Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 20, Issue 1(2018) ISSN$_{Print}$: 1816-9171

Robust Technique to Remedy of Unequal Variance Problem in the Presence of high leverage points

Mohammed A. Mohammed

Department of Materials Management Techniques/
Al-Dewanyia Technical Institute /Al-Furat Alawsat Technical University

Corresponding Author: Mohammed A. Mohammed *Email: mohammedam23@yahoo.com*

***Abstract* :** *An important assumption of linear regression model is that the variance of disturbances everywhere is equal (constant variance). However, unequal variance called heteroscedasticity does not cause biasness in estimates, but it leads to an efficient problem and the standard errors of observations will be inaccurate. Under heteroscedasticity problem, the ordinary least squares estimates (OLS) are inefficient due to it gives same weights to all observations regardless of the fact that those with large residuals contain less information about regression model. The weighted least square (WLS) is a common method for remedy the heteroscedasticity problem. Unfortunately, in the presence of high leverage points (outlier in the predictor variables), the estimates of classical method such as OLS and WLS will be damaged and an inefficient. In order to tackle the combined problem of heteroscedasticity and high leverage points, we suggested a new estimation method called robust quintile weighted least squares (RQWLS). The results of real data example and simulation study shows that the suggested method has good performance compared with the existing methods.*

***Keywords*:** *OLS, WLS, heteroskedasticity, high leverage points, quintile regression, robust standard error.*

1 INTRODUCTION

One of the usual assumptions of the classical linear regression model is that the variance of each error (u_i), is constant. Therefore, having an equal variance means that the disturbances are homoscedastic, given by;

$$var(u_i|x_1, x_2, \dots, x_k) = \sigma^2 \quad \dots (1)$$

If Eq. (1) is not true, that is, the variance of u_i is different for different values of the *x*'s, then the errors are heteroskedastic, as follows (see [8], [11] and [14]);

$$var(u_i|x_1, x_2, \dots, x_k) = \sigma_i^2 \quad \dots (2)$$

To show the differences between homoscedasticity and heteroskedasticity, assume that the following simple regression model which shows the relationship between the saving and the income variables (see [8]);

$$y_i = \beta_0 + \beta_1 x_i + u_i \dots (3)$$

where, y is a dependent variable represents the savings and *x* is an independent variable represents the income. Fig. 1 shows the relationship between the income and the savings. It is interesting to show in Fig. (1,a) that the variance of savings remains the same at all levels of income (homoscedastic), whereas in Fig. (1,b), the savings increases when the income increases (heteroskedastic) (see [9]).

Under heteroscedasticity problem, errors may increase as the value of an independent variable increases. The error terms associated with very large value might have larger variances than error terms associated with smaller value (see [1], [6] and [7]). With this issue, OLS estimates are no longer BLUE (not the best linear unbiased estimator). That is, among all the unbiased estimators, OLS does not provide the estimate with the smallest variance. Depending on the nature of the heteroskedasticity, significance tests can be too high or too low. In addition, the standard errors are biased when heteroscedasticity is presented. This in turn leads to bias in test statistics and confidence intervals. The OLS estimators are still unbiased and consistent because none of the explanatory variables is correlated with the error term. When the estimates of the distribution coefficients are affected by heteroscedasticity, the variances of the distributions will increase and therefore making the OLS estimators inefficient. So that, estimates the variances of the estimators, gives a higher values of *t* and *F* statistics (see [4], [14] and [17]).

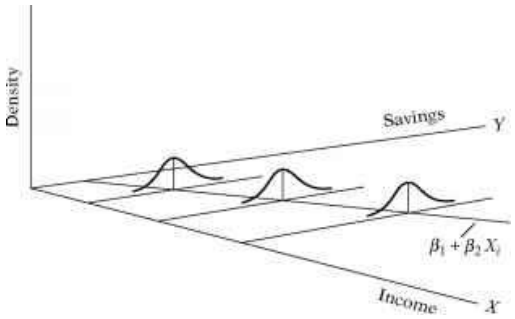

Fig.: (1.a) Homoscedastic disturbances

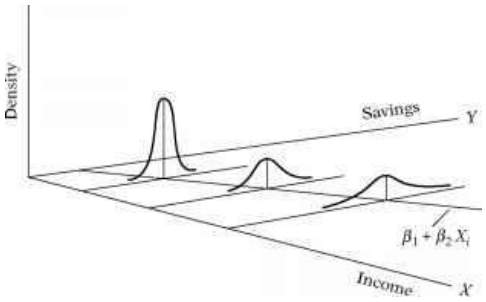

Fig.: (1.b) Heteroskedastic disturbances

Source: Gujarati et. al (2009)

This paper is organized as follows: the weighted least square method is briefly presented in Section 2. The robust quintile weighted least squares is explained in Section 3. In sections 4 and 5 the proposed method (RQWLS) were applied with real data and simulation study, sequentially. Finally, the conclusions are given in Section 6.

2 WEIGHTED LEAST SQUARE METHOD

The method of weighted least squares (WLS) can be used when the assumption of homoscedasticity is violated. For the multiple linear regression model (see [5], [8] and [13])

$$y_i = x\beta + u_i \quad \dots(4)$$

rather than assuming that errors are constant, let u_i is assumed to be (multivariate) normally distributed with mean 0 and non-constant variance-covariance matrix as follows (see[7])

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \dots(5)$$

Then, the variance is written as

$$var(y\,|x = x_i) = var(u_i) = \sigma^2/w_i \quad \dots(6)$$

where $w_i, i = 1,2,\dots,n$ are the weights and its known as positive numbers .
In matrix form, let W be an $n \times n$ diagonal matrix with the w_i on the diagonal, given as (see [1] and [7])

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \dots(7)$$

hence, the WLS estimate is a solution for the following argue,

$$\hat{\beta}_{WLS} = \arg\min_{\beta} \sum_{i=1}^{n} u_i^2 \quad \dots(8)$$
$$= \sum w_i (y - x_i^T \beta)^2$$
$$= (X^T W X)^{-1} X^T W y \quad \dots(9)$$

The Eq. (9) can be found directly, but it is suitable to transform the problem by sitting $var(u_i) = \sigma^2 {w_i}^{-1}$,as shown in Eq. (6), in to one that can be solved by OLS method. Then, all of the OLS results can be applied to WLS.

Let $W^{1/2}$ and $W^{-1/2}$ are diagonal $(n \times n)$ matrices with elements $\sqrt{w_i}$ and $1/W^{1/2}$, respectively. Then

$$W^{1/2}W^{-1/2} = I \quad \dots(10)$$

hence, the covariance matrix of $W^{1/2}u$ is given by

$$\begin{aligned} var(W^{\frac{1}{2}}u) &= W^{\frac{1}{2}}\ var(u)\ W^{\frac{1}{2}} \\ &= W^{\frac{1}{2}}(\sigma^2 W^{-1})\, W^{\frac{1}{2}} \\ &= W^{\frac{1}{2}}(\sigma^2 W^{-1/2}W^{-1/2})\, W^{\frac{1}{2}} \\ &= \sigma^2(W^{\frac{1}{2}}W^{-\frac{1}{2}})(W^{\frac{1}{2}}W^{-\frac{1}{2}}) \\ &= \sigma^2 I_{n\times n} \end{aligned}$$

It is clearly to see that the term $W^{1/2}u$ is a random vector with covariance matrix equal to $\sigma^2 I_{n\times n}$. Multiplying both sides of Eq. (4) by $W^{1/2}$ gives;

$$W^{\frac{1}{2}}y_i = W^{\frac{1}{2}}x\beta + W^{\frac{1}{2}}u_i \quad \dots(11)$$

$$\Rightarrow\ y_i^* = x^*\beta + u_i^* \quad \dots(12)$$

where, $y_i^* = W^{\frac{1}{2}}y$, $x^* = W^{\frac{1}{2}}x$ and $u_i^* = W^{\frac{1}{2}}u_i$

Eq. (12) is the OLS technique with new variables (y_i^*, x^* and u_i^*). These estimators are unbiased and have minimum variance among all unbiased estimators (see [1], [10] and [12]). If the heteroscedastic error structure of the regression model is known, it is easy to compute the weights of W matrix, and consequently the WLS would be a good solution of heteroscedastic regression model.

3 THE ROBUST QUINTILE WEIGHTED LEAST SQUARES

Weighted least squares regression, like the other least squares methods, is also sensitive to effects of outliers and high leverage points (HLPs). If potential outliers are not investigated, they will likely have a negative impact on the parameter estimation and inferences of weighted least squares analysis (see [2], [3] and [9]). To deal with the combined problem of heteroscedasticity and presence of HLPs, we suggested a new estimation technique called Robust Quintile Weighted Least Squares (RQWLS). The procedure of RQWLS is as follows (see [11] and [13]):

1) We suppose the variance – covariance matrix is unknown, then the values of $\mathrm{w_i} = \frac{1}{\sigma_i^2}$, $\mathrm{i} = 1,2,\dots,\mathrm{n}$ are also unknown.
2) Examine a plot of residuals ($\hat{u}_i$) versus predicted values ($\hat{y}_i$) by using OLS estimates. When the constant variance assumption is violated, the plot may look like megaphone form which is shown in Fig. 2.

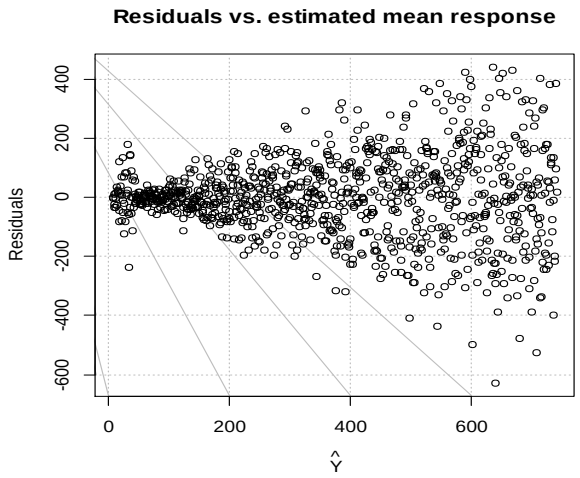

Fig. 2: plot of residuals ($\hat{u}_i$) against predicted values ($\hat{y}_i$)

3) Divide data into suitable number of groups such as 3 to 5 groups as shown in Fig. 3, then estimate the median absolute deviation (MAD) for each group. The MAD_j is computed as;

$$MAD_j = c \text{ median } |x_{ij} - \text{median } (x_j)| \quad \ldots (13)$$

where, c is a constant to prepare the MAD to be consistent for σ. c is substituted by 1.4826 for consistency (see[4]). The MAD has the best possible break down point equal to 50% and it has asymptotic efficiency.

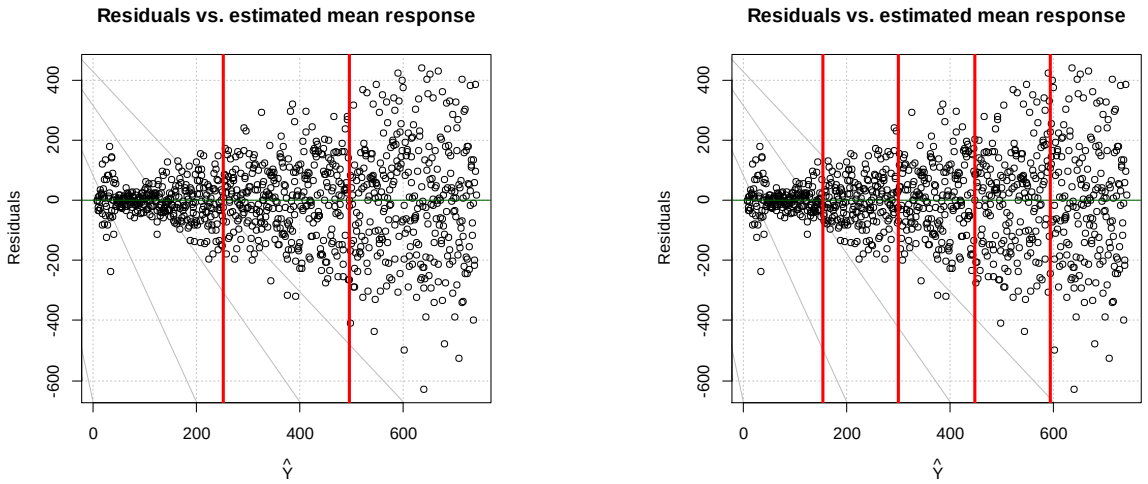

Fig. 3: divided plot of residuals versus predicted values

4) Apply the RQWLS method, the coefficients of RQWLS ($\hat{\beta}_{RQWLS}$) is given by

$$\hat{\beta}_{RQWLS} = (X^T W^* X)^{-1} X^T W^* y \quad \ldots (14)$$

where, $w_i^* = \frac{1}{Mad_j}$, and j denotes to the group number.

4 APPLIED THE RQWLS WITH REAL DATA

In this section, real data is applied to evaluation the suggested method. The R language is used to analyze the data. The data is collected from a random sample of 100 students (male and female) in Al-Diwaniya Technical Institute. The data of the random sample represent the relationship between weight and height, where high represent the independent variable and weight represent the dependent variable. To test the heteroscedasticity problem in the data set, the Breusch -Pagan test (BP) is used. The BP test is suggested in 1979 by Trevor Breusch and Adrian Pagan to test an unequal variance for residuals in linear regression model (see [8] and [10]). The BP test implies the two following hypotheses.

H_0: Data is homoscedastic.
H_1: Data is heteroscedastic.

If the p-value associated to a BP test falls below a certain threshold (0.05), we would conclude that the data is significantly heteroscedastic. The result of the BP test for our sample data is as following:

BP = 19.48, df = 1and p-value = 1.017e-05

The p-value of BP- test is less than 0.05, indicates that the null hypothesis (data is homoscedastic) can be rejected and therefore heteroscedasticity is exists. Another evidence of heteroscedasticity can be shown in Fig. 4. The plot of residuals versus fitted value is likes a megaphone shape which means the data set has heteroscedasticity problem. To make the combined problem of heteroscedasticity and high leverage points, we replace the first value on the explanatory variable by huge value as shown in the plot of residuals versus leverage in Fig. 4.

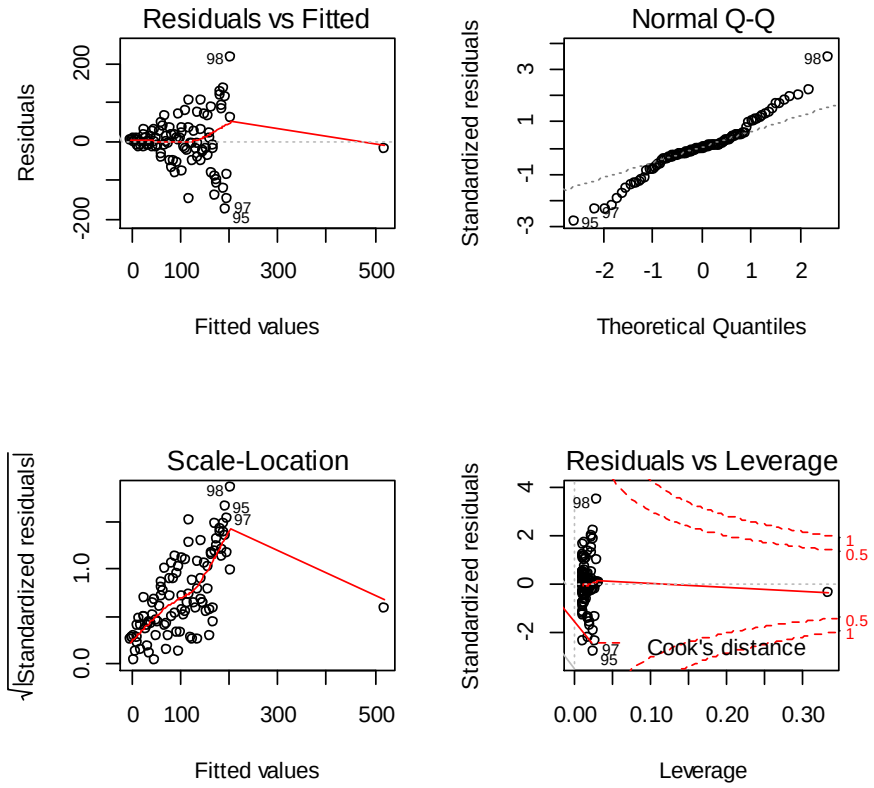

Fig. 4: plots of residuals versus fitted, normal Q-Q, fitted values versus standard residuals and leverage

The *RQWLS* were applied by divided the data into three groups. To assess the suggested method (*RQWLS*), it is compared with some existent methods such as *OLS*, robust MM-estimate (*MM-est.*) and *WLS*. Table 1 presents values of residual standard error (*Se*), *t-test* of estimator and coefficient of determination (R^2) for all methods of study.

Table 1: *values of* Se, *t-test and* R^2 *for methods of study*

Criteria	*Methods*			
	OLS	*MM-est.*	*WLS*	*RQWLS*
Se	*63.84*	*42.5*	*1.903*	*0.281*
t-test	*11,319*	*12.539*	*12.44*	*12.93*
R^2	*0.566*	*0.401*	*0.778*	*0.847*

From the results in Table 1 we can see that the *RQWLS* has a lower value of Se and higher values of *t-value* and R^2 which indicate that the *RQWLS* has superior performance followed by *WLS.*

5 SIMULATION STUDY

In this section, we present a simulation study to evaluate the performance of the suggested method (*RQWLS*) in the presence of combined problem of heteroscedasticity and HLPs (see [9] and [14]). The following simple linear regression model is considered,

$$y_i = 2 + 3x_i + u_i\,,\ \ i = 1,2,\dots,n \quad \dots(15)$$

The heteroscedastic data is created by generating *x* using a normal distribution and by adding to each value of y an additional term (error term) generated with a normal distribution with mean 0 and variance equal to $x\sigma$, where $\sigma \sim N(0,1)$. Four different sizes of samples are selected corresponding to 50, 100, 200 and 300. The contamination is done by replacing a clean datum in the explanatory variables with HLPs corresponding to two percentages of contaminations (0.05 and 0.10). Consequently, four estimation methods are applied in the simulation study, such as, *OLS*, *MM-estimator*, *WLS* and *RQWLS* (see [3], [15] and [16]).
A plot of the residuals versus the predictor values in Fig. 5 indicates that possible non-constant variance in the simulated data set since there is a very slight "megaphone" pattern.

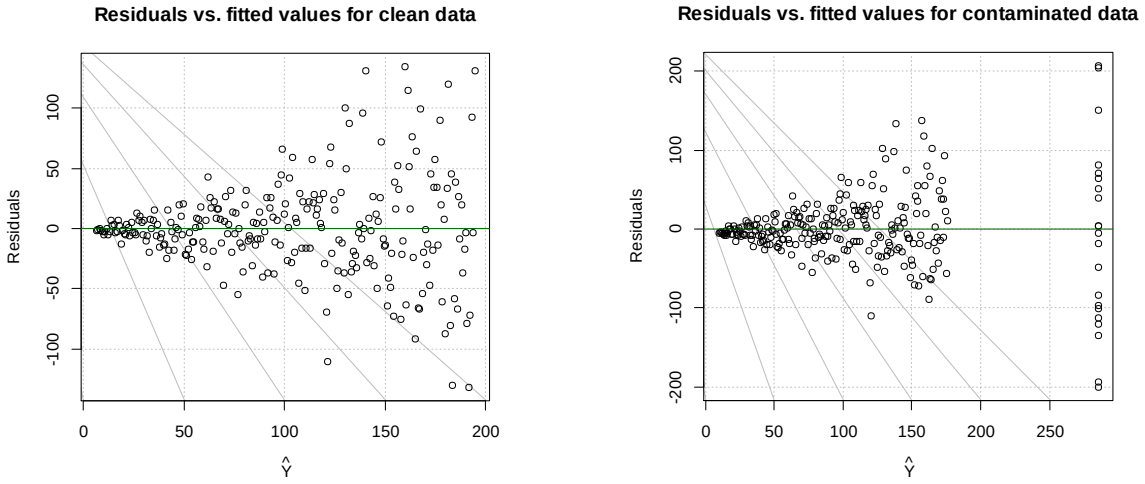

Fig. 5: plot or residuals versus fitted values for two cases; clean and contaminated data set

The results of simulation study in different sizes of samples and different percentage of contaminated are presented in tables 1 and 2. It is clearly to show that the *RQWLS* has the best performance since it has lower values of *Se* and higher values of *t-test* and R^2than other estimation methods in various sizes of samples and different percentage of contaminated.

6 CONCLUSIONS

In this study, we proposed a new estimation methods called robust quintile weighted least square (*RQWLS*) to solve the problem of heteroscedasticity in the presence of high leverage points. The procedure of *RQWLS* is done by split a data set into 5 groups then using a robust dispersion measure such as *MAD* to calculate the weights of *WLS* method. In order to evaluate the performance of the suggested method, it was compared with some existing methods by using real data set and several simulation data based on Se, *t-test and* R^2. The results indicate that the existing methods have worse performance compared with the proposed method when the unequal variance data has HLPs. It is interesting to conclude from the results of real data and simulation study that the *RQWLS* method has successfully solving for heteroscedasticity problem in the presence of HLPs.

Table 2: V*alues of* Se, *t-test and* R^2 *with 0.05 percentage of contaminated*

Methods	*Criteria*	*Sample of size*			
		$n = 50$	$n = 100$	$n = 200$	$n = 300$
OLS	Se	*72.40*	*70.00*	*53.23*	*49.19*
	t-value	*7.86*	*8.313*	*13.17*	*17.92*
	R^2	*0.559*	*0.572*	*0.577*	*0.581*
MM-est.	Se	*20.63*	*18.14*	*16.01*	*15.89*
	t-value	*7.33*	*7.47*	*13.80*	*15.21*
	R^2	*31.77*	*0.39*	*0.44*	*0.46*
WLS	Se	*13.21*	*15.99*	*15.75*	*16.93*
	t-value	*14.14*	*14.89*	*17.22*	*18.87*
	R^2	*0.40*	*0.41*	*0.46*	*0.48*
RQWLS	Se	*1.011*	*0.833*	*0.707*	*0.699*
	t-value	*18.19*	*21.15*	*23.08*	*25.79*
	R^2	*74.95*	*0.766*	*0.810*	*0.830*

Table 3: *Values of* Se*, t-test and* R^2 *with 0.10 percentage of contaminated*

Methods	*Criteria*	*Sample of size*			
		$n = 50$	$n = 100$	$n = 200$	$n = 300$
OLS	Se	*230.75*	*203.00*	*126.7*	*94.79*
	t-value	*4.59*	*7.322*	*11.17*	*17.34*
	R^2	*0.291*	*0.347*	*0.373*	*0.443*
MM-est.	Se	*27.94*	*24.29*	*20.24*	*17.33*
	t-value	*4.74*	*7.05*	*13.64*	*14.51*
	R^2	*0.342*	*0.370*	*0.430*	*0.454*
WLS	Se	*16.45*	*14.79*	*11.25*	*10.22*
	t-value	*10.68*	*10.89*	*12.44*	*13.67*
	R^2	*0.364*	*0.366*	*0.436*	*.472*
RQWLS	Se	*1.756*	*0.911*	*0.756*	*0.711*
	t-value	*13.75*	*13.85*	*15.88*	*20.74*
	R^2	*0.708*	*0.715*	*0.751*	*0.801*

REFERENCES

[1] White and Halbert. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Journal of the Econometric Society, 48(4),*1980, 817-838.
[2] Huber, Peter J., and Ronchetti E. M., *Robust statistics* (Series in probability and mathematical statistics, 1981).
[3] Carroll, R. J., & Ruppert D., Robust estimation in heteroscedastic linear models, *The annals of statistics, 10(2),* 1982, 429-441.
[4] Leroy, A. M., & Rousseeuw, P. J., *Robust regression and outlier detection* (*Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*).
[5] Scott Long, J., & Ervin, L., *Correcting for heteroscedasticity with heteroscedasticity consistent standard errors in the linear regression model, Small Sample Considerations, Indiana University Bloomington, IN, 47405*, 1998.
[6] Kiefer, N. M., & Vogelsang, T. J., Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation, *Econometrica*, *70*(5), 2002, 2093-2095
[7] Weisberg, Sanford, *Applied linear regression* (John Wiley & Sons, 2005)
[8] Hayes, A. F., & Cai, L., Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior research methods*, *39*(4), (2007), 709-722.
[9] Gujarati, Damodar N., Porter, Dawn C. (Basic Econometrics, Fifth ed., *New York: McGraw-Hill Irwin*, 2009).
[10] Midi, H., Rana, M. S., & Imon, A. R., The performance of robust weighted least squares in the presence of outliers and heteroscedastic errors. *WSEAS Transactions on Mathematics*, *8*(7), 2009, 351-361.
[11] Asteriou, Dimitros; Hall, Stephen G., *Applied Econometrics, Second ed.* (*Palgrave MacMillan*, 2011).
[12] Christopher R. Bilder, Building the regression model III: Remedial measures and validation' *http://www.chrisbilder.com/*., 2012.
[13] Fahrmeir, L., Kneib, T., Lang, S., & Marx, B., *Regression: models, methods and applications* (Springer Science & Business Media, 2013).
[14] Sheather, S. (2009). *A modern Approach to Regression with R*. Springer Science & Business Media.
[15] Midi, H., Rana, S., & Imon, A. H. M. R., On a Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers. *In Proceedings of the World Congress on Engineering* Vol. 1, 2013.
[16] King, G., & Roberts, M. E. How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, *23*(2), 2014, 159-179.
[17] Wilcox, R. R., Linear regression: robust heteroscedastic confidence bands that have some specified simultaneous probability coverage. *Journal of Applied Statistics*, *44*(14), 2017, 2564-2574.