Journal of Modern Applied Statistical Methods

Volume 17 | Issue 1 Article 17

6-29-2018

Robust Heteroscedasticity Consistent Covariance Matrix Estimator based on Robust Mahalanobis Distance and Diagnostic Robust Generalized Potential Weighting Methods in Linear Regression

M. Habshah
Universiti Putra Malaysia, habshahmidi@gmail.com

Muhammad Sani
Federal University, Dutsin-Ma, sanimksoro@gmail.com

Jayanthi Arasan
Universiti Putra Malaysia, jayanthi@upm.edu.my

Journal of Modern Applied Statistical Methods
May 2018, Vol. 17, No. 1, eP2596
doi: 10.22237/jmasm/1530279855

ISSN 1538 − 9472

Robust Heteroscedasticity Consistent Covariance Matrix Estimator based on Robust Mahalanobis Distance and Diagnostic Robust Generalized Potential Weighting Methods in Linear Regression

M. Habshah
Universiti Putra Malaysia
Selangor, Malaysia

Muhammad Sani
Federal University Dutsin-Ma
Dutsin-Ma, Nigeria

Jayanthi Arasan
Universiti Putra Malaysia
Selangor, Malaysia

The violation of the assumption of homoscedasticity and the presence of high leverage points (HLPs) are common in the use of regression models. The weighted least squares can provide the solution to heteroscedastic regression model if the heteroscedastic error structures are known. Based on Furno (1996), two robust weighting methods are proposed based on HLP detection measures (robust Mahalanobis distance based on minimum volume ellipsoid and diagnostic robust generalized potential based on index set equality (DRGP(ISE)) on robust heteroscedasticity consistent covariance matrix estimators. Results obtained from a simulation study and real data sets indicated the DRGP(ISE) method is superior.

Keywords: Linear regression, robust HCCM estimator, ordinary least squares, weighted least squares, high leverage points

Introduction

Ordinary least squares (OLS) is a widely used method for analyzing data in multiple regression models. The homoscedasticity assumption (i.e., equal variances of the errors) is often violated in most empirical analyses. As a result, the error variances tend to be heteroscedastic (unequal variances of the errors). Although OLS is still unbiased, its estimates become inefficient and will not provide reliable inference due to the inconsistency of the variance-covariance matrix estimator.

The commonly used estimation strategy for a heteroscedasticity of unknown form is to perform OLS estimation, and then employ a heteroscedasticity consistent

doi: 10.22237/jmasm/1530279855 | Accepted: August 11, 2017; Published: June 29, 2018.
Correspondence: Muhammad Sani, sanimksoro@gmail.com

covariance matrix (HCCM) estimator denoted by HC0 (see White, 1980). It is consistent under both homoscedasticity and heteroscedasticity of unknown form. The weakness of the HC0 estimator is it is biased in finite samples (MacKinnon & White, 1985; Cribari-Neto & Zarkos, 1999; Long & Ervin, 2000; Rana, Midi, & Imon, 2012). MacKinnon and White (1985) proposed another HCCM estimator referred to as HC1 and HC2. Davidson and MacKinnon (1993) slightly modified HC2 and named it HC3; it is closely approximated to the jackknife estimator. Cribari-Neto (2004) proposed another HCCM estimator where the residuals were adjusted by a leverage factor and called it HC4. Cribari-Neto, Souza, and Vasconcellos (2007) proposed HC5, wherein the exponent used in HC4 was modified to consider the effect of maximal leverage.

HCCM estimators are constructed using the OLS residuals vector. In the presence of outliers in the *X*-direction or high leverage points (HLPs), the coefficient estimates and residuals are biased. As a consequence, the inference becomes misleading. Furno (1996) proposed the robust heteroscedasticity consistent covariance matrix (RHCCM) in order to reduce the biased caused by leverage points. Residuals of a weighted least squares (WLS) regression were employed, where the weights were determined by the leverage measures (hat matrix) of the different observations. Lima, Souza, Cribari-Neto, and Fernandes (2009) built on Furno's procedure based on least median of squares (LMS) and least trimmed squares (LMS) residuals. A shortcoming of Furno's method is, in the presence of HLPs, the variances tend to be large resulting to unreliable parameter estimates which is due to the effect of swamping and masking of HLPs. The main reason for this weakness is the use of the hat matrix in determining the weight of the RHCCM algorithm of Furno (1996). Peña and Yohai (1995) showed swamping and masking results from the presence of HLPs in linear regression. It is evident that the hat matrix is not very successful in detecting HLPs (Habshah, Norazan, & Imon, 2009). Consequently, less efficient estimates are obtained by employing an unreliable method of detecting HLPs. Furno's work has motivated us to use a weight function based on a more reliable diagnostic measure for the identification of HLPs.

In this study, two new robust weighting methods are proposed based on HLPs detection measures; robust Mahalanobis distance based on minimum volume ellipsoid (RMD(MVE)) and diagnostic robust generalized potential based on index set equality (DRGP(ISE)) of Lim and Habshah (2016). The weights determined by DRGP(ISE) are expected to successfully down weight all HLPs. The DRGP(ISE) technique has been proven to be very successful in down weighting HLPs with low masking and swamping effects and less computational complexity, and the algorithm is very fast compared to DRGP(MVE).

Heteroscedasticity Consistent Covariance Matrix (HCCM) Estimators

Consider a regression model

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

where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of independent variables, $\boldsymbol{\beta}$ is a vector of regression parameters, and $\boldsymbol{\varepsilon}$ is the n-vector of random errors. For heteroscedasticity the errors are such that $E(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma_i^2$ for $i = 1,\ldots, n$, and $E(\varepsilon_i \varepsilon_s) = 0$ for all $i \neq s$. The covariance matrix of $\boldsymbol{\varepsilon}$ is given as $\boldsymbol{\Phi} = \text{diag}\{\sigma_i^2\}$. The ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$, which is unbiased, with the covariance matrix given by

$$\text{cov}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Phi}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{2}$$

However, under homoscedasticity, $\sigma_i^2 = \sigma^2$ which implies $\boldsymbol{\Phi} = \sigma^2\mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. The covariance matrix $\text{cov}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ is estimated by $\hat{\sigma}^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ (which is inconsistent and biased under heteroscedasticity), and $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/n - p$, $\hat{\boldsymbol{\varepsilon}} = \left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{y}$, where $\mathbf{H}$ is an idempotent and symmetric matrix known as a hat matrix, leverage matrix, or weight matrix (according to different authors). The hat matrix ($\mathbf{H}$) is defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and it plays great role in determining the HLPs in regression model. The diagonal elements hi = $x_i(x'x)^{-1}x_i'$ 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 $\hat{\varepsilon}_i^2$ in the covariance matrix of $\hat{\boldsymbol{\beta}}$, i.e.

$$\text{HC0} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Phi}}_0\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{3}$$

where, $\hat{\mathbf{\Phi}}_0 = \mathrm{diag}\left\{\hat{\varepsilon}_i^2\right\}$. HC0, HC1, HC2, and HC3 are generally biased for small sample size (see Furno, 1997; Lima et al., 2009; Hausman & Palmer, 2012). This paper will focus only on HC4 and HC5. The HC4 proposed by Cribari-Neto (2004) was built under HC3, and is defined as follows:

$$\mathrm{HC4} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Phi}}_4\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{4}$$

where

$$\hat{\mathbf{\Phi}}_4 = \mathrm{diag}\left\{\frac{\hat{\varepsilon}_i^2}{\left(1-h_i\right)^{\delta_i}}\right\}$$

for $i = 1,\ldots, n$ with $\delta_i = \min\{4, h_i/h\}$, which control the discount factor of the i^{th} squared residuals, given by the ratio between h_i and the average values of the h_i (h). Note that $\delta_i = \min\{4, nh_i/p\}$. Since $0 < 1 - h_i < 1$ and $\delta_i > 0$ it follows that $0 < \left(1-h_i\right)^{\delta_i} < 1$. The larger 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, $\delta_i = 4$ when $h_i > 4h = 4p/n$. The result obtained by Cribari-Neto (2004) suggested HC4 inference in finite sample size relative to HC3.

Similarly, another modification of the exponent $(1 - h_i)$ 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

$$\mathrm{HC5} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Phi}}_5\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{5}$$

where

$$\hat{\mathbf{\Phi}}_5 = \mathrm{diag}\left\{\frac{\hat{\varepsilon}_i^2}{\sqrt{\left(1-h_i\right)^{\alpha_i}}}\right\}$$

for $i = 1,\ldots, n$ with

$$\alpha_i = \min\left\{\frac{h_i}{h}, \max\left\{4, \frac{kh_{\max}}{h}\right\}\right\}$$

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 the h_i), and k is a constant $0 < k < 1$ and was suggested to be chosen as 0.7 by Cribari-Neto et al. (2007) following simulation results that lead to efficient quasi-t inference. When $h_i/h \leq 4$ it follows that $\alpha_i = h_i/h$. Also, since $0 < 1 - h_i < 1$ and $\alpha_i > 0$, it similarly follows that $0 < (1 - h_i)^{\alpha_i} < 1$.

Robust HCCM Estimators

The problems of heteroscedasticity and high leverage points were addressed by Furno (1996) to reduce the bias caused by the effect of leverage points in the presence of heteroscedasticity. It was suggested to use weighted least squares (WLS) regression residuals instead of the 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

$$\text{HC0}_{\text{W}} = (\mathbf{X'WX})^{-1}\mathbf{X'W}\hat{\mathbf{\Phi}}_{0\text{W}}\mathbf{WX}(\mathbf{X'WX})^{-1} \tag{6}$$

where $\mathbf{W}$ is an $n \times n$ diagonal matrix with

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

and c is the cutoff point $c = 1.5p/n$, p being the number of parameters in a model including the intercept and n the sample size, and $\hat{\mathbf{\Phi}}_{0\text{W}} = diag\{\tilde{\varepsilon}_i^2\}$ with $\tilde{\varepsilon}_i$ being the i^{th} residuals from weighted least squares (WLS). Note that non-leveraged observations are weighted by 1 and leveraged observations are weighted by c/h_i to reduce their intensity; w_i is considered to be the weight in this WLS regression, so that the WLS estimator of $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X'WX})^{-1}\mathbf{X'Wy} \tag{8}$$

The robust HCCM estimator for HC4 and HC5, based on Furno's weighting method considered by Lima et al. (2009), are HC4_{W} and HC5_{W}, defined as

$$\begin{aligned}\mathrm{HC4_W} &= \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'W}\hat{\mathbf{\Phi}}_{4w}\mathbf{WX}\left(\mathbf{X'WX}\right)^{-1} \\ \mathrm{HC5_W} &= \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'W}\hat{\mathbf{\Phi}}_{5w}\mathbf{WX}\left(\mathbf{X'WX}\right)^{-1}\end{aligned} \tag{9}$$

where

$$\hat{\mathbf{\Phi}}_{4\mathrm{W}} = \mathrm{diag}\left\{\frac{\tilde{\varepsilon}_i^2}{\left(1-h_i^*\right)^{\delta_i^*}}\right\}, \quad \hat{\mathbf{\Phi}}_{5\mathrm{W}} = \mathrm{diag}\left\{\frac{\tilde{\varepsilon}_i^2}{\sqrt{\left(1-h_i^*\right)^{\alpha_i^*}}}\right\}$$

for $i = 1,\ldots, n$, with

$$\delta_i^* = \min\left\{4, \frac{h_i^*}{h^*}\right\}, \quad \alpha_i^* = \min\left\{\frac{h_i^*}{h^*}, \max\left\{4, \frac{kh^*_{\max}}{h^*}\right\}\right\}$$

and h_i^* is the i^{th} diagonal element of the weighted hat matrix $\mathbf{H}_{\mathrm{W}} = \sqrt{\mathbf{W}}\mathbf{X}\left(\mathbf{X'WX}\right)^{-1}\mathbf{X'}\sqrt{\mathbf{W}}$. In this paper the Furno's weighted least square for RHCCM estimation method is denoted by WLSF.

New Proposed Robust HCCM Estimators

Consider the idea of Furno's RHCCM estimation on two new weighting methods based on HLPs identification measures: robust Mahalanobis distance based on minimum volume ellipsoid (RMD(MVE)) and diagnostic robust generalized potential based on index set equality (DRGP(ISE)). These two methods are very successful in identifying correct HLPs in a data set.

Robust HCCM Estimator based on RMD(MVE)

Mahalanobis (1936) introduced a diagnostic measure of the deviation of a data point from its center named Mahalanobis distance (MD), in which the independent variables of the i^{th} observations are presented as $\mathbf{x}_i = (1, x_{i1}, x_{i2},\ldots, x_{ik}) = (1, \mathbf{R}_i)$ so that $\mathbf{R}_i = (x_{i1}, x_{i2},\ldots, x_{ik})$ will be a k-dimensional row vector, where the mean and covariance matrix vector

$$\bar{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{R}_i, \quad \mathbf{CV} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{R}_i - \mathbf{R}\right)'\left(\mathbf{R}_i - \bar{\mathbf{R}}\right)$$

respectively. The MD for the i^{th} point is given as

$$\text{RMD}_i = \sqrt{\left(\mathbf{R}_i - \bar{\mathbf{R}}\right)'\left(\mathbf{CV}\right)^{-1}\left(\mathbf{R}_i - \bar{\mathbf{R}}\right)}, \quad i = 1, 2, \ldots, n \tag{10}$$

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

$$cd = \text{median}\left(\text{RMD}_i\right) + 3\,\text{MAD}\left(\text{RMD}_i\right) \tag{11}$$

where, MAD stands for median absolute deviation. Since the average vector $\bar{\mathbf{R}}$ and covariance matrix **CV** are not robust, Rousseeuw (1984) recommended using a minimum volume ellipsoid (MVE) estimator of $\bar{\mathbf{R}}$ and the corresponding **CV** which the ellipsoid produced. 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**) = center of the MVE covering at least h points of **X** for $h \geq (n + k + 1)/2$, where k is the number of explanatory variables (Rousseeuw & Driessen, 1999). The corresponding **CV** is provided by the ellipsoid and multiplied by a suitable factor in order to obtain consistency. The weight obtained by this RMD(MVE) method is given by

$$w_{ir} = \min\left(1, cd/\text{RMD}_i\right) \tag{12}$$

so that, HLPs are weighted by (cd/RMD_i) and non-leverage by 1. To obtain the RHCCM estimator based on RMD(MVE) weighting method denoted by WLSRMD, we replace equation (7) by (12) and adopt Furno's RHCCM estimation method as discussed above.

Robust HCCM estimator based on DRGP(ISE)

The diagnostic robust generalized potential based on minimum volume ellipsoid (DRGP(MVE)) was proposed by Habshah et al. (2009). It has been shown that this method is very successful in the detection of multiple HLPs in linear regression.

The method consists of two steps where suspects HLPs are identified in the first step by employing RMD based on MVE. The calculation of MVE involves a lot of computational effort. Due to this, the calculation of DRGP based on RMD-MVE takes too much computing time. As such, Lim and Habshah (2016) proposed an improvised DRGP based on index set equality (ISE) in order to reduce the computational complexity of the algorithm. The ISE was tested and found to execute much faster in the estimation of robust estimators of scale and location. Thus, ISE has faster running time compared to MVE (Lim & Habshah, 2016). They replaced the MVE estimator with the ISE to form DRGP(ISE).

Index set equality (ISE) was developed from the fast minimum covariance determinant (MCD) proposed by Rohayu (2013). The ISE idea is to denote the index set that corresponds to the sample of items in $\mathbf{H}_{\text{old}}$ when their Mahalanobis distance squares are arranged in ascending order by $\text{IS}_{\text{old}} = \left\{\pi_{(1)}^{\text{old}}, \pi_{(2)}^{\text{old}}, \ldots, \pi_{(h)}^{\text{old}}\right\}$ and the corresponding index set of the sample items in $\mathbf{H}_{\text{new}}$ by $\text{IS}_{\text{new}} = \left\{\pi_{(1)}^{\text{new}}, \pi_{(2)}^{\text{new}}, \ldots, \pi_{(h)}^{\text{new}}\right\}$, where π is a permutation on $\{1, 2,\ldots, n\}$. The steps to compute ISE are as follows:

Step 1: Arbitrarily selecting a subset $\mathbf{H}_{\text{old}}$ containing different h observations.

Step 2: Compute the average vector $\bar{\mathbf{R}}_{\mathbf{H}_{\text{old}}}$ and covariance matrix $\mathbf{CV}_{\mathbf{H}_{\text{old}}}$ for all observations belonging to $\mathbf{H}_{\text{old}}$.

Step 3: Compute $d_{\text{old}}^2(i) = \left(\mathbf{R}_i - \bar{\mathbf{R}}_{\mathbf{H}_{\text{old}}}\right)' \mathbf{CV}_{\mathbf{H}_{\text{old}}}^{-1} \left(\mathbf{R}_i - \bar{\mathbf{R}}_{\mathbf{H}_{\text{old}}}\right)$ for $i = 1, 2,\ldots, n$.

Step 4: Arrange $d_{\text{old}}^2(i)$ for $i = 1, 2,\ldots, n$ in ascending order, i.e. $d_{\text{old}}^2(\pi(1)) \le d_{\text{old}}^2(\pi(2)) \le \ldots \le d_{\text{old}}^2(\pi(n))$

Step 5: Construct $\mathbf{H}_{\text{new}} = \{\mathbf{R}_{\pi(1)}, \mathbf{R}_{\pi(2)},\ldots, \mathbf{R}_{\pi(h)}\}$.

Step 6: If $\text{IS}_{\text{new}} \ne \text{IS}_{\text{old}}$, let $\mathbf{H}_{\text{old}} := \mathbf{H}_{\text{new}}$ and $\mathbf{CV}_{\mathbf{H}_{\text{old}}} := \mathbf{CV}_{\mathbf{H}_{\text{new}}}$, compute $\bar{\mathbf{R}}_{\mathbf{H}_{\text{new}}}$ and let $\bar{\mathbf{R}}_{\mathbf{H}_{\text{old}}} := \bar{\mathbf{R}}_{\mathbf{H}_{\text{new}}}$ and go back to step (3). Else, the process is stopped.

The DRGP(ISE) consists 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 possessing a low leverage point will be put back to the 'R' group. This technique si continued until all points of the 'D' group have been checked to confirm whether they can be referred as HLPs. The generalized potential is defined as follows:

$$\hat{p}_i = \begin{cases} h_i^{(-\mathrm{D})} & \text{for } i \in \mathrm{D} \\ \dfrac{h_i^{(-\mathrm{D})}}{1-h_i^{(-\mathrm{D})}} & \text{for } i \in \mathrm{R} \end{cases} \tag{13}$$

The cut-off point for DRGP is given by

$$cdi = \mathrm{median}\left(\hat{p}_i\right) + 3\mathrm{Q}_n\left(\hat{p}_i\right) \tag{14}$$

Q_n, a pairwise order statistic for all distance proposed by Rousseeuw and Croux (1993), is employed to improve the accuracy of the identification of HLPs and is given by $\mathrm{Q}_n = c\{|x_i - x_j|; < j\}_{(k)}$, where $\mathrm{k} = {}^{h}C_2 \approx {}^{h}C_2/4$ and $h = [n/2] + 1$. They make used of $c = 2.2219$, as this value will provides Q_n a consistent estimator for Gaussian data. If some identified $\hat{p}_i$ did not exceed cdi 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 the final 'D' set is the DRGP(ISE) represented by $\hat{p}_i$ and the 'D' points will be declared as HLPs. Following Furno (1996), the DRGP(ISE) weight can be obtained as follows:

$$w_{id} = \min\left(1, cdi/\hat{p}_i\right) \tag{15}$$

where the HLPs are weighted by ($cdi/\hat{p}_i$) and non-leverage by 1. We also replace equation (7) by (15) and employ RHCCM estimation methods discussed above to obtain the RHCCM estimator based on DRGP(ISE) weighting method denoted by $\mathrm{WLS}_{\mathrm{DRGP}}$. The procedure for DRGP(ISE) can be summarized in the following steps:

- Step 1: For every i^{th} point, use the ISE method to compute the RMD_i.
- Step 2: Any i^{th} case having $\mathrm{RMD}_i > \mathrm{median}(\mathrm{RMD}_i) + 3\mathrm{MAD}(\mathrm{RMD}i)$ is suspected to be HLP and is assigned to the deletion set (D); the other (remaining) cases are put into the set R.
- Step 3: Compute $\hat{p}_i$ as defined in equation (14) based on the sets R and D above.

Step 4: If all the deleted cases $\hat{p}_i$ > median($\hat{p}_i$) + 3Q$_n$($\hat{p}_i$), the respective cases are declared as HLPs. Otherwise, the case with least $\hat{p}_i$ will be return to set R and repeat steps (3) and (4) until all $\hat{p}_i$ > median($\hat{p}_i$) + 3Q$_n$($\hat{p}_i$)

Simulation Study

A Monte Carlo simulation is used to assess the performance of the proposed methods under a heteroscedasticity of unknown form in a linear regression model. Following the simulation procedure used by Lima et al. (2009), 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 a standard normal distribution in which the true parameters were set at $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$ and $\varepsilon_{\mathrm{i}} \sim \mathrm{N}(0, \sigma_i^2)$. The strength (degree) of heteroscedasticity is measured by $\lambda = \max(\sigma_i^2)/\min(\sigma_i^2)$. 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., 2009) where the value of $c_1 = 0.450$ was chosen such that $\lambda \approx 43$ 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 $\lambda = 1$. For each of $x_i \sim \mathrm{N}(0, 1)$, a certain percentage of HLPs were replaced randomly with N(20, 1) at 5%, 10%, and 20% contamination levels for all the sample sizes considered at the average of 10,000 replications.

Shown in Table 1 is the performance of the proposed and existing methods in a clean simulated heteroscedastic data. Presented in Tables 2-4 are the results of both proposed and existing methods for heteroscedastic data with HLP contamination. The tables indicate, in the presence of clean heteroscedastic data, all methods are reasonably closed to each other. However, in the presence of HLPs, the proposed $\mathrm{WLS_{DRGP}}$ method based on HC4 and HC5 outperformed the existing methods as evident by having the smallest standard error of estimate. The $\mathrm{WLS_{DRGP}}$ also provides the coefficient of estimates that is closest to the true coefficient. The results which are based on HC4 are fairly close 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 two methods employed, HC4 and HC5, in which their results are very close to each other. The standard error of $\mathrm{WLS_{DRGP}}$ is the smallest. followed by $\mathrm{WLS_{RMD}}$, $\mathrm{WLS_F}$, and OLS for a heteroscedastic model in the presence of HLPs in the data set. The result can be seen clearly from the % reduction of standard errors exhibited in the tables that our proposed methods consistently have the highest reduction of standard errors irrespective of sample sizes and contamination levels.

Table 1. Regression estimates of the simulated data for n = 200, λ = 43

Con. Level	Estimator		Coefficient of estimates	Standard error of estimates	Standard error HC4	Standard error HC5
0% HLPs	OLS	b_0	1.0007	0.1613	0.1582	0.1643
		b_1	1.0013	0.1658	0.1576	0.1625
		b_2	0.9991	0.1668	0.1604	0.1639
		b_3	1.0012	0.1669	0.1610	0.1626
	WLS_F	b_0	1.0008	0.1611	0.1599	0.1659
		b_1	1.0018	0.1708	0.1634	0.1634
		b_2	0.9990	0.1718	0.1668	0.1668
		b_3	1.0014	0.1719	0.1673	0.1663
	WLS_{RMD}	b_0	1.0009	0.1713	0.1626	0.1626
		b_1	1.0016	0.1781	0.1690	0.1640
		b_2	0.9990	0.1791	0.1699	0.1649
		b_3	1.0014	0.1792	0.1704	0.1654
	WLS_{DRGP}	b_0	1.0008	0.1714	0.1621	0.1621
		b_1	1.0013	0.1787	0.1674	0.1634
		b_2	0.9990	0.1799	0.1692	0.1652
		b_3	1.0012	0.1700	0.1697	0.1652

Table 2. Regression estimates of the simulated data for n = 50, λ = 43

Con. Level	Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	Standard error HC5	Standard error HC4 % red.	Standard error HC5 % red.
5% HLPs	OLS	b_0	0.9477	0.3021	0.2906	0.3130	-	-
		b_1	0.9287	0.2658	0.2740	0.3044	-	-
		b_2	0.9404	0.2558	0.2531	0.2810	-	-
		b_3	0.9440	0.2561	0.2542	0.2836	-	-
	WLS_F	b_0	0.9866	0.2254	0.2228	0.2228	23.3495	28.8235
		b_1	0.9716	0.2370	0.2398	0.2398	12.9496	41.7297
		b_2	0.9816	0.2269	0.2265	0.2265	10.5197	40.5483
		b_3	0.9820	0.2275	0.2289	0.2290	9.9175	40.3043
	WLS_{RMD}	b_0	0.9919	0.2080	0.2058	0.2059	29.1684	34.2240
		b_1	0.9891	0.2109	0.2174	0.2174	21.4390	47.4040
		b_2	0.9889	0.2010	0.2036	0.2037	19.5464	46.5320
		b_3	0.9836	0.2015	0.2048	0.2049	19.4148	46.5873
	WLS_{DRGP}	b_0	0.9983	0.1812	0.1833	0.1833	36.9103	41.4163
		b_1	0.9978	0.1972	0.1912	0.1912	31.3409	54.0439
		b_2	0.9977	0.1876	0.1837	0.1837	27.4254	51.7841
		b_3	0.9991	0.1883	0.1854	0.1854	27.0480	51.6593

Table 2 (continued).

Con. Level	Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	HC5	HC4 % red.	HC5 % red.
10% HLPs	OLS	b_0	0.9468	0.3142	0.3063	0.3011	-	-
		b_1	0.9104	0.2661	0.2611	0.2619	-	-
		b_2	0.9048	0.2567	0.2515	0.2514	-	-
		b_3	0.8926	0.2561	0.2504	0.2510	-	-
	WLS_F	b_0	0.9736	0.2380	0.2258	0.2258	26.2968	25.0285
		b_1	0.9703	0.2330	0.2312	0.2342	11.8895	10.9897
		b_2	0.9724	0.2235	0.2217	0.2247	11.8381	10.6380
		b_3	0.9700	0.2229	0.2216	0.2236	11.4785	10.8940
	WLS_{RMD}	b_0	0.9728	0.2149	0.2132	0.2132	30.4037	29.2060
		b_1	0.9734	0.2251	0.2246	0.2246	14.5361	14.8183
		b_2	0.9721	0.2158	0.2126	0.2126	15.4513	15.4444
		b_3	0.9785	0.2150	0.2115	0.2115	15.5345	15.7371
	WLS_{DRGP}	b_0	0.9931	0.1925	0.1880	0.1880	38.6319	37.5758
		b_1	0.9900	0.2050	0.1980	0.1980	25.1278	25.3751
		b_2	0.9962	0.1959	0.1884	0.1884	25.0729	25.0667
		b_3	0.9995	0.1948	0.1861	0.1861	25.6627	25.8410
20% HLPs	OLS	b_0	0.9199	0.3345	0.3263	0.3308	-	-
		b_1	0.8432	0.2859	0.2709	0.2879	-	-
		b_2	0.7903	0.2761	0.2510	0.2661	-	-
		b_3	0.8678	0.2767	0.2517	0.2671	-	-
	WLS_F	b_0	0.9760	0.2522	0.2288	0.2288	29.8802	30.8276
		b_1	0.9716	0.2550	0.2527	0.2527	14.6493	19.8747
		b_2	0.9713	0.2450	0.2234	0.2234	11.0042	16.0615
		b_3	0.9706	0.2459	0.2244	0.2244	10.8522	15.9935
	WLS_{RMD}	b_0	0.9724	0.2373	0.2161	0.2161	33.7858	34.6805
		b_1	0.9725	0.2459	0.2262	0.2262	17.1431	22.2158
		b_2	0.9728	0.2360	0.2152	0.2152	14.2775	19.1488
		b_3	0.9710	0.2366	0.2158	0.2158	14.2465	19.1919
	WLS_{DRGP}	b_0	0.9956	0.1972	0.1806	0.1806	44.6654	45.4131
		b_1	0.9905	0.2037	0.1978	0.1978	28.0204	32.4272
		b_2	0.9906	0.1939	0.1824	0.1824	27.3150	31.4454
		b_3	0.9897	0.1943	0.1829	0.1829	27.3248	31.5160

Note: b_0, b_1, b_2, and b_3 are the estimates and % red. indicates the percentage improvement of the corresponding method over the OLS method; that is why the OLS rows are blank for % reduction

Table 3. Regression estimates of the simulated data for n = 100, λ = 43

Con. Level	Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	HC5	HC4 % red.	HC5 % red.
5% HLPs	OLS	b_0	0.9439	0.1770	0.1430	0.1440	-	-
		b_1	0.9375	0.1850	0.1511	0.1506	-	-
		b_2	0.9477	0.1748	0.1458	0.1442	-	-
		b_3	0.9476	0.1745	0.1459	0.1442	-	-
	WLS_F	b_0	0.9881	0.1508	0.1206	0.1206	15.6307	16.2319
		b_1	0.9824	0.1665	0.1320	0.1320	13.5382	13.2177
		b_2	0.9897	0.1562	0.1265	0.1265	13.2387	12.2865
		b_3	0.9707	0.1559	0.1265	0.1265	13.3019	12.3180
	WLS_{RMD}	b_0	0.9808	0.1378	0.1118	0.1118	21.7814	22.3388
		b_1	0.9831	0.1531	0.1254	0.1254	18.1884	17.8851
		b_2	0.9834	0.1429	0.1185	0.1185	18.6717	17.7791
		b_3	0.9824	0.1426	0.1186	0.1186	18.6788	17.7560
	WLS_{DRGP}	b_0	0.9974	0.1269	0.0862	0.0862	39.7108	40.1404
		b_1	0.9978	0.1432	0.1169	0.1169	24.2198	23.9389
		b_2	0.9979	0.1331	0.1048	0.1048	28.1038	27.3147
		b_3	0.9977	0.1329	0.1051	0.1051	27.9511	27.1335
10% HLPs	OLS	b_0	0.9281	0.2192	0.1584	0.1591	-	-
		b_1	0.8713	0.2667	0.1567	0.1572	-	-
		b_2	0.9050	0.2166	0.1569	0.1508	-	-
		b_3	0.9081	0.2169	0.1569	0.1507	-	-
	WLS_F	b_0	0.9970	0.1682	0.1242	0.1242	21.5785	21.8933
		b_1	0.9808	0.1745	0.1286	0.1286	18.5056	18.7799
		b_2	0.9844	0.1695	0.1255	0.1255	19.9945	16.7447
		b_3	0.9878	0.1697	0.1256	0.1256	19.9786	16.7041
	WLS_{RMD}	b_0	0.9972	0.1437	0.1127	0.1127	28.8414	29.1270
		b_1	0.9863	0.1610	0.1194	0.1194	24.5589	24.8128
		b_2	0.9860	0.1561	0.1188	0.1188	24.2512	21.1744
		b_3	0.9863	0.1563	0.1188	0.1188	24.2660	21.1670
	WLS_{DRGP}	b_0	0.9978	0.1262	0.0916	0.0916	42.1637	42.3959
		b_1	0.9987	0.1360	0.1088	0.1088	31.5746	31.8049
		b_2	0.9971	0.1313	0.0949	0.0949	39.5106	37.0536
		b_3	0.9972	0.1314	0.0949	0.0949	39.4835	37.0072

Table 3 (continued).

Con. Level	Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	HC5	HC4 % red.	HC5 % red.
20% HLPs	OLS	b_0	0.9141	0.2940	0.1715	0.1722	-	-
		b_1	0.6384	0.3155	0.1806	0.1840	-	-
		b_2	0.8576	0.3107	0.1707	0.1738	-	-
		b_3	0.7624	0.3105	0.1705	0.1735	-	-
	WLS_F	b_0	0.9787	0.1725	0.1317	0.1317	23.1937	23.4976
		b_1	0.9769	0.1793	0.1425	0.1425	21.6934	23.2176
		b_2	0.9718	0.1745	0.1335	0.1335	21.7786	23.1501
		b_3	0.9718	0.1743	0.1334	0.1334	21.7348	23.0710
	WLS_{RMD}	b_0	0.9871	0.1630	0.1201	0.1201	29.9625	30.2396
		b_1	0.9849	0.1655	0.1336	0.1336	26.7203	28.1467
		b_2	0.9803	0.1606	0.1279	0.1279	25.0967	26.4101
		b_3	0.9799	0.1604	0.1274	0.1274	25.2752	26.5509
	WLS_{DRGP}	b_0	0.9981	0.1333	0.1201	0.1042	29.9625	39.4910
		b_1	0.9980	0.1439	0.1336	0.1126	26.7203	39.8862
		b_2	0.9936	0.1390	0.1279	0.1051	25.0967	39.5181
		b_3	0.9925	0.1388	0.1274	0.1047	25.2752	39.6350

Note: b_0, b_1, b_2, and b_3 are the estimates and % red. indicates the percentage improvement of the corresponding method over the OLS method; that is why the OLS rows are blank for % reduction

Table 4. Regression estimates of the simulated data for n = 200, λ = 43

Con. Level	Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	HC5	HC4 % red.	HC5 % red.
5% HLPs	OLS	b_0	0.9481	0.1750	0.1456	0.1461	-	-
		b_1	0.8214	0.1794	0.1488	0.1490	-	-
		b_2	0.9122	0.1719	0.1431	0.1476	-	-
		b_3	0.9146	0.1718	0.1438	0.1482	-	-
	WLS_F	b_0	0.9746	0.1421	0.1252	0.1252	13.9818	14.2649
		b_1	0.9678	0.1508	0.1340	0.1340	10.2815	10.4046
		b_2	0.9723	0.1438	0.1293	0.1293	9.6916	12.4048
		b_3	0.9727	0.1435	0.1297	0.1297	9.7932	12.4811
	WLS_{RMD}	b_0	0.9855	0.1368	0.1108	0.1108	23.8683	24.1189
		b_1	0.9802	0.1448	0.1176	0.1176	21.6724	21.7799
		b_2	0.9840	0.1396	0.1107	0.1107	22.6745	24.9976
		b_3	0.9855	0.1396	0.1122	0.1122	21.9857	24.3103
	WLS_{DRGP}	b_0	0.9972	0.1128	0.0920	0.0920	36.7822	36.9903
		b_1	0.9978	0.1136	0.1003	0.1003	33.7470	33.8380
		b_2	0.9985	0.1070	0.0984	0.0984	31.2174	33.2839
		b_3	0.9986	0.1073	0.0998	0.0998	30.5851	32.6534

Table 4 (continued).

Con. Level	Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	HC5	HC4 % red.	HC5 % red.
10% HLPs	OLS	b_0	0.9230	0.2287	0.1449	0.1452	-	-
		b_1	0.8535	0.2358	0.1589	0.1580	-	-
		b_2	0.8737	0.2270	0.1458	0.1488	-	-
		b_3	0.7836	0.2261	0.1469	0.1495	-	-
	WLS_F	b_0	0.9795	0.1542	0.1288	0.1288	11.1246	11.3354
		b_1	0.9660	0.1696	0.1385	0.1385	13.7211	13.1560
		b_2	0.9651	0.1572	0.1273	0.1273	12.7049	14.4391
		b_3	0.9661	0.1599	0.1288	0.1288	12.3139	13.8354
	WLS_{RMD}	b_0	0.9879	0.1484	0.1228	0.1228	15.2752	15.4761
		b_1	0.9714	0.1546	0.1334	0.1334	17.1485	16.6059
		b_2	0.9716	0.1427	0.1222	0.1222	16.1878	17.8529
		b_3	0.9720	0.1451	0.1229	0.1229	16.2987	17.7510
	WLS_{DRGP}	b_0	0.9944	0.1277	0.1162	0.1162	19.8140	20.0041
		b_1	0.9917	0.1369	0.1248	0.1248	22.9247	22.4199
		b_2	0.9908	0.1257	0.1186	0.1186	18.6353	20.2517
		b_3	0.9911	0.1272	0.1201	0.1201	18.2379	19.6566
20% HLPs	OLS	b_0	0.9065	0.3662	0.1729	0.1733	-	-
		b_1	0.6250	0.3334	0.1831	0.1856	-	-
		b_2	0.7048	0.3234	0.1747	0.1766	-	-
		b_3	0.7927	0.3230	0.1748	0.1768	-	-
	WLS_F	b_0	0.9607	0.1940	0.1458	0.1458	15.6485	15.8484
		b_1	0.9650	0.2086	0.1549	0.1549	16.3059	17.4608
		b_2	0.9655	0.1982	0.1466	0.1466	16.0547	16.9728
		b_3	0.9669	0.1978	0.1474	0.1474	15.6639	16.5973
	WLS_{RMD}	b_0	0.9775	0.1822	0.1403	0.1403	18.8577	19.0501
		b_1	0.9715	0.1933	0.1504	0.1504	18.8900	20.0092
		b_2	0.9754	0.1836	0.1420	0.1420	18.7135	19.6025
		b_3	0.9785	0.1828	0.1420	0.1420	18.7608	19.6599
	WLS_{DRGP}	b_0	0.9926	0.1634	0.1218	0.1218	29.5194	29.6864
		b_1	0.9919	0.1707	0.1337	0.1337	28.5727	29.5584
		b_2	0.9900	0.1613	0.1244	0.1244	28.7865	29.5654
		b_3	0.9916	0.1607	0.1259	0.1259	27.9588	28.7561

Note: b_0, b_1, b_2, and b_3 are the estimates and % red. indicates the percentage improvement of the corresponding method over the OLS method; that is why the OLS rows are blank for % reduction

Numerical Examples

The performance of the proposed WLS_{DRGP} and WLS_{RMD} methods are evaluated using education expenditure data and an artificial heteroscedastic data set. Firstly, consider education expenditure data taken from Chatterjee and Hadi (2006). It represents the relationship between per capita income on an education project from 1975 and three independent variables, namely per capita income in 1973 (x_1), number of residents per thousands under 18 years of age (x_2), and number of residents per thousands under 18 years of age in 1974 (x_3). The existing methods (OLS and WLS_F) and the new proposed methods (WLS_{RMD} and WLS_{DRGP}) were then applied to the data. The data were modified by introducing HLP contamination, in which the 2nd, 27th, and 40th observations were replaced by 1323, 817, and 1605 for x_2, x_1, x_3, respectively. As noted in Figures 1 and 2, both data sets have heteroscedastic errors due to the funnel shape produced by the residuals versus fitted values plot.

Shown in Tables 5 and 6 are the results of the education expenditure and modified education expenditure data. The results indicate the proposed WLS_{RMD} and WLS_{DRGP} outperformed the existing methods (in Table 6 in the presence of heteroscedasticity and HLPs) by providing very small standard errors and also having equal performances in Table 5, in situation where only the heteroscedasticity problem is present. The WLS_{DRGP} based on both HC4 and HC5 immediately appears to be the best of all the estimators by possessing the highest percentage of reduction from the OLS.

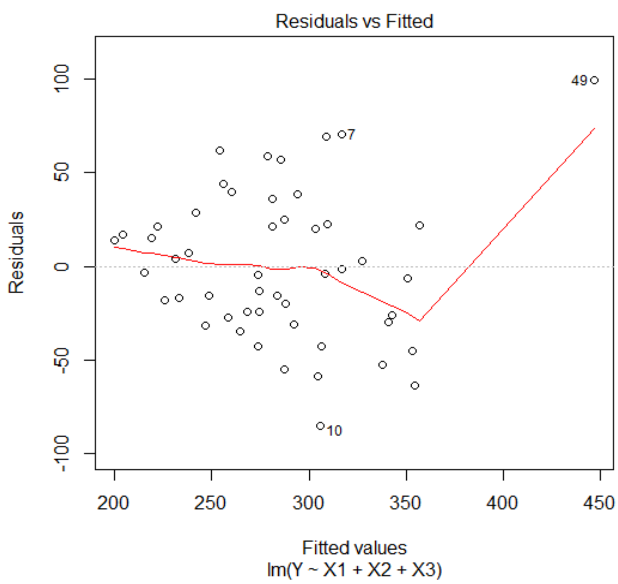

Figure 1. Plot of OLS residuals versus fitted values for education expenditure data

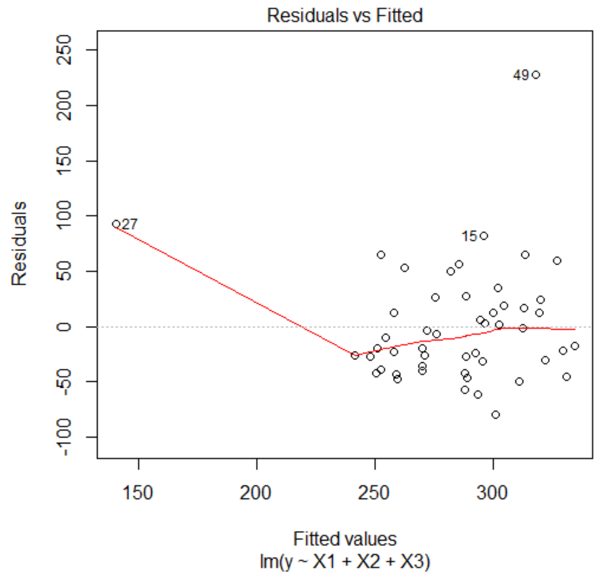

Figure 2. Plot of OLS residuals versus fitted values for modified educational expenditure data

Table 5. Regression estimates for the education expenditure data set

Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	HC5	HC4 % red.	HC5 % red.
OLS	b_0	-556.5680	123.1953	102.3823	13.7623	-	-
	b_1	0.0724	0.0116	0.0180	0.0254	-	-
	b_2	1.5521	0.3147	0.4765	0.1948	-	-
	b_3	-0.0043	0.0514	0.0623	0.1598	-	-
WLS_F	b_0	-375.7503	135.7155	0.0002	0.0002	99.9998	99.9706
	b_1	0.0591	0.0122	0.0124	0.0110	30.7942	53.7723
	b_2	1.1023	0.3502	0.0943	0.0933	80.2057	53.6256
	b_3	0.0337	0.0528	0.0517	0.0617	15.0874	57.6411
WLS_{RMD}	b_0	-485.2476	129.2253	0.0002	0.0002	99.9998	99.9703
	b_1	0.0673	0.0120	0.0118	0.0118	34.0634	56.3332
	b_2	1.3749	0.3331	0.0934	0.0924	80.3954	54.0359
	b_3	0.0096	0.0525	0.0643	0.0543	15.2662	59.7461
WLS_{DRGP}	b_0	-388.7580	134.6718	0.0002	0.0002	99.9998	99.9706
	b_1	0.0609	0.0122	0.0103	0.0105	42.4586	58.4868
	b_2	1.1327	0.3493	0.0907	0.0907	80.9705	56.4428
	b_3	0.0260	0.0528	0.0518	0.0508	16.9311	68.2448

Note: b_0, b_1, b_2, and b_3 are the estimates and % red. indicates the percentage improvement of the corresponding method over the OLS method; that is why the OLS rows are blank for % reduction

Table 6. Regression estimates for the modified education expenditure data set

Estimator		Coeff. of estimates	SE of estimates	Standard error			
				HC4	HC5	HC4 % red.	HC5 % red.
OLS	b_0	114.6463	350.0662	66.9948	60.6212	-	-
	b_1	0.0372	64.0101	32.0182	71.5216	-	-
	b_2	-0.0314	7.0530	142.0244	176.7519	-	-
	b_3	0.0130	41.0428	50.0503	147.7690	-	-
WLS_F	b_0	19.2925	270.4445	24.1656	24.1656	63.9292	60.1368
	b_1	0.0543	0.1228	0.0333	0.0333	99.8961	99.9535
	b_2	0.0790	1.2167	0.7182	0.7182	99.4943	99.5937
	b_3	-0.0206	0.5398	0.3408	0.3408	99.3191	99.7694
WLS_{RMD}	b_0	-10.9707	180.7190	13.1509	13.1509	80.3703	78.3065
	b_1	0.0460	0.1108	0.0263	0.0263	99.9177	99.9632
	b_2	0.2347	1.1806	0.6852	0.6852	99.5176	99.6124
	b_3	0.0055	0.4818	0.2982	0.2982	99.4042	99.7982
WLS_{DRGP}	b_0	-254.3590	121.1859	0.0003	0.0003	99.9995	99.9995
	b_1	0.0506	0.0108	0.0121	0.0121	99.9623	99.9831
	b_2	0.8825	0.3138	0.1257	0.1257	99.9115	99.9289
	b_3	0.0215	0.0466	0.0482	0.0482	99.9036	99.9674

Note: b_0, b_1, b_2, and b_3 are the estimates and % red. indicates the percentage improvement of the corresponding method over the OLS method; that is why the OLS rows are blank for % reduction

Secondly, an artificial heteroscedastic dataset of 100 observations were generated, where the explanatory and response variables were generated from yjr normal distribution N(20, 1) and $y_i = 1 + x_{i1} + x_{i2} + x_{i3} + \varepsilon_i$,, respectively. The heteroscedasticity was created in the same way as above, and the data was modified by introducing HLPs such that the 1st, 15th, and 70th observations were replaced by 41.0028, 40.6902, and 8.9320 for x_1, x_3, x_2, respectively. Both of Figures 3 and 4 show the presence of heteroscedasticity in the data due the funnel shape produced in the plots.

Presented in Tables 7 and 8 are the results of the artificial data and modified artificial data set, respectively. It can be observed from Table 7 that all estimators are equally good in the clean data set. Nonetheless, the OLS is much affected by HLPs, followed by the WLS_F and WLS_{RMD}.

The results indicate the superiority of WLS_{DRGP} over the rest of the methods. It can be concluded the WLS_{DRGP} is better and more efficient then WLS_{RMD}, WLS_F, and OLS in the estimation of heteroscedastic models in the presence of HLPs in a data set. As further research, we recommend investigating how this proposed methods work for both Type-I and Type-II errors using the quasi-*t* statistic.

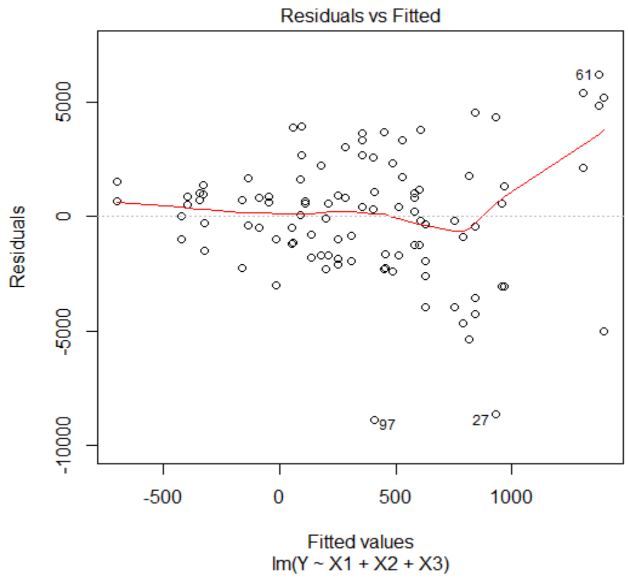

Figure 3. Plot of OLS residuals versus fitted values for artificial data

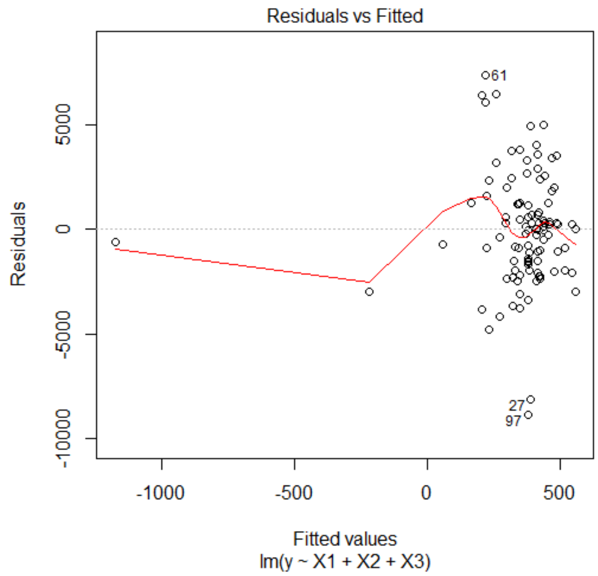

Figure 4. Plot of OLS residuals versus fitted values for modified artificial data

Table 7. Regression estimates for the artificial data set

Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	HC5	HC4 % red.	HC5 % red.
OLS	b_0	-5911.0888	9747.9701	1876.5408	1810.5194	-	-
	b_1	418.6609	351.5534	327.1687	369.4983	-	-
	b_2	-70.8100	426.8935	341.5515	352.7552	-	-
	b_3	-39.0783	394.3772	372.5343	383.8704	-	-
WLS_F	b_0	-2852.0091	9057.6309	1469.6061	1469.6061	21.6854	18.8296
	b_1	304.5275	350.0965	302.9163	302.9163	7.4128	18.0196
	b_2	-141.4352	416.5253	305.9416	305.9416	10.4259	13.2708
	b_3	-9.3324	384.7349	324.7982	324.7982	12.8139	12.3886
WLS_{RMD}	b_0	-5911.0888	8747.9732	1410.5194	1410.5194	24.8341	22.0931
	b_1	418.6609	346.5534	281.4983	281.4983	13.9593	23.8161
	b_2	-70.8100	406.8935	302.7552	302.7552	11.3588	14.1741
	b_3	-39.0783	364.3772	333.8704	333.8704	13.0629	13.0252
WLS_{DRGP}	b_0	-5911.0888	8707.9711	1410.5194	1410.5194	24.8341	22.0931
	b_1	418.6609	331.5534	279.4983	279.4983	14.5706	24.3574
	b_2	-70.8100	401.8935	292.7552	292.7552	14.2867	17.0090
	b_3	-39.0783	361.3772	321.8704	321.8704	13.5998	16.1513

Note: b_0, b_1, b_2, and b_3 are the estimates and % red. indicates the percentage improvement of the corresponding method over the OLS method; that is why the OLS rows are blank for % reduction

Table 8. Regression estimates for the modified artificial data set

Estimator		Coeff. of estimates	SE of estimates	Standard error HC4	HC5	HC4 % red.	HC5 % red.
OLS	b_0	2347.2014	13452.7878	4854.3334	4180.9672	-	-
	b_1	-32.9776	318.3895	548.5326	847.0869	-	-
	b_2	-75.1522	325.8351	558.1543	508.2788	-	-
	b_3	10.6393	393.9614	548.9331	756.4859	-	-
WLS_F	b_0	-4202.5395	8846.2733	2677.4813	2677.4813	44.8435	35.9602
	b_1	322.4820	219.3419	406.0620	436.0620	25.9730	48.5222
	b_2	-56.0296	278.6068	426.3988	426.3988	23.6056	16.1093
	b_3	-38.8055	274.4062	391.0439	391.0439	28.7629	48.3078
WLS_{RMD}	b_0	-2263.0811	8760.7035	2243.4518	2243.4518	53.7846	46.3413
	b_1	190.1914	187.8816	383.4138	383.4138	30.1019	54.7374
	b_2	-49.3891	227.0691	375.6060	375.6060	32.7057	26.1024
	b_3	-9.3389	239.9932	378.1985	378.1985	31.1030	50.0059
WLS_{DRGP}	b_0	-5719.2772	8202.9609	1463.4209	1463.4209	69.8533	64.9980
	b_1	415.0248	155.5852	292.8725	292.8725	46.6080	65.4259
	b_2	-67.9705	203.5646	282.0422	282.0422	49.4688	44.5103
	b_3	-39.0423	216.8809	299.3767	299.3767	45.4621	60.4253

Note: b_0, b_1, b_2, and b_3 are the estimates and % red. indicates the percentage improvement of the corresponding method over the OLS method; that is why the OLS rows are blank for % reduction

Conclusion

This research provides a better algorithm for estimating model parameters in linear regression when heteroscedasticity and high leverage points exist in a data set. Even though the OLS method provides unbiased estimates in the presence of heteroscedasticity, it is not efficient. The Furno's weighted least squares method based on a leverage weight function is also not efficient enough to remedy the problem of heteroscedastic errors with unknown form and high leverage point. Here, two weighting functions based on RMD and DRGP are proposed to be incorporated in the weighted least squares and Robust HCCM (HC4 and HC5) based estimators. The WLS_{DRGP} was found to be the best method as it's provides the lowest standard errors of HC4 and HC5, followed by the WLS_{RMD}, WLS_{F}, and OLS.

References

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example* (4th ed.). New York: Wiley. doi: 10.1002/0470055464

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, *45*(1), 215-233. doi: 10.1016/s0167-9473(02)00366-3

Cribari-Neto, F., Souza, T. C., & Vasconcellos, K. L. P. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics – Theory and Methods*, *36*(10), 1877-1888. doi: 10.1080/03610920601126589

Cribari-Neto, F., & Zarkos, S. (1999). Bootstrap methods for heteroskedastic regression models: Evidence on estimation and testing. *Econometric Reviews, 18*(2), 211-228. doi: 10.1080/07474939908800440

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

Furno, M. (1996). Small sample behavior of a robust heteroskedasticity consistent covariance matrix estimator. *Journal of Statistical Computation and Simulation, 54*(1-3), 115-128. doi: 10.1080/00949659608811723

Furno, M. (1997). A robust heteroskedasticity consistent covariance matrix estimator. *A Journal of Theoretical and Applied Statistics, 30*(3), 201-219. doi: 10.1080/02331889708802610

Habshah, M., Norazan, M. R., & 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*(5), 507-520. doi: 10.1080/02664760802553463

Hausman, J., & Palmer, C. (2011). Heteroskedasticity-robust inference in finite samples. *Economics Letters, 116*(2), 232-235. doi: 10.1016/j.econlet.2012.02.007

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

Leroy, A. M., & Rousseeuw, P. J. (1987). *Robust regression and outlier detection*. New York: Wiley.

Lim, H. A., & Habshah, M. (2016). Diagnostic robust generalized potential based on index set equality (DRGP(ISE)) for the identification of high leverage points in linear models. *Computational Statistics, 31*(3), 859-877. doi: 10.1007/s00180-016-0662-6

Lima, V. M. C., Souza, T. C., Cribari-Neto, F., & Fernandes, G. B. (2009). Heteroskedasticity- robust inference in linear regressions. *Communications in Statistics – Simulation and Computation, 39*(1), 194-206. doi: 10.1080/03610910903402572

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54*(3), 217-224. doi: 10.1080/00031305.2000.10474549

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*(3), 305-325. doi: 10.1016/0304-4076(85)90158-7

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science, India, 2*(1), 49-55.

Peña, D., & 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 (Methodological), 57*(1), 145-156. Available from https://www.jstor.org/stable/2346090

Rana, S., Midi, H., & Imon, A. H. M. R. (2012). Robust wild bootstrap for stabilizing the variance of parameter estimates in heteroscedastic regression models in the presence of outliers. *Mathematical Problems in Engineering, 2012*, 730328. doi: 10.1155/2012/730328

Rohayu, M. S. (2013). *A robust estimation method of location and scale with application in monitoring process variability* (Unpublished doctoral thesis). Universiti Teknologi Malaysia, Malaysia.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871-880. doi: 10.1080/01621459.1984.10477105

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273-1283. doi: 10.1080/01621459.1993.10476408

Rousseeuw, P. J., & Driessen, K. V. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212-223. doi: 10.1080/00401706.1999.10485670

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817-838. doi: 10.2307/1912934