Mathematics and Statistics 8(2A): 36-39, 2020
DOI: 10.13189/ms.2020.081306

Robust Method in Multiple Linear Regression Model on Diabetes Patients

Mohd Saifullah Rusiman[1,*], Siti Nasuha Md Nor[1], Suparman[2], Siti Noor Asyikin Mohd Razali[1]

[1]Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[2]Department of Mathematics Education, Universiti of Ahmad Dahlan, Indonesia
[*]Corresponding Author: saifulah@uthm.edu.my

Received August 3, 2019; Revised October 5, 2019; Accepted February 20, 2020

Abstract This paper is focusing on the application of robust method in multiple linear regression (MLR) model towards diabetes data. The objectives of this study are to identify the significant variables that affect diabetes by using MLR model and using MLR model with robust method, and to measure the performance of MLR model with/without robust method. Robust method is used in order to overcome the outlier problem of the data. There are three robust methods used in this study which are least quartile difference (LQD), median absolute deviation (MAD) and least-trimmed squares (LTS) estimator. The result shows that multiple linear regression with application of LTS estimator is the best model since it has the lowest value of mean square error (MSE) and mean absolute error (MAE). In conclusion, plasma glucose concentration in an oral glucose tolerance test is positively affected by body mass index, diastolic blood pressure , triceps skin fold thickness, diabetes pedigree function, age and yes/no for diabetes according to WHO criteria while negatively affected by the number of pregnancies. This finding can be used as a guideline for medical doctors as an early prevention of stage 2 of diabetes.

Keywords Multiple Linear Regression, Least Quartile Difference (LQD), Median Absolute Deviation (MAD), Least-Trimmed Squares (LTS), Mean Square Error

1. Introduction

Diabetes is defined as a disease in which the body's capability to produce or respond to the hormone insulin is reduced, resulting in unusual metabolism of carbohydrates and raised levels of glucose in the blood. Nowadays, diabetes is a common disease and is becoming more common. Age-adjusted prevalence is set to increase from 5.9% to 7.1% (246-380 million) worldwide in the 20-79 year age group [1]. 20% of case of diabetes among adults was in urban population and 10% was in rural population [2]. There are several factors that contribute to diabetes disease such as age, body mass index (BMI) and central adiposity, measured either as waist circumference (WC) [2]. Nowadays, people no longer practice physical activities but eat additional food with the high consumption of sugar causing high tendency for people in India to develop insulin resistance [3].

Multiple linear regression (MLR) model can be described as statistical approach to describe the association between two or more quantitative variables so that the dependent variable can be predicted from the others. MLR is widely used in business, the social and behavioural sciences and many other areas [4]. MLR needs the assumption of normally distributed variables and measurement errors necessarily causing underestimation of simple regression coefficients [5]. Robust regression is used to help in detection and deletion of outlier and it is an approach that provides estimation, inference and testing that are not influenced by outlying observations but described correctly the structure for the data [6]. The goal to use robust regression is to produce linear models that are not biased by few outliers [7]. There are other quite considerable studies carried out in statistical modelling [8, 9]. The objectives in this study are to identify the significant variables that affect diabetes by using multiple linear regression model, to apply the robust regression method on diabetes data and to measure the performance of robust method by comparing MLR model only and MLR model with robust method.

2. Materials and Methods

The data were collected from the US National Institute of Diabetes and Digestive and Kidney Disease webpage. It involved 332 women who were at least 21 years old of Pima India heritage and living near Phoenix, Arizona. There are 8 variables involved in this study which are 1 dependent variable and 7 independent variables. The dependent

variables are denoted by *y that* is plasma glucose concentration in an oral glucose tolerance test. This test is a common test used in general hospital in Malaysia and other countries. This test checks a standard dose of glucose ingested by mouth and blood levels that are checked two hours later where the normal reading should be below 6.1 mmol/L. While other 7 independent variables are denoted by x_1 (body mass index), x_2 (number of pregnancies), x_3 (diastolic blood pressure), x_4 (triceps skin fold thickness), x_5 (diabetes pedigree function), x_6 (age) and x_7 (yes or no for diabetes according to WHO criteria).

2.1. Multiple Linear Regression

Multiple linear regression (MLR) is one of the most commonly used of all statistical methods. MLR is known as predictive analysis that is used to explain the relationship between one dependent variable and two or more independent variables. Leona (2012) described multiple regression model as a linear regression model with two or more predictors and one response [10]. The model equation expresses the value of predictor variables as a linear model of two or more independent variables and the error terms as in (1) [4]:

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_k x_{i,k} + e_i \quad (1)$$

where,

y_i = dependent variable or real data
b_0 = constant of MLR
b_k = constant of the k^{th} independent
$x_{i,k}$ = independent variables
$e_i = y_i - \hat{y}_i$ = residual
$\hat{y}_i$ = estimated data from MLR model

Before the MLR can be done, two assumptions need to be fulfilled. Firstly, the normality test for *y* vs all $x_{i,k}$ should be done by using P-P Plot or by numerical calculation. Using P-P Plot, straight graph data indicate normality distribution. Next, multicollinearity test among $x_{i,k}$ variables should be done to identify the existence of multicollinearity in model that can affect the least square method accuracy of the estimated model. The VIFF value less than 10 indicates no multicollinearity among $x_{i,k}$ variables [4].

2.2. Robust Method

2.2.1. Least Quartile Difference (LQD) Method

LQD is a regression estimator which is highly robust since the LQD can resist up to 50% largely deviant data values without becoming too biased. Since LQD has almost 50% of breakdown point, LQD is expected to deal with unusual observation and should give the good performance when the data is not contaminated. LQD formula decreases the Q_1 of the $|\text{residual}_i - \text{residual}_k|$ as in (2) [11]. LQD estimator of β is defined by,

$$\hat{\beta}_{LQD} = argmin\ QD_n(e_1(\beta), \ldots, e_n(\beta)), \quad (2)$$

where

$$e = \text{residual}$$

$$QD_n(e_i, \ldots e_n) = \{|e_i - e_k|; 1 \le i < k \le n\}$$

In this method, the residual needs to be sorted first. Then, 25% of upper data and 25% of lower data need to be discarded. Then 50% of the remaining data need to be analysed using MLR model. This model should be applied to all data in order to find new MSE and MAE values**.**

2.2.2. Mean Absolute Deviation (MAD) Method

Median absolute deviation (MAD) is one of the most familiar robust measures. MAD is defined as the median of absolute values and overall median of the data set. MAD is also known as a robust measure of variability of univariate sample of quantitative data and also refers to parameter of population that is estimated by MAD calculated from a sample as in (3) [12].

$$\text{MAD} = \text{median}\ (|e_i - \text{median}\ (e)|) \quad (3)$$

where

$$e = \text{residual}$$

$$i = 1, 2, \ldots n$$

MAD is an estimator of the spread in data and has an approximately 50% breakdown point like the median.

2.2.3. Least- Trimmed Squares (LTS) Method

Least-Trimmed Squares (LTS) are a robust estimator with 50% breakdown point. This estimator is unaffected to exploitation due to outliers, if the outliers found not more than 50% of the data and can be represented as in (4) and (5) [13].

Ordering the squared residuals from smallest to largest:

$$(e^2)_{(1)}, (e^2)_{(2)}, \ldots, (e^2)_{(n)} \quad (4)$$

The LTS estimator chooses the regression coefficients *b* to minimize the sum of the smallest *m* of the squared residuals,

$$LTS(b) = \sum_{t=1}^{m} (e^2)_{(n)} \quad (5)$$

where,

$m = [n/2] + [(k+2)/2]$, a little more than half of the observation

$$e^2 = \text{squared residual}$$

2.3. Cross Validation Technique

Cross validation is a technique for evaluating how the results of a predicted model will predict the real data set. It is used when someone wants to evaluate how precise a predictive model is when it will be applied into real data [14]. In this study, only two methods of cross validation are used as in (6) and (7).

2.3.1. Mean Square Error (MSE)

MSE is represented as in (6),

$$\text{MSE} = \sum(y_i - \hat{y}_i)^2 / N \quad (6)$$
$$i = 1, \ldots, N$$

where y_i is the real data, $\hat{y}_i$ is the predicted data, N is the number of observations.

2.3.2. Mean Absolute Error (MAE)

MAE is represented as in (7),

$$\text{MAE} = \sum| y_i - \hat{y}_i | / N \quad (7)$$
$$i = 1, \ldots, N$$

where y_i is the real data, $\hat{y}_i$ is the predicted data, N is the number of observations.

3. Result and Discussion

3.1. Multiple Linear Regression

Referring to the Figure 1, P-P plot shows that the data are in nearly straight lines which indicate that the distribution of y vs all $x_{i,k}$ is normally distributed. Since the VIFF value for all $x_{i,k}$ is less than 10, it indicates that the multicollinearity among $x_{i,k}$ variables does not exist. So, the two early assumptions in MLR have been satisfied. The MLR model equation is given below where all variables are included without exceptional as in (8),

$$\hat{y} = 72.736 + 0.261x_1 - 1.298x_2 + 0.138x_3 + 0.149x_4 + 7.849x_5 + 0.465x_6 + 28.66x_7 \quad (8)$$

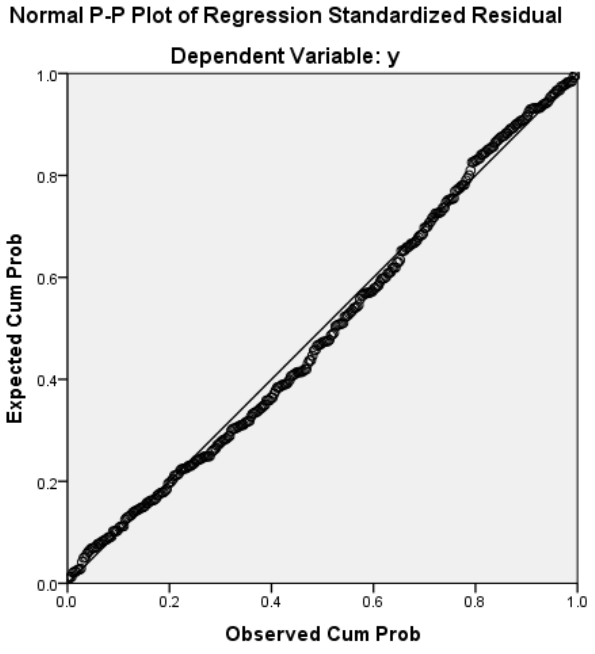

Figure 1. P-P Plot for $\hat{y}$ vs all x_i

The value of MSE of MLR model is 650.042, whereas the MAE of the MLR model is 20.5234. Using studentized residual test, it is shown that 12 points are identified as outliers. This is the reason why the usage of robust method is needed in this study.

3.2. Least Quartile Difference (LQD) Method

Firstly, sorting the residual value from smallest to highest value, then 25% of upper and 25% of lower quartile of the data will be removed. The remaining data are used to build the new model. The new model is stated as in (9).

$$\hat{y} = 56.990 + 0.311x_1 - 1.041x_2 + 0.318x_3 + 0.160x_4 + 9.711x_5 + 0.407x_6 + 28.604x_7 \quad (9)$$

The new model equation in (9) will be applied to the original data, then MSE and MAE are calculated. After using LQD method, the MSE value is 641.9068 and MAE value is 20.4288.

3.3. Mean Absolute Deviation (MAD) Method

The analysis of MADe method is as shown below,

1. MADe Method : Median ± 1 MADe = (-20.368775, 16.243715)
2. MADe Method : Median ± 2 MADe = (-38.67502, 34.54996)
3. MADe Method : Median ± 3 MADe = (-56.981265, 52.856205)

The MAD value is 18.30625. Based on the Median ± 3 MADe value obtained, 326 data have been included in building the new model and 6 data were removed. The new model equation is as shown as in (10),

$$\hat{y} = 68.604 + 0.2x_1 - 1.472x_2 + 0.176x_3 + 0.129x_4 + 5.512x_5 + 0.537x_6 + 28.147x_7 \quad (10)$$

The new model equation as in (10) is applied to the original data and the value of MSE and MAE is calculated. The values of MSE and MAE obtained are 652.9255 and 20.53357 respectively.

3.4. Least-Trimmed Square (LTS) Method

In LTS estimator method, the square of residual is sorted in ascending order in Microsoft Excel. Then 116 of the data were removed and the remaining 171 data are used to build the new model by using equation (5). The new model equation in (11) is as follows,

$$\hat{y} = 60.6 + 0.261x_1 - 0.966x_2 + 0.290x_3 + 0.172x_4 + 10.9x_5 + 0.378x_6 + 28.8x_7 \quad (11)$$

The new model in (11) then is applied to the original data and the MSE and MAE values are calculated 640.2429 and 20.4255 respectively.

3.5. Comparison of All Method

Table 1 shows the comparison among 4 models with different methods using MSE and MAE value. It indicates that MLR model with application of LTS method tends to be the best model with the smallest value of MSE and MAE.

This means y is positively affected by x_1, x_3, x_4, x_5, x_6 and x_7.Besides that, y is negatively affected by x_2.

Table 1. MSE and MAE comparison of all method

Model and method	MSE value	MAE value
MLR Model	650.042	20.523
MLR model with applied LQD method	641.907	20.429
MLR model with applied MAD method	652.926	20.534
MLR model with applied LTS method	640.243	20.426

4. Conclusions

In order to measure the effectiveness of the model, the comparison of methods is done. The value of MSE and MAE of the original data of MLR, MLR with applied LQD, MLR with applied MAD and MLR with applied LTS estimator was compared. Based on the value of MSE and MAE, it is concluded that the MLR model with applied LTS estimator is the best model since it has the lowest value of MSE and MAE. In conclusion, plasma glucose concentration in an oral glucose tolerance test is positively affected by the increasing of body mass index, diastolic blood pressure, triceps skin fold thickness, diabetes pedigree function, age and yes or no for diabetes according to WHO criteria. In fact, diabetes pedigree function and yes/no for diabetes according to WHO criteria have the highest impact on plasma glucose concentration in an oral glucose tolerance test. Besides that, plasma glucose concentration in an oral glucose tolerance test is negatively affected by number of pregnancies. This result can be used as a guideline for medical doctors as an early prevention of stage 2 of diabetes.

Acknowledgements

This research is supported by the Universiti Tun Hussein Onn Malaysia under the TIER 1 grant scheme vot number H232.

REFERENCES

[1] R. Bilous, & R. Donnelly. Handbook of Diabetes: Fourth Edition. John Wiley & Sons Ltd., 2010.

[2] A. Ramachandran & C. Snehalatha. Current scenario of diabetes in India, Journal of Diabetes, Vol. 1, No. 1, 18–28, 2009.

[3] S. Gulati & A. Misra. Sugar intake, obesity, and diabetes in India, Nutrients, Vol. 6, No. 12, 5955–5974, 2014.

[4] M. H. Kutner, C. J. Nachtsheim, J. Neter & W. Li. Applied Linear Statistical Models (fifth Edition), McGraw-Hill, 2005.

[5] M. Williams, C. A. G. Grajales & D. Kurkiewicz. Assumptions of multiple regression: Correcting two misconceptions, Practical Assessment, Research & Evaluation, Vol. 18, No. 11, 1–14, 2013.

[6] D. S. Courvoisier & O. Renaud. Robust analysis of the central tendency, simple and multiple regression and ANOVA, Journal of American Statistician, Vol. 3, No. 1, 78–87, 2011.

[7] S. Morasca. Building Statistically Significant Robust Regression Models in Empirical Software Engineering, PROMISE '09 Proceedings of the 5th International Conference on Predictor Models in Software Engineering, Vol 17, No. 1-2, 23-33, 2009.

[8] N. Che-Him, M. G. Kamardan, M. S. Rusiman, S. Sufahani, M. Mohamad & N. K. Kamaruddin. Spatio-temporal modelling of dengue fever incidence in Malaysia, Journal of Physics: Conference Series, Vol. 995, No. 1, 012003, 2018.

[9] N. Che-Him, R. Roslan, M. S. Rusiman, K. Khalid, M. G. Kamardan, F. A. Arobi, N. Mohamad. Factor Affecting Road Traffic Accident in Batu Pahat, Johor, Malaysia, Journal of Physics: Conference Series, Vol. 995, No. 1, 012033, 2018.

[10] S. A. Leona, G. W. Stephen, C. P. Steven, N. B. Amanda & C. W. Ingrid. Multiple Linear Regression - Research Methods in Psychology, Wiley Online Library, 2012.

[11] C. Croux, P. J. Rousseeuw, O. Hossjer, C. Croux, P. J. Rousseeuw & O. Hossjer. Generalized S-Estimators, Journal of the American Statistical Association, Vol. 89, No. 428, 1271–1281, 1994.

[12] M. Satyaki & S. Robert. Bahadur Representations for the Median Absolute Deviation and Its Modifications, Journal of The American Statistical Association, Vol. 88, No. 424, 1273-1283, 2009.

[13] M. M. David, S. N. Nathan, D. P. Christine, S. Ruth & Y. W. Angela. On the Least Trimmed squares Estimator, Algorithmica, Vol. 69, No. 1, 148-183, 2014.

[14] A. Khamis. Application of Statistical and Neural Network Model for Oil Palm Yield Study, Ph.D. Thesis, Universiti Teknologi Malaysia, Malaysia, 2005.