World Applied Sciences Journal 33 (7): 1234-1240, 2015
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2015.33.07.132

Combining of Response Surface Methodology (RSM), Bootstrap and Multiple Logistics Regression Method to Analyze an Aquaculture Data

[1]Wan Muhamad Amir W. Ahmad, [2]Nor Azlida Aleng,
[3]Nurfadhlina Abdul Halim, [4]Zalila Ali and [5]Syerrina Binti Zakaria

[1]School of Fisheries and Aquaculture Sciences,
University Malaysia Terengganu (UMT), Kuala Terengganu, Terengganu Malaysia
[2,3,5]School of Informatics and Applied Mathematics,
University Malaysia Terengganu (UMT), Kuala Terengganu, Terengganu Malaysia
[4]School of Mathematics Sciences, Universiti Sains Malaysia, 11800 Minden, Pulau Pinang, Malaysia

Abstract: Response Surface Methodology (RSM), Bootstrap and Multiple Logistics Regression (MLR) is a collection of statistical tool technique for analyzing data from various fields. Combining this idea is very useful for the modelling with an advanced analysis and perhaps can be an alternative method for modelling options in applied statistics scope. Combining these methods are capable to handle the case of small and limited sample size data. Using bootstrapping method, we are allowed to create a new sample based on sampling with replacement. In our case, the term "bootstrap" actually is referring to the use of the original data set to generate new ones. In this research paper, from a small and limited sample size data, we performed bootstrapping method in order to generate a new data set with a bigger sample size. After getting a new sample size, we then perform multiple logistic regression with embedding response surface methodology (RSM). This method can be used when the response the response variable, y is influenced by several variables, x's. Therefore, useful results and conclusions can be drawn from the analysis. We also provided some example of application of the method discussed by using SAS computer software.

Key words: Bootstrap · Multiple Linear Regression · Response Surface Methodology

INTRODUCTION

Logistic regression is a type of predictive models that can be used when the target variable is a categorical variable with two categories, for instance, live or die, has cancer or no cancer, coronary heart disease or not having coronary heart disease, the patient survives or dies and many more [1]. In logistic regression, the dependent variable is a binary or dichotomous; it only contains data coded as 1 or 0. The objective of logistic regression is to find the best fitting model to illustrate the relationship between the dichotomous characteristics of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Let us consider that, there are P is independent variables which will be denoted as vector $x' = \{x_1, x_2,...,x_n\}$. We assume that each of these variables is at least interval scaled. Let the conditional probability that the outcome is present denoted by [1]. Then the logit of the multiple regression is given by the formula as follows: logit $(\pi) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_px_p$. The specific form of the logistic regression model is given by:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{3}$$

where $\pi(x)$ is the probability of occurrence of the characteristic of interest [1]. Since the model produce by logistic regression is nonlinear, the equations used to describe the outcomes are slightly more complex than those multiple regression. This linear regression equation creates the logit transformation. This transformation is defined, in terms of $\pi(x)$ as:

Corresponding Author: Wan Muhamad Amir W. Ahmad, School of Fisheries and Aquaculture Sciences, University Malaysia Terengganu (UMT), Kuala Terengganu, Terengganu Malaysia.

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)=\sum_{j=1}^{p} b_j x_{ij} \quad (4)$$

The objective of this study is to discuss the improvement between the advanced multiple logistic regression with the original version of multiple logistic regression [2]. We extend the advanced of multiple logistic regressions with combining bootstrap and response surface methodology. According to Lockwood & Mackinnon in [3]; Oehlert, [4] and Wikipedia [5], the response surface methodology (RSM) was introduced by G.E.P. Box and K.B. Wilson in [6]. The response surface methodology (RSM) explores the relationships between several explanatory variables (*X*) and one or more response variables (*Y*). The main idea of RSM is to use a sequence of designed experiments to obtain an optimal response through linear model and second-degree polynomial. They acknowledge that this model is only an approximation, but use it because such a model is easy to estimate and apply, even when little is known about the process. According to Mead and Pike stated origin of RSM starts 1930s with use of *Response Curves* [7].

Bootstrap method is a statistical technique that falls under the broad heading of resampling. This method is very useful and can be used various especially in the estimation of nearly any statistics [8]. This procedure involves a relatively simple procedure, but repeated so many times depending on the need of the researcher. Bootstrap technique is heavily dependent upon computer calculation. Using the bootstrap method we are able to determine the estimating value of a parameter that presenting the whole of a population. Without using bootstrap method, the value of the parameter of a population is impossible to measure directly. So, we use statistical sampling method and we sample a population, measure a statistic of this sample and then use these statistics to say something about the corresponding parameter of the population [8].

Data and Methods: Data of this study is a sample which composed of five variables. Namely variables are as in Table 1. Multiple logistic regression technique was used in the analysis of relationship between variables. Data of 23 observations were collected on Karah Island.

Case Study I: Calculation for Logistic Regression Using Normal Procedure SAS:

```
Data Sampling;
input Temperature Salinity PH Plankton;
datalines;
0  0  0  0
0  0  0  0
0  0  0  0
1  1  1  0
1  1  1  0
1  1  1  0
0  1  0  0
0  1  0  1
0  1  0  1
0  1  0  1
1  0  1  1
1  0  1  1
1  0  1  1
1  1  0  1
1  1  0  1
1  1  0  1
0  0  1  1
0  0  1  1
0  0  1  1
0  0  1  1
;
ods rtf file='robdunc0.rtf' style=journal;
/* Logistic regression */
proc logistic data = Sampling;
model Plankton(event='1') = Temperature Salinity PH;
run;
/* Receiver Operating Characteristics Curve (ROC) */
ods graphics on;
proc logistic data=Sampling plots=EFFECT plots=ROC;
model Plankton(event='1')=Temperature Salinity PH;
output out=estimated predicted=estprob l=lower95
u=upper95;
run;
ods graphics off;
/* plots=(surface)*/
ods graphics on;
procrsreg data=Sampling plots=(surface);
model Plankton = Temperature Salinity;
run;
/* plots=(surface)*/
ods graphics on;
procrsreg data=Sampling plots=surface(3D);
model Plankton = Temperature Salinity;
run;
ods rtf close;
```

Table 1: Description of Data

Variables	Code	Explanation of user variables	Categorical
Type	y	Plankton	0 = Zooplankton and 1 = Phytoplankton
Temperature	x_1	Temperature Reading	29.31°C, 29.34°C and 29.57 coded as 0
			29.65°C, 29.66°C and 29.71°C coded as 1
Salinity	x_2	Salinity Reading	29.31, 29.21 and 29.70 coded as 0
			30.64, 30.54 and 31.69 coded as 1
PH	x_3	PH Reading	7.29, 7.42 and 7.39 coded as 0
			7.58, 7.69 and 7.75 coded as 1

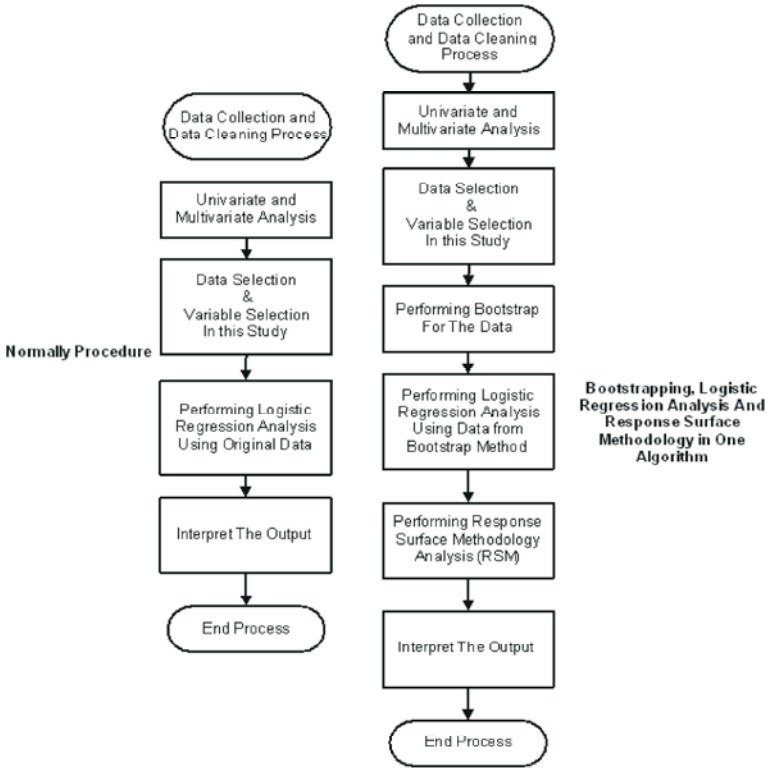

Fig. 1: Flow Chart of the Normal Procedure and Alternative Analysis

Output for the Case Study I: Figure 2. The counter and surface plots indicate that the highest value of response (Plankton) is obtained when the reading salinity, is low and the reading temperature is high. This area appears at the upper left corner of the plot.

Case Study II: 2.2 Calculation for Alternative Method for Logistic Regression Using SAS:

```
Data Sampling;
input Temperature Salinity PH Plankton;
datalines;
0 0 0 0
0 0 0 0
0 0 0 0
1 1 1 0
1 1 1 0
1 1 1 0
0 1 0 0
0 1 0 1
0 1 0 1
0 1 0 1
1 0 1 1
1 0 1 1
1 0 1 1
1 1 0 1
1 1 0 1
1 1 0 1
0 0 1 1
0 0 1 1
0 0 1 1
0 0 1 1
;
```

```
ods rtf file='robdunc0.rtf' style=journal;
/* Bootstrapping data with a case resampling */
procsurveyselect data=Sampling out=boot1
method=urssamprate=1 outhits rep=30;
run;
/* Logistic regression by using bootstrap data */
proc logistic data=boot1 outest=est1(drop=_:);
model Plankton(event='1')= Temperature Salinity PH;
run;
/* Receiver Operating Characteristics Curve (ROC) */
ods graphics on;
proc logistic data=boot1 plots=EFFECT plots=ROC;
model Plankton(event='1')=Temperature Salinity PH;
output out=estimated predicted=estprob l=lower95
u=upper95;
run;
/* plots=(surface)*/
ods graphics on;
procrsreg data=boot1 plots=(surface);
model Plankton = Temperature Salinity;
run;
/* plots=(surface)*/
ods graphics on;
procrsreg data=boot1 plots=surface(3D);
model Plankton = Temperature Salinity;
run;
ods rtf close;
```

Output for the Case Study II: Figure 3 shows the same finding as Figure 2. The counter and surface plots indicate that the highest value of response (Plankton) is obtained when the reading salinity, is low and the reading temperature is high. This area appears at the upper left corner of the plot. In addition, we can see the shape of the response surface and get a general idea of response plankton at various setting of reading salinity and reading temperature. We compared the gained results from the first case with the second case. With and without adding the bootstrapping method. For the case without using the bootstrap method, inadequate size was used in order to illustrate the importance of the bootstrap method. From the first case studied (Table 2), three variables were found with no significant and they are Temperature ($\beta_1 = 0.1944$, $se = 1.1447$, $p = 0.8651$), Salinity ($\beta_2 = -0.4143$ $se = 1.1805$, $p = 0.7256$) and PH ($\beta_3 = 0.2198$, $se = 1.1757$, $p = 0.8517$). With small sample size, three variables did not indicate the significant association to the dependent variables, using bootstrap method we replicate the data up to 50 replicates with the 1000 observations. By running the logistic regression with the alternative method using the 1000 observations we found that one out of three variables show a significant relationship to the response variables. They are Temperature ($\beta_1 = 0.3682$, $se = 0.2073$, $p = 0.0757$), Salinity ($\beta_2 = -0.6749$ $se = 0.2239$, $p = 0.0026$) and PH ($\beta_3 = -0.2929$, $se = 0.2182$, $p = 0.1795$).

Summary and Conclusion: This paper explained on how bootstrapping, method can be applied to the logistics regression. This method offered a preliminary general idea of the process that involving inadequate sample size and straightly solve the problem by bootstrapping the observations thus exceeding the minimal prerequisites of the sample size. In this paper two different methods have been used: (i) Common logistic regression model and (ii) bootstrapping logistic regression model. The first case study analyzed data using common methods and the second case study analyses using the bootstrapping method of enlarging the sample size. We can see clearly the different results from the both analyses. Result from the first case study shows that three variables were found with no significant and they are Temperature ($\beta_1 = 0.1944$, $se = 1.1447$, $p = 0.8651$), Salinity ($\beta_2 = -0.4143$, $se = 1.1805$, $p = 0.7256$) and PH ($\beta_3 = 0.2198$, $se = 1.1757$, $p = 0.8517$) but with replicating the observations we found that one variables show a significant relationship to the response variables. It is Salinity ($\beta_2 = -0.6749$, $se = 0.2239$, $p = 0.0026$) and for the Temperature ($\beta_1 = -0.2929$, $se = 0.2182$, $p = 0.0757$) and PH ($\beta_3 = -0.2929$, $se = 0.2182$, $p = 0.1795$) there no significant relationship toward dependent variables. We design the bootstrapping method of logistic regression in order to treat the modest sample size or inadequate sample. In the instance with a minuscule sample size, this proposed method is really potent to apply because it can sampling by the bootstrapping method until the sample size is decent. According to Erin and Edward [9] salinity and temperature appears to be the factor that correlates best with the shifts in many phytoplankton taxa in estuaries. Actually, it is not easier to understand the behaviour of the data in studies when it is not reaching the actual sample size needed in an analysis [10]. Some of the cases, when the sample size is too small the real factor will not disclose and some of the studied factors will not give the any significant results due to the related factors. In this paper, we approach the response surface method to the algorithm before and after bootstrapping method. This response surface method in reveals the finding with more explicitly due to the bootstrap performance in logistic regression analysis. It provides the comprehensive information and also give the general idea of dependent variables at the various setting of independent variables.

Table 2: Parameter Estimates Analysis of Maximum Likelihood Estimates

	Analysis of Maximum Likelihood Estimates				
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6402	0.9834	0.4238	0.5150
Temperature	1	0.1944	1.1447	0.0289	0.8651
Salinity	1	-0.4143	1.1805	0.1232	0.7256
PH	1	0.2198	1.1757	0.0350	0.8517

Table 3: Parameter Estimates Using Analysis of Maximum Likelihood Estimates Using Bootstrap Method with 50 Replicates (total sample size 1000)

	Analysis of Maximum Likelihood Estimates				
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.0010	0.1948	26.4097	< 0.0001
Temperature	1	0.3682	0.2073	3.1557	0.0757
Salinity	1	-0.6749	0.2239	9.0889	0.0026
PH	1	-0.2929	0.2182	1.8014	0.1795

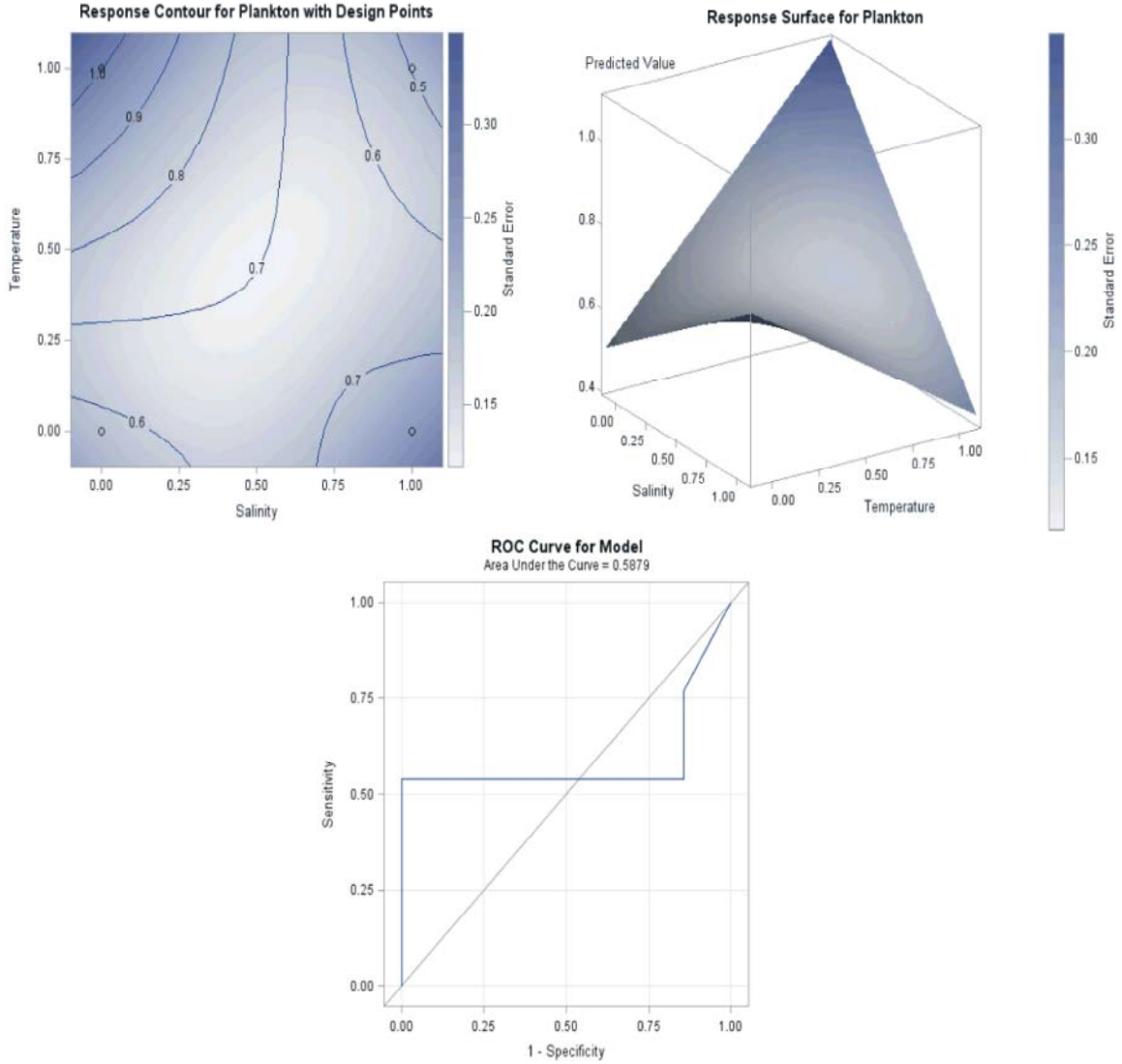

Fig. 2: Plot of Response Surface and ROC for Case Study I

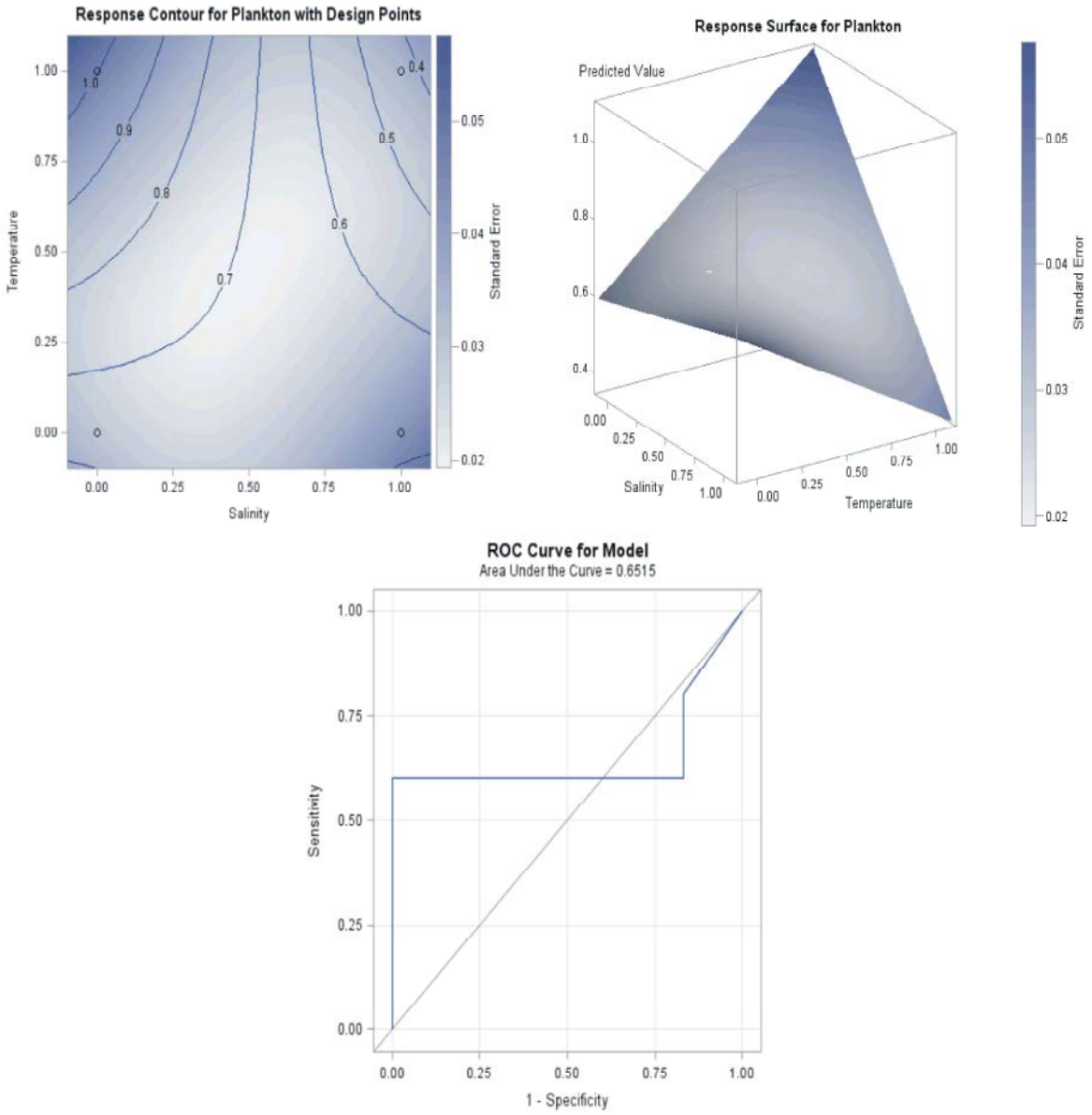

Fig. 3: Plot of Response Surface for Case Study II

REFERENCES

1. Amir, W.A., Nor Azlida Aleng and Zalila AIi, 2010. Binary Logistic Regression Analysis Technique Used in Analyzing the Categorical Data In Education Sciences: A Case Study of Terengganu State, Malaysia. World Applied Sciences Journal., 9(9): 1062-1066.
2. Hosmer, D.W. and S. Lemeshow, 2000. Applied logistic regression, second edition, John Wiley and Sons.
3. Lockwood, C.M. and D.P. Mackinnon, 1998. Bootstrapping the standard error off the mediated effect. Proceedings of the 23rd Annual Meeting of SAS users Group International (pp: 997-1002). Cary, NC: SAS Institute, Inc.
4. Oehlert, G.W., 2000. Design and analysis of experiments: Response surface design. New York: W.H. Freeman and Company.
5. Wikipedia(2014). Response surface methodology. http://en.wikipedia.org/wiki/Response_surface_methodology (Accessed Mei 17, 2014.

6. Box, G. E. P. and K.B. Wilson, 1951. On the Experimental Attainment of Optimum Conditions, Journal of the Royal Statistical Society, Series B, 13, 1-45.
7. Myers, R.H., A.I. Khuri and W.H. Carter, 1989. Response surface methodology: 1966-1988. Technometrics, 15: 301-317.
8. Cassel, D.L., 2010. Bootstrap Mania: Re sampling the SAS. SAS Global Forum 2010: Statistics and Data Analysis. Paper 268-2010: pp: 1-11.
9. Quinlan, E.L. and E.J. Phlips, 2007. Phytoplankton assemblages across the marine to low-salinity transition zone in a blackwater dominated estuary. Journal of Plankton Research, 29(5), 401-416.
10. Naing, N.N., 2003. Determination of sample size. Malaysian Journal of Medical, 10(2): 84.