International Journal of Statistics and Probability; Vol. 7, No. 6; November 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Heteroscedasticity and Model Selection via Partitioning in Fisheries Data

Morteza Marzjarani

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

Received: July 17, 2018 Accepted: August 31, 2018 Online Published: September 12, 2018

doi:10.5539/ijsp.v7n6p33 URL: https://doi.org/10.5539/ijsp.v7n6p33

Abstract

Selecting a proper model for a data set is a challenging task. In this article, an attempt was made to answer and to find a suitable model for a given data set. A general linear model (GLM) was introduced along with three different methods for estimating the parameters of the model. The three estimation methods considered in this paper were ordinary least squares (OLS), generalized least squares (GLS), and feasible generalized least squares (FGLS). In the case of GLS, two different weights were selected for improving the severity of heteroscedasticity and the proper weight (s) was deployed. The third weight was selected through the application of FGLS. Analyses showed that only two of the three weights including the FGLS were effective in improving or reducing the severity of heteroscedasticity. In addition, each data set was divided into Training, Validation, and Testing producing a more reliable set of estimates for the parameters in the model. Partitioning data is a relatively new approach is statistics borrowed from the field of machine learning. Stepwise and forward selection methods along with a number of statistics including the average square error testing (ASE), Adj. R-Sq, AIC, AICC, and ASE validate along with proper hierarchies were deployed to select a more appropriate model(s) for a given data set. Furthermore, the response variable in both data files was transformed using the Box-Cox method to meet the assumption of normality. Analysis showed that the logarithmic transformation solved this issue in a satisfactory manner. Since the issues of heteroscedasticity, model selection, and partitioning of data have not been addressed in fisheries, for introduction and demonstration purposes only, the 2015 and 2016 shrimp data in the Gulf of Mexico (GOM) were selected and the above methods were applied to these data sets. At the conclusion, some variations of the GLM were identified as possible leading candidates for the above data sets.

Keywords: heteroscedasticity, model selection, partitioning

1. Introduction:

Finding a suitable model for a given data set is a challenging method. The issue becomes more complex as the number of potential covariates increases. Although, many research articles have addressed this issue, it is still an open-ended question and every bit of progress is worthy of consideration. We may never be able to find a "perfect" model, but always attempt to find the most reliable one for representing a given data set. Model selection has been a subject of research for many years. Zucchini (2000) presented an introduction to the topic for non-specialists with basic knowledge of statistical concepts. Cherkassky and Ma (2003) presented an empirical comparison between Akaike information criterion (AIC) and the Bayesian information criterion (BIC), and the structural risk minimization (SRM). Hastie et al. (2001) claimed that the SRM method performs poorly and suggested that AIC results in a superior performance. Lubke et al. (2017) performed a simulation study for selecting a model via a bootstrap approach. In addition, Vrieze (2012) addressed the difference between the statistics AIC and BIC focusing on latent variable models.

Adding to the problem of selecting a proper model for a given data set, there are some other important issues, which play significant roles in the process. Heteroscedasticity is one of these issues where the data analyst should investigate. It is an important issue in modeling where the existence of it is often ignored by researchers. Heteroscedasticity is a statistical term meaning that the variability of a response variable is unequal across the range of its predictor and it is quite common in fishery data sets. 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.

This topic has been addressed by many authors. Breusch and Pagan (1979) addressed this issue and developed a method known as the Lagrange Multiplier (LM) for testing the existence of heteroscedasticity 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. In addition, Marzjarani (2018[a]) applied these testing methods to the shrimp data

1984-2001 in the Gulf of Mexico (GOM). In this article, to check for heteroscedasticity the Breusch-Pagan test, hereafter, called the B-P test and the White's heteroscedasticity test, hereafter, called the White test were deployed. Furthermore, since many data do not follow the assumption of normality, Box and Cox (1964) developed a method for transforming data to normal. In this article, this transformation was applied to the response variable in the data files used in this research.

There are situations where models that are more accurate can be identified. Large data sets such as the 2015 and 2016 shrimp data in the GOM provide an interesting luxury to the researchers in all disciplines. Like in an organization where each unit is responsible for a particular activity, in the case of having sufficiently large data records, the data set could be divided into two or three parts, each part responsible for a particular action. This technique is used in some areas of machine learning where a portion of the data is used to train the system. An example of this is a decision tree where after its construction, a portion of the data is devoted to train the tree and make it ready to accept and classify an incoming observation. It is relatively new to the field of statistics and statistical software packages such as SAS[(1)] have added this feature to their products. One approach here was to divide such data set into two parts, Training and Testing. Alternatively, due to the availability of a large number of records in each file, each data set was divided into three parts: Training, Validation, and Testing. Training portion was used to estimate the parameters of the model. The Testing portion of the data was used to estimate the predictive performance of the model and was not used to estimate the parameters in the model. The Validation portion was used for the purposes such as terminating the selection process or selecting the final model.

2. Methodology

Heteroscedasticity, data partitioning, and model selection have not been addressed in depth in fisheries. For this reason and for demonstration purposes only, in this article the 2015 and 2016 shrimp data in the GOM were selected for analysis. The description of each file along with the process of preparing these files for analysis is similar to the presentations given in Marzjarani (2016) and Marzjarani (2018[b]). Following the preparation phase, in the next step, each data file was checked for the existence of heteroscedasticity using the B-P and White tests. Then, in order to account for the presence of heteroscedasticity, generalized least square (GLS), weighted least square (WLS), and feasible generalized least square (FGLS) were deployed. It must be noted that the ordinary least square (OLS) does not address heteroscedasticity, but it is extended to GLS and FGLS. The OLS gives equal weight to observations regardless of the fact that the observations with large residuals contain less information about the model. The reader is referred to Fomby et al. (1984), Musau et al. (2015), Wooldridge (2002), Poloni and Sbrana (2014), and others for details on these topics.

As mentioned earlier, each data set was divided into three parts: Training, Validation, and Testing. The model considered in this research was similar to the one used in Marzjarani (2018[b]) and it is listed below with notations borrowed from the same reference.

$$towdays = exp\{\beta_0 + \beta_1\, length + \beta_2\, ln\,(totlbs) + \beta_3\, wavgppnd + \beta_4\, length * ln\,(totlbs) + \beta_5\, length * wavgppnd + \beta_6\, ln\,(totlbs) * wavgppnd + \beta_7\, area + \beta_8\, depth + \beta_9\, trimester + \varepsilon\} \quad (1)$$

or in a more convenient form

$$\underline{y} = \underline{x\beta} + \underline{\varepsilon} \quad (2)$$

where, $\underline{y}$ is a column matrix of the natural logarithm of *towdays (*hereafter called *lntd),* $\underline{x}$ is a matrix of regressors relating the vector of responses $\underline{y}$. The vector $\underline{\varepsilon}$ is the error term with $E(\underline{\varepsilon}) = \underline{0}$ and $Var(\underline{\varepsilon}) = \underline{\Omega}$.

In this model, *length* is the vessel length (size), *ln* (*totlbs*), hereafter, called *lnlbs* is the natural logarithm of aggregated *pounds* of shrimp harvested, *wavgppnd* is the weighted average price per pound of shrimp per trip, *area* (a categorical variable with four levels), and *depth* and *trimester* are categorical variables with three levels. The response variable in (1) is *towdays* and in (2) is *lntd*.

Under the assumption of finite sample properties and $\underline{\Omega} = \sigma^2\underline{I}$, where $\underline{I}$ is an identity matrix, the ordinary least square (OLS) estimate of $\underline{\beta}$ is:

$$\underline{\hat{\beta}} = (\underline{x}'\underline{x})^{-1}(\underline{x}'\underline{y}) \quad (3)$$

It can easily be shown that this an unbiased estimator of $\underline{\beta}$. Under the assumption of $\underline{\Omega} = \sigma^2\underline{I}$, the model defined by (2) is called a "Homoscedastic" model. The normality assumption on the error term in (2) is not needed when performing OLS, but is necessary to conduct statistical tests such as t-tests and F-tests on model parameters. Deviation from this requirement may be relaxed when dealing with a large sample size (The central limit theorem, CLT).

Most data sets present some degree of heteroscedasticity, that is, the diagonal elements of the matrix $\underline{\Omega}$ are not identical. Two possibilities present themselves here. The first is where these elements are known. In such cases, we can reduce a heteroscedastic model to a homoscedastic one as follows. The matrix $\underline{\Omega}$ can be decomposed into $\underline{\Omega} = (\underline{\omega}'\underline{\omega})^{-1}$ (See Ben fez, and Liu, (2013), Dereniowski, and Kubale (2004)) and multiplying both sides of (2) on the left by $\underline{\omega}$ will result

in $\underline{\omega y} = \underline{\omega x\beta} + \underline{\omega\varepsilon}$. Since $E(\underline{\omega\varepsilon})=0$, then $E[(\underline{\omega\varepsilon})(\underline{\omega\varepsilon})']=\sigma^2\underline{\omega\Omega\omega}'$, which results in a homoscedastic model. By the application of OLS, it can easily be seen that the vector $\underline{\beta}$ can be estimated by the following:

$$\underline{\hat{\beta}} = (\underline{x}' \ \underline{\Omega}^{-1}\underline{x})^{-1} (\underline{x}' \ \underline{\Omega}^{-1}\underline{y}). \quad (4)$$

This is known as the generalized least square estimate of $\underline{\beta}$ denoted by $\underline{\hat{\beta}}_{GLS}$. It can also be shown that this is an unbiased estimator of $\underline{\beta}$. A special case of GLS is where the off-diagonal elements of $\underline{\Omega}$ are 0 (no correlations among the observed variances), but the diagonal element of this matrix are not identical (heteroscedasticity). The method used to estimate $\underline{\beta}$ in this situation is known as weighted least square (WLS). There are several ways to define the weight. For example, the weight for an element could be inversely proportional to the variance of the response for this element. In this article, to correct (or at the least to improve heteroscedasticity) two weights were defined as follows:

$$\underline{res} = |\underline{y} - \underline{x\hat{\beta}}|, \underline{res} = \underline{x\gamma} + \underline{\tau}, \ \underline{w_1} = 1/(\underline{x\hat{\gamma}})$$
$$\underline{res} = (\underline{y} - \underline{x\hat{\beta}})^2, \ \underline{res} = \underline{x\gamma} + \underline{\tau}, \ \underline{w_2} = 1/(\underline{x\hat{\gamma}}), \quad (5)$$

where $\underline{res}$ is the residual of the model defined in (2) and the remaining symbols are parameters used in the models.

Another possibility is that the variance-covariance matrix of the error term is unknown. For simplicity, here it was assumed that $\underline{\Omega}$ was a diagonal matrix with fully unknown elements. In such case, the GLS does not exist. The appropriate method to use under this condition is known as feasible generalized least squares (FGLS). There are different ways to implement FGLS. The third weight considered here was the FGLS defined as follows:

$$\underline{lnressq} = \ln(\underline{y} - \underline{x\hat{\beta}})^2$$
$$\underline{lnressq} = \underline{x\gamma} + \underline{\tau}$$
$$\underline{w_3} = 1/\exp(\underline{x\hat{\gamma}}) \quad (6)$$

In these formulas, $\underline{res}$ and $\underline{lnressq}$ are vectors of the residuals and the natural logarithm of the residuals squared respectively. For simplicity and without loss of generality (WLOG), hereafter, the vectors $\underline{w_1}$, $\underline{w_2}$, $\underline{w_3}$ are labeled as w_1, w_2, and w_3 respectively.

Following the steps for dealing with the heteroscedasticity issue, the next phase was the selection of covariates for the model given in (2) via Forward and Stepwise selection procedures. In addition, some statistics or options were included in each of the two selections resulting in 8 variations of model (2), hereafter, called models 1 through 8. Table 1 displays these models along with the optional hierarchies Single (S) or None (N). In the case of a Single hierarchy, only single effects are allowed to enter the model or to be removed from the model. With the option None, all effects are allowed to enter or leave the model. In order to implement these methods and options, the statistical software package SAS(1) was used throughout this research. SAS(1) also provides a third hierarchy option called Singleclass which was not included in this article. It is the same as hierarchy Single except that only CLASS effects are subject to the hierarchy requirement.

Table 1. Selections and options/statistics for models 1 through 8.

Model	Selection	Options	Hierarchy
1	no selection	Default	Single (S)or None (N)(4)
2	forward	select= SL[1] (Significance level)	Single (S)or None (N)
3	stepwise	stop=Adj. R-Sq	Single (S)or None (N)
4	stepwise	select=AIC, stop=C_p	Single (S)or None (N)
5	stepwise	select= SBC (default), stop=validate	Single (S)or None (N)
6	forward	select= SBC (default), stop= AIC, choose=validate	Single (S)or None (N)
7	stepwise	select= SL (Significance level), sle[2] =0.15 (Default), sls[3] =0.15 (Default), choose= C_p)	Single (S)or None (N)
8	orderselect	Specifies the order in which the parameters first entered the model	Single (S)or None (N)

1: Significance level for the test statistic F for entering or departing a variable. 2: Significance level for entry. 3: Significance level for stay.4: Hierarchy=None (N) is the default and it is equivalent to no hierarchy.

Note: Defaults are only relevant to the software package used.

Since the difference between AIC and AICC, that is, $2p*(p+1)/(n-p-1)$ where p is the number of parameters in the model and n is the sample size was negligible, the latter quantity was removed from further inclusion in the model.

For the 2015 and 2016 shrimp data, out of 8 model selection procedures for each hierarchy, a model was selected based on the minimum value of ASE test, the maximum value of Adj. R-Sq, and the minimum values of the remaining criteria AIC, SBC, and ASE validate in the order listed here.

The above approach produced a few variations of (2) for the 2015 and 2016 shrimp data sets. In the next step for each data

set, these models were compared and the lists of possible candidate models for the 2015 and 2016 data were generated. The differences among these models were in the number and type of covariates. In either 2015 or 2016 data (where applicable), the model with the minimum number of parameters was called the "Initial" and the remaining ones were called "Full" models. In what follows, it was assumed that each full model was just the result of adding more variables to the initial model. Upon adding more variables to the model, the sum of the square of the error in the initial model (SSE_I) will be reduced. At the same time, the model sum of the squares will be increased by, say, SS_a. The reduction in SSE_I or increase in SS of the full model is caused by, say, q variables added to the model. In order to set up the hypothesis for testing the impact of the additional covariates, let

$$SS_a = SSE_{Full} - SSE_{I,} \tag{7}$$

where SSE_{Full} is the error sum of square term in the model with additional q variable(s). The ratio of the mean square of the difference (SS_a/q) and that of the full model (MSE_{Full}) is a proper criterion to use as the test statistic for justifying the addition (s) to the model. The hypotheses were defined as follows:

H_0: The addition (s) of covariate(s) not justified.

vs

$$H_a: \textit{The addition (s) of covariate(s) justified} \tag{8}$$

The test statistics for testing (8) can be expressed as:

$$F_{stat} = (SS_a/q)/MSE_{Full} \tag{9}$$

The degrees of freedom for this test statistic are q and $dfMSE_{Full}$. Larger F values in (9) support the rejection of the null hypothesis.

The plots of the response variable (*towdays*) in both the 2015 and 2016 shrimp data showed that the distributions were positively skewed (skewed to the right). Some authors impose the normality assumption on the response variable. The natural logarithm of this variable showed that the normality of the response variable was satisfied. Although not needed here (log transform), the reader is referred to Marzjarani (2016) where the empirical rule was deployed for checking the normality assumption of the error term in (2). The Box-Cox transformation was also applied to the original and the transformed data. This transformation as appeared in that reference can be written as:

$$y_{transformed} = (y^{**\lambda} - 1)/\lambda)*(1/g^{**}(\lambda-1))I_{\{\lambda \neq 0\}} \oplus \log(y)*g\, I_{\{\lambda = 0\}} \tag{10}$$

where g is the geometric mean of the response variable y, I_a is the indicator function, ** is the exponentiation operator, * represents the multiplication, and $\oplus$ is the exclusive OR operator.

Upon the selection of the candidate models, a file was created with the following contents: The pair of year and weight as one field (such as year2015w_1), the model (1-8), the hierarchy code SN (S or N), and the corresponding effort figures as the response variable. A two-way ANOVA model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},\ i=1, 2, 3, \ldots, j=1, 2, 3, \ldots, \tag{11}$$

was fitted to this data file and some statistical analyses were performed.

3. Analysis/Results/Discussion

It was not the intention of this article to propose any method (s) for the shrimp effort estimation in the GOM. Since the topics included in this paper have not been addressed in fisheries in depth, the 2015 and 2016 shrimp data files in the GOM were selected and the issues covered in this paper were applied to these data for demonstration purposes only. The analysis was performed on these data using the three weights given earlier in (5) and (6). Table 2 displays the results of applying the B-P and White tests to the original (raw) data as well as the results of applying the three weights to the 2015 and 2016 shrimp data files. Out of the three weights, w_2 did not improve heteroscedasticity as well as the other two weights and therefore this weight was eliminated from further consideration. For the 2015 data, both w_1 and w_3 seemed appropriate and they were selected for additional analyses. As for the 2016 data, w3 scored lower in both the B-P and White tests and therefore it was selected for additional analysis.

Table 2. Selection of weights based on the B-P and white tests.

Year	Data	B-P test	df	White test	df
2015	Original (raw)	270.9	7	1,186	45
	w_1	49.54	7	768.1	45
	w_3	98.98	7	760.0	45
	w_2	952.5	7	2,518	45
2016	Original (raw)	313.4	7	2,547	45
	w_1	48.35	7	1,322	45
	w_3	42.98	7	952.8	45
	w_2	2,627	7	6,338	45

Note: All test statistics in this table were significant at *p-value*< 0.0001.Output was generated via SAS[(1)].

For later comparisons, the original 2015 and 2016 data files were analyzed under the assumption of homoscedasticity using models 1 through 8. The results are illustratein Table 3, Table 4, Figure 1, and Figure 2.

Table 3. The ASE and the number of significant parameters for the 2015 shrimp data file using models 1-8 with hierarchy Single/None under the assumption of homoscedasticity.

Year	Model	Hierarchy	No. of params	ASE test
2015	1	S	12	0.31048
	2	S	14	0.33980
	3	S	13	0.32156
	4	S	14	0.29221
	5	S	14	0.29441
	6	S	14	0.29017
	7	S	12	0.29755
	8	S	13	0.31046
	1	N	14	0.30818
	2	N	14	0.33518
	3	N	13	0.31149
	4	N	13	0.29471
	5	N	12	0.27664
	6	N	14	0.31905
	7	N	13	0.33187
	8	N	14	0.27198

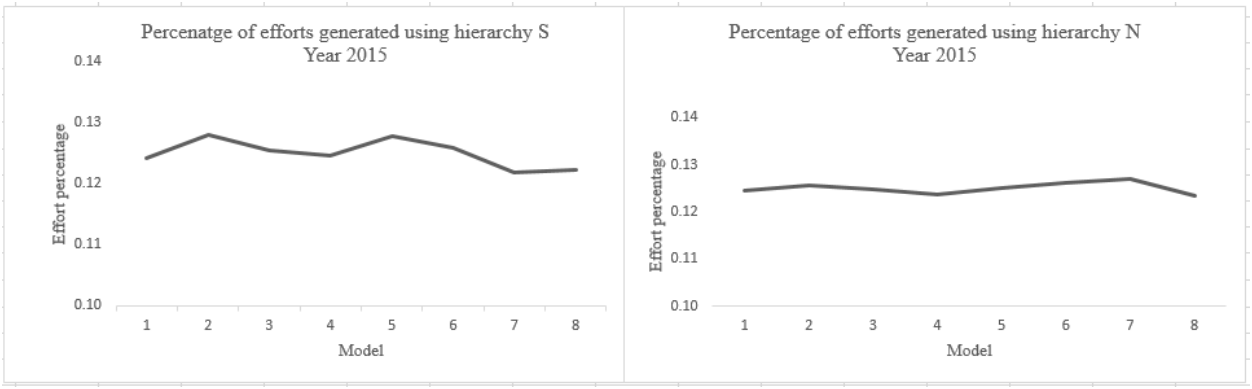


Figure 1. Percentages of effort estimates for the 2015 shrimp data file using models 1-8 with hierarchy Single/None under the assumption of homoscedasticity.

Table 4. The ASE and the number of significant parameters for the 2016 shrimp data file using models 1-8 with hierarchy Single/None under the assumption of homoscedasticity.

Year	Model	Hierarchy	No. of params	ASE test
2016	1	S	9	0.44518
	2	S	13	0.43246
	3	S	11	0.47855
	4	S	13	0.40562
	5	S	9	0.41595
	6	S	9	0.43681
	7	S	13	0.41324
	8	S	10	0.40763
	1	N	9	0.43240
	2	N	13	0.44889
	3	N	11	0.41705
	4	N	12	0.45503
	5	N	9	0.44068
	6	N	11	0.45311
	7	N	13	0.40512
	8	N	10	0.40613

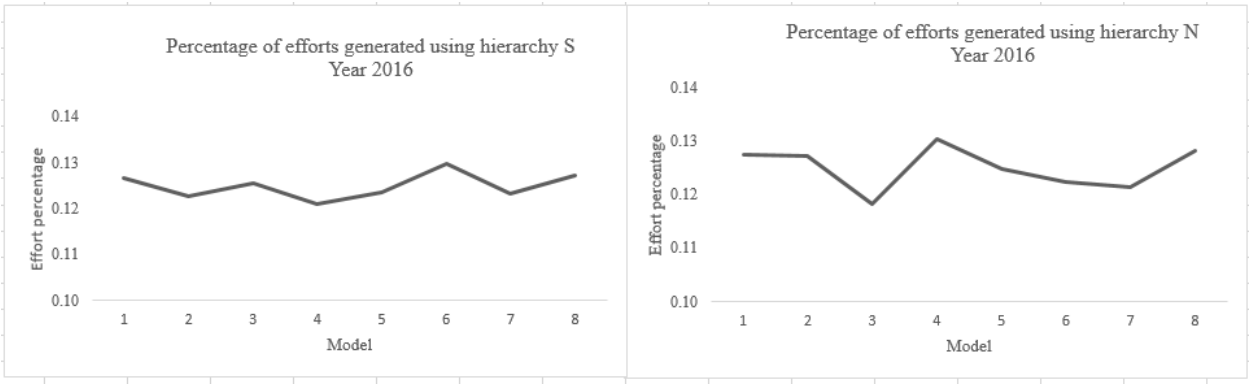


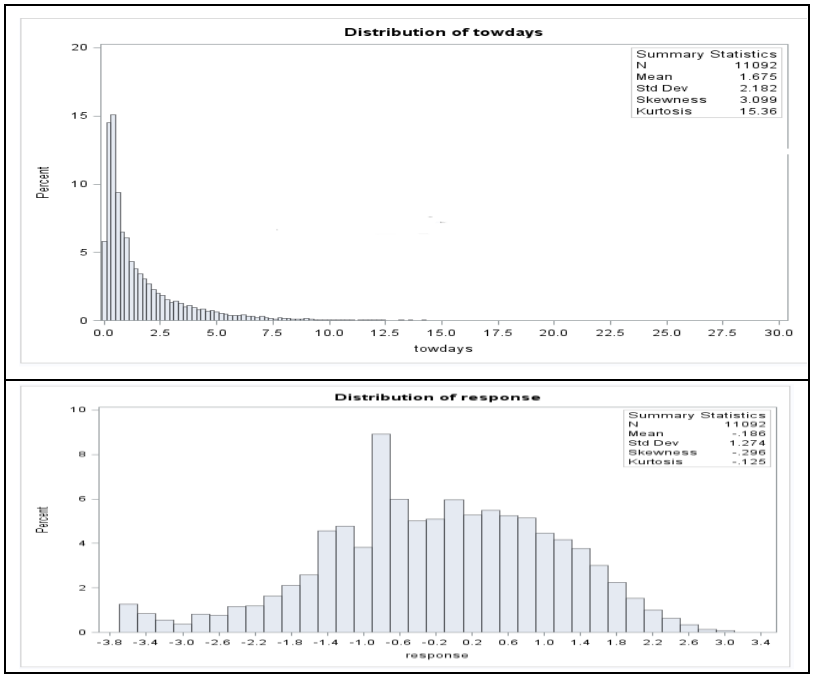
Figure 2. Percentages of effort estimates for the 2016 shrimp data file using models 1-8 with hierarchy Single/None under the assumption of homoscedasticity.

Table 5 and Figures 3a and 3b display the results of testing for normality of the 2015 and 2016 shrimp data in the GOM. In both cases, the original data were sharply right (positively) skewed. However, the logarithmic transformation solved this issue to a satisfactory point. The Box-Cox transformation did not improve the log-transformed data set towards the normality requirement. The minor deviation from normality as displayed by the skewness and kurtosis coefficients should not affect the estimation process significantly.

Table 5. Results of testing for normality of the 2015 and 2016 shrimp data before and after the application of Box-Cox transformation.

Year	Test[+]	Statistic		Decision rule	*p-value*	Skewness	kurtosis
2015							
Before transformation	Kolmogorov	D	0.22497	P_r> D	< 0.0100	3.098631	15.36343
(*towdays*)	Cramer-von Mises	W-Sq	180.5587	P_r> W-Sq	< 0.0050		
	Anderson-Darling	A-Sq	973.6835	P_r> A-Sq	< 0.0050		
2015							
Log transformation	Kolmogorov	D	0.027043	P_r> D	< 0.0100	-0.295999	-0.1254931
(*lntd*)	Cramer-von Mises	W-Sq	2.442959	P_r> W-Sq	< 0.0050		
	Anderson-Darling	A-Sq	21.32054	P_r> A-Sq	< 0.0050		
2015							
After log transformation	Kolmogorov	D	0.027043	P_r> D	< 0.0100	-0.295999	-0.1254931
(*lntd*)	Cramer-von Mises	W-Sq	2.442959	P_r> W-Sq	< 0.0050		
	Anderson-Darling	A-Sq	21.32054	P_r> A-Sq	< 0.0050		
2016							
Before transformation (*towdays*)	Kolmogorov	D	0.232403	P_r> D	< 0.0100	3.058958	13.27143
	Cramer-von Mises	W-Sq	207.1113	P_r> W-Sq	< 0.0050		
2016	Anderson-Darling	A-Sq	1111.095	P_r> A-Sq	< 0.0050		
Log transformation (*lntd*)							
	Kolmogorov	D	0.029233	P_r> D	< 0.0100	-0.261677	-0.157991
	Cramer-von Mises	W-Sq	2.134817	P_r> W-Sq	< 0.0050		
	Anderson-Darling	A-Sq	19.28038	P_r> A-Sq	< 0.0050		
2016							
After log transformation (*lntd*)							
	Kolmogorov	D	0.029233	P_r> D	< 0.0100	-0.261677	-0.157991
	Cramer-von Mises	W-Sq	2.134817	P_r> W-Sq	< 0.0050		
	Anderson-Darling	A-Sq	19.28038	P_r> A-Sq	< 0.0050		

+:Shapiro-Wilk test was not performed due to the large sample sizes (> 50).



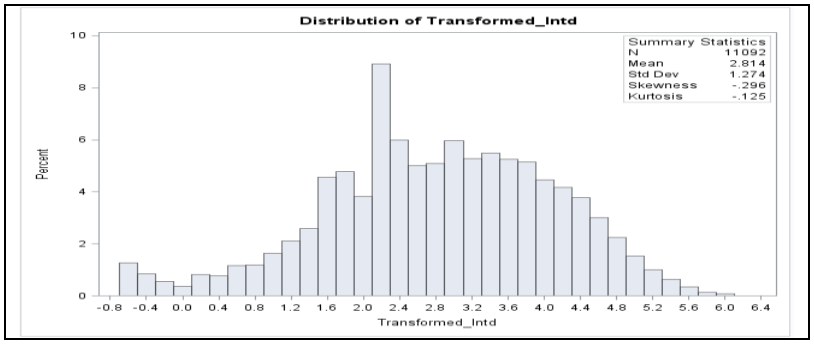


Figure 3a. Plots of the data points in shrimp data file 2015 (*towdays*, *lntd* and after the application of Box-Cox transformation to *lntd*.

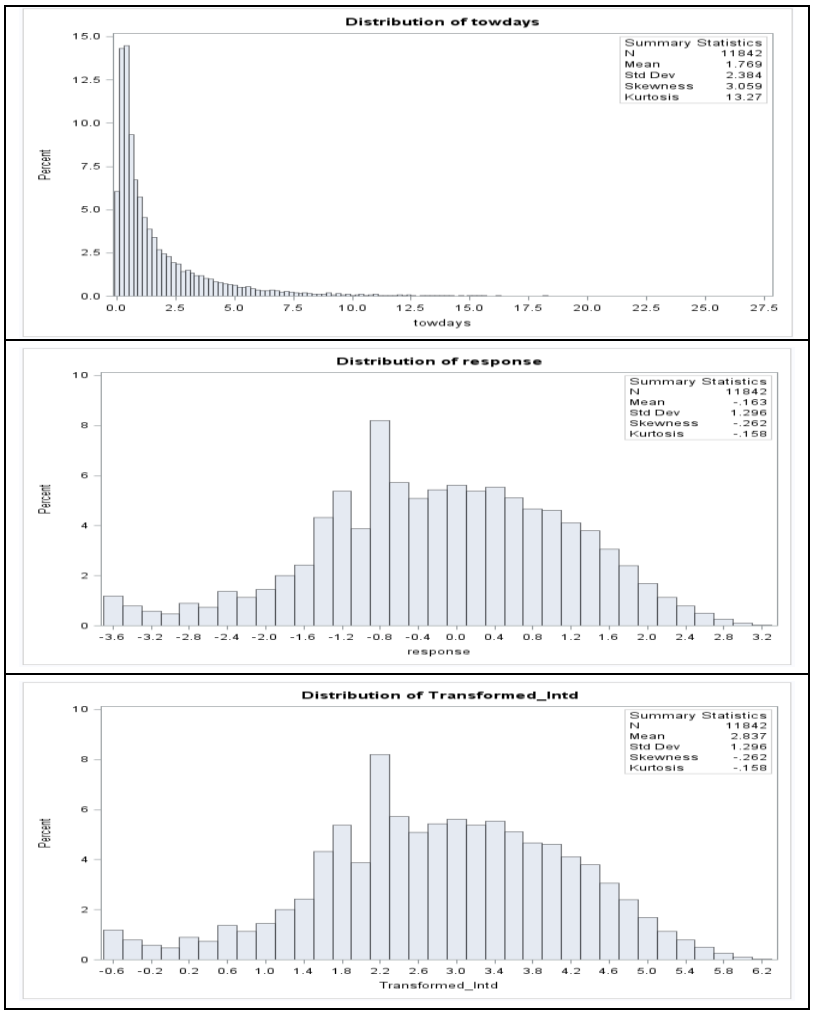


Figure 3b. Plots of the data points in shrimp data file 2016 (*towdays*, *lntd*) and after the application of Box-Cox transformation to *lntd.*

Figures 4 and 5display the results of applying models 1 through 8 to the 2015 shrimp data file using the weight w_1 and w_3 for the hierarchy Single/None respectively.

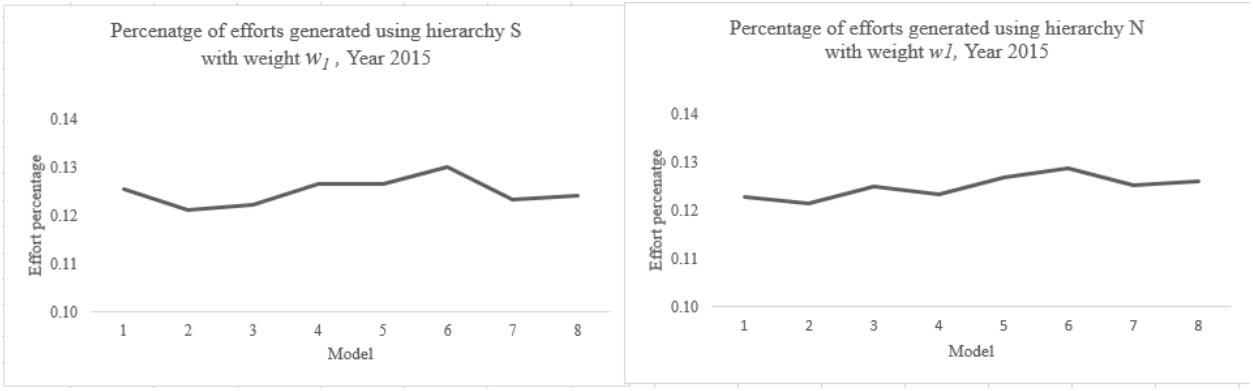


Figure 4. Percentages of effort estimates for the 2015 shrimp data file using models 1-8 with hierarchy Single/None with weight w_1.

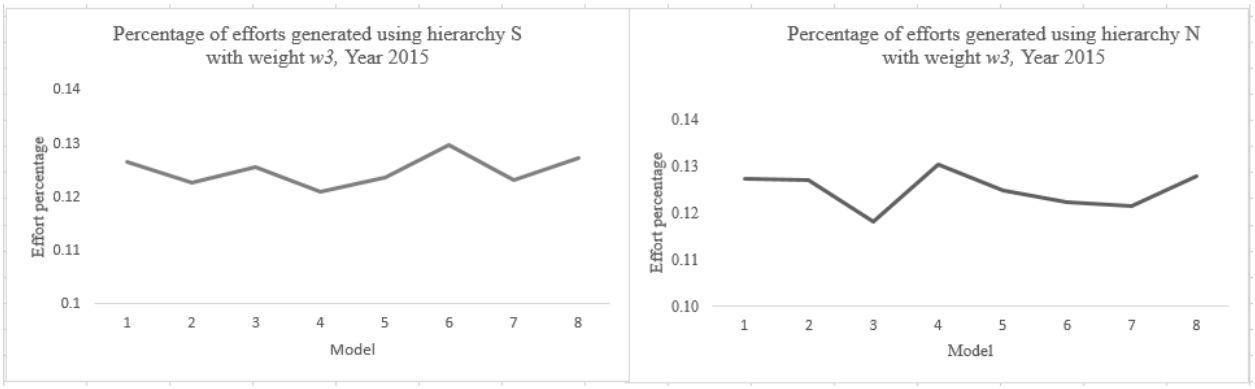


Figure 5. Percentages of effort estimates for the 2015 shrimp data file using models 1-8 with hierarchy Single/None with weight w_3.

As stated earlier, the 2016 shrimp data file was analyzed under the choice of w_3 as the weight. Figure 6 displays the results of applying models 1 through 8 to this data set for the hierarchy Single/None respectively.

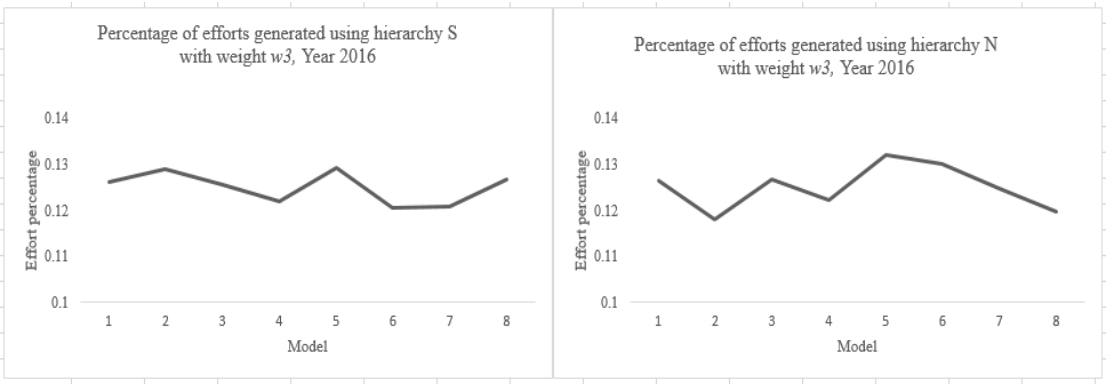


Figure 6. Percentages of effort estimates for the 2016 shrimp data file using models 1-8 with hierarchy Single/None using w_3.

Tables 6 through 8 display the values of different statistics used in selecting a model. The selection was performed based on the minimum value of ASE test, and in the case of ties, the maximum of Adj. R-Sq, followed by the minimum of the remaining criteria in the order of AIC, SBC, and ASE validate. The final selections for the 2015 and 2016 shrimp data are listed in Table 9.

Table 6. Some statistics of interest used to select the most appropriate model for the 2015 data using hierarchy Single/None with w_1 as the weight.

Year	Model	Hierarchy	No. of params	Adj. R-Sq	AIC	SBC	ASE test	ASE validate	C_p
2015	1	S	13	0.8274	4,138.45	-1,351.42	0.26125	0.27705	
	2	S	14	0.8294	3,935.79	-1,475.63	0.28889	0.26339	
	3	S	12	0.8203	4,262.01	-1,261.42	0.26047	0.27253	
	4	S	14	0.8191	4,170.08	-1,250.32	0.2415	0.2873	14
	5	S	14	0.8323	3,897.44	-1,491.04	0.2813	0.2743	
	6	S	14	0.8251	4,068.22	-1,397.06	0.27358	0.26724	
	7	S	14	0.8365	3,908.17	-1,646.89	0.29296	0.27312	14
	8	S	14	0.824	4,026.17	-1,321.42	0.25611	0.2834	
	1	N	12	0.822	4,153.48	-1,257.19	0.26858	0.26248	
	2	N	14	0.8244	4,064.98	-1,306.54	0.26265	0.27434	
	3	N	12	0.8279	3,952.38	-1,494.22	0.27825	0.27907	
	4	N	14	0.8246	4,142.41	-1,325.86	0.26757	0.26816	14
	5	N	14	0.8254	4,030.98	-1,374.46	0.26584	0.27785	
	6	N	14	0.8308	3,897.69	-1,551.63	0.28597	0.27385	
	7	N	14	0.8353	3,786.19	-1,691.07	0.27601	0.29513	14
	8	N	14	0.8309	3,950.61	-1,549.59	0.30261	0.25029	

Table 7. Some statistics of interest used to select the most appropriate model for the 2015data using hierarchy Single/None with w_3 as the weight.

Year	Model	Hierarchy	No. of params	Adj. R-Sq	AIC	SBC	ASE test	ASE validate	C_p
2015	1	S	14	0.8426	14,724.00	9,247.05	0.23937	0.28135	
	2	S	14	0.8365	14,907.00	9,423.93	0.25317	0.24661	
	3	S	14	0.8287	15,248.00	9,723.20	0.25769	0.22564	
	4	S	14	0.8325	14,949.00	9,442.58	0.24602	0.24934	14
	5	S	13	0.8371	14,879.00	9,391.66	0.26366	0.23997	
	6	S	14	0.8365	14,958.00	9,479.28	0.24751	0.24672	
	7	S	14	0.8359	14,789.00	9,343.72	0.26706	0.23639	14
	8	S	13	0.8309	14,892.00	9,481.02	0.26209	0.22444	
	1	N	14	0.8353	14,675.00	9,272.90	0.26368	0.23773	
	2	N	14	0.8417	14,538.00	9,114.42	0.25475	0.26534	
	3	N	14	0.832	14,870.00	9,478.26	0.24535	0.23551	
	4	N	14	0.8337	14,888.00	9,459.38	0.25342	0.23473	14
	5	N	12	0.8404	14,699.00	9,221.14	0.24542	0.27625	
	6	N	14	0.8376	15,138.00	9,576.73	0.24306	0.25538	
	7	N	14	0.8265	15,359.00	9,835.61	0.25298	0.21776	14
	8	N	14	0.8452	14,576.00	9,131.95	0.25596	0.26423	

Table 8. Some statistics of interest used to select the most appropriate model for the 2016 data using hierarchy Single/None with w_3 as the weight.

Year	Model	Hierarchy	No. of params	Adj. R-Sq	AIC	SBC	ASE test	ASE validate	C_p
2016	1	S	12	0.7653	15,915.00	10,094.00	0.41003	0.36476	
	2	S	11	0.7586	16,321.00	10,353.00	0.39872	0.38698	
	3	S	12	0.7659	16,034.00	10,205.00	0.37767	0.3896	
	4	S	14	0.7626	16,088.00	10,230.00	0.39097	0.37345	14
	5	S	8	0.7577	15,923.00	10,126.00	0.38007	0.40772	
	6	S	14	0.7624	16,158.00	10,267.00	0.38743	0.38521	
	7	S	14	0.7601	15,884.00	10,077.00	0.3725	0.40711	14
	8	S	10	0.7715	15,918.00	9,953.79	0.61955	0.60548	
	1	N	10	0.7515	16,223.00	10,372.00	0.3894	0.36498	
	2	N	14	0.7539	16,383.00	10,514.00	0.38331	0.34892	
	3	N	9	0.7632	15,931.00	10,094.00	0.37728	1.61618	
	4	N	12	0.768	15,754.00	9,935.69	0.40437	0.38685	11.77
	5	N	8	0.7624	16,014.00	10,147.00	0.39077	0.39472	
	6	N	8	0.7561	16,059.00	10,167.00	0.39143	0.3977	
	7	N	10	0.7623	15,912.00	10,074.00	0.91786	1.42986	8.45
	8	N	12	0.7557	16,233.00	10,350.00	0.36464	0.38006	

Table 9. Summary of the selected models and the corresponding information for the years 2015-2016.

Year		Hierarchy	Selected model	No. of params (incl. intercept)	C_p	Significant covariates (reference levels not included here)
2015	w_1	S	4	14	14	*Intercept, length, lnlbs, wavgppnd, area$_1$, area$_2$$^{(1)}$, area$_3$, depth$_1$, depth$_2$, trimester$_1$, trimester$_2$, length*lnlbs, length*wavgppnd, lnlbs*wavgppnd*
		N	2	14		*Intercept, length, lnlbs, wavgppnd, area$_1$, area$_2$, area$_3$, depth$_1$, depth$_2$, trimester$_1$, trimester$_2$, length*lnlbs, length*wavgppnd, lnlbs*wavgppnd*
	w_3	S	1	14		*Intercept, length, lnlbs, wavgppnd, area$_1$, area$_2$, area$_3$, depth$_1$, depth$_2$, trimester$_1$, trimester$_2$, length*lnlbs, length*wavgppnd, lnlbs*wavgppnd*
		N	6	14		*Intercept, length, lnlbs, wavgppnd, area$_1$, area$_2$$^{(1)}$, area$_3$, depth$_1$, depth$_2$, trimester$_1$, trimester$_2$, length*lnlbs, length*wavgppnd, lnlbs*wavgppnd*
2016	w_3	S	7	14		*Intercept, length, lnlbs, wavgppnd, area$_1$, area$_2$$^{(1)}$, area$_3$, depth$_1$, depth$_2$, trimester$_1$, trimester$_2$, length*lnlbs, length*wavgppnd, lnlbs*wavgppnd*
		N	8	12		*Intercept, length, lnlbs, wavgppnd, area$_1$, area$_2$, area$_3$, trimester$_1$, trimester$_2$, length*lnlbs, length*wavgppnd, lnlbs*wavgppnd*

(1): Subscripts represent levels of the categorical variables.

The selected models were further analyzed by deploying a two-way ANOVA with no replications/interactions where models 1 through 8 forming the first factor and the hierarchy option Single/None (SN) as the second factor with efforts as the response variable. The hypotheses were to find out if there were differences in effort estimations among the models 1 through 8 or there was a difference in effort estimates between S and N. In addition, a couple of contrasts comparing the model mean efforts and the hierarchy mean efforts were included in the analysis as listed below. These were selected arbitrarily and for demonstration purposes only.

Hypothesis: The average effort generated by models 1, 2, 3, 4 is equal to the average of models 5, 6, 7, 8.

Hypothesis: The average effort generated by hierarchies S and N are the same.

Table 10 displays the results. The selected models 1-8 in 2015 with w_1 as the weight showed a significant result. In addition, the comparison between models 1, 3, 3, 4 with 5, 6, 7, and 8 was also significant. Further analysis showed that model 6 was placed in a separate group than the remaining models. Of course, column 6 in this table is equivalent to column 3 as it is the t-test version of an F-test. Again, this test was performed for demonstration purposes only.

Table 10. Results of deploying a GLM with covariates model (1 through 8) and the options Single/None (SN) with the response variable effort.

Year	Weight	SN (1 df)	Model (7 df)	Model 1, 2, 3, 4 vs 5, 6, 7, 8	S vs N
2015	w_1	F=0.35, *p-value*= 0.57	F=4.64, *p-value*= 0.03	t=-3.64, *p-value=0.01*	t=0.60, *p-value=0.57*
	w_3	F=3.66, *p-value*= 0.10	F=0.46, *p-value*= 0.83	t=-0.05, *p-value*= 0.96	t=3.66, *p-value=0.10*
2016	w_3	F=0.14, *p-value*= 0.72	F= 0.77, *p-value*= 0.63	t=-0.44, *p-value*= 0.67	t=0.37, *p-value=0.72*

In the next step, for the 2015 data the weights w_1 and w_3 captured all the effects in the models (including the intercept). Either of these could be selected as a candidate model for this data set. However, if it would make sense at all, models 4 and 1 had lower ASE tests for the hierarchies S and N respectively. The approach was applied to the 2016 data set with model 8 as the initial model. Results are listed in Table 11. As displayed in this table, model 7 showed a minor increase in the adj. R-Sq and therefore it was selected as the possible candidate for the 2016 data set.

Table 11. Selections of final models for representing shrimp data file, 2016.

Year	Model	F_{stat} (1)	Adj. R-Sq	Increase in Adj. R-Sq
2016	8 (Initial)	616.79	0.771694	
	7	522.71	0.771916	0.000223

(1): Both significant at *p-v*alue<0.0001

Figures 1, 2, 4, 5, and 6 display the change in effort estimates by model and hierarchy and Table 12 contains the corresponding CVs for each year. In the case of 2015 data, the variations among models with hierarchy None was lower than those under the hierarchy Single. The 2016 data showed the opposite results. However, there was not sufficient evidence that the difference was statistically significant (Table 10).

Table 12. CV values corresponding to Figures 1, 2, 4, 5, and 6.

Year	Weight	Hierarchy	CV(%)
2015	No weight (Homoscedastic)	S	1.83
		N	0.99
	w_1	S	2.28
		N	1.89
	w_3	S	1.59
		N	1.01
2016	No weight (Homoscedastic)	S	2.29
		N	3.21
	w_3	S	2.82
		N	3.95

In this article, the 2015 and 2016 shrimp data files in the GOM were analyzed by considering the possibility of

heteroscedasticity and some alternative estimation models. This study extended the heteroscedasticity issue presented in Marzjarani (2018[a]) where the scope of that reference was limited to the use of the WLS. The GLS (and its special case, WLS) and FGLS were included in this study. As displayed in Table 2, both the 2015 and 2016 data files contained heteroscedasticity. The weights w_1 and w_3 reduced the severity of heteroscedasticity to some degree, but not completely. Weights that are more appropriate might be considered.

In order to get a more reliable set of parameter estimates for the model considered in this study, each of the 2015 and 2016 shrimp data in the GOM was divided into three parts (Training, Validation, and Testing). The percentages used were 25, 50, and 25 respectively. It would be at the discretion of the researcher to modify these and use different percentages as desired.

Analysis showed that the hierarchy Single/None did not play a significant role in 2015 data. The 8 variations of the GLM showed a significant difference in the 2015 data under the weight w_1 Out of 8 variations considered in this article, the selected ones are listed in Table 9. The selection was performed by considering the minimum value of ASE test, the maximum value of the Adj. R-Sq and the minimum values of the other criteria used in this study. This selection is subject to the discretion of the researcher and understanding of the differences among these criteria.

In his 2012 article, Vrieze (2012) addressed the difference between the AIC and the BIC statistics in detail. As appeared in that article, BIC is asymptotically consistent in selecting a model if the said model is among the candidate models. In that respect, AIC is not efficient. If the model is not among the candidates, AIC is more efficient as it selects the model, which minimizes the MSE of the prediction or estimation. Kuha (2004) argues, "*It is argued that useful information for model selection can be obtained from using AIC and BIC together, particularly from trying as far as possible to find models favored by both criteria.*" Hansen and Yu (2001) extensively studied another model selection method called the minimum description length (MDL) and extended the work of Rissanen (1978) where the idea was to select a model based on the shortest description of data. Through simulation, they showed that MDL outperformed AIC, AICC and BIC. They also showed that the two-stage MDL is equivalent to BIC.

The criteria such as AIC were included in this article when selecting a model. Priority was given to the ASE test followed by Adj. R-Sq and the remaining criteria listed in Tables 6 through 8. However, to limit the scope of this paper, MDL was not included in the selection process.

Rissanen (1978) addressed the information criteria in depth. As appeared in that article, "*Sometimes, the AIC-favored model might be so large as to be difficult to use or understand, so the BIC-favored model is a better choice*." The question to consider here is the "Model parsimony." As stated in the above reference, "*Model parsimony is not a motivating goal in its own right, but is a means to reduce unnecessary sampling error caused by having to estimate too many parameters relative to n*." Aho et al. (2014) state "*While some scientists feel that more complex models are always more desirable (cf. Gelman, 2009), others prefer those that balance uncertainty, caused by excessively complex models, and bias, resulting from overly simplistic models. The latter approach emphasizes parsimony.*" The first argument calls for including all the variables, which have significant effects on the model. This of course, might result in an overly complex and overfitting model especially if the sample size is small. The second argument emphasizes a balance between the model complexity and its simplicity. Needless to say that each approach has an advantage and a disadvantage. For example, a more complex model generally requires more expertise. In addition, the selected approach heavily depends on tangible resources.

Applying the log transformation to a dependent variable normalize the residuals (See Lo and Andrews (2015)).The question of the normality assumption on the dependent variable in a GLM has been raised many times in the literatures. The distinction between GLM and generalized linear mixed models (GLMM) is the fact that GLMM does not make the default assumption that this distribution is Gaussian and therefore requires that the researcher specify an appropriate distribution (See Feng, et at. (2014)). Table 5 still indicated a slight departure from normality. However, the skewness and excess kurtosis were close to 0 indicating that the normality assumption was approximately satisfied.

The transformation to normality proposed by Box and Cox (1964) did not play a role in capturing normality beyond what was achieved by the logarithmic transformation (λ= 1). That is, in both 2015 and 2016 data, skewness and kurtosis remained unchanged at the acceptable level for satisfying the normality assumption following the application of this transformation. This was expected, as generally speaking, log transformation will shrink large values much faster than small values, but it does not necessarily make the data normal (Feng et al. (2014))**.** The Box-Cox transformation in fact reduces to a log-transformation when the parameter λ=0. For additional information on the log transformation, see Koch (1966) and Koch, (1969) and the related article by McAlister (1879). Formula (10) holds only for positive values of *y* (here either *towdays* or *lntd*). Since *lntd* will be negative or zero for *towdays* less than or equal to 1, a proper number, say, λ_1, as the scale parameter must be added to it before it is entered into the equation.

Although the proposed method (s) was applied to the fisheries data only, it can be extended and deployed in analyzing

other data sets. However, the results/conclusions might vary depending on the data sets used.

In this research, the GLM consisted of both categorical and continuous variables. The decision of retaining significant parameters was performed by the selection methods and options listed in Table 1and in addition, it was governed by the selection of Training, Validation, and Testing data sets by the software, and also the percentages assigned to these defined by the user.

In order to implement a categorical variable, one must create some coding patterns. Examples of coding patterns include dummy coding, effect coding, and orthogonal coding among others. Statistical software packages have designed their own coding systems. For example, SAS [(1)] and STATA [(1)] use dummy coding whereas JMP[(1)] deploys effect coding.

Categorical variables play an important role in data analysis. However, they present some issues regarding "how" they are to be implemented. For example, multiple imputation was developed to handle this class of variables. However, one must make sure that this method is appropriate for estimating missingness in a given categorical data due to the rounding issue involved (Horton et al., 2003, Marzjarani, 2018[c]).

It must be mentioned that the user must know the default coding pattern used by the software. In SAS[(1)], for example, if the coding pattern is different from the default used by the software, the CLASS option in PROC GLM will not be needed and if it is used, the software simply uses its own default pattern. In other words, the CLASS option in this procedure uses the default coding regardless of the coding patterns selected by the user. In addition, it should be mentioned that some SAS procedures do not support the CLASS option at this time and if a categorical variable were to be passed to these routines, it would have to be passed as dummy or effect coding, for example.

Theoretically, all non-significant levels of a categorical variable should be retained if such variable is significant as a whole. This preserves the relationships among the intercept and the levels of the said categorical variable and estimates. Such relationships include, for example, in the case of only categorical variables using dummy coding, each parameter estimate equals the mean of the corresponding level minus the mean of the reference level and the intercept is equal to the mean of the reference level. In the case of using effect coding, for example, the intercept is equal to the grand mean of the said categorical variable. However, when continuous variables are also present, they will affect these relationships in the sense that one also needs to consider the contributions of these variables to the intercept. The intercept in the model represents the predicted mean value for the response when all covariates in the model are set at their "bases" or reference levels when using the dummy coding pattern, for example. Regardless of which of the coding patterns used, the predicted values must and will be the same, but some figures such as parameter estimates will be different.

If some levels of a categorical variable are non-significant, instead of retaining or removing those from the model, one approach is to combine (collapse) the said level (s) with the corresponding reference level because their distinctions from the reference level seem to have no significant effect on the response variable. Clearly, none of the retaining, removing, or collapsing the non-significant levels of a categorical variable is justified and none provides a perfect solution to the issue. Box (1979) states, "*Essentially, all models are wrong, but some are useful."*

(1). References to any software package throughout this article does not imply the endorsement of the said product.

4. Disclaimer

"The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author(s) and do not necessarily reflect those of National Oceanic and Atmospheric Administration or the Department of Commerce."

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology, 95*(3), 2014, 631–636. https://doi.org/10.1890/13-1452.1

American Journal of Mathematics and Statistics 2018, *8*(2), 36-49.

Ben ́ez, J., & Liu, X. (2013). A short proof of a matrix decomposition with applications. Linea Algebra and its Applications, 438, 1398-1414. https://doi.org/10.1016/j.laa.2012.10.002

Box, G. E. P., & Cox, D. R. (1964). An analysis of Transformation. Journal of Royal Statistical Society, 211-252.

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

Cherkassky, V., & Ma, Y. (2003). Comparison of Model Selection for Regression. *Neural Computation, 15*, 1691-1714. https://doi.org/10.1162/089976603321891864

Dereniowski, D., & Kubale, M. (2004). Choleski Factorization of Matrices in Parallel and Ranking of Graphs, 5th International Conference on Parallel Processing and Applied Mathematics, Lecture. https://doi.org/10.1007/978-3-540-24669-5_127

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). Sensitivity and specificity of information criteria. The Methodology Center, The Pennsylvania State University, Technical Report Series #12-119, 1-30.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis, *Shanghai Archives of Psychiatry, 26*(2).

Fomby, T. B., Johnson, S. R., & Hill, R. C. (1984). Feasible Generalized Least squares Estimation. Advanced Economic Methods, 147-169. https://doi.org/10.1007/978-1-4419-8746-4_8

Gelman, A. (2009). Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics1. Statistical Science, *24*(2), 176–178. https://doi.org/10.1214/09-STS284D

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of Minimum description length, Journal of the *American Statistical Association, 96*(454), 746-774. https://doi.org/10.1198/016214501753168398

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: Data mining, inference and prediction. New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-21606-5

Horton N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple Imputation, *American Statistician, 57*, 229-232. https://doi.org/10.1198/0003130032314

Koch, A. L. (1966). The logarithm in biology I. Mechanisms generating the log-normal distribution exactly. J. Theor. Biol., *12*(2), 276-90. https://doi.org/10.1016/0022-5193(66)90119-6

Koch, A. L. (1969). The logarithm in biology: II. Distributions simulating the log-normal. Journal of Theoretical Biology, *23*(2), 251-268. https://doi.org/10.1016/0022-5193(69)90040-X

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research, 2*, 188-229. https://doi.org/10.1177/0049124103262065

Lo, S., & Andrews, S. (2015). To transform or not to transform using generalized linear mixed models to analyse reaction time data. Front. *Psychol. 6*(1171). https://doi.org/10.3389/fpsyg.2015.01171

Lubke, G. H., Campbell, I., McArtor, D., Miller, P., Luningham, J., & van den Berg, S. M. (2017). Assessing Model Selection Uncertainty Using a Bootstrap Approach: An update. Structural Equation Modeling: *A multidisciplinary Journal, 24*(2), 230-245. https://doi.org/10.1080/10705511.2016.1252265

Marzjarani, M. (2016). Higher Dimensional Linear Models: An Application to Shrimp Effort in the Gulf of Mexico (Years 2007-2014), *International Journal of Statistics and Applications, 6*(3), 96-104.

Marzjarani, M. (2018[b]). 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), 42-52.

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[c]). Using Fuzzy Logic or Probability Approach in Revising Unknown, Invalid, or Missing Data Points: Application to Shrimp Data Files in the Gulf of Mexico, Years 2005 and 2006. *American Journal of Mathematics and Statistics, 8*(2), 36-49.

McAlister, D. (1879). The law of the geometric mean. *Proc R Soc London, 29*, 367-76. https://doi.org/10.1098/rspl.1879.0061

Musau, V. M., Waititu, A. G., & Wanjoya, A. K. (2015). Modeling Panel Data: Comparison of GLS Estimation and Robust Covariance Matrix Estimation. American Journal of Theoretical and Applied Statistics, 185-191. https://doi.org/10.11648/j.ajtas.20150403.25

Notes on Computer Science, 3019, Springer-Verlag, 985-992.

Poloni, F., & Sbrana, G. (2014). Feasible generalized least squares estimation of multivariate GARCH (1, 1) models. *Journal of Multivariate Analysis, 129*, 151-159. https://doi.org/10.1016/j.jmva.2014.04.015

Rissanen, J. (1978). Modeling by Shortest Data Description, *Automatica, 14,* 465–471. https://doi.org/10.1016/0005-1098(78)90005-5

Robustness in the strategy of scientific model building, in Launer, R. L., Wilkinson. (1979)

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences Between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17,* 228-243. https://doi.org/10.1037/a0027127

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

Wooldridge, J. (2002). Econometric Analysis of Cross Section and Panel Data, MIT Press.

Zucchini, W. (2000). An Introduction to Model Selection, *Journal of Mathematical Psychology, 44*, 41-61. https://doi.org/10.1006/jmps.1999.1276