Proving the Efficiency of Alternative Linear Regression Model Based on Mean Square Error (MSE) and Average Width using Aquaculture Data

Mohamad Arif Awang Nawi, Wan Muhamad Amir W Ahmad, Mohamad Shafiq Mohd Ibrahim, Mustafa Mamat, Mohd Fadhli Khamis, Mohamad Afendee Mohamed

ABSTRACT--- Multiple linear regressions (MLR) model is an important tool for investigating relationships between several response variables and some predictor variables. This method is very powerful and commonly used in finance, economic, medical, agriculture and many more. The main objective of this paper is to compare mean square error (MSE) and the average width between alternative linear regression models and linear regression model. The alternative method in this study is a combination of four methods, namely multiple linear regression method, the bootstrap method, a robust regression method and fuzzy regression through the construction of algorithms by using SAS software. Typically, the alternative method optimized by minimizing the mean square error (MSE) and average width. The results of the study showed a positive improvement for the estimation of parameters generated through these alternative methods.

Index Terms — Alternative linear regression, average width, mean square error.

I. INTRODUCTION

Multiple linear regressions (MLR) model is an important tool for investigating relationships between several response variables and some predictor variables. MLR modeling is a very powerful technique and commonly used in finance, economic, medical, agriculture and many more. The relationship is described as a model for estimating the dependent variable from independent variables. The multiple linear regression models are expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \mathrm{K} + \beta_n X_n + 1 \qquad (1)$$

where Y is the dependent variable, X is independently variable, β's are crisp parameters and X_n are the vector of crisp numbers. Usually, 1 are assumed to be independent random variables with a mean of 0 and variance σ^2. There are several weaknesses of linear regression methods [1]. For example, linear regression is sensitive to outlier and a huge effect on regression. Any kind of analysis, the sample size is very important for getting better results. A large sample size leads to increased precision in estimates of various properties of the population, though the results will become less accurate if there is a systematic error in the experiment, for example, mean square error. It will give effect to the estimation of parameters. Mean square error is the sum of squares divided by its corresponding degrees of freedom*: MSE = SSE/(n − p ^{0}) and MSR = SSR/p* [2]. It can be shown that these mean squares have the following expected values, average values in repeated sampling at the same observed X levels:

$$E\{MSE\} = \sigma^2,\ E\{MSR\} \geq \sigma^2$$

Note that when $\beta_1 = \mathrm{K}\ \beta_p = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise, $E\{MSR\} > E\{MSE\}$. A way of testing whether $\beta_1 = \mathrm{K}\ \beta_p = 0$, is by the F-test.

In [3] had written the local level model as a normal mixed effects model in order to use the restricted maximum likelihood estimator (RMLE). In [1] presented an alternative way to build the restricted likelihood function, also using mixed effects models. Another way to incorporate the uncertainty in the estimation of the parameters is through an asymptotic sampling of the maximum likelihood estimator (MLE) [4], which may be a poor approximation, especially for small samples. In [5], [6] states that bootstrap is a method for resampling the data based on random sorts of retrieval in the sample. In addition, this method also provides an estimate of the statistical distribution, the coverage probability of the confidence interval, and the probability of rejecting the hypothesis test that produces accurate results. The theoretical bootstrap model is as follows:

$$Y^* = X\hat{\beta} + u^* \qquad (2)$$

where u^* is a random term obtained from the residuals $\hat{u}$ of the initial regression at each iteration $b(b = 1,....,B)$, a sample $\{y_i^*\}_{i=1}^n$ of size (n, 1), is created from the theoretical bootstrap model. Since the OLS residuals are smaller than the errors they estimate, the random term of the theoretical

Revised Manuscript Received on July 10, 2019.

Mohamad Arif Awang Nawi, School of Dental Science, Universiti Sains Malaysia, Health Campus, Kubang Kerian, Kelantan, Malaysia.

Wan Muhamad Amir W Ahmad, School of Dental Science, Universiti Sains Malaysia, Health Campus, Kubang Kerian, Kelantan, Malaysia.

Mohamad Shafiq Mohd Ibrahim, Kulliyah of Dentistry, International Islamic University Malaysia, Kuantan Campus, Pahang, Malaysia.

Mustafa Mamat, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia.

Mohd Fadhli Khamis, School of Dental Science, Universiti Sains Malaysia, Health Campus, Kubang Kerian, Kelantan, Malaysia.

Mohamad Afendee Mohamed, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia.

bootstrap model is constructed from the following transform residuals which have the same norm as the error terms u_i:

$$\text{[illegible]} = \frac{\hat{u}_i}{\sqrt{(1-h_i)}} - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{u}_i}{\sqrt{(1-h_i)}}$$

The theoretical bootstrap model is hence expressed as:

$$y_i^*(b) = X_i\hat{\beta} + \text{[illegible]}(b), \quad \text{i= 1… n} \qquad (3)$$

where $\text{[illegible]}(b)$ is resample from [illegible].

The width of the bootstrap confidence intervals is closely related to the coverage probabilities - confidence intervals are wider for those methods where coverage probabilities are higher. The sub sampling bootstrap confidence interval with large resample size has the shortest intervals and the *m* out of *n* bootstrap with small resample size has the widest intervals. The width of all intervals increases as sample size decreases. In [2] proposed parametric and nonparametric bootstrap methods for estimating the PMSE of the state vector. Fuzzy regression method plays an important role in analyzing imprecise data. Mostly, a single independent fuzzy variable is used to analyze situation involving fuzzy data. Fuzzy Linear Regression (FLR) proposed for the first time by the Japanese researcher [7], provides the tools to study the problems that failing to the above-mentioned assumptions. A fuzzy linear regression model corresponding to multiple linear regression equations can be stated as:

$$y = A_0 + A_1x_1 + A_2x_2 + \text{K} + A_kx_k \qquad (4)$$

Previously, explanatory variables x_i's are said to be concise. However, according to (4), the response variable Y is fuzzy and not crisp and so are the parameters. It is the interest of this paper that is to estimate the values of these parameters. In coming discussion, A_i's are assumed to be symmetric fuzzy numbers which is representable by an interval. For instance, A_i can be written as $A_i = < a_{1c}, a_{1w} >$ with a_{ic} being the center, a_{iw} the reddish or vagueness associated.

Fuzzy set above reflects the confidence in the regression coefficients around a_{ic} in terms of the symmetric triangular membership function. When applying this method to fuzzy phenomenon, one should be aware that the response variable and thus the relation is also fuzzy.

This $A_i = < a_{1c}, a_{1w} >$ can be expressed as $A_1 = [a_{1L}, a_{1R}]$ with $a_{1L} = a_{1c} - a_{1w}$ and $a_{1R} = a_{1c} - a_{1w}$ [8]. In fuzzy linear regression methodology, parameters are estimated by minimizing total vagueness in the model.

$$y_j = A_0 + A_1x_{1j} + A_2x_{2j} + \text{K} + A_kx_{kj} \qquad (5)$$

Using $A_i = < a_{1c}, a_{1w} >$ we can write

$$y_j = < a_{0c}, a_{0w} > + < a_{1c}, a_{1w} > x_{1j} + \text{K} + < a_{nc}, a_{nw} > x_{nj}$$
$$= < a_{jc}, a_{jw} >$$

thus

$$y_{jc} = a_{0c} + a_{1c}x_{1j} + \text{L} + a_{nc}x_{nj}$$

$$y_{jw} = a_{0w} + a_{1w}|x_{1j}| + \Lambda + a_{nw}|x_{nj}|$$

where y_{jw} denotes the radius and thus having a positive value. From equation $y_{jw} = a_{0w} + a_{1w}|x_{1j}| + \text{L} + a_{nw}|x_{nj}|$, we shall use an absolute values of x_{ij}. Consider *m* data points, each comprising $a(n+1) - row$ vector. By minimizing the total vagueness of the model-data set combination, with condition that each data point belongs to the estimated value of the response variable, we can estimate the value for parameter A_i.

The main objective of this paper is to propose and compare mean square error and average width between alternative linear regression models and linear regression model. The secondary data was selected in this study based on zooplankton to investigate the factors contributing to the abundance of zooplankton in Malaysia. The alternative method in this study is a combination of four methods, namely multiple linear regression method, the bootstrap method, a robust regression method, and fuzzy regression. The combination of these methods is generated through the construction of algorithms through SAS software.

II. PROCEDURE METHODOLOGY OF DEVELOPING ALTERNATIVE ALGORITHM

A. This Algorithm is used to develop Multiple Linear Regression Using SAS software

```
/* First of all create multiple linear regression algorithm*/
proc reg data= for example, Data 1;
model y        = x1      x2;
run;
```

B. Approach the MM-Estimation Procedure for Robust Regression

```
/*Algorithm for Robust Regression by using MM-
estimation*/
ods graphics on;
procrobustreg method = MM fwls data= for example Data
1 plot=fitplot(nolimits)
plots=all;
model y = x1 x2 / diagnostics itprint;
output out=resids out=robout r=residual weight=weight
outlier=outlier sr=stdres;
run;
ods graphics off;
```

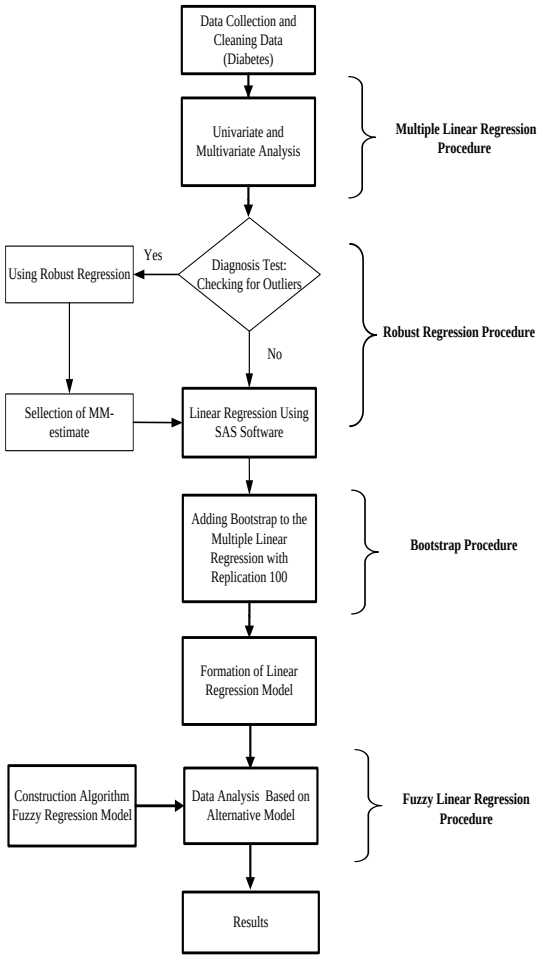

Fig. 1: Flow chart of alternative linear regression model

C. Next Step is Procedure for Bootstrap with Case Resampling

```
/* Next step we use a bootstrap with case resampling */
ods listing close;
procsurveyselect data= for example Data 1 out=boot1
method=urs
samprate=1outhits rep=100 ; /* Depending on the
researcher to do resampling data as possible */
run;
```

D. Finally, the procedure for Bootstrap into Fuzzy Regression Model (Alternative)

```
/*Combination of the Bootstrap algorithm with Fuzzy
Regression*/
ods listing close;
proc optmodel;
set j= 1..8;
number y{j}, x1{j}, x2{j};
read boot1 into [_n_] y x1 x2;
print y x1 x2;
number n init 8; /*Total of Observation*/

/*Decision Variable Bounded or Not Bounded*/
var vaw{1..3}>=0;/*bounded var*/
var vac{1..3};       /*not bounded var*/
/*Objective Function*/
min z1= vaw[1] * n + sum{i in j} x1[i] * vaw[2]+sum{i in
j} x2[i] * vaw[3];
/*Linear Constraints*/
con c{i in 1..n}:
vac[1]+x1[i]*vac[2]+x2[i]*vac[3] -vaw[1]-x1[i]*vaw[2]-
x2[i]*vaw[3]<=y[i];

con c1{i in 1..n}:
vac[1]+x1[i]*ac[2]+x2[i]*vac[3]
+vaw[1]+x1[i]*vaw[2]+x2[i]*vaw[3]>=y[i];

expand;
solve;
print vac vaw;
quit;
ods rtf close;
```

III. MEASUREMENT OF PARAMETERS

Zooplankton which is the second producer and primary consumer is a very small organism commonly located in aquatic ecosystem. It acts as the main energy transferor between primary producers and others from the same food chain. Zooplankton plays significant roles in influencing various aspects of aquatic ecosystem such as the food webs, cycling of energy and materials. At the same time, they have an undeniable role in natural and artificial fish nutrition. Zooplankton abundance and distribution are affected by environmental elements such as water temperature, the presence of nutrients and physicochemical factors [9]. As Zooplanktons respond immediately to environmental changes, their species composition is able to show any signs of pollution or any decline in the environmental qualification of ecosystems. Anzali international wetland provides various ecological values that are frequently overlooked. Physicochemical factors such as temperature, pH, Do and electrical conductivity form part of abiotic elements of an aquatic ecosystem [10]. As an example, intolerable levels of water temperature would limit the abundance of zooplanktons well as high pH levels may lead to the death of zooplankton, moreover, sensitivity to the low amount of dissolved oxygen would influence on zooplankton various life stages and different biological functions including feeding, growth, and reproduction. In this study, we used temperature, dissolved oxygen, and pH as the parameters. Water Temperature: It was determined using mercury-in-glass thermometer by dipping it into the water and allowed to stabilize for 5 seconds, removing and reading immediately recorded. pH: These were measured using pH/EC/TDS/Salinity meter by dipping the probes into the water until the screen showed a fixed reading as described by the manufacturers. Dissolved Oxygen (DO): It was determined using DO meter in which the probe was inserted into the water until DO reading in mg/l was recorded as described by the manufacturers.

IV. RESULTS AND DISCUSSION

In Table 1 show that the information about parameters Y, X^1, $X^{2,}$ and X^3.

Table 1: Description of the dependent and independent variable for zooplankton study

Variable in Zooplankton Data	
Zooplankton	Y = The number of zooplankton
Temperature	X_1 = Water temperature (C)
DO	X_2 = Dissolved oxygen (mg/l)
pH	X_3 = pH value of water

Based on the results of the multiple regression analysis, it was found that there are two variables that contribute to the abundance of zooplankton such as dissolved oxygen (β = 0.672; $p < 0.05$) and water pH value (β = -0.51; $p < 0.05$). Water temperature is not significant to the abundance of zooplankton.

Table 2: Factors contributing to the abundance of zooplankton in Malaysia by using multiple linear regressions

Independent Variable	DF	β	SE	p-Value
Temperature	1	-0.237	0.339	0.491
DO	1	0.672	0.208	0.004*
pH	1	-0.51	0.225	0.033*
R-Square	0.7924			
Adj R-Sq	0.7653			

*$p < 0.05$

Based on the results of the alternative linear regression model show that all of three variables that contribute to the abundance of zooplankton such as Water temperature (β = 0.399; $p < 0.0001$), dissolved oxygen (β = 0.480; $p < 0.0001$) and water pH value (β = -0.844; $p < 0.0001$). From the Adj R-Square obtained from the results, 96.65% of any changes in factors affecting the abundance of zooplankton can be explained by the three independent variables used in this regression such as water temperature, DO and water pH value. The Adj R-Square for this alternative model is much higher than the Adj R-Square for the linear regression model (76.53%).

Table 3: Factors contributing to the abundance of zooplankton in Malaysia by using alternative linear regression model approach (n = 100)

Independent Variable	DF	β	SE	p-Value
Temperature	1	0.399	0.011	.0001***
DO	1	0.480	0.007	.0001***
pH	1	-0.844	0.008	.0001***
R-Square	0.9665			
Adj R-Sq	0.9665			

***$p < 0.0001$

Table 4 shows the comparison of the study based on mean square error and average width. Based on the mean square error that produced by this alternative model of 0.0286. While the mean square error of the linear regression model of 0.2315. This shows that the mean square error generated from this alternative model is much more efficient than the linear regression model. In addition, the average width of an alternative model that involving 100 replication produces a small value rather than a linear regression model. Overall it can be concluded that the comparative study involving alternative model produced more efficient results compare to the linear regression model.

Table 4: Comparison between multiple linear regression model and alternative linear regression model (n = 100)

Model	Multiple Linear Regression Model	Alternative Linear Regression Model (n = 100)
Mean Square Error (MSE)	0.2315	0.0286
Average Width (AW)	67.548	0.475

V. CONCLUSION

In this study gives importance in terms of improvements to existing methods and preparing the application of parametric methods to the data of a study to be carried out more efficiently. This alternative method models are is very useful to be applied in various fields and shows a positive improvement. Computation by bootstrap method, robust regression, and fuzzy linear regression improve the efficiency of the results and can handle the problem of linear regression method. Typically, the alternative method optimized by minimizing the mean square error (MSE) and average width. The results of the study showed a positive improvement for the estimation of parameters generated through these alternative methods. This alternative method yields more efficient results compared to traditional methods of multiple linear regression. There are three special criteria for alternative methods. The first, the parameter estimates obtained very well with a minimum of sample data. In addition, this method can develop strong estimates based on standard errors and confidence intervals. Secondly, the problem of subtracting data can be addressed and resolved. Accordingly, this method can produce a robust estimator and generates a high breakdown point with high efficiency. The third privilege of an alternative method is algorithm constructed so detailed using optimization methods with emphasis on research data and consequently, the level of efficiency can be improved better. The resulting efficiency of the model is tested with significant value-p, the estimation parameters to obtain the average width interval and the mean square error value. The process of improving the existing method in this study can be seen in the positive changes in the results based on Tables 3 and 4. Based on the average width and mean square error it produces a small value. With this alternative method, it can contribute another research methodology for researchers in various fields to produce more efficient results.

REFERENCES

1. J. Tsimikas, and J. Ledolter, "REML and best linear unbiased prediction in state space models," Communications in Statistics: Theory and Methods, 23(8), 1994, pp. 2253-2268.

2. D. Pfeffermann, and R. Tiller, "A bootstrap approximation to prediction MSE for State Space models with estimated parameters," Journal of Time Series Analysis, 26, 2005, pp. 893–916.
3. L. A. Shepard, "Evaluating test validity," in Review of Research in Education, L. Darling-Hammond, Ed. Washington DC: American Educational Research Association 1993, pp. 405-450.
4. B. Queenville, and A. C. Singh, "Bayesian prediction means squared error for state-space models with estimated parameters," Journal of Time Series Analysis, 21(2), 2000, pp. 219-236.
5. B. Efron, and R. J. Tibshyrani, An Introduction to the Bootstrap. New York: Chapman and Hall, 1993.
6. P. Hall, The Bootstrap and Edgeworth Expansion. New York: Springer Verlag, 1992.
7. H. Tanaka, and H. Lee, "Interval regression analysis by quadratic programming approach," IEEE Transactions on Fuzzy Systems, 6(4), 1998, pp. 473–481.
8. J. Kacprzyk, and M. Fedrizzi, Fuzzy Regression Analysis. Warsaw: Omnitech Press, 1992.
9. U. Ahmed, S. Parveen, A. A. Khan, H. A. Kabir, H. R. A. Mola, and A. H. Ganai, "Zooplankton population in relation to physic-chemical factors of a sewage fed pond of Aligarh (UP), India," Biology and Medicine, 3, 2011, pp. 336-341.
10. P. S Verma, and V. K Agarwal, Environmental Biology: Principles of Ecology. New Delhi: S. Chand Publishing, 2007.