MJFAS Malaysian Journal of Fundamental and Applied Sciences
ISSN:2289-5981 e-ISSN:2289-599X
UTM UNIVERSITI TEKNOLOGI MALAYSIA

Standard errors estimation in the presence of high leverage point and heteroscedastic errors in multiple linear regression

Khoo Li Peng*, Robiah Adnan, Maizah Hura Ahmad

Department of Mathematical Science, Faculty of Science, UTM, 81310 UTM Skudai, Johor, Malaysia
**Corresponding Author: lipeng.khoo@gmail.com (L.P. Khoo)*

Article history :
Received 19 February 2014
Revised 1 April 2014
Accepted 5 June 2014
Available online 5 July 2014

GRAPHICAL ABSTRACT

ABSTRACT

In this study, the Robust Heteroscedastic Consistent Covariance Matrix (RHCCM) was proposed in order to estimate standard errors of regression coefficients in the presence of high leverage points and heteroscedastic errors in multiple linear regression. Robust Heteroscedastic Consistent Covariance Matrix (RHCCM) is the combination of a robust method and Heteroscedasticit Consistent Covariance Matrix (HCCM). The robust method is used to eliminate the effect of high leverage points while HCCM is mainly used to eliminate the effect of heteroscedastic errors. The performance of RHCCM was assessed through an empirical study and compared with results obtained when the original Heteroscedastic Consistent Covariance Matrix was used.

Keywords: Robust Heteroscedastic Consistent Covariance Matrix (RHCCM), High Leverage Point, Heteroscedastic Erorrs, Multiple Linear Regression

http://dx.doi.org/xx.xxx/xxx.xxx.xxx

1. INTRODUCTION

High leverage points are the observations that have extreme values in independent variables (x spaces) and will influence the intercept and slope estimation in the method of least squares. These high leverage points can be caused by a gross error in x, a unique priceless observation, or an accurate but useless observations [1].

The heteroscedastic errors will mislead the ordinary least squares estimate of regression coefficients to become inefficient which resulted in the inaccuracy conclusion [2]. Heteroscedasticity yielded hypothesis tests that fail to keep false rejections at the nominal level; and estimated standard errors as well as confidence intervals to become either too narrow or too large [3].

Ordinary Least Squares (OLS) method is a well-known method that is able to provide efficient and unbiased parameter estimation when there are no high leverage points or heteroscedastic errors in multiple linear regression. Nevertheless, the ordinary least squares method does not perform well in the presence of high leverage points and heteroscedastic errors, resulting in hypothesis tests that are liberal or conservative. Ordinary Least Squares Covariance Matrix (OLSCM) whose diagonal elements are used to estimate the standard error and regression coefficient becomes biased and inconsistent due to the effects of heteroscedastic errors.

Furthermore, the presence of high leverage points can make all estimation procedures meaningless. None of the estimation techniques work well when high leverage points and heteroscedastic errors are present at the same time in the regression model [3].

The main focus in this study is to estimate the standard errors of the regression coefficients in the presence of high leverage points and heteroscedasticity in multiple linear regression. Robust techniques and heteroscedasticity consistent covariance matrix (HCCM) will be employed for this purpose.

2. METHODOLOGY

2.1 Method

This study is focused on using Least Trimmed of Squares (LTS) and Heteroscedasticity Consistent Covariance Matrix (HCCM) methods to estimate the standard errors in the regression coefficients in the presence of high leverage points and heteroscedastic errors in multiple linear regression. LTS is used in order to eliminate the effect of high leverage points since OLS is believed to be highly influenced by high leverage points while HCCM is an alternative and high appealing method in reducing the effect of heteroscedasticity.

2.2 Robust heteroscedasticity consistent covariance matrix

Heteroscedasticity Consistent Covariance Matrix (HCCM) estimators are derived from an estimate of variance-covariance matrix of the regression coefficient $\left(\sum_{\hat{\beta}}\right)$ which does not assume homoscedasticity.

Consider a multiple linear regression model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where
$\boldsymbol{y}$ is the $n \times 1$ vector of observed values for the response variables,
$\boldsymbol{X}$ is the $n \times p$ of predictor including the intercept,
$\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters, and
$\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of errors.
The covariance matrix of regression coefficient $\boldsymbol{\beta}$ is defined as:

$$\sum_{\hat{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}\ \boldsymbol{X}\ (\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2}$$

$\boldsymbol{\Omega} = \sigma^2\boldsymbol{I}$ is define as the variance of the error $\left(var\ (\varepsilon)\right)$ and it is a $(n\ \times n)$ square matrix where $(\boldsymbol{I})$ is an identity matrix of order n and $E(\varepsilon\varepsilon^T)$ is a positive definite matrix.

In homoscedasticty assumption $(\boldsymbol{\Omega} = \sigma^2\boldsymbol{I})$, the variance-covariance matrix can be defined as:

$$\sum_{\hat{\beta}} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{3}$$

In this study, the residuals values were obtained by LTS estimator. However when the error is heteroscedastic, the variance-covariance matrix can be defined as:

$$\sum_{\hat{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Phi}\ \boldsymbol{X}\ (\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{4}$$

$\boldsymbol{\Phi} = \sigma^2\boldsymbol{V}$ is defined as the variance of error $\left(var\ (\varepsilon)\right)$ from heteroscedastic error where (V) is $(n\ \times n)$ square matrix form by heteroscedastic errors .

With heteroscedasticty assumption $(\boldsymbol{\Phi} = \sigma^2\boldsymbol{V})$ the estimators of variance-covariance matrix becomes biased and the hypothesis tests are either too liberal or conservative.

Over the last 25 years, several heteroscedasticity consistent covariance matrices have been developed. There are five HCCM estimators developed over the last 25 years which are defined as $HC0$,$HC1$, $HC2$, $HC3$and $HC4$.

White in 1980 was proposed substituting the i^{th}squares error into the i^{th} row of diagonal of the Φ matrix, making $\Phi = \text{diag}[\varepsilon_i^2]$ to be the diagonal matrix of the squares of OLS residuals [4]. However, in this study, LTS residuals replaced the OLS residuals so that the errors were not affected by high leverage points. Therefore the $HC0$ estimator is defined as:

$$HC0 = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}diag[e_i^2]\boldsymbol{X}(\boldsymbol{X'X})^{-1} \tag{5}$$

where the main diagonal of $HC0$ are the estimated squared standard errors of regression coefficients. The biasness of $HC0$ increases when the sample sizes are decreased. $HC1$ proposed by Hinkley in 1977 is a simple degree of freedom adjustment of $HC0$ and every squares residuals is multiplied by $\frac{n}{(n-p-1)}$ [5].

$$HC1 = \frac{n}{n-p-1}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}diag[e_i^2]\boldsymbol{X}(\boldsymbol{X'X})^{-1} \tag{6}$$

$HC2$ was introduced by Mackinnon and White in 1985 [6]. For $HC2$, the i^{th}squared residuals is weighted by $(1 - \boldsymbol{h_{ii}})$ instead of a degree of freedom correction. h_{ii} are the leverage values obtained from the diagonal elements in the "hat" matrix which define as

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} \tag{7}$$

However HC2 will produce bias due to the high leverage points in explanatory variables.

$$HC2 = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}diag\left[\frac{e_i^2}{1-\boldsymbol{h_{ii}}}\right]\boldsymbol{X}(\boldsymbol{X'X})^{-1} \tag{8}$$

$HC3$ was proposed by Davidson and Mackinnon in 1993 [7]. $HC3$ weighted each squares of residuals by $\frac{1}{(1-h_{ii})^2}$. Besides that, $HC3$ is always recommended because it can keep test sizes at nominal level regardless with the presence or absence of heteroscedasticity. Its performance is dependent to some extent on the presence or absence of high leverage points.

$$HC3 = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}diag\left[\frac{e_i^2}{(1-\boldsymbol{h_{ii}})^2}\right]\boldsymbol{X}(\boldsymbol{X'X})^{-1} \tag{9}$$

$HC4$ is the most recent proposal of HCCM estimator, derived by Cribari-Neto in 2004 with the explicit aim of taking high leverage points as consideration in standard errors estimation [8].

$$HC4 = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}diag\left[\frac{e_i^2}{(1-h_{ii})^{\delta_i}}\right]\boldsymbol{X}(\boldsymbol{X'X})^{-1} \tag{10}$$

where

$$\delta_i = min\left\{4, \frac{nh_{ii}}{p+1}\right\} \tag{11}$$

The exponent δ_i controls the level of "discounting" for the i^{th}observation, and with the truncation point at 4. $HC4$ outperformed $HC3$ in terms of test size control when there are high leverage points and non-normal errors [8].

3. RESULTS & DISCUSSION

3.1 Simulation example

A simulation study was designed to investigate Robust Heteroscedasticity Consistent Covariance Matrix

(RHCCM). The following multiple regression model was considered:-

$$y_i = 10 + 2x_{1i} + 2.5x_{2i} + 3x_{3i} + \varepsilon_i$$

where

x_{1i} is uniformlly distributed on [0,1], x_{2i} is normally distributed on [0,1] , and x_{3i} from chi square distribution.

The random errors ε_{ij}'s were drawn from normal distribution where $\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$, $i=1,2,...,n$ and $j = 1, 2, ...,g$ where g is the number of error groups in each sample and each group consisted of 10 random errors.

In order to generate the heteroscedastic errors, 50 random errors were generated by taking the first 10 random errors from $N(0,1)$, the second 10 random errors from $N(0,2)$, the third 10 random errors from $N(0,3)$, the fourth 10 random errors from $N(0,4)$ and the fifth 10 random errors from $N(0,5)$. Thus the errors that were generated have a zero mean and non-constant variance.

3.2 Numerical example

In this section, a modified education expenditure data taken was from Chatterjee and Price in 1997 which state at chapter 4 and page 97 [9]. The data represent the relationship between response variable and three independent variables for 30 states in United State of America. The data was used to evaluate the performance of robust heteroscedaticity consistent covariance matrix (RHCCM). The variables of the data are as shown as the followings:-

y: Per capita income on education projected for 1975
x_1: Per capita income in 1973
x_2:Number of residents per thousand under 18 years of age in 1974
x_3: Number of residents per thousand living in urban areas in 1970

The original data set is modified in order to contain high leverage points and heteroscedastic errors in the data set. Some of the explanatory variables (x) had been modify to contain high leverage point while the error of the data been modify to obtain the heteroscedatic error. Therefore the new data set is present in the high leverage point and heteroscedastic errors.

3.3 Discussion

In this paper, the performance comparison between HCCM and RHCCM was done with the presence of high leverage points and heteroscedastic errors.

Table 1 Summary of Standard Errors Estimation by using HCMM for Simulation Data

HCCM						
Methods	Statistical Analysis	Coefficients				
		β_0	β_1	β_2	β_3	SE (Res)
OLS	Values	157.626	-0.1734	-0.106	0.04856	26.9700
OLSCM	SE	4.5282	0.3188	0.8938	0.0399	
HC0	SE	4.5236	0.1572	0.66331	0.0315	
HC1	SE	4.7161	0.1639	0.66	0.03287	
HC2	SE	4.7893	0.2047	0.8349	0.0413	
HC3	SE	5.2016	0.2825	1.1257	0.05486	
HC4	SE	6.72	0.6458	2.2126	0.1007	

Table 2 Summary of Standard Errors Estimation by using Robust HCMM for Simulation Data

ROBUST HCCM						
Methods	Statistical Analysis	Coefficients				
		β_0	β_1	β_2	β_3	SE (Res)
LTS	Values	6.041	1.913	3.137	3.084	2.5160
LTSCM	SE	124.5254	8.7668	24.5797	1.0977	
HC0	SE	164.03	12.1226	44.1093	3.466	
HC1	SE	171.0131	12.6387	45.9871	3.6135	
HC2	SE	246.6036	17.9311	61.7509	5.4379	
HC3	SE	382.3588	27.4772	87.9515	8.7538	
HC4	SE	994.6921	70.71093	190.8772	23.9862	

Table 3 Summary of Standard Errors Estimation by using HCMM for Numerical Example

HCCM						
Methods	Statistical Analysis	Coefficients				
		β_0	β_1	β_2	β_3	SE (Res)
OLS	Values	215.8626	0.0017	-0.0384	0.1111	59.1000
OLSCM	SE	9.9240	0.6987	1.9589	0.0875	
HC0	SE	39.9367	0.0056	0.0316	0.0475	
HC1	SE	41.6369	0.0058	0.0330	0.0495	
HC2	SE	39.9353	0.0056	0.0316	0.0475	
HC3	SE	60.9905	0.0146	0.0479	0.0670	
HC4	SE	127.0706	0.0429	0.0745	0.1407	

Table 4 Summary of Standard Errors Estimation by using Robust HCMM for Numerical Example

ROBUST HCCM						
Methods	Statistical Analysis	Coefficients				
		β_0	β_1	β_2	β_3	SE (Res)
LTS	Values	177.2824	0.01413	-0.20811	0.14409	27.7600
LTSCM	SE	12.232	0.8612	2.4144	0.1078	
HC0	SE	51.1350	0.0119	0.1214	0.0663	
HC1	SE	53.3119	0.0124	0.1265	0.0691	
HC2	SE	47.6717	0.0088	0.0388	0.0541	
HC3	SE	116.9089	0.0348	0.2597	0.1241	
HC4	SE	127.0706	0.0429	0.0745	0.1407	

The main interest in this study was to estimate the standard errors in multiple linear regression inthe presence of high leverage points and heteroscedastic errors by using RHCCM which is the combination of LTS and HCCM method. LTS was able to estimate the parameter without the influence by high leverage point while HCCM was able to estimate the parameter in the presence of heteroscedastic error accurately.

Table 1 shows the results of standard errors estimation in simulation data by using HCCM, while Table 2 shows the standard errors estimation in simulation data by using RHCCM. Both tables show the standard errors estimation by the family of heteroscedasticity consistent covariance matrix from $HC0$ to $HC4$.

The standard error of residuals obtained from both HCCM and RHCCM in simulation data are 26.9700 and 2.5160 respectively. The estimated values of regression coefficients obtained from RHCMM are closer to the true values and the estimated standard errors are also smaller than HCCM. However, the standard errors obtained from RHCCM are larger than HCCM. Therefore the parameters that estimate from RHCCM are more accurate than the HCCM method and overall performance of RHCMM is better than HCCM in standard errors estimation.

Furthermore, the standard error of regression coefficients obtained from $HC0$ until $HC4$ through HCCM and Robust HCCM are shown in Table 1 and Table 2. The results obtained from $HC4$ in RHCCM are more reliable compared to the other HCCM estimator since it was reported that $HC4$ will be less influenced by the high leverage points and RHCCM performed better than HCCM [8].

Table 3 and Table 4 show the standard error estimations obtained by HCCM and RHCCM when used in our numerical example. The standard errors of residuals obtained from HCCM and RHCCM are 59.1000 and 27.7600 respectively. This shows that the standard error of residuals obtained from RHCCM is smaller compared to HCCM which indicates that the RHCCM is more accurate than HCCM in parameter estimation in the presence of high leverage point and heteroscedastic error and the performance of RHCCM is better in parameter estimation in the presence of high leverage point and heteroscedastic error.

The results of the standard errors that were obtained using $HC4$ in HCCM and RHCCM are the same. It shows that $HC4$ was not influenced by high leverage points as stated by Cribari-Neto [8]. Therefore, the standard errors that was obatined by $HC4$ in RHCCM is the most accurate compared to other family of heteroscedasticity consistent covarince matrix and HCCM.

As the overall conclusion, the performanceof RHCCM in standard errors estimation was better compared

to the HCCM especially using $HC4$ in the family of heteroscedasticity consistent covariance matrix.

4. CONCLUSION

The main interest in this study is standard errors estimation using HCCM and RHCCM in the presence of high leverage points and heteroscedastic errors in multiple linear regression model. According to the results obtained from HCCM and RHCCM, the RHCCM performed better than HCCM when high leverage points and heteroscedastic errors are present in the data.

ACKNOWLEDGEMENT

The authors thank the Department of Mathematical Science, Faculty of Science, Universiti Teknologi Malaysia, Johor as well as to Malaysian Government for the funding the project.

REFERENCES

[1] P. J. Huber, E. M. Ronchetti, Robust Statistics, Wiley & Sons Inc., United State of America, 2009.
[2] M. A Mukhtar, C. Giaccitto, J. Appl. Econom. 26 (1984) 355.
[3] H. Midi, S. Rana, A. H. M. Imon, J. Appl. Sci. 9 (2009) 4013.
[4] H. A. White, Econometrica. 48 (1980) 817.
[5] D. V. Hinkley, Technometrics. 19 (1977) 285.
[6] J. G. Mackinnon, H. J. White, J. Econometrics. 29 (1985) 305.
[7] R. Davidson, J. G. Mackinnon, Estimation and Inference in Econometrics, Oxford University Press, Oxford, 1993.
[8] F. Cribari-Neto, Comput. Stat. Data. An. 45 (2004) 215.
[9] S. Chatterjee, B. Price, Regression Analysis by Example, John Wiley & Sons Inc., United States of America, 1977.