© A Publication of the Centre for Research and Development (CERAD), The Federal University of Technology, Akure, Nigeria (www.jost.futa.edu.ng) JOST 2021

Type I Error Rates on Some Methods of Heteroscedasticity Detection in Linear Regression Model without Multicollinearity Problem

Alabi, O.O.
Department of Statistics, The Federal University of Technology, Akure, Nigeria
Corresponding Author: Alabi, O.O.: alabioo@futa.edu.ng

ABSTRACT: There are circumstances when assumption of constant error variance is violated in linear regression model. When this happened, heteroscedasticity problem exist. In testing heteroscedasticity, there is need for relatively large samples for us to observe the problems of such errors and evaluate their behaviors. In practice variances of the error term are unequal and unknown in nature, with this knowledge, determination of presence or absence of heteroscedasticity problem in the data set to work with is important. There are several structures of heteroscedasticity which are unknown in nature, likewise, there are several methods for heteroscedasticity detection. In order to always arrive at correct decision one needs to determine the best methods of heteroscedasticity detection to be used. The study involved the conduct of Monte Carlo experiment in a linear regression model with seven different heteroscedasticity structures and multicollinearity level of zero on Seven different sample sizes (n = 15, 20, 30, 40, 50, 100 and 250). The parameters of the model were specified as $\beta_0 = 4$, $\beta_1 = 0.4$, $\beta_2 = 1.5$, $\beta_3 = 3.6$, the various tests were examined at 0.1, 0.05 and 0.01 levels of significance with confidence interval criterion. The study concluded that Breusch and Godfrey (BG) test is the best heteroscedasticity detection method when there exist no multicollinearity in the linear regression model.

Keywords: Regression model, heteroscedasticity, significance levels, type I error rates and confidence interval.

JoST. 2021. 11(2): 78-85
Accepted for Publication: November 03, 2021

INTRODUCTION

Linear regression model postulated the relationship between dependent variable and one predictor variable. However, one of the assumptions of classical linear regression model is that the variance of the error term is constant across observations (homoscedasticity). When homoscedasticity assumption is violated, it then leads to heteroscedasticity. Heteroscedasticity is a major concern in the application of regression analysis, which always occurs in cross sectional data, when the variances of the error terms are no longer constant. The consequences of using the Ordinary Least Square (OLS) estimator to obtain estimates of the population parameters when there is heteroscedasticity includes; inefficient parameter estimates and biased variance estimates which make standard hypothesis tests inappropriate. Another problem often associated with explanatory variables in regression analysis is multicollinearity. OLS estimator in the presence of multicollinearity is unbiased but inefficient. Hawking and Pendleton, (1983) define Multicollinearity as a situation whereby there is an exact or nearly exact linear relation among two or more explanatory variables. The consequences of multicollinearity includes high standard error (Chatterjee *et al*, 2000), inability to estimate the unique effects of individual variables in the regression model, getting large sampling variances which leads to invalid hypothesis (Johnson, 1987). In 2020, Alabi *et al*, observed the effects of multicollinearity on type I error of some methods of detecting heteroscedasticity in Linear regression Model.

This paper is concerned with evaluating a computationally simple asymptotic test that was proposed by Spearman (1904), Rao (1948), Goldfeld and Quandt (1965), Breusch and Godfrey (1966), Park (1966), Glejser (1969), Breusch and Pagan (1979), Harrison and McCabe (1979) and White (1980). These tests, originally designed for structural forms and methods of detecting heteroscedasticity when there is no multicollinearity with various sample sizes under appropriate assumptions.

Methods of Heteroscedasticity

There are many methods of heteroscedasticity in linear regression model. In this work, nine different methods that are commonly used for heteroscedasticity detection were considered. The methods are:

1. Breush-Pagan Test (BPT): Breusch and pagan (1979) developed a test used in examining the presence of heteroscedasticity in a linear regression model. The variance of the error term was tested from a regression and is dependent on the value of the independent variables. Bresuch-Pagan illustrates this test by considering the following:

Given the regression model

$$Y = \beta_0 + \beta_1 x_1 + \mu \qquad (1)$$

Apply OLS on the model in equation (1) and compute the regression residuals, after that perform the auxiliary regression with the equation;

$$\mu_i^2 = y_1 + y_2 z_{2i} + \ldots + y_p z_{pi} + \eta_i \qquad (2)$$

where z could be partly replaced by independent variable x

The test statistic is the result of the coefficient of determination of the auxiliary regression in (2) and sample size n with Lagrange Multiplier (LM) , LM = nR^2. The test statistic is asymptotically distributed as χ^2_{p-1} under the null hypothesis of homoscedasticity.

2. Park Test (PT): Park (1966) propose a Lagrange Multiplier (LM) test, the test assumes the proportionality between error variance and the square of the regressor. According to Gujarati, the Park LM test formulizes the graphical method by suggesting that σ^2 is a particular function of the explanatory variable x_i. Park illustrates this test by regressing the natural log of squared residuals against the independent variable, if the independent variable has a significant coefficient, the data is likely heteroscedastic. Given the model below

$$\sigma^2 = \sigma^2 X_i^{\beta} e^{v} \qquad (3)$$

We need to find the log

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i \qquad (4)$$

Where v_i is the stochastic disturbance term, since σ_i^2 is not known, Park suggest using $\hat{u}_i^2$ as a proxy and run the following regression

$$\ln \mu_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$

$$= \alpha + \beta \ln X_i + v_i \qquad (5)$$

If β turns out to be statistically significant, we then say that heteroscedasticity is present in the data and if it turns out to be insignificant, we may accept the assumption of homoscedasticity.

3. Spearman's Rank Correlation Test (SRCT): Spearman's Rank correlation (1904) assumes that the variance of the disturbance term is either increasing or decreasing as x increases and there will be a correlation between the absolute size of the residuals and the size of x in an OLS regression. The data on x and the residuals are both ranked. The rank correlation coefficient is defined as

$$r_{X,e} = 1 - \left[\frac{6 \sum_i d_i^2}{n(n^2 - 1)} \right]; \qquad -1 \le r \le 1$$

(6)

where d_i is the difference between the rank of x and the rank of e in observations.

4. Glejser Test (GLJT): Glejser (1969) developed a test similar to the Park test, after obtaining the residual $(\hat{u}_i)$ from the OLS regression. Glejser suggest that regressing the absolute value of the estimated residuals on the explanatory variables that is thought to be closely associated with the heteroscedastic variance and attempts to determine whether as the independent variable increase in size, the variance of the observed dependent variable increases. This is done by regressing the error term of the predicted model against the independent variable. A high t-statistic (or low prob-value) for the estimate coefficient of the independent variable(s) would indicate the presence of heteroscedasticity.

5. Goldfeld-Quandt Test (GFQT): Goldfeld - Quandt (1965) developed an alternative test to LM test, applying this test requires to perform a sequence of intermediate stages. First step involves to arrange the observations either is ascending or in descending order. Another step aims to divide the ordered sequence into two equal sub-sequences by omitting an arbitrary number p of the central observation. Consequently, the two equal sub-sequences will summarize each of them a number of $\frac{(n-p)}{2}$ observations. We then compute two different OLS regression the first one for the lowest values of x_i and the second for the highest values of x_i, in addition, obtain the residual sum of squares (RSS) for each regression equation, RSS_1 for the lowest values of x_i and RSS_2 for the highest values of x_i. An F-statistic is calculated based on the following formula:

$$F = RSS_1 / RSS_2 \qquad (7)$$

The F-statistics is distributed with $(N-p-2k)/2$ degrees of freedom for both numerator and denominator. Subsequently, compare the value obtained for the F-statistic with the tabulated values of F-critical for the specified number of degrees of freedom and a certain confidence level. If F-statistic is higher than F-critical, the null hypothesis of homoscedasticity is rejected and the presence of heteroscedasticity is confirmed.

6. Breusch-Godfrey Test (BGT): Breusch-Godfrey (1978) developed a LM test of the null hypothesis of no heteroscedasticity against heteroscedasticity of the form $\sigma_t^2 = \sigma^2 h(z_t'\alpha)$, where z_t is a vector of independent variables. This vector contains the regressors from the original least square regression. The test is performed by completing an auxiliary regression of the squared residuals from the original equation on $(1, z_t)$. The test statistic follows a chi-square distribution with degrees of freedom equal to the number of z under the null hypothesis of no heteroscedasticity.

7. White's Test (WT): White (1980) proposed a statistical test that establishes whether the variance of the error in a regression model is constant. This test is generally, unrestricted and widely used for detecting heteroscedasticity in the residual from a least square regression. Particularly, White test is a test of heteroscedasticity in OLS residual. The null hypothesis is that there is no heteroscedasticity. The procedure for running the test is shows as follows:

Given the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad (8)$$

Estimate equation (8) and obtained the residual $\hat{u}_i$ we then run the following auxiliary regression

$$\hat{u}_i^2 = b_1 + b_2 X_{2i} + b_3 X_{3i} + b_4 X_{2i}^2 + b_5 X_{3i}^2 + b_6 X_{2i} x_{3i} + v_i \qquad (9)$$

The null hypothesis of homoscedasticity is $H_0 : b_1 = b_2 = \cdots = b_m = 0$ where H_0 highlights the fact that the variance of the residual is homoscedasticity i.e. $\mathrm{var}(\varepsilon_i) = Var(Y_i) = \sigma^2$.

The alternative hypothesis is H_1, it aims at the fact that the variance of the residual is heteroscedasticity $\mathrm{var}(\varepsilon_i) = Var(Y_i) = \sigma_i^2$ that is at least one of the $b_i's$ is different from zero. The LM-statistic equal to nR^2, this follows a χ^2 distribution characterized by m-1, where n is the number of observation established to determine the auxiliary regression and R^2 is the coefficient of determination. Finally, we assume to reject the null hypothesis and to highlight the presence of heteroscedasticity when LM-statistic is higher than the critical value.

8. Harrison McCabe Test (HMT): Harrison-McCabe (1979) proposes a test to check the heteroscedasticity of the residuals. The breakpoint in the variances is set by default to the half of the sample. The p-value is estimated-using simulation. If the binary quality measure is false, then the homoscedasticity hypothesis can be rejected with respect to the given level.

9. Non-Constant Variation Score Test (NCVST): Rao (1948), Cox and Hinkley (1974) develop a test of null hypothesis $H_0 : E(\varepsilon^2 / X_{1,} X_{2,} \cdots X_k) = \sigma^2$ against an alternative (H_1) hypothesis with a general functional form.

We recall the central issue is whether $E(\varepsilon^2) = \sigma^2 w_i$ is related to x and x_i. Then, a simple strategy is to use OLS residuals to estimate disturbance and check the relationship between ε_i^2 and x_i and x_i^2.

Suppose that the relationship between ε^2 and x is linear then residual is given by the expression in equation (10)

$$\varepsilon^2 = x\alpha + v \qquad (10)$$

Then, we test $H_0 : \alpha = 0$ against $H_0 : \alpha \neq 0$ and base the test on how the squared OLS residual ε correlate with x.

MATERIALS AND METHOD

Consider the multiple linear regression model of the form:

$$Y_t = \beta_0 + \beta_1 x_{1t} + \beta_{21} x_{2t} + \beta_p x_{pt} + u_t \qquad (11)$$

where, $u_t \sim N(0, \sigma_t^2)$; u_t is the error term and σ_t^2 is the heteroscedasticity variance that is considered. Y_t is the dependent variable, x_{pt} is the explanatory variables that contain multicollinearity and β_p is the regression coefficient of the model.

A Monte Carlo Experiment was performed 1000 times, in generating the data for the simulation study. The Seven (7) error variance containing heteroscedasticity structures considered are;

$$\sigma_t^2 = \sigma^2 (x_{2t}^2)^2 \qquad (12)$$

$$\sigma_t^2 = \sigma^2 (x_{2t}^2) \qquad (13)$$

$$\sigma_t^2 = \sigma^2 (x_{2t}) \qquad (14)$$

$$\sigma_t^2 = \sigma^2 \left[E(y_t)^2 \right] \qquad (15)$$

$$\sigma_t^2 = \sigma^2 \left[E(y_t) \right] \qquad (16)$$

$$\sigma_t^2 = \sigma^2 \left[\exp(\beta_0 + \delta\beta_1 x_{1t} + \delta\beta_2 x_{2t} + \delta\beta_3 x_{3t}) \right] \qquad (17)$$

where $\delta = 0$ and $\delta = 0.2$.

$$\sigma_t^2 = \sigma^2 (1 + X_{2t}^2)^2 \qquad (18)$$

Generation of error term

The error term ε_i was generated to be normally distributed with mean zero and variance σ^2, that is, $\varepsilon_i \sim N(0, \sigma^2)$.

The error term containing different explanatory variables, heteroscedasticity structure and dependent variable were generated.

Generation of explanatory variables

The procedure used by Mansson *et al* (2010), Lukman and Ayinde (2015) and Dorugade (2016), was adopted to generate explanatory variables in this study. This is given as:

$$X_{ti} = (1 - \rho^2)^{0.5} * z_{ti} + \rho * (z_{tp}) \qquad (19)$$

where t =1, 2,...,n and $i = 1, 2, \cdots, p$,

z_{ti} is the independent standard normal distribution with mean zero and unit variance.

Rho (ρ) is the correlation between any two explanatory variables in this study, the multicollinearity level $\rho = 0$, and p is the number of explanatory variables.

Generation of Dependent variables

The dependent variables was generated and was used to carry out the Monte Carlo simulation study. The true values of the model parameters were fixed as follows;

$\beta_0 = 4$, $\beta_1 = 0.4$, $\beta_2 = 1.5$, $\beta_3 = 3.6$.

The sample sizes varied from 15, 20, 30, 40, 50, 100 and 250. At a specified value of n, p, the fixed $x's$ are first generated; followed by the u_t, and the values of Y_t were then determined. Then Y_t and $x's$ were then treated as real life data set while the nine methods described earlier were then applied.

Estimated significance levels

The hypothesis about the methods of detecting heteroscedasticity under different forms of heteroscedasticity structures was tested at (10%, 5% and 1%) levels of significance to examine the (type I error rate) of each error terms. These intervals were referred to as the estimated significance level. The intervals was set to know the number of times each significance level falls between the range set for the confidence interval of each method of detecting heteroscedasticity in order to reject the hypothesis or not. At each level of significance; the interval set for α = 0.1 is (0.09 to 0.14), the interval set for α = 0.05 is (0.045 to 0.054), and the interval set for α = 0.01 is (0.009 to 0.014). Sample sizes were classified as small $(15 \le n \le 30)$, medium $(40 \le n \le 50)$ and large $(100 \le n \le 250)$.

Type I error: This is designated as α, the probability of rejecting null hypothesis (H_0) given that H_0 is true. In this study the null hypothesis (H_0) stated that there exist homoscedasticity while the alternative hypothesis stated that there exist no homoscedasticity. Thus, when the H_0 is rejected when H_0 is not true it implies that we have took a correct decision. Since we have built heteroscedasticity into the model the set H_0 is not true in this case. Hence, the higher the count number of estimated $\hat{\alpha}$ for any particular method over the level of sample sizes, then, the better the method(s).

Determination of the Best Method(s)

At a particular α level a confidence interval was set for 10%, 5% and 1%, the number of times $\hat{\alpha}$ falls in between, the set confidence interval was counted over the sample size and heteroscedasticity structures.

$$\hat{\alpha} = r / R \qquad (20)$$

Where r is the number of times $\hat{\alpha}$ falls in between the confidence interval set at a particular significance level. R is the number of times the experiment was carried out. The heteroscedasticity test with highest number of count is chosen to be the best.

Criterion: The criterion used to determine the best method(s) in this study is confidence interval criterion, it was adopted by considering the number of times the estimated alpha values $\hat{\alpha}$ obtain for any methods that falls in between the set confidence interval. The method(s) with highest total number of counts of estimated alpha values $\hat{\alpha}$ is the best method.

RESULTS AND DISCUSSION

From the simulation study results, it was observed that the number of times the estimated probability of type I error ($\hat{\alpha}$) fall in between the set confidence interval for α = 10%, 5% and 1% was counted over the sample sizes and heteroscedasticity structures for each heteroscedasticity method of detection. The summary of results for all the detection methods, the various sample sizes and the level of significance are presented in Table 1.

Nine methods of heteroscedasticity detection was investigated via hypothesized type I error rate with Monte Carlo study on linear regression model without multicollinearity problem. Here different forms of heteroscedasticity structures were considered at several sample sizes group into small, (i.e (15 ≤ n ≤ 30), medium (40 ≤ n ≤ 50) and large (100 ≤ n ≤ 250)). It was observed under the situation of committing type I error rate that has the overall highest counts when the number of time $\hat{\alpha}$ fall in between the set confidence interval for $\hat{\alpha}$ = 10%, 5% and 1% at all sample sizes and heteroscedasticity structures is Breusch-Godfrey (BGT) method. It should be noted that the bold values in Table 1 indicates the best performed methods of detection under a particular estimated significant level ($\hat{\alpha}$).

Table 1: Summary of Results

$\hat{\alpha}$	Method	Sample size (n)							
		15	**20**	**30**	**40**	**50**	**100**	**250**	**Total**
	BPT	1	0	0	0	0	0	0	1
	PT	0	0	0	0	0	0	0	0
	SRCT	0	1	1	0	0	0	0	2
	GLJT	0	0	0	0	0	0	0	0
0.1	GFQT	1	2	0	3	2	0	0	8
	BGT	5	4	3	3	7	7	6	**35**
	WT	0	0	0	0	3	1	3	7
	HMT	1	2	1	3	2	0	0	9
	NCVST	0	0	0	0	3	1	3	7
	BPT	2	0	0	0	0	0	0	2
	PT	0	0	0	0	0	0	0	0
	SRCT	0	1	1	0	0	0	0	2
	GLJT	0	0	0	0	0	0	0	0
0.05	GFQT	1	1	0	2	0	0	1	5
	BGT	3	1	0	2	0	1	4	**11**
	WT	0	0	0	1	1	0	2	4
	HMT	1	2	0	2	0	0	1	6
	NCVST	3	1	0	0	0	0	0	4
	BPT	0	1	2	0	0	0	0	3
	PT	0	0	0	0	0	0	1	1
	SRCT	0	0	1	0	0	0	0	1
	GLJT	0	0	0	0	0	0	0	0
0.01	GFQT	3	2	0	3	0	0	1	9
	BGT	3	0	2	1	1	6	3	**16**
	WT	0	0	0	0	1	1	1	3
	HMT	0	1	0	3	0	0	1	5
	NCVST	0	1	0	0	0	0	0	1

Source: computed from simulated data.

Result when $\hat{\alpha}$ = 0.1

From Table 1, when $\hat{\alpha}$ = 0.1, it was observed that BGT is generally best, the performances of BGT is better than all other methods at samples 15, 20 and 30. At sample size 40, HMT compete favorably well with BGT to performed equivalently. At sample sizes equal or greater than 50, BGT method outperformed all other methods to be the best method of heteroscedasticity detection. This shows that as sample size increases over all the structural forms of heteroscedasticity, BGT method for heteroscedasticity detection is the best. However, the results suggest that BGT is the best method for heteroscedasticity detection when there is no multicollinearity in the model.

Result when $\hat{\alpha}$ = 0.05

At $\hat{\alpha}$ = 0.05 it was observed that; BGT method is generally the best method for heteroscedasticity detection over all the heteroscedasticity structure.

NCVST compete equivalently well with BGT at sample sizes 15 and 20 to perform best. At sample size 20, HMT compete well and performed equivalently with NCSVT and BGT. At high sample sizes equal or greater than 40, BGT performed best except at sample size is 50, at this instance, all the methods of heteroscedasticity detection performed poorly. Generally, BGT method of heteroscedasticity detection outperformed all other methods of heteroscedasticity detection. However, if multicollinearity does not exist in the model, the results suggest that the performances of BGT $\succ$ HMT $\succ$ GFQT $\succ$ NCVST. Hence, at 0.05 percent level of significance it was observed that BGT is the best heteroscedasticity detection method at all levels of sample sizes and different structure or forms of heteroscedasticity in a linear regression model without multicollinearity problem.

Result when $\hat{\alpha}$ = 0.01

At $\hat{\alpha}$ = 0.01, it was observed that; BG test is generally best over all the structural forms of heteroscedasticity. At sample size 15, BGT and GFQT compete equivalently well to perform best. At sample size 20 and 40, HMT outperformed BGT. At sample size 50, BGT and WT performed equivalently well. At all sample sizes greater than 50, BGT performed best.

However, no matter the forms of heteroscedasticity structure and sample sizes, if multicollinearity does not exist between the explanatory variables of the model, BGT test is the best heteroscedasticity detection method when significance level is 0.01 percent.

Thus, this result suggests that BG test is the best method of heteroscedasticity detection when there is no multicollinearity in the model.

CONCLUSION

This study concluded that, HMT and GFQT can also detect heteroscedasticity as BGT will do in low sample sizes but not as BGT does. Hence, the best method for heteroscedasticity detection in a linear regression model without the problem of multicollinearity for all the structures of heteroscedasticity, at all categories of sample sizes and different levels of significance is BGT.

REFERENCES

Alabi, O. O., Ayinde, K., Babalola, O.E., Bello, H.A and Okon, E. C. (2020). Effects of Multicollinearity on Type I Error of Some Methods of Detecting Heteroscedasticity in Linear Regression Model. Open Journal of Statistics, 10, 664-677.

Breusch, T. S. and Godfrey, L. G. (1978). Misspecification Tests and Their Uses in Econometrics. *Journal of Statistical Planning and Inference.* 49 (2): 241–260. doi:10.1016/0378-3758(95)00039-9.

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

Charles Spearman (1904). Spearman Rank Correlation Coefficient. *International Encyclopedia of the Social Sciences. Copyright* 2008.

Chatterjee, S., Hadi, A. S. and Price, B. (2000). Regression Analysis by Example (Third ed.), John Wiley and Son, Inc. New York.

Cox, D. R and Hinkley, D. V. (1974). Theoretical Statistics. London: Chapman and Hall.

Dorugade, A. V. (2016). New Ridge Parameters for Ridge Regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 1-6.

Glejser, H. (1969). A New Test for Heteroscedasticity. *Journal of the American Statistical Association*. 64(235). 315-323 doi:10.1080/01621459. 1969. 10500976. JSTOR 2283741.

Goldfield, S. M., Quandt, R. E. (1965). Some Test for Homoscedasticity. *Journal of the American Statistical Association*. (310): 539-547

Harrison, M. J. and P. McCabe (1979). A Test for Heteroscedasticity Based on Ordinary Least Square Residuals. *Journal* *of the American Statistical Association* 74, 494-499

Hawking, R. R. and Pendleton, O. J. (1983). The regression dilemma, Commun. Stat.-Theo. Meth, 12, 497-527.

Johnson, J. (1987). Econometrics Methods McGraw-Hill, Auckland.

Lukman, A. F. and Ayinde, K. (2015). "Review and Classifications of the Ridge Parameter Estimation Techniques," *Hacettepe Journal of Mathematics and Statistics,* vol. 46(113): 1- 1. DOI: 10. 15672/HJMS.201815671

Mansson, K., Shukur, G. and Kibria, B. M. G. (2010). On Some Ridge Regression Estimators: A Monte Carlo simulation study under different error variances. *Journal of Statistics*, 17(1), 1-22.

Park, R. E. (1966). Estimation with Heteroscedasticity Error Term. Econometrica 34(4): 888. JSTOR 1910108

Rao, C. R (1948). Large sample test of Statistical Hypothesis concerning several parameters with application to problems of testing. Proc. Comb. Phil. Soc. 44, 50-7.

White, H. (1980). A Heteroscedasticity Consistent Covariance Matrix Estimation and a Direct Test of Heteroscedasticiy. Econometrica*,* 48, pp. 817-818.