16]. Alma [2] defined MMLE as the estimator with high breakdown value and high efficiency. It was the first estimator with a high breakdown point and high efficiency under normal error. MMLE has three-stage procedure described by [16] as

1. A high breakdown estimator is used to find an initial estimate, which we denote $\tilde{\beta}$. The estimator needs not to be efficient. Using this estimate the residuals, $r_i(\tilde{\beta}) = y_i - x_i^T\tilde{\beta}$, are computed.
2. Using these residuals from the robust fit and (2.9), an M-estimate of scale with 50% BDP is computed. This $s(r_1(\tilde{\beta})...r_n(\tilde{\beta}))$ is denoted s_n. The objective function used in this stage is labeled ρ_0.
3. The MM-estimator is now defined as an M-estimator of β using a redescending score function, $\psi_1(u) = \frac{\partial \rho_1(u)}{\partial u}$, and the scale estimate s_n obtained from Stage 2. So an MM-estimator $\hat{\beta}$ is defined as a solution to

$$\sum_{i=1}^{n} x_{ij}\psi_1\left(\frac{y_i - x_i^T\beta}{s_n}\right) = 0 \tag{2.11}$$

where $j = 1, ..., p$. The objective function ρ_1 associated with this score function does not have to be the same as ρ_0 but it must satisfy the following conditions:

(a) ρ is symmetric and continuously differentiable, and $\rho(0) = 0$.

(b) There exists $a > 0$ such that ρ is strictly increasing on $[0, a]$ and constant on $[a, \infty)$.

(c) $\rho_1(u) \leq \rho_0(u)$
A final condition that must be satisfied by the solution to (2.11) is that

$$\sum_{i=1}^{n} x_{ij}\psi_1\left(\frac{y_i - x_i^T\hat{\beta}}{s_n}\right) \leq \sum_{i=1}^{n} x_{ij}\psi_1\left(\frac{y_i - x_i^T\tilde{\beta}}{s_n}\right). \tag{2.12}$$

The first two stages of the MM-estimation process are responsible for the estimator having high breakdown point, whilst the third stage aims for high asymptotic relative efficiency. This is why ρ_0 and ρ_1 need not be the same, and why the estimator chosen in stage 2 can be inefficient. Yohai [16] showed that when estimating MM-estimator, using an estimator with 50% BDP at the first stage will result in the final MM-estimator has 50% BDP. The MM-estimator is very resistant to multiple leverage points and vertical outliers. The MME is also equivariant and hence it transforms 'properly' in some sense, [13].

3 Results and Discussion

To assess the sensitivity of the robust methods above, we used datasets on anthropometric measurements of patients from Komfo Anokye Teaching Hospital (KATH). The dataset is on Body fat as response and Body Mass Index (BMI), Triceps skin-fold (TS), Arm Fat as percentage composition of the body (parmfat) and Height as predictors which are measured with OMRON machine. We also use coefficients of the model for the original data with normal errors as the parameters and in conjunction with simulated predictors from log-normal distribution to simulate the response variable using R software.

3.1 Dataset with normal errors

Tables 1 and 2 contain the measures for comparing the estimators for dataset with normal errors.

Table 1. The coefficients (standard errors) of the estimators for original dataset with normal errors

Methods	Intercept	BMI	Parmfat	Height	TS
OLSE	7.362(0.945)	0.845(0.056)	0.145(0.082)	0.004(0.010)	0.288(0.018)
LTSE	7.544(0.872)	0.830(0.051)	0.204(0.076)	0.002(0.010)	0.281(0.017)
HME	7.407(1.012)	0.84(0.059)	0.163(0.088)	0.004(0.011)	0.287(0.020)
SE	7.454(1.007)	0.832(0.059)	0.173(0.088)	0.005(0.011)	0.287(0.020)
MME	7.419(1.008)	0.836(0.059)	0.165(0.088)	0.004(0.011)	0.2869(0.020)

Source: [20]

Table 2. Residual standard error, relative efficiency, coefficient of determination and the power of the test for original dataset with normal errors

Method of estimation	Standard error	Relative efficiency	Coefficient of determination	Power of the test
OLSE	1.0650	1.0000	0.9696	1.0000
LTSE	0.9641	1.2203	0.9641	1.0000
HME	1.2050	0.7811	0.9574	1.0000
SE	1.0720	0.9870	0.9586	1.0000
MME	1.0670	0.9963	0.9577	1.0000

Source: [20]

From Table 1, it is observed that all the estimators performed well, since the errors are normally distributed. The standard errors, Relative efficiencies and the coefficients of determination from Table 2, also showed that when the errors are normal, all the estimators perform well.

3.2 10% outliers

The coefficients for the different estimation methods are presented in Table 3.

Table 3. The coefficients (standard error) of the estimators for 10% outliers perturbation

Methods	Intercept	BMI	Parmfat	Height	TS
OLSE	5.642(13.035)	1.580(0.765)	-0.597(1.137)	-0.048(0.143)	0.1941(0.255)
LTSE	7.257(0.977)	0.849(0.058)	0.146(0.087)	0.003(0.011)	0.289(0.020)
HME	6.924(1.316)	0.895(0.077)	0.125(0.115)	0.0003(0.014)	0.279(0.026)
SE	7.295(1.064)	0.844(0.063)	0.158(0.093)	0.004(0.012)	0.288(0.021)
MME	7.257(1.014)	0.849(0.060)	0.1464(0.090)	0.003(0.011)	0.289(0.021)

Source: [20]

From Tables 3 and 4, it is observed that only OLSE was affected when the dependent variable has 10% atypical observations. The coefficients of the OLSE have values which differ much from when the errors were normally distributed. It also has the least coefficient of determination among all the estimators. Moreover, OLSE assuming large value for residual standard error and small value for relative makes it unreliable.

Table 4. Standard error, relative efficiency, coefficient of determination and power of the test for 10% outliers

Methods	Standard error	Relative efficiency	Coefficient of determination	Power of the test
OLSE	14.6900	1.0000	0.1702	0.9850
LTSE	1.0810	184.6682	0.9681	1.0000
HME	1.2850	130.6885	0.9674	1.0000
SE	1.3590	116.8435	0.9616	1.0000
MME	1.3460	119.1114	0.9609	1.0000

Source: [20]

3.3 20% outliers

With 20% outliers, the coefficients and measures of performance are presented in Tables 5 and 6 respectively.

Table 5. The coefficients (standard error) of the estimators for 20% outliers

Methods	Intercept	BMI	Parmfat	Height	TS
OLSE	-3.678(19.725)	2.278(1.158)	-1.268(1.720)	-0.009(0.216)	0.086(0.385)
LTSE	5.901(3.021)	0.878(0.184)	0.228(0.266)	0.007(0.031)	0.275(0.061)
HME	6.289(1.732)	0.927(0.102)	0.137(0.151)	-0.0001(0.019)	0.280(0.034)
SE	7.436(1.114)	0.836(0.065)	0.188(0.097)	0.001(0.012)	0.293(0.022)
MME	7.459(1.033)	0.838(0.061)	0.182(0.091)	0.0004(0.011)	0.294(0.021)

Source: [20]

Table 6. Standard error, relative efficiency, coefficient of determination and power of the test for 20% outliers

Methods	Standard error	Relative efficiency	Coefficient of determination	Power of the test
OLSE	22.2300	1.0000	0.1186	0.9000
LTSE	3.0940	51.6224	0.8083	1.0000
HME	1.5520	205.1613	0.9638	1.0000
SE	1.6270	186.6826	0.9621	1.0000
MME	1.6130	189.9373	0.9618	1.0000

Source: [20]

The OLSE failed on almost all the criteria. It has large residual standard error, low relative efficiency and low coefficient of determination, which undermines the usefulness of the OLSE for this perturbed dataset. On the contrary, robust methods such as MME and SE perform well, this is because, they reported estimates which are similar to estimates for the normal error dataset.

3.4 30% outliers

Tables 7 and 8 present the numerical measures (criteria) for comparing the regression estimators when there are 30% vertical outliers.

The OLSE broke down completely at this stage with some robust methods slightly affected. Some of the robust estimators have very large standard errors and small coefficients of determination

Table 7. The coefficients (standard error) of the estimators for 30% outliers perturbation

Methods	Intercept	BMI	Parmfat	Height	TS
OLSE	-18.310(19.239)	3.193(1.129)	-2.691(1.678)	0.141(0.211)	-0.408(0.376)
LTSE	-16.837(8.393)	2.641(0.502)	-1.361(0.726)	0.034(0.085)	-0.217(0.176)
HME	-13.141(13.952)	2.6703(0.819)	-1.843(1.217)	0.062(0.153)	-0.120(0.272)
SE	7.676(1.206)	0.837(0.071)	0.180(0.105)	-0.001(0.013)	0.296(0.024)
MME	7.683(1.211)	0.839(0.074)	0.176(0.1067)	-0.002(0.012)	0.296(0.024)

Source: [20]

Table 8. Standard error, relative efficiency, coefficient of determination and power of the test 30% outliers perturbation

Methods	Standard error	Relative efficiency	Coefficient of determination	Power of the test
OLSE	21.6800	1.0000	0.1132	0.9800
LTSE	8.4150	6.6376	0.5075	1.0000
HME	13.3400	2.6412	0.4934	1.0000
SE	2.0980	106.7843	0.9690	1.0000
MME	2.0470	112.1716	0.9682	1.0000

Source: [20]

with unreliable coefficients. The modified maximum likelihood estimator and the S-estimator were robust to the influence of the outliers.

3.5 20% error contamination

Tables 9 and 10 present the estimates for regression parameters from a dataset with contaminated response variable. The 20% of the observations of the response variable were replaced with observations from the Cauchy distribution.

Table 9. The coefficients (standard error) of the estimators for 20% error contamination

Methods	Intercept	BMI	Parmfat	Height	TS
OLSE	-10.326(122.134)	11.743(7.169)	-10.774(10.649)	-1.313(1.338)	-1.068(2.384)
LTSE	1.715(23.406)	5.447(1.120)	-4.589(1.510)	-0.531(0.334)	-0.410(0.322)
HME	7.533(1.579)	0.929(0.093)	-0.027(0.138)	0.002(0.017)	0.273(0.031)
SE	7.488(0.994)	0.8452(0.058)	0.136(0.087)	0.002(0.011)	0.292(0.019)
MME	7.427(0.946)	0.8528(0.057)	0.1231(0.086)	0.002(0.010)	0.292(0.020)

Source: [20]

In this section, 20% error contamination from Cauchy distribution distorted the performance of some estimators. The OLSE and LTSE performed poorly. However, estimators such as, MME, SE and HME performed well. The performances of the MME and SE are very impressive in this study. The coefficients of MME and SE are very similar to that of the normal errors. In addition, the residual standard error, coefficients of determination and power of the test are also analogous to the estimates of the original data with normal errors.

Table 10. Standard error, relative efficiency, coefficient of determination and power of the test 20% error contamination

Methods	Standard error	Relative efficiency	Coefficient of determination	Power of the test
OLSE	137.6000	1.0000	0.0261	0.2500
LTSE	15.6700	77.1079	0.2376	0.9960
HME	1.4140	9469.7399	0.9683	1.0000
SE	1.4160	9443.0081	0.9695	1.0000
MME	1.4030	9618.8140	0.9695	1.0000

Source: [20]

3.6 Log-normal contamination

Distributional robustness of the robust methods was assessed by simulating dataset from log-normal distribution. The coefficients of the model for the original data with normal errors were used as the parameters and in conjunction with simulated predictors from log-normal distribution to simulate the response variable. Tables 11 and12 below present the results for the simulated dataset.

Table 11. The coefficients (standard errors) of the estimators for non-normal distribution(lognormal)

Methods	Intercept	BMI	Parmfat	Height	TS
OLSE	207.577(429.437)	10.677(10.669)	-3.854(21.11)	-2.859(1.982)	-1.114(6.238)
LTSE	47.622(26.191)	0.4520(0.6718)	-0.453(1.274)	-0.1289(0.125)	0.478(0.379)
HME	11.269(3.373)	0.804(0.084)	0.213(0.166)	-0.0051(0.016)	0.320(0.049)
SE	8.558(1.372)	0.837(0.034)	0.134(0.067)	0.007(0.006)	0.283(0.012)
MME	7.969(1.740)	0.843(0.041)	0.152(0.075)	0.008(0.007)	0.288(0.024)

Source: [20]

Table 12. Standard error, relative efficiency, coefficient of determination and power of the test for non-normal distribution(lognormal)

Methods	Standard error	Relative efficiency	Coefficient of determination	Power of the test
OLSE	582.8000	1.0000	0.0273	0.2600
LTSE	35.1200	275.3785	0.0235	0.2200
HME	4.2420	18875.4693	0.5737	1.0000
SE	1.7270	113881.8231	0.9394	1.0000
MME	1.6550	124006.1117	0.9469	1.0000

Source: [20]

From the Tables 11 and 12, the R^2 values show that OLSE and LTSE broke down with HME slightly affected. Moreover, using the coefficients and the relative efficiency, it is also clear that OLSE and LTSE did not do well. This is because the data was simulated from a heavy tailed distribution. On the other hand, MME and SE were still resistant to the aberrations in the data. Moreover, robust methods like MME and SE are insensitive to data from fat tailed distributions, therefore, the results in this section have illustrated the distributional robustness of MME and SE.

4 Conclusion

This study compares the performance of some robust regression methods against the ordinary least squares method by using more than one criterion. The results show that the robust methods such as MME and SE are resistant to all manner of aberrations: outliers and contaminations in the datasets. The study also shows the distributional robustness of MME and SE. OLSE and LTSE performed poorly with 20% contaminations. This illustrates the vulnerability of OLSE and LTSE when the dataset to be used is highly contaminated. The results again show that HME performed averagely when there are high outliers and error contaminations.The loss of power of the OLSE and LTSE with 20% contaminations also indicates how unreliable they are. This is because the contaminated dataset report small power of the test for them as compared to the other estimators. Therefore, in this study, we realize that MME and SE perform excellently in all the datasets and across all the criteria.

Acknowledgement

We are most grateful to Dr. Kofi Dorvlo for academic guidance and support. We also express our gratitude to Mrs Ivy Mensah-Adedia for her valuable suggestions.

Competing Interests

The authors declare that there is no competing interests regarding the publication of this manuscript.

References

[1] Gauss CF. Theoria motus corporum coelestium in sectionibus conicis solem ambientium; 1809. Available: http://www.historyofinformation.com/expanded.php?id=2707

[2] Alma OG. Comparison of robust regression methods in linear regression. International Journal for contemp Maths and Science. 2011;6:409421.

[3] Tukey JW. A survey of sampling from contaminated distributions, contributions to probability and statistics, Olkin I. (ed.), Stanford, CA: Stanford University Press; 1960.

[4] Tukey JW. The future of data analysis. The Annals of Mathematical Statistics. 1962;33:167.

[5] Huber PJ. Robust estimation of a location parameter. The Annals of Mathematical Statistics. 1964;35:73101.

[6] Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions, Proceedings of the Fifth Berkeley Symposium on Mathematics and Statistics Probability. University of California Press. 1967;1:221233.

[7] Hampel FR. A general definition of qualitative robustness. The Annals of Mathematical Statistics. 1971;42:18871896.

[8] Hampel FR. The influence curve and its role in robust estimation. The Annals of Statistics. 1974;69:383393.

[9] Rousseeuw PJ. Least median of squares regression. Journal of the American Statistical Association. 1984;79:871880.

[10] Stuart C. Robust regression. Department of Mathematical Sciences, Durham University. 2011;169.

[11] Huber PJ. Robust regression: Asymptotics, conjectures and monte carlo. The Annals of Statistics. 1973;799821.

[12] Rousseeuw P, Yohai V. Robust regression by means of s-estimators. Springer. 1984;256272.

[13] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. John Wiley and Sons; 1987.

[14] Maronna RA, Martin RD, Yohai VJ. Robust statistics, theory and methods. John Wiley and Sons Ltd; 2006.

[15] Tiku ML, Akkaya AD. Robust estimation and hypothesis testing. New Age International (P) Ltd; 2004.

[16] Yohai VJ. High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics. 1987;15:642656.

[17] AL-Noor HN, Mohammad AA. Model of robust regression with parametric and nonparametric methods. Mathematical Theory and Modeling. 2013;3:2739.

[18] Matias SB, Yohai VJ. A fast algorithm for s-regression estimates. Journal of Computational and Graphical Statistics. 2006;15(2):414427.

[19] Verardi V, Croux C. Robust regression in stata. The Stata Journal. 2009;3:439453.

[20] Adedia D. Comparison of robust regression estimators, unpublished MPhil thesis, KNUST, Kumasi; 2014.
Available: http://ir.knust.edu.gh/bitstream/123456789/6629/1/david%20adedia.pdf

---

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/12717