Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

The performance comparison of two-step robust weighted least squares (TSRWLS) with different robust's weight functions

Zulkifli Mohd Ghazali [1,*], Muhammad Syawal Abd Halim [1], Jaida Najihah Jamidin [2]

[1]*Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, Tapah Campus, Perak, Malaysia*
[2]*Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, Seremban 3 Campus, Negeri Sembilan, Malaysia*

ARTICLE INFO

Article history:
Received 20 January 2017
Received in revised form
9 March 2017
Accepted 8 April 2017

Keywords:
Heteroscedasticity
Outlier
Two-step robust weighted least squares
Robust's weight function

ABSTRACT

The purpose of this paper is to compare the performance of Two-Step Robust Weighted Least Squares (TSRWLS) using three different Robust's Weight Function namely Huber, Bisquare and Hampel. Previously, the procedure of TSRWLS only used Huber's weight function as the second weight and this study serves to compare the performance of TSRWLS when the three different weight functions are used. The performance was evaluated based on real data and Monte-Carlo simulation study and the findings suggests that the performance of TSRWLS by using Huber, Bisquare and Hampel as the second weight is relatively close to one another with a fairly close standard error and almost identical values of biasness and root mean square error. Based on the result in the numerical example and simulation study, this study concluded that the performances of TSRWLS using all three weight functions performed equally. It is therefore suggested that any one of the three robust's weight function can be used as the second weight in performing TSRWLS. However, the use of Huber's weight function as the second weight in TSRWLS is recommended because of the simplicity of the function when compared against the other two weight functions.

1. Introduction

In regression analysis, the assumptions and outliers must be considered in order to ensure that the result or estimated regression model is correct. Violated assumptions and presence of outliers will lead to the estimated regression model to be imprecise. One of the violated assumptions that are commonly faced by the researcher in conducting linear regression analysis is heteroscedastics error. The heteroscedastic error and outlier are two problems that will affect the performance of Ordinary Least Square (OLS) in estimating the regression linear model. As an alternative, the Two-Step Robust Weighted Least Squares (TSRWLS) method was proposed to remedy this problem. It has been proved that this method is not affected by heteroscedastic error and outlier simultaneously (Habshah et al., 2013). Previously, however, the procedure of TSRWLS only used Huber's weight function as the second weight to perform this method. In robust statistics, there are three robust's weight functions that are widely used such as Huber, Bisquare and Hampel (Bellio and Ventura, 2005). Therefore, this study was performed to investigate the performance of TSRWLS using three different robust's weight functions (Huber, Bisquare and Hampel). The performance will be evaluated based on the error measures such as the standard error, biasness and the root mean square error.

The heteroscedasticity refers to the situation when the variance of the error terms is not constant. It has been proved that when the homoscedasticity assumption is violated, the OLS is no longer at its optimum. The OLS estimator remains unbiased, but becomes inefficient, leading to the estimates of the standard errors to be inconsistent. The statistical hypothesis tests such as the t-test, F-test, and Wald-test are then rendered invalid (Schmidheiny, 2012). Therefore, the weighted least square (WLS) based on the variance function was proposed as an alternative (Kutner et al., 2008). By using this method, the estimated parameters in linear regression model will be unbiased and efficient (Sosa-Escudero, 2009). However, due to the presence of both heteroscedasticity and outliers in the data, the WLS is no longer appropriate because the WLS estimators are affected by the outlier (Habshah et al., 2009).

* Corresponding Author.
Email Address: zulki656@perak.uitm.edu.my (Z. M. Ghazali)
https://doi.org/10.21833/ijaas.2017.05.008
2313-626X/

An outlier refers to a value that is extremely large or small compared to the other observations. Outliers can create great difficulty and in least square method for example, a fitted line may be pulled disproportionately toward an outlying observation because the sum of the squared deviations is minimized (Kutner et al., 2008). This could cause a misleading regression model. Therefore, Robust Weighted Least Square (RWLS) was put forward to remedy the effect of outliers and heteroscedastic errors simultaneously (Habshah et al., 2009). However, RWLS method can only be used for single linear regression.

Because of the limitation, another method which is called as the Two-Step Robust Weighted Least Squares (TSRWLS) was proposed (Habshah et al., 2013). This method can be used to estimate the multiple linear regression models. Besides, TSRWLS is not affected by heteroscedasticity and outliers compared to OLS and WLS methods.

2. Methodology

The procedure of the Two-Step Robust Weighted Least Square (TSRWLS) is initiated by computing the regression function based on LTS estimator and obtaining the fitted values. The next step is obtaining the residual $e_i = y_i - \hat{y}$ and regressing the absolute residual on the fitted values. From the standard deviation function, the fitted values of S_i is obtained. The estimated first weighted (w_1) is then acquired through the inverse of squared standard deviation function as in Eq. 1.

$$w_i = \frac{1}{S^2} \tag{1}$$

The second weight (w_2) from the robust's weight function can now be attained. The three robust's weight functions which are Huber, Bisquare and Hampel can be referred to in Table 1. The final weighted W is now computed as in Eq. 2.

$$W = w_1 w_2 \tag{2}$$

Next, the estimate parameters are computed as depicted in Eq. 3.

$$\hat{\beta} = (X'WX)^{-1}X'WY \tag{3}$$

To evaluate the performance of TSRWLS with three different robust's weight functions, the analysis section has been divided into two parts which are the numerical example and the simulation study. The performance will also be tested in three different conditions of data, which include heteroscedastic error, heteroscedastic error with a single outlier and heteroscedastic error with several outliers.

In the numerical example, the data is taken from Chatterjee and Price (1977). The dataset have 50 observations where education expenditure is the response variable with three independent variables comprising of income, resident under 18 and resident in urban area. The performance of this method will be evaluated based on the value of standard error.

Table 1: Robust's weight function

Robust's Weighted Function	
Huber	$w_2 = \begin{cases} 1 & if \quad \lvert e_i \rvert < 1.345 \\ \frac{1.345}{\lvert e_i \rvert} & if \quad \lvert e_i \rvert \geq 1.345 \end{cases}$
Bisquare	$w_2 = \begin{cases} \left(1 - \left(\frac{e_i}{4.685}\right)^2\right)^2 & if \quad \lvert e_i \rvert \leq 4.685 \\ 0 & if \quad \lvert e_i \rvert > 4 \end{cases}$
Hampel	$w_2 = \begin{cases} 1 & if \quad \lvert e_i \rvert < a \\ \frac{a}{\lvert e_i \rvert} & if \quad a \leq \lvert e_i \rvert < b \\ a\left(\frac{c/\lvert e_i - 1 \rvert}{c-1}\right) & if \quad a \leq \lvert e_i \rvert < c \\ 0 & if\ Otherwise \end{cases}$

In the simulation part, the Monte-Carlo simulation will be performed. The regression model from Chatterjee and Price (1977) will be adopted in Eq. 4;

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_i \varepsilon_i, \quad i = 1,2,3, \dots, n \tag{4}$$

where, $\varepsilon_i \sim N(0,1)$, $x_1 \sim U(0,1)$ and $x_2 \sim N(0,1)$.

To generate a heteroscedastic regression model, the regression parameter of $\beta_0 = \beta_1 = \beta_2 = 1$ and $\sigma_i^2 = \sigma^2\, exp(ax_{1i} + ax_{2i})$ with $\sigma^2 = 1$ and a is an arbitrary constant will be used.

In this simulation study, $a = 0.4$ will be employed. To generate a certain percentage of the outlier, $\varepsilon_i \sim N(0,1) + Cauchy(0,10)$ will be included.

The percentage of outlier may vary. Based on this regression model, data in two sample sizes will be generated. 30 and 100 observations respectively for 1000 trials will be obtained to get the summary statistics such as bias, the mean square error and the root mean square error in order to evaluate the overall performance. The summary statistics is summarized in Table 2.

This simulation study will be carried out using R-programming language.

3. Results and discussion

The performance of TSRWLS using three different robust's weight functions will be discussed based on the numerical example and simulation study.

3.1. Numerical example

The data taken from Chatterjee and Price (1977) has a heteroscedastic error with a single outlier. The heteroscedastic error was examined by using the residual plot, while the outlier was identified by using LTS method. In this part, the performance of TSRWLS using three different robust's weight functions was examined in two different conditions of data which contain heteroscedastic error and heteroscedastic error with a single outlier.

Table 2: Summary statistics

Summary Statistics	
Bias	$bias(\hat{\beta}_j) = \bar{\beta}_j - \beta_j$, where, $\bar{\beta}_j = \frac{1}{m}\sum_{k=1}^{m}\beta_j^{(k)}$
Mean Square Error	$MSE(\hat{\beta}_j) = (\hat{\beta}_j - \beta_j)^2 + \frac{1}{m}\sum_{k=1}^{m}(\beta_j^{(k)} - \bar{\beta}_j)^2$
Root Mean Square Error	$[MSE(\hat{\beta}_j)]^{1/2}$

3.1.1. Heteroscedastic error

To test the performance of TSRWLS using three different robust's weight functions in heteroscedastic error conditions, one observation which is observation 49 from the data was excluded. This is due to the observation being detected as an outlier by using the LTS method.

Table 3 shows the estimated coefficients and standard error for data with heteroscedastic error. Based on the result, the performance of TSRWLS using three different robust's weight functions are not too different since the values of the respective standard errors are fairly close to one another. The estimated coefficient values for β_1, β_2 and β_3 were also close to one another for each robust weight function.

Table 3: Estimated coefficients and standard error values for data with heteroscedastic error

Coefficient	Robust weight Function	Statistical Analysis	
		Estimated Value	Standard Error
β_0	Huber	-201.30	56.68
	Bisquare	-199.20	26.53
	Hampel	-198.90	32.37
β_1	Huber	0.0417	0.0057
	Bisquare	0.0413	0.0032
	Hampel	0.0414	0.0037
β_2	Huber	0.7421	0.1584
	Bisquare	0.7580	0.0775
	Hampel	0.7592	0.0935
β_3	Huber	0.0596	0.0269
	Bisquare	0.0509	0.0218
	Hampel	0.0487	0.0233

3.1.2. Heteroscedastic error with single outlier

In this section, the performance of TSRWLS using three robust's weight functions when the data has heteroscedastic error with a single outlier was examined. The result in Table 4 suggested that the performance of TSRWLS using three robust's weight functions performed equally since the estimated coefficient and standard error are relatively close. The Monte-Carlo simulation study was then performed to support this finding.

3.2. Simulation study

The Monte-Carlo simulation was employed to illustrate the performance of TSRWLS using three different robust's weight functions. The performance was measured by using the biasness measure and the root mean square error. The performance was also examined with two different sample sizes which are 30 and 100 observations respectively.

Table 4: Estimated coefficient and standard error values for data with heteroscedastic error and outlier

Coefficient	Robust weight Function	Statistical Analysis	
		Estimated Value	Standard Error
β_0	Huber	-214.30	57.74
	Bisquare	-199.20	26.53
	Hampel	-198.90	32.37
β_1	Huber	0.0428	0.0058
	Bisquare	0.0413	0.0032
	Hampel	0.0414	0.0037
β_2	Huber	0.7749	0.1616
	Bisquare	0.7580	0.0775
	Hampel	0.7592	0.0935
β_3	Huber	0.0552	0.0275
	Bisquare	0.0509	0.0218
	Hampel	0.0487	0.0233

Based on Table 5, the estimated coefficients for all robust's weight functions are fairly close to the actual value which is equal to one. These estimated values are consistent even when the percentage of outlier went up to 40% in the data. This result suggests that the performance of TSRWLS using three robust's weight functions are relatively close to one another.

Table 5: Estimated coefficients of tsrwls with three robust's weight functions

Sample size n =30				
Outliers	Robust weight Function	Coefficients		
		β_0	β_1	β_2
0%	Huber	1.0167	0.9626	0.9955
	Bisquare	1.0208	0.9553	0.9925
	Hampel	1.0145	0.9642	0.9947
10%	Huber	0.9877	0.9801	1.0583
	Bisquare	0.9755	1.0270	1.0141
	Hampel	0.9749	1.0257	1.0106
20%	Huber	1.0124	0.9051	1.0953
	Bisquare	1.0744	0.8360	1.0878
	Hampel	1.0719	0.8423	1.0891
30%	Huber	1.1178	0.7172	1.0757
	Bisquare	1.1239	0.7099	1.0910
	Hampel	1.1282	0.7108	1.0951
40%	Huber	1.1613	0.6145	1.0243
	Bisquare	1.0319	0.8570	1.0759
	Hampel	1.0291	0.8598	1.0806
Sample Size n = 100				
0%	Huber	1.0001	1.0015	0.9955
	Bisquare	1.0029	0.9978	0.9983
	Hampel	1.0029	0.9975	0.9961
10%	Huber	0.9958	1.0406	0.9877
	Bisquare	0.9984	1.0421	1.0052
	Hampel	1.0003	1.0361	1.006
20%	Huber	1.0100	1.0145	1.0043
	Bisquare	1.0017	1.0433	1.0183
	Hampel	0.9971	1.0472	1.0181
30%	Huber	1.0256	1.0276	1.0429
	Bisquare	1.0088	1.0781	1.0537
	Hampel	1.0098	1.0772	1.0560
40%	Huber	1.0482	1.0142	1.0328
	Bisquare	1.0540	1.0689	1.0695
	Hampel	1.0556	1.0636	1.0731

The result in Table 6 and Table 7 shows the value of biasness measure and the root mean square error in two different sample sizes. Based on the result, the values of bias and the root mean square error are not too different between robust's weight functions which are Huber, Bisquare and Hampel for both sample sizes. This indicates that the performance of TSRWLS with three robust's weight functions performed equally.

4. Conclusion

In the numerical example, the performance of TSRWLS using three different robust's weight functions are fairly close to one another since the value of the standard error for each estimated coefficient are not too different. This result has also been supported by the Monte-Carlo simulation study. In the simulation study, the values of the estimated coefficients by using Huber, Bisquare and Hampel as the second weight are fairly close to the actual value which is equal to one. The value of biasness measure and the root mean square error were also relatively close to one another.

Table 6: Biasness measure of parameters for three Robust's Weight Functions

Sample size n =30				
Outliers	Robust weight Function	Bias		
		β_0	β_1	β_2
0%	Huber	0.0167	-0.0374	-0.0045
	Bisquare	0.0208	-0.0447	-0.0075
	Hampel	0.0144	-0.0359	-0.0053
10%	Huber	-0.0123	-0.0199	0.0583
	Bisquare	-0.0245	0.0270	0.0141
	Hampel	-0.0251	0.0257	0.0106
20%	Huber	0.0124	-0.0949	0.0953
	Bisquare	0.0744	-0.1640	0.0878
	Hampel	0.0072	-0.1577	0.0089
30%	Huber	0.1178	-0.2828	0.0757
	Bisquare	0.1239	-0.2901	0.0910
	Hampel	0.1282	-0.2892	0.0951
40%	Huber	0.1613	-0.3855	0.0243
	Bisquare	0.0319	-0.1430	0.0759
	Hampel	0.0291	-0.1402	0.0806
Sample Size n = 100				
0%	Huber	0.0009	0.0015	-0.0045
	Bisquare	0.0029	-0.0022	-0.0017
	Hampel	0.0029	-0.0025	-0.0039
10%	Huber	-0.0042	0.0406	-0.0123
	Bisquare	-0.0016	0.0421	0.0053
	Hampel	0.0004	0.0360	0.0057
20%	Huber	0.0100	0.0145	0.0043
	Bisquare	0.0017	0.0433	0.0182
	Hampel	-0.0029	0.0472	0.0181
30%	Huber	0.0256	0.0276	0.0429
	Bisquare	0.0088	0.0781	0.0537
	Hampel	0.0098	0.0772	0.0559
40%	Huber	0.0482	0.0142	0.0328
	Bisquare	0.0540	0.0689	0.0695
	Hampel	0.0556	0.0636	0.0731

As a conclusion, the performance of TSRWLS using three different robust's weight functions which are Huber, Bisquare and Hampel performed equally since the value of the error measures (the standard error, biasness and the root mean square error) in the numerical example and the simulation study are relatively close to one another. Therefore, it suggested that any robust's weight function, either Huber, Hampel or Bisquare can be used as the second weight in the procedure of TSRWLS. However, the use of Huber's weight function as the second weight in the procedure of TSRWLS is recommended because the function is simpler than other two weight functions.

Table 7: Root mean square error of parameters for three robust's weight function

Sample size n =30				
Outliers	Robust weight Function	Root Mean Square Error		
		β_0	β_1	β_2
0%	Huber	0.5076	0.9236	0.3154
	Bisquare	0.5326	0.9707	0.3354
	Hampel	0.4955	0.9114	0.3195
10%	Huber	1.8400	3.2484	1.2119
	Bisquare	1.7362	3.0400	0.8958
	Hampel	1.7660	3.0399	0.8789
20%	Huber	2.0957	3.7629	2.0328
	Bisquare	2.1108	3.8216	1.2389
	Hampel	2.0797	3.7678	1.2223
30%	Huber	4.3270	7.6278	3.4126
	Bisquare	3.0468	5.5151	1.7432
	Hampel	3.0378	5.5052	1.7476
40%	Huber	4.9040	8.3541	3.1220
	Bisquare	3.9348	7.1874	2.2944
	Hampel	3.9451	7.2005	2.3020
Sample Size n = 100				
0%	Huber	0.2510	0.4670	0.1482
	Bisquare	0.2560	0.4739	0.1558
	Hampel	0.2343	0.4381	0.1384
10%	Huber	0.6320	1.1179	0.3759
	Bisquare	0.6167	1.1217	0.3512
	Hampel	0.5620	1.0223	0.3252
20%	Huber	0.8935	1.6118	0.4966
	Bisquare	1.0581	1.9058	0.5940
	Hampel	1.0128	1.8321	0.5673
30%	Huber	1.3140	2.4105	0.7167
	Bisquare	1.4873	2.6618	0.8636
	Hampel	1.4779	2.6441	0.8568
40%	Huber	1.8826	3.2483	1.3698
	Bisquare	1.9337	3.4283	1.3170
	Hampel	1.9306	3.4250	1.1650

References

Bellio R and Ventura L (2005). An introduction to robust estimation with R functions. In the 1st Conference on International Work, University of Padova, Padua, Italy: 1-57.

Chatterjee S and Price B (1977). Regression analysis by examples. Wiley, New York, USA.

Habshah M, Rana MS, and Imon AR (2009). The performance of robust weighted least squares in the presence of outliers and heteroscedastic errors. WSEAS Transactions on Mathematics, 8(7): 351-361.

Habshah M, Rana S, and Imon AHMR (2013). On a robust estimator in heteroscedastic regression model in the presence of outliers. In the Conference on Engineering (WCE'13), London, UK, 1: 280-285.

Kutner MH, Nachtsheim C, and Neter J (2008). Applied linear regression models. McGraw-Hill/Irwin, New York, USA.

Schmidheiny KURT (2012). Heteroskedasticity in the Linear Model. In: Schmidheiny KURT (Ed.), Short Guides to Microeconometris: 1-10. Univeristat Basel, Basel, Switzerland.

Sosa-Escudero W (2009). Heteroskedasticity and weighted least squares. Econ 507. Econometric Analysis. Available online at: https://www.noexperiencenecessarybook.com/QNjmX/heteroskedasticity-and-weighted-least-squares.html