2.2. Multiple Linear Regression (MLR)

MLR is a mathematical model used for modeling the relationship between multiple predictor variables and one dependent variable. The model is limited to data that follow a linear function to yield unique estimation solution for regression coefficients (Marill, 2004). The objective of using MLR analysis is to get the best model the relationship between independent and dependent variables (Adamowski *et al.*, 2012).

Many researchers used multiple linear regression analysis and the general model of equation can be formulated as follows:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e_i \qquad (2.1)$$

where y_i is the dependent variable, $x_1, \mathrm{x}_2, \ldots, \mathrm{x}_p$ are independent variables, β_0 is the constant model, $\beta_1, \beta_2, \ldots, \beta_p$ are the parameters of regression model, and e_i is the amount of random error. Equation (2.1) can be written in matrix form as follows:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathrm{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \qquad (2.2)$$

Where Y is the size $(n \times 1)$ and the matrix X of the degree $(n \times (p+1))$ and the size of β is $((p+1) \times 1)$ and the degree ε is $(n \times 1)$.

Note that the first column in the data matrix contains the value of one at all observations from (1) to (n) to estimate the constant coefficient. Using matrix symbols, the linear regression model can be written as follows:

$$Y = X\beta + \varepsilon \qquad (2.3)$$

There are a number of assumptions (analysis assumptions) about the distribution variables in the MLR such as follows.

1. The average of random error term (ε_i) should be equaled zero.

$\mathrm{E}(e_i) = 0$ for all $i = 1, 2, 3, \ldots, n$

2. The variance of the random error term must be constant in all time periods. This hypothesis is called Homoscedasticity. But when the variance of the random errors is not constant, it is called Heteroscedasticity.

$Var(e_i) = \mathrm{E}(e_i^{\,2}) = \sigma^2$ for all $i = 1, 2, 3, \ldots, n$

3. The random error follows a normal distribution by zero mean and constant variance.

$\varepsilon_i \sim N(0, \sigma^2)$ for all $i = 1, 2, 3, \ldots, n$

4. The random variable (e_i) is independent of (x_i) variables. This means that the covariance of the random error and independent variable are equal to zero.

$COV(e_i, \mathrm{x}_i) = 0$ for all $i = 1, 2, 3, \ldots, n$

5. The covariance of any random errors in different lags such as $(e_{i,} e_j)$ is equals to zero $COV(e_i, \mathrm{e}_j) = 0$ $(i \neq j)\, i, j = 1, 2, \ldots, n$

   (Gunst, 2018)

2.1.1 Estimating the parameters in the least squares method

The use of the least squares method in estimating the parameters of the model is characterized by choosing the best model corresponding to the data so that the total number of squares of error or residual is as low as possible. Thus put the form as follows:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e_i \qquad (2.4)$$

$$e_i = y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \qquad (2.5)$$

And the two sides and then collect each party produces the function of the least squares:

$$Q(\beta_0 + \beta_1 + \beta_2 + \ldots + \beta_p) = \sum e_i^{\,2} = \sum \left[y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \right]^2 \qquad (2.6)$$

When you take the partial derivative with respect to β_0 and β_j of the function and make it zero:

$$\frac{\partial Q}{\partial \beta_0} = -\sum\left(y_i - \beta_0 - \sum \beta_j x_{ij}\right) = 0 \tag{2.7}$$

$$\frac{\partial Q}{\partial \beta_j} = -2\sum\left(y_i - \beta_0 - \sum \beta_j x_{ij}\right)x_{ij} = 0 \tag{2.8}$$

After simplification we get the following natural equation:

$$\begin{aligned}
& n\beta_0 + \beta_1 \sum x_{i1} + \beta_2 \sum x_{i2} + \ldots + \beta_p \sum x_{ip} = \sum y_i \\
& \beta_0 \sum x_{i1} + \beta_1 \sum x_{i1}{}^2 + \beta_2 \sum x_{i1} x_{i2} + \ldots + \beta_p \sum x_{i1} x_{ip} = \sum x_{i1} y_i \\
& \vdots \\
& \beta_0 \sum x_{ip} + \beta_1 \sum x_{ip} x_{i1} + \beta_2 \sum x_{ip} x_{i2} + \ldots + \beta_p \sum x_{ip}{}^2 = \sum x_{ip} y_i
\end{aligned} \tag{2.9}$$

These natural equations can be written using matrices follows:

$$\begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \cdots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}{}^2 & \sum x_{i1} x_{i2} & \cdots & \sum x_{i1} x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ip} & \sum x_{ip} x_{i1} & \sum x_{ip} x_{i2} & \cdots & \sum x_{ip}{}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \vdots \\ \sum x_{ip} y_i \end{bmatrix} \tag{2.10}$$

Note that it is easy to find natural equations when we use matrices .

The model in general is:

$$\begin{aligned} y &= x\beta + e \\ e &= y - x\widehat{\beta} \end{aligned} \tag{2.11}$$

The sum of the squares of error is

$$Q_i = e'\mathrm{e} = \left(y - x\beta\right)' \left(y - x\beta\right)$$

$$= y'y - \beta' x' y - y' x \beta + \beta' x' x \beta \tag{2.12}$$

Since the $y'x\beta = \beta' x' y$

So the:

$$Q = e'e = y'y - 2\beta'x'y + \beta'x'x\beta \tag{2.13}$$

To find the β that makes $\varepsilon'\varepsilon$ as low as possible we take the partial derivative for each β_i and then make it equal to zero:

$$\frac{\partial Q}{\partial \beta} = \begin{bmatrix} \frac{\partial Q}{\partial \beta_0} \\ \frac{\partial Q}{\partial \beta_1} \\ \vdots \\ \frac{\partial Q}{\partial \beta_p} \end{bmatrix} = -2x'y + 2x'x\beta = 0 \tag{2.14}$$

After simplification we find that:

$$X'X\beta = X'Y \tag{2.15}$$

Thus the estimations of the least squares are:

$$\widehat{\beta} = \left(X'X\right)^{-1} X'Y \tag{2.16}$$ (Yuan, 2007).

2.3. Logistic Regression Model (LR)

The MLR model is applicable when the dependent variable is continuous, and not categorical, while LR is different from the MLR and used when the dependent variable is binary and the independent variables are quantity, categorical variables, or both (Marill, 2004).

Simple and multiple linear regression assumes linear relationship between the independent variables and the dependent variable, while logistic regression does not assume that. It is a useful tool for modeling and forecasting data that consists of binary dependent variable. Nowadays, the researchers state the fact that it is not appropriate to propose multiple linear regression for a binary dependent variable and the better choice institute of multiple linear regression is LR for accurate modeling and forecasting. Categorical events will be coded as binary variables with a value of one represents the