FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 3 No. 2, June, 2019, pp 173- 178

ROBUST WHITE'S TEST FOR HETEROSCEDASTICITY DETECTION IN LINEAR REGRESSION

[1*]Muhammad Sani, [2]Habshah Midi, and [3]Babangida Ibrahim Babura

[1]Department of Mathematical Sciences, Federal University Dutsin-Ma Katsina State, Nigeria.
[2]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Serdang, Malaysia.
[3]Depatment of Mathematics, Faculty of Science, Federal University Dutse, Jigawa State, Nigeria.

Corresponding Author's Email: sanimksoro@gmail.com

ABSTRACT
The existing diagnostic measures for heteroscedasticity incorrectly detect heteroscedasticity in the presence of outlying observations; usually high leverage points (HLPs). The classical White's Test (WT) is the most commonly used diagnostic method for heteroscedasticity in linear regression. The WT does not depend on either normality or prior knowledge of the source of heteroscedasticity. The shortcoming of WT is that in the presence of HLPs it incorrectly detects heteroscedasticity in a data set. In this paper, a Robust White's Test (RWT) has been proposed which is capable of detecting heteroscedasticity in the presence of HLPs. The results based on Monte Carlo simulation study and real data examples show that the proposed RWT correctly detect heteroscedasticity in the presence of HLPs.

Keywords: *Outlier, high leverage point, heteroscedasticity, linear regression.*

INTRODUCTION

The existence of atypical observations which often referred to as outliers is inevitable in real data sets (Hampel et al., 1986). Rousseeuw and Van Zomeren (1990) classified outliers into high leverage points (HLPs) and vertical outliers (VOs). Presence of anomalous observations especially HLPs in a data set invalidate classical statistical inference (Hampel et al., 1986). The Ordinary Least Squares (OLS) is inefficient and produce unreliable estimates even when a single outlying observation is added or present in a data set. Hampel et al. (1986) claimed that a routine data set typically contains about 1–10% outliers and even the highest quality data set cannot be guaranteed to be free of outliers.

Outlying observations are usually responsible for causing heteroscedasticity. Heteroscedasticity occurs when the residual variances of a linear regression model are not constant. In the presence of heteroscedasticity the OLS is still unbiased, but it estimates become inefficient and will not provide reliable inference due to the inconsistency of the variance-covariance matrix. In the literature, many diagnostic plots for heteroscedasticity are now available (Ryan, 1997; Montgomery et al., 2001; Draper and Smith, 2003; Chatterjee and Hadi, 2006; Imon, 2009). Nonetheless, the graphical methods are very subjective. Analytical methods are more effective in detecting the problem of heteroscedasticity.

Several procedures for testing the heteroscedasticity are available in the literature (Goldfeld and Quandt, 1965; Breusch and Pagan, 1980; White, 1980; Cook and Weisberg, 1983; Muller and Zhao, 1995; Diblasi and Bowman, 1997; Cai, et al., 1998). However, most of these tests depend on normality assumption and/or require prior knowledge of what might be the cause of the heteroscedasticity. The White's test (WT) does not depend on either normality or prior knowledge of the source of heteroscedasticity. The shortcoming of WT is that in the presence of outlying observations; usually high leverage points (HLPs) it incorrectly detect heteroscedasticity. HLPs here mean outlying observations in X-direction.

In this paper, we proposed a Robust White Test (RWT) by replacing some components of the WT which are very sensitive to outlying observations by robust alternative to form RWT. The RWT is expected to correctly detect heteroscedasticity in presence of HLPs.

MATERIALS AND METHODS

The White's Test (WT) of White (1980) follows a Chi-square distribution with p degree of freedom and requires two times of minimizing sum of squares residual by using OLS. Firstly, when obtaining the residual of the original regression and secondly when regressing the squared residuals in the auxiliary regression. The OLS has been reported to have been very sensitive and easily affected by outlying observations. To remedy these problems, we replaced the OLS of the original regression with MM-estimates developed by Yohai (1978) which has 50% break down point and 95% efficiency relative to OLS under Gauss-Markov assumptions. The OLS in the auxiliary regression was also replaced by a weighted least squares (WLS) based on GM-FIID weighting method of Sani (2018).

The algorithm of the proposed RWT is summarized as follows:

Step 1: Estimate the regression coefficients using MM-estimator and obtain the residuals (r_i)

Step 2: Obtain the auxiliary regression by regressing the squared residuals (r_i^2) obtained in Step 1 against the original regressors along with their squares and cross product terms using WLS based on GM-FIID weighting function.

Step 3: Obtain the coefficient of determination R_R^2 from the auxiliary regression, given by:

$$R_R^2 = \frac{SSR_R}{SSE_R + SSR_R}$$

where, SSR_R is the sum of squares regression and SSE_R is the sum of squares errors of the auxiliary regression.

Step 4: Reject the null hypothesis (H_0) that the residual variances are constant if $nR_R^2 > \chi^2_{(p,0.05)}$, where *p* is the number of regressors in the auxiliary regression and *n* is the sample size.

The distribution of the Lagrange Multiplier statistics of RWT (nR_R^2) is intractable. However, we anticipated that it's approximately follows Chi-square distribution with *p* degree of freedom (*df*). It is very important to show that the distribution of RWT statistic has similar distribution as the WT statistic. This is an important property for RWT statistic to make it comparable with the WT test in detecting the heteroscedasticity of a data set.

Distribution of RWT Statistic: The Lagrange Multiplier of WT and RWT can be specified as nR^2 and nR_R^2, respectively. To verify the distribution of the Lagrange Multiplier of RWT (nR_R^2). Consider a linear regression with three independent variables.

$$y = 1 + x_1 + 2x_2 + 3x_3 + e \quad (1)$$

where, y is the response variable, x_j, $j = 1,2,3$ are the independent variables generated from standard normal distribution and $e_i \sim N(0, \sigma_i^2)$ with $\sigma_i^2 = \sqrt{2X_2}$ where $i = 1,2,3,\dots,n$. (Draper and Smith 2003). 1000 data points were generated for the sample sizes $n = 50$, 100, 200, 300 and 500. The Lagrange Multiplier of WT and RWT are computed for each sample size. The distribution for the comparison will be Chi-square distribution with 10 degree of freedom (*df*) since our Lagrange Multiplier of WT and RWT has $p = 10$ (number of regressors in the auxiliary regression). The distribution of RWT has been verified using mean and variance, Cramer-von Mises one sample test (Choulakian et al., 1994) and Anderson-Darling test (Rahman et al., 2006).

Monte Carlo simulation: An experiment has been conducted to investigate the effect of HLPs on heteroscedasticity detection; we generate a homoscedastic data with three independent variables as,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad (2)$$

where *y* is the response variable, *X* is the (*n* x 3) vector of explanatory variables. In this simulation study, all the values of *X* and the random errors (ε) were generated from standard normal distributions. The *Y* values are obtained from Equation (2). The HLPs contamination were created by random replacement of a certain percentage of regular observations by a values generated from normal distribution *N*(10,1) for both X and *y* at 5% and 10% HLPs contamination. We run this simulation experiment for six different sample sizes *n* = 20, 50, 100, 150, 200 and 250. To measure the effect of HLPs on heteroscedasticity detection, we calculate the *p*-value of the Lagrange Multiplier for WT and RWT with different sample size and contamination level. We set the level of significance to be 0.05. Subsequently, the average value of LM and *p*-value is recorded. Results are presented in Table 4 which is based on the average of 2,000 replications.

The p-value is the probability of obtaining a test statistic at least as extreme as the one we observed from the sample, if the null hypothesis were true. For a Chi-square distribution the p-value is obtained by R package using the following formula,

$$Pvalue = 1 - pchisq(LM, df) \quad (3)$$

where LM and *df* are the Lagrange Multiplier and degree of freedom, respectively.

Numerical Example1: Education Expenditure Data
This data set is taken from Chatterjee and Hadi (2006). It contains 50 observations with three explanatory variables. Moreover, it has been identified using FIID that observation 49 Alaska(AK) is high leverage points (HLP). The classical and proposed robust White tests were applied to this data set. The result is presented in Table 5.

Numerical Example2: Housing Expenditure data
This data is given by Pindyck and Rubinfeld (1997). It contains 20 observations that give housing expenditure for four different income groups. As expected, people with higher income have relatively more variation in their expenditures on housing. Two HLPs (observation 16 and 17) were identified as HLPs using FIID. We removed these 2 HLPs as shown in Table 6 and apply WT and RWT to investigate the heteroscedasticity.

RESULTS AND DISCUSSION

This section presents the results and discussion of the analysis of this paper.

Table 1: Mean and Variance of WT and RWT Statistic

Tests	Values	Samples				
		n=50	*n=100*	*n=200*	*n=300*	*n=500*
WT	Mean	10.1774	10.1406	10.0671	10.0802	10.0738
	Variance	20.4105	20.3231	20.1052	20.1721	20.1094
RWT	Mean	9.8980	10.2064	10.1988	9.9213	10.1038
	Variance	20.1154	20.1663	20.1340	20.0930	20.1108

Table 1 shows the computed mean and variance of WT and RWT statistics for the different sample sizes considered. The mean and variance of both WT and RWT are fairly closed to the mean (≈ 10) and variance (≈ 20) of the Chi-squares distribution with 10 *df*.

Table 2: Cramer-von Mises One Sample Test for Testing the Distribution of WT and RWT Statistics

Test	Cramèr-von Mises	Samples				
		n=50	*n=100*	*n=200*	*n=300*	*n=500*
WT	*T*	0.0944	0.1030	0.0839	0.1493	0.0984
	p-values	0.1244	0.0944	0.1732	0.0617	0.1279
RWT	*T*	0.0409	0.0939	0.0391	0.0698	0.0975
	p-values	0.6465	0.1266	0.6803	0.2672	0.2306

Table 3: Anderson-Darling Test for Testing the Distribution of WT and RWT Statistics

Test	Anderson-Darling	Samples				
		n=50	*n=100*	*n=200*	*n=300*	*n=500*
WT	A^2	0.0872	0.0944	0.1030	0.0839	0.1493
	p-values	0.1797	0.1244	0.0944	0.1732	0.0217
RWT	A^2	0.0871	0.0409	0.0939	0.0391	0.0698
	p-values	0.1793	0.6465	0.1266	0.6803	0.2672

Table 2 and 3 shows the result of Cramer-von Mises and Anderson-Darling test of no difference between WT and RWT statistic following Chi-square distribution with 10 *df.* It is very interesting to see that all the *p*-values are greater than 0.05 significance level for all the sample sizes considered. This finding shows that WT and RWT statistic are following Chi-square distribution with 10 degree of freedom.

Table 4: Average Lagrange Multiplier (LM) for WT and RWT in a simulated Homoscedastic Data

Samples	Contamination Levels	WT		RWT	
		LM (18.307)	*p-value* (0.05)	LM (18.307)	*p-value* (0.05)
n = 20	Without BLPs	5.5423	0.0849	5.2900	0.1905
	5% BLPs	21.7264	0.0130	5.6196	0.1302
	10% BLPs	20.1148	0.0174	6.3921	0.1971
n = 50	Without BLPs	6.7743	0.06540	5.7681	0.1482
	5% BLPs	19.0929	0.0111	6.0550	0.2933
	10% BLPs	18.8878	0.0226	6.7342	0.1157
n = 100	Without BLPs	6.6951	0.1174	5.4209	0.1890
	5% BLPs	18.7821	0.0230	5.8611	0.1730
	10% BLPs	21.4894	0.0169	6.5379	0.1897
n =150	Without BLPs	6.8358	0.1012	5.4441	0.1886
	5% BLPs	23.7780	0.0216	5.9235	0.3409
	10% BLPs	25.4055	0.0079	6.4090	0.5163

n = 200	Without BLPs	6.9337	0.1735	5.4810	0.3282
	5% BLPs	26.6135	0.0011	5.7335	0.4605
	10% BLPs	24.0140	0.0312	6.3793	0.3496
n = 250	Without BLPs	6.9336	0.5517	5.4333	0.4147
	5% BLPs	31.0287	9.99e-16	5.9703	0.0955
	10% BLPs	27.2550	1.16e-07	6.2696	0.4573

Table 4 shows that in the absence of HLPs, both WT and RWT indicate that the variances of the errors are homoscedastic since all the p-values are greater than $\alpha = 0.05$. Looking at 5% and 10% HLPs contamination for all sample sizes considered, the WT shows presence heteroscedasticity ($p < 0.05$) and RWT still shows the data is homoscedastic. This implies that, with 5% and 10% of HLPs contamination, the WT cannot correctly detect heteroscedasticity in a data set.

Table 2: Heteroscedasticity Diagnostics for Education Expenditure Data

Tests	**WT**		**RWT**	
	$LM = nR^2$ (18.307)	*p*-values (0.05)	$LM = nR^2$ (18.307)	*p*-values (0.05)
Without AK(HLP)	5.7978	0.1219	5.2922	0.1516
With AK(HLP)	22.7817	4.48e-05	5.4339	0.1426

Table 6: Heteroscedasticity Diagnostics for Housing Expenditure Data

Tests	**WT**		**RWT**	
	$LM = nR^2$ (3.841)	*p*-values (0.05)	$LM = nR^2$ (3.841)	*p*-values (0.05)
Without 2 HLPs	3.1454	0.0761	3.0705	0.0797
With 2 HLPs	7.1892	0.0073	3.7557	0.0664

It can be observed that in the absence of HLPs in both Table 5 and 6 the data is homoscedastic. However, the classical WT indicates heteroscedasticity in the presence of HLPs and RWT still shows homoscedasticity which clearly indicates its robustness against the effect HLPs.

CONCLUSION

This paper provides a robust method for the detection of heteroscedasticity in linear regression. The existing method (White's Test) fail to correctly detect heteroscedasticity in the presence of high leverage points (HLPs). Therefore, a Robust White's Test (RWT) to detect the presence of heteroscedasticity in the presence of HLPs has been proposed. A Monte carlo simulation study and real data examples were used to evaluate the performance of the proposed (RWT) method. The results based on the numerical examples and simulation study signifies that the RWT is resistance to the effect of HLPs.

REFERENCES

Breusch, T. and Pagan, A. (1980). A simple test for heteroscedasticity and random coefficient variation, *Econometrica*, 47: 1287–1294

Chatterjee, S. and A. S. Hadi (2006). *Regression Analysis by Example*, Fourth Edition.

Cai, Z., Hurvich, C.M. and Tsai, C.L. (1998). Score tests for heteroscedasticity inwavelet regression, Biometrika. 85: 229–234

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994). Cramér-von Mises statistics for discrete distributions. *Canadian Journal of Statistics*, 22(1), 125-137

Cook, R.D. and Weisberg S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*.70: 1–10

Diblasi, A. and Bowman, A.W. (1997). Testing for constancy of variance. Statistics and Probability,*Letters*. 33: 95–103

Draper, N. R. and Smith, H. (2003). *Applied Regression Analysis.* New York:Wiley.

Goldfeld, S.M. and Quandt, R.E. (1965). Some tests for homoskedasticity.*Journal of the American Statistical Association.* 60: 539–547

Hampel, F. R., Elvezio M. R., Peter J., Rousseeuw and Werner A. S. (1986). Robust Statistics: The Approach Based on Influence Functions. John Wiley and Sons.

Imon, A.H.M.R. (2009). Deletion residuals in the detection of heterogeneity ofvariances in linear regression. *Journal of Applied Statistics*, 36: 347–358

Montgomery, D. C., Peck, E. A. and Viving, G.G. (2001). *Introduction to linear regression Analysis.* 3rd edition. New York: John Wiley and sons

Müller, HG. and Zhao, P.L. (1995). On a semiparametric variance function modeland a test for heteroscedasticity. *Annals of Statistics*. 23: 946–967

Pindyck, S.R and Rubinfeld, L.D. (1997). *Econometric Models and Econometric Forecasts*, 4th Edition. New York: Irwin/McGraw-Hill.

Rahman, M., Pearson, L. M., and Heien, H. C. (2006). A modified anderson-darling test for uniformity. *Bulletin of the Malaysian Mathematical Sciences Society*, 29(1)

Rousseeuw, P. J. and B. C. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of American Statistical Association* 85(411): 633-639

Ryan, T. P. (1997). *Modern Regression Methods*. NewYork: Wiley.

Sani M. (2018) Robust diagnostic and parameter estimation for multiple linear and panel data regression models. Unpublished Ph.D. dissertation, Institute for Mathematical Research, Universiti Putra Malaysia.

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

Yohai, V. J. (1987). "High breakdown-point and high efficiency robust estimates for regression." The Annals of Statistics: 642-656