

A Computational Statistics: Multilayer Feed-Forward Neural Network Approach To Two-Way Anova Toward Linear Model

Wan Muhamad Amir W Ahmad^{1*}, Nor Azlida Aleng², Nur Fatiha Ghazalli¹, Nurfadhlin Abdul Halim³, Nor Farid Mohd Noor⁴, Mohamad Shafiq Mohd Ibrahim⁵, Gobikrishnan Veluplay⁶ and Farah Muna Mohamad Ghazali¹, Mohamad Nasarudin Adnan¹

¹ School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM)

16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia

² Faculty of Ocean Engineering Technology and Informatics,

Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia

³ Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM)

Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia

⁴ Faculty of Medicine, Universiti Sultan Zainal Abidin (UniSZA),

Medical Campus, Jalan Sultan Mahmud, 20400 Kuala Terengganu, Terengganu, Malaysia

⁵ Kuliyyah of Dentistry, International Islamic University Malaysia, (IIUM) Kuantan Campus

Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia

⁶ UniKL Mimet Dataran Industri Teknologi Kejuruteraan Marin Bandar Teknologi Maritim Jalan Pantai Remis, 32200 Lumut, Perak

ABSTRACT: This study investigates the connection between Two-Factor Analyses of Variance (Two-way ANOVA) and multiple linear regression (MLR). This paper introduces straightforward instructions and a template for data transformation, making it an essential and mandatory step in constructing multiple linear regression models. The primary concept involves converting Two-way ANOVA into an applied linear model, providing a valuable framework for establishing a comprehensive connection between Two-way ANOVA and linear models. The step-by-step process required to fit the linear model is illustrated in this paper, along with a validation procedure using a Multilayer Feed-Forward Neural Network (MLFFNN). The application of a response surface plot aimed to elucidate the interplay between smoking and gender factors concerning uric acid characteristics. Two-way ANOVA and multiple linear regression (MLR) are interrelated; therefore, the multiple linear regression method is an alternate data analysis strategy.

Keywords: Two-way ANOVA; linear model, Multilayer Feed-Forward Neural Network.

INTRODUCTION

Two-way analysis of Variance (ANOVA) is a statistical technique used to analyze the variance in a dataset when there are two independent categorical variables. When considering the perspective of a linear model, a Two-Way ANOVA can be interpreted as fitting a linear model to the data. Combining a multilayer feedforward neural network with linear regression (which derives from Two-Way Analysis of Variance) can offer several advantages in the context of predictive modeling. The data obtained from the study utilizing two-way analysis of variance will be transformed into linear regression in this work [11]. Prior to that, a comprehensive discussion will be provided on the Two-Way ANOVA Toward Linear Model and the Multilayer Feed-Forward Neural Network.

TWO-WAY ANOVA TOWARD LINEAR MODEL

Two-way ANOVA is an advanced statistical method that extends from one-way ANOVA. The two-way ANOVA method aims to compare the means of the response variable across different groups specified by the factor variable [6, 10, 11]. In 1920, Sir Ronald Fisher developed Two-way ANOVA [9]. At the basic level, Two-way ANOVA can explain the nature of the statistical relation between the mean response and the level(s) of the predictor variable(s). Besides that, Two-way ANOVA provides a method of data analysis that is motivated by consideration of the DOE [5]. Two-way ANOVA can be extended to MLR by creating the independent variables and dependent variables [7]. Table 1 gives the template of Two-Factor Analyses of Variance toward Linear Model. The Factor Effect of Two-Factor Analyses of Variance model for two-factor studies is given in (1.1).

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (1.1)$$

where

$\mu_{..}$ is a constant

$$\begin{aligned}
&\alpha_i \text{ are constantly subject to the restriction } \sum \alpha_i = 0 \\
&\beta_j \text{ are constantly subject to the restriction } \sum \beta_j = 0 \\
&(\alpha\beta)_{ij} \text{ constant subject to the restriction:} \\
&\sum_i (\alpha\beta)_{ij} = 0 \quad j = 1, \dots, b; \quad \sum_j (\alpha\beta)_{ij} = 0 \quad i = 1, \dots, a \\
&\varepsilon_{ijk} \text{ are independent } N(0, \sigma^2) \quad i = 1, \dots, a; \quad j = 1, \dots, b; \quad k = 1, \dots, n
\end{aligned}$$

In this present study, we developed the regression model for the uric acid example that is equivalent to the factor effects Two-Factor ANOVA model (1.1) by first defining the 1, -1, 0 indicator variables for the factor A and Factor B main effect as follows:

$$\begin{aligned}
&1 \quad \text{If the case from level 1 for factor A} & 1 \quad \text{If the case from level 2 for factor A} \\
X_1 = -1 & \text{If the case from level 3 for factor A;} & X_2 = -1 \quad \text{If the case from level 3 for factor A} \\
0 & \text{Otherwise} & 0 \quad \text{Otherwise} \\
&1 \quad \text{If the case from level 1 for factor B} \\
X_3 = -1 & \text{If the case from level 2 for factor B} \\
0 & \text{Otherwise}
\end{aligned}$$

The multiple regression model that is the equivalent of the two-factor ANOVA model for the uric acid example therefore is:

$$Y_{ijk} = \mu.. + \underbrace{\alpha_1 X_{ijk1} + \alpha_2 X_{ijk2}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk3}}_{\text{B main effect}} + \underbrace{(\alpha\beta)_{11} X_{ijk1} X_{ijk3} + (\alpha\beta)_{21} X_{ijk2} X_{ijk3}}_{\text{AB main effect}} + \varepsilon_{ijk}$$

$$i = 1, 2, 3; \quad j = 1, 2; \quad k = 1, 2$$

where

$$X_1, X_2, X_3 \text{ are indicator variables}$$

Here X_{ijk1} denotes the value of indicator variable X_1 for the k -th case from the treatment for which factor A is at the i th level and factor B is at the j th level, and X_{ijk2} and X_{ijk3} have corresponding meanings.

MULTILAYER FEED-FORWARD NEURAL NETWORK (MLFFNN)

In this study, MLFFNN is generally grouped into layers that are divided into input layers, hidden layers, and output layers. In this research, the output node is fixed at one since there is only one dependent variable [2, 3, and 4]. In MLFFNN the values

$$\hat{y} \text{ are given by } \hat{Y} = g_i \left(\sum_{j=1}^H w_j h_j + w_o \right) \text{ where } w_j \text{ an output weight from hidden node } j$$

to the output node is w_o the bias for the output node, g is an activation function. The values of the hidden node $h_j, j=1 \dots H$

are given by $h_j = g_i \left(\sum_{i=1}^H v_{ji} x_i + v_{j0} \right)$. Here, v_{ji} the output weight from input node i to hidden node j , v_{j0} is the

bias for hidden node j where $j=1, \dots, H$ x_i is the independent variables where $i=1, \dots, N$, and g is an activation function [8]. The general architecture of the MLFFNN model is illustrated in Figure 1.

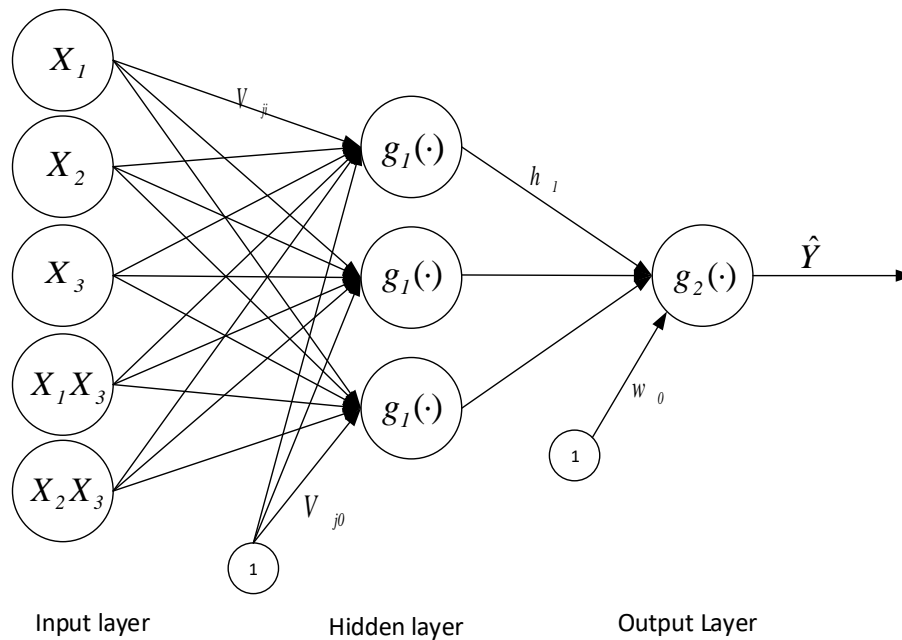


Figure 1. The proposed architecture of the best (MLFFNN) model with five input variables, one hidden layer, and one output node

Five selected variables, which as X_1 , X_2 , X_3 , X_1X_3 , and X_2X_3 are treated as input for MLFFNN. Therefore, the available data set was partitioned into a training set and a testing set with 60% and 40% of the available experimental measurements selected for the training and testing phases, respectively [2, 3, 4]. The computation node in MLFFNN is also stated to as the “hidden neuron of the hidden unit”, the function of the hidden neuron is to intervene between external input and network output. In an analysis having more hidden layers, the network is enabled to extract higher-ordered statistics [1, 9]. Theoretical works have shown that a single hidden layer is sufficient for MLFFNN to approximate any complex non-linear function [8]. Therefore, in this study, one-hidden-layer MLFFNN is proposed and tested with the selected case study. This conducted study initiates with a fitting to Multiple Linear Regression (MLR) toward Two-Factor ANOVA. Then, in the second phase, it proceeds with the MLFFNN procedure. This is to obtain the mean square error for forecasting (MSE-F). The smallest error indicates the obtained model has high accuracy and supports the linear fitting model.

Response surface plot (RSP)

A surface plot is a visualization technique in three dimensions that illustrates the connection between two independent variables and a dependent variable. It presents a continuous surface where the height or color intensity corresponds to the value of the dependent variable at different combinations of the independent variables. Surface plots enable the visualization of intricate relationships and interactions among variables[6]. Typically, the two independent variables are represented on the X_1 and X_2 axes, while the dependent variable is displayed on the Y -axis.

$$Y = f(X_1, X_2) + \varepsilon \quad (3.1)$$

The independent variables, X_1 , and X_2 , play a crucial role in determining the response variable, Y . The dependent variable, Y , is influenced by the values of X_1 and X_2 , as well as the presence of an experimental error term denoted as ε . The error term ε captures various sources of variation, including measurement errors in the response variable and other factors not accounted for by the function f . In statistical terms, the error term ε is assumed to follow a normal distribution with a mean of zero and a variance of σ^2 . To obtain an accurate approximation for the function f , researchers typically begin by considering a low-order polynomial within specific regions of interest. When the response can be represented by a linear relationship with the independent variables, a first-order model is employed [6].

THE STUDY DESIGN, SYNTAX, AND RESULTS.

The output node in this study is one node that refers to one dependent variable (Y). To find the appropriate number of hidden nodes and the best combination of input variables, the model was constructed based on the proposed architecture of the best (MLFFNN), to ensure that this model fits the data. In our case, the best number of hidden nodes is one node. The architecture of the MLFFNN neural network model of Y is composed of five input variables, one hidden node, and one output node, as presented in Figure 1. The performance of MLFFNN was evaluated through sum square error for forecasting (MSE-F) and R^2 which was obtained from the fitting model. Table 1 shows the sample data with A levels of factor A and B levels of factor B, and then each replicate contains all AB treatment combinations.

2.1 The Tabulation of Study Design Data

A table containing the data layout for the two-factor analysis of variance was presented as Table 1.

Table 1. Sample Data on The Reading of Acid Uric and Notation For Two-Factor Study

		Factor II: (Gender)	
Factor I : (Smoking Status)	Factor A(<i>i</i>)	Factor B (<i>j</i>)	
		B ₁ (Male)	B ₂ (Female)
	A ₁ : Status Smoker	Y ₁₁₁ = 550, Y ₁₁₂ = 424	Y ₁₂₁ = 446, Y ₁₂₂ = 440
	A ₂ : Status Ex-smoker	Y ₂₁₁ = 398, Y ₂₁₂ = 387	Y ₂₂₁ = 367, Y ₂₂₂ = 371
	A ₃ : Status Non-smoker	Y ₃₁₁ = 319, Y ₃₁₂ = 273	Y ₃₂₁ = 233, Y ₃₂₂ = 262

2.2 Transforming Data

Table 2 indicates the template of the two-factor study before conducting the linear model analyses. The coding for the input was set at the standard level by defining the 1, -1, and 0 indicator variables for factor A and Factor B main effect.

Table 2. Data Arrangement For The Regression Methodology Building

	(1)	(2)	(3)	(4)	(5)	(6)		
<i>i</i>	<i>j</i>	<i>k</i>	<i>Y</i>	<i>X</i> ₁	<i>X</i> ₂	<i>X</i> ₃	<i>X</i> ₁ <i>X</i> ₃	<i>X</i> ₂ <i>X</i> ₃
1	1	1	550	1	0	1	1	0
1	1	2	424	1	0	1	1	0
1	2	1	446	1	0	-1	-1	0
1	2	2	440	1	0	-1	-1	0
2	1	1	398	0	1	1	0	1
2	1	2	387	0	1	1	0	1
2	2	1	367	0	1	-1	0	-1
2	2	2	371	0	1	-1	0	-1
3	1	1	319	-1	-1	1	-1	-1
3	1	2	273	-1	-1	1	-1	-1
3	2	1	233	-1	-1	-1	1	1
3	2	2	262	-1	-1	-1	1	1

2.3 The Developed Syntax

This section focuses on the methodology building using the R syntax. The full syntax of the methodology is given as follows.

#STEP 1: Data input for the analysis

```
Input =("
Y X1 X2 X3 X1X3 X2X3
550 1 0 1 1 0
424 1 0 1 1 0
446 1 0 -1 -1 0
440 1 0 -1 -1 0
398 0 1 1 0 1
387 0 1 1 0 1
367 0 1 -1 0 -1
371 0 1 -1 0 -1
319 -1 -1 1 -1 -1
273 -1 -1 1 -1 -1
233 -1 -1 -1 1 1
262 -1 -1 -1 1 1
")
data = read.table(textConnection(Input),header=TRUE)
```

Multiple Linear Regression

STEP 2- Develop the regression model

```
Regression_Model <- lm(Y~X1+X2+X3+X1X3+X2X3, data=data)
summary(Regression_Model)
```

MultiLayer Feedforward Neural Network

#STEP 3-Install the Neuralnet Package

```
if(!require(neuralnet)){install.packages("neuralnet")}
library("neuralnet")
```

```

#STEP 4- Checking For the Missing Values/
apply(data, 2, function(x) sum(is.na(x)))

#STEP 5 - Max-Min Data Normalization/
normalize <- function(x) {return ((x - min(x)) / (max(x) - min(x)))}
maxmindf <- as.data.frame(lapply(data, normalize))

#STEP 6-Determine the Training and Testing of the Dataset/
#/60% for Training and 40% For Testing/
Training <- maxmindf[1:10, ]
Testing <- maxmindf[11:12, ]

#STEP 7-Print Dataset -Training and Testing Data set/
print(Training)
print(Testing )

#STEP 8-Plotting the Architecture of MLFFNN Neural Network/
nn <- neuralnet(Y~X1+X2+X3+X1X3+X2X3, data=Training, hidden=3,
  linear.output = F, stepmax = 1000000)
plot(nn)
options(warn=-1)
nn$result.matrix

#####Testing The Accuracy of The Model- Predicted Mean Square Error#####

#STEP 9-Predicted Results are Compared To The Actual Results/.
Temp_test <- subset(Testing, select = c("X1","X2","X3","X1X3","X2X3"))
head(Temp_test)
nn.results <- compute(nn, Temp_test)
results <- data.frame(actual = Testing$Y, prediction = nn.results$net.result)
results

#STEP 10 Use The Predicted Mean Squared Error NN (MSE forecasts the Network) as a
#Measure of How Far the Predictions Are From The Real Data/
predicted <- compute(nn,Testing[,1:5])
MSE.net <- sum((Testing$Y - predicted$net.result)^2)/nrow(Testing)

#STEP 110-Printing the Mean Square Error Forecasting/
MSE.net

#####The End#####

```

2. 4 Results

A summary of the findings from the illustrated case study is presented in this section. Table 3 displays the output of the model's parameter regression analysis.

Table 3 Coefficients Table For Multiple Linear Regression

Model	Unstandardized Coefficients		Standardized Coefficients		t	Sig.	95% CI for B		Collinearity Statistics	
	B	Std. Error	Beta				Lower Bound	Upper Bound	Tolerance	VIF
(Constant)	372.50	11.49			32.42	0.00	344.39	400.61		
X1	92.50	16.25	0.88		5.69	0.00*	52.74	132.26	0.75	1.33
X2	8.25	16.25	0.08		0.51	0.63	-31.51	48.01	0.75	1.33
X3	19.33	11.49	0.22		1.68	0.14	-8.78	47.45	1.00	1.00
X1X3	2.67	16.25	0.03		0.16	0.88	-37.09	42.42	0.75	1.33
X2X3	-7.58	16.25	-0.07		-0.47	0.66	-47.34	32.17	0.75	1.33

Dependent Variable: Y; Adjusted R² 0.894, ANOVA [F(5,6) = 10.094; $p < 0.05$]; Note: Significant Levels: * $p < 0.05$; Normality assumptions were fulfilled; Multiple Linear Regression was applied.

Table 3 gives the result of the linear model fitting with the interaction effect. The full model can be written with the interaction effect as given as follows: (1.2)

$$\hat{Y} = 372.50 + 92.50X_1 + 8.25X_2 + 19.33X_3 + 2.67X_1X_3 - 7.58X_2X_3 \quad (1.2)$$

Equation (1.2) stands out as the optimal model for representing the linear aspects of the two-way ANOVA. The acid uric level prediction can be derived using this particular equation (1.2).

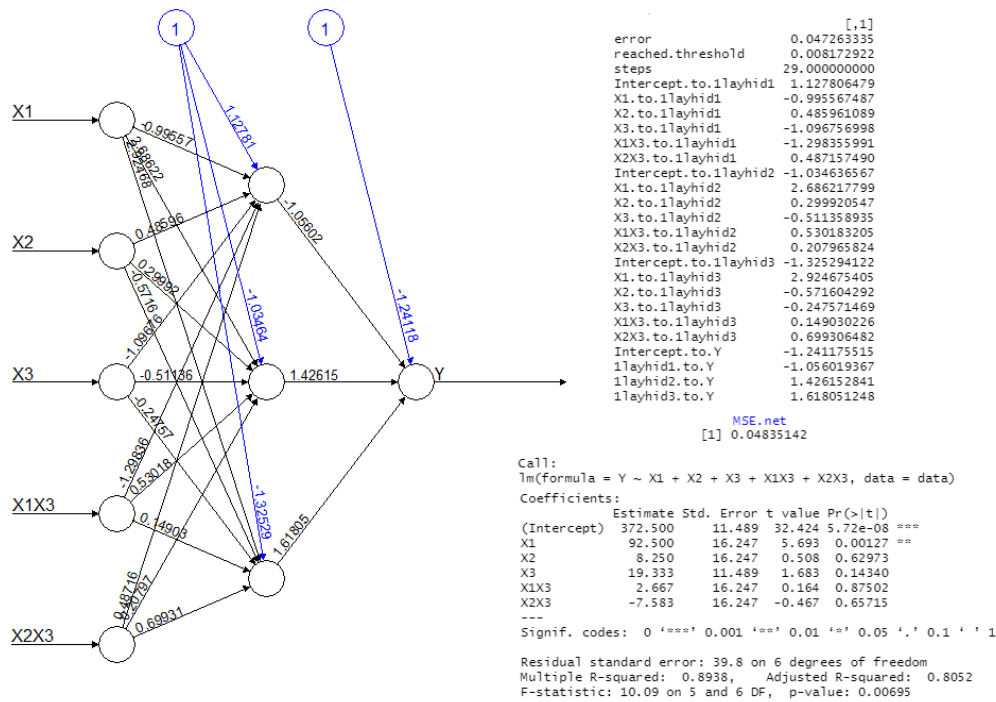


Figure 2. Multilayer Feed-Forward Neural Network (MLFFNN) and Multiple Linear Regression (MLR)

Figures 3 through Figures 5 present the results of the response plot surface (RPS) analysis. This figure provides a summary of the surface plot for all of the possible inputs on the variables that are dependent. The properties of uric acid can be easily inferred from this graph.

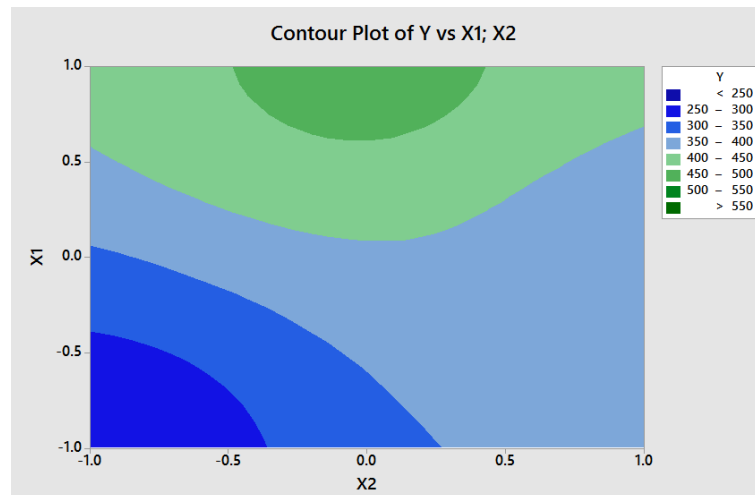


Figure 3. Contour plot of Y versus X₁ and X₂

In Figure 3, the levels of uric acid are displayed for both current smokers (level 1) and former smokers (level 2). (level 2). Individuals who have a propensity for smoking were shown to have a greater level of uric acid, as indicated by the contour. In this particular instance, it was discovered that nonsmokers had lower levels of uric acid. The area of the plot may be found in the lower-left corner of the plot.

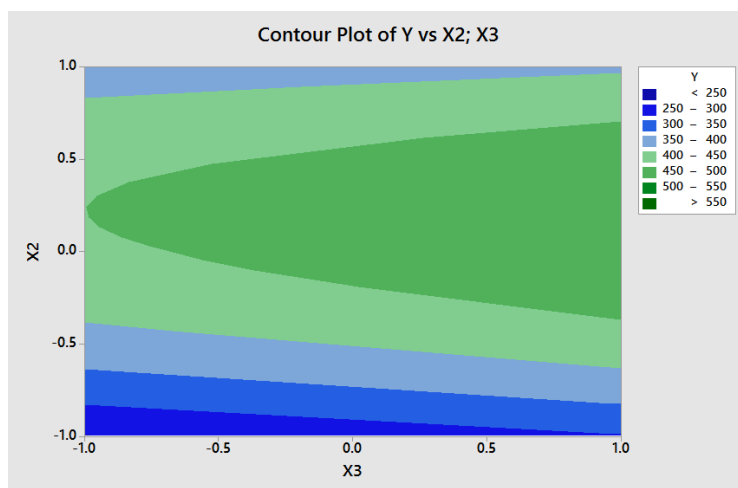


Figure 4. Contour plot of Y versus X_2 and X_3

Figure 4 displays the uric acid levels in ex-smokers (X_2) considering the gender factor (X_3). As indicated by the contour lines, male patients exhibit higher uric acid levels than female patients. However, no distinct pattern of uric acid is observed among ex-smokers in general.

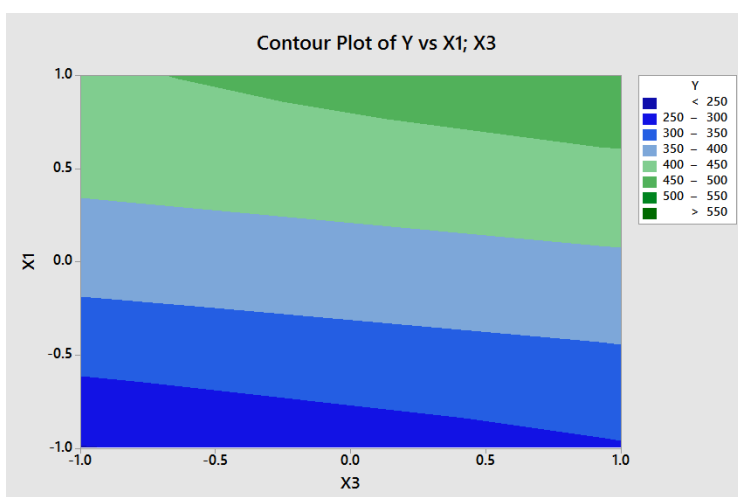


Figure 5. Contour plot of Y versus X_1 and X_3

In Figure 5, the visualization presents the uric acid levels specifically within the cohort of individuals identified as smokers (level 1), taking into account the gender factor. The contour analysis within this context reveals a discernible trend where the uric acid levels are notably elevated in male patients as opposed to their female counterparts.

CONCLUSION

The primary purpose of this paper is to demonstrate two-factor analyses of variance (ANOVA) techniques that can be employed to explain such relationships through multiple linear regression. It has been shown that, from the design of the experiment (DOE) of two-factor analyses of variance (ANOVA), the multiple linear regressions can be well fit according to the template in Table 2. In the second phase of the study, all the independent variables in the regression model are further investigated by performing the MLFFNN procedure. The performance of MLFFNN was evaluated through the result of the sum square error and relative error of testing/out-sample. From the analysis conducted, it was found that only (X_1) contributes significantly to the MLR model while (X_2), (X_3), (X_1X_3), and (X_2X_3) are not directly significant to MLR. In this case, the fitting of the regression model should comprise all the factor levels; therefore, this allows us to investigate the importance of the interaction effect between the factors.

ACKNOWLEDGMENT

The authors express gratitude to the Ministry of Higher Education Malaysia for the Fundamental Research Grant Scheme, which was awarded under Project Code: FRGS/1/2022/STG06/USM/02/10, as well as to Universiti Sains Malaysia (USM).

REFERENCES

1. Neter, J., Kutner, M. H., Wasserman, W., Nachtsheim C.J. *Applied Linear Statistical Models, Regression Analysis, Analysis of A Variance and Experimental Designs*. 4th Edition. Richard D. Irwin, Inc., Homewood, IL, 1990.
2. Norizan M., Maizah H. A., Suhartono, and W. M. A. W. Ahmad. Forecasting Short-Term Load Demand using Multilayer Feed-forward (MLFF) Neural Network Model. *Applied Mathematical Sciences*, **2012**; Vol. 6, No. 108, pp. 5359-5368.
3. Norizan M., Nor Azlida A., W. M. A. W. Ahmad, and Maizah H. A. Multilayer Feed-Forward Neural Network Approach to Lymphoma Cancer Data. *Int. J. Contemp. Math. Sciences*, **2012**; Vol. 7, No. 35, pp. 1749-1756.
4. Norizan, M., W. M. A. W. Ahmad, Nor Azlida A., and Maiza H. A. Modelling Multi-Layer Feed-forward Neural Network Model on the Influence of Hypertension and Diabetes Mellitus on Family History of Heart Attack in Male Patients. *Applied Mathematical Sciences*, **2013**; Vol. 7, No. 41, pp. 2047-2053.
5. Ahmad., W. M. A. W., Nyi Nyi Naing, Zalila Ali, and Mustafa Mamat. Approximation of Randomization Block Design to Linear Model. *UltraScientist*, **2020**; Vol 22(1) M, pp. 241-246
6. Montgomery, D.C. *Design and Analysis of Experiments*, 5th Edn. John Wiley, & Sons, Inc, New York, 2009.
7. Ahmad W. M. A. W., Shafiq, M. M. I., Hanafi A. R., Puspa, L., & Nor Azlida, A. Algorithm for Combining Robust and Bootstrap In Multiple Linear Model Regression. *Journal of Modern Applied Statistical Methods*, **2016**; 15(1), pp. 884-892. DOI: 10.22237/jmasm/1462077900
8. Dulkadiroglu H., Galip Seckin & Derin Orhon. Modeling nitrate concentrations in a moving bed sequencing batch biofilm reactor using an artificial neural network technique, *Desalination and Water Treatment*, **2015**; 54:9, pp. 2496-2503, DOI: 10.1080/19443994.2014.902336
9. Armstrong, R. A., et al. An introduction to the analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry. *Ophthalmic and Physiological Optics*, **2000**; 20(3): 235-241.
10. Churchill, G. A. Using ANOVA to analyze microarray data. *Biotechniques*, **2004**; 37(2), pp. 173-177.
11. Shahbaba, B. Analysis of Variance (ANOVA) *Biostatistics with R*, **2012**; pp. 221-234: Springer.