International Journal of Statistics and Probability; Vol. 9, No. 3; May 2020
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

A Comparison of a General Linear Model and the Ratio Estimator

Morteza Marzjarani

Correspondence: National Marine Fisheries Service, Southeast Fisheries Science Center, Galveston Laboratory, 4700AvenueU, Galveston, Texas 77551, USA

Received: March 8, 2020 Accepted: April 14, 2020 Online Published: April 15, 2020

doi:10.5539/ijsp.v9n3p54 URL: https://doi.org/10.5539/ijsp.v9n3p54

Abstract

In data analysis, selecting a proper statistical model is a challenging issue. Upon the selection, there are other important factors impacting the results. In this article, two statistical models, a General Linear Model (GLM) and the Ratio Estimator will be compared. Where applicable, some issues such as heteroscedasticity, outliers, etc. and the role they play in data analysis will be studied. For reducing the severity of heteroscedasticity, Weighted Least Square (WLS), Generalized Least Square (GLS), and Feasible Generalized Least Square (FGLS) will be deployed. Also, a revised version of FGLS is introduced. Since these issues are data dependent, shrimp effort data collected in the Gulf of Mexico for the years 2005 through 2018 will be used and it is shown that the revised FGLS reduces the impact of heteroscedasticity significantly compared to that of FGLS. The data sets will also be checked for the outliers and corrections are made (where applicable). It is concluded that these issues play a significant role in data analysis and must be taken seriously. Further, the two statistical models, that is, the GLM and the Ratio Estimator are compared.

Keywords: heteroscedasticity, outliers, model comparison

1. Introduction and Background

Selecting a model which satisfactorily represents a given data set is very challenging. Despite of advances in science, we are still far away from the point we can claim that we have found a “Perfect” model. Meanwhile, several statisticians have proposed criteria for selecting a model out of many choices. Each criterion suffers in one way or another. For example, AIC has no upper or lower bound. Adjusted R-Squared has an upper limit, but a good value for it does not necessarily mean that the model is good and goodness of fit is satisfied. As a result, at this time we heavily depend on personal judgment and preferences play a key role in the process. Selecting a proper model is definitely not the only concern. There are other issues which impact the analysis upon the selection of the model. To name a few, collinearity, outliers, acceptable range for the predictor (s), and heteroscedasticity are among those to mention here. If the purpose of selecting a model is prediction (which is usually the case), these issues play a role in the outcome. The interesting question here is this: How important are these factors to your analysis and how much do you want to expand the scope of the problem?

Formally, in the equation $y=a+bx$, the ration estimator (RE) method assumes that the parameter a does not exists or it is negligible. Therefore it is absolutely necessary to check and make sure that the condition stated by Snedecor and Cochran (1967) is satisfied before the method is applied to a data set.

Let’s look at some examples to show the significance of the condition stated above. In the first example we see two data sets where the RE is suitable for one but not the other. In the second example, we see a comparison of the RE and a simple linear regression applied to a data set where theoretically either model is suitable and we will see how these two methods behave under ideal situations.

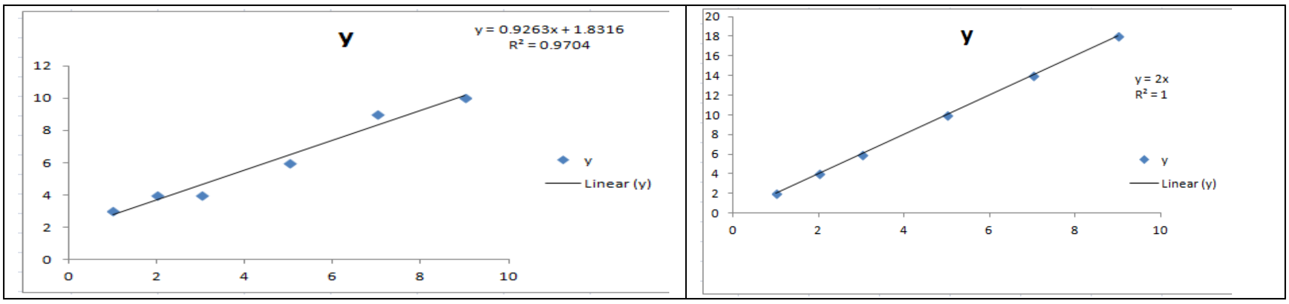
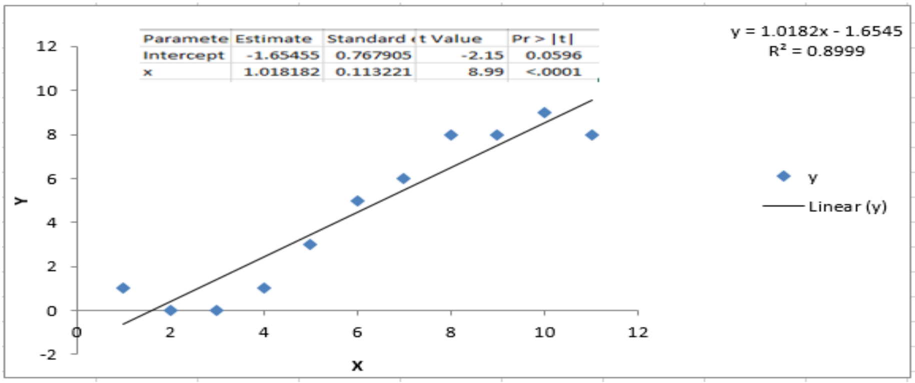


Figure 1. Fitted regression lines

Figure 2 displays the plot of the second data set with a high R-Sq. value.



Paramete	Estimate	Standard ‹	t Value	Pr > |t|
Intercept	-1.65455	0.767905	-2.15	0.0596
x	1.018182	0.113221	8.99	<.0001

Figure 2. Fitted regression line

Figure 2 displays a data set where either a ratio estimator (RE) or a linear regression could be used. It can be seen that the results don’t look the same even under this ideal condition. These differences will be much higher when applied to a very large data file. *Karl Pearson said in 1897 that the ratio estimates are biased and cautioned against their use.” (Wikipedia).*

Ratio estimates are biased. The order of bias is O (1/n) meaning that as the sample size increases, the method become more robust (unbiased).

2. Data Files

The major data contributors to effort estimation modeling are the following files: Shrimp (Analyst) data files in the GOM and two additional files. The shrimp data files included several fields of interest to this research. The fields of interest here are the same as those listed in Marzjarani (2019). Beginning with the year 2015, some records were recorded as 5555 in the *priceppnd* field. It was assumed that this value was a code and was converted using the ratio of the *value* and *pounds* fields (all these records had non-zero values in the *value* and *pounds* fields). Overall, due to the low percentages, the number of missing data points in the fields of interest to this research were not of concern (nevertheless, these missing values must and were addressed and handled here).

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 the designated port- the concatenation of these three was called *edate*
Daysfished (effort)	Actual hours of fishing per trip (24 hours per day)
pounds	Pounds of shrimp harvested
priceppnd	Average real price per pound of shrimp in the year data was collected
shore	1=offshore, 2=inshore

Table 2. Total number of records and total catch in the Analyst files

Year	Total offshore records	Total offshore catch	No. Missing or 0 in the *pounds* field	No. Missing or 0 in the *priceppnd* field
2005	74,202	86,622,326	1	20
2006	70,522	120,348,871	1	12
2007	65,709	83,765,251	0	14
2008	53,808	74,846,032	0	63
2009	56,527	100,844,994	1	6
2010	50,061	68,516,663	0	9
2011	55,649	87,025,166	1	17
2012	54,292	85,794,811	2	1
2013	52,243	77,580,822	1	4
2014	46,805	69,939,290	0	78
2015	42,759	67,819,919	0	4 missing, 5 with code 5555
2016	46,675	71,258,910	0	4 “, 13 “
2017	49,083	89,590,005	0	12 “ 26 “
2018	43,253	83,701,512	0	16 “ 23 “

The U.S. Gulf of Mexico is divided into 21 statistical subareas. Each statistical subarea is further divided into five-fathom depth increments. The 21 statistical subareas are placed into four areas 1 through 4, and twelve-fathomzones are placed into three depths 1 through 3. For further details on these, see Marzjarani (2019).

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

3. Method

3.1 Model Selection

The proposed model was a General Linear Model (GLM). The acronym GLM is used to represent either a general linear model with normally distributed dependent variable or a generalized linear model which is an extension of it where the dependent variable is allowed to have different distributions. Here, the first definition was deployed since the dependent variable was a log-transformed making this variable normal or normally distributed (approximately). The model considered here was in the form:

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

where *length*, *lands,* and *wavgppnd* are continuous variables, *area*, *depth*, and *tri* are categorical variables with 4, 3, and 3 levels respectively, and the second order terms denote the pair wise interactions between continuous variables. In the above the prefix *ln* stands for the natural logarithm. In literature, the model is generally written in matrix form:

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

3.2 The Process of Testing and Selecting the Model Covariates

For the purpose of demonstration, the shrimp data from 2005 through 2018 were included in this research. As expected, several issues had to be addressed. To name a few, the issue of handling missing data points, how to adjust the unusual values (influential, leverage, and outliers), etc. Several methods have been developed which help to select a covariate from the list of potential candidates. Here, a backward elimination approach was deployed. The process started with the "Full" model and the most promising ones were selected based on the minimum AIC (SBC), BIC, or the optimum Adj. R-Sq. In sum, all three criteria agreed with the full model for the years 2005, 2007, 2015, and 2017. In a few cases, up to two covariates were dropped. The ideal case is where all criteria agree to the full model. That is, the model is perfect. This is very hard to achieve especially if the number of covariates is large. Up to know, statisticians have not been able to find a way to locate the “perfect” model for representing a given data set. In the words of George Box (1976):“*Essentially, all models are wrong, but some are useful*.” He further comments : “*Since all models are wrong, the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad*”

In the next step, two additional files called Match and Trips were generated using the data files mentioned (Marzjarani, 2019).

3.3 Collinearity

Collinearity plays a significant role in data analysis. When building a model, one must make sure that the covariates used are not correlated with each other. It is almost impossible to find a list of variables which are totally independent. Therefore, in most cases, there would be some degree of compromise involved. In this research the issue of collinearity was examined through the Condition Index, Eigenvalue, and Variance Inflation Factor, VIF (or its inverse Tolerance). There is no doubt that interactions were correlated with the corresponding main effects. Other than that, it seemed that collinearity was not of major concern. To remove collinearity, one must carefully review the list of covariates and remove those causing this issue. In this research, this issue was not further addressed.

3.4 Heteroscedasticity

Heteroscedasticity is also a very important topic in data analysis where it refers to the circumstance in which the variability of a variable is unequal across the range of values of the variables that predict it. Its definition, its cause, and its impact on the analysis have been addressed in literatures in depth (See Marzjarani, 2019 for example). Among the impacts of heteroscedasticity we can name a few here. Parameter estimates are no longer BLUE, statistical tests of significance where it is assumed that the error terms are uncorrelated and uniform and therefore their variances do not vary with the effects might be invalid.

The difficulty is not in finding whether or not a given data set is heteroscedastic (which in almost all cases is). But, rather the challenging issue is how to handle it. That is, how to remove or at the least reduce it severity. Several methods including Weighted Least Square (WLS), Generalized Least Square (GLS), or Feasible GLS (FGLS) have been developed. Marzjarani (2018[b, d,] and 2019) has addressed some of these techniques. Each method involves finding a weight for reducing heteroscedasticity. Unfortunately, heteroscedasticity goes along with the data. The FGLS and WLS methods were deployed to determine the weight for the model as follows:

$$\begin{gathered}
\underline{res} = |\underline{y} - \underline{x\hat{\beta}}|, \\
\underline{res} = \underline{x\gamma} + \underline{\tau}, \\
\underline{weight_1} = 1/\,(\underline{x\hat{\gamma}}). \\
\underline{lnressq} = ln\,(\underline{y} - \underline{x\hat{\beta}})^2, \\
\underline{lnressq} = \underline{x\gamma} + \underline{\tau}, \\
\underline{weight} = 1/exp(\underline{x\hat{\gamma}})
\end{gathered} \tag{3}$$

In these formulas, $\underline{res}$ and $\underline{lnressq}$ are vectors of the residuals and the natural logarithm of the residuals squared respectively. A revised version of the FGLS was considered by defining the weight through the following:

$$\underline{weight} = 1/exp(\underline{lnressq} - \underline{x\hat{\gamma}}) \tag{4}$$

The raw data from 2005 through 2018 were checked for heteroscedasticity using the two well-known statistical methods, Breusch-Pagan (B-P), White, and also the proposed revised version of the FGLS. It was observed that these data sets (like others) were heteroscedastic. Due to the existence of heteroscedasticity, the weights defined above were used and the impact of each method was reviewed.

In the case of 2018 data, both B-P and White’s tests showed that the level of heteroscedasticity was relatively low. Further, generally the application of the revised FGLS generated an almost a homoscedastic data set.

The following table shows the RMSE values after applying the model to the raw data, then applying the weights using the FGLS and its revised version introduced here. Generally, RMSE ranged from 0.36 to 0.64 in the first column, for the GFLS, the range was from 1.93 to 2.31 (excluding year as a covariate with a very high number), and from 0.17 to 0.32 when the revised version of FGLS was deployed.

Table 3. RMSE and percentage increase/decrease

Year	Raw data	FGLS	Revised FGLS w.r.t. raw data	FGLS %increase w.r.t. raw data	Revised FGLS %decrease w.r.t. raw data
2005	0.36	2.22	0.17	517	53
2006	0.53	2.11	0.25	298	53
2007	0.48	2.05	0.23	327	52
2008	0.48	2.06	0.24	329	50
2009	0.52	2.08	0.25	300	52
2010	0.50	2.08	0.25	316	50
2011	0.50	2.12	0.24	324	52
2012	0.48	2.20	0.22	358	54
2013	0.64	1.93	0.32	202	50
2014	0.57	2.23	0.25	291	56
2015	0.54	2.31	0.24	328	56
2016	0.62	2.24	0.27	261	56
2017	0.61	2.13	0.29	249	52
2018	0.58	2.18	0.27	276	53
Year as covariate	0.57	Very high	0.27	Very high	53

It can be concluded that if heteroscedasticity is to be handled, the FGLS method is not helpful (especially when applied to the model with year as a covariate). The revised version is a much preferred method for reducing the impact of heteroscedasticity. As mentioned earlier, heteroscedasticity is highly data dependent and the major issue in handling it is to find the right weight which works for a given data set. In general, finding the right weight for improving heteroscedasticity is an open-ended issue at this time.

3.5 Residual Analysis

An important investigative process in model development/goodness of fit/selection is the review of the residuals. In what follows, we will study the residuals generated using the raw data from 2005 through 2017 and also all years combined (that is, year as a covariate).

3.6 Outliers

Outliers also play an important role in data analysis. There are a few issues of concern when dealing with outliers. First, we must establish an acceptable method for identifying the outliers. Second, we must investigate the source generating these data points. This is not always an easy issue to address. An outlier could have been the result of misreading/misreporting information due to the human errors. It could have been the fault of the software generating such points. Or, an outlier simply could be a valid data point. Third, if it turns out that an outlier is not a valid value, then the next question would be what to do with it. Some authors suggest the removal of an outlier from the data. This is a very risky action since it could impact the results significantly. So, removing outliers is not a good idea. Some authors suggest replacing outliers with the average of existing data points. Perhaps the most commonly used method is to flag any data point outside the interval (Q_1-x*IQR, Q_3+x*IQR) as an outlier where x can take a value between 1 and 3. In this interval, Q_1 and Q_3 are the first and the third quartiles and IQR is the interquartile range. Clearly, neither of the ideas presented for identifying/handling the outliers is perfect. As Cox states "*There are no routine statistical questions*; *only questionable statistical routines.*" The best approach is to absolutely make sure that in the case the outliers are valid and if so, check to find out the source generating the data for possible bugs. In the absence of any valid argument, we must assume that every data point is valid. In this article, outliers were handled using the approach of subtracting/adding 1.5 times IQR from/to the first and third quartiles mentioned above.

3.7 Model Predictability Feature

To measure the model predictability feature, some simulations were performed as follows:

Simulation I: This simulation was performed on the Match files. The simulation began with the initial sample size of 50 records, incrementing it by 50 each time up to and including 4000 records or the maximum number of available records (whichever was higher). Comparisons were made between the actual towdays and the corresponding predicted values by the model using the raw data. The analyses showed that there were no significant differences between the actual and predicted values. The conclusion was that the model performs satisfactorily. When plotting the actual versus predicted values, the ideal situation is where all the points on the graph are on the line *y*= *x*. That is, the model is perfect. However, we are still far away to claim that a perfect model have been located. To some degree of compromise, it seemed this condition was satisfied in 2005 and 2018 better than it did in 2017. But, all looked satisfactory.

Simulation II: Since the Analyst files were much larger than the corresponding Match files, an alternative simulation was performed by generating random samples (without replacement) from the Analyst files and checking the total of the actual input to the model (*lntd*) and the total of the output from the model (predicted values). This provided with the opportunity of using larger sample sizes. The initial and final values for the sample size were 500 and 40,000 with the increment of 500. No significant differences were observed between the corresponding actual and predicted *towdays*. This was consistent with what was expected when OLS was used to estimate the parameters of the model. It was not possible to apply this simulation to the case where *year* was included in the model as a covariate since the Match file was created after the trips were formed. The conclusion was that the model performed satisfactorily.

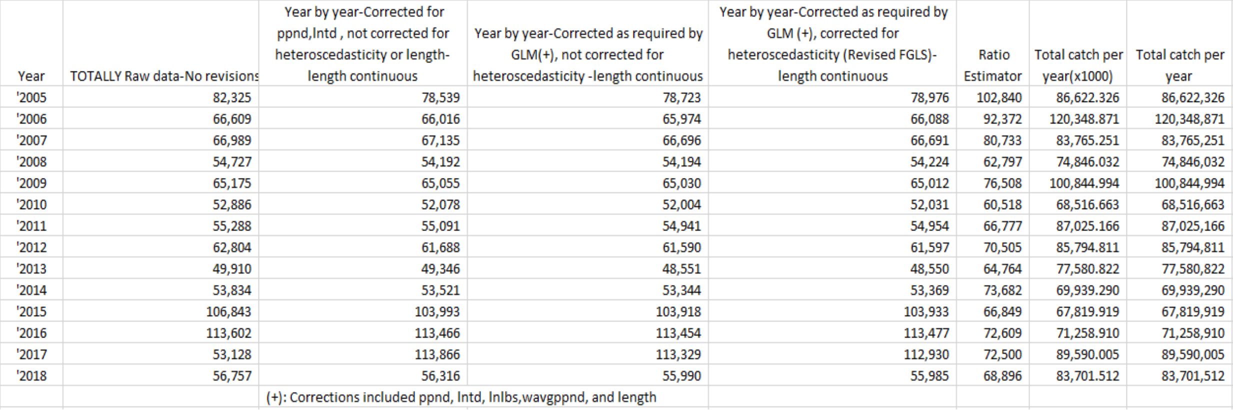
It was noticed that the corresponding actual and predicted values looked almost the same and it should be the case as expected. However, in practice, they were slightly different. But, the difference was very small (in some cases up to 10 zeros to the right of the decimal points).

Similar to the discussion presented at the end of simulation I, the magnitudes of the residuals were not considered in this simulation. The simulation simply showed that how the model deviates from being perfect. Next, the results of comparing the actual (input to the model) and predicted (output from the model) values within each sample taken from the Analyst files were performed. Due to the length concern, this analysis was done on 2005, 2017, and 2018 data only. Further, the simulation was performed on a limited basis starting with a sample size 2,000 up to including 40,000 with an increment of 2,000. Again, the plots of the actual versus predicted values looked satisfactory. The 2005 data points were closer to the line *y* = *x*. As mentioned earlier, the model predictability was satisfactory. The points on each plot were close to the line *y*=*x*. That is, the model represented the data fairly well.

3.8 Ratio Estimator Efforts and Its Comparison to the GLM Estimates

In this section, the model predictability was examined through a comparison with the Ratio Estimator. For such a comparison, several scenarios were considered (years 2005 through 2017 and year as a covariate). The results for some are listed in the table and figure below.

Table 4. Effort estimates using different scenarios



Year	TOTALLY Raw data-No revisions	Year by year-Corrected for ppnd,lntd , not corrected for heteroscedasticity or length-length continuous	Year by year-Corrected as required by GLM(+), not corrected for heteroscedasticity -length continuous	Year by year-Corrected as required by GLM (+), corrected for heteroscedasticity (Revised FGLS)-length continuous	Ratio Estimator	Total catch per year(x1000)	Total catch per year
'2005	82,325	78,539	78,723	78,976	102,840	86,622.326	86,622,326
'2006	66,609	66,016	65,974	66,088	92,372	120,348.871	120,348,871
'2007	66,989	67,135	66,696	66,691	80,733	83,765.251	83,765,251
'2008	54,727	54,192	54,194	54,224	62,797	74,846.032	74,846,032
'2009	65,175	65,055	65,030	65,012	76,508	100,844.994	100,844,994
'2010	52,886	52,078	52,004	52,031	60,518	68,516.663	68,516,663
'2011	55,288	55,091	54,941	54,954	66,777	87,025.166	87,025,166
'2012	62,804	61,688	61,590	61,597	70,505	85,794.811	85,794,811
'2013	49,910	49,346	48,551	48,550	64,764	77,580.822	77,580,822
'2014	53,834	53,521	53,344	53,369	73,682	69,939.290	69,939,290
'2015	106,843	103,993	103,918	103,933	66,849	67,819.919	67,819,919
'2016	113,602	113,466	113,454	113,477	72,609	71,258.910	71,258,910
'2017	53,128	113,866	113,329	112,930	72,500	89,590.005	89,590,005
'2018	56,757	56,316	55,990	55,985	68,896	83,701.512	83,701,512
		(+): Corrections included ppnd, lntd, lnlbs,wavgppnd, and length					

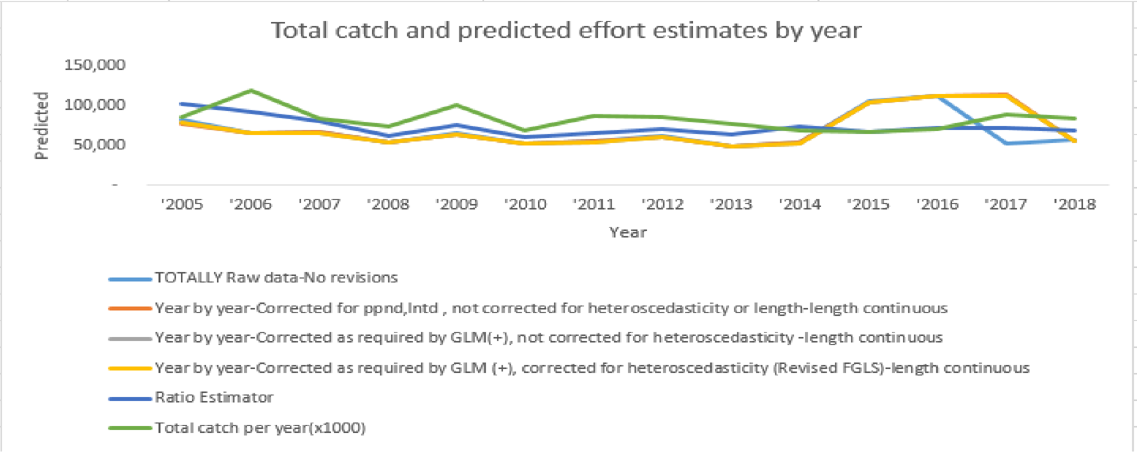


Figure 3. Effort estimates using different scenarios

Analysis showed that the 2018 data did not show much sensitivity towards heteroscedasticity and/or unusual data points. We will revisit this issue shortly.

Simulation III: Again, although obvious, the following simulation was run on the raw data and the predicted values generated by the Ratio Estimator and the GLM were compared. Due to the extensive calculations requiring significant effort and time, the simulation was limited to 60 times (beginning with the sample size 500 randomly taken from the Analyst files (without replacement) and incrementing it by 500 each time up to and including 33,000) applied to 2005, 2017, and 2018 data sets. As mentioned before, the reasons for selecting these files were as follows: First, these files are at the two ends of the list one being the oldest and the other two being the most recent. Second, analyses indicated that the 2005 and 2018 data were less prone to the outliers and/or heteroscedasticity than the 2017 data.

Generally, for the 2005, 2017, and 2018, the efforts generated by the RE were higher than those predicted by the proposed GLM. As we know, the ratio estimates in general increase as the sample size grows and obviously, they do not show any sensitivity with respected to the environmental changes (A smoother curve). Even with a limited simulation, the two models predicted differently.

Simulation IV: A similar simulation was performed comparing the predicted values by the Ratio Estimator and those generated by the GLM this time using the proposed revised version of the FGLS along with the corrections for outliers and the restrictions imposed by the GLM.

It was observed that the predicted values by the Ratio Estimator generally increased, but smoothly without regard to the changes in the covariate included in the study. The GLM on the other hand, showed sensitivity with respect to such changes. Statistically, the predicted values by the two models differed significantly.

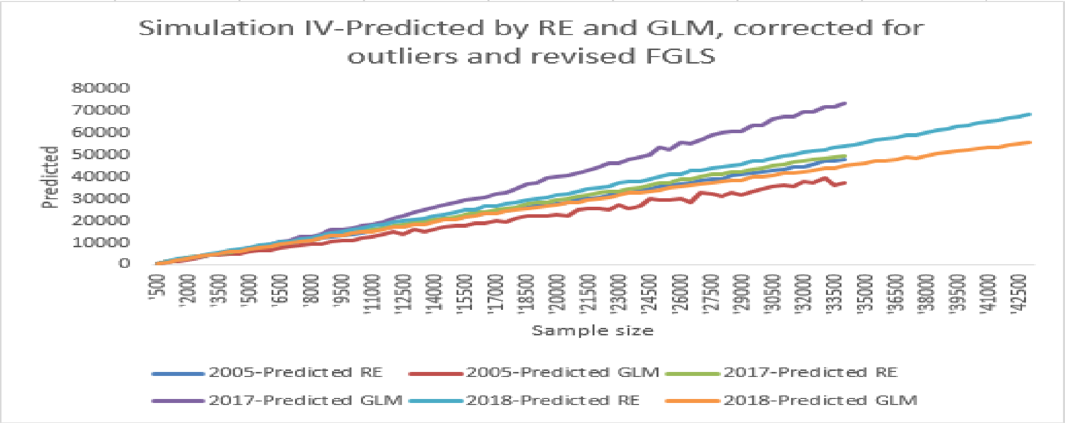


Figure 4. Predicted by Ratio Estimator and the Revised FGLS

3.9 Comparison of Simulations III and IV

Simulation III was performed on the raw data and IV was performed on the revised version of the data sets corrected for unusual data points and heteroscedasticity. In III, predicted values by the GLM were generally lower than those by the Ratio Estimator. In Simulation IV, the pattern for the 2005 and 2018 data remained the same (compared the Ratio

Estimator). However, the situation with respect to the 2017 data was reversed. This is an indication of either a high number of unusual data points and/or a high level of heteroscedasticity. According to Table 4, it seems the impact of outliers outweighed the impact of heteroscedasticity. The kurtosis coefficients for the 2005 and 2017 data were 1.8 and 5.4 respectively. That is, there were more outliers in 2017 than they were in 2005 and the plot of residuals was in favor of this claim. Further, the Breusch-Pagan test-statistic for the 2005 was 111.9 and the corresponding value for the 2017 data was 122.8. That is, one could conclude that the heteroscedasticity level in 2017 was higher than that of 2005 and resulting in the parameter estimates not being BLUE (in the case of the raw data). According to the Breusch-Pagan test, the Revised FGLS almost removed heteroscedasticity, but the White's test still persisted on its existence in the said data set. One reason for this difference could have been the inclusion of non-linear heteroscedasticity by the White's test. Also, the RMSE values before and after corrections for heteroscedasticity were 0.36 and 0.17 for the 2005 data and 0.62 and 0.29 for the 2017 data respectively. These might provide a clue for the reverse situation observed in the 2017 data.

As for the 2018 data, the kurtosis coefficient was 5.3 which indicate the existence of outlier (s). The plot of the residuals also displayed at least one such point. Further, the B-P test statistic for this data was 63.18 which translate into a lower heteroscedasticity level compared to the 2005 or 2017 data. In short, very likely a different method should be used in identifying the unusual data points in the 2018 data set. The following ranking of the three data sets 2005, 2017, and 2018 in terms of heteroscedasticity and unusual data points might be of interest.

Heteroscedasticity: 2018, 2005, 2017 (low to high).

Unusual data points: 2005, 2018, 2017 (low number to high).

Of course, we must realize that the above rankings depend on the methods deployed in handling either issue. Roughly speaking, the two data sets 2005, and 2018 represented two examples where correcting for either unusual data points or heteroscedasticity may not have been critical and could have been disregarded. In the 2017 on the other hand, revising such data for either outliers or heteroscedasticity made a difference in the predicted values by the GLM. Shortly, we will revisit this issue by analyzing the data for only outliers and again only for heteroscedasticity.

As mentioned above, of interest would be to perform at the minimum, two additional simulations as follows:

Simulation V: Perform a similar simulation on the 2005, 2017, and 2018 data this time correct the outliers only. Figure below summarizes the results.

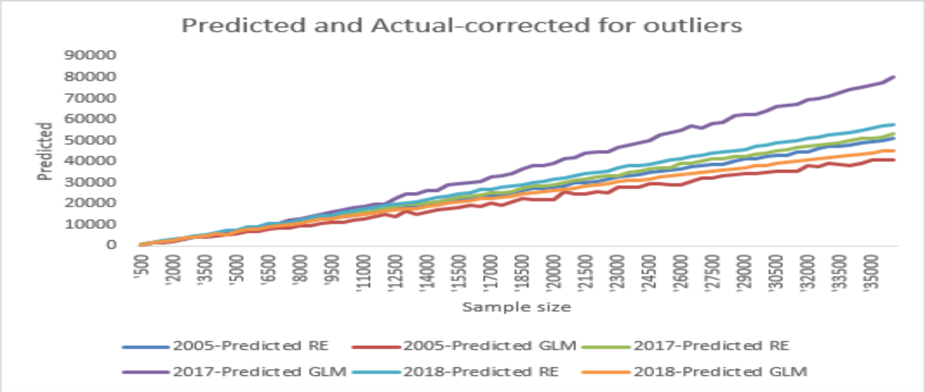


Figure 5. Predicted by Ratio Estimator and GLM data corrected for outliers

Simulation VI: A similar simulation was performed on these data sets, but this time the said data were corrected (relatively speaking) for heteroscedasticity using the revised version of FGLS.

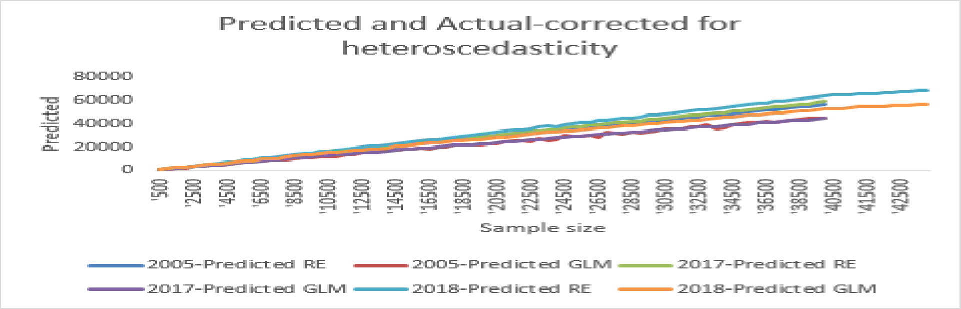


Figure 6. Predicted by Ratio Estimator and GLM, data corrected for heteroscedasticity

3.10 A Comparison of Simulations III, IV, V and VI

It would of interest to take look at the simulations III through VI and draw some conclusions. First, we will display the following figures for each year.

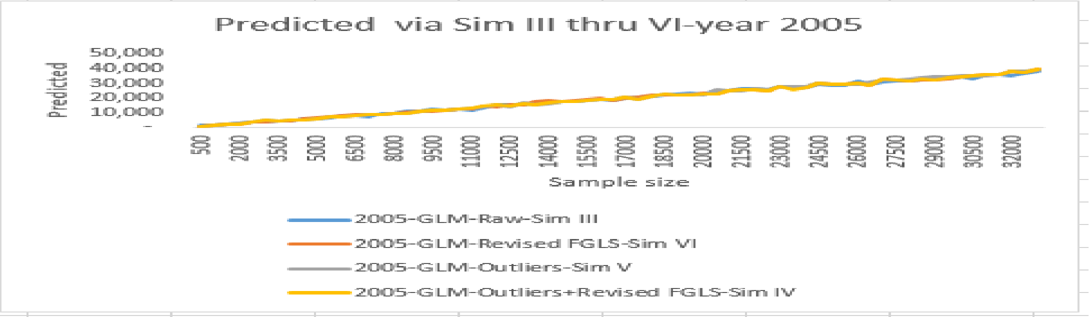


Figure 7a. Predicted GLM values considering different scenarios, year 2005

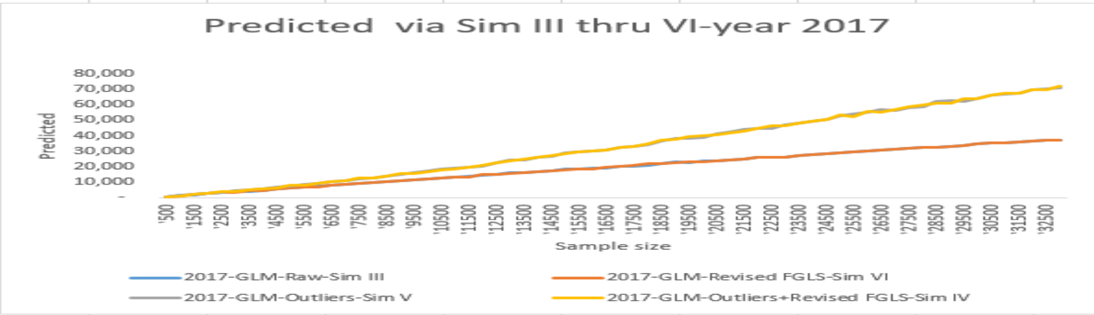


Figure 7b. Predicted GLM values considering different scenarios, year 2017

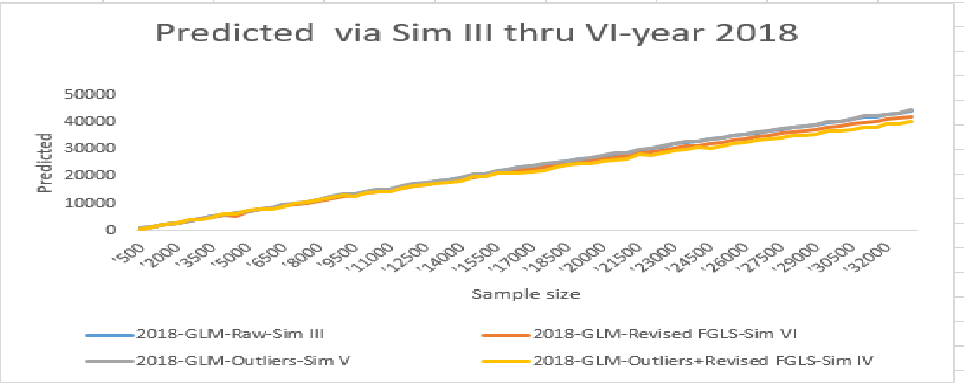


Figure 7c. Predicted GLM values considering different scenarios, year 2018

Through visual inspections of the above graphs and analyses, one can easily conclude that there are no statistically significant differences among the simulations under consideration here for the years 2005 and 2018. These two data sets (especially 2018) were well-behaved. Statistical analyses placed all four simulations in one category in each of 2005 and 2018 year. For the 2017 data on the other hand, the statistical analysis placed simulations III and VI in one category and IV and V in another category. Table 5 summarizes the results. In this table, simulations with no significant differences were assigned the same letters.

Table 5. Comparison of simulations III, IV, V, and VI

Simulation/Year	III	IV	V	VI
2005	A	A	A	A
2017	A	B	B	A
2018	A	A	A	A

We notice that III and VI do not involve the corrections for the outliers (more generally, unusual data points). In simulations IV and V on the other hand, the data were corrected for the outliers. In other words, as stated earlier, the outliers were mostly responsible for causing the divergence we observed in Figure 5.

4. Alternative Models

The other alternative models considered in the research were a generalized linear mixed model (GLMM) and a Bayesian generalized linear mixed model (BGLMM) both of the form:

$$y=X\beta+Z\Upsilon+\varepsilon, \tag{5}$$

where, the first and the second parts formed the fixed and random effects respectively. Using the default link functions, the GLMM, for example, generated a predicted value of 79,498 with the Pearson Chi Sq/DF = 0.13 for the 2005 data and 56,398 with Pearson Chi Sq/DF = 0.34 for the year 2018 using column 3 of Table 4. In the opinion of most statisticians a value close to 1 for the Pearson Chi Sq/DF is desirable.

5. Discussion/Conclusion

Statisticians are not in a complete agreement whether or not to include some non-significant parameters in the model (for estimation purposes) and what *p-value* should be used as the threshold to decide if something to be considered statistically significant. As suggested by Fisher, the *p-value* is generally compared to 0.05. But, the validity of the *p-value* has raised some concerns in recent years. Specifically where you are using multiple variables to predict an outcome through model building, Akaike Information Criteria (AIC) can take the place of the *p*-value, providing quantified information on what model is best (Quote from an article by Lewis G. Halsey, 2019).

It seems the versions of the model listed in columns 3 (assuming homoscedasticity) and 6 (reducing heteroscedasticity through the proposed revised FGLS) in Table 4 and Figure 3 are satisfactory. Nevertheless, the model showed sensitivity with the respect to the changes made. That is, one must investigate and take the best course of action. For example, it seems the model responded better when the data were corrected (relatively speaking) for heteroscedasticity using the proposed revised version of FGLS.

As displayed above, heteroscedasticity played a role in some years. But, the major concern about reducing the severity of it is the selection of proper weights. This is of course still an issue among the statisticians as it is the selection of a proper model.

There was a sharp decrease in 2017 compared to the other columns in Table 4 when the raw data were used with no revisions. Model diagnosis did not reveal any issues. The decrease was due to the corrections needed for the outliers or heteroscedasticity, most likely the outliers. Also, when forming the Match file, the width of the interval (*edate-x*, *edate+x*) and the Vessel file mentioned earlier contributed to the process significantly. Throughout these analyses, the interval (*edate-1*, *edate+1)* was deployed.

In regard to adding year to the model, such addition did not seem justified. The impact of year was highly significant and perhaps overshadowing the impacts of the other important covariates such as the vessel size. Intuitively, years are independent of efforts. That is, efforts vary from year to year mainly on the basis of environmental factors and not the year. If year is to be included in the model, a much better approach is time series and not a linear model. But, that raises other concerns. The concerns include but not limited to the way where important continuous and discrete environmental covariates and their interactions could be incorporated into such model. This of course, would make the process much more challenging. What happened in 2005 virtually has nothing to do with what is happening in 2019. There no convincing argument for adding this as a covariate to the model. This is also consistent with the approach taken by Griffin, et al. (1997) and Nance, et al. (2008) where in the first reference year was included in the model predicting CPT, but later it was removed when estimating the efforts (see equations 7 and 9).

Further, when adding years to the model as a covariate, one must consider issues such as a significant increase in heteroscedasticity, temperature and global warming, pollution, advances in technology, natural and human created events, and price inflation. Factors like these have direct impact on the fishery.

As displayed in Table 4, the model presented an increase in efforts during the most recent years (2015, 2016, and 2017). The plots of the residuals and the distribution of the error terms ε in the model presented earlier did not reveal any issues with the model. Further, in the case of 2018 data, there was a decrease in effort estimates.

In regard to issue of comparing the Ratio Estimator with the GLM or others, such comparison does not provide any useful information since the models are different and behave differently. Previous research results are also consistent with the above. See for example, Griffin, et al. (1997).

Through visual inspection, it was noticed that the 2005 and especially 2018 data did not show much sensitivity with respected to either unusual data points or heteroscedasticity. The 2017 data on the other hand showed somewhat more sensitivity with respect to the unusual data points than it did for heteroscedasticity. Further, in the 2017 data, the presence of a high number of unusual data points (that is, leverage, influential, or outliers) overshadowed the impact of heteroscedasticity. There might be a need to deploy a different method for handling the unusual data points in the 2018 data if more accurate results were needed.

No claim is made about the perfection of the model used in the research. Certainly, there are other models which could better fit the shrimp data. Therefore, the search for a better one must continue. A quote from Singer, J. D., Willett, J. B (2003) is an appropriate one for the conclusion of this research.

"We believe no statistical model is ever final; it is simply a placeholder until a better model is found."

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

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association, 76*, 791-799.

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.

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

Marzjarani, M. (2018a). 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. (2018b). 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. (2018c). 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.

Marzjarani, M. (2018d). Heteroscedasticity and Model Selection via Partitioning in Fisheries Data. *International Journal of Statistics and Probability, 7*(6). https://doi.org/10.5539/ijsp.v7n6p33

Marzjarani, M. (2019). Iterative Approaches to Handling Heteroscedasticity with Partially Known Error Variances. *International Journal of Probability and Statistics, 8*(2), 159-171. https://doi.org/10.5539/ijsp.v8n2p159

Nance, J., Keithly, W., Caillouet, C., Cole, J., Gaidry, W., Gallaway, B., … 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, 71 p.

Singer, J. D., & Willet, J. B. (2003). *Applied Longitudinal Data Analysis, Modeling Change and Event Occurrences*. Oxford University Press, New York.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical Methods*. Iowa State University Press.

Copyrights