British Journal of Mathematics & Computer Science
15(3): 1-14, 2016, Article no.BJMCS.24281
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Leverages, Outliers and the Performance of Robust Regression Estimators

David Adedia[1*], Atinuke Adebanji[2], Eric Okyere[1] and James Kwaku Agyen[1]

[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.*

Authors' contributions

This work was carried out in collaboration between all authors. Authors DA and AA designed the study, performed the statistical analysis, wrote the protocol, and wrote the first draft of the manuscript and managed literature searches. Authors EO and JKA managed the analyses of the study and literature searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJMCS/2016/24281
Editor(s):
(1) Raducanu Razvan, Department of Applied Mathematics, Al. I. Cuza University, Romania.
Reviewers:
(1) P. E. Oguntunde, Covenant University, Nigeria.
(2) Radosaw Jedynak, Kazimierz Puaski University of Technology and Humanities in Radom, Poland.
Complete Peer review History: http://sciencedomain.org/review-history/13653

Original Research Article

Received: 12th January 2016
Accepted: 17th February 2016
Published: 12th March 2016

Abstract

This study compared the performance of some robust regression methods and the Ordinary Least Squares Estimator (OLSE). The estimators were compared using varied levels of leverages and vertical outliers in the predictors and the dependent variables. An anthropometric dataset on total body fat with height, Body Mass Index (BMI), Triceps Skin-fold(TS), and arm fat as percent composition of the body (parmfat), as the predictors. The effects of outliers and leverages on the estimators, were investigated at (5% leverages and 10% vertical outliers, 5% leverages with 15% vertical outliers). The criteria for the comparison: coefficients, Root Mean Square Error (RMSE), Relative Efficiencies (RE), coefficients of determination (R-squared) and power of the test. The

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

findings from this study revealed that, OLSE was affected by both outliers and leverages whilst Huber Maximum likelihood Estimator (HME) was affected by leverages. The Least Trimmed Squares Estimator (LTSE) was slightly affected by high perturbations of outliers and leverages. The study also showed that Modified Maximum likelihood Estimator (MME) and S Estimator (SE) were robust to all levels of outliers and leverage perturbations. Therefore leverages and outliers in datasets do affect the post hoc power analysis of the methods which cannot resist them.

Keywords: Ordinary least squares estimator; Huber maximum likelihood estimator; least trimmed squares estimator; S-estimator; modified maximum likelihood estimator; power of the test; leverages; outliers.

2010 Mathematics Subject Classification: 62G35, 62F35.

1 Introduction

Robust methods were introduced as alternatives to the OLSE when assumptions of classical OLSE are violated [1]. The OLSE has a breakdown point of $\frac{1}{n}$, thus being overly sensitive to outliers. Again, good and bad leverages also limit the performance of the OLSE. Thus, robust methods were introduced as modifications to the OLSE as they resist the influences of the outliers in a dataset [2].

2 Linear Regression Model

The matrix representation of the multiple linear regression model expressed as

$$y = X\beta + e$$

where y is an $n \times 1$ vector of observed response values, X is the $n \times p$ matrix of the predictor variables, β a $p \times 1$ vector contains the unknown parameters and has to be estimated, and e is the $n \times 1$ vector of random error terms. An estimate of β is $\hat{\beta}^T = (\hat{\beta}_1.....\hat{\beta}_P)$, which gives fitted values $\hat{y}_i = X_i^T\hat{\beta}$. The estimated residuals are computed as $e_i = y_i - \hat{y}_i$

where $i = 1.....n$ and n is the sample size. According to [3], 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.

3 The Ordinary Least Squares Estimator

The least squares estimator aims to minimize the sum of the square residuals as:
$\sum_{i=1}^{n} e_i^2 = (Y - X\beta)^T(Y - X\beta)$. Therefore using the OLSE to estimate the regression parameters in the model $Y = X\beta + e$, we have:

$$\hat{\beta} = (X^TX)^{-1}X^TY. \tag{3.1}$$

Hence, one can compute the least squares estimates directly from any dataset when X^TX is nonsingular. According to Adedia et. al. in [2] and Stuart in [3], the ordinary least squares estimator is very sensitive to the outliers in a dataset with a breakdown Point $(BDP) = \frac{1}{n}$, as a result, there are robust methods which can resist the effect of the outliers.

4 Robust Estimators

When the assumptions of the OLSE are not met, the OLSE cannot be used to estimate the regression parameters, [4]. As a result, some robust methods were introduced by Huber in [5], Rousseeuw in [6], Rousseeuw and Yohai in [7], and Yohai in [8] as cited by Maronna et. al. in [9].

4.1 Huber maximum likelihood estimator

The most common method of robust regression is M-estimation and is almost as efficient as OLSE, developed by [5]. According to Alma in [10], the HME is computed by minimizing the objective function,

$$\sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right). \tag{4.1}$$

The "s" is the scale estimate computed from a linear combination of the residuals, ($e_i = y_i - x^T\beta$) [10]. Moreover, the function ρ gives the contribution of the individual residuals to the objective function of the Huber m-estimator. According to Alma in [10], 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 $|e_i| \geq \left|e_i^{'}\right|$, and ρ is continuous.

The objective function of the least squares estimation is given by $\rho(e_i) = e_i^2$. The system of normal equations to solve this minimization problem is found by taking partial derivatives of 4.1 with respect to β and setting them equal to 0. So we minimize equations 4.1 with respect to each of the p parameters.

A weight function is defined as $w(u) = \frac{\psi(u)}{u}$, where $u = \frac{y_i - x_i^T}{s}$ and $\psi(u) = \frac{\partial\rho}{\partial u}$ is the score function, which results in $w_i = w\left(\frac{e_i}{s}\right)$ for $i = 1, 2, ..., n$ with $w_i = 1$ if $e_i = 0$. With $s \neq 0$, we define the weight matrix $W = diag(w_i : i = 1.....n)$. Solving for β results in the equation

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y. \tag{4.2}$$

The HME is 95% efficient under normal errors and it is robust to vertical outliers in the data, however, it is influenced by leverages and therefore has a breakdown point of 0 for leverages as the sample size approaches infinity.

4.2 Least trimmed squares estimator

The least trimmed squares estimator of [6], is computed by minimizing the h ordered squared residuals,

$$\hat{\beta} = min \sum_{i=1}^{h} (e_i^2) \tag{4.3}$$

where e_i are the residuals, n and h are the sample size and the trimming constant, respectively [11]. According to [12], the h trimmed dataset ensures that estimates have a high breakdown point of 50% but a low efficiency of 7.13%. In a study conducted by [13], they 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. When the data trimming is done well, this method is computationally equivalent to OLSE. 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 [10].

4.3 S-estimator

According to Alma in [10], the SE is a high breakdown estimator discovered by [7] which minimizes the standard deviation of the residuals. The S-estimator addresses the low breakdown point of the M-estimators. The high breakdown SE possesses a desirable property, that is it is affine, scale and regression equivariant, [12]. The SE minimizes the dispersion of the scaled residuals, that is, S-estimator is the $\hat{\beta}$ that makes $s(e_1, ..., e_n)$ minimal. The robust S-estimation minimizes a robust M-estimate of the residual scale

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{s}\right) = k. \tag{4.4}$$

Differentiating 4.4 we obtain the estimating equations for S-estimator, where ψ is replaced with an appropriate weight function. Although S-estimates have a BDP=0.5, it comes at the cost of a very low relative efficiency [14].

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 unfortunately 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.

The objective function ρ satisfies the following conditions, ρ is symmetric, continuously differentiable and $\rho(0) = 0$; there exists $a > 0$ such that ρ is strictly increasing on [0, a] and constant on $[a, \infty)$; $\frac{k}{\rho(a)} = \frac{1}{2}$. To obtain a breakdown point of 50%, the third condition is required even though it is not strictly necessary, [3]. The choice of k is done so that the resulting s is can estimate for σ when the errors are normally distributed. To do this, we set k such that $k = E_\phi(\rho(u))$, which is the expected value of the objective function if it is assumed that u has a standard normal distribution [13]. To use the Tukey bisquare objective function, [7] stated that if we set the tuning constant a = 1.547, the third condition is satisfied, and hence makes the S-estimator has 50% BDP.

4.4 Modified maximum likelihood estimator

The MME is introduced by [8], and it is a type of M-estimators which is both a high breakdown and efficient estimator [10]. It was the first estimator with a high breakdown point and high efficiency under normal error. An MME $\hat{\beta}$ is defined as a solution to

$$\sum_{i=1}^{n} x_{ij}\psi_1\left(\frac{y_i - x_i^T\beta}{s_n}\right) = 0 \tag{4.5}$$

where $j = 1, ..., p$. Yohai in [10] explained that the objective function ρ_1 associated with this score function, $\psi_1(u) = \frac{\partial\rho_1}{\partial u}$, must satisfy the following conditions,

ρ is symmetric and continuously differentiable, and $\rho(0) = 0$; there exists $a > 0$ such that ρ is strictly increasing on $[0, a]$ and constant on $[a, \infty)$; $\rho_1(u) \leq \rho_0(u)$. Lastly, the solution to 4.5 must satisfy this condition,

$$\sum_{i=1}^{n} x_{ij}\psi_1\left(\frac{y_i - x_i^T\hat{\beta}}{s_n}\right) \leq \sum_{i=1}^{n} x_{ij}\psi_1\left(\frac{y_i - x_i^T\tilde{\beta}}{s_n}\right). \tag{4.6}$$

Where $\tilde{\beta}$ is the initial estimate with 50% breakdown point. 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 [8]. The MME is very resistant to multiple leverage points and vertical outliers and also equivariant [13].

5 Results and Discussion

To demonstrate the performance of the robust methods above, we used dataset from Komfo Anokye Teaching Hospital (KATH). This secondary data was collected on the patients who patronize the KATH facility. The data was collected on the anthropometric measurements of the patients. The total body fat was regressed on the predictor variables such as BMI, arm fat as a percentage of the body, height and triceps skin fold. The results according to [15] are presented graphically using Stata.

5.1 Data with normally distributed residuals

All the procedures discussed were applied to the anthropometric data set with normally distributed residuals and the estimated coefficients, standard errors, coefficient of determination and power the test are presented in Figs. 1 and 2.

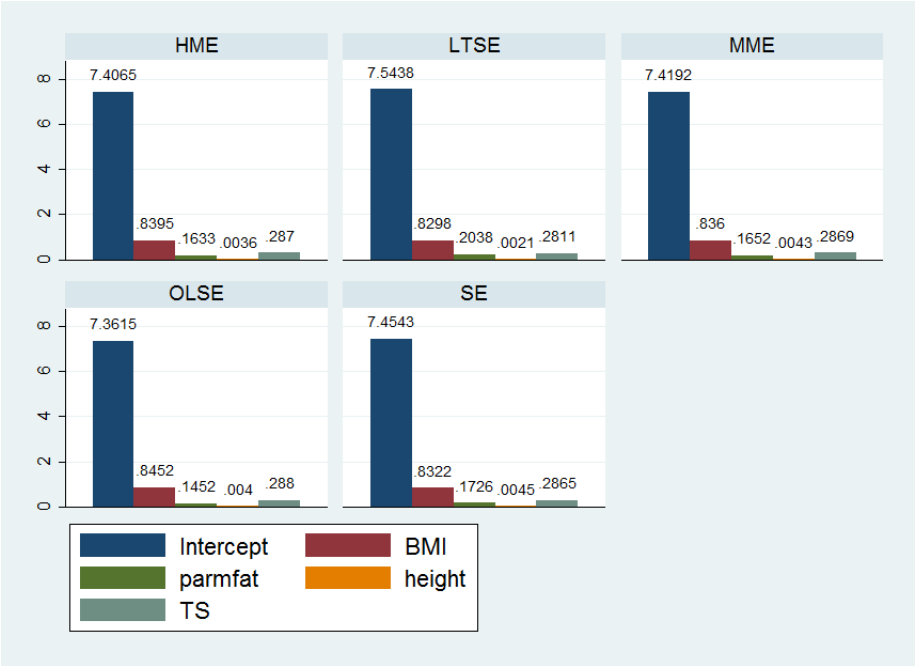

Fig. 1. The coefficients of the estimators for dataset with normal errors

Fig. 1 shows that all the estimators performed well when errors are normally distributed. The standard errors, relative efficiencies and the coefficients of determination from Fig. 2, also showed that robust methods performing as efficiently as the OLSE.

5.2 5% leverages in BMI, parmfat and 10% outliers

The results presented in the Figs. 3, 4 show the performances of the estimators when there are 5% leverages in BMI and parmfat, and 10% vertical outliers.

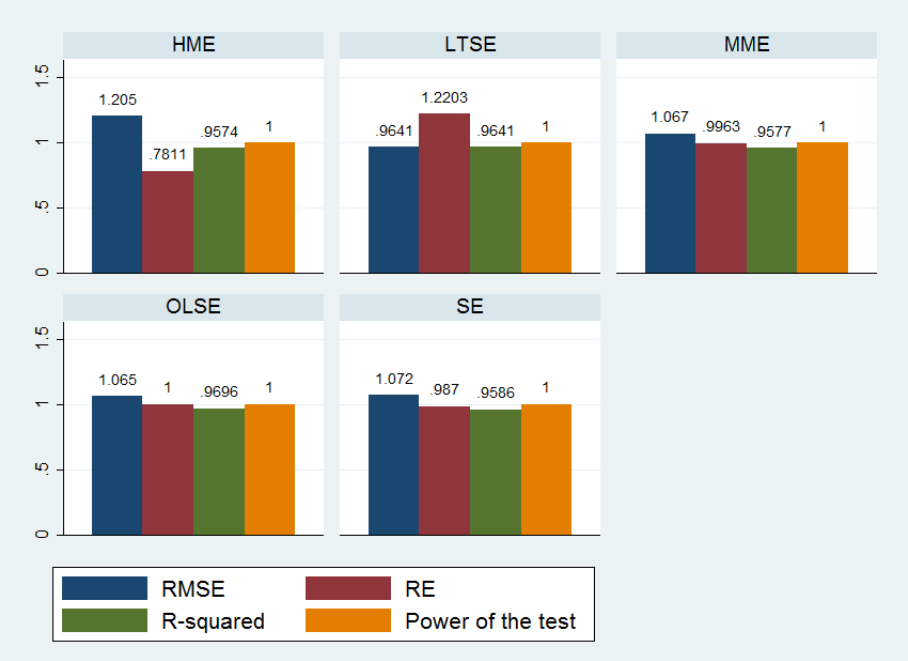

Fig. 2. Residual standard error, relative efficiency, coefficient of determination and the power of the test for original dataset with normal errors

Fig. 3. The coefficients of the estimators for 5% leverages in BMI, parmfat and 10% outliers

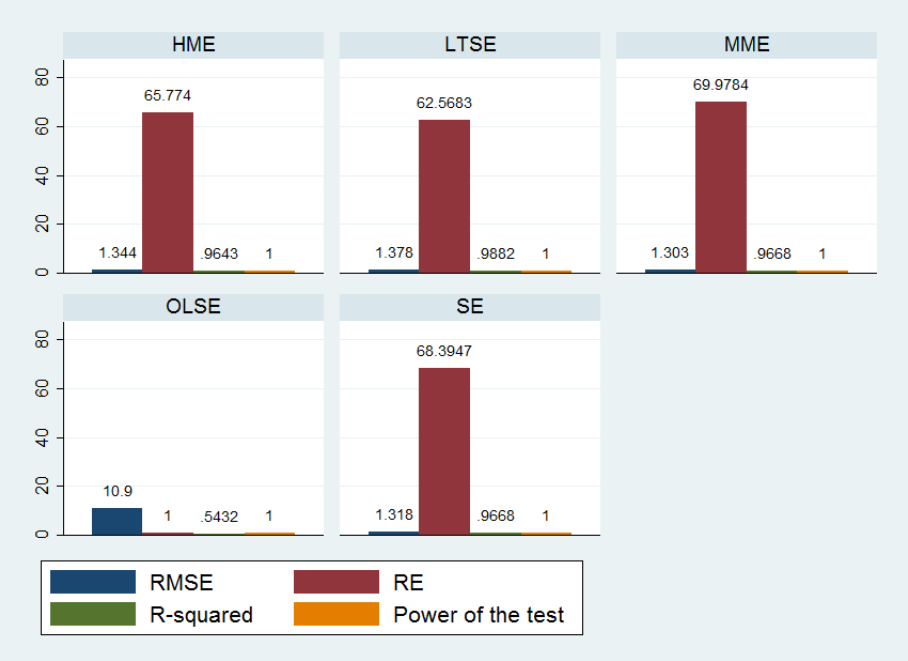

Fig. 4. Standard error, relative efficiency, coefficient of determination and power of the test for 5% leverages in BMI, parmfat and 10% outliers

The OLSE recorded higher standard error than the other estimators. Methods such SE and MME were able to bound the effects of both leverages and outliers. HME was robust to the influence of outliers but not the effects of leverages.

5.3 5% leverages in height, TS and 10% outliers

The various numerical measures computed for estimators when there are 5 % leverages in height and TS, and 10% outliers are presented in the Figs. 5 and 6.

MME and SE were resistant to both outliers and leverages. These robust methods try to get models that fit the majority of the data, whilst OLSE provides models that fit the average of the data. As a result, OLSE is always affected by few unusual observations. The coefficients of OLSE and HME were largely influenced by the aberrations in the data at this level, whilst LTSE was slightly affected. Comparing the model fit produced at this stage with when the model fit has normal residuals shows that the fitted model of the OLSE differed a lot from when the residuals were normal due to outliers and leverages.

5.4 5% leverages in BMI, parmfat, height and TS and 10% outliers

The Figs. 7 and 8 display how the estimators faired in the presence of 5% perturbations in all predictor variables and also 10% vertical outliers.

From Figs. 7 and 8, LTSE in addition to other robust methods performed well. LTSE does perform well when the trimming is done properly. The robust methods also performed well using; relative efficiencies, coefficients of determination, residual standard errors and power of the tests. The intercepts of all robust methods are similar and are close to that of the original data with

normal errors. However, the intercept for OLSE was largely affected and the residual standard error assumed a very large value.

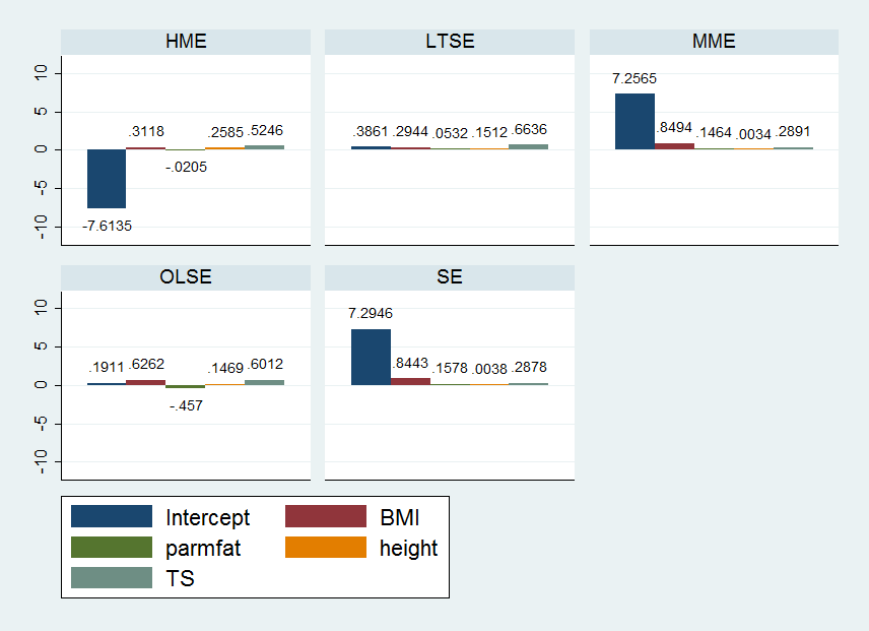

Fig. 5. The coefficients of the estimators for 5% leverages in height, TS and 10% outliers

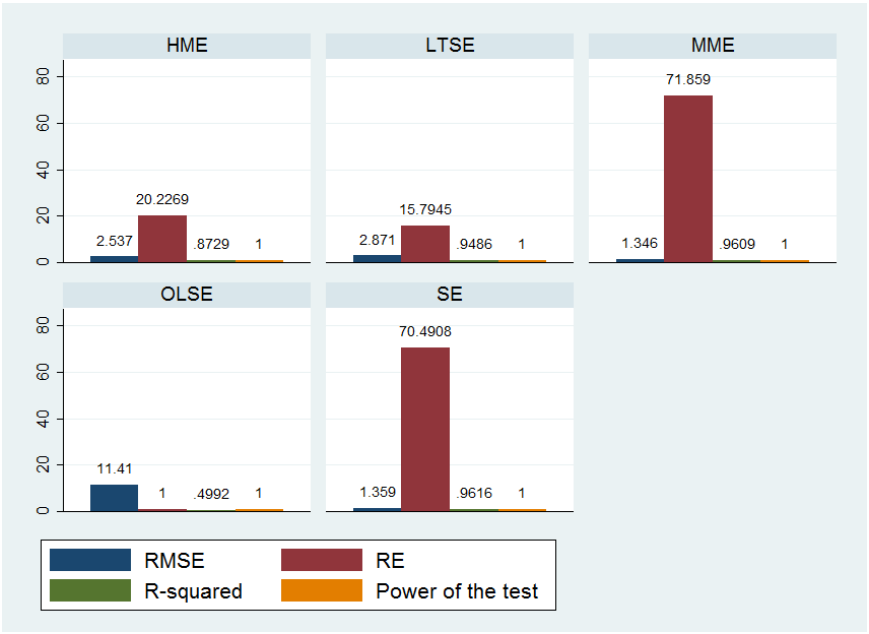

Fig. 6. Standard error, relative efficiency, coefficient of determination and power of the test for 5% leverages in height, TS and 10% outliers

Fig. 7. The coefficients of the estimators for 5% leverages in BMI, parmfat, height and TS and 10% outliers

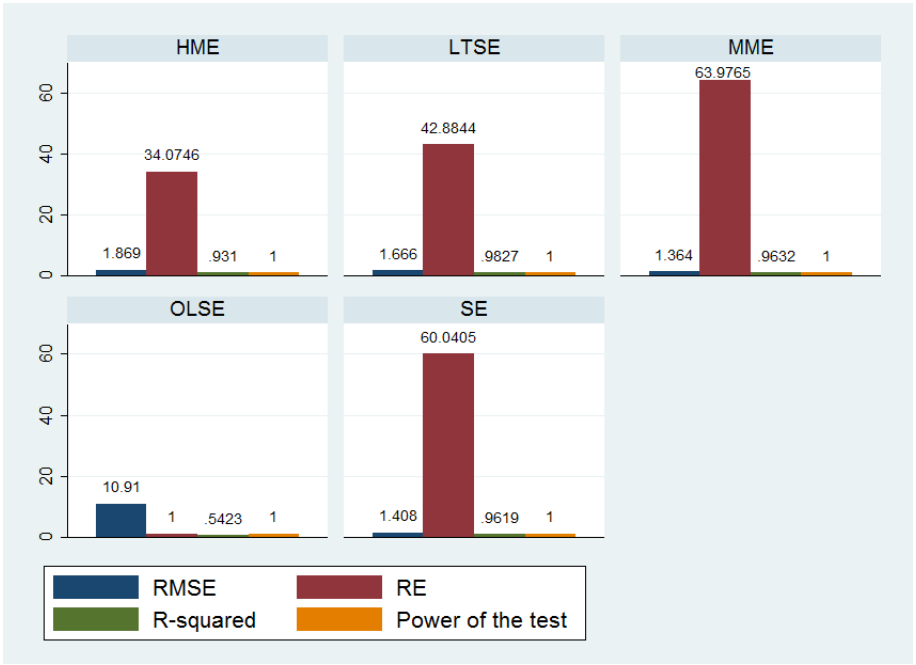

Fig. 8. Standard error, relative efficiency, coefficient of determination and power of the test for 5% leverages in BMI, parmfat, height and TS and 10% outliers

5.5 5% leverages in BMI and parmfat and 15% outliers

The numerical measures computed for the estimators with percentage of vertical outliers increased to 15% and 5% leverages in predictor variables; BMI and parmfat.

When the percentage of outliers increased, only OLSE is very much affected by using Coefficient of determination from Fig. 10. Moreover, using other criteria, the other estimators are also affected but not as compared to the OLSE. The coefficient of parmfat for LTSE, HME and SE were assumed negative values. Also, the coefficient of height for OLSE, LTSE and HME assumed different values due to the perturbations in the dataset. By comparing the estimators using coefficients of determination, relative efficiencies, standard error and power of the tests, LTSE performed better than the other estimators.

5.6 5% leverages in height and TS and 15% outliers

Figs. 11 and 12 show the results for the comparison of the estimators, when there are 5% leverages in height and TS and 15% outliers.

The observed R^2 for OLSE in Fig. 12 is quite low. The coefficients of parmfat for OLSE, LTSE and HME differed a lot from that observed for normal errors models. Also, vertical outliers in the data has influenced the standard error of the OLSE. MME and SE perform better than the other estimators.

5.7 5% leverages in BMI, parmfat, height and TS and 15% outliers

The Figs. 13 and 14 list the numerical measures (criteria) for comparing the ordinary least squares estimator to the robust methods, when all predictors contain 5% leverages and the response variable contains 15% vertical outliers.

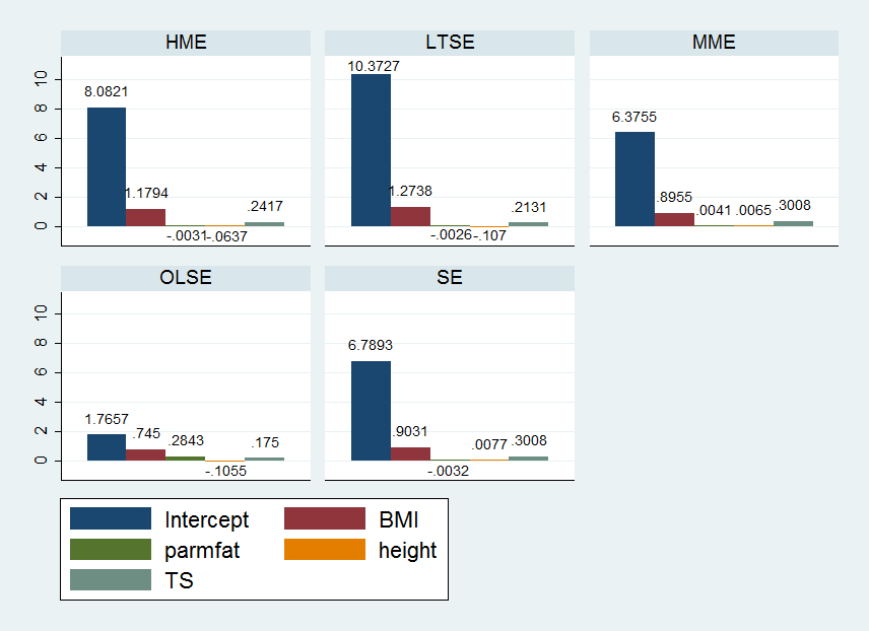

Fig. 9. The coefficients of the estimators for 5% leverages in BMI and parmfat and 15% outliers

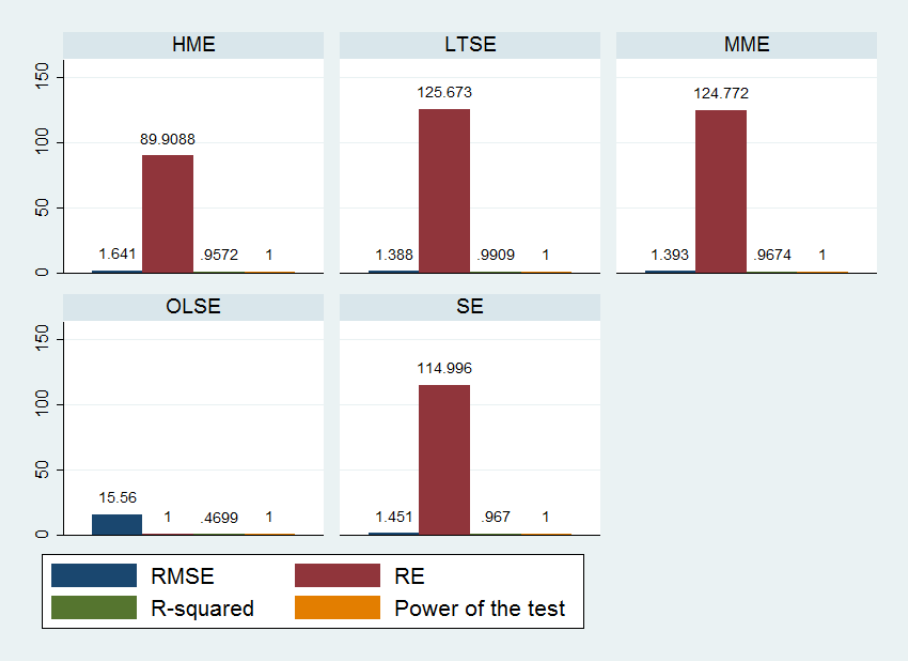

Fig. 10. Standard error, relative efficiency, coefficient of determination and power of the test 5% leverages in BMI and parmfat and 15% outliers

Fig. 11. The coefficients of the estimators for leverages in height and TS and 15% outliers

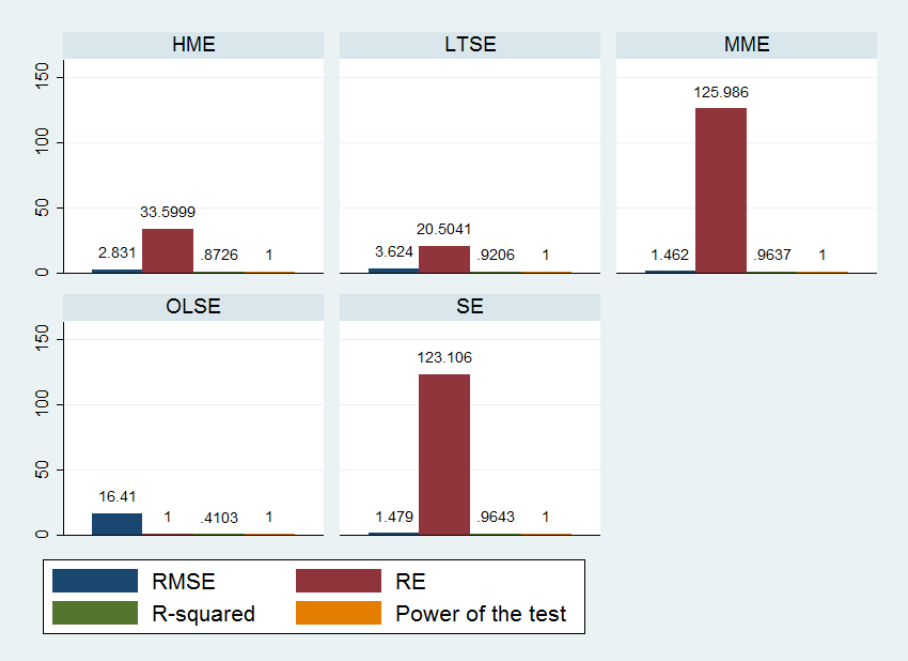

Fig. 12. Standard error, relative efficiency, coefficient of determination and power of the test for leverages in height and TS and 15% outliers

Fig. 13. The coefficients of the estimators for 5% leverages in BMI, parmfat, height and TS and 15% outliers

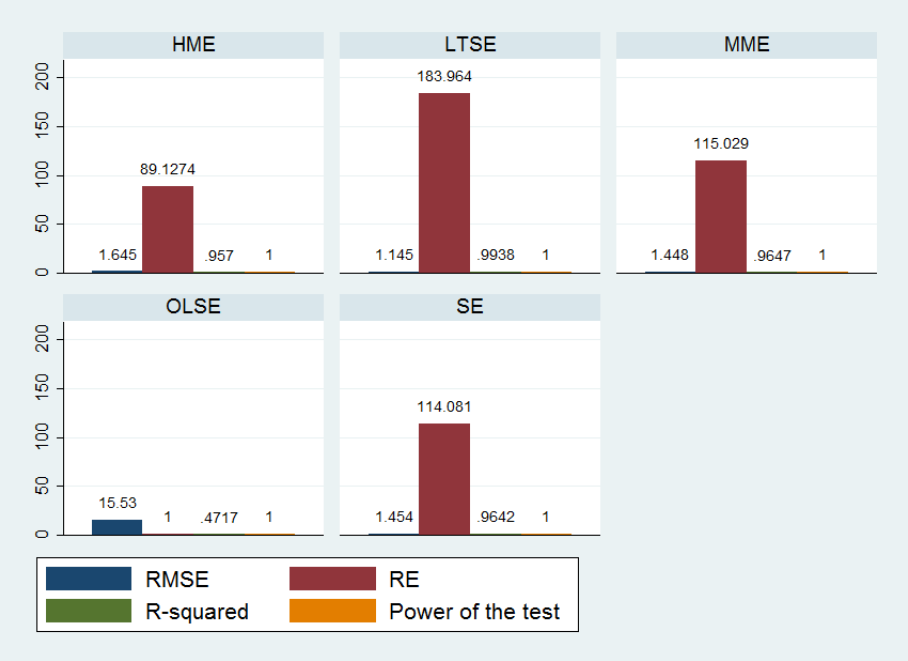

Fig. 14. Standard error, relative efficiency, coefficient of determination and power of the test for 5% leverages in BMI, parmfat, height and TS; and 15% outliers

The Figs. 13 and 14 show that the LTSE performed better than the other estimators when all independent variables are perturbed with 5% leverages and 15% outliers in the response variable. Using the relative efficiency and the standard errors, we observed that the OLSE did not perform as compared to the other estimators.

6 Concluding Remarks

In this study, we applied some robust regression methods and the ordinary least squares method on a dataset from KATH with different levels of perturbations (outliers and leverage points). The results showed that HME breaks down when there are leverage points in the data. This is because, HME was resistant to only vertical outliers; the effects of leverages and vertical outliers on OLSE increases as the percentage of leverages and outliers perturbations increase. The LTSE performed well against leverage points and vertical outliers in this study. Moreover, the SE also performed well and was able to resist the effects of outliers and leverage points. Though the SE is known to be less efficient, it is a robust method that is capable of counteracting the influence of the influential observations in various levels of perturbations in the dataset above. Finally, the most efficient and the robust method, MME, was able to display its performance in this study.

Acknowledgement

We express our gratitude to our families and friends for their support.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Muthukrishnan R, Radha M. Comparison of robust estimators in regression. 2014;1-12. Available: http://interstat.statjournals.net/YEAR/2010/articles/1005003.pdf

[2] Adedia D, Adebanji A, Labeodan M, Adeyemi S. Ordinary least squares and robust estimators in linear regression: Impacts of outliers, error and response contaminations. British Journal of Mathematics & Computer Science. 2015;13(4):1-11. DOI: 10.9734/BJMCS/2016/22876

[3] Stuart C. Robust Regression. Department of Mathematical Sciences. Durham University. 2011;169.

[4] Gschwandtner M, Filzmoser P. Computing robust regression estimators: Developments since dutter (1977). Austrian Journal of Statistics. 2012;41:45-58.

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

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

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

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

[9] Maronna RA, Martin RD, Yohai VJ. Robust Statistics, Theory and Methods. John Wiley and Sons Ltd; 2006.

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

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

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

[13] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. JOHN WILEY and SONS; 1987.

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

[15] 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/13653