IJHRIM

Intl J Health Recs & Info Mgt. Jan – Dec 2023;6(1):18-21

Research article

Print ISSN 2645-2464; E ISSN 2695-1770

Performance of the ordinary least squares estimator method of estimating regression parameters and some robust regression methods

Abayomi Ope-Oluwa Kehinde[1], Kazeem Kehinde Adesanya[1], Moriam Adeyinka Onafowokan[1]

[1]Department of Health Information Management, Ogun State Polytechnic of Health and Allied Sciences, Ilese-Ijebu, Nigeria

Corresponding author*: E-mail: *abayomikehinde1982@gmail.com*

ABSTRACT

Background/Objectives: Ordinary least squares (OLS) estimation of regression parameters is a popular technique. It is susceptible to outliers and high-leverage spots in the data, though. A set of methods known as robust regression methods are less susceptible to the impact of outliers and high-leverage points. This study evaluated the performance of the Ordinary Least Squares Estimator (OLSE) method of estimating regression parameters and some robust regression methods. The Least-Trimmed Squares Estimator (LTSE), Huber Maximum likelihood Estimator (HME), S-Estimator (SE) and Modified Maximum likelihood Estimator (MME) were considered in this study. ***Design/Methods:*** Criteria for the comparison were: coefficients and their standard errors, relative efficiencies, Root Mean Square Errors, coefficients of determination and the power of the test. The sensitivity of these robust methods were considered using Anthropometric data from Olabisi Onabanjo University Teaching Hospital in Sagamu, Ogun state. The dataset was on Total Body fat and Body Mass Index, Triceps skin-fold, Arm Fat as percent composition of the body and Height as predictors. Leverages were introduced first into two variables, and into all predictors. ***Results/Conclusion:*** Results showed that robust methods are as efficient as the OLSE if the assumptions of OLSE are met.

Keywords: Ordinary Least Squares (OLS); Robust regression; Least-trimmed squares (LTS); Huber maximum likelihood estimation (HME); S-estimation (SE)

Edited by *IT Adeleke; submitted on 30.08.2023; peer reviewed by AA Adebisi, TC Anamah; U Isah accepted 12.09.2023; published 21.09.2023.*

Please cite as: *Kehinde AO, Adesanya KK, Onafowokan MO. Performance of the ordinary least squares estimator method of estimating regression parameters and some robust regression methods Int J Health Recs & Info Mgt. 2023;6(1):18-21.*

Conflict of interest: None declared.

Funding disclosure: No funding was solicited for nor obtained for this study

INTRODUCTION

Since ancient times, linear regression analysis has been used as a statistical tool to determine a linear relationship between variables. Additionally, it is applicable to every subject of study, including engineering, physical science, social science, health science, and a wide range of others. For instance, banks utilize regression analysis to calculate their earnings, so they are aware of the factors that favorably and negatively impact their profitability. By taking into account elements that can influence or cause an increase or decrease in the quantity of fat in the body, healthcare professionals can also utilize regression to estimate the total fat in the bodies of their patients.

Hospital statisticians use regression analysis to examine the lifestyles that might be contributing factors in the development of specific diseases, such as high blood pressure. Regression analysis is a crucial statistical tool as a result. In this paper, we examine alternative techniques for estimating regression parameters that are distributionally robust to small deviations. Robust methods of regression analysis have reportedly been studied since the nineteenth century, according to Maronna[1]. They continued by stating that while the eighteenth century saw the realization of much robust estimating knowledge, the first significant advancements were made in the 1960s and early 1970s. Thanks to the essential work of John Tukey (1960, 1962), Peter Huber (1964, 1967), and Frank Hampel (1971, 1974). Huber (1981), Hampel, Ronchetti, Rousseeuw and Stahel (2016), and Rousseeuw and Leroy (2017) have all published several studies and significant works.

If the assumptions of the regression model, variables, and error terms are met, the application of the ordinary least squares approach in regression analysis works well. The ordinary least squares method of estimation however becomes problematic in the presence of outliers, leverage points, or failure of the assumptions. This is due to the fact that both good and bad leverage points, as well as vertical outliers, can have an impact on the model's residuals, coefficients, and standard errors.

Bhar compared the Huber M-estimator to the ordinary least square's estimator, comparing robust regression methods like M-estimator, W-estimators, R-estimators, Least median of squares estimator, Least trimmed of squares estimator, and Re-weighted least squares estimator[2]. M-estimation is the most efficient method, minimizing standardized residuals and giving smaller weights to unusual observations. W-estimators show the importance of each observation, while R-estimators compute data ranks. L-estimators compute linear combinations of order statistics, including least trimmed squares and least median squares. The Huber M-estimator outperforms the ordinary least squares estimator in both standard error and coefficient of determination.

Fox and Weisberg studied robust regression methods on Duncan's occupational-prestige data, revealing that least squares estimator broke down due to vertical outliers[3]. Robust methods however effectively bound the influence of unusual observations, making them powerful statistical tools for identifying unusual observations. Least trimmed squares performed better than ordinary least squares estimator when no outliers were removed.

Al-Noor and Mohammad conducted a simulation study to compare Ordinary Least Squares Method, Least Absolute Deviations Method, M- Estimators, Trimmed Least Squares estimators, and Non-parametric Regression[4]. They found that the Ordinary least squares method performed better without contamination. When outliers were however introduced in the dependent and independent variables, the method broke down. Non-parametric methods performed better when outliers were present in both X-dimension and Y-dimension. The study recommends future studies to consider the presence of outliers and the underlying assumptions before selecting an estimator.

Schumacker *et al*. compared ordinary least squares and robust regression using the S-PLUS statistical package[5]. They found that the MM-estimator performed better than LTS and OLS estimators, indicating the best method. Comparing robust estimators is however challenging due to the presence of few statistical packages with robust methods.

Yohai developed the MM-estimator, the most efficient with a high breakdown point. It uses S-estimator as initial estimates, achieving high breakdown point properties. The robust estimator was compared to ordinary least squares, showing no influence from outliers[6].

Cetin and Toka compared robust estimation methods to Ordinary least squares (OLSE) using a datasets with weak multi-collinearity[7]. They found that OLSE is inefficient when outliers are introduced, with S-estimator performing better. Simulation studies showed OLSE inefficient when contaminated with outliers[7].

Rousseeuw *et al*. found that robust regression methods face challenges due to factors like large sample data clustering and outliers[8]. The OLSE method is not efficient for large datasets, and some people struggle to interpret results. Rousseeuw *et al*. emphasized the advantages of using the LTSE as a robust method, stating that deleting outlying points can significantly improve regression results and model fit[8].

Ruppert *et al*. compared Iteratively Reweighted Least Squares (OLSE) with other estimators, finding Huber has leverage points issues but OLSE performed poorly overall[9].

Verardi and Croux (2009) define vertical outliers as observations with outlying y-dimension values but not in predictor variables, impacting ordinary least squares estimation[10].

As a result, this study aimed at assessing the effectiveness of a few resilient strategies that mitigate the effects of a dataset's flaws.

The study compares robust regression estimators against the conventional least squares estimator in order to determine how robust they are to datasets aberrations. To determine how consistently the different estimators can withstand the limitations of the datasets, they will be compared for diverse datasets. Olabisi Onabanjo

University Teaching Hospital in Sagamu, Ogun State provided secondary information on the following variables: height, triceps skin-fold, body fat percentage, body mass index, and arm fat percentage. We made use of the R statistical software to contaminate and analyze the data.

Table 1: The coefficients (standard errors) of the estimators with normal errors

Methods	Intercept	BMI	parmfat	height	TS
OLSE	7.3615(0.9448)	0.8452(0.0555)	0.1452(0.0824)	0.0040(0.0104)	0.2880(0.0184)
LTSE	7.5438(0.8715)	0.8298(0.0514)	0.2038(0.0763)	0.0021(0.0095)	0.2811(0.0171)
HME	7.4065(1.0123)	0.8395(0.0594)	0.1633(0.0883)	0.0036(0.0111)	0.2870(0.0198)
SE	7.4543(1.0070)	0.8322(0.0591)	0.1726(0.0878)	0.0045(0.0110)	0.2865(0.0197)
MME	7.4192(1.0084)	0.8360(0.0592)	0.1652(0.0879)	0.0043(0.0111)	0.2869(0.0197)

METHODS

This study examined the impact of outliers, leverage points, non-normality, and contamination on classical least squares estimation in linear regression analysis. Robust methods like MM-estimator, Huber M-estimator, least trimmed squares estimator, and S-estimator were compared to ordinary least squares estimator. Secondary data from Olabisi Onabanjo University Teaching Hospital was collected, analyzed using R statistical package and information from the internet and related articles.

The study compared robust regression estimators like LSE, HME, SE, and MME with OLSE. It analyzes Total Body Fat and four independent variables: BMI, Triceps skin-fold, arm fat, and height.

RESULTS

The results are presented in tables, with coefficients and standard errors reported as one set and residual standard errors, efficiencies and test power as another set.

Original datasets with normal Errors

The estimated model parameters and standard errors when the errors are normally distributed are shown on Table 1. Table 2 shows the normally distributed residuals' residual standard errors, relative efficiency, coefficients of determination, and test power

As can be seen from Table 1, all estimators work well because the errors are normally distributed. This supported the adage that all estimators function well with typical errors. When the errors are normal, all the estimators perform well, according to Table 2's RMSE, relative efficiency, and coefficients of determination.

CONCLUSION

The study demonstrates that robust methods are efficient as OLSE if basic assumptions are met. Small deviations from normality do not significantly impair these methods and MME and SE do not breakdown completely. Turbulence with leverages however causes the Ordinary least squares estimator and Huber Maximum likelihood estimator to breakdown, while the S-estimator and Modified Maximum Likelihood Estimator perform well.

REFERENCES

1. Maronna RA, Martin RD, Yohai VJ. Robust Statistics, Theory and Methods. John Wiley and Sons Ltd, 2006.
2. Bhar, L. Robust regression. http://www.iasri.res.in/ebook/EBADAT/3- Diagnostics, 2014.
3. Fox J, Weisberg S. An appendix to an r companion to applied regression second edition. pp1–17.
4. Al-Noor, H. N. and MoMohammad, A. (2013). Model of robust regression with parametric and non-parametric methods. *Mathematical Theory and Modeling*, 3:27–39.
5. Schumacker RE, Monahan MP, Mount RE. A comparison of OLS and Robust regression using S-Plus. 2002;28(2):10–13.
6. Yohai VJ. High breakdown-point and high-efficiency robust estimates for regression. *The Annals of Statistics.* 1987;15:642–656.
7. Cetin M, Toka O. The comparing of s-estimator and m-estimators in linear regression. *Gazi University Journal of Science.* 2011;24(4):747–752.
8. Rousseeuw PJ, Zaman A, Orhan M. Econometric applications of high-breakdown robust regression techniques. *Economic Letters*. 2001;71:1–8.
9. Ruppert D, Street JO, Carroll RJ. A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician.* 1988;42:152–154.
10. Verardi V, Croux C. Robust regression in Stata. *The Stata Journal.* 2009;3:439–453.

.

Authors Contribution:

KAO conceived of the study, initiated the design, participated in literature search, data collection, analysis and coordination. AKK and OMA participated in the design, literature search, technical process, data analysis and coordination and reviewed the final manuscript.

Table 2: The Root Mean Square Error (RMSE), Relative Efficiency, Coefficient of Determination and the Power of the test for original datasets with normal errors

Method of estimation	RMSE	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