Professor Habshah MIDI, PhD
Senior Lecturer Sohel RANA, PhD*
Email: srana_stat@yahoo.com
Department of Mathematics / Institute for Mathematical Research
University Putra, Malaysia
Professor A.H.M. Rahmatullah IMON, PhD
Department of Mathematical Sciences
Ball State University, Muncie, U.S.A.

TWO-STEP ROBUST ESTIMATOR IN HETEROSCEDASTIC REGRESSION MODEL IN THE PRESENCE OF OUTLIERS

Abstract. *Although the ordinary least squares (OLS) estimates are unbiased in the presence of heteroscedasticity, these are no longer efficient. This problem becomes more complicated when the violation of constant error variances comes together with the existence of outliers. The weighted least squares (WLS) procedure is often used to estimate the regression parameters when heteroscedasticity occurs in the data. But there is evidence that the WLS estimators suffer a huge set back in the presence of outliers. Moreover, the use of the WLS requires a known form of the heteroscedastic errors structures. To rectify this problem, we proposed a new method that we call two-step robust weighted least squares (TSRWLS) method where prior information on the structure of the heteroscedastic errors is not required. In the proposed procedure, the robust technique is used twice. Firstly, the robust weights are used for solving the heteroscedasic error and secondly, the robust weighting function is used for eliminating the effect of outliers. The performance of the newly proposed estimator is investigated extensively by real data sets and Monte Carlo simulations.*

Keywords: *Heteroscedasticity, Weighted least squares, Two-step robust weighted least squares, Outliers, Monte Carlo simulation.*

JEL Classification: C12, C22, C52, C63

1. INTRODUCTION

The linear regression model is commonly used by statistics practitioners in many different fields like engineering, physics, medicine, biology, chemistry, social science and economics. The regression parameters are often estimated by the ordinary least squares (OLS). Under the usual assumptions, the least-squares estimators possess many

desirable properties. In particular, these assumptions imply that the estimators of the parameters will be unbiased, consistent, and efficient in the class of linear unbiased estimators. A commonly used assumption is the constancy of error variances or homoskedasticity, mainly because of which the OLS estimators retain the minimum variance property. In a real life situation it is really hard to believe that the error variances will remain constant and that is why the violation of this assumption which causes the heterogeneity of error variances or heteroskedasticity is more prevalent in nature. The main problem with the violation of homoskedaticity assumption is that the usual covariance matrix estimator of the OLS becomes biased and inconsistent.

A large body of literature is now available [1-3,7-10,16,20,21,23,25,26] for correcting the problem of heteroscedasticity. The correction for heteroscedasticity is very simple by means of the weighted least squares (WLS) if the form and magnitude of heteroscedasticity are known. The WLS is equivalent to perform the OLS on the transformed variables. 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 [4] or by transforming both sides [2]. However, the transformation procedure might be complicated when dealing with more than one explanatory variable. Montgomery *et al*. [20], Kutner *et al.* [16], and others have tried to find the appropriate weight to solve the heteroscedastic problem when the form of heteroscedasticity is unknown. White [28] proposed the heteroskedasticity-consistent covariance matrix (HCCM) estimators in this regard. Different forms of HCCM estimators such as the HC0, HC1, HC2, HC3 and HC4 have been proposed [5,6,12,13,18,28]. However, there is no general agreement among statisticians about which of the five estimators of the HCCM (HC0, HC1, HC2, HC3, HC4) should be used [5,6,17,18]. Chatterjee and Hadi [3] proposed an estimator which is weight based, but these weights depend on the known structure of the heteroscedastic data. Montgomery *et al*. [20] and Kutner *et al*. [16] proposed estimators which do not depend on the known structure of the heteroscedastic data. But the main limitation of the Montgomery *et al*. [20] estimator is that it cannot be applied to more than one regressor situation. The estimator proposed by Kutner *et al.* [16] can be applied to more than one variable and it does not depend on the known form of heteroscedasticity, but we suspect this estimator is not outlier resistant.

It is now evident that a few atypical observations (outliers) can make the entire inferential procedure meaningless [2,19,24]. The weighted least squares also suffer the same problem in the presence of outliers [19]. We also believe that the HCCM estimators should suffer from the same problem, as they are based on the OLS residuals. Generally speaking, none of the estimation techniques work well unless the effect of outliers in a heteroscedastic regression model is eliminated or reduced by robustifying the WLS or HCCM. Therefore, in this article we address the following

question: which robust WLS or HCCM procedure should be used when heteroscedasticity and outliers occur at the same time? This problem motivates us to develop a new and more accurate estimation technique. However, in this article, our study is only confined to the development of the robust WLS. In the presence of outliers we have some robust techniques for the detection of heteroscedasticity [15, 22]. Unfortunately, there is not much work in the literature that deals with the estimation of the regression parameters in the presence of both heteroscedasticity and outliers when the structure of heteroscedasticity is unknown. Although Habshah *et al.* [9] has proposed this type of robust estimation procedure, but their procedure can be applied to only one regressor.

In this article, we propose a two-step robust weighted least squares (TSRWLS) estimator which can be applied for more than one regressor when the form of the heteroscedasticity is not known. Firstly, for solving the heteroscedastic problem we estimate the robust initial weights following the idea of Kutner *et al*. [16] and secondly, we estimate the parameters of the model based on Huber's [14] weight function in order to reduce the effect of outliers. Our results show, as expected, that the existing estimators are very sensitive to outliers whereas our proposed estimator is less sensitive to outliers. The proposed TSRWLS estimator is described in section 2. Section 3 provides an illustrative example to show the better performance of the proposed method. Section 4 reports the results of a Monte Carlo simulation study which is designed to investigate the performance of the proposed method and, section 5 contains the concluding remarks.

2. 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

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

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. From a standard adaptation of the Gauss-Markov theorem, one can easily prove that, if the W matrix is known, the WLS provides the best linear unbiased estimator. Moreover, under normality of the errors, it is the best unbiased estimator ever. But this situation almost never exists in real applications and the estimated weights are used instead. Although it is difficult to assess the effect of using estimated weights, but it is generally believed that small variations in the weights due to estimation do not often affect a regression analysis or its interpretation much. But the presence of outliers should have an adverse effect on the determination of weights. Likewise the OLS method, the WLS regression is also sensitive to the presence of outliers. If potential outliers are 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 in which the heteroscedastic error structure is unknown. It is worth mentioning here that the W matrix should perform well in the presence of heteroscedasticity and outliers. To find the robust weight matrix W, we propose a two-step robust weighted least squares (TSRWLS) estimator. The TSRWLS is an extension of works of Habshah *et al*. [9] and Kutner *et al* [16]. Habshah *et al*. [9] proposed a robust weighted least squares (RWLS) estimator to solve the heteroscedatic and outlying problem by developing robust weighting technique. Instead of fitting regression with all the data, they suggested finding several "near-neighbor**"** groups in the explanatory variable. The group medians represent the explanatory variable (X) and the groups in the response variable Y are formed in accordance with the groups formed in X. The sample median absolute deviations (MAD) of each groups of Y and the median of each group of X are then computed. The square of group MADs in Y are then regressed on the corresponding group medians of X by the least trimmed of squares (LTS) [24] method and the regression coefficients from this fitting are computed. Using these coefficients and full X values, the fitted values are obtained. The inverse of these absolute fitted values then form the initial weights. The final weights are obtained after multiplying these weights by Huber's weight [14].

The main limitation of this procedure is that it cannot be applied to more than one regressor. To overcome this problem we incorporate Habshah *et al.* [9] and Kutner *et al.* [16] estimators. Hereafter we will refer to the Kutner *et al.* [16] estimator as KNN (Kutner, Nachtsheim and Neter) estimator. The KNN estimator starts with fitting a linear regression by the ordinary least squares and conducting some preliminary analysis of the residuals. It is obvious that the megaphone shape of absolute residuals of the OLS against the fitted values may confirm the non-constancy of error variances. Therefore, Kutner *et al.* [16] suggested regressing the absolute residuals against the fitted values and obtain a standard deviation regression function. To obtain the weights, the fitted values from this standard deviation regression function are computed and the inverse of the square fitted values are considered as the desirable weights. 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 [14], the Bisquare function [27] and the Hampel function [11] can be used to obtain the final weight. However, in this study, we will use the Huber's [14] weights function which is defined as

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

The constant 1.345 is called the tuning constant and e_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.

3. EXAMPLE

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

3.1 Education Expenditure Data

This data is taken from Chatterjee and Hadi [3] 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 plotted in Fig.1. Fig.'s 1(a) - 1(c) display the residuals-fitted plots without considering Alaska. If the 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 indicates a violation of the constant variance assumption. This signifies that the OLS fit is inappropriate here, as there is a clear indication of heterogeneous error variances. However, the KNN and TSRWLS fit, presented in:

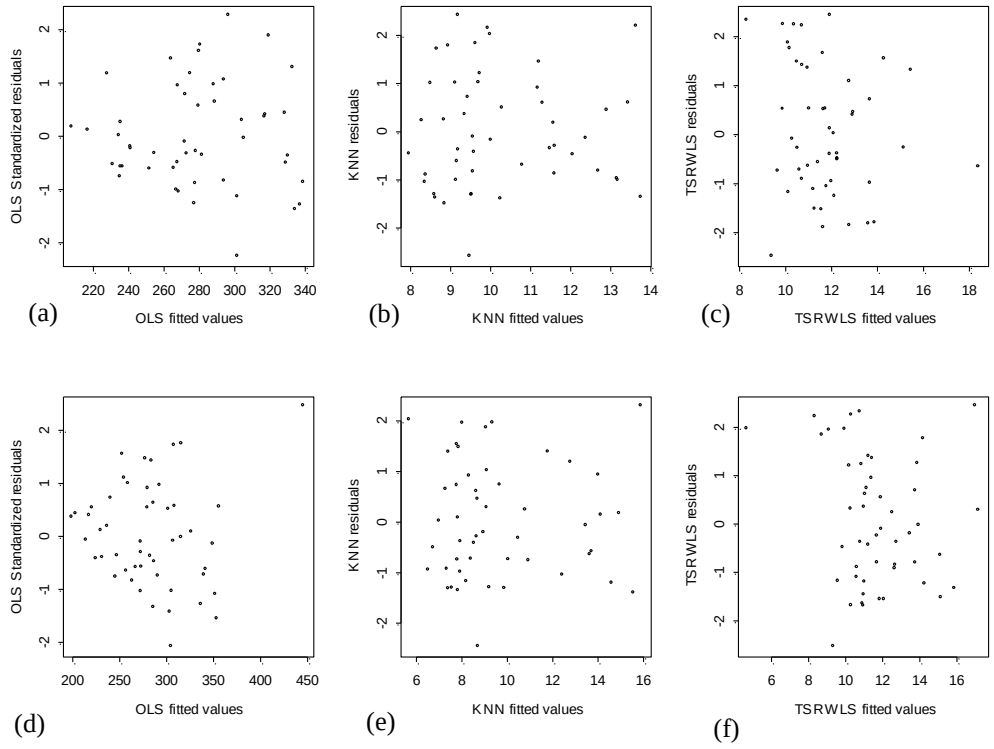

Figure 1. The OLS, KNN and TSRWLS fitted values vs. residuals plots without AK, (a)-(c); with AK, (d)- (f)

Fig.1(b) and Fig.1(c) respectively, do not show any symmetrical shape like the OLS fit. It shows that for this 'clean' data (without AK) the non-constancy of error variances is not reflected in KNN and TSRWLS. To see the effect of outliers, we include the observation Alaska 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.

3.2 Modified Education Expenditure Data

In reality we often have to deal with multiple outliers. For this reason, we deliberately change four data points to generate big 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. In fact, 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 mean and standard deviation of *Y*. In this situation, it is more likely that these points would become big outliers. 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.

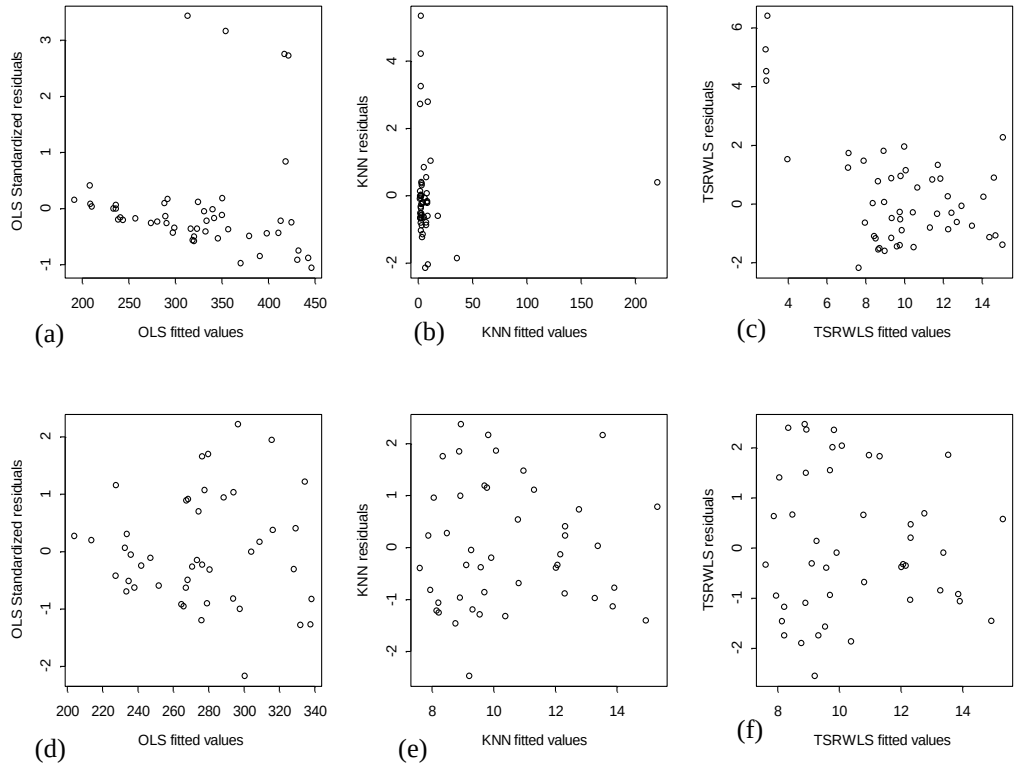

Figure 2. The OLS, KNN and TSRWLS fitted values vs. residuals plots with 10% outliers, (a)-(c) ; without 10% outliers, (d)-(f)

The plots of the residuals against the fitted values of the OLS, KNN and TSRWLS for the modified data are illustrated in Fig.'s 2(a)-2(f). It is observed from Fig.'s 2(a) and 2(b) that in the presence of outliers the patterns of residuals are completely destroyed. 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. It is interesting to note that in Fig. 2(c), the TSRWLS shows the scatter plot of the residuals 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 proposed TSRWLS are successful as it is expected. It re-emphasizes our concern that the KNN might good in the absence of outliers whereas our proposed TSRWLS might be good in the presence or absence of outliers since it is keeping the scatter plot in both situations. In particular, the residuals plots of Fig.1 and Fig.2 show that the TSRWLS estimator is successful to cope with the problem of heteroscedsaticity and outliers.

We know that graphical displays are always very subjective and that is why we would like to present some numerical summaries of the examples considered above. Here, we compare the performance of the proposed TSRWLS estimator with the existing estimators, such as the OLS, KNN and five versions of the HCCM estimators. Table 1 displays the summary statistics such as estimates of the parameters and their standard errors. It also considers three different situations: when there are no outliers, with only one outlier (AK), and with 5 outliers. 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 together with the five versions of HCCM 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.

Table1: 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
	HC0	100.5722	0.0098	0.2590	0.0396
	HC1	109.5119	0.0106	0.2821	0.0431
	HC2	105.5744	0.0103	0.2733	0.0421
	HC3	111.0343	0.0108	0.2891	0.0449
	HC4	101.1556	0.0098	0.2609	0.0399
	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
	HC0	172.6703	0.0157	0.4242	0.0559
	HC1	187.6852	0.0170	0.4611	0.0607
	HC2	222.4836	0.0199	0.5415	0.0673
	HC3	290.5865	0.0257	0.7025	0.0834
	HC4	192.3270	0.0173	0.4700	0.0600
	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
	HC0	400.5560	0.0257	1.0027	0.1379
	HC1	435.3869	0.0279	1.0898	0.1499
	HC2	446.6578	0.0304	1.1106	0.1503
	HC3	513.9515	0.0372	1.2681	0.1671
	HC4	415.541	0.0274	1.0371	0.1411
	TSRWLS	161.8082	0.0170	0.3932	0.0630

4. 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 [5]. 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

$$\lambda = \max(\sigma_i^2) / \min(\sigma_i^2), \; i = 1,2,\ldots,n.$$

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 robustness measures and standard errors of the parameters of the OLS, KNN, and TSRWLS methods are investigated by considering the samples of size 50, 100 and 150. We performed 10,000 simulations using the S-Plus programming language. Summary values such as the mean estimated values

$$\bar{\beta}_j = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}_j^{(k)} \tag{6}$$

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

Table 2: Robustness measure of the parameters of the different estimators, λ =2

%OT[a]	Estimators	Bias			Relative measure of RMSE		
		OLS	KNN	TSRWLS	OLS	KNN	TSRWLS
	Coeff.	Sample Size n= 50					
0%	beta0	-0.0059	-0.0013	-0.00198	–	103.8433	101.2776
	beta1	0.0140	0.0060	0.006226	–	100.6190	96.7113
	beta2	-0.0008	-0.0027	-0.00227	–	99.9619	95.5360
10%	beta0	0.9357	0.1849	0.005338	0.2806	1.2529	68.1788
	beta1	-0.3898	0.1293	-0.01026	0.3632	1.3702	76.1437
	beta2	-2.1139	-1.1011	-0.00035	0.4208	1.0991	74.1417
15%	beta0	-85.7342	-36.0054	0.011383	0.0049	0.0122	71.4086
	beta1	82.7205	33.1180	-0.0172	0.0079	0.0204	83.3345
	beta2	32.2882	9.1732	-0.00564	0.0155	0.0437	71.9914
20%	beta0	-195.6400	-87.9323	-0.00422	0.0028	0.0056	52.1659
	beta1	-372.1610	-146.0850	0.00136	0.0016	0.0045	59.9413
	beta2	494.4786	181.1661	0.012381	0.0015	0.0041	56.5258
		Sample Size n= 100					
0%	beta0	0.0005	-0.0018	-0.0005	–	103.0542	100.6648
	beta1	0.0004	0.0035	0.0011	–	101.8250	98.3953
	beta2	-0.0014	-0.0009	-0.0016	–	101.6148	97.7441
10%	beta0	10.5640	1.3684	-0.0007	0.0342	0.2374	81.5563
	beta1	-15.3713	-2.4119	0.0032	0.0339	0.1863	83.6645
	beta2	-9.1550	-1.9725	-0.0043	0.0460	0.2411	84.4919
15%	beta0	0.0707	0.8919	-0.0046	0.0970	0.3056	77.1535
	beta1	0.9458	-1.4343	0.0033	0.1534	0.3415	79.3901
	beta2	1.7874	0.9745	0.0045	0.0827	0.3002	81.0387
20%	beta0	3.5876	-0.8170	-0.0050	0.0605	0.1566	76.7077
	beta1	-5.4541	-0.7692	0.0059	0.0733	0.2030	77.2481
	beta2	-7.0086	-1.8175	0.0014	0.0487	0.1171	77.1561
		Sample Size n= 150					
0%	beta0	0.0016	0.0027	0.0029	–	103.0130	99.8505
	beta1	0.0002	-0.0015	-0.0023	–	101.6369	96.9257
	beta2	-0.0041	-0.0048	-0.0046	–	102.2114	97.8810
10%	beta0	-6.0098	-1.2097	-0.0026	0.0327	0.1486	80.4602
	beta1	0.1340	-0.1719	0.0031	0.0619	0.3243	88.3611
	beta2	7.7121	1.7646	0.0056	0.0482	0.2295	79.0424
15%	beta0	-28.7559	-5.8467	-0.0078	0.0104	0.0556	69.3154
	beta1	22.4220	4.0933	0.0019	0.0204	0.1298	81.2431
	beta2	32.6964	7.1617	0.0141	0.0156	0.0819	71.1146
20%	beta0	-1.9031	-0.2281	-0.0043	0.0804	0.2444	64.2251
	beta1	7.1205	1.1036	0.0056	0.0609	0.3344	77.9002
	beta2	-0.6418	0.0808	0.0049	0.0558	0.2371	64.3352

a. Percentages of outliers

Table 3: Robustness measure of the parameters of the different estimators, λ =4

%OT[a]	Estimators	Bias			Relative measure of RMSE		
		OLS	KNN	TSRWLS	OLS	KNN	TSRWLS
	Coeff.	Sample Size n= 50					
0%	beta0	-0.0105	0.0039	0.0025	–	112.4879	125.1855
	beta1	0.0152	0.0020	-0.0028	–	101.3282	109.5216
	beta2	0.0157	-0.0013	0.0028	–	88.8705	105.7927
10%	beta0	-3.5890	-0.7598	-0.0043	0.1139	0.5216	34.7474
	beta1	6.1317	1.6857	0.0042	0.1631	0.6527	53.7951
	beta2	-0.9029	-0.9975	0.0112	0.1860	0.4935	33.8907
15%	beta0	1.3242	0.3470	0.0316	0.4223	1.1818	18.9640
	beta1	0.6738	0.3530	0.0084	0.4953	1.2997	42.5771
	beta2	0.0245	0.8270	-0.0787	0.4228	0.6019	13.5809
20%	beta0	-5.0442	-1.9398	0.0122	0.1050	0.2184	34.7901
	beta1	-11.4162	-4.4248	-0.0129	0.0678	0.1864	44.8670
	beta2	15.4846	5.8186	-0.0215	0.0692	0.1736	37.7802
		Sample Size n= 100					
0%	beta0	-0.0035	-0.0048	-0.0041	–	111.8690	118.3724
	beta1	0.0076	0.0185	0.0122	–	95.1401	107.5837
	beta2	-0.0002	-0.0029	-0.0039	–	84.9755	108.4458
10%	beta0	21.0349	5.9017	-0.0032	0.0300	0.1167	93.8616
	beta1	18.5522	4.5442	0.0009	0.0469	0.1668	91.8525
	beta2	-46.2991	-11.0444	0.0077	0.0252	0.1145	91.7733
15%	beta0	-22.9743	-5.2984	0.0010	0.0473	0.1843	89.0858
	beta1	28.5573	7.3703	-0.0056	0.0472	0.1543	87.4307
	beta2	22.0493	6.4229	-0.0093	0.0701	0.2321	89.0389
20%	beta0	6.6902	2.4548	0.0109	0.0737	0.2251	86.4168
	beta1	0.7136	0.3268	-0.0034	0.0393	0.1187	83.4477
	beta2	-11.0030	-3.5738	-0.0058	0.0888	0.3042	82.9778
		Sample Size n= 150					
0%	beta0	-0.0008	-0.0080	-0.0023	–	107.5062	116.1884
	beta1	0.0045	0.0137	0.0055	–	99.3414	106.0485
	beta2	0.0006	0.0092	0.0006	–	101.1024	109.6160
10%	beta0	-6.2692	-2.1037	-0.0004	0.0587	0.1424	84.3554
	beta1	47.2234	9.9197	-0.0022	0.0145	0.0732	93.4204
	beta2	-52.8924	-12.6937	-0.0015	0.0114	0.0507	72.4760
15%	beta0	3.1242	1.1391	-0.0052	0.1338	0.5144	66.6502
	beta1	-1.0545	-0.4605	0.0090	0.2199	0.7523	81.2984
	beta2	-9.7318	-4.4918	0.0017	0.0727	0.2105	66.9573
20%	beta0	-5.6191	-0.4311	0.0102	0.0450	0.1757	61.5970
	beta1	4.1717	1.0792	-0.0219	0.0548	0.1820	77.9695
	beta2	17.2251	4.9047	0.0038	0.0447	0.1585	60.9549

a. Percentages of outliers

Table 4: Standard errors of the parameters of the different estimators, λ =2

%OT[a]	Estimators	Standard errors of the Parameters							
		OLS	KNN	HC0	HC1	HC2	HC3	HC4	TSRWLS
	Coeff.	Sample Size n=50							
0%	beta0	0.5146	0.3886	0.4282	0.4555	0.4458	0.4642	0.4296	0.3924
	beta1	0.8891	0.6044	0.6339	0.6743	0.6594	0.6861	0.4296	0.5738
	beta2	0.7555	0.6148	0.6451	0.6863	0.6708	0.6978	0.6471	0.5737
10%	beta0	17.7305	7.2738	12.7274	13.5398	13.2629	13.8246	12.7736	1.0664
	beta1	23.3970	10.4365	17.1504	18.2451	17.7991	18.4765	12.7736	1.5815
	beta2	22.5848	12.0001	17.5826	18.7049	18.1509	18.7429	17.6240	1.5880
15%	beta0	99.5604	52.3143	115.5253	122.8992	118.8438	122.2770	115.7264	1.3825
	beta1	131.3787	52.4478	116.7274	124.1781	120.0780	123.5445	115.7264	1.9774
	beta2	126.8181	40.8369	78.8850	83.9202	81.1562	83.5130	79.0293	1.9604
20%	beta0	461.5606	119.7004	231.0199	245.7658	238.8034	246.9060	231.5823	1.8428
	beta1	609.0697	189.9633	487.4313	518.5440	501.7849	516.6386	231.5823	2.6153
	beta2	587.9272	206.0266	513.1987	545.9561	527.6012	542.4713	514.0404	2.5808
		Sample Size n=100							
0%	beta0	0.4458	0.3648	0.3969	0.4092	0.4048	0.4128	0.3973	0.3551
	beta1	0.7449	0.4964	0.5188	0.5348	0.5291	0.5397	0.3973	0.4657
	beta2	0.6107	0.4793	0.5117	0.5275	0.5213	0.5311	0.5121	0.4470
10%	beta0	34.5617	11.2185	36.9714	38.1149	37.6980	38.4399	37.0007	1.0382
	beta1	42.2446	15.8414	45.6371	47.0485	46.5277	47.4371	37.0007	1.3507
	beta2	39.5932	14.9017	41.7365	43.0273	42.5197	43.3189	41.7669	1.3046
15%	beta0	42.3308	16.4106	37.6879	38.8535	38.3326	38.9895	37.7111	1.3154
	beta1	51.7408	19.4518	42.2961	43.6042	43.0333	43.7847	37.7111	1.7094
	beta2	48.4933	21.1654	48.0123	49.4972	48.8209	49.6446	48.0409	1.6383
20%	beta0	49.3378	18.8709	37.3376	38.4924	37.9208	38.5145	37.3571	1.5341
	beta1	60.3054	24.3215	47.7927	49.2708	48.5488	49.3186	37.3571	2.0007
	beta2	56.5204	25.3062	49.1922	50.7136	49.9345	50.6894	49.2160	1.9201
		Sample Size n=150							
0%	beta0	0.2972	0.2506	0.2702	0.2757	0.2734	0.2766	0.2703	0.2422
	beta1	0.6195	0.4124	0.4257	0.4344	0.4315	0.4373	0.2703	0.3780
	beta2	0.5844	0.3632	0.3881	0.3961	0.3931	0.3980	0.3883	0.3379
10%	beta0	24.2137	10.4550	26.7381	27.2838	27.0830	27.4326	26.7472	0.7400
	beta1	34.1740	11.4259	29.3331	29.9317	29.7236	30.1194	26.7472	1.1437
	beta2	31.3201	14.7913	36.6304	37.3780	37.0851	37.5457	36.6419	1.0459
15%	beta0	0.2972	0.2506	0.2702	0.2757	0.2734	0.2766	0.2703	0.2422
	beta1	0.4195	0.4124	0.4257	0.4344	0.4315	0.4373	0.2703	0.3780
	beta2	0.3844	0.3632	0.3881	0.3961	0.3931	0.3980	0.3883	0.3379
20%	beta0	40.5311	19.1497	35.1320	35.8490	35.5474	35.9684	35.1426	1.1576
	beta1	57.2035	20.1988	49.2925	50.2985	49.9564	50.6300	35.1426	1.7229
	beta2	52.4263	25.1124	58.2668	59.4559	58.9954	59.7339	58.2860	1.6235

a. Percentages of outliers

Table 5: Standard errors of the parameters of the different estimators, λ =4

%OT[a]	Estimators	Standard errors of the Parameters							
		OLS	KNN	HC0	HC1	HC2	HC3	HC4	TSRWLS
	Coeff.	Sample Size n=50							
0%	beta0	0.7841	0.3701	0.6144	0.6536	0.6396	0.6661	0.6165	0.4456
	beta1	1.9347	0.8175	0.9655	1.0271	1.004	1.0447	0.6165	0.7656
	beta2	1.4988	0.8770	1.0245	1.0899	1.0661	1.1096	1.0279	0.7956
10%	beta0	35.6725	11.7913	24.8378	26.4232	25.8755	26.9634	24.9266	1.2850
	beta1	47.0730	18.1229	34.6098	36.8189	35.9004	37.2473	24.9266	2.4265
	beta2	45.4390	21.9545	36.0837	38.3869	37.2307	38.4245	36.1657	2.5319
15%	beta0	35.8009	14.8161	23.1191	24.5948	23.9461	24.8099	23.1831	1.6114
	beta1	47.2424	20.1513	31.2127	33.2050	32.2503	33.3304	23.1831	2.7513
	beta2	45.6025	23.8200	32.6295	34.7123	33.5511	34.5078	32.6899	2.8267
20%	beta0	62.2517	28.3594	42.9578	45.6997	44.3926	45.8858	43.0599	2.1857
	beta1	82.1466	35.3214	58.6896	62.4358	60.5577	62.4975	43.0599	3.6111
	beta2	79.2951	38.2728	59.2919	63.0765	60.9630	62.6927	59.3943	3.6728
		Sample Size n=100							
0%	beta0	0.6887	0.3983	0.5922	0.6105	0.6039	0.61588	0.5927	0.4347
	beta1	1.0141	0.6564	0.7964	0.8210	0.8121	0.8282	0.5927	0.6320
	beta2	0. 9778	0.6713	0.8224	0.8478	0.8377	0.8534	0.8229	0.6246
10%	beta0	73.8892	21.7184	68.7456	70.8717	70.0285	71.3372	68.7943	1.3103
	beta1	90.3146	28.0017	77.1267	79.5120	78.4760	79.8510	68.7943	1.8999
	beta2	84.6461	30.8492	93.2082	96.0909	94.9240	96.6736	93.2725	1.8854
15%	beta0	71.6299	24.5261	65.2019	67.2185	66.2860	67.3900	65.2394	1.6604
	beta1	87.5531	33.5901	81.1361	83.6454	82.4972	83.8834	65.2394	2.3836
	beta2	82.0579	31.4783	75.9616	78.3110	77.1846	78.4293	76.0027	2.3416
20%	beta0	94.1637	28.8749	64.8414	66.8468	65.8750	66.9275	64.8766	1.9582
	beta1	115.0960	46.7776	104.9376	108.1831	106.5877	108.2669	64.8766	2.7995
	beta2	107.8721	38.3557	80.8973	83.3992	82.2061	83.5387	80.9417	2.7655
		Sample Size n=150							
0%	beta0	0.44058	0.2661	0.3851	0.3930	0.3899	0.3947	0.3853	0.2896
	beta1	0.9817	0.5672	0.6501	0.6634	0.6591	0.6683	0.3853	0.5202
	beta2	0.7998	0.4776	0.5963	0.6085	0.6040	0.6117	0.5965	0.4498
10%	beta0	57.0626	13.8041	31.3104	31.9494	31.6907	32.0759	31.3199	0.9061
	beta1	80.5352	24.1373	80.8688	82.5192	81.9391	83.0239	31.3199	1.6290
	beta2	73.8095	28.1042	86.6113	88.3789	87.7312	88.8662	86.6402	1.4442
15%	beta0	43.3293	15.7365	32.5162	33.17987	32.8919	33.2723	32.5254	1.1761
	beta1	61.1527	20.0178	42.6891	43.5603	43.2259	43.7700	32.5254	1.9996
	beta2	56.0457	27.0094	59.6327	60.8497	60.2767	60.9283	59.6475	1.7925
20%	beta0	60.2810	25.2604	51.2814	52.3280	51.8535	52.4326	51.2949	1.4753
	beta1	85.0775	30.5681	67.0884	68.4575	67.8914	68.7049	51.2949	2.4898
	beta2	77.9724	37.3080	84.9627	86.6966	85.9239	86.8970	84.9854	2.2724

a. Percentages of outliers

$$MSE(\hat{\beta}_j) = (\bar{\beta}_j - \beta_j)^2 + \frac{1}{m}\sum_{k=1}^{m}(\hat{\beta}_j^{(k)} - \bar{\beta}_j)^2 \quad (7)$$

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 of contaminated models 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 2 and 3. Several interesting points appear from Tables [2- 3]. For 'clean' data, all the three estimators considered here are fairly close to one another with respect to the values of the robustness measure. By inspecting the bias and the values of robustness measures in Table 2 and 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. Irrespective of the percentages of outliers it maintains producing low bias and small RMSE.

Tables 4 and 5 present the standard errors of the parameter estimates of the OLS, KNN, five versions of HCCM, and TSRWLS estimators. We observe that the standard errors of the five versions of HCCM estimates also reasonably close to the KNN and TSRWLS, for the 'clean' data. If the form of heteroscedasticity is unknown, many authors recommend using the HCCM based estimators [5,6,12,13,17,18, 28]. But these results clearly show that likewise the OLS and KNN, HCCM based estimators may breakdown even in a very small percentage of contamination and their performances also tend to deteriorate with the increase in the percentage of outliers. Nevertheless, the TSRWLS are not much affected by outliers. The biases and robustness measure of the TSRWLS are consistently small and deteriorate slightly as the percentage of outliers increases.

5. 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, HCCMs 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 HCCMs are much inferior to the TSRWLS when contamination occurred in the data. The empirical study reveals

that the proposed estimator is outlier(s) resistant. Larger bias in estimates and standard errors, and smaller values of robustness measures clearly prove that the OLS, KNN and the five versions of HCCM are easily get affected by outliers. To the contrary, both graphical and numerical evidences signify that the TSRWLS is capable of rectifying the problems of heteroscedasticity and outliers at the same time. Thus, the TSRWLS estimates emerge to be conspicuously more efficient and more reliable in comparison with other estimators considered in this article.

REFERENCES

[1] **Carroll, R.J. & Ruppert, D. (1982),** ***Robust Estimation in Heteroscedastic Linear Models;*** *Annals of Statistics*, 10, 429-441;

[2] **Carroll, R.J. & Ruppert, D. (1988),** ***Transformation and Weighting in Regression*****;** New York: Chapman and Hall;

[3] **Chatterjee S. & Hadi, A.S. (2006),** ***Regression Analysis by Examples***; *New York: Wiley;*

[4] **Cook R.D. & Weisberg, S. (1983),** ***Diagnostics for Heteroscedasticity in Regression*****;** *Biometrika***,** 70, 1-10;

[5] **Cribari-Neto, F. (2004),** ***Asymptotic Inference under Heteroskedasticity of Unknown Form*****;** *Computational Statistics and Data Analysis*, 45, 215-233;

[6] **Davidson, R. & MacKinnon, J.G.(1993),** ***Estimation and Inference in Econometrics*****;** New York: *Oxford University Press;*

[7] **Greene, W. (2008),** ***Econometric Analysis*****;** *New York: Pearson;*

[8] **Habshah, M. (2000),** ***Heteroscedastic Nonlinear Regression by Using Tanh Phi Function,*** *Sains Malaysiana*, 29,103-118;

[9] **Habshah, M., Rana, S. & Imon, A.H.M.R. (2009),** ***The Performance of Robust Weighted Least Squares in the Presence of Outliers and Heteroscedastic*****;** *WSEAS Transition of Mathematics*, 8, 351 – 361;

[10] **Habshah, M., Rana, S. & Imon, A.H.M.R. (2009),** ***Estimation of Parameters in Heteroscedastic Multiple Regression Model Using Leverage Based Near-Neighbors;*** *Journal of Applied Sciences*, 9, 4013-4019;

[11] **Hample, F.R. (1974),** ***The Influence Curve and its Role in Robust Estimation,*** *Journal of the American Statistical Association*,69, 383-393;

[12] **Hinkley, D.V. (1977),** ***Jackknifing in Unbalanced Situations*****;** *Technometrics*, 19, 285-292;

[13] **Horn, S.D., Horn, R.A. & Duncan, D.B. (1975),** ***Estimating Heteroscedastic Variances in Linear Model,*** *Journal of the American Statistical Association,* 70, 380-385;

[14] **Huber, P.J. (1981),** ***Robust Statistics;*** *New York: Wiley;*

[15] **Imon, A.H.M.R. (2009),** ***Deletion Residuals in the Detection of Heterogeneity of Variances in Linear Regression*****;** *Journal of Applied Statistics,* 36, 347-358;

[16] **Kutner, M.H., Nachtsheim, C.J. & Neter, J. (2004),** ***Applied Linear Regression Models*****;** New York: *McGraw- Hill/Irwin*;

[17] **Long, J.S. & Ervin, L.H. (2000),** ***Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model*****;** *The American Statistician,* 54, 217-224;

[18] **MacKinnon, J.G. & White, H. (1985),** ***Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties*****;** *Journal of Econometrics,* 29, 53-57;

[19] **Maronna, R.A., Martin, R.D. & Yohai, V.J. (2006),** ***Robust Statistics -Theory and Methods*****;** *New York: Wiley*;

[20] **Montgomery, D.C., Peck, E.A & Vining, G.G. (2001),** ***Introduction to Linear Regression Analysis*****;** *New York: Wiley;*

[21] **Rana, S., Habshah, M. & 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,* Article ID 730328, 2012, 14 pages;

[22] **Rana, S., Habshah, M. & Imon, A.H.M.R. (2008),** ***A Robust Modification of the Goldfeld-Quandt Test for the Detection of Heteroscedasticity in the Presence of Outliers;*** *Journal of Mathematics and Statistics,* 4, 277-283;

[23] **Robinson, P.M. (1987),** ***Asymptotically Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form;*** *Econometrica,* 55, 875-891;

[24] **Rousseeuw, P.J. Leroy, A. (1997),** ***Robust Regression and Outlier Detection,*** *New York: Wiley;*

[25] **Ryan, T.P. (1997),** ***Modern Regression Methods*****;** *New York: Wiley*;

[26] **Siraj-ud-doulah, M, Rana, S., Midi, H. & Imon, A.H.M.R (2012),** ***New Robust Tests for Detection of ARCH Effect;*** *Economic Computation and Economic Cybernetics Studies and Research,* 46, 251- 259;

[27] **Tukey, J.W. (1977),** ***Exploratory Data Analysis*****;** Cyprus: *Addison-Wesley Publishers;*

[28] **White, H. (1980),** ***A Heteroskedastic- consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity*****;** *Econometrica,* 48, 817-838.