On a Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers

Habshah Midi, Sohel Rana, and A.H.M.R. Imon

***Abstract*—The ordinary least squares (OLS) procedure is inefficient when the underlying assumption of constant error variances (homoscedasticity) is not met. As an alternative, we often used weighted least squares (WLS) procedure which requires a known form of the heteroscedastic errors structures, to estimate the regression parameters when heteroscedasticity occurs in the data. It is now evident that the WLS estimator is easily affected by outliers. To remedy the problem of heteroscedasticity and outliers simultaneously, we proposed a new method that we call two-step robust weighted least squares (TSRWLS) where prior information on the structure of the heteroscedastic errors is not required. The performance of the newly proposed estimator is investigated extensively by real data sets and Monte Carlo simulations.**

***Index Terms*— Heteroscedasticity, Monte Carlo simulation, Outliers, Two-step robust weighted least squares, Weighted least squares.**

I. Introduction

THE ordinary least squares (OLS) method is widely used to estimate the parameters of the linear regression model. Under the usual assumptions, the least-squares estimators possess many desirable properties. Among the assumptions, the assumption of constancy of error variances or homoskedasticity is difficult to achieve which causes the heterogeneity of error variances or heteroskedasticity. The main problem with the violation of homoskedaticity assumption is that the usual covariance matrix estimator of the OLS becomes biased and inconsistent.

There are abundant literatures which deal with heteroscedasticity problems [1-4]. The weighted least squares (WLS) is the most popular method to correct the problem of heteroscedasticity. Unfortunately, in practice, the form of heteroscedasticity is unknown, which makes the weighting approach impractical. When heteroscedasticity is caused by an incorrect functional form, it can be corrected by making variance-stabilizing transformations of the dependent variables [5-6] or by transforming both sides [7]. However, the transformation procedure might be complicated when dealing with more than one explanatory variable. Montgomery *et al*. [2], Kutner *et al.* [1], and others have tried to find the appropriate weight to solve the heteroscedastic problem when the form of heteroscedasticity is unknown. Chatterjee and Hadi [8] proposed an estimator which is weight based, but these weights depend on the known structure of the heteroscedastic data. Montgomery *et al*. [2] and Kutner *et al*. [1] proposed estimators which do not depend on the known structure of the heteroscedastic data. Hereafter we will refer to the Kutner *et al.* [1] estimator as KNN (Kutner, Nachtsheim and Neter) estimator. Nonetheless, the shortcoming of the Montgomery *et al*. [2] estimator is that it cannot be applied to more than one regressor situation. The advantage of the KNN estimator is that it can be applied to more than one variable and it does not depend on the known form of heteroscedasticity.

It is now evident that outliers can make the entire inferential procedure meaningless [7,9,10]. The KNN method can only remedy the problem of heteroscedasticity but not both problems of heteroscedasticity and outliers. Habshah *et al.* [11] has proposed robust estimation procedure to rectify both problems simultaneously, but their procedure can be applied to only one regressor. Not much work in the literature that devoted to estimation of the multiple regression parameters in the presence of both heteroscedasticity and outliers when the structure of heteroscedasticity is unknown. This has motivated us to propose a two-step robust weighted least squares (TSRWLS) estimator which is outlier resistant and at the same time can be applied to more than one regressor when the form of the heteroscedasticity is not known.

Manuscript received 6 March, 2013; revised 2 April, 2013.

Habshah Midi is with the Department of Mathematics/ Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia.(corresponding author: phone: 0603-8946-6606; fax: 0603-8943-7958; e-mail: habshahmidi@gmail.com).

Sohel Rana is with the Department of Mathematics/ Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia (e-mail: sohel@science.upm.edu.my).

A.H.M.R. Imon is with the Department of Mathematical Sciences, Ball State University, *Muncie, IN 47306, U.S.A.* (e-mail: rimon@bsu.edu).

II. Two-Step Robust Weighted Least Squares (TSRWLS)

Consider the general multiple linear regression model:

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

where $y = (y_1, y_2, \cdots, y_n)^T$ is an $n \times 1$ vector of response variable, $X = (x_1, x_2, \cdots, x_n)^T$ is an $n \times p$ fixed design matrix including the intercept, β is an $p \times 1$ vector of unknown linear parameters, and ε is an $n \times 1$ vectors of errors. The traditionally used OLS estimator of β is $\hat{\beta} = (X^T X)^{-1} X^T y$. It has mean β (i.e., it is unbiased) and covariance matrix

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

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

where $E(\varepsilon\varepsilon^T) = \Omega$, a positive definite matrix. Under homoscedasticity, we have $\Omega = \sigma^2 I_n$, and it follows that the $\text{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$, which can be estimated by $\hat{\sigma}^2 (X^T X)^{-1}$, where $\hat{\sigma}^2 = \hat{\varepsilon}^T \hat{\varepsilon} / (n-p)$, $\hat{\varepsilon} = (\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)$ being the *n*- vector of OLS residuals.

Under heteroscedasticity, that is, $\Omega = \sigma^2 Z$, where *Z* is a diagonal matrix, equation (2) becomes

$$V(\hat{\beta}) = \sigma^2 (X^T X)^{-1} X^T Z X (X^T X)^{-1} \quad (3)$$

Define $W = Z^{-1}$, where *W* is a diagonal matrix with diagonal elements or weights $w_1, w_2, ..., w_n$. It can be easily proved that the weighted least squares estimator is $\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W y$ and $\text{cov}(\hat{\beta}_{WLS}) = \sigma^2_{WLS} (X^T W X)^{-1}$. $\text{cov}(\hat{\beta}_{WLS})$ also can be estimated by $\hat{\sigma}^2_{WLS} (X^T W X)^{-1}$ where $\hat{\sigma}^2_{WLS} = \sum w_i \hat{\varepsilon}_i^2 / (n-p)$. It is not difficult to compute the weights of the *W* matrix, if the heteroscedastic error structure of the regression model is known. However, it is believed that the determination of weight is much affected by outliers and if not properly addressed, they will definitely affect the parameter estimation and other aspects of a weighted least squares analysis.

In this paper, our initial goal is to find an appropriate weight matrix *W* that should perform well in the presence of heteroscedasticity and outliers in which the heteroscedastic error structure is unknown. To find the robust weight matrix *W*, we propose a two-step robust weighted least squares (TSRWLS) estimator by adapting Kutner *et al* [1] and Habshah *et al*. [11] procedure. We use the LTS estimator, instead of the OLS in the KNN algorithm to get the initial robust weights. The TSRWLS consists of the following two steps. In step 1 we form the initial weight and in step 2 we obtain the final weight.

Step1:

(i) Find the fitted values $\hat{y}_i$ and the residuals $\hat{\varepsilon}_i$ from the regression model in equation (1), by using the least trimmed of squares (LTS) method.

(ii) Regress the absolute residuals, denoted as s_i where $s_i = |\hat{\varepsilon}_i|$, on $\hat{y}_i$ also by using the LTS method.

(iii) Find the fitted values $\hat{s}_i$ from step 1(ii).

(iv) The square of the inverse fitted values would form the initial robust weights, i.e., we obtain $w_{1i} = 1/(\hat{s}_i)^2$.

Step2:

The robust weighting function such as the Huber function [12], the Bisquare function [13] and the Hampel function [14] can be used to obtain the final weight. However, in this study, we will use the Huber's [12] weights function which is defined as

$$w_{2i} = \begin{cases} 1 & |\varepsilon_i| \le 1.345 \\ \dfrac{1.345}{|\varepsilon_i|} & |\varepsilon_i| > 1.345 \end{cases}$$

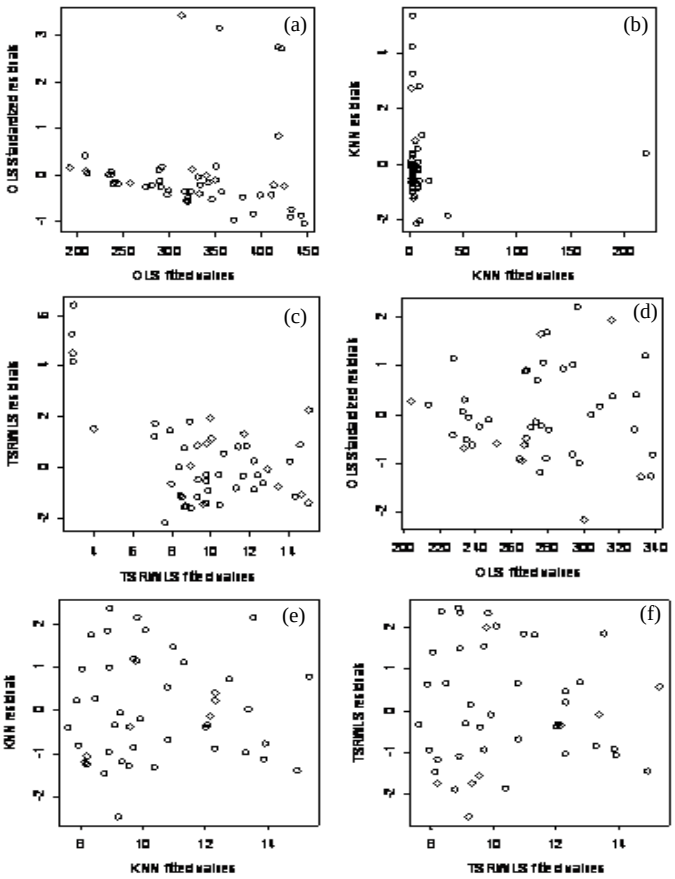
The constant 1.345 is called the tuning constant and ε_i is the *i*-th standardized residuals of the LTS obtained from step 1(i). We multiply the weight w_{1i} with the weight w_{2i} to get the final weight w_i. Finally we perform a WLS regression using the final weights w_i. The regression coefficients obtained from this WLS are the desired estimate of the heteroscedastic multiple regression model in the presence of outliers.

III. EXAMPLES

In this section, we consider a real data to evaluate the performance of the proposed TSRWLS method.

Education Expenditure Data

This data is taken from Chatterjee and Hadi [8] which consider the per capita income on education projected for 1975 as the response variable (*Y*) while the three explanatory variables are X_1, the per capita income in 1973; X_2, the number of residents per thousand under 18 years of age in 1974, and X_3, the number of residents per thousand living in urban areas in 1970 for all 30 states in USA. According to geographical regions based on the pre-assumption, the states are grouped in a sense that there exists a regional homogeneity. The four geographic regions (i) Northeast, (ii) North centre, (iii) South, and (iv) West. The LTS estimator detected that the observation 49 [Alaska (AK)] is an outlier. The residuals vs. fitted values of OLS (Standardized), KNN and TSRWLS are presented in Fig.1. Fig.'s 1(a)-1(c) display the residuals-fitted plots without considering Alaska. If the



ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Fig.1. The OLS, KNN and TSRWLS fitted values vs. residuals plots without AK, (a)-(c); with AK, (d)- (f)

variances of the error terms are constant then one can expect that the residuals are randomly distributed around zero residual, without showing any systematic pattern. Fig.1 (a) clearly shows that the OLS fit is inappropriate here, as there is a clear indication of heterogeneous error variances. However, Fig.1(b) and Fig.1(c) suggest that the KNN and TSRWLS fit are appropriate for this 'clean' data (without AK). We purposely include the observation Alaska to see the effect of outliers and the resulting residuals and fitted values are plotted in Fig.'s 1(d)-1(f). We see that OLS residuals are affected in the presence of outliers, but the effect of AK observation is not substantial on KNN and TSRWLS estimators.

Modified Education Expenditure Data

We then deliberately change four data points to generate big outliers to investigate the effect of multiple outliers. Our changed data points are cases 46, 47, 48 and 50 by taking the value from outside the well known 3-σ sigma normal distance in *Y* direction. We replace the data points of *Y* for observations 46, 47, 48 and 50 by $|y_{cont.}|$ where $y_{cont.}$ are generated as $\bar{y} \pm 9s_y$, with $\bar{y}$ and s_y as the respective

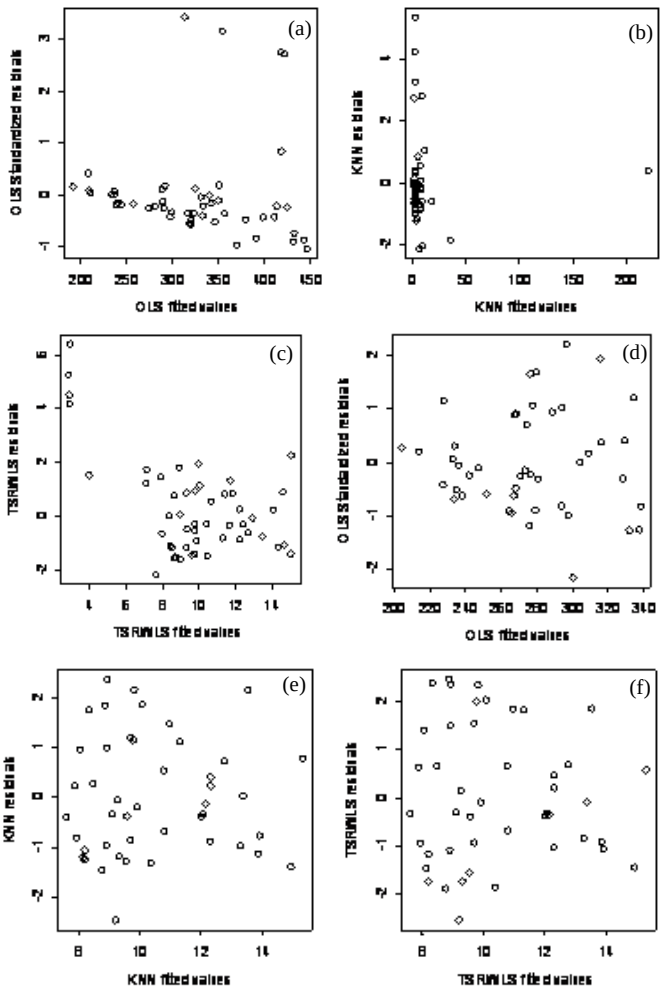


Fig. 2. The OLS, KNN and TSRWLS fitted values vs. residuals plots with 10% outliers, (a)-(c); without 10% outliers, (d)-(f).

mean and standard deviation of *Y*. With this modified data, now we have five outliers (since this data already contained one outlier, i.e., Alaska). When the LTS is employed to the data, all 5 outliers are identified.

Fig.'s 2(a)-2(f) present the plots of the residuals against the fitted values of the OLS, KNN and TSRWLS for the modified data. It is observed from Fig.'s 2(a) and 2(b) that the patterns of residuals are completely destroyed in the presence of outliers. That is, the OLS and KNN are greatly affected by outliers and so they are not good estimators for the remedy of the heteroscedastic problem when outliers are present. On the other hand, the TSRWLS in Fig. 2(c) shows the scatter plot of the residuals for the 'good' data except the data points which are outliers. Like as Fig.1, the residual-fitted plots without the 10% outliers for the OLS, KNN and the TSRWLS are shown in Fig.'s 2(d)-2(f). Fig. 2(d) signifies that the OLS cannot remedy the problem of heteroscedasticity but the KNN and the proposed TSRWLS are successful in this regard. From Fig's 2(a)-2(f), it suggest that the KNN is good in the absence of outliers whereas our proposed TSRWLS is good in the presence or absence of outliers.

Since graphical displays are always very subjective, we would like to present some numerical summaries of the examples considered above. Table 1 displays the summary statistics such as estimates of the parameters and their standard errors when there are no outliers, with only one outlier (AK), and with 5 outliers.

TABLE 1
REGRESSION ESTIMATES OF THE EDUCATION EXPENDITURE DATA

		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Without outliers	OLS	-277.5773	0.0483	0.8869	0.0668
	KNN	-334.4223	0.0550	0.9809	0.0599
	TSRWLS	-283.2395	0.0508	0.8827	0.0573
With AK outlier	OLS	-556.5680	0.0724	1.5521	-0.0043
	KNN	-423.7212	0.0620	1.1782	0.0519
	TSRWLS	-365.4785	0.0543	1.0779	0.0633
With multiple Outliers	OLS	-452.0702	0.0821	0.8200	0.1936
	KNN	-536.6901	0.1219	1.0639	-0.0983
	TSRWLS	-391.5358	0.0605	1.0815	0.0626
Standard Errors of Estimators					
Without outliers	OLS	132.4229	0.0121	0.3311	0.0493
	KNN	108.2248	0.0111	0.2642	0.0419
	TSRWLS	105.9811	0.0106	0.2732	0.0422
With AK outlier	OLS	123.1953	0.0116	0.3147	.0514
	KNN	96.8830	0.0107	0.2313	0.0405
	TSRWLS	102.6924	0.0105	0.2486	0.0402
With multiple Outliers	OLS	464.4632	0.0437	1.1864	0.1938
	KNN	182.0470	0.0204	0.4591	0.0397
	TSRWLS	161.8082	0.0170	0.3932	0.0630

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

In the absence of outliers, all estimators perform equally in terms of parameter estimates and their standard errors and the resulting values are relatively close. But things change dramatically when outliers are present in the data. All estimators except the TSRWLS are strongly affected by outlier(s). We observe that the OLS and the KNN estimators not only have more bias in comparison to the TSRWLS, but also the sign of $\hat{\beta}_{3OLS}$ and $\hat{\beta}_{3KNN}$ have been changed in some occasions. By looking at the results of standard errors it is clear that both the OLS and the KNN estimators break down easily even in the presence of a single outlier. They produce much higher standard errors as compared with the TSRWLS estimator and things deteriorate when multiple outliers are present in the data. It can be concluded from Table 1 that the proposed TSRWLS is the best overall estimator as it possesses less bias and standard errors as compared to other estimators in the presence of heteroscedasticity and outliers.

IV. SIMULATIONS

In this section, we report a Monte Carlo simulation study which is designed to compare the performance of the proposed TSRWLS estimator with the OLS, KNN and five versions of HCCM estimators. We re-use a design of Cribari-Neto [15]. In this simulation study the 'good' observations are generated according to linear regression model:

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

where $\varepsilon_i \sim N(0,1)$ and $E(\varepsilon_i \varepsilon_j) = 0 \forall i \neq j$. To generate a heteroscedastic regression model, we consider $\sigma_i^2 = \sigma^2 \exp(a x_{1i} + a x_{2i}^2)$ with $\sigma^2 = 1$ and a is an arbitrary constant. The covariate values are selected as random draws from the $U(0,1)$ distribution. The level of heteroscedasticity is measured as $\lambda = \max(\sigma_i^2)/\min(\sigma_i^2)$, i = 1,2,…,n. For each sample sizes we set a =.4 and a =.8, which yield $\lambda \approx 2$ and $\lambda \approx 4$, respectively. The values of the regression parameters used in the data generation scheme are $\beta_0 = \beta_1 = \beta_3 = 1$. Then we generate the contaminated model. At each step, one 'good' observation is substituted with an outlier. We focus on the situation where the errors are contaminated normal distribution. To generate a certain percentages of outliers, we use the regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_i \varepsilon_{i(cont.)}, \; i = 1,2, \ldots,n. \tag{5}$$

where $\varepsilon_{i(cont.)} \sim N(0,1) + Cauchy(0,10)$. The percentages of outliers can be varied. Since Cauchy is a longer tailed distribution, we are convinced that the contaminated normal errors would produce outliers. The values of the regression estimates $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ for the OLS, KNN, and TSRWLS methods are obtained when certain percentages of 'good' observations are replaced by outliers and these results are presented in Table 2.

It is observed that all these three estimators are fairly close to

TABLE 2
THE VALUES OF THE $\hat{\beta}_0, \hat{\beta}_1$ AND $\hat{\beta}_2$ FOR N=100

		λ =2			
Coeff	Outliers	OLS	KNN	TSRWLS	True Value
beta0	0%	0.51381	0.75832	0.78255	1
	10%	20.6403	7.95988	1.05986	1
	20%	6.23238	8.27183	0.74946	1
	30%	5.59469	7.07112	0.91234	1
	40%	2.00477	-3.0608	0.77635	1
	50%	74.5531	30.3342	-12.085	1
beta1	0%	1.23603	1.10946	1.16784	1
	10%	-4.4502	-13.545	0.95281	1
	20%	-13.74	-7.8405	1.13331	1
	30%	-16.945	-16.305	0.94242	1
	40%	-3.4652	16.6149	1.67675	1
	50%	-76.907	-59.5	18.0249	1
beta2	0%	1.78597	1.43032	1.20845	1
	10%	-6.1737	-12.353	1.16009	1
	20%	11.4026	-0.3606	0.76126	1
	30%	6.6363	5.35328	1.62855	1
	40%	-0.8861	-10.214	1.12264	1
	50%	-34.105	-29.176	6.23046	1
		λ =4			
Beta0	0%	0.59506	1.05948	0.92545	1
	10%	14.0779	1.81036	0.5168	1
	20%	6.398	8.44143	0.98938	1
	30%	3.40944	-3.8921	0.42918	1
	40%	-5.68	-3.9599	0.29887	1
	50%	-35.518	-34.901	-8.1603	1
beta1	0%	1.56987	0.78473	1.00136	1
	10%	-14.15	4.42855	0.95224	1
	20%	-0.8509	-6.6388	1.43201	1
	30%	-22.439	-23.168	1.25833	1
	40%	4.403	5.18559	1.9236	1
	50%	57.5132	45.9695	10.278	1
beta2	0%	1.53569	1.39424	1.28584	1
	10%	-12.651	-5.0678	1.4004	1
	20%	-5.0988	-2.6546	1.05175	1
	30%	18.2148	33.1469	1.14573	1
	40%	20.1455	5.29954	1.0685	1
	50%	-82.94	-49.466	7.94492	1

It is observed that all these three estimators are fairly close to the true values in the absence of outliers. However, the OLS estimates tend to move away from the true value rigorously, followed by the KNN with the increase in the percentage of outliers. The KNN can tolerate slightly over 1% outliers. But the TSRWLS appears as fairly robust in attaining almost the highest possible 50% break down for both $\lambda = 2$ and $\lambda = 4$. These results also confirm that the OLS and KNN estimators cannot retain their unbiased properties in the presence of outliers in heteroscedastic

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

model, whereas the proposed TSRWLS estimators can successfully retain the unbiasedness properties.

The breakdown properties of the OLS, KNN, and TSRWLS methods are investigated further by considering the samples of size 50, 100 and 150. Simulation studies were performed exactly in the same way as described earlier. 10,000 simulations are carried out using the S-Plus programming language. Summary values such as the mean estimated values

$$\bar{\beta}_j = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_j^{(k)}$$

are computed over m = 10,000 replications. This also yields the bias $\bar{\beta}_j - \beta_j$. The mean-squared error (MSE) is given by

$$MSE(\hat{\beta}_j) = (\hat{\beta}_j - \beta_j)^2 + \frac{1}{m}\sum_{k=1}^{m}(\hat{\beta}_j^{(k)} - \bar{\beta}_j)^2$$

Therefore, the root mean squared error (RMSE) is given by $[MSE(\hat{\beta}_j)]^{1/2}$. As a measure of robustness, we compute the 'relative measure of RMSE' which is the ratio of the RMSEs of the estimators compared with the least-squares estimators for good data. The relative bias and relative measure of RMSE of the OLS, KNN, and TSRWLS methods are presented in Tables 3 and Table 4 respectively. Due to space limitation only results for λ = 2 , n=50 and 100 are reported. But for the other sample sizes the results were consistent for λ = 2 and λ = 4. Several interesting points appear from Tables 3. For 'clean' data, all the three estimators considered here are fairly close to one another with respect to the values of the biasness measure. By inspecting the bias in Table 3, it is observed that the performance of both the OLS and the KNN tends to deteriorate with the increase in the percentage of outliers and they produce poor estimates at both levels ($\lambda \approx 2$ and $\lambda \approx 4$) of heteroscedasticity. The performance of the TSRWLS is very satisfactory here. In Table 4 values of robustness measures is justified by Relative measure of RMSE as compared to the OLS estimate with no outliers. We see that the relative efficiency of the all three methods is very satisfactory when there are no outliers. However, the scenario has changed when there are outliers. It is seen that in the presence of outliers the efficiency of the TSRWLS is very reasonable in this regard which not in the case of other two methods. Irrespective of the percentages of outliers it maintains producing low bias and small RMSE and does not break down before 5% contamination.

TABLE 3
BIASNESS MEASURE OF THE PARAMETERS OF THE DIFFERENT ESTIMATORS FOR λ =2

Outliers (%)	Estimators	Bias		
		OLS	KNN	TSRWLS
	Coeff.	Sample Size n = 50		
0%	beta0	-0.0059	-0.0013	-0.00198
	beta1	0.0140	0.0060	0.006226
	beta2	-0.0008	-0.0027	-0.00227
5%	beta0	-17.5758	-3.5790	-0.00329
	beta1	16.7337	3.2144	0.009711
	beta2	13.7552	2.6234	0.000918
10%	beta0	0.9357	0.1849	0.005338
	beta1	-0.3898	0.1293	-0.01026
	beta2	-2.1139	-1.1011	-0.00035
15%	beta0	-85.7342	-36.0054	0.011383
	beta1	82.7205	33.1180	-0.0172
	beta2	32.2882	9.1732	-0.00564
20%	beta0	-195.6400	-87.9323	-0.00422
	beta1	-372.1610	-146.0850	0.00136
	beta2	494.4786	181.1661	0.012381
		Sample Size n = 100		
0%	beta0	0.0005	-0.0018	-0.0005
	beta1	0.0004	0.0035	0.0011
	beta2	-0.0014	-0.0009	-0.0016
5%	beta0	1.9708	0.3723	-0.0009
	beta1	0.4748	0.1227	0.0005
	beta2	-3.6606	-0.7952	0.0001
10%	beta0	10.5640	1.3684	-0.0007
	beta1	-15.3713	-2.4119	0.0032
	beta2	-9.1550	-1.9725	-0.0043
15%	beta0	0.0707	0.8919	-0.0046
	beta1	0.9458	-1.4343	0.0033
	beta2	1.7874	0.9745	0.0045
20%	beta0	3.5876	-0.8170	-0.0050
	beta1	-5.4541	-0.7692	0.0059
	beta2	-7.0086	-1.8175	0.0014

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE 4
RELATIVE MEASURE OF RMSE OF THE PARAMETERS OF THE DIFFERENT ESTIMATORS FOR λ =2

Outliers(%)	Estimators	Relative measure of RMSE		
		OLS	KNN	TSRWLS
	Coeff.	Sample Size *n*= 50		
0%	beta0	–	103.8433	101.2776
	beta1	–	100.6190	96.7113
	beta2	–	99.9619	95.5360
5%	beta0	0.0265	0.1367	64.8578
	beta1	0.0439	0.2614	74.6672
	beta2	0.0544	0.3143	69.2143
10%	beta0	0.2806	1.2529	68.1788
	beta1	0.3632	1.3702	76.1437
	beta2	0.4208	1.0991	74.1417
15%	beta0	0.0049	0.0122	71.4086
	beta1	0.0079	0.0204	83.3345
	beta2	0.0155	0.0437	71.9914
20%	beta0	0.0028	0.0056	52.1659
	beta1	0.0016	0.0045	59.9413
	beta2	0.0015	0.0041	56.5258
Sample Size *n*= 100				
0%	beta0	–	103.0542	100.6648
	beta1	–	101.8250	98.3953
	beta2	–	101.6148	97.7441
5%	beta0	0.2752	1.0516	97.3127
	beta1	0.2010	0.6944	93.2256
	beta2	0.2165	0.9376	91.0025
10%	beta0	0.0342	0.2374	81.5563
	beta1	0.0339	0.1863	83.6645
	beta2	0.0460	0.2411	84.4919
15%	beta0	0.0970	0.3056	77.1535
	beta1	0.1534	0.3415	79.3901
	beta2	0.0827	0.3002	81.0387
20%	beta0	0.0605	0.1566	76.7077
	beta1	0.0733	0.2030	77.2481
	beta2	0.0487	0.1171	77.1561

V. CONCLUSIONS

In this article, we propose a two-step robust weighted least squares estimator which is designed for handling the problem of heteroscedasticity and outliers in multiple regression when the form of the heteroscedasticity is unknown. We have examined the performance of the proposed TSRWLS estimator and compare its performance with other existing estimators. Although the KNN and TSRWLS estimators are reasonably close to one another in the presence of heteroscedasticity with clean data, but the TSRWLS is the most reliable estimator as it possesses the least bias and standard errors. However, the performance of KNN and OLS are much inferior to the TSRWLS when contamination occurred in the data evident by having larger bias in estimates and standard errors, and smaller values of robustness measures.

REFERENCES

[1] M. H. Kutner, C. J. Nachtsheim, J. Neter, *Applied Linear Regression Models*, McGraw- Hill/Irwin, New York, 2004.
[2] D. C. Montgomery, E. A. Peck, G. G. Vining, *Introduction to Linear Regression Analysis*, Wiley, New York, 2001.
[3] P. M. Robinson, "Asymptotically Efficient Estimation in the Presence of Heteroscedasticity of Unknown form," *Econometrica*, 55 (1987), pp. 875-891.
[4] T.P. Ryan, *Modern Regression Methods*, Wiley, New York, 1997.
[5] R. D. Cook, S. Weisberg, "Diagnostics for heteroscedasticity in regression," *Biometrika*, 70 (1983), pp. 1-10.
[6] S. Weisberg, *Applied Linear Regression*, Wiley, New York , 1980.
[7] R. J. Carroll, D. Ruppert, *Transformation and Weighting in Regression*, Chapman and Hall, New York, 1988.
[8] S. Chatterjee, A.S. Hadi, *Regression Analysis by Examples*, Wiley, New York, 2006.
[9] R. A. Maronna, R. D. Martin, V.J. Yohai, *Robust Statistics -Theory and Methods*, Wiley, New York, 2006.
[10] P. J. Rousseeuw, A. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1997.
[11] M. Habshah, S. Rana, A. H. M. R. Imon, "The Performance of Robust Weighted Least Squares in the Presence of Outliers and Heteroscedastic," *WSEAS Transition of Mathematics*, 8 (2009), pp. 351 – 361.
[12] P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.
[13] J.W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishers, Cy, 1977.
[14] F. R. Hample, "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, 69 (1974), pp. 383-393.
[15] F. Cribari-Neto, "Asymptotic inference under heteroskedasticity of unknown form," *Computational Statistics and Data Analysis*, 45 (2004), pp. 215-233.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)