International Journal of Probability and Statistics 2018, 7(1): 19-30
DOI: 10.5923/j.ijps.20180701.03

# Heteroscedastic and Homoscedastic GLMM and GLM: Application to Effort Estimation in the Gulf of Mexico Shrimp Fishery, 1984 through 2001

**Morteza Marzjarani**

NOAA, National Marine Fisheries Service, Southeast Fisheries Science Center, Galveston Laboratory, Galveston, USA

**Abstract** This article presents an overview of the homoscedastic and heteroscedastic Generalized Linear Mixed Model (GLMM) and General Linear Model (GLM). Mathematical relations are defined which map categorical variables onto continuous covariates. It is shown that these relations can be used for different purposes including the addition of pairwise interactions and higher order terms (such as nested terms) to the models. The impacts of these relations on 1984 through 2001 shrimp efforts data in the Gulf of Mexico (GOM), year by year or all years together are compared in the paper. These data sets are also checked for possible heteroscedasticity using Breusch-Pagan and the White's test. It was observed that these data sets show some degree of the heteroscedasticity. The method of weighted least square (WLS) was applied to these data sets and shrimp efforts were estimated before and after the corrections for the heteroscedasticity were made. In addition, it was shown that both the GLMM and the GLM represent these data sets in a satisfactory manner. Efforts generated via a GLMM and a GLM for both homoscedastic and heteroscedastic models showed that although each data set was heteroscedastic, the severity of the heteroscedasticity was compromised when the data sets 1984 through 2001 were compared.

**Keywords** Estimation, General linear models, Generalized linear mixed models, Heteroscedasticity

## 1. Introduction

The purpose of this article was to study the statistical issues related to the effort data files in the northern Gulf of Mexico (GOM). It was not intended to recommend any method for the effort estimation, but rather to examine the 1984 through 2001 shrimp data files for the heteroscedasticity and the possible impact of it on the shrimp effort estimation. The readers are encouraged to read the literature and the current method used to estimate shrimp effort in the GOM by the National Marine Fisheries Service [1].

The General Linear Mixed Model (GLMM) is an extension of the General Linear Model (GLM) which in addition to the fixed portion, it also includes a random part. In the GLMM the word "Generalized" indicates that the response variable is not necessarily normal and the word "Mixed" refers to the random effects in addition to the fixed effects in the GLM. The GLMM has been addressed by many authors. Authors in [2] cover a large number of the applications of this model in social sciences. References [3-9] have addressed the generalized linear mixed models extensively.

The general linear model has been used to estimate shrimp effort in the Gulf of Mexico (GOM) [10]. Reference [11] extensively applied different aspects of the GLM to estimate the shrimp effort in the GOM for the years 2007 through 2014 including the introduction of a mathematical relation to the model. This article presents an overview of homoscedastic and heteroscedastic models. The GLMM and GLM will be applied to the shrimp data files 1984 through 2001. The mathematical relations presented in [11] are extended and applied to different scenarios including pairwise interactions and nested models. Since the data sets used in the article extend over a period of time (18 years), from a statistical perspective, it is of interest to measure the impact of year on the analysis performed on the data sets.

National Marine Fisheries Service (NMFS) is responsible for shrimp effort estimation in the Gulf of Mexico (GOM). NMFS port agents and state trip tickets record the daily operations and shrimp production of the commercial fisheries fleet operating within the boundaries of the U.S. GOM [12]. For assigning fishing activity to a specific geographical location, scientists have subdivided the continental shelf of U.S. Gulf of Mexico into 21 statistical *subarea*s [13]. Subareas 1-9 represent areas off the west coast of Florida, 10-12 represent Alabama/Mississippi, 13-17 represent Louisiana, and 18-21 are designated to Texas (Figure 1). These subareas are further subdivided into

* Corresponding author:
morteza.marzjarani@noaa.gov (Morteza Marzjarani)
Published online at http://journal.sapub.org/ijps

5-fathom depth increments from the shoreline out to 50 fathoms ([1], Page 5). These divisions are used by port agents and the state trip ticket system to assign the location of catches and fishing effort expended by the shrimp fleet on a trip-by-trip basis The shrimp data files include 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* |
| *subarea* | Division of the GOM into 21 statistical *subareas* (1 to 9, 10 to 12, 13 to 17, and 18 to 21) |
| *fathomzone* | *Depth* of water where the shrimp was caught (1 to 2, 3 to 6, and 7 to 12 fathoms) |
| *daysfished* | On 24 hours/day fished. Included all interviewed and non-interviewed records |
| *pounds* | *Pounds* of shrimp harvested |
| *priceppnd* | Average real price per pound of shrimp |
| *shore* | Code for identification of inshore/offshore records (inshore=2, offshore=1) |

An additional file called the Vessel file used in this research was the US Coast Guard file containing vessel id number (*vessel*) and the corresponding vessel lengths (*length*) among other pertaining information not used in this research.

## 2. Methodology

This study focused only on the offshore records in the shrimp data files 1984 through 2001 in the northern GOM. Since the shrimp data files contained both inshore and offshore data, the first step was to identify and then remove the inshore records from these files. All the records in the shrimp data files 1984 through 2001 with the shore code 2 were removed from these files.

Scientists have divided the Gulf of Mexico into 21 statistical *subarea*s as shown (Figure 1).

Since the focus of this study was on the northern portion of the GOM, the *subarea* data points were limited to the values satisfying the following relation and all others including the corresponding records from the shrimp files were deleted.

$$subarea \in \{subarea \mid subarea_i \leq 21,\ i=1,2,3,\ldots\} \quad (1)$$

Similarly, the values in the field *fathomzone* were limited to those satisfying the following relation.

$$fathomzone \in \{fathomzone \mid fathomzone_i,\ i \leq 12,\ i=1,2,3,\ldots\} \quad (2)$$

For each record, the *fathomzone* values falling outside of the interval [0, 12] were revised using the middle column of Table 2. That is, for example, if in a shrimp data file, a record had a value 22 in its *fathomzone* field, this value was changed to 5 using Table 2 (columns 1 and 2).

Upon the completion of this step, all data points in this field in each shrimp file satisfied the following relation.

$$fathomzone \in \{fathomzone \mid fathomzone \in \{1,2,\ldots,12\}\} \oplus \{fathomzone \mid fathomzone \in \{interval\ (i,j) given\ in\ the\ middle\ column\ of\ Table\ 2\}\} \quad (3)$$

where the first set gives the *fathomzone*s up to including 12 and the second set generates the equivalent *fathomzone*s between 1 and 12. The symbol $\oplus$ stands for "Exclusive or Exclusive Disjunction or XOR operator."

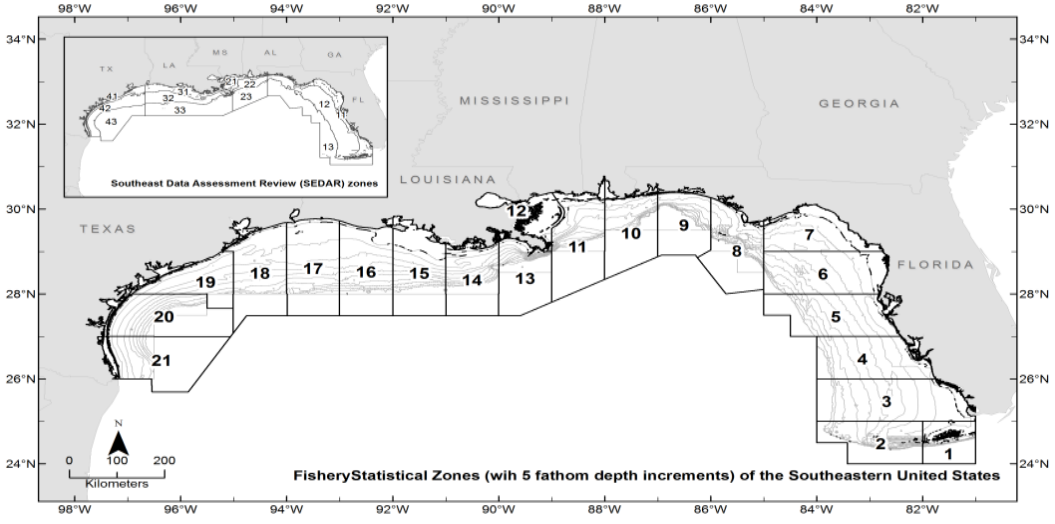


**Figure 1.** The Gulf of Mexico is divided into twenty-one statistical *area*s (1-21) as shown

**Table 2.** *Fathomzone*s (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 |

As an alternative to using Table 2 to revise *fathomzone* values above 12, the modes of *fathomzone* values below 12 for each vessel were found (see formula (4)). Then all *fathomzone*s above 12 per vessel were replaced with the corresponding mode.

$$fathomzone = L_m + C_m * \frac{a}{a+b}, \quad (4)$$

In this formula, $L_m$ is the lower limit of the modal class (class containing the mode), $C_m$ is the class width of the modal class, $a$ is the absolute value of the difference between the frequencies of the modal class and the preceding class, and $b$ is the absolute values of the difference between frequency of the modal class and the next class.

In the next step, using Table 2 and Table 3, *fathomzone* and *subarea* fields were converted to two categorical variables *depth* (a categorical variable with 3 levels) and *area* (with 4 levels) respectively.

The calendar year was placed into three trimesters (January-April, May-August, and September-December) and was used as another categorical variable, *trimesters*, with 3 levels.

**Table 3.** Conversion of *subarea* field in the shrimp file to *area*

| *Subarea* | *Area* |
|---|---|
| 1 through 9 | 1 |
| 10 through 12 | 2 |
| 13 through 17 | 3 |
| 18 through 21 | 4 |

In order to further prepare the data for the estimation purposes, the fields *yearU*, *monthU*, and *dayU* in each shrimp file were concatenated and the result was called the *edate* (unload or end date). As explained in more detail in [11], each shrimp file was converted to trips using vessel id (*vessel*), *edate*, and *port*. The weighted average price per pound, *wavgppnd* for each trip was also computed and used as the price per pound per trip. To perform analysis with *year* as a covariate later, the resulting trips files were then appended creating a large file with 785,459 trips. Since the vessel length (*length*) was considered as an independent covariate, the next step was to match all these files against the United State Coast Guard file using the vessel id and locate as many missing vessel lengths as possible.

For the purpose of model development, in cases of year by year or year as a covariate (*year*), in each data file, all records with either missing vessel lengths or missing or 0 *daysfished* were removed. The resulting file was called the "Match" files. All data points in this file were actual values and none was estimated. The natural logarithm of the variables *pounds* (*lnlbs*) and *daysfished* (*lntd*) were used due to their high variability.

For estimating shrimp effort later, the missing vessel lengths in the trips files generated above were estimated using a monotone imputation method [11].

In addition, the mathematical relations proposed by [11] were developed establishing relationships among continuous and categorical variables. This concept provides some interesting flexibilities to the user.

Perhaps, the most interesting example of such relations, as appeared in [11], is what was called "nested" models in that paper. In such models, categorical variables were nested within continuous ones. A case study regarding this scenario has been included in this paper. Additional examples of the mathematical relations mentioned above appear below.

- Suppose the researcher knows from experience that the smaller vessels generally fish in *area*s 3 or 4, and generally fish at a shallower water. In addition, this happens mostly in trimester 2. These vessels catch more shrimp and sell them at a lower price. One way to implement this scenario is the following algorithm:

IF *area* $\geq$ 3 AND *depth* = 1 AND trimester = 2 THEN
*length* = *l* * *length*
*lnlbs* = *p***lnlbs*
*wavgppnd* = *w** *Wavgppnd*
END.

where $l<1$, $p>1$, $w<1$.

- To present a more challenging scenario, suppose the experimenter knows from the past that vessels beginning with vessel id number 12345 are considered at least 60 feet long and they usually fish in *area* 4, at *depth* 2, and in trimester 1. He/she wishes to set the vessel *length* at 60 if a *length* is less than this number. Below is the algorithm for this hypothetical situation.

IF *area* = *4* AND *depth* = *2* AND *trimester* = 1 THEN
IF vessel id begins with *12345* THEN
IF *length*< *60* THEN *length=60*
END
END.

Although, all the examples mentioned above are functions, such relations do not have to be defined as functions and this provides even a greater flexibility to the researcher.

In order to develop a GLMM or a GLM for shrimp effort estimation using the 1984 through 2001 data sets individually or collectively, the variables *length*, *lnlbs*, *wavgppnd*, *area*, *depth*, and trimester and also the first order

interactions between continuous variables were included in the models as described below.

### 2.1. The Model

The model considered in this research is a well-known generalized linear mixed model (GLMM). Algebraically, the model can be written as follows:

$$y_{ij} = \mu + x_{ik}\beta_k + z_{ip}\gamma_p + \varepsilon_{ij} \quad (5)$$

$i=1, 2, \ldots, n,\ j=1, 2, \ldots, n_2,\ k=1, 2, \ldots, n_3,\ p=1, 2, \ldots, n_4$

where $y_{ij} \in \mathbb{R}^+$ is the response, $\mu$ is the overall mean, $x_{ik}$'s $\in \mathbb{R}$ are constant observations, $\varepsilon_{ij}$ 's are error terms, and $z_{ip} = 1$ if $i=p$; 0 otherwise. It is more convenient to write the equation given in (6) in matrix form.

$$\underline{y} = \underline{x}\,\underline{\beta} + \underline{z}\,\underline{\gamma} + \underline{\varepsilon} \quad (6)$$

where $\underline{y} \in \mathbb{R}^+$ is a column n x1 vector of observations, $\underline{x} \in \mathbb{R}$ is an n x $n_1$ matrix relating $\underline{\beta}$ to $\underline{y}$, $\underline{\beta} \in \mathbb{R}$ is a $n_1$ x 1 column vector of fixed portion of the model. Also, $\underline{z} \in \mathbb{R}$ is an n x $n_2$ identity matrix relating $\underline{\gamma}$ to $\underline{y}$, $\underline{\gamma} \in \mathbb{R}$ is an $n_2$ x 1 column vector of random portion, and $\underline{\varepsilon}$ is n x 1 column vector of the error terms, the variability in y not explained by the portion $\underline{x}\,\underline{\beta} + \underline{z}\,\underline{\gamma}$. In (6), $\underline{x}\,\underline{\beta}$ is the non-random portion of the model, $\underline{z}\,\underline{\gamma}$ is the random effect, and $\underline{\varepsilon}$ is the random error part of the model. As is usually the case, it is assumed that $\underline{\gamma}$ is normally distributed with mean 0 and variance covariance $\underline{G}$. It is also assumed that $\underline{\varepsilon}$ is normally distributed with mean 0 and variance-covariance $\underline{R}$. In this model, the fixed part includes the categorical variables *area*, *depth*, *trimester*, and *year* (where applicable). The random portion of the model includes the vessel length (*length*), *lnlbs*, *wavgppnd*, and their pairwise interactions or a mathematical relation.

The matrices $\underline{G}$ and $\underline{R}$ are commonly known as the G-side and the R-side of the model. The random effect determines the G-side of the model variance and it is defined by the RANDOM portion in the model. Furthermore, it is assumed that $\underline{\gamma}$ and $\underline{\varepsilon}$ are uncorrelated. Therefore, one can easily observe that

$$var(\underline{y}) = \underline{z}\,\underline{G}\,\underline{z}' + \underline{R} \quad (7)$$

We notice that the quantity on the right hand side of (7) contains the effects of variations on both sources of randomness. The GLMM includes a link function that relates the mean of the response to the linear predictors in the model. The general form of the link function g (.) is as follows:

$$g(E(\underline{y}|\underline{\gamma})) = \underline{x}\,\underline{\beta} + \underline{z}\,\underline{\gamma} \quad (8)$$

where g(.) is a monotone differentiable function. For a Gaussian response variable $\underline{y}$, the *Identity* function is selected for g (.). The reason for having the link function in GLMM is the fact that unlike GLM, the response variable in GLMM does not need to be normally distributed and its range does not have to be in the interval $(-\infty, +\infty)$. Furthermore, the relationship between predictors and response does not have to be simple relationship. The link function establishes a relationship between these components of the model in such a way that the range of the non-linearly transformed mean $g(.)$ ranges from $-\infty$ to $+\infty$. The model defined in (6) becomes equivalent to a randomized block design if the non-random portion of the model is dropped. Clearly, in the absence of the random portion of the model, equation (6) reduces to a GLM. In the case model (6) contains only the random effects, the $\underline{G}$ matrix is of interest. On the other hand, in a model with fixed effects only the $\underline{R}$ matrix is of interest.

In the absence of the random effects and if $\underline{R}$ in the form $\sigma^2 I$, where I is an *identity* matrix (that is, a homoscedastic model), the GLMM reduces to a GLM or a GLM with overdispersion (this term will be defined later in this article). Using a similar approach as in GLM, that is, maximum likelihood estimation (MLE), the parameters in (6) can be estimated as:

$$\hat{\underline{\beta}} = (\underline{x}'(\underline{z}\,\underline{G}\,\underline{z}' + \underline{R})^{-1}(\underline{z}\,\underline{G}\,\underline{z}' + \underline{R})^{-1}\underline{x})^{-1}\underline{x}'(\underline{z}\,\underline{G}\,\underline{z}' + \underline{R})^{-1}\underline{y}$$

$$\hat{\underline{\gamma}} = \underline{G}\,\underline{z}'(\,\underline{z}\,\underline{G}\,\underline{z}' + \underline{R})^{-1}\,(\underline{y} - \underline{x}\,\hat{\underline{\beta}}) \quad (9)$$

In this article, a simple form of the matrix $\underline{R}$ was considered as follows:

$$\underline{R} = \sigma^2\,\underline{r}, \quad (10)$$

where $\underline{r} = (r_{kk})$ are diagonal matrix. In the case of a homoscedastic model, $\underline{r}$ is an identity matrix. Some authors have considered some special cases. For example, [14] assumed that the $\underline{R}$ matrix for a fixed effect model with two covariates was in the form $\underline{R} = \sigma^2\,\underline{r}$, where $\underline{r} = (r_{kk})$, with $r_{kk} = (1/x_{1k}\,x_{2k})^{\delta}$, $k$=1,2,…, n. With such choice for the $\underline{R}$ matrix, the parameter $\delta$ measures the strength of the heteroscedasticity in the model: the lower its magnitude, the smaller the differences between individual variances. Of course, when $\delta = 0$, all the variances are identical (homoscedastic model). In practice, the three values 1/2, 1, and 2 for $\delta$ are of particular interest.

### 2.2. Investigating the Heteroscedasticity in the Data sets 1984 through 2001

The word heteroscedastic (sometimes written as heteroskedastic) has a root from the Ancient Greek hetero and skedasis meaning different and dispersion. The term refers to the variability of a variable being unequal across the variable(s) predicting it. It is a major concern in data analysis (such as regression) as it could invalidate the test results among others. It is not surprising to say that to some degree, most (if not all) data sets are heteroscedastic. In most cases like the normality assumptions, researchers take it for granted by assuming that the data sets are homoscedastic. Nevertheless, before fitting a model to a data set and estimating parameters, it seems logical and rather necessary to check for possible heteroscedasticity especially if the sample size is small. To check for possible heteroscedasticity, Breusch, and Pagan, [15], hereafter called Breusch-Pagan, proposed a method known as Lagrange Multiplier (LM). The method was modified by [16] where the normality assumption of the error term was relaxed. The

Breusch-Pagan tests the linear form of heteroscedasticity. The White's test extends this to include non-linear heteroscedasticity and therefore is more general than the Breusch -Pagan test. The issue with the White's is the addition of new covariates. The test statistic for testing the heteroscedasticity using LM test is:

$$n*R^2 \approx \chi^2_{p\text{-}l}, \quad (11)$$

where $n$ is the sample size and $R^2$ is obtained by fitting the square of residuals to the regressors.

Alternatively, one can use the following to test for the heteroscedasticity.

$$F_{p\text{-}1,n\text{-}p\text{-}1} = ((SST\text{-}SSE)/SSE)/((n\text{-}p\text{-}1)/(p\text{-}1)) \quad (12)$$

where, *SST* and *SSE* are total sum square and sum square error, $n$ is the sample size, and $p$ is the number of parameters in the model including the intercept. To correct the heteroscedasticity, in this article, the weighted least square (WLS) was selected. In this article, three weight functions were defined as follows:

$$w_{1i} = 1/residual_i^2,\ w_{2i} = 1/|residual_i|,\ w_{3i} = 1/(1\text{-}residual_i)^2, \quad (13)$$

where, $residual_i$ is the residual of the regression model fitted to the residuals. In the case of a large sample size, the limiting distribution for detecting the heteroscedasticity [15-16] can asymptotically be approximated with a Chi-square distribution with *p-1* degrees of freedom.

$$lim_{n-p\to\infty}(p-1)\ F_{p\text{-}1,n\text{-}p} \approx \chi^2_{p\text{-}l} \quad (14)$$

For a large sample size, the heteroscedasticity of the variances may or may not severely affect the estimates. As the sample size increases, so does the degree of difficulty in checking for the heteroscedasticity. Nevertheless, each data set ought to be checked to make sure that the data are homoscedastic. We notice that as the sample size increases, the test statistic given in (11) increases. For large samples, the approximate distribution given in (14) might be used. In this article, the data files 1984 through 2001 were checked for the heteroscedasticity. To avoid redundancy and for brevity, and for measuring the impact of sample size on the heteroscedasticity, a few random samples from the data file 1984 were selected and additional analysis were performed on these samples. Alternatively, the model was fitted to all (including the combined) data files 1984 through 2001 and efforts were estimated using the WLS method. In addition, efforts were estimated for the 1984 through 2001 data sets described earlier under the assumption of homoscedasticity. In case of a homoscedastic model, it was further assumed that the link function was *Identity* and the distribution of the response variable was *Gaussian*. The following alternative forms of GLMM were used in this paper.

- It was assumed that the random portion of (6) consisted of all first order terms of continuous variables *length*, *lnlbs*, and *wavgppnd*, also the pairwise interaction terms of these variables. The remaining variables consisting of the categorical variables formed the fixed effects portion of the model. Under these assumptions, the vector $\underline{\gamma}$ contained all first order terms of continuous variables and their pairwise interactions and the vector $\underline{\beta}$ included all the categorical variables.
- For a comparison, a GLM with all first order terms of continuous variables *length*, *lnlbs*, *wavgppnd* and categorical variables *area, depth*, *trimester*, and the pairwise interactions of the continuous variables was applied to the Match files.
- Both GLMM and GLM were fitted to the 1984 through 2001 and also the combined data files under the assumption of heteroscedasticity and were compared to those under the assumption of homoscedastic models.

# 3. Analysis/Results

The objective of this paper was to review the heteroscedastic generalized linear mixed model, and the general linear model, then apply both to 1984 through 2001 shrimp data in the GOM, extend and give some examples of the mathematical relations proposed by [11].

Prior to these, the issue of heteroscedasticity in the shrimp data files 1984 through 2001 was examined. Table 4 displays the results of applying either Formula (11), (12), or (14) to the data sets 1984 through 2001.

**Table 4.** Test statistics generated by applying Formulas (11), (12), or (14) to the 1984 through 2001 shrimp data sets

| Year | $F_{p\text{-}1,n\text{-}p\text{-}1}$* | $nR^2 \approx \chi^2_{p\text{-}l}$* | $lim_{n-p\to\infty}(p-1)\ F_{p\text{-}1,n\text{-}p} \approx \chi^2_{p\text{-}l}$* |
|---|---|---|---|
| 1984 | 16.33 | 207.60 | 212.29 |
| 1985 | 29.61 | 373.57 | 384.93 |
| 1986 | 39,50 | 494.39 | 513.5 |
| 1987 | 39.41 | 489.76 | 512.33 |
| 1988 | 11.38 | 238.72 | 147.94 |
| 1989 | 11.38 | 145.29 | 147.94 |
| 1990 | 7.28 | 93.29 | 94.64 |
| 1991 | 9.30 | 118.87 | 120.9 |
| 1992 | 12.60 | 159.36 | 163.8 |
| 1993 | 11.94 | 151.03 | 155.22 |
| 1994 | 11.64 | 147.41 | 151.32 |
| 1995 | 20.38 | 153.98 | 264.94 |
| 1996 | 12.73 | 161.25 | 165.49 |
| 1997 | 13.98 | 175.52 | 181.74 |
| 1998 | 13.36 | 168.17 | 173.68 |
| 1999 | 16.25 | 201.64 | 211.25 |
| 2000 | 19.13 | 236.09 | 248.69 |
| 2001 | 19.10 | 236.85 | 248.0 |
| 1984-2001 | 116.9 | 3,411.81 | 3,507.03 |

*: All significant at 0.05 level

All test statistics displayed in Table 4 were statistically significant at $\alpha = 0.05$. To further research the issue of the heteroscedasticity and the effects of the sample size on it, a few random samples from the data file 1984 were selected. Table 5 displays the results of testing for the possible heteroscedasticity using the Breusch-Pagan and the White's test in these samples and the whole data set. The

Breusch-Pagan test showed the heteroscedasticity at 5% level in 4 out of 6 samples and in the whole data set. The White's test on the other hand displayed a significant result only in the whole data set. This is due to the fact that this test is more general and covers non-linear heteroscedasticity. It is evident from the table that the sample size directly influences the heteroscedasticity.

Tables 6a and 6b display the results of improving for the heteroscedasticity using the three weights defined earlier. In most cases, the weights corrected the heteroscedasticity in the data set to a point, but the heteroscedasticity remained in the data set. The weight $w_2$ lowered the test statistic more the other two weights. This weight was selected used for the rest of this research.

As displayed in Tables 6a and 6b, the concern about the White's test is the fact that the test generates a large number of covariates and sometimes presents an issue to the analyst. For example, for 13 independent variables, the method generates an additional 91 covariates. For this reason, a simple and approximate form of this test has been proposed [16]. The application of this approach to the 1984 data file generated the following values:

$F_{2,\ 8671}$ =46.10 ($F_{\ 2,\ 8671,\ 0.05}$ = 3.00) or $\chi^2_{stat,\ 2}$ = 45.88 ($\chi^2_{\ 0.05,\ 2}$ = 5.99). Both of these statistics were significant and consistent with the results given in Table 6b. Formulas (11) and (12) were directly applied to the 1984 data set and produced the following results: $F_{13,\ 8660,\ 0.05}$ = 212.34 ($F_{crit\ ,\ 0.05}$= 1.75) or $\chi^2_{13}$ = 207.59 ($\chi^2_{\ crit}$ = 22.36). Again, both of these test statistics were statistically significant.

**Table 5.** Results of testing for the heteroscedasticity using the Breusch-Pagan and White's tests to the 1984 shrimp data file

| Sample | Sample | Sample size | $t_{stat}$ | df | *p-value* |
|---|---|---|---|---|---|
| Breusch-Pagan | 1 | 82 | 25.12 | 13 | 0.0223 |
| | 2 | 98 | 22.38 | 13 | 0.0498 |
| | 3 | 155 | 18.02 | 13 | 0.1569 |
| | 4 | 192 | 23.81 | 13 | 0.0330 |
| | 5 | 264 | 23.81 | 13 | 0.0329 |
| | 6 | 281 | 22.18 | 13 | 0.0527 |
| | 1984 data set | 8674 | 207.60 | 13 | < 0.0001 |
| White's test | 1 | 82 | 74.28 | 75 | 0.5019 |
| | 2 | 98 | 81.28 | 74 | 0.2630 |
| | 3 | 155 | 87.70 | 79 | 0.2354 |
| | 4 | 192 | 69.53 | 81 | 0.8145 |
| | 5 | 264 | 97.75 | 81 | 0.0992 |
| | 6 | 281 | 86.09 | 86 | 0.4769 |
| | 1984 data set | 8674 | 1008 | 87 | < 0.0001 |

**Table 6a.** Results of applying Breusch-Pagan to the 1984 shrimp data after the application of $w_1$, $w_2$, or $w_3$ to detect for the heteroscedasticity

| Sample | Sample size | Weight | $t_{stat}$ | df | *p-value* |
|---|---|---|---|---|---|
| 1 | 82 | $w_1$ | 17.39 | 13 | 0.1823 |
| | | $w_2$ | 13.04 | 13 | 0.4449 |
| | | $w_3$ | 20.09 | 13 | 0.0905 |
| 2 | 98 | $w_1$ | 0.68 | 1 | 0.4090 |
| | | $w_2$ | 0.69 | 1 | 0.4054 |
| | | $w_3$ | 7.22 | 13 | 0.8904 |
| 3 | 155 | $w_1$ | 20.73 | 13 | 0.0783 |
| | | $w_2$ | 5.08 | 13 | 0.9733 |
| | | $w_3$ | 53.01 | 13 | < 0.0001 |
| 4 | 192 | $w_1$ | 10.51 | 13 | 0.6818 |
| | | $w_2$ | 7.24 | 13 | 0.8894 |
| | | $w_3$ | 19.90 | 13 | 0.0978 |
| 5 | 264 | $w_1$ | 16.71 | 13 | 0.2129 |
| | | $w_2$ | 6.67 | 13 | 0.9185 |
| | | $w_3$ | 24.18 | 13 | 0.0300 |
| 6 | 281 | $w_1$ | 4.34 | 13 | 0.9870 |
| | | $w_2$ | 6.01 | 13 | 0.9456 |
| | | $w_3$ | 28.47 | 13 | 0.0078 |
| 1984 data set | 8674 | $w_1$ | 1683 | 13 | < 0.0001 |
| | | $w_2$ | 46.48 | 13 | < 0.0001 |
| | | $w_3$ | 2563 | 13 | < 0.0001 |

**Table 6b.** Results of applying the White's test to the 1984 shrimp data after the application of $w_1$, $w_2$, or $w_3$ to detect for the heteroscedasticity and displaying test statistics for $w_2$

| Sample | Sample size | Weight | $t_{stat}$ | df | *p-value* | $(p-1)F_{p-1,n-p-1}$ | $nR^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 82 | $w_1$ | 75.00 | 74 | 0.4457 | | |
| | | $w_2$ | 68.85 | 75 | 0.6780 | 30.02 | 25.11 |
| | | $w_3$ | 60.00 | 59 | 0.4392 | | |
| 2 | 98 | $w_1$ | 4.29 | 12 | 0.9777 | | |
| | | $w_2$ | 4.77 | 12 | 0.9653 | 24.86 | 22.38 |
| | | $w_3$ | 66.00 | 65 | 0.4421 | | |
| 3 | 155 | $w_1$ | 104.6 | 79 | 0.0284 | | |
| | | $w_2$ | 78.18 | 79 | 0.5050 | 18.54 | 18.02 |
| | | $w_3$ | 96.88 | 77 | 0.0624 | | |
| 4 | 192 | $w_1$ | 66.23 | 81 | 0.8823 | | |
| | | $w_2$ | 57.65 | 81 | 0.9769 | 25.19 | 23.81 |
| | | $w_3$ | 66.79 | 78 | 0.8133 | | |
| 5 | 264 | $w_1$ | 94.53 | 81 | 0.1444 | | |
| | | $w_2$ | 84.35 | 81 | 0.3776 | 24.79 | 23.81 |
| | | $w_3$ | 109.00 | 73 | 0.0041 | | |
| 6 | 281 | $w_1$ | 75.36 | 86 | 0.7869 | | |
| | | $w_2$ | 68.20 | 86 | 0.9211 | 22.89 | 22.18 |
| | | $w_3$ | 105.40 | 86 | 0.0763 | | |
| 1984 data set | 8674 | $w_1$ | 2,954 | 87 | < 0.0001 | | |
| | | $w_2$ | 5,73.3 | 87 | < 0.0001 | 212.34 | 207.59 |
| | | $w_3$ | 3,581 | 87 | < 0.0001 | | |

Again, as displayed above (Tables 6a and 6b), out of the three weights for detecting the heteroscedasticity, $w_2$ performed better and it was selected and used in the remaining of this article. Tables 7a and 7b show the results of before and after the application of $w_2$ to the 1984 through 2001 data sets.

The Breusch-Pagan and White's tests showed that the heteroscedasticity was improved by 100% and 77% respectively over the period of 1984 through 2001 using $w_2$ as the weight (Tables 7a and 7b).

It was difficult to measure the severity of the heteroscedasticity in a given data set. The tests presented and implemented above, simply tell us whether the data set is heteroscedastic. However, it was of great interest to measure the impact of the heteroscedasticity on the shrimp effort estimation. As an alternative to the above tests, efforts were estimated using the model given in (6) under the assumption of heteroscedasticity, corrected using the weight $w_2$, and compared to those under homoscedasticity. The results are displayed in Table 8.

**Table 7a.** Test statistics generated by the Breusch-Pagan test before and after the application of $w_2$ to the 1984-2001 shrimp data sets

| | Before | | | After | | |
|---|---|---|---|---|---|---|
| Year | $t_{stat}$ | df | *p-value* | $t_{stat}$ | df | *p-value* |
| 1984 | 207.6 | 13 | <0.0001 | 46.48 | 13 | <0.0001 |
| 1985 | 373.6 | 13 | <0.0001 | 106.6 | 13 | <0.0001 |
| 1986 | 494.4 | 13 | <0.0001 | 73.41 | 13 | <0.0001 |
| 1987 | 489.8 | 13 | <0.0001 | 78.55 | 13 | <0.0001 |
| 1988 | 238.7 | 13 | <0.0001 | 38.99 | 13 | <0.0001 |
| 1989 | 145.3 | 13 | <0.0001 | 23.84 | 13 | <0.0001 |
| 1990 | 93.29 | 13 | <0.0001 | 26.49 | 13 | 0.0146 |
| 1991 | 118.9 | 13 | <0.0001 | 26.91 | 13 | 0.0128 |
| 1992 | 159.4 | 13 | <0.0001 | 40.96 | 13 | <0.0001 |
| 1993 | 151.0 | 13 | <0.0001 | 34.35 | 13 | 0.0011 |
| 1994 | 147.4 | 13 | <0.0001 | 74.57 | 13 | <0.0001 |
| 1995 | 254.0 | 13 | <0.0001 | 83.84 | 13 | <0.0001 |
| 1996 | 161.2 | 13 | <0.0001 | 41.32 | 13 | <0.0001 |
| 1997 | 175.5 | 13 | <0.0001 | 99.49 | 13 | <0.0001 |
| 1998 | 168.2 | 13 | <0.0001 | 49.26 | 13 | <0.0001 |
| 1999 | 201.6 | 13 | <0.0001 | 61.68 | 13 | <0.0001 |
| 2000 | 236.1 | 13 | <0.0001 | 35.04 | 13 | 0.0008 |
| 2001 | 236.9 | 13 | <0.0001 | 21.32 | 13 | 0.0669 |
| 1984-2001 | 2,785 | 13 | <0.0001 | 930.2 | 13 | <0.0001 |

**Table 7b.** Test statistics generated by the White's test before and after the application of $w_2$ to the 1984-2001 shrimp data sets

| | Before | | | After | | |
|---|---|---|---|---|---|---|
| Year | $t_{stat}$ | df | *p-value* | $t_{stat}$ | df | *p-value* |
| 1984 | 1008 | 87 | <0.0001 | 573 | 87 | <0.0001 |
| 1985 | 1071 | 87 | <0.0001 | 1573 | 87 | <0.0001 |
| 1986 | 1059 | 87 | <0.0001 | 604 | 87 | <0.0001 |
| 1987 | 1242 | 87 | <0.0001 | 786.4 | 87 | <0.0001 |
| 1988 | 742.5 | 87 | <0.0001 | 652.9 | 87 | <0.0001 |
| 1989 | 513.6 | 87 | <0.0001 | 386.5 | 87 | <0.0001 |
| 1990 | 326.6 | 86 | <0.0001 | 265.7 | 86 | <0.0001 |
| 1991 | 778.5 | 87 | <0.0001 | 529.7 | 87 | <0.0001 |
| 1992 | 1042 | 87 | <0.0001 | 404.4 | 87 | <0.0001 |
| 1993 | 609 | 87 | <0.0001 | 391.7 | 87 | <0.0001 |
| 1994 | 545.9 | 87 | <0.0001 | 1052 | 87 | <0.0001 |
| 1995 | 723.3 | 87 | <0.0001 | 955.8 | 87 | <0.0001 |
| 1996 | 538.7 | 87 | <0.0001 | 425.4 | 87 | <0.0001 |
| 1997 | 674.2 | 87 | <0.0001 | 547.9 | 87 | <0.0001 |
| 1998 | 685.9 | 87 | <0.0001 | 507.4 | 87 | <0.0001 |
| 1999 | 511.1 | 87 | <0.0001 | 517.1 | 87 | <0.0001 |
| 2000 | 606.3 | 86 | <0.0001 | 417.2 | 86 | <0.0001 |
| 2001 | 707.5 | 87 | <0.0001 | 505.7 | 87 | <0.0001 |
| 1984-2001 | 8,151 | 87 | <0.0001 | 7,816 | 87 | <0.0001 |

**Table 8.** Efforts generated using equation (4) and under homoscedasticity or heteroscedasticity assumptions for the years 1984 through 2001 using $w_2$

| | Homoscedastic | | Heteroscedastic | |
|---|---|---|---|---|
| | GLMM | GLM | GLMM | GLM |
| 1984 | 123,931 | 124,180 | 123,422 | 123,814 |
| 1985 | 131,455 | 131,779 | 131,906 | 131,556 |
| 1986 | 221,309 | 142,624 | 199,900 | 143,064 |
| 1987 | 169,422 | 204,299 | 170,563 | 182,337 |
| 1988 | 127,808 | 113,550 | 125,856 | 114,264 |
| 1989 | 160,419 | 129,820 | 139,755 | 114,716 |
| 1990 | 150,782 | 150,495 | 150,799 | 150,937 |
| 1991 | 173,281 | 172,918 | 170,975 | 234,089 |
| 1992 | 171,495 | 171,777 | 171,405 | 171,743 |
| 1993 | 157,562 | 156,947 | 295,533 | 156,057 |
| 1994 | 159,544 | 159,097 | 158,181 | 158,259 |
| 1995 | 148,866 | 172,178 | 147,586 | 147,887 |
| 1996 | 148,526 | 148,972 | 149,723 | 148,966 |
| 1997 | 157,804 | 157,426 | 154,985 | 155,422 |
| 1998 | 162,849 | 162,227 | 162,607 | 162,565 |
| 1999 | 161,337 | 161,558 | 167,468 | 166,505 |
| 2000 | 130,547 | 176,769 | 134,625 | 132,279 |
| 2001 | 145,054 | 132,135 | 163,177 | 185,918 |
| 1984-2001 | 2,839,124 | 2,841,334 | 2,847,640 | 2,844,519 |

Analysis showed that excluding the last row, there was no significant difference between the means of columns 2 and 4 ($t_{stat}$ = -0.81, *p-value* = 0.43, $t_{creit}$ = 2.11) and 3, and 5 ($t_{stat}$ = -1.12, *p-value* = 0.28, $t_{creit}$ = 2.11). The heteroscedastic GLMM generated a slightly higher effort estimates than the homoscedastic GLMM. The homoscedastic GLM on the other hand, generated slightly higher estimates than the heteroscedastic GLM.

Further analysis showed that the overall effort means were statistically different ($F_{20, 51}$ =3.09, *p-value* = 0.0006). The effort means within year were statistically placed in three groups (Table 9a), and the effort means within the model were all placed in one category (Table 9b). Therefore, for the remaining analyses in the paper, it was assumed that the data sets 1984 through 2001 were homoscedastic.

**Table 9a.** Grouping of the mean efforts per year (sorted by group means-descending)

| Year | Group B | GROUP D | Group A | Group C |
|---|---|---|---|---|
| 1993 | | | A | |
| 1991 | B | | A | |
| 1987 | B | | A | C |
| 1986 | B | D | A | C |
| 1992 | B | D | A | C |
| 1999 | B | D | A | C |
| 1998 | B | D | A | C |
| 1994 | B | D | A | C |
| 2001 | B | D | A | C |
| 1997 | B | D | A | C |
| 1995 | B | D | A | C |
| 1990 | B | D | A | C |
| 1996 | B | D | A | C |
| 2000 | B | D | A | C |
| 1989 | | D | A | C |
| 1985 | | D | | C |
| 1984 | | D | | C |
| 1988 | | D | | |

**Table 9b.** Grouping of the mean efforts per model (sorted by group means-descending)

| Model | Group |
|---|---|
| GLMM (corrected for heteroscedasticity) | A |
| GLMM (not corrected for heteroscedasticity) | A |
| GLM (corrected for heteroscedasticity) | A |
| GLM (not corrected for heteroscedasticity) | A |

Table 10 shows some statistics of interest including Pearson dispersion parameter and Adj. R-Squared when fitting both GLMM and GLM to the data sets under the assumption of homoscedasticity.

**Table 10.** Some statistics of interest when fitting both GLMM and GLM to shrimp data sets 1984 through 2001 year by year under the assumption of homoscedasticity

| year | Chi-Sq/DF GLMM | R-Sq GLM |
|---|---|---|
| 1984 | 0.38 | 0.61 |
| 1985 | 0.40 | 0.55 |
| 1986 | 0.51 | 0.57 |
| 1987 | 0.39 | 0.65 |
| 1988 | 0.38 | 0.61 |
| 1989 | 0.35 | 0.64 |
| 1990 | 0.32 | 0.68 |
| 1991 | 0.35 | 0.65 |
| 1992 | 0.28 | 0.79 |
| 1993 | 0.27 | 0.78 |
| 1994 | 0.38 | 0.66 |
| 1995 | 0.39 | 0.60 |
| 1996 | 0.37 | 0.64 |
| 1997 | 0.39 | 0.68 |
| 1998 | 0.53 | 0.56 |
| 1999 | 0.55 | 0.56 |
| 2000 | 0.30 | 0.71 |
| 2001 | 0.53 | 0.41 |
| 1984-2001 | 0.42 | 0.62 |

Table 11, contains efforts generated via GLMM and GLM for the years 1984 through 2001 depending on whether the middle column of Table 2 or formula (4) were used to revise *fathomzone* over 12 (under the assumption of homoscedasticity). Figure 2 is the display of the same. A simple analysis (for example, ANOVA with a single factor) showed that there was no significant difference among the mean columns in this table ($F_{stat}$ = 0.16, $F_{crit,\ 3,\ 68}$= 2.74, *p-value*= 0.92).

**Table 11.** Efforts generated by GLMM and GLM for years 1984 through 2001 using Table 2 or Formula (4) to revise *fathomzone*s over 12 (under the assumption of homoscedasticity)

| Year | GLMM effort using Table 2 | GLM effort using Table 2 | GLMM effort using mode | GLM effort using mode |
|---|---|---|---|---|
| 1984 | 123,761 | 124,001 | 123,931 | 124,180 |
| 1985 | 131,455 | 131,779 | 131,455 | 131,779 |
| 1986 | 222,170 | 142,624 | 221,309 | 142,624 |
| 1987 | 169,417 | 204,057 | 169,422 | 204,299 |
| 1988 | 127,227 | 113,115 | 127,808 | 113,550 |
| 1989 | 160,413 | 129,814 | 160,419 | 129,820 |
| 1990 | 150,654 | 150,351 | 150,782 | 150,495 |
| 1991 | 173,577 | 173,021 | 173,281 | 172,918 |
| 1992 | 171,827 | 171,792 | 171,495 | 171,777 |
| 1993 | 157,485 | 157,002 | 157,562 | 156,947 |
| 1994 | 159,471 | 159,364 | 159,544 | 159,097 |
| 1995 | 149,815 | 172,045 | 148,866 | 172,178 |
| 1996 | 148,684 | 148,927 | 148,526 | 148,972 |
| 1997 | 156,683 | 157,462 | 157,804 | 157,426 |
| 1998 | 162,886 | 162,332 | 162,849 | 162,227 |
| 1999 | 162,245 | 161,581 | 161,337 | 161,558 |
| 2000 | 176,230 | 176,805 | 130,547 | 176,769 |
| 2001 | 147,230 | 133,789 | 145,054 | 132,135 |

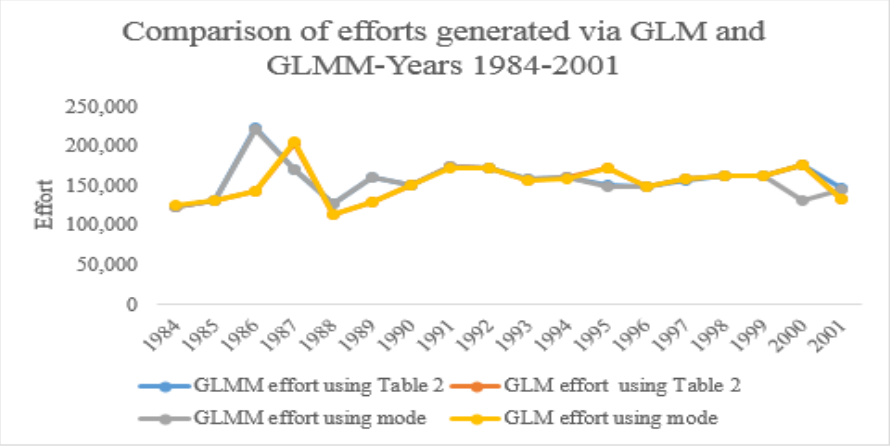


**Figure 2.** Efforts generated year by year via a GLMM or a GLM using Table 2 or Formula (4) to revise *fathomzone*s over 12 (under the assumption of homoscedasticity)

**Table 12.** Efforts generated by GLMM and GLM for years 1984 through 2001 with and without nested terms, *fathomzone* determined via Table 2 or equation (4) (under the assumption of homoscedasticity)

| Model | GLMM Chi-Sq/DF | GLM R-Sq | Effort GLMM | Effort GLM | Average over 18 years-GLMM | Average over 18 years-GLM |
|---|---|---|---|---|---|---|
| First order and interactions (using Table 2) | 0.42 | 0.62 | 2,724,900 | 2,716,193 | 151,383 | 150,900 |
| First order and nested terms (using Table 2) | 0.41 | 0.63 | 2,674,305 | 2,777,294 | 148,572 | 154,294 |
| First order and interactions (using equation (4) | 0.43 | 0.61 | 2,839,124 | 2,841,334 | 157,729 | 157,852 |
| First order and nested terms (using equation (4) | 0.42 | 0.63 | 2,782,133 | 2,736,638 | 154,563 | 152,035 |

In the next step, again both GLMM and GLM were applied to the combined shrimp file with year as a covariate in the model defined in (6). Table 12 displays the total efforts generated by GLMM and GLM for this large data set where either the middle column of Table 2 or equation (4) were used to revise *fathomzone* values over 12 (under the assumption of homoscedasticity).

A comparison of the sum of the actual towdays and the sum of the predicted values by the models in the Match file for the years 1984 through 2001 with year as a covariate and using Table 2 to revise *fathomzone*s showed that the two models have been fitted reasonably well (Table 13).

**Table 13.** Results of fitting GLMM or GLM to the combined Match file using Table 2 to revise *fathomzone*s above 12

| Model | GLMM | GLM |
|---|---|---|
| Sum of actual lntd | 147,099 | 147,099 |
| Sum of predicted lntd | 147,422 | 147,045 |
| Sum of actual *daysfished* | 639,731 | 639,731 |
| Sum of predicted *daysfished* | 544,290 | 542,447 |

## 4. Discussion and Concluding Remarks

The goal of this research was to present an overview of the GLMM and the GLM for both homoscedastic and heteroscedastic, to further study the mathematical relations defined in [11] and then to apply these models to the 1984 through 2001 shrimp data files in the GOM. The heteroscedasticity issue is a challenging one to a point that authors usually assumes that the data sets are homoscedastic.

Authors in [15] addressed this issue and developed a method for testing the existence of the heteroscedasticity in a data set. Reference [16] modified the method by assuming that the error terms were not necessarily normal also included the non-linear heteroscedasticity in his approach. Clearly, the White's test is more general, but generates many covariates. The test generally requires higher degrees of freedom and not necessarily consistent with Breusch-Pagan test (Tables 7a and 7b). Reference [17] showed that under certain conditions the two tests are algebraically equivalent.

Reference [14] developed an iterative method for estimating the parameters for a heteroscedastic linear regression model with two covariates in a special case. This approach was difficult to deploy here due to a relatively large number of covariates. In this article, the methods proposed by [15, 16] were deployed for testing the existence of the heteroscedasticity in the 1984 through 2001 shrimp data files. It was observed that these data sets include heteroscedasticity. The weighted least square (WLS) method with the inverse of the absolute value of the predicted values from the regression with the residuals as the response ($w_2$) was deployed to correct for the heteroscedasticity. It was concluded that the weight $w_2$ reduced the heteroscedasticity (Tables 7a and 7b).

To measure the impact of the heteroscedasticity on the shrimp effort estimation, alternatively efforts were estimated for the years 1984 through 2001 and all years combined

under the assumptions of both heteroscedasticity and homoscedasticity. The results showed that over this period, the heteroscedasticity did not cause a significant difference in effort estimation. Clearly, the severity of the heteroscedasticity was compromised by year since the data for each year individually showed the existence of the heteroscedasticity.

Under the assumption of homoscedasticity, in the case of the GLMM, the Pearson Chi-Sq/DF ranged from 0.27 to 0.53 where the analysis was performed on the year by year basis. The numerator of this ratio is a quadratic form in the marginal residuals that takes correlations among the data into account. This ratio is known as "overdispersion" which means that the variability in the data is greater than that predicted by the model. Statisticians agree to the definition of this term, but there is no general agreement on the precise interpretation of this quantity (see for example, http://davidakenny.net/cm/fit.htm). Here, I used 1, that is, residual deviance equals its residual degrees of freedom, as the cutoff point to indicate that the variability in the data set has been modeled properly. However, this can easily be challenged. Using this criterion, the GLMM overall represents the data set satisfactorily. Equivalently, GLM produced an R-Sq ranging from 0.41 to 0.79 on the year by year basis. Again, this could be interpreted as a satisfactory model.

Efforts for both GLMM and GLM were estimated using nested models with year as a variable. A word of caution is in place here. Both models are very computer extensive and require a relatively high computing power due to a large number of parameters in the models and also the volume of the data set generated when the data for the years 1984 through 2001 are combined (the number of significant parameters in GLMM and GLM were 78 and 108 respectively).

Further review of the estimates given in Table 11 showed that there was no significant difference between efforts generated using Table 2 or equation (4). That is, revising the *fathomzone*s above 12 using either Table 2 or equation (4) produce the same or equivalent results.

A few possibilities for the models as well as the mathematical relations were considered in this paper. Examples for the relations were not intended to generate efforts, but to show how these relations can be used in case, for example, the experimenter wishes to modify the model by adding arbitrary terms to it or to handle other issues such as the outliers in the data file.

The heteroscedasticity issue is a challenging one and most data analysts do not check the data sets in that respect. Like to normality assumptions, it is usually taken for granted and data sets are assumed homoscedastic. However, most data sets are heteroscedastic to some degree and need to be checked and proper adjustments (if needed) be made before any analysis is performed. The issue becomes more challenging when dealing with large data sets as the test statistics tend to increase along with the sample size. Once the heteroscedasticity was detected, the next step is to develop a proper weight function to correct the heteroscedasticity, which in turn presents a challenge.

It was concluded that a GLMM with first order of categorical variables as the fixed effect portion and first order and pairwise interactions of continuous variables or a GLM with the same categorical and continuous terms present this data file adequately with a slight edge given to the GLM (based on my interpretation of the dispersion parameter). However, one must always be cautious not to claim that the proposed models are "perfect." A quote from a well-known statistician [18] seems appropriate here. "*Essentially, all models are wrong, but some are useful.*"

Although, the application was limited to the shrimp data in this article, it could be easily applied to other data sets. Heteroscedasticity is common among most data sets and this paper should be helpful in other areas of research.

(1) References to any software packages throughout this article do not imply the endorsement of the said products.

## 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

[1] Nance, J., Keithly, W., Caillouet, C., Cole, J., Gaidry, W., Gallaway, B., Griffin, W., Hart, R., Travis, M. (2008). Estimation of Effort, Maximum Sustainable Yield, and Maximum Economic Yield in the Shrimp Fishery of the Gulf of Mexico, NOAA Technical Memorandum NMFS-SEFSC-570.

[2] Agresti, A., Booth, J. G., Hobart, J. P., Caffo, B. (2000). Random-effects modeling of categorical response data, Sociological Methodology 30, 27–80.

[3] Laird, N. M., Ware, J. H. (1982). Random-effects models for longitudinal data, Biometrics 38, 963–974.

[4] Fahrmeir, L., Tutz, G. T. (2001). Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd Edition, Springer-Verlag, New York.

[5] McCulloch, C.E., Searle, S.R. (2001). Generalized, Linear, and Mixed Models, Wiley, New York.

[6] McCullagh, P., Nelder, J. A. (1989). Generalized Linear Models, 2nd ed. Chapman & Hall/CRCPress.

[7] Nelder, J.A., Wedderburn, R. W. M. (1972). Generalized linear models, Journal of the Royal Statistical Society, Series A 135, 370–384.

[8] Nelder, J. A., Verrall, R. (1997). Credibility Theory and Generalized Linear Models, ASTIN Bulletin 27(1), 71-82.

[9] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., Schabenberger O. (2006). SAS for Mixed Models, 2nd ed.,

Cary, NC: SAS Institute, 2006.

[10] Griffin, W., Shah, A. K., Nance, J. M. (1997). Estimation of Standardized Effort in the Heterogeneous Gulf of Mexico Shrimp Fleet, Marine Fisheries Review, 59(3), 23-33.

[11] 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 Application, 6(3), 96-104.

[12] Hart R. A., Nance, J. M. (2013). Three Decades of U.S. Gulf of Mexico White Shrimp, Litopenaeus setiferus, Commercial Catch Statistics, Marine Fisheries Review, 75 (4), 43-47.

[13] Patella, F. (1975). Water surface area within statistical subareas used in reporting Gulf Coast Shrimp Data. Marine Fisheries Review, 37(12), 22–24.

[14] Marzjarani, M. (2010). Logarithmic Transformation of Raw Data, International Journal of Science, Technology and Management, Vol. 2, No. 3-4, Oct. 2009-Mar. 2010, 37-42.

[15] Breusch, T. S., Pagan, A. R. (1979). A Simple Test for Heteroskedasticity and Random Coefficient Variation. Econometrica. 47 (5), 1287–1294. JSTOR 1911963. MR 545960.

[16] White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica. 48 (4): 817–838. JSTOR 1912934. MR 575027. G. N., Robustness in Statistics, Academic Press, 201–236.

[17] Waldman, D. M. (1983). A note on algebraic equivalence of White's test and a variation of the Godfrey/Breusch-Pagan test for heteroscedasticity. Economics Letters, 13 (2–3), 197.

[18] Box, G. E. P. (1979). Robustness in the strategy of scientific model building, in Launer, R. L.; Wilkinson.