Sapporo Medical Journal
Volume 55, Issue 10, October 2021

THE DEVELOPMENT OF REGRESSION MODELING THROUGH DESIGN OF EXPERIMENT (DOE) METHODOLOGY: AN OVERVIEWTHE EFFECT OF BOOTSTRAPPING DATA

Soban Qadir[1,2], Wan M.A.W. Ahmad[1], Nor A. Aleng[3]

[1]Unit of Biostatistics, School of Dental Sciences, UniversitiSains Malaysia (USM), Malaysia
[2]Lecturer of Biostatistics, Department of Dental Education, College of Dentistry, Imam Abdulrahman Bin Faisal University, Kingdom of Saudi Arabia
[3]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu (UMT), Malaysia

Abstract— The focus of the study was to evaluate the effect of the bootstrapping method on an applied sciences study. In this case study, a set of data from the design of an experiment (DOE) was taken which consisted of 20 observations, which is specifically design forTwo-way Analysis of Variance (ANOVA). After reorganizing the data according to the specific template, a regression approach to one-factorAnalysis of variance was used for the best model fitting which considering all the factor levels.At first, the regression was fit through the origin data. Second, the bootstrapped method was applied towards the origin data through SAS syntaxin order to obtained large sample data, then the obtained data was used for regression fitting.The regression results with and without bootstrap was compared according to the parameter of estimated coefficient beta, standard error and the obtained *p*-value.To validate the obtained model, a multilayer perceptron neural network was applied, the Mean Error (ME), Mean Absolute Error (MAE) and Maximum Error were compared towards training, testing and validation procedure. As a conclusion, the regression approach through bootstrapping to one-factor Analysis of variance was a significant methodology. The mean error (ME) and mean absolute error (MAE) of the measurement of the model accuracy are relatively similar to the training, testing, and validation. Small error measurement was showing the high accuracy of the developed model. Hence, it can be concluded that bootstrapping is a very useful tool to draw precise and reflective inference especially in case of small datasets.

Keywords: Design of experiment, bootstrapping, multilayer perceptron, Analysis of variance, multiple linear regression

1. Introduction:

Statistical significance testing is a vital corner of every research methodology, especially in case of medical researches. Acceptance or rejection of null hypothesis is associated with statistical significance of the performed test. In the case of studies with small sample sizes, chances become higher that true null hypothesis will be proved as false. Hence, studies with large enough sample size are recommended to obtained statistical significance and true reflection of calculated results towards actual population parameters. However, due to obvious reasons it is not always possible for researchers to get large sample size. Then question arises that what would be the solution of the problem, how strong inference and better estimation can be drawn by having small sample size.

Bootstrapping is a technique which can solve the problem of drawing strong inference with small data and get the estimated parameters which truly reflects the population.Bootstrapping is a simulation technique which helps to evaluate the standard error of the estimation of a parameter is called Bootstrapping [1]. It is a methodology which enables researchers to estimate statistics on a population by sampling datasets with replacement [2]. Bootstrap method is used to calculate pseudo population by using small sample andafter that statistical analysis can be performed over booted data which help to draw results and check the validity of null hypothesis. Chances become less to reject null hypothesis if it is true, in case of using bootstrapping over small datasets. Furthermore, multilayer perceptron neural network (MLPNN) is a method which can help to validate the drive model through regression analysis [3]. Simplest and commonly used MLPNN has an input layer, one hidden layer and one output layer andthe same structure was used in the current study. However, there are various

approaches and many possible structures to run MLPNN. In this method statistical software automatically calculate the weights and value of bias to run activation function. In case of simple MLPNN which includes one hidden layer and one output layer, two activation functions involved. First function runs before creating hiding layer and second before output layer. Software calculates weights and bias value for each function separately.

Design of experiment (DOE) data can only be used to test statistical significances and cannot be used for prediction purposes or constructing models directly. Transformation of DOE data is always required to use it for regression analysis. In the present study DOE data set used towards regression analysis. Let us consider the two factor ANOVA model as follows

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,b \\ k = 1,2,\ldots,n \end{cases}$$

where

Y_{ijk} = is the kth value of the response variable in the j^{th}trial for the i^{th} factor level or treatment

$\mu_{ij} =$ are parameters

ε_{ijk} = random error $\sim i.i.d\ N(0, \sigma^2)$

In order to present ANOVA model in terms of liner regression model, adjustment of the parameters is required which was explained in detailed in previously published article [4]. The derived linear regression model after transformation of DOE data was

$$Y_{ij} = \mu. + \tau_1 x_{ij1} + \tau x_{ij2} + \tau x_{ij3} + \varepsilon_{ij}$$

The derived regression model was used for parameter estimations with and without bootstrapping. Hence, the purpose of the study was to describe the procedure of transforming two-factor analysis of variance (ANOVA) into liner regression model and to evaluate the effectiveness of using bootstrapping method in order to draw more validated inference from regression model.Furthermore, multilayer perceptron neural network was used to study the strength of the derived model.

2. Material and Methodology:

2.1 DataSelection

A set of data was taken from an experiment which focused on the average of plasma calcium concentration on birds. This data was related to the changes in mean plasma calcium concentration after hormone treatment in birds, and the data is related to two-factor analysis of variance in which first factor was about hormone treatment and second factor was the gender of the birds. There were total of 20 observations in the given example.Table 1 summarized the four groups of an average plasma calcium concentration on birds (mg/100 ml) according to gender and type of the hormone treatment.

Table 1.1: Four groups of an average plasma calcium concentration on birds (mg/100 ml)

Group 1	**Group 2**	**Group 3**	**Group 4**
Factor 1: **Female**	**Factor 1**: Male	**Factor 1:** Female	**Factor 1:** Male
Factor 2 : **Without Hormone Treatment**	**Factor 2 :**Without Hormone Treatment	**Factor 2 :** With Hormone Treatment	**Factor 2 :** With Hormone Treatment
16.5	14.5	39.1	32
18.4	11.0	26.2	23.8
12.7	10.8	21.3	28.8

14.0	14.3	35.8	25.0
12.8	10.0	40.2	29.3

(Sources :Zar, J.H.1999 "Biostatistical Analysis", Prentice Hall, New Jersey).

2.2 Performing Bootstrap, Fitting Multiple Linear Regression and Multilayer Perceptron.

2.2.1 Bootstrap Methodology Approach

One of the property of bootstrapping is it develops a large sample size from original sample that may include an observation several times while omitting other observations. Theoretical sampling distribution is not followed in the case of bootstrapping. Rather, sampling with replacement allows it to develop an empirical distribution [4,5]. Series of steps involved in the process of running bootstrap technique [6]. First, bootstrap develops a mega file by coping original dataset. Secondly, bootstrap draws samples with replacement from that mega file. Third, bootstrap method can calculate and store results by using the drawn sample. Fourth, this method can repeat the process desire number of times. Fifth, averages, standard errors and confidence intervals can be calculated from the stored results.

2.2.2 Fitting Multiple Linear

As it is mentioned in Table 1.1 that data belonged to DOE hence the transformation of data from DOE to regression was required to run the SAS syntax for multiple linear regression model (MLR). Consider the independent variables as single factor with r=4 and n=5.

$$\mathbf{Y}=\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{21} \\ \vdots \\ Y_{45} \end{bmatrix};\ \mathbf{X}=\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & -1 \end{bmatrix};\boldsymbol{\beta}=\begin{bmatrix} \mu. \equiv \beta_0 \\ \tau_1 = \beta_1 \\ \tau_2 = \beta_2 \\ \tau_3 = \beta_3 \end{bmatrix};\ \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{45} \end{bmatrix}$$

There were two variables included in the study and in regression model. Dependent variable was the calcium concentration and independent variables were hormone treatment and gender. Hence, transformation of DOE data was required to fit and run the regression analysis and the method of transformation of data was defined in previously published article [7]. Coding of data is most important part of analysis and it must be done carefully. Indicator variables were required and that takes values 0, 1 and -1. Coding of independent variables will help in obtaining the regression coefficients [8]. Let X_{ij1} denote the value of indicator variable X_1, X_{ij2} denote the value of indicator variable X_2 and X_{ij3} denote the value of indicator variable X_3. Using t-1 indicators in the model and multiple linear regression model for the study can be stated as

$$Y_{ij} = \mu. + \tau_1 x_{ij1} + \tau x_{ij2} + \tau x_{ij3} + \varepsilon_{ij}$$

where,

$$x_{ij1} = \begin{cases} 1 & \text{if the case from factor level 1} \\ -1 & \text{if the case from factor level 4} \\ 0 & \text{Otherwise} \end{cases}$$

$$x_{ij2} = \begin{cases} 1 & \text{if the case from factor level 2} \\ -1 & \text{if the case from factor level 4} \\ 0 & \text{Otherwise} \end{cases}$$

$$x_{ij3} = \begin{cases} 1 & \text{if the case from factor level 3} \\ -1 & \text{if the case from factor level 4} \\ 0 & \text{Otherwise} \end{cases}$$

To illustrate how a linear model was developed with this approach. A portion of the data in the Table 1.1 repeated in Table 1.2, together with the coding of the indicator variables X_{ij1}, X_{ij2} and X_{ij3}.

Table 1.2: Regression approach to the analysis of variance

Y_{ij}	X_{ij1}	X_{ij2}	X_{ij3}
16.5	1	0	0
18.4	1	0	0
12.7	1	0	0
14.0	1	0	0
⋮			
25.0	-1	-1	-1
29.3	-1	-1	-1

After arranging the original data, data were tabulated as Table 1.2. At this stage, the data now were ready for the bootstrapping method. Part II (Section 2.2.4) provided the SAS syntax for the bootstrapping method.

2.2.3 Multilayer Perceptron Neural Network

Multilayer perceptron model is a well-known neural network model, which consists of an input layer, one or severalhidden layers and an output layer. The neurons in multilayer perceptron neuralnetwork are generally grouped into layers. Signals flow in one direction fromthe input layer to the next, but not within the same layer [9]. An essentialfactor of successes of the neural networks depends on the training network.Basically, the multilayer perceptron trainingalgorithm with three-layer architecture means that, the networkhas an input layer, one hidden layer and an output layer.Thus, for the multilayer perceptron network with input nodes, ahidden nodes and one output node which given as follows:

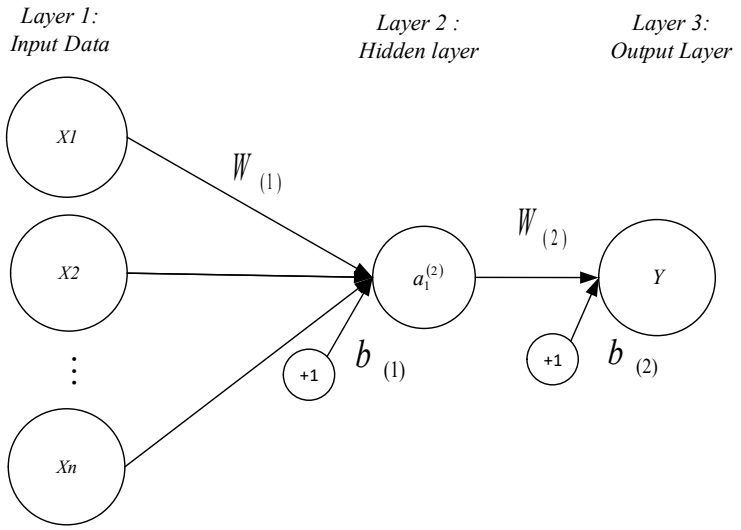

Figure 1.1The architecture of the multilayer perceptron neural network model with one hidden layer, *n* input nodes, a hidden nodes and one output node.

2.2.4 Methodology Building through SAS Syntax

Part I: A SAS Syntax Regression Modeling (Original Data)

```
/* INPUT DATA */
/*Title 'Effect of hormone treatment on the mean plasma calcium concentration of birds of both sexes'*/

Datahealth1;
input y x1 x2x3;
cards;
```

16.5	1	0	0
18.4	1	0	0
12.7	1	0	0
14	1	0	0
12.8	1	0	0
14.5	0	1	0
11	0	1	0
10.8	0	1	0
14.3	0	1	0
10	0	1	0
39.1	0	0	1
26.2	0	0	1
21.3	0	0	1
35.8	0	0	1
40.2	0	0	1
32	-1	-1	-1
23.8	-1	-1	-1
28.8	-1	-1	-1
25	-1	-1	-1
29.3	-1	-1	-1

```
run;
odsrtffile='MLR_data1.rtf';

/* performing regression model without bootstrapping */
procregdata=health1;
model y= x1 x2x3;
run;
odsrtfclose;
run;
```

Part II : A SAS Syntax Regression Modeling (Bootstrap Data)

```
/* INPUT DATA */
/*Title 'Effect of hormone treatment on the mean plasma calcium concentration of birds of both
sexes' */

Datahealth1;
input y x1 x2 x3;
cards;
                    16.5    1    0    0
                    18.4    1    0    0
                    12.7    1    0    0
                      14    1    0    0
                    12.8    1    0    0
                    14.5    0    1    0
                      11    0    1    0
                    10.8    0    1    0
                    14.3    0    1    0
                      10    0    1    0
                    39.1    0    0    1
                    26.2    0    0    1
                    21.3    0    0    1
                    35.8    0    0    1
                    40.2    0    0    1
                      32   -1   -1   -1
                    23.8   -1   -1   -1
                    28.8   -1   -1   -1
                      25   -1   -1   -1
                    29.3   -1   -1   -1
;
run;
odsrtffile='MLR_data1.rtf';

/**BOOTSTRAPPING WITH CASE RESAMPLING**/
procsurveyselectdata=health1 out=boot1 method=urssamprate=1outhitsrep=1000;
run;

/* performing regression model with bootstrapping */
procregdata=boot1;
model y= x1 x2 x3;
run;
Proc print data= Boot1;
```

```
run;
odsrtfclose;
run;
```

Part III: Multilayer Perceptron Approach for Assessingan Accuracy of the Developed Model (Using Bootstrapping Data)

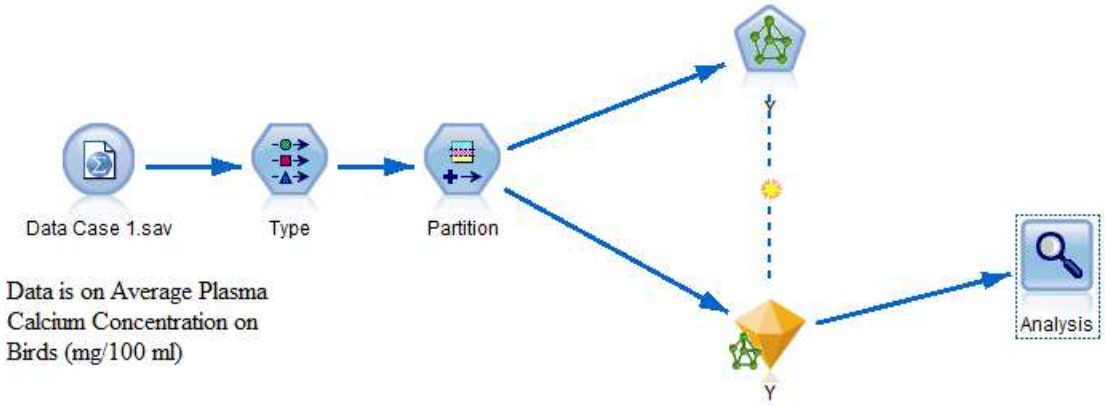

Figure 1.2 Conceptual framework of the analysis using SPSS Modeler.

Figure 1.2 has proposed the structural framework of the analysis. There are four variables that have been identified, one as a dependent variable and rest independent variables. The aim of this research is to find out the accuracy of the developed model. The structural framework of the proposed analysis is given in Figure 1.1. The MLP architecture is consisted of input, hidden and output nodes. There were three independent variables that were considered as input for this analysis. Figure 1.2 shows the architecture of the best (MLP) model with one hidden layer and one output layer which is referred to as the average plasma calcium concentration on birds. Data were partitioned into three which were training (80%), testing (15%) and validation (5%).

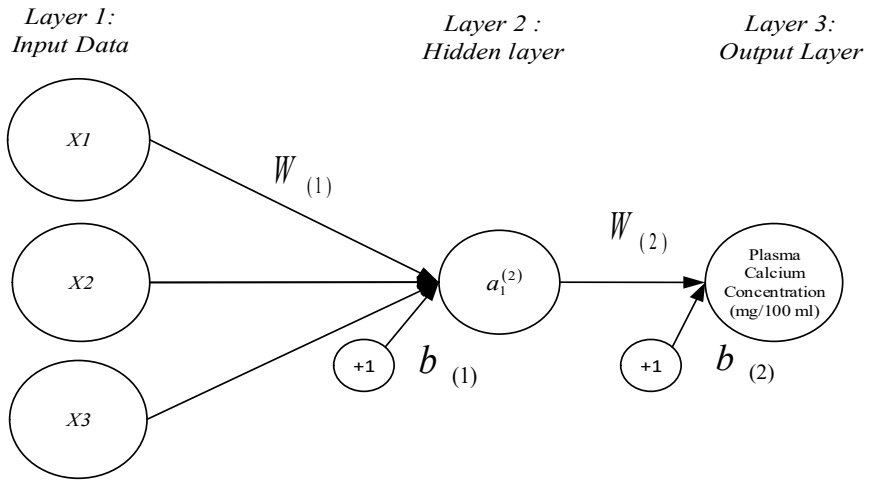

Figure 1.3: The architecture of the best (MLP)model with three input variables, one hidden layer and one output node.

Figure 1.3 shows the architecture of the best (MLP) which is considering three independent variables as inputs and one dependent variable which is average plasma calcium concentration as the output (output node).

3. Results

Fitted regression model was yield when SAS syntax ran for the multiple linear regression of Y on x1, x2 and x3. Output from analysis was summarized under part I and II. Tables have regression output for small dataset with and without bootstrapping

PhaseI: Result for Regression Modeling (using original data)

Table 1.3 Analysis of Variance

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1461.32550	487.10850	21.27	<.0001
Error	16	366.37200	22.89825		
Corrected Total	19	1827.69750			

Table 1.4 Model Fitting

Root MSE	4.78521	R-Square	0.7995
Dependent Mean	21.82500	Adj R-Sq	0.7620
Coeff Var	21.92537		

Table 1.5 Parameter Estimates of Regression Modeling

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21.82	1.07001	20.40	<.0001
x1	1	-6.945	1.85330	-3.75	0.0018
x2	1	-9.705	1.85330	-5.24	<.0001
x3	1	10.695	1.85330	5.77	<.0001

Part I presented the estimates for regression model when small data was used without bootstrapping. Mostly, intercept and slopes were statistically significant and one-way ANOVA results to test the significance of the model was also provided significant results with p-value <.0001. Fitted regression model was given as follow

$$\text{Calcium concentration} = 21.82 - 6.945X_1 - 9.705X_2 + 10.695X_3 \quad (1)$$

Phase II: Result for Regression Modeling (using bootstrapped data)

Table 1.6 Analysis of Variance

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1445440	481813	25923.9	<.0001
Error	19996	371639	18.58568		
Corrected Total	19999	1817079			

Table 1.7 Model Fitting

Root MSE	4.31111	R-Square	0.7955
Dependent Mean	21.77177	Adj R-Sq	0.7954
Coeff Var	19.80138		

Table 1.8 Parameter Estimates of Regression Modeling

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21.788	0.03049	714.69	<0.0001
x1	1	-6.93	0.05249	-132.05	<0.0001
x2	1	-9.657	0.05299	-182.26	<0.0001
x3	1	10.586	0.05278	200.55	<0.0001

When the same data set was used (n=20) but SAS syntax for bootstrapping was employed before running the regression model (Part II). It was found that ANOVA result to test the model fitting was significant (p-value < 0.001) and all slopes and intercept became statistically significant. Hence, regression model can be written as;

$$\text{Calcium concentration} = 21.788 - 6.93X_1 - 9.657X_2 + 10.586X_3 \quad (2)$$

After comparison of part I and II results, especially the tables for parameter estimations, it would be observed that change in magnitude of estimated parameters was not considerable and values of parameters before and after bootstrapping were very close. However, change in standard error was quite significant before and after bootstrapping.

Phase III: Multilayer Perceptron Approach

Multilayer perceptron (MLP) was created based on the four selected variables and also the recommendation proposed by IBM SPSS Modeler 18.0. The accuracy of MLP was 79.3%, which was good level of accuracy (See Figure 1.2).

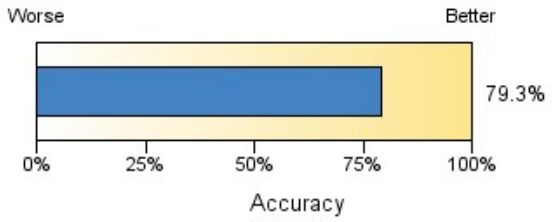

Figure 1.2 The Multilayer Perceptron (MLP)

Table 1.9. The results of Correct, Wrong and Mean Correct for Training, Testing and Validation

Input Variables :**X1,X2,X3**	**Training**	**Testing**	**Validation**
Mean Error (ME)	0.138	0.031	0.048
Mean Absolute Error (MAE)	3.411	3.48	3.479
Maximum Error	7.927	7.927	7.927
Linear Correlation	0.896	0.89	0.892

The performance of the MLP was evaluated through training, testing and validation as shown in Table 1.9. The mean error (ME) and mean absolute error (MAE) of the measurement of the model accuracy were relatively similar to the training, testing and validation.Small error measurement showing the high accuracy of the developed model. These three variables were coming fromDOE transformation toward regression and these variables were being examined from the view of MLP model towards average plasma calcium concentration (mg/100 ml)

4. Conclusion:

The main purpose of this paper was to demonstrate different techniques that can be employed to explain such relationships. In this paper, three different methods had been used: (i) Transforming one-way ANOVA to multiple linear regression (ii) Bootstrap method and, (ii) Multilayer Perceptron Neural Network. The first method was the regression approach to the single-factor analysis of variance. At this stage, regression modeling was designed, and the data needed some modifications as shown in Table 1.2. The second approach was, applying bootstrap to the data and that increased the sample size and lead to the high accuracy of the parameter estimate. It was found through the analysis that bootstrapping was a very useful tool in case of having small data sets. The precision of regression parameters obtained through bootstrapping was higher compare to one without bootstrapping, and precision was evaluated through standard error (SE). It was found that SE reduced 100 times in the regression model derived through booted data compare to the outcome from the original data set. The third approach was to assess the independent variables towards dependent variables. By applying this method, the contribution for each variable toward model building will be clearly seen. The result from the regression model validated through the MLPNN procedure. It is surprising that all these variables appeared to be very significant and mean square error showed high accuracy of the derived model through the testing, training and validation procedure.

5. REFERENCES

[1] Efron, B. (1981). Censored data and the bootstrap. Journal of the American Statistical
a. Association, 76(374):312-9.

[2] Tonidandel, S. & Overall J.E. (2004). Determining the number of clusters by sampling with replacement. Psychological Methods, 9(2):238.

[3] Norizan, M., Maizah, H. A., Suhartono & Ahmad, W.M.A.W. (2012). Forecasting Short Term Load Demand UsingMultilayer Feed-forward (MLFF) Neural Network Model, Applied Mathematical Sciences, 6(108): 5359 – 5368

[4] Bradley, E. &Tibshirani, R.J. (1993). An Introduction to the Bootstrap. New York, NY:
a. Chapman and Hall.

[5] Cassel, D.L. (2010). Bootstrap Mania: Re sampling the SAS. SAS Global Forum 2010:
a. Statistics and Data Analysis,2010: 1-11.

[6] Higgins, G.E. (2005). Statistical significance testing: the bootstrapping method and an application of self-control theory. Southwest journal of criminal justice, 2(1): 54-76.

[7] Ahmad, W.M.A.W., Naing, N.N., Ali, Z. &Mamat, M. (2020). Approximation of Randomization Block Design to Linear Model. UltraScientist, 22(1); 241-246.

[8] Ahmad, W.M.A.W., Khan, S.Q., Rohim, R.A.A., Aleng, N.A. & Ghazali, F.M.M. (2020). Approximation of Randomized Block Design Towards Fuzzy Multiple Linear Regression: A Case Study In Health Sciences. IJSTR, 9(1): 1303-8.

[9] Pham, T.D. & Liu X. (1995). Neural Networks for Identification, Prediction and Control, Great Britain, 1995.