

Introdu Research on the Influencing Factors of Diabetes Prediction Model among Pima Indian Heritages Female Data

Yaxuan Huang *

University of Connecticut, Connecticut, 06269, United State

* Corresponding Author Email: yaxuan.huang@uconn.edu

Abstract. Diabetes, as the chronic disease that has the greatest impact on the health of the world, has caused an immeasurable impact on people's lives. This research employed Multiple Linear Regression to explore various factors contributing to the incidence of diabetes. The data contains multiple variables that may predict diabetes. The goal was to assess their predictive value on the diabetes binary outcome through a multiple linear regression model. Diagnostic plots indicated potential model violations, with the residuals vs. fitted values plot suggesting non-linear relationships and variable residual variance. The normal Q-Q plot revealed minor departures from normality, principally at the data extremes. The Scale-Location plot raised questions about the consistency of variance across the data, while leverage analysis did not identify any significantly influential data points. The MLR model yielded an R-squared value of 0.3033, suggesting that approximately 30% of the variability in diabetes outcomes was accounted for by the model. There were some variables that were found to be very significant: Glucose and BMI, which showed p-values that suggested strong associations with diabetes. The findings from this study have identified important determinants of health and have also indicated reasons why there is a need for improvements in the model diagnostics for diabetes. For better health care outcomes, these findings show that diabetes risks are complex and that predictive capacities could benefit from more sophisticated analytical models.

Keywords: Diabetes; multiple linear regression; influencing factors.

1. Introduction

Diabetes mellitus is a serious worldwide health issue that is having an impact on healthcare systems and is rising quickly. This study examines the significance of statistical models in predicting diabetes due to its increasing prevalence and health implications [1]. The complexity of the disease, which involves genetic, environmental, and lifestyle factors, necessitates a comprehensive approach to effective management and prevention. The use of statistical and machine learning techniques for diabetes prediction holds potential for early detection and intervention [2].

Diabetes has been extensively investigated in research, which includes genetic susceptibility [3], global trends and the impact of lifestyle changes [4, 5]. These studies emphasize various causes of diabetes and the requirement for predictive models that can explain such factors. The use of Advanced Statistical Techniques in predicting diabetes has helped to uncover important variables such as BMI, age, and genetics. During these years, many researchers have used Pima women as study population in their works to illustrate how predictable analysis can be. The authors were able to show that among other types of classification done by the Multilayer Perceptron (MLP), it had an accuracy of 86.08%. Concerning Long Short Term Memory (LSTM) model's enhancement on prediction, this was observed by having an accuracy level of 87.26% in diagnosing diabetes [6]. Topmost among them has been the application and comparison of Linear Regression and Multiple Linear Regression models which have shown a remarkable degree of accuracy in predicting new active cases COVID-19 [7]. Such models help to identify those at high risk, hence personalized medications are possible.

In investigating the factors that lead to type 2 diabetes, study proposed that the classic factor cannot explain the incidence of diabetes. In addition to some classic possible causes of disease, infection is listed as one of the causes of type 2 diabetes [8]. According to the dataset from BRFSS, the study reveals that the social factors such as race, income, age, or gender can affect the risk of chronic diseases [9]. In this group of studies, we also tried to find out the relationship between diabetes and

other possible pathogenic factors. This will have a huge impact on the prediction and prevention of diabetes.

However, challenges in statistical modeling, like data quality and the need for extensive data for training, are noted [10]. These issues emphasize the need for thorough development and validation of models. This research aims to add to the understanding of diabetes prediction through statistical models, examining their effectiveness, challenges, and potential for future advancements in healthcare and public health. The study's findings demonstrate the critical role of comprehensive data analysis in understanding complex health conditions like diabetes. Addressing issues such as data quality and the need for large datasets is essential in refining these models for more accurate predictions. The healthcare practices and strategies of public health can be totally transformed by these models, which offer opportunities for personalized treatments and improved management of diseases. The models need to be developed and updated continually to keep pace with the changing face of diabetes and its impact on the global population.

2. Methodology

2.1. Data Source

The original source of the diabetes patient's data in this study is from the National Institute of Diabetes and Digestive and Kidney Diseases. The original data was saved in .CSV format.

2.2. Variable Introduction

The data used in this paper contains 769 instances with 9 variables (Table 1). The dataset does not contain any missing values. Each column has complete data with no null entries. The dataset is combined with 8 independent numeric variables (medical predictor variables) and one target dependent variable (Outcome).

Table 1. Different types of variables

Term	Type	Meaning
Pregnancies	numeric	Number of pregnancies
Glucose	numeric	Glucose level in blood
BloodPressure	numeric	Blood pressure measurement
SkinThickness	numeric	Thickness of the skin
Insulin	numeric	Insulin level in blood
BMI	numeric	Body mass index
DiabetesPedigreeFunction	numeric	Diabetes percentage (Genetic factors)
Age	numeric	Age
Outcome	numeric	Final result where 1 is Yes and 0 is No (indicating the presence or absence of diabetes)

Although there are total of 9 variables in the dataset, there is no clear correlation between the 8 independent variables. Therefore, the simplification is not significant and meaningful. The dataset can be used to do the next step directly.

2.3. Method Introduction

The method used in this study are Binary Logistic Regression Model and the Multiple Linear Regression Model. The Binary Logistic Regression Model is a specialized form of regression used when the dependent variable is binary (in this case, indicating diabetes or non-diabetes). It estimates the probability of an event occurring, making it suitable for classification tasks. The Binary Logistic Regression Model predicts the likelihood of an individual having diabetes based on various input features.

One statistical method for analyzing the relationship between a dependent variable and multiple independent variables is the Multiple Linear Regression Model. This model aims to predict a continuous outcome, such as blood glucose level, based on several predictor variables, including multiple other factors that likely effect the possibility of suffer from diabetes.

3. Results and Discussion

3.1. Binary Logistic Regression Model

To begin with, data preprocessing is deeded for conducting the following process. Some physiologically implausible zeros in certain columns were replaced with median values. This preprocessing step ensures that the dataset is more accurate and suitable for training a Binary Logistic Regression Model. After that, the train-test split was made for reproducibility by randomly divided into two parts: 20% of the data is used as a test set, which is used to assess the model's performance on untested data, while the remaining 80% of the data is used as the training set. Then, the next step goes into the model training. we are fitting a logistic regression model, which is suitable for binary classification problems like predicting diabetes (binomial outcome). Ultimately, the model's performance on the test data is evaluated by creating a confusion matrix (Table 2, 3).

Table 2. Model training results

name	Accuracy	Precision	Recall	F1-Score
data	0.7727273	0.7073171	0.5576923	0.62365

Table 3. Confusion Matrix

Actual\Predicted		
	0	1
0	90	12
1	23	29

The numbers "0" and "1" in this confusion matrix represent the presence and absence of diabetes, respectively. The model correctly predicted non-diabetic instances (true negatives: 90) and diabetic instances (true positives: 29). However, it also made some errors, predicting non-diabetic instances as diabetic (false positives: 12) and failing to predict some diabetic instances (false negatives: 23).

The accuracy of 77.27% indicates the overall correctness of the model's predictions. The precision of 70.73% suggests that when the model predicts diabetes, it is correct about 70.73% of the time. The recall of 55.77% indicates the model's ability to capture actual instances of diabetes. The F1-Score, a balance between precision and recall, is 62.37%.

These results highlight the model's capability to identify non-diabetic cases more accurately than diabetic cases. The trade-off between precision and recall should be considered based on the specific goals of the application. Further optimization or exploration of feature engineering could enhance the model's performance. Additionally, clinical validation and consideration of the model's impact in a healthcare context are essential for practical implementation.

3.2. Multiple Linear Regression Model

First of all, the diabetes dataset is loaded into R environment using read.csv function in R Language, which is a standard method of importing CSV files in R. Next, a Multiple Linear Regression model is created using the "lm" function. In this model, the criterion variable (Outcome), which indicates whether or not an individual has diabetes, is set as the dependent variable while all other columns in the data set are considered independent variables. This is indicated by the formula Outcome. where the dot stands for all other columns in the data frame. After building this model, we immediately call summary function to provide a detailed summary about it. This information includes

such things as coefficients of independent variables; R-squared value, and p-values that are important for assessing how well the model performs, thus its every predictor's meaning respectively (Table 4).

Table 4. Linear model Results

Coefficients:	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.853	0.085	-9.989	< 2e-16
Pregnancies	0.0205	0.005	4.014	6.56e-05
Glucose	0.005	0.0005	11.493	< 2e-16
BloodPressure	-0.002	0.0008	-2.873	0.00418
SkinThickness	0.0001	0.0011	0.139	0.88954
Insulin	-0.0001	0.0001	-1.205	0.22857
BMI	0.013	0.002	6.344	3.85e-10
DiabetesPedigreeFunction	0.147	0.045	3.268	0.00113
Age	0.002	0.0015	1.693	0.09092

The significance codes at the bottom explain the symbols associated with p-values: " for $p < 0.001$, " for $p < 0.01$, " for $p < 0.05$, and '.' for $p < 0.1$. These codes help quickly identify which predictors are statistically significant. The residual standard error is approximately 0.4002 on 759 degrees of freedom, which gives an idea of the typical deviation of the observed outcomes from the fitted values. With a Multiple R-squared value of 0.3033, the model can be responsible for about 30.33% of the variance in 'Outcome'. When penalized for the number of variables, the Adjusted R-squared, which takes into account the number of predictors in the model, is somewhat lower at 0.2959 and gives a more accurate picture of the explanatory power of the model. Lastly, the null hypothesis that all regression coefficients are equal to zero is tested using the F-statistic and its corresponding p-value ($p < 2.2e-16$). The extremely low p-value suggests that the model is statistically significant and suggests that at least one of the variables is likely related to the outcome.

Additionally, diagnostic plots are generated using the plot function to visually assess various assumptions of the linear regression model, such as linearity, homoscedasticity, and normality of residuals. Finally, the script optionally includes a step for making predictions using the constructed model. This is achieved through the predict function, which can be applied to either the same dataset or ideally to a separate testing dataset for evaluating the model's predictive accuracy (Fig 1).

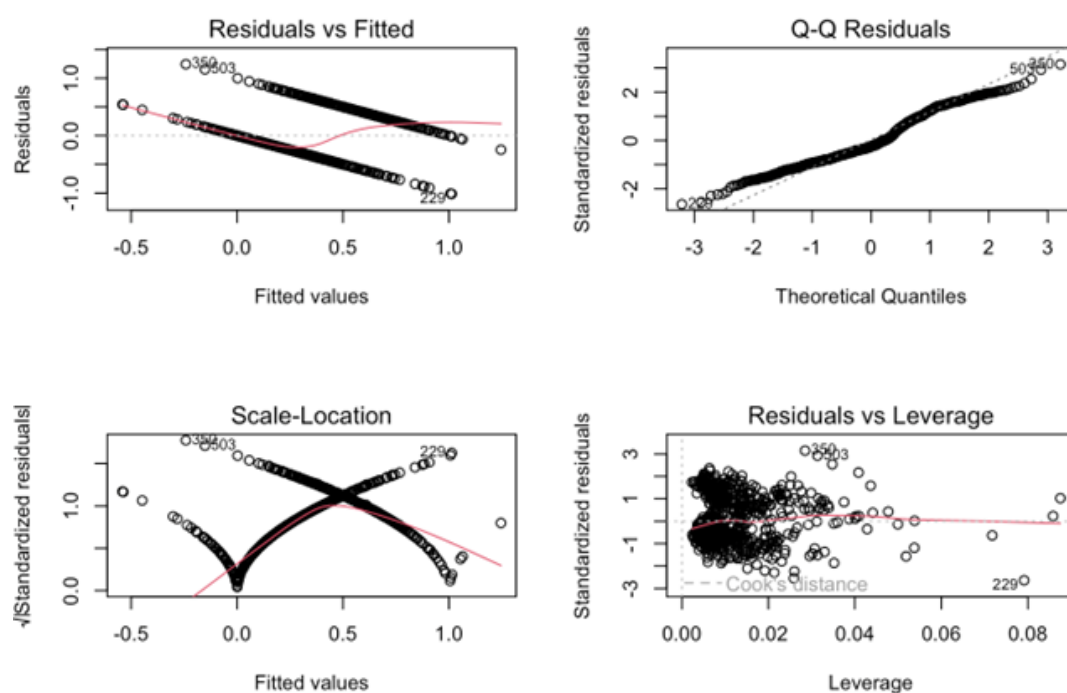


Fig 1. Model test results

Base on the results shows above, the Binary Logistic Regression Model demonstrates a more effective performance for the specific task of predicting diabetes in the Pima heritage female cohort. This model is tailored for binary outcomes, such as classifying individuals as diabetic or non-diabetic, and it aligns well with the nature of the data and the research objective in this particular study. It effectively handles the dichotomy in the dataset, making it a more suitable choice.

4. Conclusion

Finally, this study has successfully used Binary Logistic Regression and Multiple Linear Regression Models to identify and assess critical variables of diabetes incidence in a cohort of women of Pima descent. The study has shown that BMI and glucose are important variables that have strong predictive power for the binary diabetes outcome. While the Binary Logistic Regression model had an impressive accuracy rate of 77.27%, the Multiple Linear Regression model explained around 30% of the outcome variation. Indicating a significant association between the predictors and diabetes status. These findings not only make a significant contribution to the field of predictive analytics in diabetes research but also indicate that the customized medical intervention has entered a new era.

References

- [1] Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Prev Chronic Dis*, 2019, 16: 190109.
- [2] Ambhrish G, Bharathi G, Anitha G, Chetana S, Dhanraj, Kiran M. Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, 2020, 3(1), 127-130.
- [3] Wu Y, Ding Y, Tanaka Y, Zhang W. Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *Int J Med Sci.*, 2014, 11(11): 1185-1200.
- [4] Chang V, Bailey J, Xu Q A, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput Appl.*, 2022, 24: 1-17.
- [5] Pierannunzi C, Hu S S, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011. *BMC Med Res Methodol.*, 2013, 13: 49.
- [6] Butt U M, Letchmunan S, Ali M, Hassan F H, Baqir A, Sherazi H H R. Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *J Healthc Eng.*, 2021.
- [7] Rath S, Tripathy A, Tripathy A R. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes Metab Syndr.*, 2020, 14(5): 1467-1474.
- [8] Chakraborty S, Bhattacharyya R, Banerjee D. Infections: A Possible Risk Factor for Type 2 Diabetes. *Adv Clin Chem*, 2017, 80: 227-251.
- [9] Massouh N, Jaffa A A, Tamim H, Jaffa M A. Social and racial inequalities in diabetes and cancer in the United States. *Front Public Health*, 2023, 11: 1178979.
- [10] Tan S Y, et al. Type 1 and 2 diabetes mellitus: A review on current treatment approach and gene therapy as potential intervention. *Diabetes Metab Syndr*, 2019, 13(1): 364-372.