Pure and Applied Mathematics Journal
2021; 10(6): 139-149
http://www.sciencepublishinggroup.com/j/pamj
doi: 10.11648/j.pamj.20211006.13
ISSN: 2326-9790 (Print); ISSN: 2326-9812 (Online)

A Modified Breusch–Pagan Test for Detecting Heteroscedasticity in the Presence of Outliers

Bolakale Abdul-Hameed, Oyeyemi Gafar Matanmi

Department of Statistics, University of Ilorin, Ilorin, Nigeria

Email address:
jembaed@gmail.com (B. Abdul-Hameed), gmoyeyemi@gmail.com (O. G. Matanmi)

To cite this article:
Bolakale Abdul-Hameed, Oyeyemi Gafar Matanmi. A Modified Breusch–Pagan Test for Detecting Heteroscedasticity in the Presence of Outliers. *Pure and Applied Mathematics Journal*. Vol. 10, No. 6, 2021, pp. 139-149. doi: 10.11648/j.pamj.20211006.13

Received: November 1, 2021; **Accepted**: November 26, 2021; **Published**: December 29, 2021

Abstract: Heteroscedasticity is a problem that arises in regression analysis for a variety of causes. This problem impacts both the estimation and test procedures and it is therefore critical to be able to detect the problem and address it. The presence of outliers is a regular occurrence in data analysis and the detection of heteroscedasticity in the presence of outliers poses lots of difficulty for most of the existing methods. In this paper, a modified Breusch-Pagan test for heteroscedasticity in the presence of outliers was proposed. The modified test is obtained by substituting non-robust components in the Breusch-Pagan test with robust procedures which makes the modified Breusch-Pagan test to be unaffected by outliers. Monte Carlo simulations and real data sets were used to investigate the performance of the newly proposed test. The probability value (p–value) and power of all methods considered in this study were computed and the results indicate that the modified robust version of Breusch-Pagan test outperformed the previous tests significantly. The proposed modified Breusch-Pagan test is therefore recommended for testing for heteroscedasticity in linear regression diagnosis, especially when the data sets evidently contain outliers.

Keywords: Heteroscedasticity, Outliers, Cook’s Distance, S-estimation, Modified Breusch-Pagan Test, Monte Carlo Simulations

1. Introduction

For decades, heteroscedasticity (or time-dependent volatility) has been recognized in economic and financial time series. If researchers ignore this issue (heteroscedasticity), they may end up with inefficient approaches. Ordinary Least Squares (OLS) has been widely used as an inferential tool in regression over the years. The OLS has some nice and appealing qualities under the normal assumptions. Homogeneity of error variances (homoscedasticity) is one of them, and the OLS estimators have the minimal variance property for it. That is,

$$E(\mu_i{}^2) = \sigma^2 \; i = 1,2,\dots,n$$

However, there are times when the assumption of homoscedastic error variance is violated. When looking at a cross section of enterprises in a single industry, for example, error terms associated with very large firms may have more variance than error terms associated with smaller firms. The condition is known as heteroscedastic of error terms when the error variance changes. When there is a considerable disparity in the sizes of the observations, heteroscedasticity is common. It is critical to identify and address this issue. Otherwise, least squares estimators will still be unbiased, but will not have the minimum variance property, and as a result, the standard errors of the regression coefficients will be greater than necessary. There are various reasons why μ_i's variances may not be constant, which includes asymmetry in the distribution of one or more of the model's regressors, presence of outliers, incorrect data transformation, reduction in experimenter’s error, among others.

A vast variety of diagnostic plots for diagnosing heteroscedasticity are available in the literature. However, because graphical approaches are highly subjective, analytical methods are most often preferred in detection heteroscedasticity. There are also rigorous processes for assessing the homoscedasticity of data in the literatures. Kutner et al. [7] and Chatterjee and Hadi [3] provided comprehensive reviews of various analytical tests for the detection of heteroscedasticity. The majority of these methods rely on least squares residuals, but there is evidence that if outliers are present in the data, these residuals may not exhibit a heteroscedastic pattern. The presence of 1 to 10% outliers in routine data, according to Hampel et al. [6], is more of a norm than an exception. As a result, in the presence of outliers, many analytical tests may suffer from a lack of power.

The formal methods of detecting heteroscedasticity includes the Park test, Spearman's Rank Correlation test (Spearman, [12]), Goldfeld-Quandt test, Breusch–Pagan–Godfrey (BPG) test, White's General Heteroscedasticity test, among others. Among these methods, the Goldfeld-Quandt test has been modified by two authors to detect heteroscedasticity in the presence of outliers. This article proposed and compared the Modified Breusch–Pagan (MBP) test with the famous Goldfeld-Quandt test, Breusch–Pagan–Godfrey (BPG) test, White's General Heteroscedasticity test, the Modified Goldfeld–Quandt (MGQ) test of Rana et al [10] and the Robust Goldfeld–Quandt test (MGQ) of Alih and Ong [1].

2. Methodology

2.1. Goldfeld-Quandt Test

The Goldfeld-Quandt test is applicable if one assumes that the heteroscedastic variance, σ_i^2, is positively related to one of the explanatory variables in the regression model. For simplicity, consider the two-variable model:

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

Suppose σ_i^2 is positively related to X_i as

$$\sigma_i^2 = \sigma^2 X_i^2 \tag{1}$$

where σ^2 is a constant.

Assumption (1) postulates that σ_i^2 is proportional to the square of the X variable.

If (1) is appropriate, it would mean σ_i^2 would be larger, the larger the values of X_i. If that turns out to be the case, heteroscedasticity is most likely to be present in the model. To test this explicitly, Goldfeld and Quandt suggest the following steps:

i. Order or rank the observations according to the values of X_i, beginning with the lowest X value.

ii. Omit c central observations, where c is specified a priori, and divide the remaining (n − c) observations into two groups each of $(\mathrm{n} - \mathrm{c})/2$ observations.

iii. Fit separate OLS regressions to the first $(\mathrm{n} - \mathrm{c})/2$ observations and the last $(\mathrm{n} - \mathrm{c})/2$ observations, and obtain the respective residual sums of squares RSS_1 and RSS_2, RSS_1 representing the RSS from the regression corresponding to the smaller X_i values (the small variance group) and RSS_2 that form the larger X_i values (the large variance group). These RSS each have

$$\frac{(n-c)}{2} - k \ or \ \frac{(n-c-2k)}{2} \ df$$

where k is the number of parameters to be estimated, including the intercept.

iv. Compute the ratio

$$\lambda = \frac{RSS_2/df}{RSS_1/df} \tag{2}$$

If μ_i are assumed to be normally distributed (which we usually do), and if the assumption of homoscedasticity is valid, then λ of (2) follows the F distribution with numerator and denominator df each of $(n - c - 2k)/2$.

If in an application the computed λ (= F) is greater than the critical F at the chosen level of significance, the hypothesis of homoscedasticity is rejected, that is, heteroscedasticity is very likely.(Goldfield and Quandit, [5])

2.2. Breusch–Pagan–Godfrey (BPG) Test

The success of the Goldfeld–Quandt test depends not only on the value of *c* (the number of central observations to be omitted) but also on identifying the correct *X* variable with which to order the observations. The limitation of the Goldfeld–Quandt test can be avoided if we consider the Breusch–Pagan–Godfrey (BPG) test.

To illustrate this test, consider the *k*-variable linear regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i \tag{3}$$

Assume that the error variance σ_i^2 is described as

$$\sigma_i^2 = f\ (\alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}) \tag{4}$$

that is, σ_i^2 is a function of the nonstochastic variables Z's; some or all of the X's can serve as Z's. Specifically, assume that

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} \tag{5}$$

that is, σ_i^2 is a linear function of the Z's. If $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$, $\sigma_i^2 = \alpha_1$, which is a constant. Therefore, to test whether σ_i^2 is homoscedastic, one can test the hypothesis that $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$. This is the basic idea behind the Breusch–Pagan test. The actual test procedure is as follows.

i. Estimate $Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i$ by OLS and obtain the residuals $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n$.

ii. Obtain $\tilde{\sigma}^2 = \sum \hat{\mu}_i{}^2/n$, the maximum likelihood (ML) estimator of σ^2.

iii. Construct variables p_i defined as

$$p_i = \hat{\mu}_i{}^2/\tilde{\sigma}^2$$

which is simply each residual squared divided by $\tilde{\sigma}^2$.

iv. Regress p_i constructed on the Z's as

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} + v_i \tag{6}$$

where v_i is the residual term of this regression.

v. Obtain the ESS (explained sum of squares) from (6) and define

$$\emptyset = \frac{1}{2}(ESS) \tag{7}$$

Assuming μ_i are normally distributed, if there is homoscedasticity and if the sample size n increases indefinitely, then

$$\emptyset \ \widetilde{asy} \ \chi^2_{m-1}$$

that is, ∅ follows the chi-square distribution with (m − 1) degrees of freedom.

Therefore, if in an application the computed ∅ (= χ^2) exceeds the critical χ^2 value at the chosen level of significance, one can reject the hypothesis of homoscedasticity; otherwise one does not reject it. (Breusch and Pagan, [2])

2.3. White's General Test of Heteroscedasticity

Unlike the Goldfeld–Quandt test, which requires reordering the observations with respect to the X variable that is thought to cause heteroscedasticity, or the BPG test, which is sensitive to the normality assumption, White's general test of heteroscedasticity does not rely on the normality assumption and is simple to use.

Consider the following three-variable regression model as an example of the basic concept:

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

The White test procedure is as follows:

i. Given the data, estimate the parameters in (8) and obtain the residuals, $\hat{\mu}_i$.

ii. Run the following (auxiliary) regression:

$$\hat{\mu}_i{}^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}{}^2 + \alpha_5 X_{3i}{}^2 + \alpha_6 X_{2i} X_{3i} + v_i \qquad (9)$$

That is, the squared residuals from the original regression are regressed on the original X variables or regressors, their squared values, and the cross product(s) of the regressors. Higher powers of regressors can also be introduced. Note that there is a constant term in the equation even though the original regression may or may not contain it. Obtain the R^2 from the (auxiliary) regression.

Under the null hypothesis that there is no heteroscedasticity, it can be shown that sample size (n) times the R^2 obtained from the auxiliary regression asymptotically follows the chi-square distribution with degrees of freedom, df equal to the number of regressors (excluding the constant term) in the auxiliary regression. That is,

$$n.R^2 \ \widetilde{asy} \ \chi^2_{df} \qquad (10)$$

where df is as defined previously (i.e., df = *m – 1*). In this case, df = 5, since there are 5 regressors in the auxiliary regression.

If the chi-square value obtained in (10) exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroscedasticity. If it does not exceed the critical chi-square value, there is no heteroscedasticity, which is to say that in the auxiliary regression equation (10), $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$. (White, [13])

2.4. Modified Goldfeld-Quandt Test

Rana et al. [10] presented a novel test that is based on the Goldfeld-Quandt test. The parts of the Goldfeld-Quandt test that are influenced by outliers are first identified and then replaced by more reliable measurements. This test is known as the Modified Goldfeld-Quandt (MGQ) test. The following is the procedure of the modified Goldfeld-Quandt test.

Step 1: Order or rank the observations according to the value of X, starting with the lowest X values, as in the standard Goldfeld-Quandt test.

Step 2: Exclude the centre c observations, where c is predetermined, and divide the remaining (n-c) observations into two groups, each with (n-c)/2 observations.

Step 3: Fit the regression line using Rousseeuw and Leroy [11]'s robust Least Trimmed of Squares (LTS) approach to look for outliers. The deletion residuals for the whole data set should next be computed using a fit that excludes the points indicated as outliers by the LTS fit.

Step 4: Calculate the MSDR (Median of Squared Deletion Residuals) and the ratio for each groups as:

$$\text{MGQ} = \frac{\text{MSDR}_2}{\text{MSDR}_1} \qquad (11)$$

where, MSDR_1 and MSDR_1 are the median of the squared deletion residuals for the smaller and the larger group variances respectively. Under normality, the MGQ statistic follows an F distribution with numerator and denominator degrees of freedom each of (n-c-2k)/2.

2.5. Robust Goldfeld-Quandt Test

Alih and Ong [1] also developed a robust test for heteroscedasticity detection that is resistant to outliers and inherits the OLS's high efficiency. The proposed test procedure which was also a modification of the GQ test was called robust GQ test (RGQ test). The method used an outlier search algorithm to robustify the non-robust component of the GQ test before computing the RGQ score. The RGQ test is a two-phase technique that starts with an outlier identification procedure and then moves on to RGQ-score estimation. Algorithm 1 searches the data for outliers and subsequently removes the data point from the set of observations while algorithm 2 partition the remaining *n* observations in into two groups with each group containing *n/*2 observations, then fit separate OLS to the two groups based on the data set without points identified as outliers in algorithm 1 and finally calculates the RGQ score which is the ratio of the prediction residuals for the two groups based on the OLS fit as

$$\text{RGQ} = \frac{\text{PRESS}_2}{\text{PRESS}_1} \qquad (12)$$

where PRESS_1 and PRESS_2 are the prediction residuals sum of squares for the smaller and the larger group variances, respectively.

The RGQ score follows the *F*-distribution with numerator and denominator degrees of freedom each equals *(n* − 2*k)/*2.

2.6. The Proposed Procedure

Summarily, the proposed modified Breusch–Pagan (MBP) test for detecting heteroscedasticity in the presence of outliers is given in the following steps.

Step 1: Estimate

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i \qquad (13)$$

by OLS and obtain the regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$.

Step 2: Detect the presence of outliers in the data using the Cook's distance

$$D_i \equiv \frac{(\hat{\beta}_{(-i)} - \hat{\beta})' \boldsymbol{X}'\boldsymbol{X}(\hat{\beta}_{(-i)} - \hat{\beta})}{ps^2} \qquad (14)$$

for each of the response variable.

Where $\hat{\beta}$ = (Xl X) – 1 Xl Y (for the entire data set);

$\hat{\beta}_{(-i)}$ = (Xl X) – 1 Xl Y (without the data point of the outlying response variable);

s2 = R′R/(n - p);

R = (R_i) = Y – $\hat{Y}$ = Y – X$\hat{\beta}$ = (I – X (X^l X) $^{-1}$ X^l)Y [the corresponding residual vector]; and

p is the number of regressors. (Cook, [4])

Step 3: Obtain the residual values $e_i = y_i - \hat{y}_i$.

Step 4: Obtain the scaling values

$$\hat{\sigma}_i = \begin{cases} \frac{median|e_i - median(e_i)|}{0.6745}, iteration = 1; \\ \sqrt{\frac{1}{nK}\sum_{i=i}^{n} w_i {e_i}^2}, iteration > 1 \end{cases} \qquad (15)$$

where K = 0.199 and $w_i = w_\sigma(e_i) = \frac{\rho(e)}{e_i^2}$

Step 5: Obtain the scaled residuals μ_i, where

$$\mu_i = \frac{e_i}{\hat{\sigma}_i} \qquad (16)$$

Step 6: Obtain

$$\hat{\varphi}^2 = \sum \hat{\mu}_i{}^2/n \qquad (17)$$

the maximum likelihood (ML) estimator of φ^2.

Step 7: Construct variables f_i defined as

$$f_i = \frac{\hat{\mu}_i{}^2}{\hat{\varphi}^2} - 1 \qquad (18)$$

Step 8: Regress f_i thus constructed on the Z's as

$$f_i = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} + v_i \qquad (19)$$

where v_i is the residual term of this regression and Z's are some or all of the explanatory variables.

Step 9: Compute $\emptyset^*$ from (18), where

$$\emptyset^* = k\left(\frac{n\hat{\alpha}'\boldsymbol{D}\hat{\alpha}}{\xi}\right) \to \chi^2_p\,(\eta) \qquad (20)$$

where $k = \frac{n+p}{\Lambda n}$, Λ is the percentage of outliers;

$$\hat{\alpha} = (Z'Z)^{-1}Z'\hat{f};$$

$$D = Z(Z'Z); \text{ and}$$

$$\xi = 8\sigma^4$$

If the computed $\emptyset^*$ value is greater than the critical $\chi^2_{p,1-\alpha}$ value, then we reject the hypothesis of homoscedasticity; otherwise we do not reject it.

2.7. Criteria for Comparison of Test Statistics

To compare the proposed modified procedure with existing tests for heteroscedasticity, the probability value (p-value) of all test statistics were computed and used to determine which of the test statistic outperforms the other.

3. Data Analysis and Discussion of Results

3.1. Simulation Study

3.1.1. Simple Linear Regression Case

A simple but interesting heteroscedastic variance problem where the variance is the square of the mean of the response variable is considered. Consider a simple linear model:

$$Y = 7 + 3X + \epsilon \qquad (21)$$

The values of X are being taken equally spaced such as 1, 2, …, 10 and these values are replicated two times to get a sample sizes of 20. The random errors from Normal distributions with mean 0 and standard deviations X, 2X and 3X were generated. The outliers in the error term were introduced in every 20th, 10th and 5th positions to generate 5, 10 and 20% outliers respectively. The magnitude of the outlier is 4 times the standard deviation of the original errors.

The six methods of detecting heteroscedasticity under study were then applied to the simulated data. The methods compared are the Modified Breusch–Pagan (MBP) test, the conventional Goldfeld–Quandt (GQ) test, Breusch–Pagan–Godfrey (BPG) test, White's General Heteroscedasticity (WGH) test and two other robust methods called the Modified Goldfeld–Quandt (MGQ) test of Rana et al [10], and the Robust Goldfeld–Quandt test (MGQ) of Alih and Ong [1]. The results obtained are presented in Tables 1 – 3.

Table 1. *Heteroscedasticity diagnostics for simple linear regression simulated data ($\sigma = X$).*

Test procedure	Without outliers		With 5% outliers		With 10% outliers		With 20% outliers	
	Test statistic	P–value	Test statistic	P–value	Test statistic	P–value	Test statistic	P–value
MBP	5.4064	0.0349	14.035	0.0007	9.4303	0.0009	13.939	0.0002
RGQ	5.4401	0.0124	3.8700	0.1128	7.3348	0.0054	3.1189	0.3433
MGQ	9.3250	0.0008	8.0806	0.0254	4.1022	0.4389	3.1249	0.3311
BPG	6.1617	0.0131	7.495	0.0062	6.9559	0.0084	3.209	0.2416
GQ	0.8309	0.6002	0.6207	0.7424	0.4261	0.8755	0.7324	0.6649
WGH	24.258	0.0188	21.942	0.0382	16.272	0.1791	13.845	0.2013

Table 2. *Heteroscedasticity diagnostics for simple linear regression simulated data* ($\sigma = 2X$).

Test procedure	Without outliers		With 5% outliers		With 10% outliers		With 20% outliers	
	Test statistic	**P–value**	**Test statistic**	**P–value**	**Test statistic**	**P–value**	**Test statistic**	**P–value**
MBP	6.1127	0.0134	16.721	0.0012	17.221	0.0004	9.0282	0.0027
RGQ	5.2209	0.0093	7.8767	0.0408	4.2248	0.0784	3.0909	0.4583
MGQ	3.1750	0.3248	9.1806	0.0144	3.1055	0.1889	2.1229	0.0213
BPG	8.3359	0.0039	5.6617	0.0173	4.9478	0.0261	2.0786	0.7801
GQ	0.6573	0.7167	0.6652	0.7112	2.5335	0.1051	1.1837	0.4086
WGH	32.141	0.0013	26.316	0.0097	18.549	0.1000	7.1472	0.1266

Table 3. *Heteroscedasticity diagnostics for simple linear regression simulated data* ($\sigma = 3X$).

Test procedure	Without outliers		With 5% outliers		With 10% outliers		With 20% outliers	
	Test statistic	**P–value**	**Test statistic**	**P–value**	**Test statistic**	**P–value**	**Test statistic**	**P–value**
MBP	6.1127	0.0134	12.215	0.0005	11.426	0.0007	12.739	0.0037
RGQ	14.436	0.0082	10.876	0.0098	4.1228	0.0973	3.7681	0.0986
MGQ	5.8450	0.0248	10.806	0.0034	2.1011	0.0789	1.1109	0.2113
BPG	10.236	0.0014	5.5736	0.0182	8.4362	0.0036	4.8526	0.1276
GQ	0.9876	0.4231	0.7832	0.631	0.321	0.9357	1.1261	0.4353
WGH	41.876	0.0008	21.840	0.0394	25.464	0.0128	2.3219	0.2259

From the results obtained in Tables 1 – 3, the BPG, WGH test, MGQ and RGQ tests were able to detect heteroscedasticity as much as the proposed (MBP test) method could detect it when the data were free of outliers, except the GQ test. However, when applied to data with regression outliers, the conventional BPG and WGH tests are inconsistence in detecting heteroscedasticity at 5 and 10 percent levels of outliers, and could not detect heteroscedasticity at 20 percent. The proposed method was able to detect heteroscedasticity at all levels of outliers, while the RGQ and MGQ tests were able to detect heteroscedasticity in the presence of 5% outliers but became inconsistent at 10% and failed outrightly when the percentage level of outliers is at least 20 percent.

3.1.2. Multiple Linear Regression Case

The study further draws a sample of size 30 for two explanatory variables and a response variable to show the inconsistencies and ineffectiveness of existing and previously modified tests for heteroscedasticity, in the presence of outliers. The data were drawn as follows.

Two predictors were originally generated from a uniform distribution as $x_{1i} \sim U(30,3,6)$ and $x_{2i} \sim$ U(30, 6,12). Their means were obtained as $\bar{x}_{1i} = 4.3898$ and $\bar{x}_{2i} = 8.4754$. The target predictors were then simulated from a normal distribution as $xi_1 \sim N(30, 4.3898, \sigma^2 = x_{11}) + e_i$; $xi_2 \sim N(30, 8.4754, \sigma^2 = x_{21}) + e_i$. The response variable was generated as $y_i = xi_1 + xi_2 + e_{yi}$, where $e_i \sim N(30, 10, 1)$ and $e_{yi} \sim N(30, 0{,}1)$. It is obvious that the data set is heteroscedastic since the variance of the target predictors xi_1 and xi_2 varies from x_{1i} to x_{1n} and x_{2i} to x_{2n} for x_{1i} and x_{2i}, respectively. Then, k-outliers at $xi_1 = x_{i2} = y_i = 10 - 0.5(j - 1), j = 1,\ldots, k$, were planted in the data. The performance of the MBP test, RGQ test, MGQ test, GQ test, BPG test and whites test were evaluated at k = 0%, 10%, 20% and 30% of the sample size and the results are recorded in Table 4.

Table 4. *Heteroscedasticity diagnostics for multiple linear regression simulated data.*

Test procedure	Without outliers		With 10% outliers		With 20% outliers		With 30% outliers	
	Test statistic	**P–value**	**Test statistic**	**P–value**	**Test statistic**	**P–value**	**Test statistic**	**P–value**
MBP	11.123	0.0038	11.091	0.0039	11.041	0.0041	11.002	0.0043
RGQ	7.0099	0.0004	4.0419	0.0698	3.5124	0.0561	2.4621	0.1276
MGQ	12.142	0.0023	1.8098	0.1398	2.3659	0.5132	1.7128	0.1636
BPG	7.3746	0.0044	5.2437	0.0727	5.5885	0.0612	1.3503	0.5091
GQ	12.131	0.0022	1.2702	0.3426	0.6650	0.7548	1.3587	0.3019
Whites	31.443	0.0051	27.471	0.8455	25.939	0.8921	33.986	0.5647

From the results in Table 4, the conventional tests, the MGQ and RGQ tests were able to detect heteroscedasticity as much as the proposed (MBP test) method could detect it when the data were free of outliers. However, when applied to data with regression outliers, the conventional methods as well as the MGQ and RGQ tests failed to detect heteroscedasticity at all levels of outliers, whereas our proposed method was able to detect it.

3.1.3. Simulation Experiment for Level of Heteroscedasticity

In the simulation study, the 'good' observations are generated according to linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \sigma_i \varepsilon_i \; i = 1, 2, \ldots, n$$

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 $\sigma_i^2 = \sigma^2 \exp(a\, x_{1i} + a x_{21}^2)$ where $\sigma^2 = 1$ 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 a is set as a = 0.2, 1.0 and 1.8, which yield $\lambda \approx 0.1$, 0.5 and 0.9 respectively. The values of the regression parameters used in the data generation scheme are $\beta_0 = \beta_1 = \beta_2 = 1$. The contaminated model was then generated. At each step, one 'good' observation is substituted with an outlier. The study focused on the situation where the errors are contaminated normal distribution. To generate a certain percentages of outliers, we use the regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \sigma_i \varepsilon_{i(cont)} \; i = 1, 2, \dots, n$$

where $\varepsilon_{i(cont)} \sim N(0,1) + \mathit{Cauchy}\ (0,10)$. The percentages of outliers were varied, and since Cauchy is a longer tailed distribution, it is believed that the contaminated normal errors produced outliers.

All simulations were performed for m = 1000 replicates for sample sizes n = 20, 40, 60 and 100. A power analysis of the results obtained is presented in Tables 5 to 8.

Table 5. *Simulation results of heteroscedasticity tests for sample size 20 (n = 20).*

Test	$\lambda = 0.1$			$\lambda = 0.5$			$\lambda = 0.9$		
	***k* = 5%**	***k* = 10%**	***k* = 20%**	***k* = 5%**	***k* = 10%**	***k* = 20%**	***k* = 5%**	***k* = 10%**	***k* = 20%**
MBP	0.9929	0.9762	0.9852	0.9987	0.9758	0.9500	0.9858	0.9930	0.9988
RGQ	0.5287	0.2354	0.6745	0.6204	0.5580	0.1234	0.7623	0.4144	0.3230
MGQ	0.8453	0.9526	0.2187	0.8102	0.4351	0.1008	0.8842	0.0322	0.0827
BPG	0.0041	0.0000	0.0000	0.2132	0.0189	0.0000	0.4321	02367	0.0000
GQ	0.8450	0.2824	0.0020	0.8500	0.1423	0.0032	0.8628	0.2451	0.0098
Whites	0.0000	0.0000	0.0000	0.0002	0.0018	0.0000	0.0003	0.0345	0.0000

Table 6. *Simulation results of heteroscedasticity tests for sample size 40 (n = 40).*

Test	$\lambda = 0.1$			$\lambda = 0.5$			$\lambda = 0.9$		
	***k* = 5%**	***k* = 10%**	***k* = 20%**	***k* = 5%**	***k* = 10%**	***k* = 20%**	***k* = 5%**	***k* = 10%**	***k* = 20%**
MBP	0.9856	0.9930	0.8938	0.9750	0.9910	0.9176	0.9772	0.9912	0.9084
RGQ	0.9299	0.8073	0.5342	0.7967	0.8137	0.3125	0.8792	0.6997	0.1023
MGQ	0.8329	0.7761	0.2041	0.9720	0.3824	0.3680	0.9746	0.3680	0.2456
BPG	0.0023	0.0000	0.0000	0.0000	0.0000	0.0065	0.5461	0.0014	0.0000
GQ	0.5783	0.0589	0.0023	0.5836	0.0578	0.0034	0.5712	0.0630	0.0007
Whites	0.0004	0.0000	0.0002	0.0000	0.0003	0.0000	0.0012	0.0000	0.0000

Table 7. *Simulation results of heteroscedasticity tests for sample size 60 (n = 60).*

Test	$\lambda = 0.1$			$\lambda = 0.5$			$\lambda = 0.9$		
	***k* = 5%**	***k* = 10%**	***k* = 20%**	***k* = 5%**	***k* = 10%**	***k* = 20%**	***k* = 5%**	***k* = 10%**	***k* = 20%**
MBP	0.9952	0.9988	0.9998	0.9970	0.9992	0.9871	0.9966	0.9994	1.0000
RGQ	0.9576	0.6781	0.3241	0.8998	0.2313	0.8765	0.8145	0.7915	0.3344
MGQ	0.9456	0.5876	0.0021	0.8000	0.0058	0.3998	0.6314	0.5678	0.0057
BPG	0.2256	0.0031	0.0000	0.0712	0.0000	0.0000	0.3834	0.0094	0.0000
GQ	0.9320	0..0229	0.0051	0.1220	0.0452	0.0039	0.4453	0.3240	0.0098
Whites	0.0000	0.0000	0.0408	0.0000	0.0980	0.0000	0.0000	0.0000	0.0200

Table 8. *Simulation results of heteroscedasticity tests for sample size 100 (n = 100).*

Test	$\lambda = 0.1$			$\lambda = 0.5$			$\lambda = 0.9$		
	***k* = 5%**	***k* = 10%**	***k* = 20%**	***k* = 5%**	***k* = 10%**	***k* = 20%**	***k* = 5%**	***k* = 10%**	***k* = 20%**
MBP	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RGQ	0.9116	0.8003	0.4712	0.9328	0.8145	0.5012	0.7915	0.5801	0.2241
MGQ	0.9507	0.8123	0.0198	0.8714	0.0320	0.0415	0.8801	0.0421	0.0002
BPG	0.0000	0.0043	0.6512	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GQ	1.0000	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000	0.0000	0.0000
Whites	0.0009	0.0000	0.0000	0.0000	0.2020	0.0000	0.0000	0.2014	0.0000

From the results in Tables 5 to 8, all the three conventional methods considered performed poorly in the simulation study; the GQ test performed relatively well for 5% outliers, while the performance of the BPG and WGH tests were very poor. The MGQ and RGQ tests performs very well for 5% level of outliers, inconsistent when k = 10% and very poor when k = 20%.

But the proposed modified procedure, the Modified Breusch-Pagan test performs superbly well throughout. For small sample size (n = 20) and lower contamination (5%), its performance was similar to the GQ, MGQ, and RGQ tests. Also, the MBP test performed like the MGQ and RGQ tests at 10% level of contamination, but its power tends to increase with increase in sample size. It was also observed that the test is robust in the sense that it performs exactly in the same way when outliers occur in a data with different levels of error variances. Thus, the Modified Breusch-Pagan test outperformed the conventional and previously modified tests considered in this study in every respect and is proved to be best overall.

3.2. Numerical Illustrations

Two data sets that have been widely utilized to test for heteroscedasticity in the literature are used for comparison purpose in this section. The proposed procedure is evaluated and compared to other existing methods for detecting heteroscedasticity in regression models using these two data sets as a benchmark. Variables that are heteroscedastic in nature define these data sets. The proposed method was compared with the various approaches under this study before and after outliers were introduced into the data.

3.2.1. Housing Expenditure Data

Pindyck and Rubinfeld [9] described the data in Table 9 in Econometric Models and Economic Forecasts as detailing housing expenditure for four different groups, each with five sample points, resulting in the 20 observations that make up the study data. These data were also utilized to investigate the performance of heteroscedasticity diagnostics in the presence of outliers by Rana et al., [10] and Alih and Ong [1].

Table 9. *Housing expenditure data.*

Index (i)	Income (x_i)	Housing Exp. (y_i)	Index (i)	Income (x_i)	Housing Exp. (y_i)
1	5	1.8		15	4.2
2	5	2		15	4.2
3	5	2		15	4.5
4	5	2		15	4.8
5	5	2.1		15	5
6	10	3.1		20	4.8
7	10	3.2		20	5
8	10	3.5		20	5.7
9	10	3.5		20	6
10	10	3.6		20	6.2

A residual plot of the data in Table 9 as shown in Figure 1 was able to discover heteroscedasticity, since the graph mirrors a typical megaphone shape.

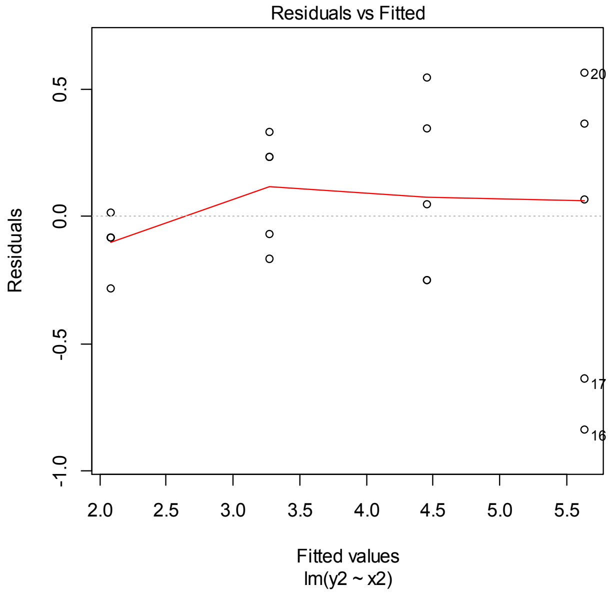

Figure 1. *Fitted values vs. residual plot for housing expenditure original data.*

Now, as indicated in Table 10, observations 1 and 20 were replaced with regression outliers.

Table 10. *Housing expenditure data with 10% outliers.*

Index (i)	Income (x_i)	Housing Exp. (y_i)	Index (i)	Income (x_i)	Housing Exp. (y_i)
1	5	1.8 (4.9)		15	4.2
2	5	2		15	4.2
3	5	2		15	4.5
4	5	2		15	4.8
5	5	2.1		15	5
6	10	3.1		20	4.8
7	10	3.2		20	5
8	10	3.5		20	5.7
9	10	3.5		20	6
10	10	3.6		20	6.2 (2.0)

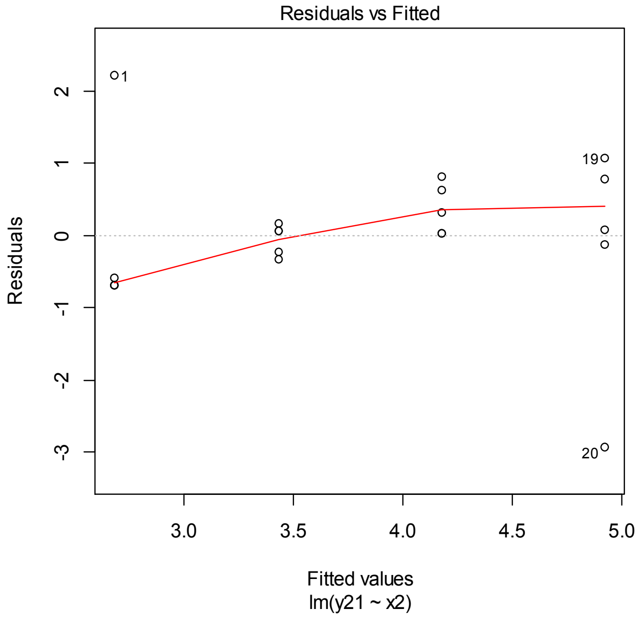

Figure 2. *Fitted values vs. residual plot for housing expenditure data with 10% outliers.*

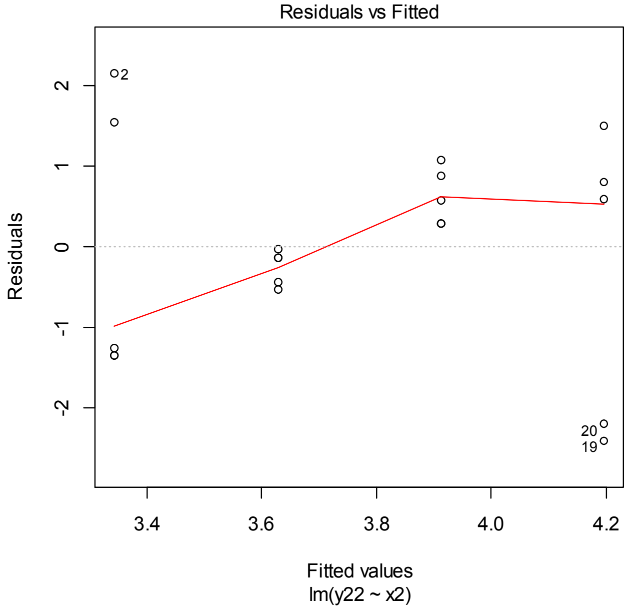

Figure 3. *Fitted values vs. residual plot for housing expenditure data with 20% outliers.*

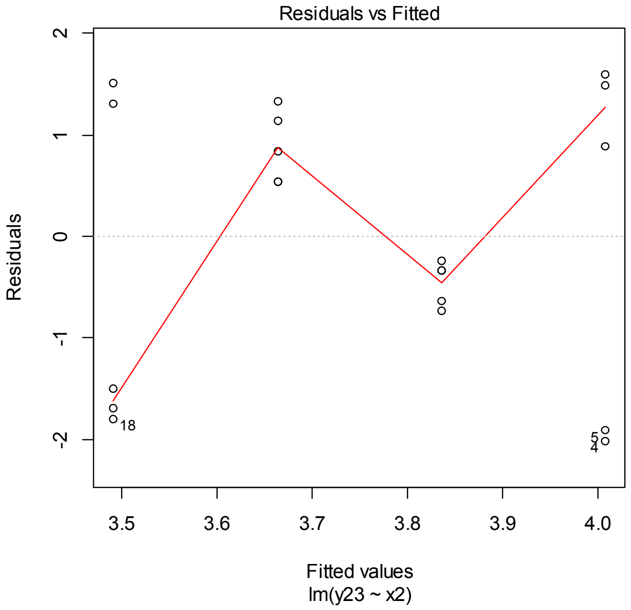

Figure 4. *Fitted values vs. residual plot for housing expenditure data with 30% outliers.*

Due to the existence of outliers planted at index numbers 1 and 20, the megaphone shape in Figure 1 dwindled as shown in Figure 2. As a result, it is reasonable to conclude that the presence of outliers has an impact on diagnostic plots for heteroscedasticity.

Furthermore, two more outliers (for 20% outliers) and another two (for 30% outliers) were planted in the data set in Table 10. Their residual plots are shown in Figures 3 and 4 respectively.

The data was then subjected to the six approaches for detecting heteroscedasticity under investigation, both before and after the outliers were introduced.

Table 11 shows that when the data were free of outliers, the conventional tests, the MGQ and the RGQ tests were all able to detect heteroscedasticity as well as the proposed (MBP test) technique. All conventional tests, however, failed to detect heteroscedasticity at all levels of outliers when applied to data with regression outliers, while the RGQ and MGQ tests were able to detect heteroscedasticity in the presence of only 10% outliers, but failed at 20 and 30 percent levels. But the proposed method (MBP test) was able to detect heteroscedasticity at all levels of outliers.

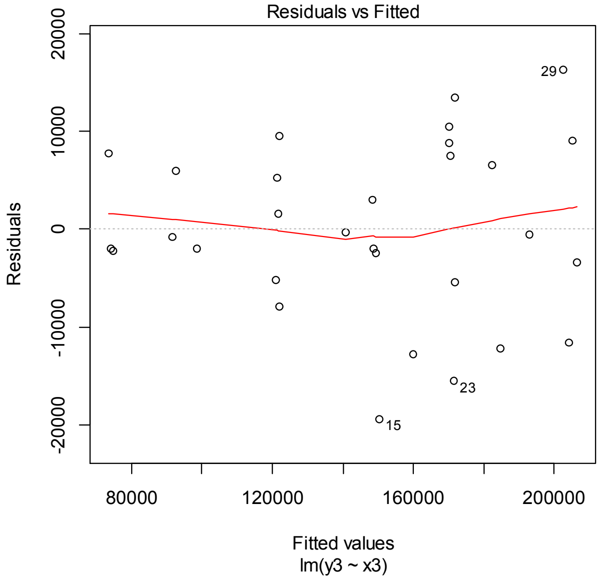

Figure 5. *Fitted values vs. residual plot for Restaurant food sales original data.*

Table 11. *Heteroscedasticity diagnostics for housing expenditure data.*

Test procedure	Without outliers		With 10% outliers		With 20% outliers		With 30% outliers	
	Test statistic	**P–value**	**Test statistic**	**P–value**	**Test statistic**	**P–value**	**Test statistic**	**P–value**
MBP	9.5814	0.0020	9.1395	0.0025	8.4533	0.0036	8.0076	0.0047
RGQ	4.4436	0.0184	5.8767	0.0108	3.1548	0.0954	2.0089	0.4286
MGQ	5.8450	0.0248	10.806	0.0034	2.1011	0.0789	1.1109	0.2113
BPG	7.1921	0.0073	0.8529	0.3557	0.1099	0.7403	0.0002	0.9895
GQ	8.5763	0.0032	1.5558	0.2731	1.1109	0.4427	0.8843	0.5669
Whites	28.929	0.0039	23.073	0.0868	20.642	0.0559	17.497	0.1318

3.2.2. Restaurant Food Sales Data

Montgomery et al. [8]'s data on restaurant food sales as shown in Table 12 in the Introduction to Linear Regression Analysis shows that there is a relationship between revenue and advertising expense. The variance variability is clearly seen in a residual plot of the data in Table 12 (as illustrated in Figure 5). By substituting the income of the cases indexed by 1, 29, and 30 with regression outliers as shown in Table 13, three outliers were purposefully introduced into the data set (modified values are presented within the parentheses).

Table 12. *Restaurant food sales data.*

Index (*i*)	Adv. Exp. (x_i)	Income (y_i)	Index (*i*)	Adv. Exp. (x_i)	Income (y_i)
1	3000	81464		12310	146630
2	3150	72661		13700	147041
3	3085	72344		15000	179021
4	5225	90743		15175	166200
5	5350	98588		14995	180732
6	6090	96507		15050	178187
7	8925	126574		15200	185304

Index (i)	Adv. Exp. (x_i)	Income (y_i)	Index (i)	Adv. Exp. (x_i)	Income (y_i)
8	9015	114133		15150	155931
9	8885	115814		16800	172579
10	8950	123181		16500	188851
11	9000	131434		17830	192424
12	11345	140564		19500	203112
13	12275	151352		19,200	192482
14	12400	146926		19000	218715
15	12525	130963		19350	214317

Table 13. *Restaurant food sales data with 10% outliers.*

Index (i)	Adv. Exp. (x_i)	Income (y_i)	Index (i)	Adv. Exp. (x_i)	Income (y_i)
1	3000	81464 (**814644**)		12310	146630 (**546630**)
2	3150	72661		13700	147041
3	3085	72344		15000	179021
4	5225	90743		15175	166200
5	5350	98588		14995	180732
6	6090	96507		15050	178187
7	8925	126574		15200	185304
8	9015	114133		15150	155931
9	8885	115814		16800	172579
10	8950	123181		16500	188851
11	9000	131434		17830	192424
12	11345	140564		19500	203112
13	12275	151352		19,200	192482
14	12400	146926		19000	218715
15	12525	130963		19350	214317 (**21431**)

The effect of outliers planted at index 1, 16 and 30 in the above data shows that there is no more heteroscedasticity in the data, from the residual plot obtained, as shown in Figure 6.

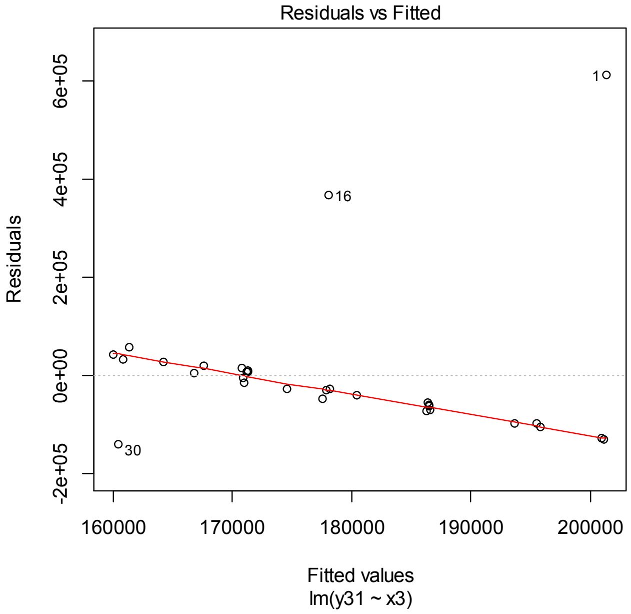

Figure 6. *Fitted values vs. residual plot for Restaurant food sales data with 10% outliers.*

Three more outliers (for 20% outliers) were further planted in the data set in Table 13 and another three (for 30% outliers). Their residual plots are shown in Figures 7 and 8 respectively.

The results obtained in Table 14 shows that the conventional tests, the MGQ and RGQ tests were able to detect heteroscedasticity as much as the proposed MBP test could detect it when the data were free of outliers. However, when applied to data with regression outliers, the GQ test failed to detect heteroscedasticity at all levels of outliers. The proposed method and the BPG tests were able to detect heteroscedasticity at all levels of outliers, while the RGQ and MGQ tests were able to detect heteroscedasticity in the presence of 10% outliers, but failed at 20 and 30 percent levels. The WGH test was able to detect heteroscedasticity at 20 and 30 percents level of outliers, but failed at 10%.

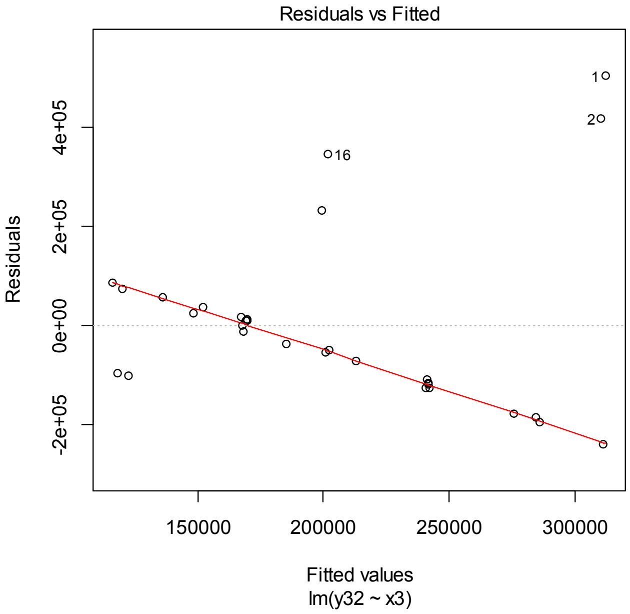

Figure 7. *Fitted values vs. residual plot for Restaurant food sales data with 20% outliers.*

Table 14. *Heteroscedasticity diagnostics for Restaurant food sales data.*

Test procedure	Without outliers		With 10% outliers		With 20% outliers		With 30% outliers	
	Test statistic	P–value	Test statistic	P–value	Test statistic	P–value	Test statistic	P–value
MBP	8.8731	0.0028	8.7421	0.0031	8.1956	0.0041	8.0896	0.0045
RGQ	5.1179	0.0021	4.2389	0.0080	3.1194	0.0838	2.1019	0.1230
MGQ	4.9917	0.0090	10.457	0.0005	1.1189	0.3501	2.1231	0.2416
BPG	25.534	4.3e-07	10.737	0.0011	11.281	0.0007	14.067	0.0002
GQ	14.878	1.0e-05	0.3641	0.9601	0.2627	0.9889	0.3697	0.9578
Whites	34.629	0.0005	20.356	0.0606	36.433	0.0003	39.160	0.0001

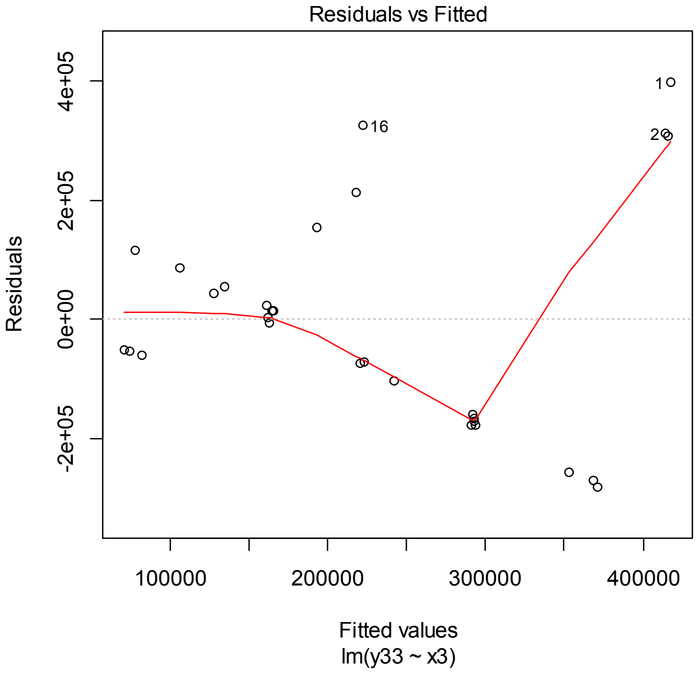

Figure 8. *Fitted values vs. residual plot for Restaurant food sales data with 30% outliers.*

4. Conclusion

The impact of outliers on the detection of heteroscedasticity in a data set by existing methods in literature was addressed in this study. The study further examined two modified procedures that have significantly improved the detection of heteroscedasticity in the presence of outliers but, however found them to be inconsistence or unable to detect heteroscedasticity especially when the level of outliers is greater than 10%. Thus, the research developed an algorithm by modifying the Breusch-Pagan test to suitably detect heteroscedasticity whether or not outliers are present in a data set at any level of contamination. The method was compared with existing and previously modified methods using Monte Carlo simulations and real life data, and a significant improvement was achieved.

References

[1] Alih, E., & Ong, H. C. (2015): An outlier-resistant test for heteroscedasticity in linear models. Journal of Applied Statistics, 42 (8), 1617–1634.

[2] Breusch, T. S. and Pagan A. R. (1979): A simple test for heteroscedasticity and random coefficient variation, Econometrica 47, 1287-1294.

[3] Chatterjee, S. and Hadi A. S. (2006): Regression Analysis by Examples. 4th Edition, Wiley, New York.

[4] Cook, R. D. (1977): Detection of Influential Observations in Linear Regression, Technometrics. American Statistical Association. 19 (1): 15–18. doi: 10.2307/1268249. JSTOR 1268249. MR 0436478.

[5] Goldfeld, S. M. and Quandt R. E., (1965): Some tests for homoskedasticity. J. Am. Stat. Assoc., 60: 539-547. http://www.belkcollege.uncc.edu/cdepken/econ6090/readings/goldfeld-quandt-1965.pdf

[6] Hampel, F. R., Ronchetti E. M., Rousseeuw P. J. and Stahel W. (1986): Robust Statistics: The Approach Based on Influence Function. 1st Edition, Wiley, New York, pp: 536. ISBN: 0471735779.

[7] Kutner M. H., Nachtsheim C. J. and Neter J. (2004): Applied Linear Regression Models. 4th Edition, McGraw-Hill/Irwin, New York, pp: 701. ISBN: 0-07-301344-7.

[8] Montgomery D., Peck E., and Vining G. (2001): Introduction to Linear Regression Analysis, Student Solutions Manual, Wiley Series in Probability and Statistics, Wiley, New York.

[9] Pindyck R. and Rubinfeld D. (1998): Econometric Models and Economic Forecasts (Text Alone), Econometric Models and Economic Forecasts, McGraw-Hill Companies, New York.

[10] Rana M. S., Midi H. and 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, 2008.

[11] Rousseeuw, P. J. and Leroy A. (1987): Robust Regression and Outlier Detection, 1st Edition, Wiley, New York, pp: 329. ISBN: 0471852333.

[12] Spearman C. (1904): The proof and measurement of association between two things, American Journal of Psychology. 15 (1): 72–101. doi: 10.2307/1412159.

[13] White H. (1980): A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econ.: J. Econ. Soc., pp. 817–838.