Malaysian Journal of Mathematical Sciences 13(S) April: 71–88 (2019)
Special Issue: 3rd International Conference on Mathematical Sciences and Statistics (ICMSS2018)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

The Performance of Robust Heteroscedasticity Consistent Covariance Matrix Estimator

Sani, M.[1,3], Midi, H. [*1,2], and Arasan, J.[1,2]

[1]*Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia*
[2]*Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Malaysia*
[3]*Department of Mathematical Sciences, Federal University Dutsinma, Katsina State, Nigeria*

E-mail: habshahmidi@gmail.com
[*] *Corresponding author*

Abstract

The weighted least squares (WLS) method together with heteroscedasticity consistent covariance matrix (HCCM) estimator is often used to estimate the parameters of a heteroscedastic regression model when the form of errors structure is unknown. However, WLS based on weight determined by hat matrix suffers much set back in the presence of high leverage points (HLPs) in a data set. Moreover, the use of WLS requires an efficient weighting method that will successfully down weight HLPs. In this paper, we proposed new weighting method based on HLPs detection measure in which the good leverage points are allowed to contribute in the estimation of parameters and the bad leverage points are down weighted as they are responsible for the deviation of the model fit. In the proposed method we employed modified generalized studentized residuals (MGt) with diagnostic robust generalized potentials based on index set equality (DRGPISE) termed FMGt on HCCM estimator. The performance of the proposed weighting method is assessed by generated artificial data set.

Keywords: Ordinary least squares, weighted least squares, linear regression, robust HCCM estimator, high leverage points.

1. Introduction

The commonly used method for the analysis of a regression model is the ordinary least squares (OLS). Under the violation of the assumption of equal variances of the errors (homoscedasticity), the covariance matrix becomes inconsistent. White (1980) suggested replacing inconsistent OLS covariance matrix with heteroscedasticity consistent covariance matrix (HCCM) estimator denoted by HC0. This estimator is consistent under both homoscedasticity and heteroscedasticity and does not require the structural form of model heteroscedasticity White (1980). Different adjustments of HC0 were made to increase its efficiency (Cribari-Neto, 2004, Cribari-Neto et al., 2007, Cribari-Neto and Zarkos, 2009, Davidson and MacKinnon, 1993, Long and Ervin, 2000, MacKinnon and White, 1985).

The construction of HCCM estimator is based on OLS residuals vector. In the presence of anomalous observation called outliers the coefficient estimates and residuals of OLS estimate are biased. As a consequence, the inference becomes misleading. Furno[6] proposed using residuals of weighted least squares (WLS) regression in construction of robust HCCM (RHCCM) estimator, whereby the weight used by Furno is determined by the leverage measures (hat matrix) of the different observations. Lima et al. (2016) considered least median of squares (LMS) and least trimmed squares (LTS) residuals. However, both Furno's and Lima's methods were inefficient as they suffer much from the effect of swamping and masking. As the consequence, the variances tend to be large resulting to unreliable parameter estimates. The main reason for this weakness is the use of hat matrix (that is unable to discriminate between good and bad leverage points) which down weight both good and bad leverage points in RHCCM. Pena and Yohai (1995) had shown swamping and masking resulted from the presence of HLPs in linear regression. Habshah et al. (2009) also proven that hat matrix is not very successful in detecting HLPs. Consequently, less efficient estimates can be obtained by employing unreliable method of detecting HLPs. This shortcoming motivated us to use weight function based on more reliable diagnostic measure for the identification of HLPs.

In this paper, we proposed new robust weighting methods based modified generalized studentized residuals (MGt) with diagnostic robust generalized potentials based on index set equality (DRGP_{ISE}) which is also known as fast modified generalized studentized residuals (FMGt) on HCCM estimator. The

FMGt method identifies the regular observations, vertical outliers, good and bad leverage points. But, only bad leverage points and vertical outliers will be down weighted. The weight determined by FMGt is expected to successfully down weight all bad influential observations.

The article is arranged as follows: Section 2 describes the classical heteroscedasticity consistent covariance matrix (HCCM) estimators. The robust HCCM estimator based on Furno's and RMD weighting method is described in section 3 and 4 respectively. Section 5 presents the new proposed estimator. Section 6 presents examples using real data set. The last section provided the conclusion of the study.

2. The Classical HCCM Estimators

The linear regression model is given by:

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

where, y is an $n \times 1$ vector of response variables, X is an $n \times p$ matrix of explanatory variables, β is a vector of regression parameters, and ε is the n-vector of random errors. For a model with heteroscedastic errors the $E(\varepsilon_i) = 0$, $var(\varepsilon_i) = \sigma_i^2$ for $i = 1, \ldots, n$ and, $E(\varepsilon_i \varepsilon_s) = 0$ for all $i \neq s$. The covariance matrix of ε is given as $\varphi = \text{diag}\left\{\sigma_i^2\right\}$. The ordinary least squares (OLS) estimator of β is $\widehat{\beta} = (X'X)^{-1}X'y$ which is unbiased, with the covariance matrix given by

$$cov\left(\widehat{\beta}\right) = (X'X)^{-1}X'\varphi X(X'X)^{-1}, \tag{2}$$

under homoscedasticity $\sigma_i^2 = \sigma^2$ such that $\varphi = \sigma^2 I_n$, where I_n is an $n \times n$ identity matrix. The covariance matrix $cov\left(\widehat{\beta}\right) = \sigma^2(X'X)^{-1}$ is estimated by $\widehat{\sigma}^2(X'X)^{-1}$ (which is inconsistent and biased under heteroscedasticity) and $\widehat{\sigma}^2 = (\widehat{\varepsilon}'\widehat{\varepsilon})/(n-p)$, $\widehat{\varepsilon} = (I_n - H)\,y$, where H is an idempotent and symmetric matrix known as hat matrix. The hat matrix (H) is defined as $H = X(X'X)^{-1}X'$, and it plays great role in determining the HLPs in regression model. The diagonal elements $\mathrm{h_i} = x_i(x'x)^{-1}x_i'$ for $i = 1, \ldots, n$ of the hat matrix are the values for leverage of the i^{th} observations.

White (1980) proposed the most popular HCCM estimator known as HC0 where he replaced the σ_i^2 with $\widehat{\varepsilon}_i^2$ in covariance matrix of $\widehat{\beta}$ as:

$$HC0 = (X'X)^{-1}X'\widehat{\varphi}_0 X(X'X)^{-1} \tag{3}$$

where, $\widehat{\varphi}_0$= diag $\{\widehat{\varepsilon}_i^2\}$. HC0, HC1, HC2, and HC3 are generally biased for small sample size (see [6, 8, 12]). This research will focus only on HC4 and HC5. The HC4 proposed by [3] was build under HC3, which is defined as follows:

$$HC4 = (X^{'}X)^{-1}X^{'}\widehat{\varphi}_4 X(X^{'}X)^{-1} \tag{4}$$

where, $\widehat{\varphi}_4$= diag $\left\{\frac{\widehat{\varepsilon}_i^2}{(1-\mathrm{h_i})^{\delta_\mathrm{i}}}\right\}$ for $i = 1, \ldots, n$ with δ_i= min $\left\{4, \frac{\mathrm{h_i}}{\mathrm{h}}\right\}$, which control the discount factor of the i^{th} squared residuals, given by the ratio between $\mathrm{h_i}$ and the average values of $\mathrm{h_i}$'s (h). Note that, δ_i= min $\left\{4, \frac{n\mathrm{h_i}}{p}\right\}$. Since $0 < 1 - \mathrm{h_i} < 1$ and $\delta_i > 0$ it follows that $0 < (1 - \mathrm{h_i})^{\delta_i} < 1$. The larger $\mathrm{h_i}$ is relative to h, the more the HC4 discount factor inflates the i^{th} squared residual. The truncation at 4 amounts to twice what is used in the definition of HC3; that is, $\delta_\mathrm{i} = 4$ when $\mathrm{h}_i > 4h = 4p/n$. The result obtained by Cribari-Neto (2004) suggested HC4 inference in finite sample size relative to HC3.

Similarly, another modification of the exponent $(1-\mathrm{h_i})$ of HC4 was proposed by Cribari-Neto et al. (2007) to control the level of maximal leverage. The estimator was called HC5 and defined as

$$HC5 = (X^{'}X)^{-1}X^{'}\widehat{\varphi}_5 X(X^{'}X)^{-1} \tag{5}$$

where, $\widehat{\varphi}_5$= diag $\left\{\frac{\widehat{\varepsilon}_i^2}{\sqrt{(1-\mathrm{h_i})^{\alpha_\mathrm{i}}}}\right\}$ for $i = 1, \ldots, n$ with α_i= min $\left\{\frac{\mathrm{h_i}}{\mathrm{h}}, \max\left\{4, \frac{\mathrm{kh}_{max}}{\mathrm{h}}\right\}\right\}$, which determine how much the i^{th} squared residual should be inflated, given by the ratio between h_{max} (maximal leverage) and h (mean leverage value of $\mathrm{h_i}$'s). when $\frac{\mathrm{h_i}}{\mathrm{h}} \leq 4$ it follows that $\alpha_i = \frac{\mathrm{h_i}}{\mathrm{h}}$. Also, since $0 < 1 - \mathrm{h}_i < 1$ and $\alpha_i > 0$, it similarly follows that $0 < (1 - \mathrm{h}_i)^{\alpha_i} < 1$ and k is a constant ranges between $0 < \mathrm{k} < 1$ and was suggested to be chosen as 0.7 by Cribari-Neto et al. (2007) following his simulation result that leads to efficient quasi-t inference.

3. Robust HCCM Estimators based on Furno's Weighting Method

The problem of heteroscedasticity and high leverage points was addressed by Furno (1996) in order to reduce the bias caused by the effect of leverage points in the presence of heteroscedasticity. He suggested using weighted least squares (WLS) regression residuals instead of OLS residuals used by White (1980) in HCCM estimator. The weight is based on the hat matrix (h_i) and the robust (weighted) version of HC0 is defined as:

$$HC0_W = (X^{'}WX)^{-1}X^{'}W\widehat{\varphi}_{0\mathrm{w}}WX(X^{'}WX)^{-1} \tag{6}$$

where, W is an $n \times n$ diagonal matrix with,

$$w_i = \min\left(1, \quad c/\mathrm{h_i}\right), \tag{7}$$

and c is the cutoff point, $c = 1.5p/n$, p being the number of parameters in a model including the intercept and n is the sample size, $\widehat{\varphi}_{0\mathrm{w}}$= diag $\{\widetilde{\varepsilon}_i^2\}$ with $\widetilde{\varepsilon}_i$ being the i^{th} residuals from weighted least squares. Note that, non-leveraged observations are weighted by 1 and leveraged observations are weighted by $(c/\mathrm{h_i})$ to reduce their intensity and w_i is considered as the weight in this weighted least squares (WLS) regression, so that the WLS estimator of β is:

$$\widetilde{\beta} = \left(X^{'}WX\right)^{-1}X^{'}Wy. \tag{8}$$

The robust HCCM estimator for the HC4 and HC5 based on Furno's weighting method considered by Lima et al. (2016) are $HC4_W$ and $HC5_{??}$ defined as:

$$HC4_W = \left(X^{'}WX\right)^{-1}X^{'}W\widehat{\varphi}_{4\mathrm{w}}WX\left(X^{'}WX\right)^{-1} \tag{9}$$

where, $\widehat{\varphi}_{4\mathrm{w}}$= diag $\left\{\frac{\widetilde{\varepsilon}_i^2}{\left(1-\mathrm{h_i^*}\right)^{\delta_\mathrm{i}^*}}\right\}$ for $i = 1, \ldots, n$ with δ_i^*= min $\left\{4, \frac{\mathrm{h_i^*}}{\mathrm{h^*}}\right\}$, and $\mathrm{h_i^*}$ is the i^{th} diagonal of the weighted hat matrix $\mathrm{H_w} = \sqrt{W}X\left(X^{'}WX\right)^{-1}X^{'}\sqrt{W}$. And,

$$HC5_W = \left(X^{'}WX\right)^{-1}X^{'}W\widehat{\varphi}_{5\mathrm{w}}WX\left(X^{'}WX\right)^{-1} \tag{10}$$

where, $\widehat{\varphi}_{5\mathrm{w}}$= diag $\left\{\frac{\widetilde{\epsilon}_i^2}{\sqrt{\left(1-\mathrm{h_i^*}\right)^{\alpha_\mathrm{i}^*}}}\right\}$ for $i = 1, \ldots, n$ with α_i^*= min $\left\{\frac{\mathrm{h_i^*}}{\mathrm{h^*}}, \max\left\{4, \frac{\mathrm{kh}^*_{max}}{\mathrm{h^*}}\right\}\right\}$. In this paper the Furno's weighted least square for RHCCM estimation method is denoted by WLS_F.

4. Robust HCCM Estimator based on Robust Mahalanobis Distance with Minimum Volume Ellipsoid (RMD(MVE)) Weighting Method

The diagnostic measure of the deviation of a data point from its center named Mahalanobis Distance (MD) was introduce by Mahalanobis (2000), in which the independent variables of the i^{th} observations are presented as $x_i =$

$(1,\ x_{i1},\ x_{i2}, \ldots,\ x_{ik}) = (1,\ R_i)$ so that $R_i = (x_{i1},\ x_{i2}, \ldots,\ x_{ik})$ will be k-dimensional row vector, where the mean and covariance matrix vector are $\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$ and $\vartheta = \frac{1}{n-1}\sum_{i=1}^{n}\left(R_i - \overline{R}\right)^{'}(R_i - \overline{R})$ respectively. The MD for the i^{th} points is given as:

$$RMD_i = \sqrt{\left(R_i - \overline{R}\right)^{'}\vartheta^{-1}(R_i - \overline{R})}\, i = 1,\ 2, \ldots, n \tag{11}$$

Leroy and Rousseeuw (1987) recommended $\sqrt{\chi^2_{k,\ 0.5}}$ as the cutoff point of MD_i whereby, any observation that exceeds this cutoff point is considered as HLP. Imon (2002) suggested another cutoff point (cp) for RMD_i given by:

$$cp = \text{median}\,(RMD_i) + 3\text{MAD}(RMD_i) \tag{12}$$

where, MAD stands for median absolute deviation. Since, the average vector $\overline{R}$ and covariance matrix ϑ are not robust, Rousseeuw (1984) recommended using minimum volume ellipsoid (MVE) estimator of $\overline{R}$ and the corresponding ϑ produced by the ellipsoid. This technique of MVE is to produce the smallest volume ellipsoid among all the ellipsoids of at least half of the data. The MVE estimator of the average vector is T (X) = centre of the MVE covering at least h points of X, for the value of $h \geq \frac{n+k+1}{2}$ and, k is the number of explanatory variables Rousseeuw and Driessen (1999). The corresponding ϑ is provided by ellipsoid and multiplied by a suitable factor in order to obtain consistency. The weight obtained by this RMD(MVE) method is given by:

$$w_{ir} = \min\,(1,\quad cp/RMD_i) \tag{13}$$

so that, HLPs are weighted by (cp/RMD_i) and non leverage by 1. To obtain the RHCCM estimator based on RMD(MVE) weighting method denoted by WLS_{RMD}, we replace equation (7) by (13) and adopt Furno's RHCCM estimation method as discussed in Section **3**.

5. New proposed Robust HCCM Estimators

In this study, we employed the idea of Furno's RHCCM estimation on new weighting method based on modified generalized studentized residuals (MGt) and diagnostic robust generalized potential based on index set equality (DRGP (ISE)) in order to identify good and bad HLPs. We anticipate that our method will be more efficient than the existing method as only bad leverage observations

(BLOs) will be down weighted and good leverage observations (GLOs) will be allowed to contribute to the estimation. The DRGP(ISE) consist of two steps, whereby in the first step, the suspected HLPs are determined using RMD based on ISE. The suspected HLPs will be placed in the 'D' set and the remaining in the 'R' set. The generalized potential ($\hat{p}_i$) is employed in the second step to check all the suspected HLPs, those possess a low leverage point will be put back to the 'R' group. This technique continued until all points of the 'D' group has been checked to confirm whether they can be referred as HLPs. The generalized potential is defined as follows:

$$\hat{p}_i = \begin{cases} h_i^{(-D)} & for \;\; i \in D \\ \frac{h_i^{(-D)}}{1-h_i^{(-D)}} & for \;\; i \in R \end{cases} \tag{14}$$

The cut-off point for DRGP is given by,

$$C_{DRGP} = \text{median } (\hat{p}_i) + 3 \; Q_n \; (\hat{p}_i) \tag{15}$$

Q_n is employed to improve the accuracy of the identification of HLPs. $Q_n = c\{|x_i - xj \; |; < j\}_{(k)}$ is a pair wise order statistic for all distance proposed by Rousseeuw and Driessen[19] where $k = {}^{h}C_2 \approx {}^{h}C_2/4$ and $h = [n/2] + 1$. They used c = 2.2219 as this value will provide a consistent estimator Q_n for gaussian data. If some values of $\hat{p}_i$ did not exceed C_{DRGP} then, the case with the least $\hat{p}_i$ will be returned to the estimation subset for re-computation of $\hat{p}_i$. The values of generalized potential based on final 'D' set is DRGP(ISE) represented by $\hat{p}_i$ and the 'D' points will be declared as HLPs. The modified generalized studentized residuals (MGt) proposed by Mohammed et al. (2015) is given by,

$$MGt_i = \begin{cases} \frac{\widehat{e}_{i(R^*)}}{\widehat{\sigma}_{R^*-1}\sqrt{1-h^{**}_{i(R^*)}}}, & for \; i \in R^* \\ \frac{\widehat{e}_{i(R^*)}}{\widehat{\sigma}_{R^*}\sqrt{1+h^{**}_{i(R^*)}}}, & for \; i \notin R^* \end{cases} \tag{16}$$

where $\widehat{e}_{i(R^*)}$, $\widehat{\sigma}_{(R^*)}$ are the OLS residuals and residuals standard error for remaining set R, respectively. The observations are called influential observation when their values of MGti greater than its cut-off point (C_{MGti}). The C_{MGti} is calculated as follows:

$$C_{MGti} = \text{median}\,(MGti) + c\text{MAD}(MGti) \tag{17}$$

To classify HLPs, we plot MGt versus DRGP(ISE) and follows the procedure given by Mohammed et al. (2015) of classification of HLPs.

1. Regular observation (RO): An observation is declared as RO if ;

$|MGti| \leq C_{MGti}$ and $|DRGPi| \leq C_{DRGPi}$

2. Vertical outlying observation (VO): An observation is declared as VO if ;

$|MGti| > C_{MGti}$ and $|DRGPi| \leq C_{DRGPi}$

3. Good leverage observation (GLO): An observation is declared GLO if ;

$|MGti| \leq C_{MGti}$ and $|DRGPi| > C_{DRGPi}$

4. Bad leverage observation (BLO): An observation is declared BLO if ;

$|MGti| > C_{MGti}$ and $|DRGPi| > C_{DRGPi}$

This proposed method (MGt-DRGP$_{ISE}$) down weight only BLOs and employed RHCCM estimation methods discussed in section **3** to obtain the RHCCM estimator based on MGt-DRGP$_{ISE}$ weighting method denoted by WLS$_{FMGt}$.

6. Monte Carlo Simulation Study

In this section, we use monte carlo simulation to assess the performance of our new proposed methods under a heteroscedasticity of unknown form in linear regression model. Following Lima et al. (2016) simulation procedure, we consider a linear relation $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, $i = 1, 2, \ldots, n$. Three explanatory variables (x_1, x_2, x_3) are generated from standard normal distribution, in which the true parameters were set at $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, and $\varepsilon_i \sim N\left(0, \sigma_i^2\right)$. The strength of heteroscedasticity is measured by $\lambda = \max\left(\sigma_i^2\right) / \min(\sigma_i^2)$. Three sample sizes $n = 25$, 50 and 100 were replicated twice to form sample sizes of 50, 100 and 200, respectively. The skedastic function is defined as $\sigma_i^2 = \exp\{c_1 x_{i1}\}$ (Lima et al., 2016) where the value of $c_1 = 0.25$ was chosen such that $\lambda \approx 27$ and will be constant among the sample sizes. The value of λ indicates the degree of the heteroscedasticity in the data, whereby for homoscedasticity the value of $\lambda = 1$. The regular observations are generated according to standard normal ($x_i \sim N\,(0,\ 1)$), a certain percentage of regular observations were replaced by $N(10{,}1)$ observations in X and y at

different percentages level " $100\alpha\%$" (α = 0, 0.05, 0.10) of contamination for all the sample sizes considered at the average of 2,000 replications.

Table 1-4 exhibits the results of the proposed method together with other methods. A good method is one that has the lowest value of the standard error of estimates, bias, and variance of HC4 and HC5. It can be seen from Table 1 that for clean simulated heteroscedastic data (without contamination) the performance of all methods is reasonably closed to each other. However, for heteroscedastic data with HLPs (Tables 2-4).

The proposed WLS_{FMGt} method based on HC4 and HC5 outperformed the existing methods as evident by having the smallest standard error of estimates. The WLS_{FMGt} also provides a smallest bias which result to the coefficient of estimates that is closest to the true coefficient. The results which are based on HC4 are fairly closed to the results which are based on HC5. The standard error of the estimates will only be good and efficient when the form of heteroscedasticity is known. In this case when the structure of heteroscedasticity is unknown the estimation will lie on the HCCM estimator based on the employed HC4 and HC5 methods in which their results are very close to each other. Nonetheless, the OLS is much affected by HLPs followed by the $\mathrm{WLS}_{RMD}$and $\mathrm{WLS}_{F.}$

The results clearly indicate the robustness of WLS_{FMGt}over the rest of the methods. It can be concluded that the $\mathrm{WLS}_{FMGt}$is better and more efficient then $\mathrm{WLS}_{RMD,}\mathrm{WLS}_{F}$and OLS in the estimation of heteroscedastic model in the presence of HLPs in a data set.

Table 1: Regression estimates of the simulated data for n = 100.

Con. Level	Estimator		Coeff. of Estimates	Standard Error of Estimates	Bias	Variance	
						HC4	HC5
0 % HLPs	OLS	b_0	1.0523	0.6394	-0.0523	0.4564	0.7099
		b_1	1.0447	0.6645	-0.0447	0.6402	0.5238
		b_2	1.0227	0.6006	-0.0227	0.7884	0.6246
		b_3	1.0458	0.6128	-0.0458	0.3158	0.8730
	WLS_F	b_0	1.0628	0.5565	-0.0628	0.5408	0.5408
		b_1	1.0132	0.5891	-0.0132	0.2264	0.2264
		b_2	1.0482	0.5565	-0.0482	0.7324	0.7324
		b_3	0.9729	0.5597	0.0271	0.5418	0.5418
	WLS_{RMD}	b_0	1.0514	0.5394	-0.0514	0.4945	0.4945
		b_1	1.0237	0.5646	-0.0237	0.6845	0.6845
		b_2	1.0198	0.5176	-0.0197	0.9525	0.9525
		b_3	1.0290	0.5221	-0.0290	0.4974	0.4974
	WLS_{FMGt}	b_0	1.0135	0.4137	-0.0135	0.5106	0.5106
		b_1	1.0069	0.4270	-0.0069	0.4315	0.4315
		b_2	1.0142	0.4334	-0.0142	0.0690	0.0690
		b_3	0.9948	0.4312	0.0052	0.0376	0.0376

Table 2: Regression estimates of the simulated data for n = 50.

Con. Level	Estimator		Coeff. of Estimates	Standard Error of Estimates	Bias	Variance	
						HC4	HC5
5 % HLPs	OLS	b_0	0.5975	1.8466	0.4025	1.7365	2.3910
		b_1	1.1147	0.8288	-0.1147	0.6537	1.8383
		b_2	0.5483	1.7505	0.4517	3.5641	5.5306
		b_3	0.7315	1.7746	0.2685	3.2291	4.8675
	WLS_F	b_0	0.8168	1.5498	0.1832	1.4042	1.4042
		b_1	1.0836	0.9950	-0.0836	0.9526	0.9526
		b_2	0.2240	1.5407	0.7760	2.9485	2.9485
		b_3	0.9080	1.5432	0.0920	2.7190	2.7190
	WLS_{RMD}	b_0	0.6996	1.7490	0.3004	1.9700	1.9700
		b_1	1.1239	1.1207	-0.1239	1.1956	1.1956
		b_2	0.2649	1.6899	0.7351	4.3665	4.3665
		b_3	0.8456	1.7039	0.1544	3.9115	3.9115
	WLS_{FMGt}	b_0	0.9997	0.5569	0.0003	0.3970	0.4430
		b_1	0.9990	0.2521	0.0010	0.5472	0.9522
		b_2	0.9957	0.5934	0.0043	0.6297	0.7111
		b_3	1.0044	0.5954	-0.0044	0.6310	0.7046
10 % HLPs	OLS	b_0	1.0330	1.6102	-0.0330	1.4716	1.8409
		b_1	1.0198	0.4874	-0.0198	0.3973	0.5950
		b_2	0.6523	1.5128	0.3477	2.7754	3.7858
		b_3	1.1080	1.5245	-0.1080	2.6763	3.5727
	WLS_F	b_0	1.0247	1.4394	-0.0247	1.3473	1.3473
		b_1	1.0238	0.4943	-0.0238	0.4115	0.4115
		b_2	0.8151	1.3957	0.1849	2.6340	2.6340
		b_3	1.1127	1.4031	-0.1127	2.5523	2.5523
	WLS_{RMD}	b_0	1.0244	1.5469	-0.0244	1.7646	1.7646
		b_1	1.0345	0.6885	-0.0345	0.5900	0.5900
		b_2	0.6341	1.4907	0.3659	3.6167	3.6167
		b_3	1.1317	1.5075	-0.1317	3.4076	3.4076
	WLS_{FMGt}	b_0	0.9890	0.5833	0.0110	0.3899	0.3899
		b_1	0.9973	0.1754	0.0027	0.1955	0.1957
		b_2	0.9940	0.6048	0.0060	0.6350	0.6354
		b_3	1.0242	0.6044	-0.0242	0.6352	0.6353

Table 3: Regression estimates of the simulated data for n = 100.

Con. Level	Estimator		Coeff. of Estimates	Standard Error of Estimates	Bias	Variance	
						HC4	HC5
5 % HLPs	OLS	b_0	1.0219	1.6861	0.0219	1.4375	1.6655
		b_1	0.9995	0.6880	0.0005	0.5862	0.8796
		b_2	2.2098	1.5964	-1.2098	3.5353	4.3942
		b_3	1.7271	1.6427	-0.7271	3.3059	4.1136
	WLS_F	b_0	0.9998	1.3423	0.0002	1.0297	1.0297
		b_1	1.0464	0.7609	-0.0464	0.5909	0.5909
		b_2	1.5927	1.3166	-0.5927	2.4610	2.4610
		b_3	1.3311	1.3492	-0.3311	2.3307	2.3307
	WLS_{RMD}	b_0	0.9990	1.6090	0.0010	1.5361	1.5361
		b_1	1.0974	0.9469	-0.0974	0.9077	0.9077
		b_2	2.0654	1.5469	-1.0654	3.9607	3.9607
		b_3	1.6527	1.5865	-0.6527	3.6748	3.6748
	WLS_{FMGs}	b_0	1.0059	0.3906	-0.0059	0.2479	0.2482
		b_1	0.9957	0.1596	0.0043	0.1667	0.1772
		b_2	0.9983	0.4118	0.0017	0.3906	0.3935
		b_3	1.0017	0.4139	-0.0017	0.3915	0.3943
10 % HLPs	OLS	b_0	1.2386	1.5798	-0.2386	1.4261	1.6702
		b_1	1.0037	0.4741	-0.0037	0.3739	0.4715
		b_2	1.0316	1.4924	-0.0316	3.0567	3.7482
		b_3	0.9478	1.5028	0.0522	3.2812	4.0674
	WLS_F	b_0	1.1314	1.2960	-0.1314	1.0617	1.0617
		b_1	1.0092	0.4425	-0.0092	0.2957	0.2957
		b_2	1.0396	1.2638	-0.0396	2.2962	2.2962
		b_3	0.9412	1.2718	0.0588	2.4088	2.4088
	WLS_{RMD}	b_0	1.2042	1.4917	-0.2042	1.5576	1.5576
		b_1	1.0132	0.6584	-0.0132	0.4970	0.4970
		b_2	1.1029	1.4489	-0.1029	3.5068	3.5068
		b_3	0.8949	1.4601	0.1051	3.7089	3.7089
	WLS_{FMGs}	b_0	1.0101	0.4090	-0.0101	0.2545	0.2545
		b_1	0.9865	0.1224	0.0135	0.1124	0.1124
		b_2	1.0088	0.4215	-0.0088	0.4044	0.4045
		b_3	1.0092	0.4226	-0.0092	0.4031	0.4032

Table 4: Regression estimates of the simulated data for n = 200.

Con. Level	Estimator		Coeff. of Estimates	Standard Error of Estimates	Bias	Variance	
						HC4	HC5
	OLS	b_0	1.3482	1.2366	-0.3482	1.2390	1.3701
		b_1	0.9033	0.5053	0.0967	0.3740	0.4477
		b_2	1.1001	1.2204	-0.1001	2.2709	2.5126
		b_3	1.5135	1.1950	-0.5135	3.1822	3.5853
	WLS_F	b_0	1.1377	0.9441	-0.1377	0.7613	0.7613
5 % HLPs		b_1	0.9835	0.5407	0.0165	0.3203	0.3203
		b_2	1.0510	0.9604	-0.0510	1.4912	1.4912
		b_3	1.1242	0.9473	-0.1242	1.8965	1.8965
	WLS_{RMD}	b_0	1.2904	1.1851	-0.2904	1.2652	1.2652
		b_1	0.9395	0.6992	0.0605	0.5133	0.5133
		b_2	1.1172	1.1864	-0.1172	2.3594	2.3594
		b_3	1.4092	1.1658	-0.4092	3.3192	3.3192
	WLS_{FMGt}	b_0	0.9852	0.2685	0.0148	0.1697	0.1698
		b_1	1.0004	0.1102	-0.0004	0.0976	0.1013
		b_2	0.9730	0.2863	0.0270	0.2662	0.2670
		b_3	0.9805	0.2857	0.0195	0.2670	0.2678
	OLS	b_0	0.8707	0.9112	0.1293	0.8084	0.8596
		b_1	0.9689	0.2728	0.0311	0.2319	0.2610
		b_2	0.3923	0.8742	0.6077	1.8714	2.0404
		b_3	1.1659	0.8619	-0.1659	1.7460	1.8900
	WLS_F	b_0	0.9116	0.7598	0.0884	0.6161	0.6161
		b_1	0.9911	0.2592	0.0089	0.1725	0.1725
		b_2	0.6651	0.7490	0.3349	1.3629	1.3629
		b_3	1.1726	0.7382	-0.1726	1.3141	1.3141
10 % HLPs	WLS_{RMD}	b_0	0.8692	0.8652	0.1308	0.8457	0.8457
		b_1	0.9738	0.3820	0.0262	0.2843	0.2843
		b_2	0.4271	0.8556	0.5729	1.9366	1.9366
		b_3	1.2272	0.8429	-0.2272	1.8616	1.8616
	WLS_{FMGt}	b_0	0.9945	0.2664	0.0055	0.1737	0.1737
		b_1	0.9905	0.0795	0.0095	0.0704	0.0704
		b_2	0.9743	0.2761	0.0257	0.2673	0.2674
		b_3	0.9724	0.2734	0.0276	0.2695	0.2696

7. Numerical Example

In this section, we consider artificial data sets to assess the performance of the proposed weighting method (MGt-DRGP$_{ISE}$) in robust heteroscedasticity consistent covariance matrix (RHCCM) estimator

7.1 Artificial Data Set

An artificial heteroscedastic data of 100 observations was generated. Following Lima et al. (2016) simulation procedure, three explanatory variables were generated with n=50 from uniform distribution $\sim$ U(5,30) in order have average values of 20 and replicated twice to form the sample 100 each for x_1, x_2 *and* x_3. The response variable is given by; $y_i = 1+x_{i1}+x_{i2}+x_{i3}+\varepsilon_i$ with $\varepsilon_i \sim N\left(0, \sigma_i^2\right)$. The strength of heteroscedasticity is measured by $\lambda = \max\left(\sigma_i^2\right) / \min(\sigma_i^2)$. The skedastic function is defined as $\sigma_i^2 = \exp\{c_1 x_{i1}\}$ Lima et al. (2016) where the value of $c_1 = 0.15$ was chosen such that $\lambda \approx 141.68$. The value of λ indicates the degree of the heteroscedasticity in the data, whereby for homoscedasticity the value of λ will be equal to 1. Figure 1 indicates that there is heteroscedasticity in the data set due to a systematic funnel shaped pattern observed in the first plot and the second plot shows that there is no HLPs in this data set.

The proposed and existing methods were evaluated based on the standard error of HC4 and HC5. Table 5 shows the result of uncontaminated heteroscedastic artificial data which indicates that all the methods performed fairly the same. However, the results of HC4 are fairly closed to the results which are based on HC5. The standard error of the estimates will only be good and efficient when the form of heteroscedasticity is known. In this case the structure of heteroscedasticity is unknown. So, the estimation will lie on the HCCM estimator based on the employed methods HC4 and HC5, which their results are very close to each other. Nonetheless, the OLS is much affected by HLPs followed by the $\text{WLS}_{RMD}$and WLS_F.

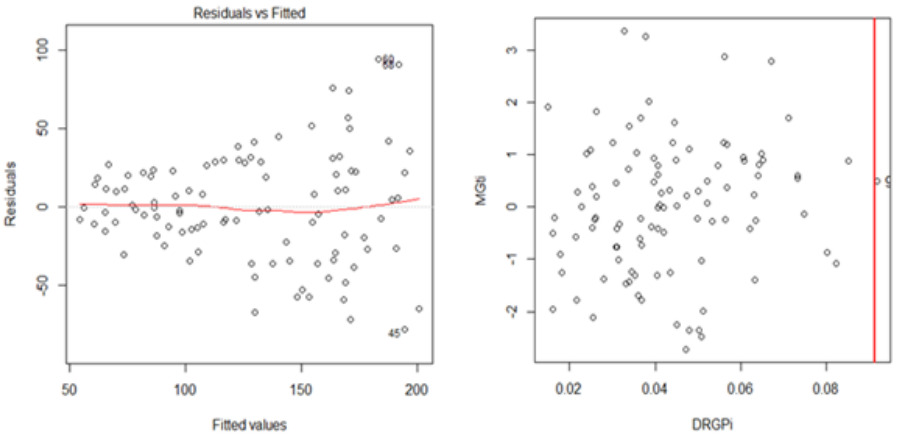

Figure 1: OLS residual vs fitted value and MGt vs DRGP for artificial data set

Table 5: Regression estimates for the artificial data set

Estimator		Coeff. of Estimates	Standard Error of Estimates	Standard Error	
				HC4	HC5
OLS	b_0	9.0649	14.6176	12.8353	12.8943
	b_1	5.4117	0.4649	0.4960	0.4979
	b_2	0.7480	0.5294	0.4858	0.4893
	b_3	0.5722	0.4509	0.4349	0.4379
WLS_F	b_0	9.0859	14.6174	12.8987	12.8987
	b_1	5.4139	0.4652	0.4982	0.4982
	b_2	0.7459	0.5296	0.4898	0.4898
	b_3	0.5704	0.4511	0.4383	0.4383
WLS_{RMD}	b_0	9.0649	14.6176	12.8943	12.8943
	b_1	5.4117	0.4649	0.4979	0.4979
	b_2	0.7480	0.5294	0.4893	0.4893
	b_3	0.5722	0.4509	0.4379	0.4379
WLS_{FMGt}	b_0	9.0649	14.6176	12.8943	12.8943
	b_1	5.4117	0.4649	0.4979	0.4979
	b_2	0.7480	0.5294	0.4893	0.4893
	b_3	0.5722	0.4509	0.4379	0.4379

7.2 Modified Artificial Data Set

The artificial data was modified by introducing two HLPs, the first and last observations were incremented by 10 for x_1 *and* x_2, respectively. The first plot in Figures 2 shows the presence of heteroscedasticity in the data due the funnel shape produced in the plot and and second indicated the presence of one GLO (observation number 100) and one BLO (observation number 1) in the data set. Table 6 shows the performance of the proposed (WLS_{FMGt}) and existing (WLS_{RMD}, WLS_F and OLS) methods in the modified artificial data. The result shows that WLS_{FMGt} has the least values of standard errors of HC4, HC5 and coefficient of estimate. This indicates that, the proposed method is more efficient and robust against the effect of bad leverage observations. The results clearly indicate the robustness of WLS_{FMGt} over the rest of the methods. It can be concluded that the WLS_{FMGt} is the most efficient method followed by WLS_F, WLS_{RMD} and OLS in the estimation of heteroscedastic model in the presence of HLPs in a data set.

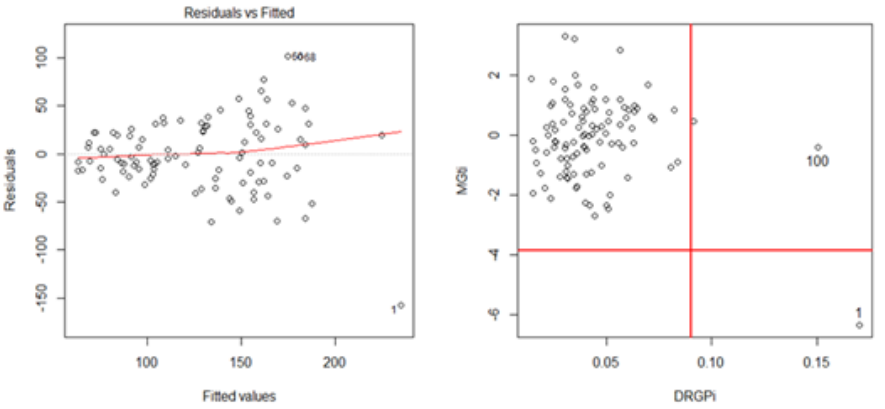

Figure 2: OLS residual vs fitted value and MGt vs DRGP for modified artificial data set

Table 6: Regression estimates for the modified artificial data set

Estimator		Coeff. of Estimates	Standard Error of Estimates	Standard Error	
				HC4	HC5
OLS	b_0	31.4707	14.3632	23.2959	27.9408
	b_1	6.8120	0.4952	0.6385	0.7156
	b_2	0.1742	0.5652	0.6584	0.7456
	b_3	0.3571	0.4939	0.5251	0.5853
WLS_F	b_0	22.5122	14.2502	17.8576	17.8576
	b_1	4.9999	0.4807	0.5606	0.5606
	b_2	0.3752	0.5467	0.5650	0.5650
	b_3	0.5030	0.4769	0.4672	0.4672
WLS_{RMD}	b_0	29.0537	14.2965	24.5806	24.5806
	b_1	4.8623	0.4908	0.6598	0.6598
	b_2	0.2284	0.5601	0.6828	0.6828
	b_3	0.3970	0.4891	0.5427	0.5427
WLS_{FMGt}	b_0	9.5442	13.8763	12.8817	12.8817
	b_1	5.1226	0.4667	0.5005	0.5005
	b_2	0.5093	0.5319	0.4996	0.4996
	b_3	0.6034	0.4630	0.4238	0.4238

8. Conclusion

This paper provides a robust method for estimating model parameters in linear regression when heteroscedasticity and high leverage points exist in a

data set. The proposed method WLS_{FMGt} down weight only bad leverage observations (BLOs) and allowed good leverage observations (GLOs) to contribute to the parameter estimation, as GLOs may contribute to the precision of the estimates.

The OLS method provides unbiased estimates in the presence of heteroscedasticity, but it is not efficient. The Furno's weighted least squares based on leverage weight function and RMD(MVE) are not efficient enough to remedy the problem of heteroscedastic errors with unknown structure and high leverage point. In this research, the weighting method which is based on MGt-DRGP_{ISE} is proposed to be incorporated in the weighted least squares and robust HCCM (HC4 and HC5) estimators. The WLS_{FMGt} was found to be the more efficient method as it provides the lowest bias, lowest standard errors of estimates, and lowest variance of HC4 and HC5 estimators.

Acknowledgement

The authors would like to thank Universiti Putra Malaysia for the financial support of this research (The GP-IRS Research Grant, Vot No. 9645900).

References

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, 45:215–233.

Cribari-Neto, F., Souza, T. C., and Vasconcellos, K. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics-Theory and Methods*, 36:1877–1888.

Cribari-Neto, F. and Zarkos, S. (2009). Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. *Econometric Reviews*, 18:211–228.

Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York.

Furno, M. (1996). Inference under heteroskedasticity and leveraged data. *Journal of Statistical Computation and Simulations*, 36.

Habshah, M., Norazan, M. R., and Imon, A. H. M. R. (2009). The performance of diagnostic-robust generalized potentials for the identification of

multiple high leverage points in linear regression. *Journal of Applied Statistics*, 36:507–520.

Imon, A. H. M. R. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Study*, 3:207–218.

Leroy, A. M. and Rousseeuw, P. (1987). *Robust regression and outlier detection Wiley series in probability and mathematical statistics*. Wiley, New York.

Lima, V. M. C., C.Souza, T., et al. (2016). Communications in statistics-simulation and computation. *Computational statistics*, 39:194–206.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 1:217–224.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample propertiesl. *Journal of Econometrics*, 29:305–325.

Mahalanobis, P. C. (2000). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample propertiesl. *Proceedings of the Indian National Science Academy*, 2:49–55.

Mohammed, A., Habshah, M., and Imon, A. H. M. R. (2015). A new robust diagnostic plot for classifying high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*, 1:27–42.

Pena, D. and Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix journal of the royal statistical society series b. *Statistical Methodology*, 57:145–156.

Rousseeuw, P. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 48:817–838.