JP Journal of Biostatistics
© 2025 Pushpa Publishing House, Prayagraj, India
https://pphmjopenaccess.com
https://doi.org/10.17654/0973514325029
Volume 25, Number 3, November 2025, Pages 533-543 **P-ISSN: 0973-5143**

MODELING AND VALIDATING QUALITATIVE PREDICTORS IN MULTIPLE LINEAR REGRESSION USING NEURAL NETWORKS

Nor Azlida Aleng[1,*], Wan Muhamad Amir W Ahmad[2,*], Nurfadhlina Abdul Halim[3], Syerrina Zakaria[1], Norsamsu Arni Samsudin[2], Mohamad Nasarudin Adnan[2] and Farah Muna Mohamad Ghazali[4]

[1]Faculty of Computer Sciences and Mathematics
Universiti Malaysia Terengganu (UMT)
Malaysia

[2]School of Dental Sciences
Universiti Sains Malaysia (USM)
Malaysia

[3]Risk and Analytics Research Group
Faculty of Science and Technology
Universiti Sains Islam Malaysia (USIM)
Malaysia

Received: June 7, 2025; Accepted: July 25, 2025
Keywords and phrases: linear regression, MLFF, qualitative predictor, blood pressure.
*Corresponding authors

How to cite this article: Nor Azlida Aleng, Wan Muhamad Amir W Ahmad, Nurfadhlina Abdul Halim, Syerrina Zakaria, Norsamsu Arni Samsudin, Mohamad Nasarudin Adnan and Farah Muna Mohamad Ghazali, Modeling and validating qualitative predictors in multiple linear regression using neural networks, JP Journal of Biostatistics 25(3) (2025), 533-543. https://doi.org/10.17654/0973514325029

Published Online: September 5, 2025

[4]Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM)
Malaysia

Abstract

This study aims to establish the fundamental concepts of multiple linear regression models involving qualitative predictor variables and to validate their associations using a Multilayer Feed Forward (MLFF) neural network. A simple guide is introduced to separate qualitative variables based on the number of classes, ensuring they meet the assumptions of multiple linear regression. The approach provides a basic and practical template for integrating qualitative predictors into applied linear models. Validation is carried out using the sum of square error and relative error obtained from the MLFF neural network. The low error values produced by the MLFF model highlight the effectiveness and superiority of the proposed methodology.

1. Introduction

Qualitative predictor variables with categories are represented using indicator variables, where each indicator is assigned a value of 0 or 1. These indicator variables, often called *dummy* or *binary variables*, allow the inclusion of categorical data in regression models. If a qualitative variable has more than two categories, then additional indicator variables are needed to represent each category in the model [1, 5-7].

Case I. Qualitative predictor variable with two classes

If a qualitative predictor variable has two categories, then only one indicator variable is needed in the regression model [1]. For example, consider the regression of Y_{i1} on X_{i1} and X_{i2}, where X_{i2} is qualitative with two classes. Define the qualitative variable as follows:

Qualitative predictor variable with two classes:

$$X_{i2} = \begin{cases} 0, & \text{if case from Type I,} \\ 1, & \text{if case from Type II.} \end{cases}$$

Case II. Qualitative predictor variable with three classes

If a qualitative predictor variable has three classes, then we require two indicator variables in the regression model [1]. Consider the regression of Y_{i1} on X_{i1} and X_{i2}, where X_{i2} is qualitative with three classes. For example, let us define X_{i2} : 1 = Type I, : 2 = Type II, : 3 = Type III. Therefore, X_{i2} is being assigned into two different groups, let us define the variable as follows:

$$X_{i21(1)} = \begin{cases} 1, & \text{if case from Type II,} \\ 0, & \text{otherwise,} \end{cases}$$

$$X_{i21(2)} = \begin{cases} 1, & \text{if case from Type III,} \\ 0, & \text{otherwise.} \end{cases}$$

2. Materials and Methods

Multilayer perceptron neural network

A basic Multilayer Feed Forward (MLFF) neural network is typically organized into layers, consisting of an input layer, one or more hidden layers, and an output layer [2-4]. Since there is only one dependent variable in this analysis, the output node is fixed at one. For an MLFF neural network with *N* input nodes, *H* hidden nodes, and a single output node, the structure of values $\hat{y}$ can be represented as (1):

$$\hat{Y} = g_i\left(\sum_{j=1}^{H} w_j h_j + w_0\right), \tag{1}$$

where

w_j represents the output weight from hidden node *j* to the output node,

w_0 denotes the bias for the output node,

g is an activation function applied to the weighted sum of inputs from the hidden nodes.

The values of the hidden node h_j, $j = 1, \ldots, H$ are given by

$$h_j = g_i\left(\sum_{j=1}^{H} v_{ji}x_i + v_{j0}\right), \tag{2}$$

where

v_{ji} is the output weight from input node i to hidden node j,

v_{j0} is the bias for hidden node j, where $j = 1, \ldots, H$,

x_i is the independent variables, where $i = 1, \ldots, N$ and k is an activation function [8].

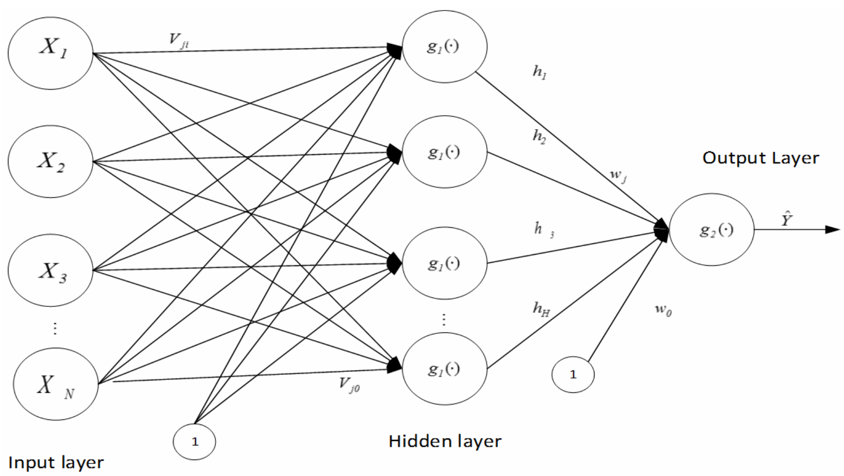

Figure 1. The general architecture of the Multilayer Feed Forward (MLFF) neural network with one hidden layer, N input nodes, H hidden nodes, and one output node.

The MLFF neural network consists of a set number of input, hidden, and output nodes. In this case, there are three selected variables, which define as X_1, X_2, X_3. All the variables were tested using MLFF neural network. Data-set was partitioned into a training set and a testing set (two subsets) which were set at 70% and 30% [2-4, 8]. A model of MLFF neural network was distinguished itself by the presence of one or more hidden layers. The computation nodes in an MLFF neural network are referred to as 'hidden neurons' or 'hidden units'. These hidden neurons serve as intermediaries between the external inputs and the network's output. With more hidden layers, the network can capture higher-order statistics [5]. In this study, an MLFF neural network modeling technique was employed. However, theoretical research has demonstrated that a single hidden layer is sufficient for an MLFF neural network to approximate any complex non-linear function [7, 8]. However, MLFF neural network also can have more than one hidden layer. In this case, the MLFF neural network model was constructed based on the recommendations proposed by SPSS. One-hidden-layer MLFF neural network is found to be the most suitable model for the studied case. This paper is structured as follows: at first, it initiates with an introduction to Multiple Linear Regression (MLR) toward the qualitative predictor variable. Secondly, it validates the model through the multilayer perceptron neural network.

3. Result and Discussion

In this section, the procedure of regression modeling will be discussed further based on the consideration of the qualitative predictor variable. At first, the modeling procedure will be illustrated based on one qualitative variable with two classes and in the second case, the modeling procedure will emphasize one qualitative variable with three classes.

Case I

Table 1. Data description of the selected variable in the study

Num.	Variables	Explanation of user variables
1	BMI	BMI reading of selected patient
2	Waist	Waist circumference of patient
3	Hypertension status	0 = Normal blood pressure 1 = High blood pressure

Let us define the variable as follows:

$$X_2 = \begin{cases} 0, & \text{Normal blood pressure,} \\ 1, & \text{Hypertension,} \end{cases}$$

where X_2 is the variable with two classes. Therefore, we require only one indicator variable. A first-order regression model is given as follows:

$$\text{BMI} = 7.776 + 0.244 \text{ Waist} - 3.533 \text{ Hyper status.}$$

Table 2. Parameter estimate for regression model

	Estimate	**Std. error**	***t*-value**	***p*-value**	**Tolerance**
(Constant)	7.776	8.215	0.947	0.37	
Waist	0.244	0.076	3.188	0.01	0.995
Hyper_Status	-3.533	1.901	-1.859	0.09	0.995

Simple linear regression was applied

Multilayer feed forward neural network for Case I

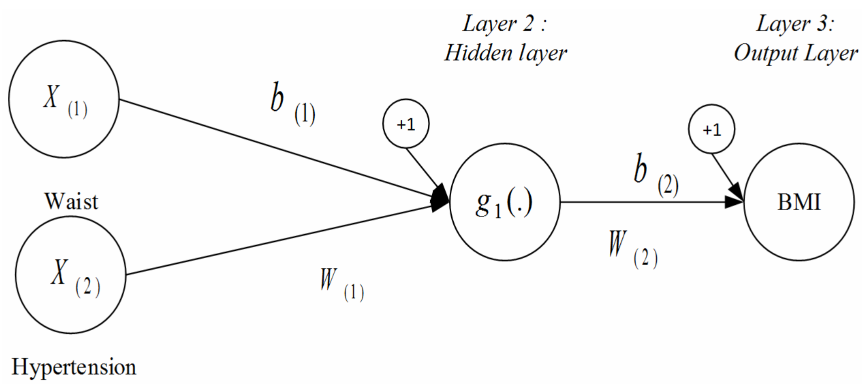

Figure 2. The architecture of the best MLFF neural network model with two input variables, one hidden layer, and one output node.

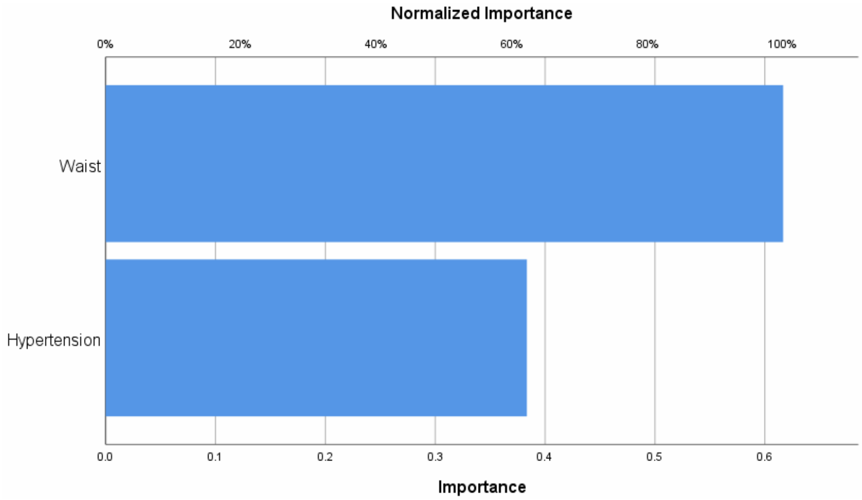

Figure 3. Arrangement of the predictor significant towards Y using multilayer feed forward neural network.

Table 3. Independent variable importance

Input	Importance	Normalized importance
Waist	0.681	100.0%
Hypertension	0.319	46.9%

According to Figure 1, the result of MLFF neural network was evaluated through the predictor significance. All the related variables are ranked from the most important to the least important using probability calculation,

Table 4. The testing and training performance of the MLFF neural network model

	Training phase		Testing phase	
Method	SSError	RError	SSError	RError
MLFF neural network	0.097	0.230	0.076	0.511

SSError: Sum of squares error

RError: Relative error

In this study, the output node corresponds to a single dependent variable, Y. To determine the optimal number of hidden nodes and the best combination of input variables, the model was constructed following the guidelines of MLFF neural network to ensure a good fit with the data. The optimal number of hidden nodes was found to be one. The architecture of the MLFF neural network model for Y consists of two input variables, one hidden node, and one output node, as shown in Figure 3. The performance of the MLFF neural network was assessed using the sum of squared errors and the relative error of the testing/out-of-sample data [8].

Case II

Table 5. Data description of the selected variable in the study

Num.	Variables	Explanation of user variables
1.	BMI	BMI reading of selected patient
2.	Waist	Waist circumference of patient
		1 = Normal blood pressure
3.	Hypertension status	2 = Borderline blood pressure
		3 = Hypertension

Therefore, we require two indicator variables. Define

$$X_{21} = \begin{cases} 0, & \text{Otherwise,} \\ 1, & \text{Normal blood pressure,} \end{cases}$$

$$X_{22} = \begin{cases} 0, & \text{Otherwise,} \\ 1, & \text{Borderline blood pressure,} \end{cases}$$

where X_2 is the variable with two classes. Therefore, we require only one indicator variable.

A regression model without interaction is given as follows:

$$\text{BMI} = -8.088 + 0.367\ \text{Waist} - 1.624\ \text{Normal} - 2.081\ \text{Borderline}.$$

Table 6. Parameter estimate for final multiple linear regression analysis

	Estimate	**Std. error**	***t*-value**	***p*-value**	**Tolerance**
(Constant)	-8.09	4.10	-1.97	0.08	
Waist	0.37	0.04	9.98	0.00	0.89
Normal BP	-1.62	0.90	-1.80	0.11	0.75
Borderline BP	-2.08	0.95	-2.19	0.05	0.68

Multilayer feed forward neural network for Case II

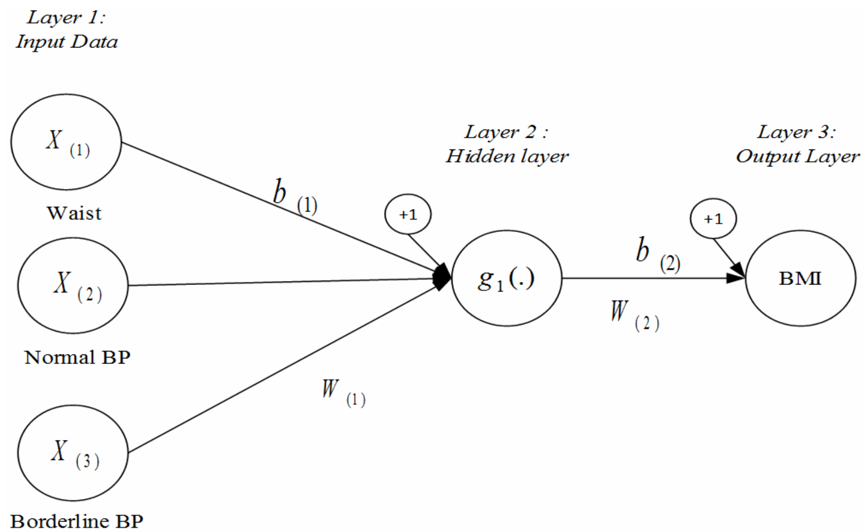

Figure 4. The architecture of the best MLFF neural network model with three input variables, one hidden layer, and one output node.

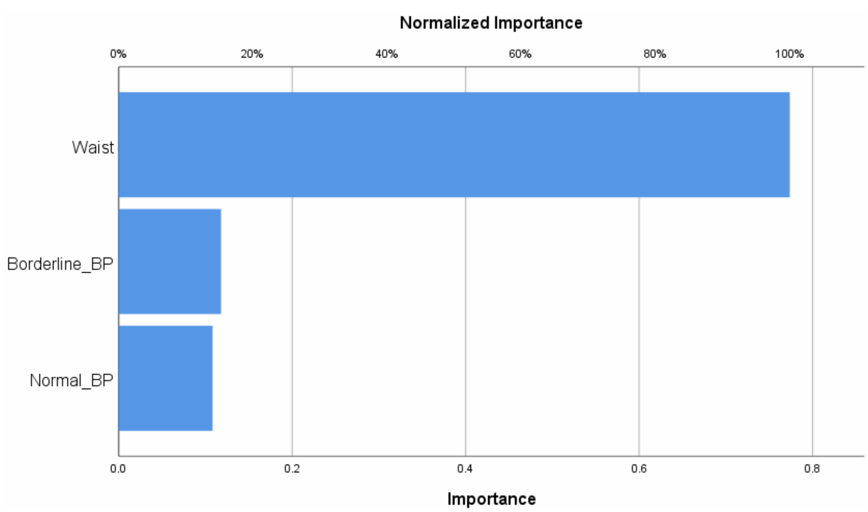

Figure 5. Arrangement of the predictor significant towards Y using multilayer perceptron neural network.

Table 7. Independent variable importance

Input	Importance	Normalized importance
Waist	0.774	100.0%
Normal BP	0.108	14.0%
Borderline BP	0.118	15.2%

According to Figure 1, the result of MLFF neural network was evaluated through the predictor significance. All the related variables are ranked from the most important to the least important using probability calculation.

Table 8. The testing and training performance of the MLFF neural network model

Method	Training phase		Testing phase	
	SSError	RError	SSError	RError
MLFF neural network	0.095	0.062	0.101	0.478

SSError: Sum of squares error

RError: Relative error

4. Conclusion

The primary purpose of this paper is to demonstrate the relationship of the fundamentals concept for multiple linear regression models with qualitative predictor variables techniques that can be employed to explain such relationships through multiple linear regression. It has been shown that a linear regression model with qualitative predictor variables can well fit according to the template in Case I and Case II. In the second phase of the study, all the independent variables and dependent variables in the regression model are further investigated by performing the MLFF neural network procedure. Using the continuous variables and qualitative variables of the linear regression model, the performance of MLFF neural network is tested for Case I and Case II. The performance of MLFF neural network was evaluated through the sum of squares error and relative error of testing/out-sample.

Acknowledgments

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] W. M. A. W. Ahmad, F. Ahmed and M. N. Adnan, Coordinating qualitative predictor variables in an applied linear model: analysis and application for applied sciences, Cureus Journal of Medical Science 16 (2024), e59151.

[2] W. M. A. W. Ahmad, M. N. Adnan, M. S. M. Ibrahim, N. A. Samsudin, N. F. M. Noor, N. A. Aleng and S. N. Hassan, Developing a hybrid linear model with a multilayer feed-forward neural network for HbA1c modeling among diabetes patients, Asian Journal of Fundamental and Applied Sciences 4(1) (2023), 41-49.

[3] Q. Zhang, Z. Li, S. Snowling, A. Siam and W. El-Dakhakhni, Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network, Water Science and Technology 80(2) (2019), 243-253.

[4] O. Üstün, E. Bekiroğlu and M. Önder, Design of highly effective multilayer feed forward neural network by using genetic algorithm, Expert Systems 37(4) (2020), 1-19.

[5] F. Li, Y. Lokhnygina, D. M. Murray, P. J. Heagerty and E. R. DeLong, An evaluation of constrained randomization for the design and analysis of group-randomized trials, Stat. Med. 35(10) (2016), 1565-1579.

[6] J. P. Kleijnen, Design and Analysis of Simulation Experiments, Springer International Publishing, 2018, pp. 3-22.

[7] W. M. A. W. Ahmad, M. S. M. Ibrahim, A. R. Hanafi, L. Puspa and N. A. Aleng, Algorithm for combining robust and bootstrap in multiple linear model regression, Journal of Modern Applied Statistical Methods 15(1) (2016), 884-892.

[8] L. Knoll, L. Breuer and M. Bach, Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning, Science of the Total Environment 668 (2019), 1317-1327.