Jurnal
Teknologi

Full Paper

THE PERFORMANCE OF LEVERAGE BASED NEAR NEIGHBOUR-ROBUST WEIGHT LEAST SQUARES IN MULTIPLE LINEAR REGRESSION IN THE PRESENCE OF HETEROSCEDASTIC ERRORS AND OUTLIER

Khoo Li Peng*, Robiah Adnan, Maizah Hura Ahmad

Department of Mathematical Science, Faculty of Science, Universiti Teknologi Malaysia, 81310, UTM Johor Bahru, Johor Darul Ta'azim, Malaysia

Article history
Received
24 March 2015
Received in revised form
12 April 2015
Accepted
2 August 2015

*Corresponding author
lipeng.khoo@gmail.com

Abstract

In this study, Leverage Based Near Neighbour–Robust Weighted Least Squares (LBNN-RWLS) method is proposed in order to estimate the standard error accurately in the presence of heteroscedastic errors and outliers in multiple linear regression. The data sets used in this study are simulated through monte carlo simulation. The data sets contain heteroscedastic errors and different percentages of outliers with different sample sizes. The study discovered that LBNN-RWLS is able to produce smaller standard errors compared to Ordinary Least Squares (OLS), Least Trimmed of Squares (LTS) and Weighted Least Squares (WLS). This shows that LBNN-RWLS can estimate the standard error accurately even when heteroscedastic errors and outliers are present in the data sets.

Keywords: Heteroscedastic errors, outliers, Leverage Based Near Neighbour–Robust Weighted Least Squares, Monte Carlo simulation, standard errors

Abstrak

Kaedah Leveraj Berdasarkan Near-Neighbour-Robust Pemberat Kuasa Dua Terkecil (LBNN-RWLS) telah dikemukan untuk menganggar ralat piawai dengan tepat semasa kesilapan heteroscedastic dan titk terpencil dalam multiple linear regression. Sets data yang digunakan dalam kajian ini disimulasi dengan mengguna keadah Simulasi Monte Carlo. Set data tersebut mengandungi kesilapan heteroscedastic dan peratus titik terpencil yang berbeza dengan saiz sample yang berbeza. Kajian ini menunjukan LBNN-RWLS dapat menghasilkan ralat paiwaian yang terkecil banding dengan Persamaan Kuasa Dua Terkecil (OLS), Kuasa Dua Papasan Terkecil (LTS), dan Pemberat Kuasa dua Terkecil (WLS). Ini menujukan bahawa LBNN-RWLS dapat menganggar ralat piawai yang tepat walaupun kesilapan heteroscedastic dan titik terpencil terdapat dalam set data.

Kata kunci: Kesilapan heteoscedastic, Titik terpencil, Leveraj Berdasarkan Near-Neighbour-Robust Pemberat Kuasa Dua Terkecil, simulasi Monte Carlo, ralat piawai

1.0 INTRODUCTION

Outliers are defined as extreme observations in data sets. The presence of outliers can dramatically change the magnitude of regression coefficient and even the direction of coefficient sign (from positive to negative or vice versa). Outliers have adverse effect on ordinary least squares (OLS) method. Furthermore, the estimates obtained from OLS in data sets that contain outliers are not efficient and may cause swamping and masking effect [1]. This will make the results that are obtained from OLS to be no longer reliable.

Rousseeuw and Leroy (1987) proposed the use of robust statistics in standard error and parameter estimations [2]. Robust statistics are able to provide reliable results even when the outliers are present in the data sets. Least Trimmed of Squares (LTS) is a robust statistic that has a high breakdown point of 50%.

Therefore, LTS can be considered as effective and efficient robust statistics as recommended by Ryan [3].

Heteroscedastic errors occur when the variance of errors for a data sets are not constant. Furthermore, heteroscedatic errors will also produce bias in the estimation of parameter and lead to inaccurate data analysis results. Heteroscedastic errors can also cause the hypothesis testing to fail. Weighted Least Squares (WLS) has been used to solve the heteroscedasticity problem. However, the problem becomes more complicated when both of the heteroscedastic errors and outliers are present in the data sets. Currently, there are no reliable methods to solve both problems effectively and efficiently [4].

In this study, Leverage Based Near Neighbour-Robust Weighted Least Squares (LBNN-RWLS) is proposed in order to estimate parameters reliably when the data sets contain outliers and heteroscedastic errors in multiple linear regression. The performances of LBNN-RWLS have been investigated using simulated data with heteroscedastic errors and different percentages of outliers with different sample size.

2.0 METHODOLOGY

2.1 Near-Neighbour Group

Near-Neighbour Method was proposed by Montgomery *et al*. in 2001 to estimate a heteroscedastic model [6]. In this method, the several near-neighbour data were groups by explanatory variables X. The group mean would represent the explanatory variable X and the variance Y are regressed corresponding to group mean of X. However this method is only valid on simple linear regression. So in this paper, the idea is extended to multiple linear regression.

2.2 Huber Weight Function

Huber function is the most widely used weight function which can be used for damping the influence of outlying cases [6].

Huber weight function is defined as:

$$w = \begin{cases} 1 & |e_i| \le 1.345 \\ \dfrac{1.345}{|e|} & |e_i| > 1.345 \end{cases}$$

Where w denotes the weight, e_i denotes the scaled residual and 1.345 is the turning constant. The Huber function is believed to be capable of making the weighted least squares procedure 95% efficient for data generated by normal error regression model [6].

2.3 Leverage Based Near Neighbour – Robust Weight Least Squares (LBNN-RWLS)

LBNN-RWLS which is the combination of Leverage Based Near-Neighbour and Robust Weighted Least Squares is used to handle heteroscedastic errors and outliers simultaneously in multiple linear regression. In this paper, the use of LBNN-RWLS will be demonstrated to solve heteroscedastic errors and outliers simultaneously in multiple linear regression.

The algorithm for LBNN-RWLS can be defined as:

1. Finding the near-neighbour
   - Compute the leverage value for the explanatory variables (diagnol hat matrix (h_{ii}))
   - Correspond the h_{ii} with y_i and x_{ij}, where $i = 1,2,\dots,n; j = 1,2,\dots,p$.
   - Sort h_{ii} from smaller to larger, carrying along y_i and x_{ij}
   - Cluster the nearby leverage h_{ii}, and obtain the $Y_{(i)jk}$ and $X_{(i)jk}$ where$(i) = 1,2,\dots,g; j = 1,2,\dots,p$, where p is the number of parameters and $k = 1,2,\dots,n_i$, where n_i is the number of observations for each cluster.

2. Determining the weight
   - After forming the $Y_{(i)jk}$ and $X_{(i)jk}$ groups, calculate $Med\,(X_{(i)}), j = 1,2,\dots,g$ and $(Y_{(i)}) = Median\{|Y_j - Median(Y_j)|\}$.
   - Regress $\{MAD(Y_j)\}^2$ on $Med\,(X_{(j)})$ by LTS and compute the regression coefficient.
   - Calculate the fitted value $y(\hat{y})$ based on the variable $X's$ by using the regression coefficients obtained by using LTS.
   - Define the weight value according to
     i. $W_{1i} = \dfrac{1}{|\hat{y}|}$
     ii. $W_{2i} = \begin{cases} 1 & |e_i| \le 1.345 \\ \dfrac{1.345}{|e_i|} & |e_i| > 1.345 \end{cases}$

     where 1.345 is the turning constant for Huber weight function
     iii. Final weight

3. Heteroscedasticity corrections
   - Perform the WLS by using the weight values obtained in step 2.
   - Use the regression coefficients obtained by WLS to estimate the parameters and standard error.

2.4 Data Simulation

The performance of LBNN-RWLS is investigated by using data simulated through the Monte Carlo simulation.

The multiple linear regression model of the simulated data set is as follow:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_{ij} \quad (1)$$

Where

$\beta_0 = 10, \beta_1 = 2, \beta_2 = 2.5, and\ \beta_3 = 3$. The numerical values for β can be chosen ambiguously.

x_{1i} is uniformly distributed in [0,1], x_{2i} is normally distributed in [0,1] , and x_{3i} is from chi square distribution [n, n-1] with sample sizes 30, 60, and 120 respectively. In order to generate heteroscedastic errors, $\varepsilon_{ij} \sim N(0, \sigma_j^2), i = 1,2, \dots, n\ and\ j = 1,2, \dots, g$ where g is the number of error groups in each corresponding j. In this paper, three sample sizes of 30, 60, and 120 which are kept fixed over repeated samples.

The random errors are simulated in two ways as shown below:

Group A: In group A , the heteroscedastic errors are categorized into 10 per group.

For generating 30 random errors, third ten-random-errors were generated from *N(0,k)* where *k=1,2,3*. For generating 60 random errors, six-ten-random-errors was generated from *N(0,k)* where *k=1,2,3,4,5,6* whereas for generating 120 random errors, twelfth ten-random-errors were generated from *N(0,k)* where *k=1,2,3,4,5,6,7,8,9,10,11,12*.

Group B: In group B, the heteroscedastic errors are categorized into 3 groups.

For generating 30 random errors, the random errors of $\frac{n}{3}$ were generating from *N(0,k)* where *k=1,2,3*. While For generating 60 random errors, the random errors of $\frac{n}{3}$ were generated from *N(0,k)* where *k=1,2,3,4,5,6*. whereas for generated 120 random errors, random errors of n/3 were generated from *N(0,k)* where k=1,2,3,4,5,6.

3.0 RESULTS AND DISCUSSION

3.1 Results

Table 1 Standard errors for different estimation techniques and different types of data sets for Group A

Sample size	Method	Different Types of data sets							
		Clean data	Heteroscedastic error + 0% outliers	Heteroscedastic error + 5% outliers	Heteroscedastic error + 10% outliers	Heteroscedastic error + 15% outliers	Heteroscedastic error + 20% outliers	Heteroscedastic error + 25% outliers	Heteroscedastic error + 30% outliers
30	OLS	2.0400×10^{-06}	2.1370	150.6000	213.1000	277.1000	312.7000	357.4000	358.6000
	LTS	7.9200×10^{-06}	1.1700	1.1680	1.1680	1.8810	2.2270	2.3570	2.5380
	WLS	0.0181	0.2173	12.3131	14.5168	19.8223	21.5038	28.4616	23.2357
	LBNN-RWLS	0.0037	0.0049	0.4329	0.4535	0.1551	0.0036	0.0679	1.7649
60	OLS	3.0510×10^{-05}	4.2320	153.9000	180.8000	938.0000	970.6000	1038.0000	1406.0000
	LTS	3.1000×10^{-05}	2.1160	1.9820	2.2850	2.7260	3.1390	3.5120	3.9410
	WLS	0.0400	0.3178	10.5067	12.2964	43.2757	44.2954	43.9348	46.4542
	LBNN-RWLS	0.0111	0.0773	0.0097	0.2508	0.5950	0.8233	0.1744	0.2068
120	OLS	3.0770×10^{-08}	7.0100	387.8000	541.0000	613.7000	699.6000	745.0000	779.3000
	LTS	3.3850×10^{-08}	3.9530	3.9770	4.2650	4.9090	4.9980	5.5200	6.6690
	WLS	0.0982	0.3681	27.2163	27.7821	23.5381	27.7427	28.5495	29.7944
	LBNN-RWLS	0.4482	0.0691	0.5219	0.4786	0.8614	0.9721	0.9852	0.9881

Table 2 Standard errors for different estimation techniques and different types of data sets for Group B

Sample size	Method	Different Types of data sets							
		Clean data	Heteroscedastic error + 0% outliers	Heteroscedastic error + 5% outliers	Heteroscedastic error + 10% outliers	Heteroscedastic error + 15% outliers	Heteroscedastic error + 20% outliers	Heteroscedastic error + 25% outliers	Heteroscedastic error + 30% outliers
30	OLS	2.0400×10^{-06}	2.1370	150.6000	213.1000	277.1000	312.7000	357.4000	358.6000
	LTS	7.9200×10^{-06}	1.1700	1.1680	1.1680	1.8810	2.2270	2.3570	2.5380
	WLS	0.0181	0.2173	12.3131	14.5168	19.8223	21.5038	28.4616	23.2357
	LBNN-RWLS	0.0037	0.0049	0.4329	0.4535	0.1551	0.0036	0.0679	1.7649
60	OLS	3.0510×10^{-05}	2.007	152.7	192.4	212.9	236.5	282.3	1025
	LTS	3.1000×10^{-05}	1.355	1.413	1.614	1.65	1.856	2.272	2.543
	WLS	0.0400	0.162667	10.4426	12.7906	13.795	15.30481	17.30314	33.49464
	LBNN-RWLS	0.0111	0.00016	0.009884	0.0106149	0.013883	0.01618	0.021732	0.436676
120	OLS	3.0770×10^{-08}	1.9770	216.9000	273.5000	316.4000	386.4000	491.5000	442.5000
	LTS	3.3850×10^{-08}	1.3970	1.6080	1.6900	1.7970	1.9330	2.3130	2.0400
	WLS	0.0982	7.6202	10.3464	13.3173	15.0084	23.1686	21.4723	20.1297
	LBNN-RWLS	0.4482	0.0019	0.0017	0.0024	0.0169	0.0240	0.0240	0.0221

3.2 Discussions

Table 1 shows the standard errors that are obtained from sample sizes 30, 60 and 120 where the percentages of outliers are 0%, 5%, 10%, 15%, 20%, 25% and 30%. There are 10 heteroscedastic errors per group. Table 1 shows that the standard errors for the data sets increase as the sample sizes increase. This is because as the samples size increase, the variations of heteroscedastic errors increase. Therefore, the effect of heteroscedastic errors increase as the samples size increases.

From the table, OLS is show to be the best method for the clean data implies that the data did not contain any outliers and heteroscedastic errors. However the standard errors of the parameter estimates of OLS gets larger as the percentage of outliers and heteroscedastic errors increase in the data sets.

LTS and WLS performed much better than OLS in the presence of outliers and heteroscedastic errors. However, the results obtained by LTS and WLS also larger when there are outliers and heteroscedastic errors in the data sets. Therefore, LTS and WLS did not perform when there are outliers and heteroscedastic errors occur in the data sets.

The results obtained from LBNN-RWLS show that it can perform well when there are outliers and heteroscedastic errors in the data sets. However, the standard errors that are obtained from LBNN-RWLS are larger in the clean data. As the percentage of the outliers increase and there are heteroscedastic errors in the data sets, the standard errors that are obtained by LBNN-RWLS is smaller compared to OLS, LTS and WLS. Therefore, LBNN-RWLS estimates have better performed compared to estimate from other methods where outliers and heterscedastic errors exits in data sets.

Table 2 shows the standard errors that are obtained for sample sizes 30, 60 and 120 where the percentage of outliers are 0%, 5%, 10%, 15%, 20%, 25% and 30% for the data of group B. The heteroscedastic errors are categorized into 3 groups. As in Table 1, OLS is the most performing method in clean data and it produces bias when the percentage of outliers and heteroscedastic errors increase in the data sets. LTS and WLS did not performed well when the percentage of outliers increases and there are heteroscedsatic errors in the data sets. Meanwhile LBNN-RWLS performed very well compared to the other methods when there are outliers and heteroscedastic errors in the data sets.

Table 1 and Table 2 present the standard errors that are obtained from different methods such as OLS, LTS, WLS and LBNN-RWLS with different types of data sets. The results show that the standard errors of OLS get larger as the heteroscedastic errors and the percentage of outliers increase in the data sets. Furthermore when the sample size increases, the standard errors obtained from OLS also getting larger. Least Trimmed of Squares also did not perform well when there are heteroscedastic errors and outliers in the data set. The standard errors that are obtained from LTS get large when there are heteroscedasric errors and the percentage of outliers increase. Similarly the standard errors from WLS show that when the heteroscedastic and percentage of outliers increase, it also increases. Therefore, WLS also did not perform well too when heteroscedastic errors and outliers are present in the data sets.

However, the LBNN-RWLS performed the best compared to OLS, LTS and WLS when there are heteroscedastic errors and outliers in the data sets. Furthermore, LBNN-RWLS also performed well when the outliers and heteroscedastic errors are presented even the percentage of outliers and sample size increase in data sets.

In addition, Table 2 shows the smaller standard errors results compared to Table 1. This is because the variation of heteroscedasticity for Group A data set is larger compared to Group B data sets. The effect of heteroscedastic errors in Group A is bigger than that in Group B which lead to more bias in estimation. However, LBNN-RWLS is able to produce smaller standard errors in both Group A and Group B.

Thus, LBNN-RWLS is a more reliable method compared to OLS, LTS and WLS when there are heteroscedasticity errors and outliers in the multiple linear regression method.

4.0 CONCLUSION

The main purpose of this paper is to investigate the performance of LBNN-RWLS estimates in the presence of heteroscedastic errors and outliers. From the results obtained from a simulation study, LBNN-RWLS estimators performed the best compared to OLS, LTS and WLS estimators especially the data contains high percentage of outliers and heteroscedatic errors with large sample size. Therefore, LBNN–RWLS can be concluded to be a reliable and efficient method in estimating the parameters in the presence of heteroscedastic errors and outliers in multiple linear regression.

Acknowledgement

We acknowledge the financial support from Universiti Teknologi Malaysia for the Research Grant (QJ130000.2526.06H68) and Ministry of Higher Education (MOHE) of Malaysia.

References

[1] Habshah, M., Noraznan, M. R., Imon, A. H. M. R. 2009. The Performance of Diagnostic-robust Generalized Potential for the Identification of Multiple High Leverage Points in Linear Regression. *Journal of Applied Statistics*. 36(5): 507-520.

[2] Rousseeuw, P. J. and Leroy, A. 1987. *Robust Regression and Outliers Detection*. Wiley, New York.
[3] Ryan T. P. 1997. *Modern Regression Methods*. Wiley, New York.
[4] Habshah, M., Rana M. S., Imon, A. H. M. R. 2009. The Performance of Robust Weighted Least Squares in the Presence of Outliers and Heteroscedastic Errors. *WSEAS Transactions on Mathematics*. 7(8): 351-361.
[5] Montgomery, D. G., Peck, D. E. and Vining, G. G. 2001. *Introduction to Linear Regression Analysis*. 3rd ed. John Wiley and Sons, New York.
[6] Kutner, M. H., Nacthsheim, C. J., Neter, J. *et al*. 2005. *Applied Linear Statistical Model*. 5th ed. McGraw-Hill Irwin, United State of America.