INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 15, Issue 5, (2025) E-ISSN: 2222-6990

# Using Applied Statistics to Support Regional Planning: Statistical Approaches to Data Prediction: From Traditional Models to Modern Techniques in Analyzing Key Drivers (Household Income in Baghdad)

Salih Sufian Munther

Department of Management Economics Banking, College of Bussiness Economics, AlNahrain University, Baghdad, Iraq

Email: sufian.m.salih@nahrainuniv.edu.iq

**To Link this Article:** http://dx.doi.org/10.6007/IJARBSS/v15-i5/25529 DOI:10.6007/IJARBSS/v15-i5/25529

***Published Date:*** 19 May 2025

**Abstract**

"This study applies advanced statistical methods to support regional planning by comparing the performance of logistic regression and artificial neural networks in classifying observations related to the key factors influencing household income in the city of Baghdad. By accurately identifying these factors, the study provides data-driven insights that can guide effective policy-making, promote equitable resource allocation, and strengthen the foundations of sustainable regional development. “It is worth mentioning that, the logistic regression model is one of statistical models that used when the dependent variable is a dichotomous or polychotomous. It is a special case of linear regression model, and hence restrictively relevant in the sense that results obtained from it may be useless if linearity does not hold. On the other hand, "Artificial Neural networks" is a method of analysis based on both linear and non-linear relationships, which makes it more relevant in such circumstances. This study presents a real life application of these two methods in order to compare the performance of the two models.

**Keywords**: Logistic Regression Model, Artificial Neural Networks, Household Income

## Introduction

The process of classification between observations is a commonly used method, due to the multitude of phenomena that can be analyzed and interpreted through various classification techniques. Numerous methods and techniques have been used in the classification process. Artificial Neural Networks (ANNs) are considered one of the statistical methods used in the classification and separation of observations. They are also used in other applied aspects such as predicting investment behavior, detecting medical phenomena, simulating robotic operations, and image analysis, due to the flexibility of the neural network approach and its

ability to deal with nonlinear functions, as it does not depend on the type of probability distribution followed by the variables of the studied phenomenon. In this study, the researcher attempts to conduct a comparison between the logistic regression model and the neural network approach, in order to classify observations into their respective groups when some variables do not follow a normal distribution.

The applied aspect of the study examines the significant factors affecting The sufficiency of household income in Baghdad. The family has received the attention and care of researchers in various economic and social sciences, and this attention stems from the significant importance of the family in the system of economic and social relations.

*Study Problem*

With the multiplicity of statistical methods and techniques used in the process of classification and separation of observations, it was Essential to identify the Optimal approach that can Be applied in classifying observations to achieve the highest possible classification accuracy, taking into account the conditions of application of each method and its assumptions on the data under study. The percentage of correctly classified observations (correct classification rate) will be relied upon as a comparison criterion.

*Study Objectives*

- To identify the concept of artificial neural networks as one of the modern classification methods.
- To determine the statistical method that achieves the highest possible classification accuracy by comparing the logistic regression model and the neural network approach.
- To identify the most important factors that significantly affect the adequacy of family income in Baghdad.

*Study Importance*

- The need to use one of the modern classification methods that do not require specific conditions for use, such as the artificial neural network approach.
- Applying the logistic regression model as well as the neural network approach to some economic applications (family income adequacy) gives importance to the study, especially since these models are commonly used in medical and social studies.
- Studying the most important factors affecting the adequacy of family income in Baghdad from a statistical point of view, while determining the extent of the impact and significance of these factors, contributes to the development of comprehensive development plans and the achievement of economic and social stability for the family.

*Study Hypotheses*

In light of the study problem and after reviewing the results of previous studies, the study hypotheses can be formulated as follows:

The use of artificial neural networks in the classification process between observations achieves high efficiency.

Relying on the use of the neural network approach in the classification process when the independent variables do not follow a normal distribution achieves a higher classification rate than the logistic regression approach.

There is no statistically significant relationship between the adequacy of family income as a dependent variable and each of the independent variables (number of family members, average monthly family income, employment status of the head of the family, educational level of the head of the family, housing type, presence of university students in the family, and presence of family members with chronic diseases).

*Study Limitations*
Spatial Limitations: iraq - Baghdad City.
Temporal Limitations: From Jan. 2023 to Dec. 2024.

*Study Terms*
Family Income: The total income of all family members in cash, in kind, or services annually or at shorter intervals, including all forms Of earnings, encompassing salaries, hourly wages, pensions,, and government financial aid, and investment gains. Family income is considered a criterion for establishing the family's Quality of life.

The field study is the primary source of data for the study, where the study variables data was obtained through personal interviews with the heads of families (male or female) selected within the sample. The data was collected and the interviews were conducted by the researcher as the principal investigator, with the assistance of a research team that included a group of teaching assistants and postgraduate students.

| Allocation | Number of Households in Each District | Neighborhood |
|---|---|---|
| $n_1 = \frac{153000}{273213} \times 384 = 215$ | 153000 | Een Shams |
| $n_2 = \frac{36884}{273213} \times 384 = 52$ | 36884 | Al-Waley |
| $n_3 = \frac{43714}{273213} \times 384 = 61$ | 43714 | Al-Zaytoun |

*Study Population and Sample*
The study population consists of all families within the four districts (Ain Shams - El Waili - El Zeitoun - El Khalifa) that were randomly selected from the four regions of Baghdad (East - West - North - South). The total number of families in the four districts reached 273,213 families. The study involved determined at a significance level of 5% and a confidence Rate of 95%, resulting in 384 families. The sample size was increased by 77 families, representing 20% of the sample size, bringing the total sample size to 461 families. This increase was to account for non-response or the absence of some families during interviews. The response rate from the studied families was 89.8%, equivalent to 414 families. 30 families were allocated as a pilot sample, and 384 families as the main study sample. The study sample of 384 was distributed proportionally among the four districts according to the number of families within each district, as follows:

Table (1)
*Distribution of Sample Size Across the Four Districts*

| | | |
|---|---|---|
| Al-Khaleafa | 39615 | $n_4 = \frac{39615}{273213} \times 384 = 56$ |
| Total | 273213 | n = 384 |

Source: Prepared by the researcher According to data from the Central Agency For the Directorate of Public Mobilization and Statistics in 2023.

Table (2)
*Shows the distribution of sample members based on to gender, Life stage, and marital status*

| Districts | Gender | | Age | | | | Social Status | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Less than 20 | 20 - 30 | 30 -40 | 40 and above | Single | Married | Divorced | Widowed |
| Een Shamas | 168 | 47 | 10 | 56 | 60 | 89 | 11 | 143 | 36 | 25 |
| Al-Waley | 41 | 11 | 2 | 11 | 16 | 23 | 3 | 35 | 9 | 5 |
| Al-Zaytoun | 47 | 14 | 4 | 10 | 22 | 25 | 4 | 36 | 12 | 9 |
| Al-Khaleafa | 44 | 12 | 2 | 9 | 20 | 25 | 3 | 35 | 12 | 6 |
| Total | 300 | 84 | 18 | 86 | 118 | 162 | 21 | 249 | 69 | 45 |
| Percentage | 78,1% | 21,9% | 4,7% | 22,4% | 30,7% | 42,2% | 5,5% | 64,8% | 17,9% | 11,8% |

Source: Prepared by the Researcher

The sample size was calculated according to the following equation

$$n = \frac{z^2 p(1 - p)}{\epsilon^2} = \frac{(1.96^2)(0.5)(0.5)}{(0.5^2)} = 384$$

*Questionnaire Design*
The initial design of the questionnaire was based on previous studies and the researcher's experience, and the questionnaire was modified based on a pilot sample of (30) families to gather opinions on the possibility of adding, deleting, or modifying some phrases due to their lack of clarity. The researcher used the personal interview method during this stage, which resulted in the current form of the questionnaire, containing two parts. The first part contains general information about the respondent, and the second part contains eight questions related to the most important variables affecting family income adequacy.
Study Variables:
The study included one dependent variable and seven independent variables, as follows:

*Dependent Variable (Y)*
The dependent variable represents family income adequacy, which is a binary variable that takes the value (0) if the family income is insufficient, and the value (1) if the family income is sufficient. Independent Variables (X'S)

Since The primary goal of the aim of the study is to compare the use of the logistic regression model and the neural network approach in classifying observations, the independent variables consisted of only seven variables among those that can affect the dependent variable (family income adequacy). The selected variables are based on real-life observations and the opinions of many economists, namely:

| | | | |
|---|---|---|---|
| X1 | Number of Family Members (Household Size) (5) | 4 individuals or fewer = 0 | More than 4 individuals = 1 |
| X2 | Average Monthly Household Income (6) | 4000 pounds or less = 0. 4o mini | More than 4000 pounds = 1 |
| X4 | Occupation status of the household head | Does not work = 0 | Working = 1 |
| X4 | Highest level of education of the household head | Uneducated = 0 | Educated = 1 |
| X5 | Type of Housing | Rent = 0 | Owner = 1 |
| X6 | Families with university students | None = 0 | Exists = 1 |
| X7 | The presence of individuals in the family with a chronic illness | None = 0 | Exists = 1 |

*Statistical Models Used in the Study*
The study relied on the use of:
First: Artificial Neural Networks.
Second: Logistic Regression Model.
The following is a simplified presentation of the theoretical background of each method separately.

*First: Artificial Neural Networks (ANN)*
Artificial Neural Networks are a field of artificial intelligence, representing mathematical formulas based on models that simulate the human brain's problem-solving and computational processes. They are also known as parallel distributed processing systems, connectionist systems, or adaptive systems.

*Components of ANN*
An ANN consists of interconnected processing units called neurons. The output of one neuron serves as the input to another. Neurons are arranged in layers: the input layer, hidden layers, and the output layer. Connections between neurons are weighted, representing the strength of the connection (wij). (Abdul Aal, 2004)

*Types of Artificial Neural Networks*
ANNs can be classified based on data flow: Feed Forward Networks, Feed Back Networks, and Self-Organizing Networks. (Issa, 2000)

*Structural Design of Artificial Neural Networks: (Al-Abbasi, 2013)*
The structural design of building Artificial Neural Networks (ANN) includes the following steps:
- Data Collection: Gathering the data used to train or select the network.
- Data Definition: Defining the data for network training and establishing a training and learning plan.
- Structure Building: Building the network structure and determining the type of network and its components in terms of the number of layers.
- Learning Method Selection: Choosing a learning method based on available network development tools.

- Weight Value Setting: Setting values for weights and variables, then adjusting weight values through backpropagation.
- Data Conversion: Converting data to the appropriate type for the network through equations.
- Training and Testing: Conducting training and testing by repeatedly presenting the desired inputs and outputs to the network, comparing actual values with calculated values, calculating the error, and adjusting weights to reduce the error until it becomes acceptable.
- Result Achievement: Achieving the targeted results by using training inputs, allowing the network to be used as an independent system or as part of a system.

*Essential Elements for Building a Neural Network*

(a) Data:

The data (observations) are divided into three categories: training set, validation set, and finally the test set.

(b) Input Functions and Transfer Functions (Summation Function):

Input Functions (Summation Function):

The first operation performed by a neuron is to calculate the input value using the following summation function:

$s_j = \sum_{i=1}^{n} x_i w_{ij}$.....(1)

Where:
Sum_j: The result of the summation process for neuron (j).
Out_i: The outputs coming from neuron (i) and directed to neuron (j).
Wij: The weight connecting neuron (j) to neuron (i).
Transformation Functions:

$F(x) = \frac{1}{1+e^{-s}}$.....(2)

After the summation process, the process of converting the summation result to one of the values that should fall within the desired network outputs begins. This step is done using a function called the transformation function. The logistic function is considered one of the most used transformation functions, where the outputs are numbers confined between zero and one, and takes the following form:

$s = \sum_{i=1}^{n} x_i w_i + \theta$.....(3)

Other transformation functions include the linear function, step function, and sign function.

*Learning Methods (Training Methods)*

Learning methods are used to give the network the ability to learn until it reaches the targeted outputs with the least error. There are two types of learning: supervised learning and unsupervised learning.

*Learning Parameters*

Learning parameters refer to the tools used to improve the performance of the neural network. There are three learning parameters:

Learning Rate: Represents the limits of weight adjustment. The higher the learning rate, the greater the network's ability to learn. Momentum Factor: Represents the bias ratio in weights from one stage to another. To obtain a stable neural network, this factor must be less than one.

Training Tolerance: This value represents the allowable error during the comparison between the network outputs and the actual outputs. If this value reaches zero, it means that the network outputs are identical to the actual outputs. If the value increases, it means a decrease in the accuracy of the predictions. This factor is determined by trial and error and based on the researcher's experience and the nature of the data.

*Characteristics of Artificial Neural Networks*

The most important characteristics of artificial neural networks are:

- Independence from Assumptions: Neural networks are applied regardless of whether certain assumptions about the nature of the variables used in the analysis and the nature of their relationships with each other are met. This justifies the use of neural networks over other traditional statistical methods. (Abdul Aal, 2004)
- Wide Range of Applications: The possibility of applying neural networks in many scientific fields, such as economic and financial fields, as well as the field of the stock market and investment, such as predicting financial failure, predicting stock and bond prices, lending risks, etc., as well as their use in vital and medical fields.
- Alternative Analysis Tool: Neural networks can be used as an analysis tool instead of the following statistical methods: classification (discrimination), cluster analysis, prediction.
- Strong Mathematical Basis: Neural networks are built on a strong mathematical basis and deal with quantitative and qualitative data.
- Statistical Inference: Statistical inference is evident in neural network models through the training or learning process.

Second: Binary Logistic Regression Model

$$y|x = \hat{B}_o + \hat{B}_1 x \text{ .....(4)}$$

Logistic regression is used when the dependent variable (Y) is binary, taking the value (1) if the event of interest occurs with a probability of (P), and the value (0) if the event of interest does not occur with a probability of (1-P). Logistic regression does not impose restrictions on the types of independent variables (Xs), which can be continuous, categorical, or a mixture of both (Lea, 1997; Pample, 2000; King, 2003). It is known that the simple linear regression equation is in the form y|x, where y|x means the dependent variable Y given the independent variable X. Assuming that the random error (e) follows a normal distribution with a mean of (0) and a standard deviation of σ, then the dependent variable Y follows a normal distribution with a mean of β0 + β1X and a standard deviation of σ, i.e., Y ~ N(β0 + β1X, σ^2) for each value of the independent variable X.

$E(ylx) = \hat{B}o + \hat{B}_1 x$.....(5)

Since E(e) = 0, the expected value of the variable Y at a specific value of the variable x is in the following form:

$$\log e\left\langle \frac{P}{1-p} \right\rangle = \hat{B}_0 + \hat{B}_1 x$$

Where : E (e) = 0.....(6)

$$\frac{P}{1-p} = e^{\hat{B}_0 + \hat{B}_1 x}$$

However, due to the inability to apply simple linear regression when the dependent variable is categorical, the logistic model is used to address the previous problem. It can be written in the case of one independent variable as follows:

$$\log_e \left\langle \frac{P}{1-p} \right\rangle = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \cdots + \hat{B}_k x_k$$

.....(7)

Where:

P: The probability of the event of interest occurring, i.e., the probability of success.
1-P: The probability of the event not occurring, i.e., the probability of failure.
P / (1-P): The odds ratio of the event of interest.
logit(P): The natural logarithm of the odds ratio.
In another form:

$$\frac{P}{1-P} = e^{\hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \cdots + \hat{B}_k x_k}$$

.....(8)

Thus, the regression equation can be written in the case of (k) independent variables as follows:

The parameters of the logistic model are estimated using the Maximum Likelihood method, and then the model evaluation stage begins after the model is estimated (Hosmer & Lemshow, 2000). There are two methods to verify the suitability of the model, which are:

**A - Goodness-of-Fit Tests:**

This is done through the Likelihood Ratio Test, Hosmer-Lemeshow test, Classification Table, and ROC (Receiver Operating Characteristic) curve analysis.

**B - Significance Tests of Coefficients:**

The Wald Test is usually used to test the significance of each independent variable individually and to clarify the degree of importance of each independent variable. The higher the Wald statistic, the more important the variable, and vice versa (King, 2002; Menard, 2002).

*Statistical Analysis of Study Data*

The study data was analyzed using SPSS V24 statistical software, and the analysis concluded with the following results:

The validity and reliability of the questionnaire were verified through Cronbach's Alpha test. The results confirmed that the alpha value exceeded 0.70, which means that the questionnaire has a high degree of validity and reliability.

The Normality Test was performed for all independent variables. The Kolmogorov-Smirnov results indicated that (Sig=0.000 < 0.05) for all independent variables, which means that the independent variables do not follow a normal distribution.
The percentages of the study variables are shown in Table (3).

Table (3)
*Relative Distribution of Study Variables*

| Variable | Classification | Percentages |
|---|---|---|
| Number of household members X1 | (4) individuals or less | 47.7% |
| | More than (4) individuals | 52.3% |
| Average monthly household income X2 | 4,000 pounds or less | 51% |
| | More than 4,000 pounds | 49% |
| Employment status of household head X3 | Unemployed | 13.5% |
| | Employed | 86.5% |
| Educational level of household head X4 | Uneducated | 30.5% |
| | Educated | 69.5% |
| Housing type X5 | Rent | 66.4% |
| | Owned | 33.6% |
| Presence of university students in the household X6 | None | 62.5% |
| | Available | 37.5% |
| Presence of family members with chronic illness X7 | None | 77.9% |
| | Available | 22.1% |
| Average monthly household income adequacy Y | Insufficient | 70.8% |
| | Sufficient | 29.2% |

Source: Prepared by the research

*The correlation coefficient matrix between each pair of study variables was calculated, as follows:*

Correlation matrix between study variables

| Varlables | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|---|---|
| y | 1 | -.136 | .575 | .187 | .051 | .429 | -.047 | -.121 |
| X1 | -.136 | 1 | .066 | .115 | -.054 | .026 | .356 | .166 |
| X2 | .575 | .066 | 1 | .205 | .230 | .473 | .188 | -.070 |
| X3 | .187 | .115 | .205 | 1 | .135 | .136 | .008 | -.027 |
| X4 | .051 | -.054 | .230 | .135 | 1 | .267 | -.013 | -.042 |
| X5 | .429 | .026 | .473 | .136 | .267 | 1 | .019 | -.060 |
| X6 | -.047 | .356 | .188 | .008 | -.013 | .019 | 1 | .002 |
| X7 | -.121 | .166 | -.070 | -.027 | -.042 | -.060 | .002 | 1 |

It became clear from the correlation coefficient matrix that there was no significant correlation between the independent variables. On the other hand, the value of the determinant of the correlation matrix between the independent variables reached = 0.529, i.e. it is not equal to zero, which indicates that there is no problem of multicollinearity. To confirm this, the variance inflation factor was calculated for the independent variables, and it was as shown in Table No. (4).

Table No. (4)
*Variance Inflation Factor (VIF) Values*

| Variables | Tolerance | VIF |
|---|---|---|
| X1 | 828. | 1.208 |
| X2 | 714. | 1.401 |
| X3 | 932. | 1.073 |
| X4 | 902. | 1.109 |
| X5 | 744. | 1.344 |
| X6 | 833. | 1.200 |
| X7 | 961. | 1.040 |

The table shows a decrease in the variance inflation factor (VIF) values for all independent variables, as none of the VIF values exceeded (10), confirming the absence of multicollinearity.

First: Results of applying the artificial neural network (ANN) method

*Characterization of the Neural Network Model*
Since the objective of the study is to classify observations, the Step Function will be used as one of the transformation functions. This is because it is suitable for the classification process between observations, as it yields only two results (0) and (1), as shown by the following formula:

……..(9)

$$f(x) = \begin{cases} 0 \text{ when}, & x \leq t \\ 1 \text{ when}, & x > 0 \end{cases}$$

The formula for the function or model becomes:

$$F(x) = \frac{1}{1 + e^{-s}}$$

$$S = \sum_{i=1}^{n} x_i w_i + \theta$$

Where:
$w_i$ weights (relative importance of variables), $x_i$: independent variables, $\theta$: bias term. Thus, the model takes the following form:

$$F(x) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+\cdots+w_7x_7+\theta}} \qquad \text{……..(10)}$$

*Network Training and Testing*

Table No. (5) shows a summary of the number of cases used in the network training and testing processes.

Table No. (5)

*Number of cases used in the network training and testing processes*.

| Case Processing Summary | | | |
|---|---|---|---|
| | | N | Percent |
| Sample | Training | 258 | %67.2 |
| | Testing | 126 | %32.8 |
| Valid | | 384 | %100.0 |
| Excluded | | 0 | |
| Total | | 384 | |

The previous table shows that (258) observations, representing 67.2% of the total observations, were used in the network training process, and (126) observations, representing 32.8% of the total observations, were used in the network testing process.

*Neural Network Information*

Table No. (6) shows the information of the neural network used, which consists of three parts:

Table No. (6)

*Neural network information*

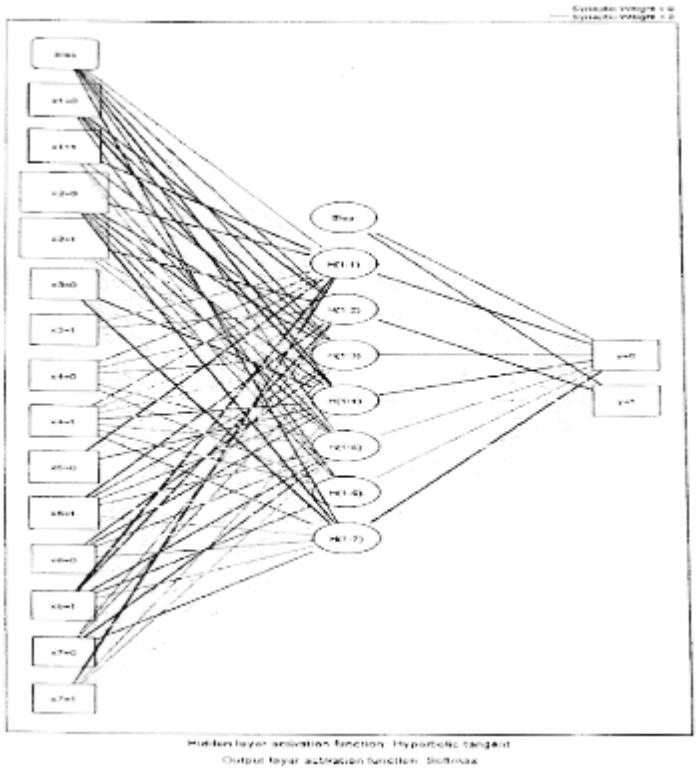
| Network Information | | | |
|---|---|---|---|
| Input Layer | Factors | 1 | Number of family members |
| | | 2 | Average monthly household income |
| | | 3 | Employment status of household head |
| | | 4 | Educational level of household head |
| | | 5 | Housing type |
| | | 6 | Presence of university students in the household |
| | | 7 | Presence of family members with chronic illness |
| | Number of Units' | | 14 |
| Hidden Layer(s) | Number of Hidden Layers | | 1 |
| | Number of Units in Hidden Layer 1[a] | | 7 |
| | Activation Function | | Hyperbolic tangent |
| Output Layer | Dependent Variables | 1 | Sufficient income. |
| | Number of Units | | 2 |
| | Activation Function | | Softmax |
| | Error Function | | Cross-entropy |
| a. Excluding the bias unit | | | |

Part One: Shows the information of the input layer represented by (7) independent variables, each variable has two levels (0), (1), and thus the number of units within the input layer is (14) units.

Part Two: Shows the information of the hidden layer (Hidden Layers) represented by a single layer. The number of units within the hidden layer is (7) units, and the activation function used is hyperbolic tangent.

Part Three: Shows the information of the output layer represented by a single dependent variable, which is the adequacy of household income (sufficient - insufficient). Therefore, the number of units processed in this layer is (2) units. The activation function used in this layer is Softmax, also known as the sigmoid function.

Figure (1) shows the hierarchical structure (architecture) of the network used, which consists of the input layer, located on the left, consisting of (14) units, in addition to the bias unit. The second layer is the hidden layer, located in the middle, consisting of (7) units, in addition to the bias unit. The final layer, located on the right, is the output layer (output). It shows that there are two outcomes: (insufficient income = (0), sufficient income = (1)



*Summary of the Neural Network Model*

Table No. (7) shows a Overview of the neural network model employed

Table No. (7)

*Overview of the Neural Network Model*

| | | Model Summary |
|---|---|---|
| Training | Cross Entropy Error | 70,523 |
| | Percent Incorrect Predictions | %12.0 |
| | Stopping Rule Used | consecutive step(s) with no decrease in error* |
| | Training Time | 0:00:00.45 |
| Testing | Cross Entropy Error | 35.117 |
| | Percent Incorrect Predictions | %11.9 |
| Sufficient income Dependent Variable | | |
| .a. Error computations are based on the testing sample | | |

The previous table shows the following:

- The error rate in the training dataset was 12%, while The classification error rate Within the test dataset was 11.9%, which is very close. This indicates that the network was well trained to classify observations.
- The stopping rule used for the network is when the error rate becomes constant, i.e., when the error rate stops increasing.
- The network training time is (45) seconds.

*Classification Results*

Table (8) Presents the classification outcomes using the proposed neural network.

Table (8)

*Classification Results Using the Neural Network*

| | | Classification | | |
|---|---|---|---|---|
| Sample | Observed | Predicted | | |
| | | Not enough | Enough | Percent Correct |
| Training | Not enough | 161 | 20 | %89.0 |
| | Enough | 11 | 66 | %85.7 |
| | Overall Percent | %66.7 | %33.3 | %88.0 |
| Testing | Not enough | 81 | 10 | %89.0 |
| | Enough | 5 | 30 | %85.7 |
| | Overall Percent | %68.3 | %31.7 | %88.1 |

The following is evident from the previous table:

- The correct classification of insufficient income was 89% in the training sample, while it was also 89% in the test sample.
- The correct classification of sufficient income was 85.7% in the training sample, while it was also 85.7% in the test sample.

- The correct classification rate of observations using the proposed neural network was 88.1%, which represents a very good rate for predicting the classification of new observations. The above confirms the validity of the study's first hypothesis, namely the possibility of using artificial neural networks in the classification process with high efficiency.

*Relative Importance of Study Variables*
Table (9) shows the relative importance of the independent variables.

Table (9)
*Relative significance of independent variables*

| **Importance of Independent Variables** | | |
|---|---|---|
| | Significance | Standardized importance |
| X1 | 159. | %65.0 |
| X2 | 245. | %100.0 |
| X3 | 128. | %52.4 |
| X4 | 132. | %53.8 |
| X5 | 140. | %57.0 |
| X6 | 101. | %41.3 |
| X7 | 094. | %38.3 |

The above table shows the following:
- The most influential variable in classifying observations using the proposed neural network is the average monthly household income (X2), with an importance rate of 24.5%, which is largely consistent with reality. This is followed by the number of household members (X1) with an importance rate of 15.9%, followed by the housing type (X5) with an importance rate of 14%, followed by the educational level of the household head (X4) with an importance rate of 13.2%, followed by the employment status of the household head (X3) with an importance rate of 12.8%, followed by the presence of university students in the household (X6) with an importance rate of 10.1%, and the least influential variable in classifying observations is the presence of family members with a chronic illness (X7) with an importance rate of 9.4%.
- The last column (Normalized Importance) represents the relative importance of each independent variable to the most influential variable in classifying the observations, which is (X2). For example, we find that (X1) represents 65% of the relative importance of variable (2) 65.0 = (159.245), and so on for the remaining variables.
- The appropriate transformation function for the study data when using the neural network model is the logistic function, as it yields only two results (10), and takes the following form:

$$F(x) = \frac{1}{1 + e^{-(.159x1+.245x2+128x3+132x4+.140x5+.101x6+.094x7)}}$$

Second: Results of applying the binary logistic regression model
(1) Description of the binary logistic regression model

Since the estimated logistic model equation takes the following form:

$$\frac{\hat{p}(x)}{1 - \hat{p}(x)} = e^{\hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \ldots + \hat{B}_k x_k}$$

Since we have seven independent variables for the phenomenon under study, the model takes the following form:

$$\frac{\hat{p}(x)}{1-\hat{p}(x)} = e^{\hat{B}_0+\hat{B}_1x_1+\hat{B}_2x_2+\hat{B}_3x_3+\hat{B}_4x_4+\hat{B}_5x_5+\hat{B}_6x_6+\hat{B}_7x_7} \ldots\ldots\ldots\ldots(11)$$

(2) Data analysis using the logistic model

The results of the Backward Stepwise Likelihood Ratio method were used, and the results were as follows: Statistical tests indicated the significance of the model as a whole, using the Likelihood Ratio Test and the Hosmer-Lemeshow Test, as shown in the following tables:

Table No. (10)

*Results of the significance tests for the model as a whole*

| | LL2- | Chi-Square | P value |
|---|---|---|---|
| Model | 264.507 | 199.086 | 0.000 |

It is clear from the table that the value of (0.000<0.05 = P_value) confirms that the model as a whole is significant and represents the data well.

Table No. (11)

*Hosmer-Lemeshow test results*

| | Chi-Square | d.f | P_value |
|---|---|---|---|
| Model | 14.399 | 8 | 0.072 |

The table shows that the P-value is 0.072 > 0.05, confirming that there is no difference between the actual observation values and the estimated values, indicating the model's good fit.

Figure (2), which represents the ROC curve, shows that the model performs better at classifying observations than the chance factor. The curve appears to move away from the diameter of the chance, which encloses 50% of the area, to provide a larger area than the chance factor covers. The efficiency of the model-based classification should exceed 72%.

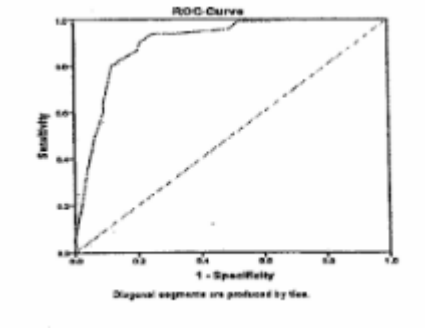



Figure (2) ROC curve

The following table shows the value of the area under the ROC curve for the fitted model.

Table No. (12)
*Area beneath the ROC curve*

| Area | S.E | Asymptotic Sig | Asymptotic 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower Bound | Upper Bound |
| 0.900 | 0.016 | 0.000 | 0.868 | 0.932 |

The table shows that the value of the area under the curve is equal to 0.90 at a significance level of 0.000% with a confidence interval of 95% (0.868 - 0.932), which means that the model helps in predicting the classification of the dependent variable cases more than chance does.

Statistical tests indicated the significance of only five independent variables, which represent the most important variables affecting the adequacy of household income in Baghdad. These variables were as follows:
X1: Number of household members X2: Average monthly household income
X4: Educational level of the household head X5: Housing type
X6: Presence of university students in the household
The significance of these variables was found at a 5% significance level. Therefore, X3 and X7 were excluded, as shown in the following table:

Table No. (13)
*Results of applying the logistic regression model*

| variables | B | S.E | Wald | d.f | Sig | Exp(B) | %95C.I.for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| X1 | -1.687- | 407. | 17.160 | 1 | 000. | 185. | 083. | 411. |
| X2 | 4.057 | 64.221 | 506. | 1 | 000. | 57.815 | 21.433 | 155.951 |
| X4 | -1.592- | 421. | 14.305 | 1 | 000. | 204 | 089. | 464. |
| X5 | 1.401 | 339. | 17.051 | 1 | 000. | 4.060 | 2.088 | 7.894 |
| X6 | -764.- | 331. | 5.314 | 1 | 021. | 466. | 243. | 892. |
| Constant | -2.447- | 417. | 34.413 | 1 | 000. | 087. | | |

The significance of the variables' coefficients was tested, and their significance was demonstrated using the Wald test as well as the Sig value, as the value (0.05 > Sig) for the five variables. In light of the Wald test, we find that the average monthly income of the family (X2) is the most influential and important variable in classifying the family income as sufficient or insufficient, followed by the variable related to the number of family members (X1), then the variable of housing quality (5X), then the variable related to the educational level of the head of the family (4), and finally the variable of the presence of university students in the family (6X), as the higher the value of (Wald), the more important the variable is. The results indicated the lack of significance of the variable related to the functional status of the head of the family (3X), as well as the variable related to the presence of family members with a chronic disease (X7), and thus they were deleted from the model.

The value of the odds ratio (Exp(B) odds ratio) refers to the value of the exponential function of the regression coefficient. It expresses the multiplier by which the odds ratio (the value of the dependent variable) changes, i.e. the change from probability (1 = Y) to probability (0 = Y). It is calculated from the formula:

$$EXP(B) = e^{Bi} = e^{-1.687} = 0.185$$

This means that as the value of the independent variable (X1) increases, the probability of income adequacy decreases by approximately 18.5%, and so on for the remaining variables. The last column in the table represents the confidence limits for the calculated exponential function value.

Table (14) indicates that the use of logistic regression achieved an overall correct classification rate of 85.7%. This is a high Percentage, confirming that the model aligns well with the data

Table No. (14)
*Classification Table*

| Observed | | Sufficient income | | Percentage Correct |
|---|---|---|---|---|
| | | Not enough | Enough | |
| Sufficient income | Not enough | 239 | 33 | 87.9 |
| | Enough | 22 | 90 | 80.4 |
| Overall Percentage | | | | 85.7 |

From all of the above, we can conclude that the logistic regression model used to classify household income in Baghdad as sufficient or insufficient takes the following form:

$$\text{Log}(\frac{\hat{p}}{1-\hat{p}}) = -2.447 - 1.687\ X1 + 4.057\ X2 - 1.592\ X4 + 1.401\ X5 - 0.764\ X6$$

**Conclusions and Recommendations**

*Conclusions*

Through analysis using the proposed models, it was found that the use of artificial neural networks provides a better classification rate than the logistic regression model. The correct classification rate using neural networks was 88.1%, while the correct classification rate using the logistic regression model was 85.7%. The significance of the variables' coefficients was tested, and their significance was demonstrated using the Wald test as well as the Sig value, as the value (0.05 > Sig) for the five variables. In light of the Wald test, we find that the average monthly income of the family (X2) is the most influential and important variable in classifying the family income as sufficient or insufficient, followed by the variable related to the number of family members (X1), then the variable of housing quality (5X), then the variable related to the educational level of the head of the family (4), and finally the variable of the presence of university students in the family (6X), as the higher the value of (Wald), the more important the variable is. The results indicated the lack of significance of the variable related to the functional status of the head of the family (3X), as well as the variable related to the presence of family members with a chronic disease (X7), and thus they were deleted from the model

The results of the used models coincided in terms of the importance of the independent variables that significantly affect the observation classification process. In both models, the monthly average family income (X2) was found to be the most influential variable in the classification process, followed by the number of family members (X1), then the housing type (X5), and then the educational level of the head of the family (X4). The results of the logistic regression model indicated that the variable related to the employment status of the head of

the family (X3) and the variable related to the presence of family members with chronic diseases (X7) were not significant.

Despite the ability of neural network models to classify observations with high efficiency, they do not allow us to perform statistical tests related to the significance of the relationship, while this is available when using the logistic regression model.

The statistical model proposed for classifying family income in Baghdad in terms of being sufficient or insufficient when using the logistic regression model takes the following form:

$$\text{Log}(\frac{\hat{p}}{1-\hat{p}}) = -2.447 - 1.687\ \text{X1} + 4.057\ \text{X2} - 1.592\ \text{X4} + 1.401\ \text{X5} - 0.764\ \text{X6}$$

5- The transformation function used in classifying family income in Baghdad in terms of being sufficient or insufficient when using the neural network model is the logistic function and takes the following form:

$$F(x) = \frac{1}{1 + e^{-(159x_1+245x_2+128x_3+132x_4+140x_5+101x_6+094x_7)}}$$

*Recommendations*

If the objective of the observation classification process is to achieve the highest classification accuracy only, it is best to use the neural network approach. However, if the objective is to achieve the highest classification accuracy with the explanation and interpretation of the model parameters used, both regression and neural network models can be used together. Conduct more studies on the nature of the data and the field of application to compare different networks with other analysis methods.

Rely on the neural network approach instead of the logistic regression model if some independent variables do not follow a normal distribution or if the distribution they follow is unknown.

Expand the use of methods and techniques used in the classification process in economic and social fields and do not limit their use to medical fields only as was the case previously. Conduct more studies on family income and spending, introducing more independent variables affecting family income and spending, such as commodity prices and exchange rates, etc., as insufficient income directly affects the quality of education and health of family members and all other aspects of life, and can also affect the size and rates of crime and the level of national security.

**References**

Hamid, Al-A. A. (2013). "Introduction to Artificial Neural Networks and Their Applications in Social Sciences Using SPSS," Baghdad University, Institute of Statistical Studies and Research, Department of Biostatistics and Population Statistics.

Bassam, Al-G. H. (2004). "Using Financial Ratios to Predict Corporate Distress," Master's Thesis, Gaza Islamic University.

Al-Mutawa, B. M., Al-Aqil, A. H. (1996). "Using Discriminant Analysis and Neural Networks to Predict the Degree of Banking Client Reliability," Arab Journal of Administrative Sciences, Kuwait, Volume (3), Issue (2), pp: 315-295.

Naji, J. H. (2010). "Comparing Methods of Estimating Economic Functions with Qualitative Dependent Variables," Iraq, Tikrit Journal of Administrative and Economic Sciences, Volume (6), Issue (18).

Risan, D. T. (2008). "Using Neural Networks for Discrimination Purposes," Al-Qadisiyah University, College of Management and Economics, Department of Statistics, Volume (52/14), pp: 246 - 256.

Denosern, R. D. (1997) - Translated by Abdel Rahman Hamed Azam, "Multivariate Statistical Analysis from an Applied Perspective," Dar Al-Merikh Publishing, Riyadh.

Aal, A., Ahmed, M. M. (2004). "Neural Networks and Business Management Applications," Baghdad, Scientific Journal of Economics and Commerce, Issue 1, pp: 465 - 494.

Qader, A. A. (2004). "Modern Econometrics between Theory and Practice," Al-Dar Al-Jami'iya for Publishing and Distribution, Alexandria, Egypt.

Adam, A. F. (2016). "Using Linear Discriminant Function to Identify the Most Important Economic and Social Factors Affecting Household Income Sufficiency in North Kordofan State," Master's Thesis, College of Sciences, Sudan University of Science and Technology.

Zaki, I. A. (2000). "Neural Networks - Engineering Structure, Algorithms, and Applications," First Edition, Shuaa Publishing and Science, Syria, Aleppo.

Al-Jaouni, G. A. F. (2011). "Using Binary Logistic Regression in Studying the Most Important Economic and Social Determinants of Household Income Sufficiency - An Empirical Study on a Random Sample of Households in Damascus Governorate," Syria, Damascus University Journal of Economic and Legal Sciences, Volume (27), Issue 1.

Al-Jaouni, G. A. F. (2007). "Multivariate Statistical Analysis (Discriminant Analysis) in Describing and Distributing Households within the Socio-Economic Structure of Society," Syria, Damascus University Journal of Economic and Legal Sciences, Volume (23), Issue 2.

Al-Taher, N. A. (2015). "Classification and Analysis of Income Categories in Sudan Using Discriminant Function Compared to Neural Network Models," PhD Thesis, Sudan University of Science and Technology, Graduate College.

Central Agency for Public Mobilization and Statistics. (2016). "Annual Statistical Book."

Central Agency for Public Mobilization and Statistics. (2016). "Results of the Income, Expenditure, and Consumption Survey."

Bandy, H. (1994). "Thoughts on Desirable Features for a neural Network - Based Financial Trading System", NUROVST Journal, 2(3): 19-22

-Hosmer, W. & Lemeshow, S. (2000). "Applied Logistic Regression", 2nd edition New York: Johnson Wiley & Sons, Inc.

King, J. E. (2002). "Logistic Regression: Going beyond point-and-click", Paper Presented at the annual Meeting of the American Educational Research Association, New Orleans, LA, April.

King, J. E. (2003). "Running A Best-Subsets Logistic Regression: An Alternative to Stepwise Methods", Educational and Psychological Measurement, Vol. 63, No. 3, June, 392-403.

Krose, B. & Smagt, P. (1996), "An introduction to neural networks", Eighth edition, The University of Amsterdam.

Lea, S. (1997). "Multivariate Analysis II: Manifest Variables analysis Topic 4: Logistic Regression and Discriminant Analysis", University of EXETER Department of Psychology Available at: www.exeter. Ac.uk/_SEGL ea./multivar2/diclogi.htm1

Lea, S. (2004). "Application of Likelihood ratio and Logistic Regression models to Landslide susceptibility mapping using GLS". Environmental Management, Vol. 34, No. 2, 223-232

Menard, S. W. (2002). "applied Logistic Regression analysis", 2nd edition Sage Publication Series Quantitative Application in the Social Sciences, No. 106, Thousand Oaks, CA: Sage.

Pample, F. (2000). " Logistic Regression a Primer". Sage Publication Series Quantitative Application in the Social Sciences, No.07- 132, Thousand Oaks, CA: Sage.

Rahman, A. (2009)."statistical Analysis of the Different Socioeconomic Factors Affecting the Education of N.W.F.P (PAKISTAN), "PhD, Institute of Mathematical methods in Economic (EOS), University of Technolgy Vienna, Austria, Journal of Applied Quantitative Methods, Vol (4), No (1).

Tang, M. (2001). "Exact Goodness of Fit test for Binary Logistic Model". Statistica Sinica (11), The Chinese University of Hong Kong.

Yoon, Y., Swales, G. (1993). "A Comparison of Discriminant Analysis Versus Artificial Neural Network", Journal of Operational Research, 44: 51-60.