**Sapporo Medical Journal**
Volume 54, Issue 10, October 2020

# Ordered logistic regression with artificial neural network models for variable selection for prediction of hypertension patient outcomes

Farah Muna Mohamad Ghazali[1], Wan Muhamad Amir W Ahmad[1*], Mohamad Arif Awang Nawi[1], Nor Farid Mohd Noor[1],Nur FatihaGhazalli[1], NorAzlida Aleng[2],Mohamad Shafiq Mohd Ibrahim[3], Nurfadhlina Abdul Halim[4]

[1]School of Dental Sciences, Health Campus, UniversitiSains Malaysia, Kelantan, Malaysia
[2]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Malaysia
[3]Kulliyyah of Dentistry, International Islamic University Malaysia, Pahang, Malaysia
[4]Faculty of Science and Technology, UniversitiSains Islam Malaysia, Negeri Sembilan, Malaysia

**Abstract**— The purpose of this study is to demonstrate the best strategy for the variable selection, using the developed Ordered Logistic Regression (OLR) and Multilayer Perceptron Neural Network (MLP). At the first stage, all the selected variables will be a screen for their important relationship point of view through ordered logistics regression and bootstrap methodology. After considering for 1500 of the bootstrapping methods, it was found that smoking factor, total cholesterol factor, and triglycerides come to a significant relationship to the level of hypertension. By considering the level of significance of 0.25 for ordered logistic regression, these three variables are being selected and used for the input of the MLP model. The performance of MLP was evaluated through the Predicted Mean Square Error (PMSE) of the neural network for the (MSE-forecasts the Network). PMSE is used as a measurement of how far away from our predictions are from the real data. The smallest MSE from MLP, indicate the best combination of variables selection in the model. In this research paper, we also provide the R syntax for OLR and MLP better illustration.

**Keywords—** Multilayer Perceptron Neural Network (MLP), Mean Square Error (MSE), hypertension

## 1. Introduction

The factors which cause hypertension have been studying intensively in the medical and scientific field. The existence of these factors has precipitated or aggravated the condition of hypertension. Scientists throughout the world have extensively identified factors which is can be modifiable for the hypertensive patients. An example of these factorsis overweight, obesity, diabetes mellitus, and lack of physical activities [5]. In certain place hypertension is considered epidemic such in southern China[14]. Here, living standard improvement is speculated as the main reasons of other factors exists [14]. Overweight and weight gain are the highly related to the hypertension [15]. The genetic of these uncontrolled weight problems is regarded linked to the parent's genetic [15]. Any amount of alcohol consumption is linked to hypertension especially in male compared to female [20]. In underdeveloped community such as indigenous people, alcohol is an important cause [11]. Presence of alcoholism within developed countries clearly doubling the risk factors in addition to weight problems. In addition to these risk factors is smoking. The smoking is demonstrated to induce oxidative stress to the mitochondia in the cells [7]. Both smoking and alcohol are modifiable risk factors that closely related to blood pressure. In this study, three major approach were used before performing analysis which bootstrapping, Ordered Logistic Regression (OLR) and Multilayer Perceptron Neural Network (MLP).

Bootstrap is a resampling technique proposed by Efron *et al*. [10]. The idea behind bootstrap is to use a sample in hand as a population to take a sample (by replacement) of the sample in hand and make a large

number of"which consist of case resampling samples" known as bootstrap samples. The bootstrap method begins with an original sample taken from a specific population under consideration. The next step is to duplicate the original sample several times to create a new population taking into account the original population. In that case, the bootstrap draws several samples with replacement by random sampling approach, and as a result, it provides a different sample from the original sample. These technique stores new data sets and create new distributions for further analysis [9,10]. The advantage of using bootstrap is its ability to extend the sample to the same size as the original, which may include some observations while eliminating other observations.

Ordinal logistic regression or ordered logistic regression is a type of logistic regression analysis when the response variable has more than two categories with having natural order or rank [1,3,13]. The Maximum Likelihood (ML) method will be applied for the value of the parameter estimate. To apply this methodology, the hypertension factor is being converted into an ordinal scale [8,16]. The model for ordinal is given by $y_i^* = x_i\beta + \varepsilon_i$ , however, since the dependent variable is categorized, we must instead use: $C_x(x) = \ln\left[\frac{P(Y \le j \mid x)}{P(Y > j \mid x)}\right]$ and $\ln\frac{\sum pr(event)}{1-\sum pr(event)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \ldots + \beta_k X_k$ .Itcanbe summarized as $\ln\left(\frac{P(Y \le j \mid x)}{1-P(Y \le j \mid x)}\right) = \alpha_j + \beta_i X_k, i = 1...k, \quad j = 1,2,..., p-1$ (1) where $\alpha_j$ = called threshold or intercept, $\beta_i$ = Parameter in the model and $X_i$ = Set of factors or independent variables. The equation (1) above is an ordinal logistic model for *k* predictors with the *p*-1 levels response variable [1,2,4]. The ordered logistic regression model is fitting through the R software. The model is fitted through the procedure of Maximum Likelihood Estimation (MLE).

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN) with one or more layers between the input, hidden, and output layer. In the research study, the output node of this analysis is fixed at one since there is only one dependent variable. Equation (1) gives the MLP with N input nodes, H hidden nodes, and one output node. The values of $\hat{y}$ are given as follows $\hat{Y} = g_i\left(\sum_{j=1}^{H} w_j h_j + w_0\right)$ where $w_j$ an output weight from hidden node *j* to the output node, $w_0$ the bias for the output node, and *g* is an activation function. The values of the hidden node $h_j$, $j$ =1…H are given by $h_j = g_i\left(\sum_{j=1}^{H} v_{ji} x_i + v_{j0}\right)$ where $v_{ji}$ the output *weight* from input node *i* to hidden node *j*, $v_{j0}$ is the bias for hidden node *j* where *j* =1, …, H and $x_i$ are the independent variables where *i* =1, …, N and *k* is an activation function [17,18,19]. The general architecture of the MLP model is illustrated in Figure 1.

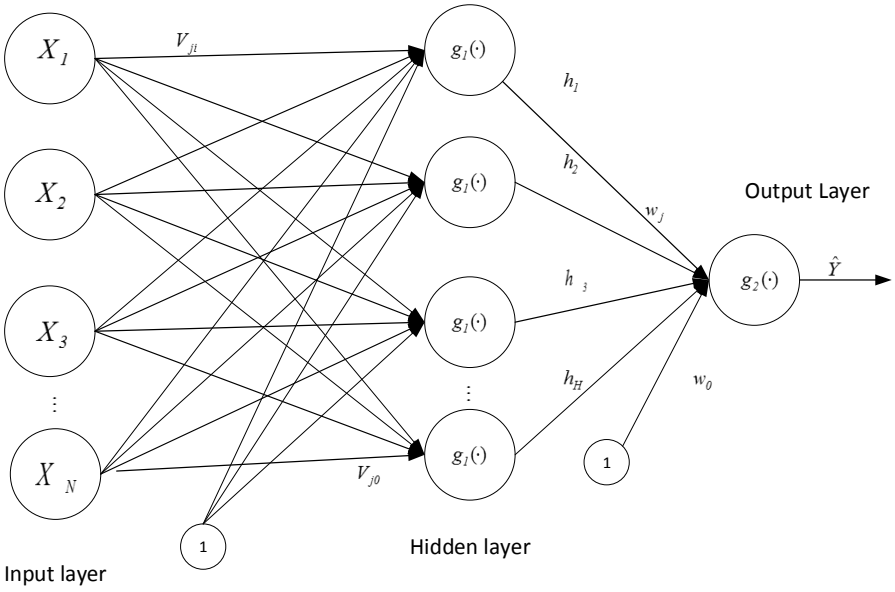



Fig. 1: The general architecture of the MLP with one hidden layer, N input nodes, H hidden nodes, and one output node

### *1.1 Method selection and combination*

Firstly, the collected data will be the screen for the outlier. In the second step, the data will be bootstrap for 1500 times. This is to make sure that the obtained sample is large enough for the best parameter estimate. Data through the bootstrap procedure will be used for the ordered logistics regression modeling. Using bootstrapping data for the modeling purpose usually will help the researcher to obtain the real relationship between studied variables. Third, the selected variable from the ordered logistic regression procedure will be the input for the multilayer perceptron analysis. The MLP analysis through this procedure will help to determine the mean error of the prediction using the selected variable from the ordered logistic regression. The smallest mean square error indicates the better results obtained for the conducted analysis.

## 2. Data and the R syntax

Data from the medical unit record, Hospital USM were collected, reviewed and the related information was extracted. The sampling frame was the list of patients, which diagnosed with hypertension. All related variable is being collected and summarized in Table 1.

Table 1: Data Description

| Variable | Code | Description |
|---|---|---|
| Hypertension | Y | 0 = Normal, 1= Borderline, 2 = Definite |
| Smoke | X1 | 1= Never, 2 = Former, 3 = Current |
| Alcohol | X2 | 1= No, 2 = Yes |
| Choltot | X3 | Total Cholesterol measurement |
| Glucos | X4 | Serum Fasting Glucose measurement |
| Trig | X5 | Triglycerides measurement |

### *2.1 The R syntax for the analysis (method selection and combination)*

```
#/Complete Dataset for a Hypertension Patient/
Input = ("
Hyper FHHA Smoke AlcohCholtGlucos Trig
```

```
3 1 1 1 228 153 304
1 1 2 1 200 185 182
1 1 2 1 207 246 112
2 2 3 2 263 153 287
3 1 1 1 183 215 195
1 1 2 1 165 168 69
1 1 3 2 184 161 373
1 1 2 1 230 356 710
3 1 2 2 254 249 137
3 1 2 1 286 203 171
⋮ ⋮ ⋮ ⋮
2 1 2 1 215 127 86
3 1 2 1 185 175 164
3 1 2 1 213 227 97
3 1 1 2 174 126 217
1 1 2 2 197 165 83
3 1 2 2 194 148 194
2 1 2 1 172 140 267
1 1 2 2 145 147 115
1 1 2 1 218 133 154
3 1 2 1 180 139 240
2 1 1 1 186 217 89
1 2 1 1 160 157 61
3 1 2 1 207 168 168
3 1 1 2 230 144 163
3 1 3 1 192 140 146
3 1 2 2 193 130 214
")
data = read.table(textConnection(Input),header=TRUE)

#####################################OLR#########################################

#/Performing Bootstrap for 1000
mydata<- rbind.data.frame(data, stringsAsFactors = FALSE)
iboot<- sample(1:nrow(mydata),size=1500, replace = TRUE)
bootdata<- mydata[iboot,]

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

bootdata$Hyper<-factor(bootdata$Hyper)
model <- polr(Hyper ~ FHHA+Smoke+Alcoh+Cholt+Glucos+Trig, data=bootdata, Hess=TRUE)
summary(model)

## Store Table
(ctable<- coef(summary(model)))

## Calculate and Store p Values
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
## Combined Table
(ctable<- cbind(ctable, "p value" = p))

##################################### MLPNN#########################################
```

```
#/Install the Neuralnet Package/
if(!require(neuralnet)){install.packages("neuralnet")}
library("neuralnet")
#/Checking For the Missing Values/.
apply(data, 2, function(x) sum(is.na(x)))

#/Scaling the data for Normalization
# Method (usually Called Feature Scaling) to get all The Scaled Data
# in the Range [0,1]/
max_data<- apply(data, 2, max)
min_data<- apply(data, 2, min)
data_scaled<- scale(data,center = min_data, scale = max_data - min_data)

#/Randomly Split the Data Into 70:30
#70 Percent of the Data at Our Disposal to Train the Network
#30 Percent to Test the Network/
index = sample(1:nrow(data),round(0.70*nrow(data)))
train_data<- as.data.frame(data_scaled[index,])
test_data<- as.data.frame(data_scaled[-index,])

# Print Data
print(train_data)
print(test_data )

#/Build the Network
#There 3 Hidden Layers Have 3 and 2 Neurons Respectfully
#Input Layer = 2
#Output Layer = 1/

nn<- neuralnet(Hyper ~Glucos+Trig+Alcoh, data=train_data, hidden=c(4,2),
linear.output = F, stepmax = 1000000)
plot(nn)
options(warn=-1)

#/30 Percent of The Available Data to Do This:
#Using Only the First 2 Columns Representing the Input Variables
#of The Network and 1 is The Output For NN/
predicted <- compute(nn,test_data[,1:3])
#/Use the Mean Squared Error NN (MSE-forecasts the Network) as a Measure of How Far
#Away Our Predictions Are from The Real Data/
MSE.net <- sum((test_data$Hyper - predicted$net.result)^2)/nrow(test_data)
MSE.net
```

## 3. Results and discussion

### *3.1 The result from ordered logistics procedure (using R syntax)*

The first section is the result which gains from the ordered logistic procedure. Table 2 gives the complete result of the analysis.

Table 2: Analysis of Maximum Likelihood Estimates

| Parameter | Coefficient | Std. Error | t-value | *p*-value |
|---|---|---|---|---|
| Threshold 1\|2 | -0.220 | 0.267 | -0.824 | 0.410 |
| Threshold 2\|3 | 0.479 | 0.267 | 1.794 | 0.073 |

Table 2: Analysis of Maximum Likelihood Estimates

| Parameter | Coefficient | Std. Error | t-value | *p*-value |
|---|---|---|---|---|
| Threshold 1\|2 | -0.220 | 0.267 | -0.824 | 0.410 |
| Threshold 2\|3 | 0.479 | 0.267 | 1.794 | 0.073 |
| FHHA | -0.114 | 0.078 | -1.456 | 0.145 |
| Smoke | 0.061 | 0.059 | 1.042 | 0.298 |
| Alcoh | 0.174 | 0.077 | 2.251 | 0.024* |
| Cholt | 0.001 | 0.001 | 1.306 | 0.191* |
| Glucos | 0.000 | 0.001 | 0.355 | 0.723 |
| Trig | 0.001 | 0.000 | 2.584 | 0.010* |

*Ordered logistic regression was applied.*

**Significant at the level of 0.05*

The association between hypertension factor with a family history of heart attack (FHHA), smoking (Smoke), consumption of alcohol (Alcoh), total cholesterol (Choltot), glucose reading (Glucos), and Triglycerides (Trig) is being presented in Table2. The result was considered statistically significant for a p-value $\leq 0.25$. All statistical analyses in this study were performed using the statistical R Software. According to Table 2, there is a significant association between hypertension status with the consumption of alcohol. The estimated coefficient for alcohol is 0.174. The value of (Odd Ratio) OR is obtained exp (0.174) = 1.2. This can be explained that consumption of alcohol factors increases one time the odd of hypertension status. The second factor related to hypertension is the level of total cholesterol level. The estimated coefficient for total cholesterol reading is 0.001. The value of OR is obtained exp (0.001) = 1.00. This indicates that one unit increase in total cholesterol (mg/dL) reading, will most probably increase the blood pressure around one time. The third factor related to hypertension is the level of triglycerides. The estimated coefficient for triglycerides reading is 0.001. The value of OR is obtained exp (0.001) = 1.00. This indicates that one unit increase in triglycerides (mg/dL) reading, will most probably increase the blood pressure around one time.

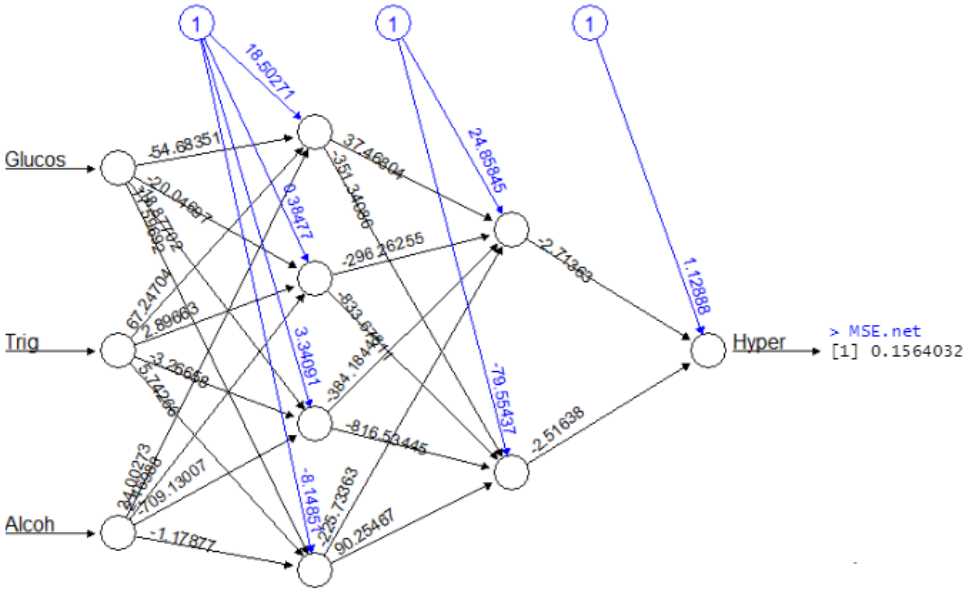
A study conducted by Briasoulis*et al.*[6] has proven that heavy alcohol intake > 20 g / d is associated with a higher risk of developing hypertension in women and men. While with light to moderate alcohol intake (<20 g / d), women have the potential to reduce the risk of hypertension, while men have the risk of increased high blood pressure[6]. A study was conducted by Hong [12], his study was to find an association between obesity levels and the risk of getting the disease. It was found that the abdominal obesity group had a 1.59-fold higher risk of high cholesterol, the risk of hypertension is 1.26 times higher and the risk of hyperglycemia is 1.54 times higher than the normal group. In 2013, the obese group with a higher BMI had a risk of high cholesterol 1.72 times higher, and the risk of hypertension 1.43 times higher than the normal group [9,11,12]. Increased triglyceride (TG) levels are a key feature of lipids closely related to hypertension. Higher TG level readings with the development of central obesity and insulin resistance are important factors towards the status of high blood pressure. The risks associated with triglycerides are more common in women than in men [12].

### *3.2 The result from multilayer perceptron neural network (using R syntax)*

Figure 2 shows the architecture of the MLP with one hidden layer, 4 input nodes, 2 hidden nodes, and one output node. It was observed that the accuracy of the model is being evaluated through the value of Predicted Mean Square Error (PMSE). The obtained value of PMSE 0.1564. This value is considered small

and the MLP model is considered as a good model.

```
MSE.net
#/Use The Mean Squared Error NN (MSE-forecasts the Network) as a Measure of How Far
#Away Our Predictions Are From The Real Data/
```




```
# Ordered Logistic Regression
bootdata$Hyper<-factor(bootdata$Hyper)
model <- polr(Hyper ~ FHHA+Smoke+Alcoh+Cholt+Glucos+Trig, data=bootdata, Hess=TRUE)
summary(model)
```

| | Value | Std. Error | t value | p value |
|---|---|---|---|---|
| FHHA | -0.1138669152 | 0.0782109210 | -1.4558953 | 0.145421555 |
| Smoke | 0.0614470212 | 0.0589921490 | 1.0416135 | 0.297590888 |
| Alcoh | 0.1738143767 | 0.0772096291 | 2.2512008 | 0.024372824 |
| Cholt | 0.0012684171 | 0.0009709859 | 1.3063188 | 0.191444167 |
| Glucos | 0.0002725227 | 0.0007687614 | 0.3544959 | 0.722967305 |
| Trig | 0.0010275463 | 0.0003975947 | 2.5844066 | 0.009754672 |
| 1\|2 | -0.2198881486 | 0.2669000503 | -0.8238595 | 0.410019381 |
| 2\|3 | 0.4789817859 | 0.2670445087 | 1.7936403 | 0.072870604 |

Fig. 2: The architecture of the MLP with one hidden layer, 4 input nodes, 2 hidden nodes, and one output node

## 4. Summary and conclusion

This paper is focusing on two major parts, which emphasized methodology based on the bootstrapping method, ordered logistics regression, and multilayer perceptron neural network. The idea of a methodology based is being synchronized with the application through the R syntax algorithm. In this study, R syntax is being developed by emphasizing the step-by-step calculation procedure. The first step of an algorithm is emphasized on the ordered logistic regression syntax with the combining methodology of bootstrap. After the bootstrapping the data, the basic ordered regression was fixed according to the bootstrap data, and the level of significance was fixed at the level of 0.25. In the second stage, the significant variable was being analyzed through a multilayer perceptron neural network. The significant variable from the ordered logistic regression will be treated as an input for the MLP procedure, and the accuracy of the model will be evaluated through the value of predicted mean square error (PMSE). The smallest value of MSE, the better result achieved. From the analysis, we found a strong relationship between hypertension status with alcohol consumption, total cholesterol, and the level of triglycerides. This technique had led to successful research and give the best results for decision making, especially for the decision-maker.

Identifying these associated factors by various method will be help the medical practitioners in the future. Inhibiting these factors will benefit the patient's life with slowing the damage done by the hypertension such as cardiovascular accident [5]. Knowledge of the risk factors should be introduced to the teenage and young ages person. The awareness of hypertension itself is one of the keys that can control the danger of hypertension [14]. Increasing public awareness is crucial in preventing the hypertension[15]. The education is a key to this public awareness. This should be educated early on especially to the young person who has parents as hypertensive patients where alcohol is a risk factor [20]. Controlling the hypertension even in underdeveloped communities will benefit to all [11]. Another added benefit is the controlling of the smoking which is damaging to the mitochondrial system [7]. In conclusion, identifying the risk factors is endless work which greatly help in managing the hypertension.

## 5. Acknowledgments

The authors would like to express their gratitude to UniversitiSains Malaysia (USM) for providing the research funding (Short Term Grant No.304/PPSG/6315410, School of Dental Sciences,Health Campus, UniversitiSains Malaysia, Kelantan, Malaysia).

## 6. References

[1] Adeleke, K. A. and Adepoju, A. A. Ordinal logistic regression model: An application to pregnancy outcomes, J Math Stat. 2010; 6: 279–285.

[2] Adepoju, A. A. and Adegbite, M. Application of ordinal logistic regression model to occupational data, J. Sci. Ind. 2009; Stud 7: 39–49.

[3] Agresti, A. An Introduction to Categorical Data Analysis, 2nd Edition, Wiley, New York; 2007.

[4] Ananth, C. V. and Kleinbaum, D. G. Regression model for ordinal responses: A review of methods and applications, Int. J. Epidemiol. 1997; 26: 1323–1332.

[5] Booth, J. N., Li, J., Zhang, L., Chen, L., Muntner, P. and Egan, B. Trends in prehypertension and hypertension risk factors in us adults: 1999-2012, Hypertension. 2017; 70(2): 275–284. DOI: https://doi.org/10.1161/HYPERTENSIONAHA.116.09004.

[6] Briasoulis, A., Agarwal, V. and Messerli, F. H. Alcohol consumption and the risk of hypertension in men and women: a systematic review and meta-analysis, The Journal of Clinical Hypertension. 2012; 14(11): 792–798. DOI: 10.1111/jch.12008.

[7] Dikalov, S., Itani, H., Richmond, B., Vergeade, A., Rahman, S. M. J., Boutaud, O., Blackwell, T., Massion, P. P., Harrison, D. G. and Dikalova, A. Tobacco smoking induces cardiovascular mitochondrial oxidative stress, promotes endothelial dysfunction, and enhances hypertension, Am J Physiol Heart Circ Physiol. 2019; 316(3): H639–H646. DOI: 10.1152/ajpheart.00595.2018.

[8] Dong, Y. Logistic regression models for ordinal response: A study of self-efficacy in colorectal cancer screening, PhD thesis, The University of Texas School of Public Health, Texas Medical Center Dissertations (via ProQuest) AAI1444593; 2007.

[9] Efron, B. The jackknife, the bootstrap, and other resampling plans, Philadelphia, Pa., Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104); 1982.

[10] Efron, B. and Tibshirani, R. An introduction to the bootstrap, Chapman and Hall, New York; 1994.

[11] Ferreira, A. A., Souza-Filho, Z. A., Gonc¸alves, M., Santos, J. and Pierin, A. Relationship between alcohol drinking and arterial hypertension in indigenous people of the mura ethnics, brazil, PloS One. 2017; 12(8): e0182352. DOI: https://doi.org/10.1371/journal.pone.0182352.

[12] Hong, M. H. Relationships of obesity, total-cholesterol, hypertension and hyperglycemia in health examinees with disabilities, Journal of the Korea Academia-Industrial cooperation Society. 2016; 17(10): 591–599.

[13] Hosmer, D. W. and S., L. Applied Logistic Regression, 2nd Edition, Wiley, New York; 2000.

[14] Hu, L., Huang, X., You, C., Li, J., Hong, K., Li, P., Wu, Y., Wu, Q., Bao, H. and Cheng, X. (2017). Prevalence and risk factors of prehypertension and hypertension in southern china, PloS One. 2017; 12(1): e0170238. DOI: https://doi.org/10.1371/journal.pone.0170238.

[15] Khader, Y., Batieha, A., Jaddou, H., Rawashdeh, S. I., El-Khateeb, M., Hyassat, D., Khader, A. and Ajlouni, K. Hypertension in jordan: Prevalence, awareness, control, and its associated factors, International Journal of Hypertension. 2019; 8 pages. DOI: https://doi.org/10.1155/2019/3210617.

[16] McCullagh, P. Regression models for ordinal data, Journal of the Royal Statistical Society. 1980; 42: 109–142.

[17] Mohamed, N., Ahmad, M. and Ahmad, W. M. A. W. Forecasting short term load demand using multilayer feed-forward (MLFF) neural network model, Applied Mathematical Sciences. 2012; 6(108): 5359–5368.

[18] Mohamed, N., Ahmad, W. M. A. W., Aleng, N. N. and Ahmad, M. Modeling multi-layer feed-forward neural network model on the influence of hypertension and diabetes mellitus on a family history of a heart attack in male patients, Applied Mathematical Sciences. 2013; 7(41): 2047–2053.

[19] Mohamed, N., Aleng, N. N., Ahmad, W. M. A. W. and Ahmad, M. Multilayer feed-forward neural network approach to lymphoma cancer data, International Journal of Contemporary Mathematical Sciences. 2012; 7(35): 1749–1756.

[20] Roerecke, M., Tobe, S. W., Kaczorowski, J., Bacon, S. L., Vafaei, A., Hasan, O. S. M., Krishnan, R. J., Raifu, A. O. and Rehm, J. Sex-specific associations between alcohol consumption and incidence of hypertension: A systematic review and meta-analysis of cohort studies, Journal of the American Heart Association. 2018; 7(13): e008202. DOI: 10.1161/JAHA.117.008202.