Academia Journal of Scientific Research 5(2): 776-783, December 2017
DOI: 10.15413/ajsr.2018.0128
ISSN 2315-7712
©2017 Academia Publishing

Academia Publishing

CrossMark
click for updates

Research Paper

Regression methods in the presence of heteroscedasticity and outliers

Accepted 27th June, 2017

Adedayo Adepoju A.*, Tayo P. Ogundunmade and Kayode B. Adebayo

Department of Statistics, University of Ibadan, Ibadan, Nigeria.

*Corresponding author. E-mail: pojuday@yahoo.com.

ABSTRACT

It has been observed over the years that real life data are usually non-conforming to the classical linear regression assumptions. One of the stringent assumptions that is unlikely to hold in many applied settings is that of homoscedasticity. When homogenous variance in a normal regression model is not appropriate, invalid standard inference procedure may result from the improper estimation of standard error when the disturbance process in a regression model present heteroscedasticity. When both outliers and heteroscedasticity exist, the inflation of the scale estimate can deteriorate. This study identifies outliers under heteroscedastic errors and seeks to study the performance of four methods; ordinary least squares (OLS), weighted least squares (WLS), robust weighted least squares (RWLS) and logarithmic transformation (Log Transform) methods to estimate the parameters of the regression model in the presence of heteroscedasticity and outliers. Real life data obtained from the Central Bank of Nigeria Bulletin and Monte Carlo simulation were carried out to investigate the performances of these four estimators. The results obtained show that the transformed logarithmic model proved to be the best estimator with minimum standard error followed by the robust weighted least squares. The performance of OLS is the least in this order.

Keywords: Heteroscedasticity, outliers, iteratively reweighted least square, robust weighted least squares, Monte Carlo Simulation.

INTRODUCTION

Heteroscedasticity may arise as a result of the presence of outliers, the inclusion or exclusion of such an observation, especially where the sample size in small, can radically alter the results of the regression analysis. In linear regression analysis the ordinary least squares (OLS) technique is widely used to fit the model mainly because of tradition and ease of computation. Under certain assumptions the OLS estimators possess some very nice and desirable properties. Among the assumptions of the OLS regression model, homoscedasticity is a rather stringent one that is unlikely to hold in many applied settings. Researchers often encounter situations in which the variance of the dependent variable is related to the values of one or more explanatory variables, resulting in heteroscedasticity (Midi et al., 2009; Chatterjee and Hadi, 2006). In such a situation, a variance model based on the explanatory variables can produce weights for the weighted least squares estimator. Weighted least squares, which is a special case of the generalized least square estimator is optimal if the covariance structure of the errors is known, but usually, the error covariance structure is not known in advance. In that case, researchers can use estimated generalized least squares instead.

In the presence of heteroscedasticity, the OLS estimators are still unbiased. However, the most damaging consequence of heteroscedasticity is that the OLS estimator of the parameter covariance matrix (OLSCM), whose diagonal elements are used to estimate the standard errors of the regression coefficients, becomes biased and inconsistent. As a consequence, the *t*-tests for individual coefficients are either too liberal or conservative, depending on the form of heteroscedasticity. White (1980) proposed a heteroscedasticity consistent covariance matrix (HCCM) to solve the consistency problem of the estimator. Theoretically, the use of HCCM allows a researcher to avoid the adverse effect of heteroscedasticity on hypothesis testing even when nothing is known about the form of heteroscedasticity. This powerful method introduced by White (1980) in his classic paper can be traced to the work of Eicker (1963, 1937), Huber (1967), Hartley et al. (1969), Hinkley (1977) and Horn et al. (1975). White's (1980) paper presented the asymptotically justified form of the HCCM later referred to as HC0. In a later paper, MacKinnon and White (1985) raised concerns about the performance of HC0 in small samples and presented three alternative estimators known as HC1, HC2 and HC3. While these estimators are asymptotically equivalent to HC0, they were expected to have superior properties in finite samples, but there is evidence that a few atypical observations (outliers) can make all the estimation and procedures meaningless. In the presence of outliers we have some robust techniques for the detection of heteroscedasticity. Unfortunately, we do not have much robust techniques available in the literature for the estimation of parameters in the presence of heteroscedasticity and outliers. Although heteroscedasticity does not cause any biasness problem to the OLS estimators, the OLS can easily be affected by the presence of outliers. The weighted least squares also suffer the same problem in the presence of outliers and can make a huge interpretive problem in the estimation technique. Generally speaking, none of the estimation techniques work well unless the effect of outliers is eliminated or reduced in a heteroscedastic regression model. Therefore, this problem motivates us to examine the performance of four estimation techniques when heteroscedasticity and outliers occur at the same time in a regression model.

DETECTION OF HETEROSCEASTICITY

Plot the residuals (Gujarati, 2004)

If there is no *a priori* or empirical information about the nature of heteroscedasticity in practice, a regression analysis can be carried out on the assumption that there is no heteroscedasticity and then post-mortem examination of the residuals is done to see if they exhibit in any systematic pattern.

The residual for the *t*th observation $\hat{\mu}_t$ is an unbiased estimate of the unknown and unobservable error for that observation, μ_t. Thus, the squared residuals, $\hat{\mu}_t^2$ can be used as an estimate of the unknown and unobservable error variance, $\sigma_t^2 E\left(\mu_t^2\right)$. The squared residuals can be calculated and then plotted against an explanatory variable that is believed to be related to the error variance. If the error variance is believed to be related to more than one of the explanatory variables, the squared residuals may be plot against each one of the variables. Alternatively, the squared residuals may be plot against the fitted value of the dependent variable obtained from the OLS estimates. Most statistical programs have a command to do these residual plots. It must be emphasized that this is not a formal test for heteroscedasticity. It will only suggest whether heteroscedasticity exist and should therefore not be substituted for a formal test.

Breusch-Pagan/Harvey-Godfrey Test (Breusch and Pagan, 1979; Godfrey, 1978)

Suppose the usual linear relation is given by:

$$y_t = x_t^1 \beta + u_t \quad (1)$$

$$\text{for } t = 1, 2, \ldots, n$$

Where:

$$x_t' = \begin{bmatrix} 1 & x_{2t} & x_{3t} & \cdots & x_{kt} \end{bmatrix}$$

We postulate that all of the assumptions of classical linear regression model are satisfied, except for the assumption of constant error variance, that is, the error variance is non-constant. It is thus assumed that heteroscedasticity takes the form:

$$E(u_t) = 0, \text{ for } \textit{all t}$$
$$\sigma_t^2 = E\left(\mathbf{u}_t^2\right) = h(z_t'\alpha) \quad (2)$$

Where $z_t' = \begin{bmatrix} 1 & z_{2t} & z_{3t} & \cdots & z_{pt} \end{bmatrix}$ is a vector of known variables, $\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_p \end{bmatrix}$ is a vector of unknown coefficients, $h(\cdot)$ is some unspecified function that must take on only positive values.

The null-hypothesis of constant error variance (no heteroscedasticity) can then be expressed as:

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_p = 0$$

White's test (White, 1980)

Suppose that the regression model is given by:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$
(3)

for $t = 1, 2, \ldots, n$

Then, postulating that all of the assumptions of classical linear regression model are satisfied, except for the assumption of constant error variance. For the White's test, assume the error variance has the following general structure:

$$\hat{u}_t^2 = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{2t}^2 + \alpha_5 X_{3t}^2 + \alpha_6 X_{2t} X_{3t} + v_t \qquad (4)$$

for $t = 1, 2, \ldots, n$

Note that all of the explanatory variables are included in the function that describes the error variance and a general functional form is used to describe the structure of the heteroscedasticity, if it exists. The null-hypothesis of constant error variance (no heteroscedasticity) can be expressed as the following restriction on the parameters of the heteroscedasticity equations:

$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_p = 0$ H₀: $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6$ = 0

To test the null-hypothesis of constant error variance (no heteroscedasticity), a Lagrange multiplier test is used.

Park test (Park, 1966)

Park (1966) formalises the graphical method by suggesting that σ_t^2 is some function of the explanatory variable X_t. The functional form suggested was:

$$\sigma_t^2 = \sigma^2 X_t^{\beta} e^{v_t}$$

$$\ln \sigma_t^2 = \ln \sigma^2 + \beta \ln X_t + v_t \qquad (5)$$

Where v_t is the stochastic disturbance term. Since σ_t^2 is generally not known, Park (1966) suggests using $\hat{u}_t^2$ as a proxy and running the following regression:

$$\ln \hat{u}_t^2 = \ln \sigma^2 + \beta \ln X_t + v_t$$

$$= \alpha + \beta \ln X_t + v_t \qquad (6)$$

If β turns out to be statistically significant, it would suggest that heteroscedasticity is present in data. If it turns out to be insignificant, the assumption of homoscedasticity is accepted.

ESTIMATION TECHNIQUES

Weighted Least Squares (WLS) Estimator

The GLS estimator is the same as a weighted least squares estimator. The WLS estimator is the OLS estimator applied to a transformed model that is obtained by multiplying each term on both sides of the regression equation by a weight denoted w_t.

For the given model:

$$w_t Y_t = w_t \beta_1 + \beta_2 (w_t X_{12}) + \beta_3 (w_t X_3) + w_t \mu_t$$

Where $w_t = 1/\sigma_t^2$.

Thus, each observation on each variable is given a weight w_t that is inversely proportional to the standard deviation of the error for that observation. This means that observations with a large error variance are given less weight and observations with a smaller error variance are given more weight in the GLS regression. Therefore, OLS estimation was used for the model earlier mentioned:

$$\beta = (XWX)^{-1} X^{-1} WX$$

Procedure for WLS

The procedure for WLS is given as:

*Regress Y against predictor variable(s) using OLS and obtain error and fitted values of Y;

*Regress absolute value of the error against the predictors or fitted values of Y;

*let s be the fitted values for the regression;

*Define $w_t = 1/s^2{}_t$ for i= l... n;

*Use $\beta = (XWX)^{-1} X^{-1} WX$ as estimated coefficients.

Robust regression

The Robust Weighted Least Squares (RWLS) method is based on the Iteratively Reweighted Least Squares. Iteratively Reweighted Least Squares (IRLS) robust regression uses the weighted least squares procedures to dampen the influence of outlying observations. Instead of weight based on the error variances, IRLS robust regression uses weights based on how far outlying a case is, as measured by the residual for that case. The weights are revised with each iteration until a robust fit was obtained.

Iteratively Reweighted Least Square (IRLS)

The IRLS estimation is computed using the following steps:

1) Choose a weight function for weighting the cases;
2) Obtain a starting weight for all cases;
3) Use the starting weights in weighted least squares and obtain the residuals from the fitted regression function;
4) Use the residuals in step3 to obtain revised weights;
5) Use the iterations until convergence is obtained.

The Transformed Logarithmic Model (LogTransform)

Data do not always come in a form that is immediately suitable for analysis. We often transformed the variables before carrying out the analysis. Transformations are applied to accomplish certain objectives such as to ensure linearity, achieve normality or stabilize the variance. It often becomes necessary to fit a linear regression model to the transformed rather than the original variables. The necessity for transforming the data arises because the original variables or the model in terms of the original variables violates one or more of the standard regression assumptions. The most commonly violated assumptions are those concerning the linearity of the model and the constancy of the error variance. The response variable Y which is analysed may have a probability distribution whose variance is related to the mean. If the mean is related to the value of the predictor variable x, then, the variance of Y will change with X and will not be constant. The distribution of Y will usually also be non-normal under these conditions. Non-normality invalidates the standard tests of significance (although not in a major way with large samples) since they are based on the normality assumptions. The unequal variance of the error terms will produce estimates that are unbiased, but are no longer best in the sense of having the smallest variance. In these situations data are often transformed so as to ensure normality and constancy of error variance. In practice, the transformations are chosen to ensure the constancy of variance. It is a fortunate coincidence that the variance stabilizing transformations are also good normalizing transforms.

DATA ANALYSIS AND SIMULATION STUDY

In order to compare the preceding methods, data obtained from the Central Bank of Nigeria using the three-variable regression model; government expenditure on economic growth disaggregated into recurrent and capital expenditure from 1981 to 2011 were used. Figure 1 and Table 1 shows that the CBN data contain outliers which make it appropriate for this study. The OLS, WLS, Robust regression and Logarithmic transformation were applied to the data. Some results are not presented here due to space limitations. The OLS residual plots of the original data against the fitted values clearly indicate 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. The WLS, Robust regression and Logarithmic transformation methods were applied to this data in order to reduce the effect of the problem of heteroscedasticity. Since the data also contain outliers in addition to the problem of non-constant errors, the study thus examined the behaviours of the estimators to varying degrees of outliers.

The transformed logarithmic model proved to be the best estimator with the minimum standard error followed by the robust weighted least squares. The performance of OLS is the least in this order as expected. The inclusion of the OLS in this study is simply for comparison purpose and to determine whether the OLS can be said to be completely inferior to the other methods to warrant its exclusion from the analysis. Incidentally, the performance of OLS is not different from WLS and RWLS estimators.

Simulation study

Here, a simulation study is presented to assess the performance of these methods. We reuse a model proposed by Lipsitz et al. (1999) and Midi et al. (2009) based on a fixed design matrix. Fifty (50) observations were generated according to linear relation:

$$Y_i = 3 + 2\, x_i + \varepsilon_i$$

Where x_i is uniformly distributed. The first 10 random samples were generated from:

Uniform (10, 1, 9), the second 10 from;
Uniform (10, 10, 19), the third 10 from;
Uniform (10, 20, 29), the fourth 10 from;
Uniform (10, 30, 39) and the fifth 10 from;
Uniform (10, 40, 49).

The error terms were generated such that they will induce heteroscedasticity. In this respect, ε_i is generated according to this relation, $\varepsilon_i = x_i \varepsilon^*$ where ε^* were drawn from standard normal distribution with mean zero and variance. For *n* = 100, we doubled the fix X sample size. We increase the sample size four times to produce sample of size 200. To create outliers in the simulated data, we take the values which are well outside the 3σ distance of the standard normal distribution. For this particular study, we considered the 12σ distance. In this situation, it is more likely that these points would produce big residuals indicating outliers in the data set. The OLS,

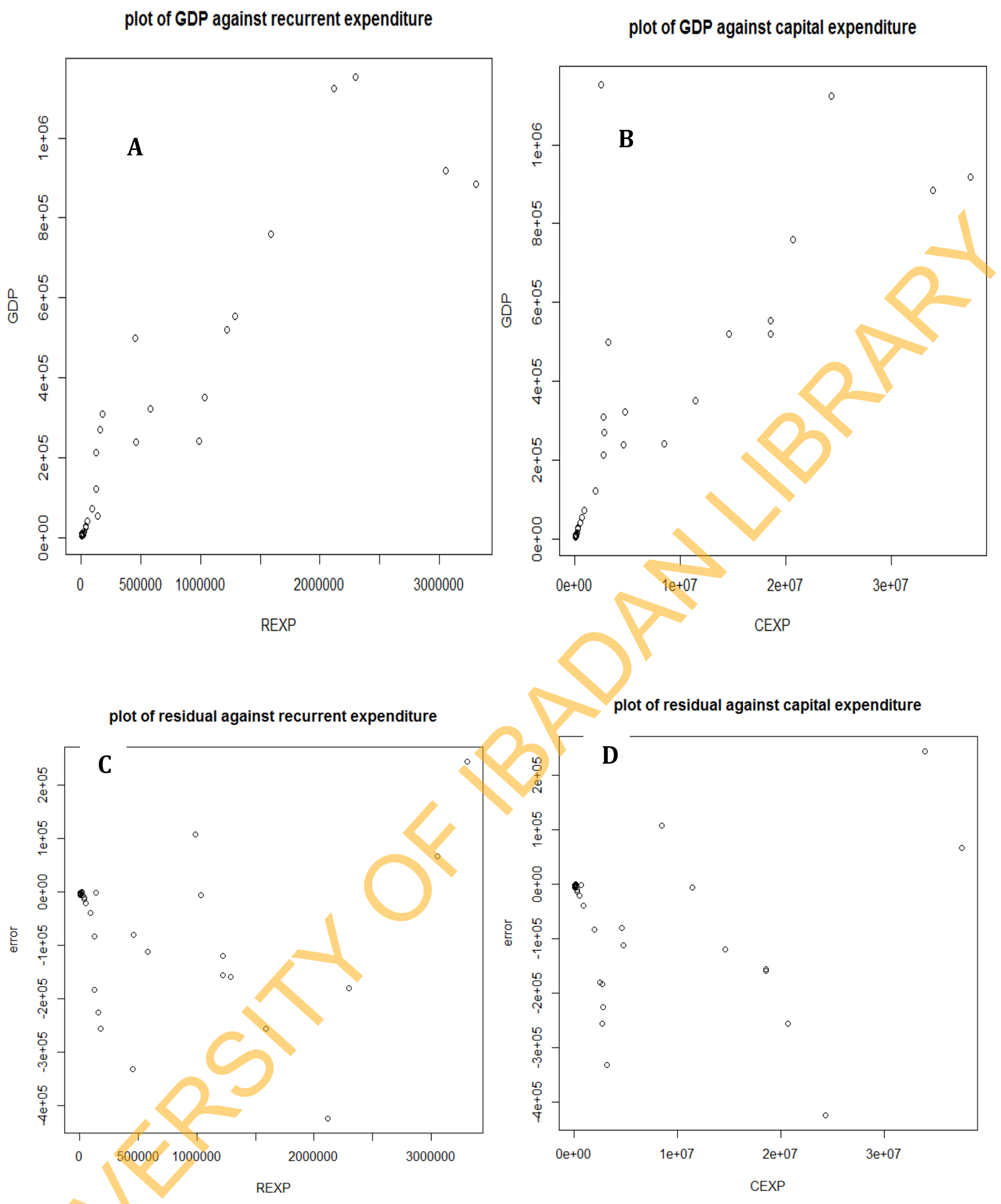

Figure 1: Scatter plots of the CBN data.

WLS, Robust regression and Logarithmic transformation are then applied to the simulated data.

Table 2 presents the average measures of the regression coefficients and their corresponding standard errors, t-statistics and mean square errors for different percentage of outliers and different sample sizes. Several interesting points emerge from this table. The results of WLS suggest that the estimates are not affected by the presence of both heteroscedasticity and outliers and the estimates remain the same at various percent outliers. The presence of heteroscedasticity is expected to retain the unbiased property of the OLS estimates, however, the problem of outliers distort the performance of OLS. As earlier mentioned, our prime interest is to investigate the effect of both outliers and heteroscedasticity on the regression coefficients, standard errors and the t-values.

As the percentage of outliers increases, the log-transform regression estimates increase steadily but not close to the true values and the robust regression estimates are better at 10% outliers. The results also show that the standard errors of the OLS and WLS estimates are larger than for RWLS and log-transform and their *t*-values

Table 1: Summary statistics for the CBN data on government expenditure.

Methods	Parameters	Estimates	SE	t-value	sig
OLS	β_0	-143121	3629420	-0.3943	0.6963
	β_1	10.8087	0.924850	14.91	0.000
	β_2	1.54765	1.93406	0.8002	0.4303
WLS	β_0	-71324	126999	-0.56	0.5788
	β_1	10.28198	0.80274	-12.81	0.0001
	β_2	2.15029	1.57585	1.36	0.1833
RWLS	β_0	-85886	97498	-0.88	0.3859
	β_1	10.02969	0.78988	12.7	0.001
	β_2	2.72219	1.65295	1.65	0.1108
Log linear	β_0	1.43816	0.259353	5.545	0.0000
	β_1	0.706881	0.0831846	8.498	0.0000
	β_2	0.381481	0.0928359	4.109	0.0003

Table 2: Simulated summary statistics for coefficient β (True value = 2).

Methods	Sample Size	Measures	Percentage of Outliers			
			0%	5%	10%	20%
OLS	N= 20	Estimates	1.538	1.05328	1.9923	0.008971
		SE	0.5373	0.000102	0.1143	0.1215
		t-value	2.862	10324.8	17.429	0.074
		MSE	1.031735	0.896273	0.013124	3.978959
	N= 50	Estimates	1.8959	1.7124	2.0178	0.7684
		SE	0.3263	0.1453	0.1966	0.1405
		t-value	5.81	11.787	10.263	5.468
		MSE	0.117309	0.103826	0.038968	0.038968
	N= 100	Estimates	1.8517	1.2657	2.0015	0.77085
		SE	0.2273	0.1167	0.1368	0.09891
		t-value	8.145	10.846	14.626	7.793
		MSE	0.073658	0.552815	0.018716	1.520593
WLS	N= 20	Estimates	2.0872	2.0872	2.0872	2.0872
		SE	0.3738	0.3738	0.3738	0.3738
		t-value	5.584	5.584	5.584	5.584
		MSE	0.007604	0.007604	0.007604	0.007604
	N= 50	Estimates	1.9105	1.9105	1.9105	1.9105
		SE	0.2042	0.2042	0.2042	0.2042

Table 2: Conts. Simulated summary statistics for coefficient β (True value = 2).

		t-value	9.357	9.357	9.357	9.357
		MSE	0.049708	0.049708	0.049708	0.049708
	N= 100	Estimates	1.9682	1.9682	1.9682	1.9682
		SE	0.1307	0.1307	0.1307	0.1307
		t-value	15.056	15.056	15.056	15.056
		MSE	0.018094	0.018094	0.018094	0.018094
RWLS	N= 20	Estimates	1.5576	1.0533	2.2112	0.2113
		SE	0.4379	0.0001	0	0.0001
		t-value	3.5566	13231.35	58928.26	3153.548
		MSE	0.195718	0.909305	0.027306	3.199448
	N= 50	Estimates	1.8439	1.7094	2.2195	0.8691
		SE	0.2235	0.1027	0.0829	0.0371
		t-value	8.2511	16.6377	23.455	26.7879
		MSE	0.074319	0.094996	0.055053	1.280311
	N= 100	Estimates	1.6839	1.6007	2.1654	0.8841
		SE	0.1614	0.0738	0.0596	0.0232
		t-value	10.4311	21.6791	36.3633	38.1198
		MSE	0.125969	0.164887	0.030909	1.245771
Log Transform	N= 20	Estimates	0.609	0.8928	0.94876	0.94061
		SE	0.2366	0.1231	0.5554	0.09395
		t-value	2.574	10.012	11.915	17.084
		MSE	1.9909	1.4136	1.404457	1.1311
	N= 50	Estimates	0.882	0.9409	1.055	1.08343
		SE	0.1214	0.07493	0.1139	0.09334
		t-value	7.262	7.646	9.261	11.607
		MSE	1.264662	1.2315	0.905998	0.848813
	N= 100	Estimates	0.7671	0.81663	0.90372	0.93914
		SE	0.0632	0.06397	0.05804	0.04663
		t-value	12.137	12.766	15.57	20.142
		MSE	0.003994	1.136846	0.230669	0.118874

are relatively small. We also observed that the HCCM estimators suffer the same problem but the results are not presented for brevity.

It is interesting to note that, the RWLS produces unbiased estimates, smaller standard errors and larger t values when compared to the OLS and WLS estimates irrespective of sample sizes and the percentage of outliers in the data. The SE and MSE of the log-transform decrease consistently as sample size increases for all the percentage of outliers considered. The *t-values* of RWLS and log-transform increase as the percentage outlier increases.

CONCLUSION

The main focus of this paper was to investigate the performances of four estimation techniques when both heteroscedastic errors and outliers are present in the data available for analysis. The empirical study reveals that the OLS estimates are easily affected by the presence of outliers and non-constant errors; the WLS has the worst outing because the estimates are the same at all the percentage of outliers considered. Hence, their estimates are not reliable. On the other hand, the log transform estimates emerges to be conspicuously more efficient and more reliable as it is less affected by the effect of outliers and non-constant errors. The results seem to suggest that the log transform method offers a substantial improvement over the other existing methods for handling the problems of outliers and heteroscedastic errors.

REFERENCES

Breusch T, Pagan A (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*. 47:1287–1294.

Chatterjee S, Hadi AS (2006). Regression Analysis by Examples. 4th ed., Wiley, New York.

Cook RD, Weisberg S (1983). Diagnostics for Heteroscedasticity in Regression. Biometrika. 70:1-10.

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

Eicker F (1963). Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear regressions. Annals Math. Stat. 34: 447 – 456.

Eicker F (1967). Limit Theorems for Regressions with Unequal and Dependent Errors. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1, Berkeley University of California Press. 59 – 82.

Godfrey L (1978) Testing for Multiplicative Heteroscedasticity. J. Econometr. 8:227–236.

Gujarati DN (2004). Basic Econometrics. The McGraw-Hill Companies, New York, 4ed.

Hartley HO, Rao JNK, Keifer G (1969). Variance Estimation with on Unit per Stratum. J. Am. Stat. Assoc. 64:841 – 851.

Hinkley DV (1977). Jackknifing in Unbalanced Situations. Technometrics. 19:285- 292.

Horn SD, Horn RA, Duncan DB (1975). Estimating Heteroscedastic Variances in Linear Model. J. Am. Stat. Assoc. 70:380 – 385.

Huber PJ (1967). Robust Statistics, Wiley, New York. 1981.

Lipsitz SR, Ibrahim JG, Parzen M (1999). A Degrees-of-Freedom Approximation for a t-statistic with Heterogeneous Variance. Statistician. 48:495-506.

Long JS, Ervin LH (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. Am. Stat. 54:217-224.

MacKinnon G, White H (1985). Some Heteroscedasticity Consistent Covariance Matrix Estimator with improved finite sample properties. J. Econometr. 29:53- 57.

Maronna RA, Martin RD, Yohai VJ (2006). Robust Statistics -Theory and Methods. Wiley, New York. 2006.

Midi H, Rana MS, Imon AHMR (2009). Robust Estimation of Regression Parameters with Heteroscedastic Errors in the Presence of Outliers. *Proceedings of the 8th WSEAS International Conference on Applied Computer and Applied Computational Science.*

Midi H, Rana S, Rahmatullah A (2009). The Performance of Robust Weighted Least Squares in the Presence of Outliers and Heteroscedasticity Errors. WSEAS Transactions on Mathematics. 8(7):351–360.

Montgomery DC, Peck EA, Vining GG (2001). Introduction to Linear Regression Analysis, 3rd ed, Wiley, New York.

Park RE (1966). Estimation with Heteroscedastic Error Terms, *Econometrica*. 34:4-888.

Rana MS, Midi H, Imon AHMR (2008). A Robust Modification of the Goldfeld-Quandt Test for the Detection of Heteroscedasticity in the Presence of Outliers. J. Math. Stat. 4(4):277-283.

White H (1980). A Heteroskedastic Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity. *Econometrica*. 48:817-838.

Cite this article as:

Adedayo AA, Tayo P (2018). Regression methods in the presence of heteroscedasticity and outliers. Acad. J. Sci. Res. 5(12): 776-783.

Submit your manuscript at
http://www.academiapublishing.org/ajsr