International Journal of Statistics and Probability; Vol. 8, No. 2; March 2019
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Iterative Approaches to Handling Heteroscedasticity With Partially Known Error Variances

Morteza Marzjarani

Correspondence: Morteza Marzjarani, National Marine Fisheries Service, Southeast Fisheries Science Center, Galveston Laboratory, 4700 Avenue U, Galveston, Texas 77551, USA.

Received: January 30, 2019 Accepted: February 20, 2019 Online Published: February 22, 2019

doi:10.5539/ijsp.v8n2p159 URL: https://doi.org/10.5539/ijsp.v8n2p159

Abstract

Heteroscedasticity plays an important role in data analysis. In this article, this issue along with a few different approaches for handling heteroscedasticity are presented. First, an iterative weighted least square (IRLS) and an iterative feasible generalized least square (IFGLS) are deployed and proper weights for reducing heteroscedasticity are determined. Next, a new approach for handling heteroscedasticity is introduced. In this approach, through fitting a multiple linear regression (MLR) model or a general linear model (GLM) to a sufficiently large data set, the data is divided into two parts through the inspection of the residuals based on the results of testing for heteroscedasticity, or via simulations. The first part contains the records where the absolute values of the residuals could be assumed small enough to the point that heteroscedasticity would be ignorable. Under this assumption, the error variances are small and close to their neighboring points. Such error variances could be assumed known (but, not necessarily equal).The second or the remaining portion of the said data is categorized as heteroscedastic. Through real data sets, it is concluded that this approach reduces the number of unusual (such as influential) data points suggested for further inspection and more importantly, it will lowers the root MSE (RMSE) resulting in a more robust set of parameter estimates.

Keywords: data partitioning, partial heteroscedastic data, handling heteroscedasticity

1. Introduction

Heteroscedasticity is an important issue in modeling where the existence of which is often ignored by researchers. Generally, it is the result of violating other assumptions. Heteroscedasticity gives the same weight to all the observations disregarding the possibility of some observations having larger error variances and containing less information about the predictor (s). Because of heteroscedasticity, least square estimates are no longer BLUE, significant tests will run either too high or too low, and standard errors and confidence intervals will be biased.

Several authors have addressed this issue, some at depth. Breusch and Pagan (1979) addressed heteroscedasticity and developed a method known as Breusch-Pagan, hereafter called the B-P test, where a Lagrange Multiplier (LM) generates the test statistic for testing its existence in a data set. White (1980) modified the method by assuming that the error terms were not necessarily normal also included the non-linear heteroscedasticity in his approach. As a result, this method generates a larger number of terms such as cross multiplication of the terms and higher degrees of freedom. This topic was addressed in Kalirajan (1989) for the usual regression model without replication giving a diagnostic test for heteroscedasticity based on the score statistic. In that article, the author also discussed an alternative and a relatively easier test for heteroscedasticity and non-normality of regression residuals without having a priori information on the source of heteroscedasticity and non-normality.

Koenker (1981) published a note on studentizing a test for heteroscedasticity. This note derives the asymptotic distribution of the B-P test under sequences of contiguous alternatives to the null hypothesis of homoscedasticity. A nonparametric hypothesis test for heteroscedasticity in multiple regressions was developed in Zambom and Kim (2017). Through extensive simulations, they concluded that while commonly used methods fail in some cases, the proposed test detects heteroscedasticity in all models under consideration. Zhou, et al. (2015) addressed the covariates associated with heteroscedastic error variances. A local polynomial estimation of heteroscedasticity in a multivariate linear regression model and its application to economics data was presented in Su, et al. (2012). Homogeneity of variances is a standard assumption in regression analysis. However, this assumption is not always true or appropriate. Spiegelman, et al. (2011) proposed an estimator for correcting regression coefficients for covariate measurement error with heteroscedastic variance and derived point and interval estimates. A score test for heteroscedasticity in linear regression model was discussed in Cook and Weisberg (1983).

More recently, Marzjarani (2018[a]) addressed heteroscedasticity in the shrimp data in the Gulf of Mexico (GOM) for the years 1984 through 2001 using weighted least square (WLS) and feasible generalized least square (FGLS) methods. Heteroscedasticity can be improved by giving a proper weight to the model. Methods for such assignment include WLS, generalized least square (GLS) and FGLS. Some authors have extended WLS to an iterative WLS (IRLS) (See Hooper, 1993, for example).

In this article, a few approaches for handling heteroscedasticity are presented. First, an iterative weighted least square (IRLS) is deployed and proper weights for reducing heteroscedasticity are determined. In addition, an iterative feasible generalized least square (IFGLS) is used to determine the weight (s) for improving heteroscedasticity. Next, a new approach for reducing heteroscedasticity is proposed where a given data set is divided into two sets, one with known (not necessarily equal) error variances and the other with unknown error variances. The effectiveness of this approached is measured through the change (reduction) in the root MSE. The root MSE (RMSE) is the square root of the variance of the residuals. It indicates how close the observed data points are to the model's predicted values. RMSE can be thought of the standard deviation of the residuals, it is in the same units as the response variable, and easier to interpret. Unlike MSE, RMSE is linear and it is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction. Models with lower values of RMSE are better-fit or preferred representations of given data sets.

2. Methodology

Large data sets provide a luxury to the researchers. When fitting a multiple linear regression model (MLR) or a general linear model (GLM), a large data set could be divided into two parts, but still analyzed as one set: The part where the absolute values of the residuals could be assumed small enough to the point that heteroscedasticity would be ignorable. That is, residuals in this part are small and close to their neighboring points and therefore, the error variances could be assumed small and/or known, hereafter called "known variances" portion. This portion if not homoscedastic, it will handled through the WLS method. The second portion would consist of all the remaining data where error variances are unknown. To support this theoretical concept, real data sets are used and it is shown that this new approach constructively helps with the way heteroscedasticity is handled.

The following two issues related to the heteroscedasticity are considered.

1. The model considered in this research is an MLR or a GLM of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i,\ i=1, 2, 3, \ldots, n, \tag{1}$$

where, y_i is the response, $\beta_0$is the constant term, X_{ij}'s are the regressors, and ε_i's are the error terms. Although, not customary for MLR, just for the purpose of simplicity and convenience, the above model is written in a matrix form as:

$$\underline{y} = \underline{x}\underline{\beta} + \underline{\varepsilon} \tag{2}$$

where, $\underline{y}$ is a *n x 1*column matrix of the response variable and $\underline{x}$ is a *n x p* matrix of repressors relating the vector of responses $\underline{y}$, and $\underline{\varepsilon}$is a *n x 1* matrix of the error term. The vector $\underline{\varepsilon}$ is assumed to have $E(\underline{\varepsilon}) = \underline{0}$ and $Var(\underline{\varepsilon}) = \underline{\Omega}$.

It is assumed that some of the observations have known (but not necessarily equal) error variances. In other words, a subset of the data set (s) under consideration with low heteroscedasticity is selected and treated as homoscedastic or heteroscedastic with known variances. In what follows, it is assumed that the matrix $\underline{\Omega}$ is a diagonal matrix with unknown and known elements. Under this assumption, and WOLG through rearranging the records, the vector $\underline{\varepsilon}$can be divided into two vectors $\underline{\varepsilon}_1$ and $\underline{\varepsilon}_2$, where the first corresponds to the known and the second corresponds to the unknown error variances. That is,

$$\underline{\varepsilon} = \begin{pmatrix} \underline{\varepsilon 1} \\ \underline{\varepsilon 2} \end{pmatrix}, \tag{3}$$

Under the above assumption, the model given in (2) can be rewritten as:

$$\begin{pmatrix} \underline{y1} \\ \underline{y2} \end{pmatrix} = \begin{pmatrix} \underline{x1} \\ \underline{x2} \end{pmatrix} \underline{\beta} + \begin{pmatrix} \underline{\varepsilon 1} \\ \underline{\varepsilon 2} \end{pmatrix}, \tag{4}$$

where the terms with the subscripts 1and 2 represent the parts corresponding to the known and unknown error variances respectively. In this representation,$\underline{y}_1$ and $\underline{y}_2$ are *q x 1* and *n-q x 1* column vectors, $\underline{x}_1$ and $\underline{x}_2$ are *q x p* and *n-q x p* matrices, $\underline{\beta}$is a *p x1* column matrix, where q is the number of records assumed to be homoscedastic or heteroscedastic with known error variances (not necessarily equal). Also, $\underline{\varepsilon}_1$ and $\underline{\varepsilon}_2$ represent the decomposition of the error term $\underline{\varepsilon}$ in the model. As shown in Marzjarani (2018[b]), heteroscedasticity due to the known or unknown error variances can be handled through the applications of WLS or FGLS respectively.

Since the elements of $\underline{\varepsilon}_1$ have known variances (equal or unequal), a WLS method is applied. The weight for this

method is similar to the one used by Marzjarani (2018[b]). That is:

$$\underline{res} = |\underline{y} - \underline{x}\underline{\hat{\beta}}|,$$
$$\underline{res} = \underline{x}\underline{\gamma} + \underline{\tau},$$
$$\underline{weight} = 1/(\underline{x}\underline{\hat{\gamma}}). \quad (5)$$

For the second part or the unknown portion of the model in (2) or (4), the weight is determined through the application of the FGLS method defined below.

$$\underline{lnressq} = ln(\underline{y} - \underline{x}\underline{\hat{\beta}})^2,$$
$$\underline{lnressq} = \underline{x}\underline{\gamma} + \underline{\tau},$$
$$\underline{weight} = 1/exp(\underline{x}\underline{\hat{\gamma}}). \quad (6)$$

In these formulas, $\underline{res}$ and $\underline{lnressq}$arevectors of the absolute value of the residuals and the natural logarithm of the residuals squared respectively.

2. The IRLS and IFGLS are applied to the model defined in (2) or (4) iteratively and the average of the weights over iterations are calculated and used to reduce the model heteroscedasticity. The iterative algorithms for IRLS and IFGLS are given below:

begin

1. $\underline{res} = |\underline{y} - \underline{x}\underline{\gamma}|$
2. $\underline{w}(\underline{\beta}_{(1)}) = 1/(\underline{x}\underline{\hat{\gamma}})$
3. $\underline{res}_t = |\underline{w}(\underline{\hat{\beta}}_{(t-1)})(\underline{y} - \underline{x}\underline{\beta})|$
4. $\underline{w}(\underline{\hat{\beta}})_{(t)} = 1/(\underline{res}_t - \underline{x}\underline{\hat{\gamma}})$

Exit if convergence is satisfied, else goto step 3

end.

begin

1. $\underline{lnressq} = ln(\underline{y} - \underline{x}\underline{\hat{\beta}})^2$
2. $\underline{lnressq} = \underline{y} - \underline{x}\underline{\gamma}$
3. $\underline{w}(\underline{\beta}_{(1)}) = 1/exp(x \quad \hat{\gamma})$
4. $\underline{lnressq}_t = (\underline{w}(\beta_{(t-1)})(\underline{y} - \underline{x}\underline{\beta}))^2$
5. $\underline{w}(\underline{\hat{\beta}})_{(t)} = 1/exp(\underline{lnressq}_t - \underline{x}\underline{\hat{\gamma}})$

Exit if convergence is satisfied, else goto step 4

end. (7)

The convergence is checked through the linear convergence. By definition, the sequence $x_1, x_2, x_3, \ldots, x_n$ converges linearly to the value a if there exists a real number b such that

$$lim_{n\to\infty}\ (|x_{n+1} - a|/|x_n - a|) = b \quad (8)$$

Assuming that there is a sufficiently large data set, the following scenarios are considered:

a. All the data are analyzed under the assumption of heteroscedasticity.

b. All the data are analyzed assuming that all the error variances are known, but not necessarily equal (heteroscedastic).

c. All the data are analyzed assuming that all error variances are unknown (heteroscedastic).

d. The data set is divided into two parts: the first with known error variances and the second with unknown error variances.

Figure 1 is a hierarchical graph displaying the flow of the process.

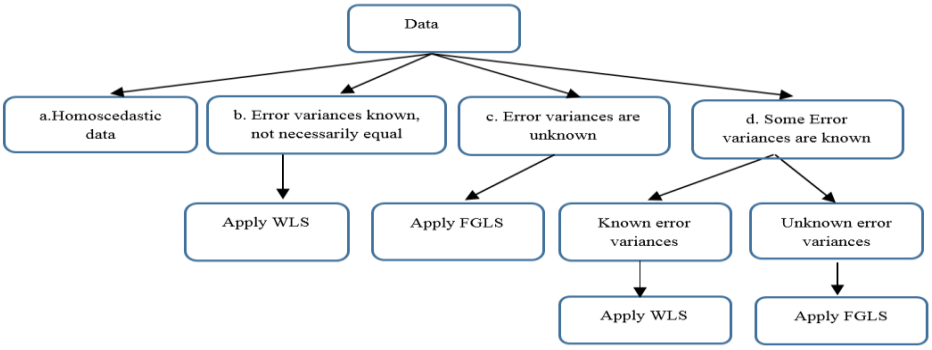

Figure 1. The diagram representing the flow of the process

2.1 Simulation

For each scenario listed above, a number of iterations are performed and the average weight over these iterations (where applicable) are selected and used as the weight to improve heteroscedasticity.

3. Numerical Example: Application to Shrimp Data in the Gulf of Mexico

For the demonstration of the above, the 2012 through 2016 shrimp data in the GOM were selected and analyzed. The major data contributors to this research were the following files: Shrimp data files in the GOM (2012-2016) and two additional files to be named later. The shrimp data files included several fields of interest to this study. Table 1 gives the fields used in this research and the corresponding descriptions.

Table 1. Description of fields in the shrimp data file used in this research

Field name	Description
port	The shrimp port of delivery
vessel id	US Coast Guard vessel identification number
yearU, monthU, dayU	Date of unloading shrimp at a designated port. The concatenation of these three was generated and call *edate*
daysfished	Actual hours of fishing per trip (24 hours per day)
pounds	Pounds of shrimp harvested
priceppnd	Average real price per pound of shrimp in the year data was collected
shore	1=offshore, 2=inshore

The U.S. Gulf of Mexico is divided into 21 statistical subareas (Figure 2). Statistical subareas 1–9 represent areas off the west coast of Florida, 10–12 represent Alabama/Mississippi, 13–17 denote Louisiana, and 18–21 represent Texas. Each statistical subarea is further divided into five-fathom depth increments (Table 2). This table also includes fathomzone and the corresponding depth zone. The 21 statistical subareas are placed into four areas 1 through 4, and twelve-fathomzones are placed into three depths 1 through 3. Figures 2 and Table 3display the 21 statistical subareas (1 through 21) and the conversion of subarea to the categorical variable area and fathomzone to the categorical variable depth respectively.

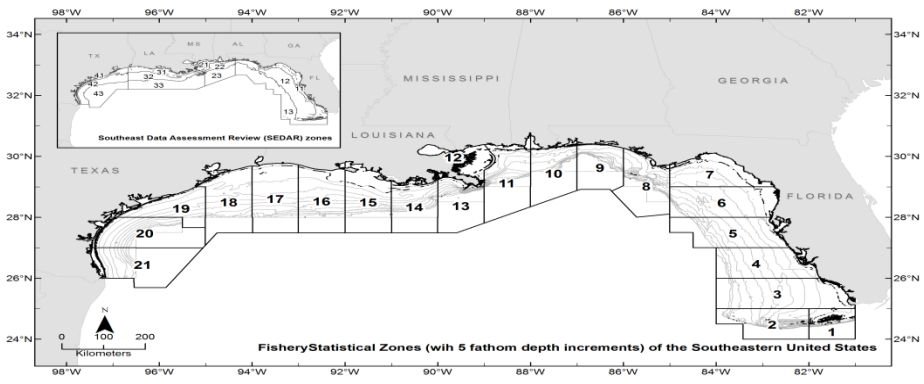

Figure 2. The Gulf of Mexico is divided into twenty-one statistical subareas (1-21) as shown

Table 2. Fathomzones (1-12), fathom, and corresponding depth zones (1-3) in the Gulf of Mexico

Fathomzone	Fathom	Depth zone(*depth*)
1	00-05	1
2	06-10	1
3	11-15	2
4	16-20	2
5	21-25	2
6	26-30	2
7	31-35	3
8	36-40	3
9	41-45	3
10	46-50	3
11	51-55	3
12	>55	3

Table 3. Conversion of statistical subareas (1-21) and fathomzones (1-12) in the Gulf of Mexico to *areas* (1-4) and *depths* (1-3) respectively

Statistical subarea	*area*	Fathomzone	*depth*
1 through 9	1	1 through 2	1
10 " 12	2	3 " 6	2
13 " 17	3	7 " 12	3
18 " 21	4		

The additional files used in this research included the AllocZoneLands (2012-2016) and another file, hereafter called the Vessel files. The first file called AllocZoneLands consisted of the electronic logbook box number (ELB), *edate*, a combination of statistical subarea and fathomzone (*zone*), actual days fished (*towdays*), shrimp landings (*landings*), and *port*. The data points in these files were interviewed and recorded by the port agents at the designated ports. The second file consisted of the vessel id number (*vessel*), vessel size (*length*) from the US Coast Guard file, and the four digit number assigned to each ELB unit.

For each year, the offshore data (that is, *shore*= 1) in the shrimp data files were converted to "Trips" based on vessel id number (*vessel*), edate (*edate*), and port (*port*) along with the weighted average price per pound per trip (*wavgppnd*) and total *pounds* per trip (*totlbs*). In the next step, the three files Trips, AllocZoneLands, and Vessel were matched based on the common fields listed in Table 4 grouped by the *zone* field from the AllocZoneLands file to create the "Match" file. In this research, the calendar year was also placed into three trimesters (January-April, May-August, and September-December).The reader is referred to Marzjarani (2018[a, c]) for additional information and detailed description on the preparation of these data files for analysis. The important issue of handling missing data points via multiple imputation is also presented in those references.

Table 4. Common fields used in creating the Match file

Files	Common field (s)
Shrimp(Trips), Vessel	*vessel*
Shrimp(Trips), AllocZoneLands	*port*
AllocZoneLands, Vessel	box (ELB)

For the purpose of further analysis, for each of 2012 through 2016 shrimp data, the statistical models were fitted to the Match files created above. The predictors in models (2) and (4) were as follows: vessel size (length), the natural logarithm of totlbs (lntotlbs or lnlbs for short), the weighted average price per pound per trip (wavgppnd), the variables area (4 levels), depth (3 levels), trimester (3 levels), and pairwise interactions of length, lnlbs, and wavgppnd. The response variable was the natural logarithm of the towdays from the AllocZoneLands file (lntd).Table 5 displays these variables, their types, and their roles in the model.

Table 5. Response and covariates used in the multiple linear regression model

Variable	Role	Continuous/discrete	Name used in the model
ln(*towdays*) or *lntowdays*	Response	Continuous	*lntd*
length (size)	Predictor	Continuous/categorical	*length*
ln(*totlbs*)	"	Continuous	*lnlbs*
weighted average price per pound/trip	"	"	*wavgppnd*
area	"	Discrete (4 levels)	*area*
depth	"	" (3 levels)	*depth*
trimester	"	" (3 levels)	*trimester*
interaction between length and ln(*towdays*)	"	Continuous/categorical	*lenlnlbs*
interaction between length and weighted average price per pound/trip	"	Continuous/categorical	*lenwavgppnd*
interaction between ln(*totlbs*) and weighted average price per pound/trip	"	Continuous	*lnlbswavgppnd*

3.1 Analysis/Results

As mentioned above, the 2012 through 2016 shrimp data in the GOM were selected and the model described in (2) or (4) with the variables listed in Table 5 were applied to these data sets. Through visual inspection, the plots of the residuals for the 2014 and 2015 data showed that roughly about 2% of such data points were close to 0 and therefore, the cutoff points ±0.01 were used to divide each data set into two, one assumed to have known error variances (equal or unequal) and the other unknown error variances.

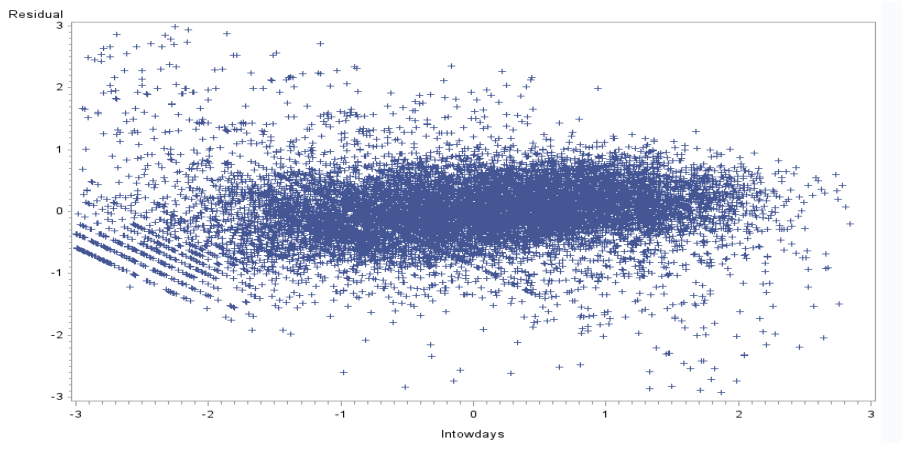

Figure 3. The plot of residuals vs predicted natural log of towdays under the assumption of homoscedasticity (Year 2014)

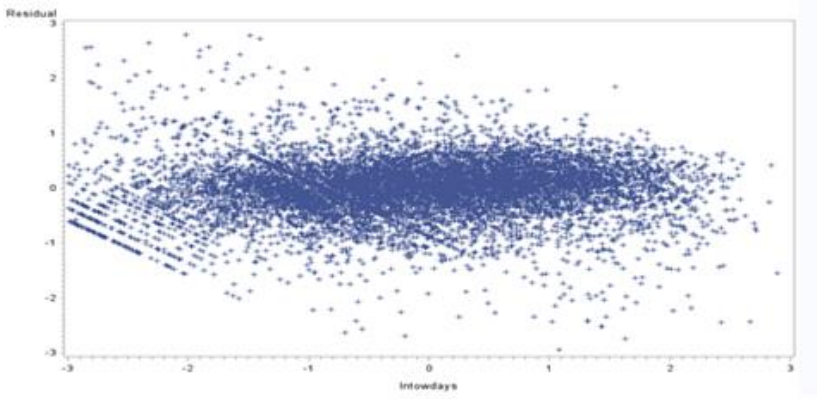

Figure 4. The plot of residuals vs predicted natural log of towdays under the assumption of homoscedasticity (Year 2015)

In the next step, the B-P and White tests were applied to several subintervals of the residuals. Tables 6 and 7 display the selection of the subintervals, the sample size and the corresponding results of the B-P and White tests. In both data files, the interval (-0.01, 0.01) generated non-significant B-P results and therefore, this interval was used to select the samples satisfying known and unknown error variances. This selection was consistent with the visual inspection and selection of ± 0.01as the cutoff points for splitting each data set into known and unknown error variances stated earlier.

It was noticed that the White test showed heteroscedasticity in all cases. This could be due the generality nature of this test as it relaxes the normality requirement and includes non-linear heteroscedasticity.

Table 6. Results of applying the B-P and White tests to the subintervals of the residuals in the 2014 data in the GOM

Residual interval	Sample size	White	df	*p-value*	B-P	df	*p-value*
(-0.01, 0.01)	229	69.57	45	0.0026	11.04	7	0.1369
(-0.05, 0.05)	1,157	103.0	45	<0.0001	18.32	7	0.0106
(-0.10, 0.10)	2,385	100.3	45	"	12.84	7	0.0762
(-0.28, 0.28)	6,259	115.7	45	"	20.43	7	0.0047
No limit	13,566	2,929	45	"	973.6	7	<0.0001

Table 7. Results of applying the B-P and White tests to the subintervals of the residuals in the 2015 shrimp data 2015 in the GOM

Residual interval	Sample size	White	df	*p-value*	B-P	df	*p-value*
(-0.01, 0.01)	206	74.71	45	0.0014	8.51	7	0.2894
(-0.05, 0.05)	1,117	87.02	45	0.0002	14.38	7	0.0448
(-0.10, 0.10)	2,209	86.66	45	0.0002	15.42	7	0.0310
(-0.28, 0.28)	5,578	176.3	45	<0.0001	34.02	7	<0.0001
No limit	11,092	1,186	45	<0.0001	279.9	7	<0.0001

In the following step, upon the selections of the sample size used in dividing each data set, IRLS and IFGLS were performed. Using the average over 200 iterations for the scenarios *b* through *d* mentioned earlier, it was observed that the ratio defined by (8) converged to 1 in less than 30 iterations as the number of iterations increased. Exception to this was the case where it was assumed that all variances were unknown and therefore a (an) FGLS was deployed. Although, the ratio did not converge to 1 as in the other cases did, it was satisfactorily close enough to assume that the average of the weights over 200 iterations was a good candidate to use and reduce the severity of heteroscedasticity in this case.

In the next step, analysis was performed on each 2014 and 2015 data set assuming that each consists of partially known error variances. Figures 5 and 7 show that the selected samples could have been assumed homoscedastic as supported by the B-P test (or heteroscedastic with known error variances which would have been converted to homoscedastic). Figures (3, 6) and (4, 8)are of interest to examine pairwise carefully. Due to a very large sample size in each data set, 3 and 6 look similar at a glance, so do 4, and 8. However, they are somewhat different. Figures 6and 8are slightly narrower and darker, meaning that the residuals are closer. That is, more points are overlapped and the magnitude of heteroscedasticity is slightly lower in these cases.

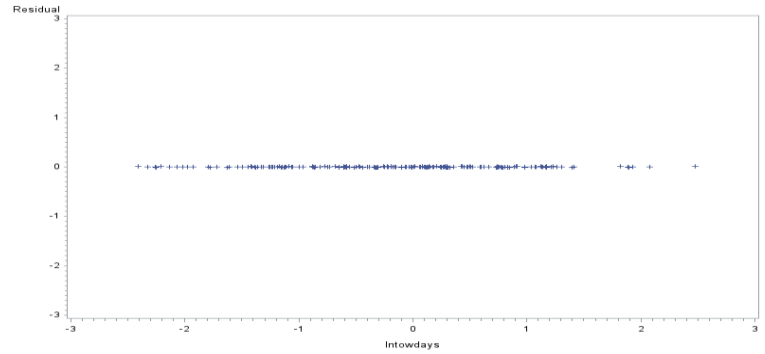

Figure 5. The plot of residuals vs predicted lntowdays for the sample (-0.01, 0.01) (Year 2014)

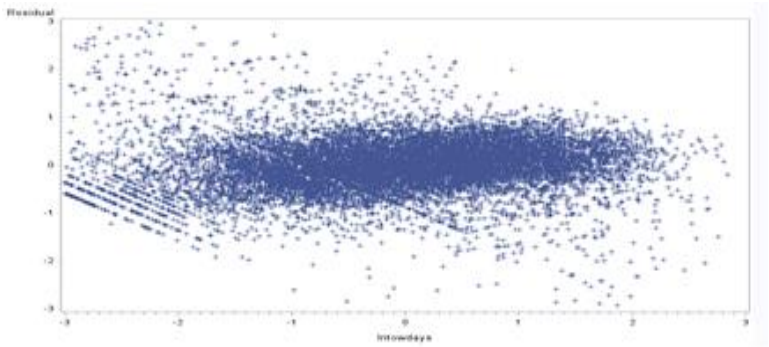

Figure 6. The plot of residuals vs predicted *lntowdays* after applying the average weight (Year 2014)

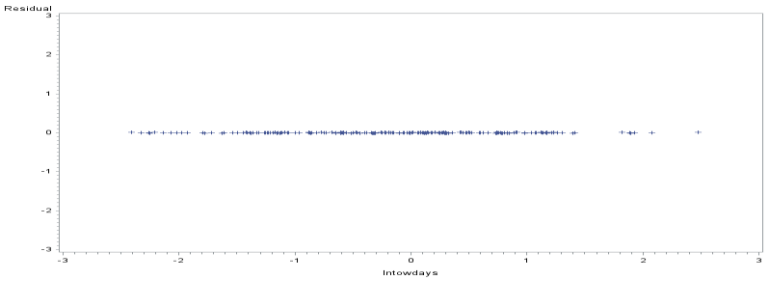

Figure 7. The plot of residuals vs predicted *lntowdays* for the sample (-0.01, 0.01) (Year 2015)

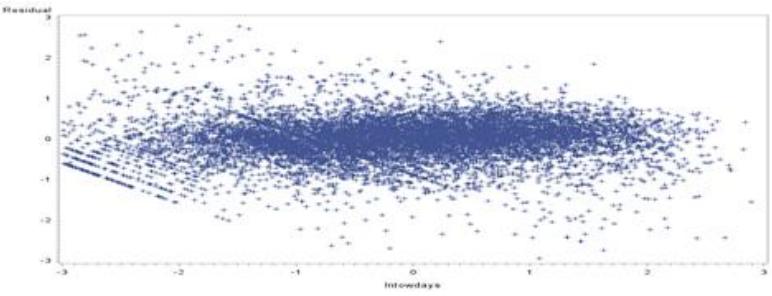

Figure 8. The plot of residuals vs predicted *lntowdays* after applying the average weight (Year 2015)

While pictures provide useful information, they could also be deceiving. Following are a few mathematically oriented arguments, which support the above visual inspections. Authors in Belsley, et al. (1980) proposed a statistic to measure the influence of an observation on the predicted value. It was suggested that any point with a DFFITS (Difference in Fits) greater than $2\sqrt{p/n}$, where p is the number of parameters including the intercept and n is the sample size, should be flagged as influential point and further be investigated. The DFFITS statistic is a scaled measure of the change in the predicted value for a data point and is calculated by deleting the said data point. Data points with large DFFITS values are very influential in their neighborhoods and should be inspected. For the 2014 data, the number of such points before and after the implementation of the weights selected via iterations were 349 and 306. The corresponding numbers for

the 2015 data were 222 and 178 respectively.

Furthermore, using the distance formula proposed by Cook (1977)with the cutoff point *4/n*, the number of points for further investigation in the 2014 data before and after the implementation of the weights were 659 and 538. The corresponding numbers for the 2015 data were 562 and 461 respectively. As an alternative, the Mahalanobis Distance formula (Mahalanobis, 1936) was applied to the 2014 and 2015 shrimp data. It was noticed that in the 2014 data, out of 13,566 records, 9,713 resulted in shorter distance from the centroid with 9 undefined distance values (diagonal elements in the hat matrix were 0). The corresponding numbers for the 2015 data were 11,092, 9,577, and 0 respectively. Another criterion used here was the studentized residuals (*Cook and Weisberg, 1982)*. This statistic was applied to the 2014 data and the number of points greater than 2 before and after the implementation of weights were 635 and 582. For the 2015 data, the corresponding numbers were 527 and 491 respectively. That is, in either data set the number of data points needing further inspection was reduced when the said data was split into two parts.

Following the diagram given in Figure 1 and upon the selection of the weights via iterations,

Table 8 displays the results of the B-P and White tests applied to the 2014 and 2015 shrimp data

in the GOM under different scenarios. First, it was assumed that these data sets were homoscedastic (no weight used). Second, it was assumed that the error variances in these data sets were all known, but not necessarily equal. As shown in Marzjarani (2018[b]), a WLS could be used to improve the heteroscedasticity. Third, it was assumed that the error variances were unequal and FGLS was deployed. The last possibility was to use the new approach. That is, dividing each data set into two parts: The first portion with small and/or known (but not necessarily equal) error variances and therefore, WLS method was deployed, and the second portion with unknown error variances, which required the deployment of the FGLS method. As displayed in this table, the B-P and White tests reduced the test statistic (increased the *p-value*) when the data were split into two parts as described above. That is, if the goal were to claim that there was no sufficient evidence for the existence of heteroscedasticity, splitting the data set would have moved in that direction.

Table 8. Test statistics for the B-P and White tests for the four possibilities described above (White's df=45, B-P's df=7)

Year	White				B-P			
	Without weight	All known error variances	All unknown error variances	Partially known error variances	Without weight	All known error variances	All unknown error variances	Partially known error variances
2014	2,929	1,387	1,322	1,099	973.6	102.9	238.2	72.52
2015	1,186	768.1	760	678.8	270.9	49.54	98.98	70.51

In the above, the B-P test statistic or visual inspection was used to argue that the partitioning method described would help with reducing the severity of heteroscedasticity. However, the visual approach is subjective and may not be sufficiently accurate. A more reliable measure is the RMSE as a way to show that the approach taken in this paper is effective in reducing the heteroscedasticity level. In the next step, in addition to the 2014 and 2015 shrimp data, out of several choices, the 2012, 2013, and 2016 shrimp data in the GOM were added and in each case, a simulation was used to select the cutoff points for splitting the data into two portions beginning with the interval (-0.01, 0.01) and incrementing it by 0.01 each time. Again, the process of preparing these two data files for analysis is similar to those presented in Marzjarani (2018a, c).

In the cases of 2013 and 2016, the vessel size (length) previously implemented as a continuous predictor, was defined as a categorical variable with three levels. The three levels of this predictor were less that 33thpercentile rank representing small vessels, between 33th and 67thpercentile rank representing mid-size vessels, and greater than 67th percentile rank for large vessels. Due to the inclusion of this and the other categorical variables (area, depth, and trimester), a GLM represented by (1), (2), or (4) was deployed, though throughout the process, an MLR could have also been used along with an appropriate coding pattern.

Two random samples of sizes 5,000 and 10,000 were selected from each of 2012 through 2016shrimpdata using simple random sampling (SRS), 100 simulations were performed on these samples, and the whole data set. The results were then checked for the first occurrences of the non-significant B-P test statistic using a backward approach. In other words, the largest samples with non-significant B-P tests at the threshold of 0.05 were selected as the portion with known error variances. In the following step, in order to determine the weights for each of five data sets, 200 simulations were performed on each data set beginning with the interval (-0.01, 0.01) and incrementing it each time by 0.01. Again, the largest interval with non-significant B-P test was selected as the cutoff point and each data set was split accordingly.

Next, 2000 simulations were performed on each sub-sample generated above, the weights generated at the i^{th} simulation were used for the $(i+1)^{th}$ simulation. Each simulation reached a steady state and the last weights were selected and used to lower the severity of heteroscedasticity. The results are summarized in Tables 9a and 9b. The first column in this table displays the year and the methods deployed throughout the article. First, it was assumed that the data were heteroscedastic and the IFGLS was deployed. Second, a combination of IRLS and IFGLS was used under the assumption of both portions were heteroscedastic, but the error variances in the first part were known (but not necessarily equal). Last, for the first portion, it was assumed that the known error variances were equal (i.e., homoscedastic). In these tables, the last two columns display the RMSE and the corresponding percentage decrease values. From the reduction in RMSE values and also the similarities in the corresponding percentage listed in these columns, it could be concluded that the assumption of homoscedastic or heteroscedastic with known error variances for the first part when dividing a data set into two is justified. In these tables, the Adj. R-Sq ranged from 0.80 to 0.99.

Table 9a. Results of analyzing the 2012, 2014, and 2015 shrimp data in the GOM with *length* as a continuous covariate

Year/Method	Sample size	Selected interval	RMSE*	Percentage decrease in RMSE
2012/				
IFGLS	5,000	No partitions	2.39	
IRLS, IFGLS	"	(-0.24, 0.24)	0.99	59
Assuming equal error variances in the first portion	"	"	0.96	60
IFGLS	10,000	No partitions	2.22	
IRLS, IFGLS	"	(-0.09, -0.09)	1.41	36
Assuming equal error variances in the first portion	"	"	1.40	37
IFGLS	All data	No partitions	2.60	
IRLS, IFGLS	"	(-0.23, 0.23)	1.04	60
Assuming equal error variances in the first portion	"	"	1.02	61
2014/				
IFGLS	5,000	No partitions	2.29	
IRLS, IFGLS	"	(0.25, -0.25)	1.05	54
Assuming equal error variances in the first portion	"	"	1.02	55
IFGLS	10,000	No partitions	2.32	
IRLS, IFGLS	"	(-0.11, 0.11)	1.49	36
Assuming equal error variances in the first portion	"	"	1.49	36
IFGLS	All data	No partitions	2.33	
IRLS, IFGLS	"	(-0.26, 0.26)	1.11	52
Assuming equal error variances in the first portion	"	"	1.08	54
2015/				
IFGLS	5,000	No partitions	2.10	
IRLS, IFGLS	"	(-0.25, 0.25)	1.06	50
Assuming equal error variances in the first portion	"	"	1.03	51
IFGLS	10,000	No partitions	2.12	
IRLS, IFGLS	"	(-0.19, 0.19)	1.18	44
Assuming equal error variances in the first portion	"	"	1.17	45
IFGLS	All data	No partitions	2.33	
IRLS, IFGLS	"	(-0.19, 0.19)	1.19	49
Assuming equal error variances in the first portion	"	"	1.18	49

*: Rounded to two decimal places

Table 9b. Results of analyzing the 2013, and 2016 shrimp data in the GOM with *length* as a categorical covariate with 3 levels

Year/Method	Sample size	Selected interval	RMSE*	Percentage decrease in RMSE
2013/				
IFGLS	5,000	No partitions	2.04	
IRLS, IFGLS	"	(-0.07, 0.07)	1.80	12
Assuming equal error variances in the first portion	"	"	1.76	14
IFGLS	10,000	No partitions	2.05	
IRLS, IFGLS	"	(-0.13, 0.13)	1.93	6
Assuming equal error variances in the first portion	"	"	1.86	9
IFGLS	All data	No partitions	2.07	
IRLS, IFGLS	"	(-0.37, 0.37)	1.12	46
Assuming equal error variances in the first portion	"	"	1.06	49
2016/				
IFGLS	5,000	No partitions	2.27	
IRLS, IFGLS	"	(-0.10, 0.10)	1.61	29
Assuming equal error variances in the first portion	"	"	1.60	29
IFGLS	10,000	No partitions	2.30	
IRLS, IFGLS	"	(-0.07, 0.07)	1.74	24
Assuming equal error variances in the first portion	"	"	1.74	24
IFGLS	All data	No partitions	2.31	
IRLS, IFGLS	"	(-0.12, 0.12)	1.55	33
Assuming equal error variances in the first portion	"	"	1.54	33

*: Rounded to two decimal places

4. Discussion

The main focus of this article was to develop a more effective method for handling heteroscedasticity in a given data set. Marzjarani (2018[a]) addressed heteroscedasticity in general linear models (GLM) and generalized linear mixed models (GLMM). Marzjarani (2010) developed an iterative method for estimating the parameters for a heteroscedastic linear regression model with two covariates. In that article, the covariance matrix of the error term was assumed to be in the form $\underline{\Omega} = \sigma^2 \underline{r}$, where $\underline{r} = (r_{kk})$, with $r_{kk} = (1/x_{1k}x_{2k})^{\delta}$, k=1,2,…, n. This approach is computationally extensive especially when dealing with a large number of covariates and for this reason it was not considered in this research. Phillips (2010) compared the finite sample properties of the iterated feasible generalized least square estimator to that of general mixed model estimators using both simulated and real data and claimed that the IFGLS estimator compares favorably. Previously, Hooper (1993) developed an iterative method for the WLS by which a more appropriate weight could be located. More recently, Marzjarani (2018[b]) implemented WLS, GLS, and FGLS in improving heteroscedasticity through partitioning a given data set into Training, Validation and Testing.

In this research, the works of Hooper (1993) and Marzjarani (2018[b]) were extended to include the IRLS and IFGLS. Furthermore, an attempt was made to develop a new method for handling heteroscedasticity by dividing a given data set into two parts: homoscedastic (or almost homoscedastic) and heteroscedastic. The partitioning was performed via the plot of the residuals and visual inspection or the application of the B-P test first and through simulations later. Through the applications of these approaches, it was shown that the method proposed in this article as shown in Table 8, generally increases the *p-value* (in other words, lowers the test statistic) and more importantly and more formally, it reduces the RMSE (Tables 9a and 9b). Furthermore, the assumption of known or equal error variances in the first portion of each data set was justified by the close RMSE values as displayed in Tables 9a and 9b.

5. Conclusion

Heteroscedasticity is a complex issue and is not something researchers are looking forward to seeing in their data sets. In fact, most researchers assume that their data sets are homoscedastic without checking (See Marzjarani, 2018[c], for example). For this reason, the goal might be to find a way to show that there is not sufficient evidence to claim the existence of it. Therefore, unlike the classical approach in setting up the null and the alternative hypotheses where one defines the alternative to be the favorite and attempts to accept it, it seems the null (homoscedastic data) becomes more

attractive when facing the issue of heteroscedasticity.

In this research, it was concluded that having a large data set, one might think of dividing the said data into two parts: The first part consists of the points with small residuals and the second portion with large residuals. Of course, such partitioning is only possible when the sample size is reasonably large. Factors determining a large sample include the number of covariates, the desired statistical power, the α-level, etc. In this research, it was shown that the partitioning a data set into homoscedastic (or heteroscedastic with known error variances) and heteroscedastic parts would reduce the MSE and the severity of the heteroscedasticity. Out of the methods presented in this article for splitting a data set, the simulation approach is a preferred since it is built on a more mathematical foundation.

Acknowledgement

The author would like to thank Mr. James Primrose of NMFS for providing data for this research, the Editorial Team of the journal, and a referee for their valuable and constructive suggestions.

Disclaimer

The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author and do not necessarily reflect those of NOAA or the Department of Commerce.

References

Belsley, D. A., Kuh, K., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity, John Wiley & Sons, New York. https://doi.org/10.1002/0471725153

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica, 47*(5), 1287–1294. JSTOR 1911963. MR 545960. https://doi.org/10.2307/1911963

Cook, D. R. (1977). Detection of Influential Observations in Linear Regression. *Technometrics, American Statistical Association, 19*(1), 15-18.

Cook, D. R., & Weisberg, S. (1982). Residuals and Influence in Regression (Repr. ed.). New York: Chapman and Hall. ISBN 041224280X.

Cook, D. R., & Weisberg, S. (1983).Diagnostics for heteroscedasticity in regression, *Biometrika*, *70*(1), 1–10. https://doi.org/10.1093/biomet/70.1.1

Hooper, P. M. (1993). Iterative Weighted Least Squares Estimation in Heteroscedastic Linear Models. *Journal of the American Statistical Association, 88*(421), 179-184.

Kalirajan, K. P. (1989).A test for heteroscedasticity and non-normality of regression residuals: A practical approach. *Economics Letters, 30*(2), August 1989, Pages 133-136. https://doi.org/10.1016/0165-1765(89)90050-5

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics, 17*(1), 107-112. https://doi.org/10.1016/0304-4076(81)90062-2

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*(1), 49–55.

Marzjarani, M. (2010). Logarithmic Transformation of Raw Data, International Journal of Science, *Technology and Management, 2*(3-4), 37-42.

Marzjarani, M. (2018[a]). Heteroscedastic and homoscedastic GLMM and GLM: application to effort estimation in the Gulf of Mexico shrimp fishery, 1984 through 2001. *International Journal of Probability and Statistics, 7*(1), 19-30.

Marzjarani, M. (2018[b]). Heteroscedasticity and model selection via partitioning in fisheries data, *International Journal of Statistics and Probability, 7*(6). https://doi.org/10.5539/ijsp.v7n6p33

Marzjarani, M. (2018[c]). Estimating Missing Values via Imputation: Application to effort estimation in the Gulf of Mexico Shrimp Fishery, 2007-2014. *International Journal of Statistics and Applications, 8*(2), 40-50.

Phillips, R. F. (2010). Iterated Feasible Generalized Least-Squares Estimation of Augmented Dynamic Panel Data Models. *Journal of Business & Economic Statistics, 28*(3). 2010 American Statistical Association. https://doi.org/10.1198/jbes.2009.08106

Spiegelman, D., Logan, R., & Grove, D. (2011). Regression Calibration with Heteroscedastic Error Variance, *The International Journal of Biostatistics, 7*(1), 1-34. https://doi.org/10.2202/1557-4679.1259

Su, L., Zhao, Y., Yan, T., & Li, F. (2012). Local Polynomial Estimation of Heteroscedasticity in a Multivariate Linear Regression Model and Its Applications in Economics. *PLoS ONE 7*(9), e43719. https://doi.org/10.1371/journal.pone.0043719

White, H. (1980). A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica, 48*(4), 817–838. JSTOR 1912934. MR 575027. https://doi.org/10.2307/1912934

Zambom, A. Z., & Kim, S. (2017). A nonparametric hypothesis test for heteroscedasticity in multiple regression. *The Canadian Journal of Statistics, 45,* 425-441. https://doi.org/10.1002/cjs.11333

Zhou, Q. M., Song, P. X. K., & Thompson, M. E. (2015). Profiling heteroscedasticity in linear regression models. *The Canadian Journal of Statistics, 43,* 358-337. https://doi.org/10.1002/cjs.11252