Risking Factors of Type II Diabetes Mellitus Using Regression analysis to test significance of every factors

Cindy Gao[1, †], Yanrong Huo[1, †], Yingzhe Liu[2, *, †]

[1]Department of Statistical Science & Department of Economics University of Toronto Toronto, Ontario, Canada

[2]Department of Biology Johns Hopkins University Baltimore Maryland United State

*Corresponding author: yliu292@hu.edu

†These authors contributed equally.

Keywords: Risking, diabetes, mellitus.

Abstract: This study is conducting quantitative research exploring if the disease of diabetes will be affected by several factors including age, weight, and BMI. In this study, data transformation by data cleansing, formatting were performed for cleaner results. Multiple logistic regression models were built to find the best one to predict one's chances of encountering diabetes. To assess these models, evaluations of Akaike information criterion (AIC), Mean Squared Error (MSE), and p-values were used. Diagnostic plots were constructed to further evaluate the models. In the end, it was found that one's age, gender, race, and BMI level were the best mix of predictors of one's chances of coming across diabetes. This study's results are noteworthy as there is an increasing number of people suffering from diabetes. Finding these factors will promote more precaution in each individual, and hopefully reduce the risk of encountering diabetes.

1. Introduction

Type II diabetes mellitus (T2DM) is becoming a significant public health concern worldwide. It is a chronic disease impairing the ability of patients to regulate blood sugar (glucose) levels and their circulations. Other complications include kidney failure, vision loss, heart disease, stroke, premature death, and amputation of limbs. Such disease will cause disorders related to in patients' circulatory, nervous, and immune systems. In 2017, there were 450 million people diagnosed with diabetes and 1.37 million deaths due to the disease worldwide. It used to be regarded as adult-onset diabetes and it is more common in elders. However, T2DM can also be found in children. There is an increased number of children diagnosed with T2DM due to obesity.

The causes of the disease can be environmental factors such as obesity, unhealthy diets, or physical inactivity. Genetic factors also contribute to the disease by resulting in multiple pathophysiological disturbances that are responsible for impaired glucose homeostasis in T2DM. It is also important to note that there is currently no cure for T2DM.

Dr. Ram D. Joshi and Dr. Chandra K. Dhakal predicted type II diabetes for Indian women utilizing a logistic regression model and a decision tree that they developed to improve the understanding of risk factors that cause diabetes in 2021 [1]. Their analysis finds five main predictors of T2DM, which are glucose, pregnancy, body mass index (BMI), diabetes pedigree function, and age. Dr. Yanling Wu, Dr. Yanping Ding, Dr, Yoshimasa Tanaka, and Dr. Wen Zhang studied the epidemiology of T2DM and found that the roles of genes, lifestyle, and other factors contribute to a rapid increase in the incidence of T2DM in 2014 [2]. Dr. Aftab Ahmad, Dr. Alamgir Khan, and Dr. Salahuddin Khan reviewed the available literature which shows that heredity, radiation, disorder, and diseases of the pancreas are the main causes of diabetes [3]. They also concluded that regular physical activities, proper intake of diet, and different curative medicine can help one to manage and reduce the symptoms of diabetes mellitus. Dr. Onyencho V. Chidi and Dr. Asagba R. Bolaji examined psycho-demographic factors as predictors of depression among persons with diabetes mellitus in the Ibadan metropolis in

2016 [4]. They applied a cross-sectional survey method to collect data from two hundred and thirty-eight respondents and concluded that psycho-demographic variables such as gender, hopelessness, and perceived life purpose have a significant influence in predicting depression among persons with diabetes mellitus. In 2017, Dr. Anggraini Dwi Kurniaa, Dr. Anchaleeporn Amatayakulb, and Dr. Sirikul Karuncharernpanitc identified factors predicting diabetes self-management among adults with T2DM in Malang City, East Java, Indonesia [5].They found that diabetes self-management among adults with T2DM could be improved by enhancing their perceived self-efficacy to achieve their self-management behavior, such as having a healthy diet, exercising regularly, actively monitoring blood glucose level, taking medication and foot care, and providing support to promote good situational influence. In 2017, Dr. Ruth A. Hackett and Dr. Andrew Steptoe reviewed the influence of psychological stress as a risk factor of T2DM highlight the physiological responses to stress that are probably related to T2DM, drawing on evidence from animal work, large epidemiological studies, and human laboratory trials [6].In 2002, Dr. Barbara Fletcher found that insulin resistance increases a person's risk for developing impaired glucose tolerance and T2DM and share many of the same risk factors as those with T2DM [7].In 2018, Dr. Beth H. Rice Bradley found that the consumption of fish and marine n-3 fatty acids among Asian populations and regular-fat dairy foods and trans-palmitoleic acid among Western populations may be associated with reduced risk for T2DM [8]. In 2017, Dr. Surajit Chakraborty found that infections may precipitate insulin resistance via multiple mechanisms, such as the proinflammatory cytokine response, the acute-phase response, and the alteration of the nutrient status [9]. They argued that infections that are known to contribute to insulin resistance should be considered as risk factors for T2DM. In 2018, Dr. Tashi Dendup reviewed different elements of the environment that have been posited to influence T2DM. They found that air pollution, food, and physical activity, environment, and roadways proximity were the most common environmental characteristics [10].

The survey data collected by the US National Center for Health Statistics (NCHS) which has conducted a series of health and nutrition surveys since the early 1960’s was analyzed using R. Since 1999 approximately 5,000 individuals of all ages are interviewed in their homes every year and complete the health examination component of the survey. The health examination is conducted in a mobile examination center (MEC). This dataset is in a built-in package of R called NHANES, the title of this dataset is “Data from the US National Health and Nutrition Examination Study” with version 2.1.0. The collecting date of this dataset is July 2nd, 2015.

As the US National Health and Nutrition Examination Survey provides body shape and related measurements, we aim to analyze if there is a relationship between these factors that can reflect patients’ living habits and the chance of getting diseases.

From the basic cleaning procedures, the chosen independent variables, Xi, included gender, age, race, education, marital status, income, weight, height, BMI (body mass index), state of depression, sleep hours, physical activities, and smoking habits.

The dependent variable, Yi, is Diabetes, and we explore if any factors could influence the chance of getting diabetes by exploring the correlations. For this part, because for some of the variables, there were too many levels, some of them were combined to facilitate future analysis and manipulations. The variables that were combined were age, which is a numerical variable and the household income, HHIncome, which is a categorical variable.

The first variable is Age, is the interviewee’s age. It is continuous and ranges from 20 to 80.

The second variable is Gender, which is the interviewee’s gender. It is categorical, which has two levels: female and male.

The third variable is Race3, which is the interviewee’s ethnicity. It is also categorical, with the six following levels: Black, White, Hispanic, Mexican, Asian, and Other.

The fourth variable is Education, which is the interviewee’s education level. Also a categorical variable, with the following five categories: High School, College Grad, Some College, 8th Grade, and 9-11th Grade.

The fifth variable is HHIncome, which is the interviewee’s household income level. It is categorical, and has four categories splitting into four different salary ranges.

The sixth variable is BMI, which is the interviewee's Body Mass Index (BMI) level. It is continuous and ranges from 16.7 to 59.1, which has a range of 42.4.

The seventh variable is Depressed, which is the interviewee's depression level. It is categorical, which has 3 levels: None, Several, and Most.

The eighth variable is SleepHrsNight, which is the amount of sleep one receives per night. It is continuous and ranges from 2 to 12, with a range of 10.

The ninth variable is PhysActive, which measures whether one is active or not. It is categorical with 2 levels: Yes and No.

Table 1. Independent Variables, Xi

Variables	Type	Range
Age	Continuous	60
Gender	Categorical	0:Female, 1: Male
Race	Categorical	Black, White, Hispanic, Mexican, Asian and Other
Education	Categorical	High School, College Grad, Some College, 8th Grade, and 9-11th Grade
Household Income (HHIncome)	Categorical	Less than 25000, 20000-54999, 55000-99999, more than 99999
BMI	Continuous	42.4
Depressed	Categorical	None, Several and Most
SleepHrsNight	Continuous	10
PhysActive	Categorical	Yes, No

2. Statistical Methods

At first sight, the dataset is quite clean and organized. The sample size before any manipulation is 1756. After ridding any NA variables in the dataset, the size remained the same. The variables Gender, Age, Race3, Education, HHIncome, BMI, Depressed, SleepHRsNight, PhysActive, and Diabetes, were chosen for further investigation. Before any data analysis or model fitting, data transformations were performed on each variable. The main data transformations done in this study were data cleansing, and formatting, where all the chosen variables were slightly manipulated for a cleaner result that can facilitate later research processes. Among the variables, age, BMI and sleep hours are continuous numerical variables, while others are categorical variables. All the categorical variables, including, Gender, Depressed, PhysActive, and Diabetes were either converted into double (0 or 1) formats or revised into combined group levels for easier and cleaner access and usage.

Interesting characteristics that we noticed when evaluating the dataset before modeling the data is that there is almost an ideal even distribution among the gender variable, with 50.17% being females and 49.83% being males. Along with that, plots were drawn to evaluate whether there could be any evident relationship seen between BMI and Diabetes through their scatterplots. From the plot we draw describing the relationship between the respondent and the independent variable BMI, it can be seen that, except for some outliers, the distributions of dots around 0 and 1 (which means non-diabetes and diabetes) are the same, which means that the variable BMI doesn't influence the respondent a lot.

3. Methods

To proceed with this investigation, multiple logistic regression models were built in order to find the best one to predict one's chances of encountering diabetes. The first model built included all of the nine predictor variables and one response listed above. The regression model is as of the following:

$$\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n \quad (1)$$

Where p represents the variable of interest, Diabetes, β0 represents the intersection, and $\beta_1 \ldots \beta_n$ represents the coefficient for each respective predictor variable. The predictor variables that were first used in the original model were Gender, Age, Race3, Education, HHIncome, BMI, Depressed, SleepHrsNight and PhysActive. Following that, other models were created with the same predictor variables above but with different mixtures of predictor variables in the model. After multiple tests, the final model came to be only with the predictor variables Gender, Age, Race3, and BMI.

To assess the models, evaluations of Akaike information criterion (AIC), Mean Squared Error (MSE), and p-values were used. AIC is an evaluation format in which it assesses how well the model fits the data it is working with. AIC uses a model's maximum likelihood estimation to assess the model's fit. Typically, a low AIC score indicates a good level of fit. However, it is important to separate between underfitting and overfitting, and AIC accounts for both. MSE is the next evaluation method used. In simple words, MSE generates the information on how well the fitted line is; it gives the average measurement of how far the actual points are from the estimated points. Again, similar to AIC, the lower the MSE the better, as a lower MSE means a better estimation of the fitted model to the actual points. Lastly, p-values were also used to evaluate the fitted model. P-values tell us which variables in the model are significant at certain levels. In this study, the 95% significance level was used to assess all models.

To further evaluate the models, diagnostic plots were constructed. This includes a Residual vs Fitted plot, Normal Q-Q plot, Scale-Location plot, and Residual vs Leverage plot. The Residual vs Fitted plot, tests to see whether the relationship between the response and predictors is linear. Ideally, a randomly scattered plot centering around a mean of zero is desired. The Normal Q-Q plot assesses normality within a model, and the desired results are for the points to lie on the 45 degrees (y=x) threshold. The Scale-Location plot assesses homoscedasticity, meaning it is in search of constant variance. Similar to the Residuals vs Fitted plot, the desired result is once again an evenly randomly distributed dataset of points. Lastly, the Residuals vs Leverage plot is extremely useful in identifying outliers, extreme values, or influential points. Anything beyond Cook's distance would be classified as an influential point.

4. Result

The first model fitted, with the predictor variables Gender, Age, Race3, Education, HHIncome, BMI, Depressed, SleepHrsNight and PhysActive, the deviance residual resulted in a minimum value of -1.6881, median of -0.3172 and a maximum value of 3.0850. With each coefficient for each random variable determined, it can be seen that only five of them are statistically significant among all 28 levels. The first one is an intercept of this model; it has an estimated value of -7.141525 with p-value 1.85 x 10-12. This value is less than 0.05 which indicates that this variable is statistically significant at the 95% significance level. The second one is the variable Gender at the level male, its estimated coefficient value is 0.393244 with a p-value of 0.0160. The third one is the variable Age, a numerical variable, and its estimated coefficient value is 0.057564 with a p-value less than 2 x 10-16. It should also be noted that the variable Age is the best performing variable in this model. The fourth one is the variable Race3 at the level White, its estimated coefficient value is -0.743397 with a p-value of 0.0204. And the last one is the variable BMI, which is also a numerical variable, and its estimated coefficient value is 0.085177 with a p-value of 1.71 x 10-13. For this model, the AIC result is 1141.2 and the MSE is 43787.72.

Overall, holistically, this model lacks in many aspects. Firstly, the proportion of significant p-values are quite low. Secondly, there may be better models with both lower AIC and MSE test results. After many attempts in building better models with different combinations of the original predictor variables, the final model came to be with only the variables Gender, Age, Race3 and BMI.

For this model, the deviance residual has a minimum value of -1.5539, median of -0.3395 and a maximum value of 2.9939. For the coefficients of variables, this model still also has five significant values. The variables that are significant are exactly the same as the previous model, however, the p-values have changed immensely. For the intercept, its p-value changes to less than 2 x 10-16. For the variable Gender at Male level, its p-value changed to 0.0377. For the variable Race3 at the White level, its p-value changed to 0.0118. Lastly, for the variable BMI, its p-value changed to 6.66x 10-15. From the perspective of significance, this model has a much stronger results than the previous. Checking the model's AIC and MSE, the results did indeed lower. The AIC for this model is 1130.6 and the MSE is 35186.58.

After fitting this model, VIF was used to assess multicollinearity. After checking, it turns out that both models have no multicollinearity and all predictors in these two models work well together, therefore, there is no need to remove any variables from this model.

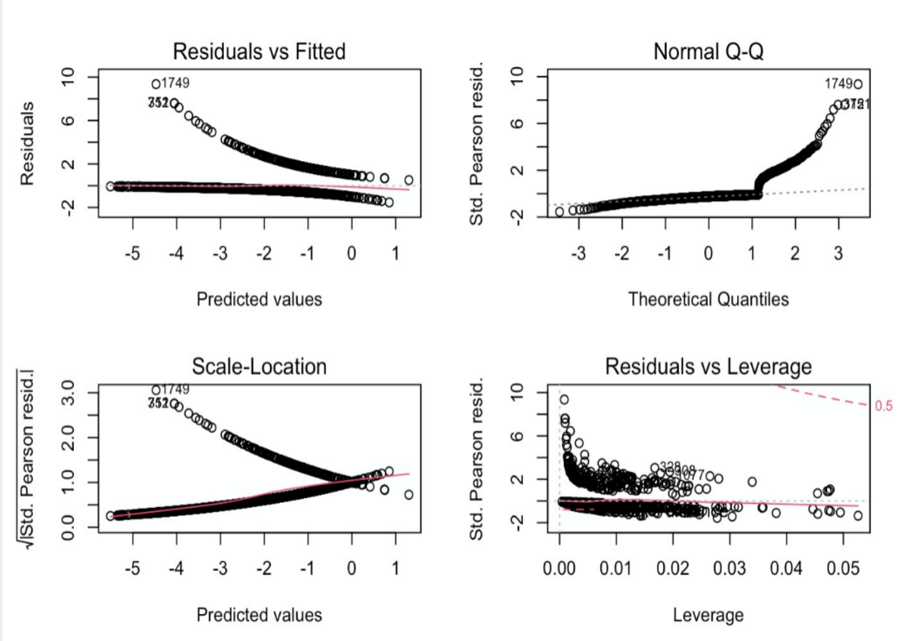

Figure 1. Diagnostic Plots

After deciding on a final model, model diagnostics were used to check the validity of this model. The Residual vs Fitted plot, tests to see whether the relationship between the response and predictors is linear. From our Residual vs Fitted plot, the residuals spread in distinct patterns, therefore the linearity assumption is not appropriate. In the Normal Q-Q plot, there are many residual points that do not rest on the diagonal threshold line, therefore meaning that the data is not distinctively normally distributed. The Scale-Location plot tests constant variance. The desired plot is a plot with residuals that are equally spread out, and a horizontal line. However, it can be seen that the red line is not horizontal in the final model's plot, meaning that the model fails constant variance assumption. Lastly, a Residual vs Leverage plot is supposed to help in identifying extreme values in the models. In the final model's Residual vs Leverage plot, points are floating around, thus again drawing the conclusion that our model is not strong enough, and that there is still room for improvements.

5. Discussion

It is evident that there are some errors in the final model as well. This could potentially be due to a few reasons. First, the method used to fit the model may not have been a good choice; there could be other model types that are able to be a stronger model. Second, the variables chosen were not perfectly related to the dependent variable. Our first step for variable selection was based on our common sense, and literature analysis. However, sometimes common senses are not efficacious in statistical analysis, thus resulting in wrong answers. Along with that, there may be other potential problems such as omitted variable bias. Next, the final model was fitted with only the original predictor variables used in the original model, therefore no new variables were studied or included. If different or new variables were chosen, perhaps a stronger model could have been created.

There are some methods to improve our model. First, we can do some variance stabilizing transformations or any transformations. Second, we can try to increase the sample size to improve randomness. Third, we can try fitting other models using different variables. Lastly, we can split the original dataset into two datasets, one with more observations (normally 80% of all observations) used as a training set, one with fewer observations used as a test set (normally 20% of all observations). We can build models on the training set, and use the test set to test their validities. It should be noticed that when we try to split the original dataset, the selection process should be random, the training set and test set should have the same characteristics including all kinds of observations, the only difference between these two sets should be population size.

6. Conclusion

After thorough literature analysis and model testing, it was found that one’s age, gender, race, and BMI level were the best mix of predictors of one’s chances of coming across diabetes. Additionally, it was found that gender, BMI level, and one’s age has the most effects on one’s chances of encountering diabetes; specifically, men, those with higher BMI levels, and the elders have a higher chance of facing diabetes. In this study specifically, the results show that those who classify as White have the lowest chances of encountering diabetes when keeping all other factors in this study constant.

It is important to keep in mind that these are solely the findings for this specific investigation; results may differ in other circumstances. With that said, as it can be seen, there exist certain issues within the model diagnostics. Although the VIF tests show no multicollinearity within the models, the diagnostic plots show that there could be improvements in the model’s linearity, normal distribution, homoscedasticity, and riding more outliers and influential points.

The next steps for this research should be to first fix the limitations listed above. Some potential resolvers are variance-stabilizing transformations or looking to expand the sample size to reach a normal distribution. Increasing the sample size could also assess whether points are truly influential points or not. Next, there are only a few predictor variables in the final model, however, there are most likely many other factors that play a part in one’s chances of encountering diabetes; these other potential factors should be sought out. Lastly, in our technologically developed world, the possible next step is to analyze these findings deeper with more advanced equipment and techniques, such as machine learning.

References

[1] Joshi RD, Dhakal CK. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. Int J Environ Res Public Health. 2021 Jul 9; 18(14): 7346. doi: 10.3390/ijerph18147346. PMID: 34299797; PMCID: PMC8306487.

[2] 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 Sep 6; 11(11): 1185-200. doi: 10.7150/ijms.10001. PMID: 25249787; PMCID: PMC4166864.

[3] Ahmad, Aftab & Khan, Alamgir & Khan, Salahuddin. (2017). Causes, Complications and Management of Diabetes Mellitus. Chronicle Journal of Food and Nutrition. 1. 1-3.

[4] Chidi, Onyencho & Asagba, Rb & Pindar, Sadique & Said, Jidda & Ibrahim, Abdu & Wakil, Musa & Isa, Rabbebe & Makput, Duwap. (2016). Psycho-Demograhic Factors as Predictors of Depression Among Person With Diabete Mellitus. International Journal Of Scientific Research And Education. 10.18535/ijsre/v4i04.12.

[5] Anggraini Dwi Kurnia, Anchaleeporn Amatayakul, Sirikul Karuncharernpanit, Predictors of diabetes self-management among type 2 diabetics in Indonesia: Application theory of the health promotion model, International Journal of Nursing Sciences, Volume 4, Issue 3, 2017, Pages 260-265, ISSN 2352-0132, https://doi.org/10.1016/j.ijnss.2017.06.010. (https://www.sciencedirect.com/science/article/pii/S2352013217300844)

[6] Hackett RA, Steptoe A. Type 2 diabetes mellitus and psychological stress - a modifiable risk factor. Nat Rev Endocrinol. 2017 Sep; 13(9):547-560. doi: 10.1038/nrendo.2017.64. Epub 2017 Jun 30. PMID: 28664919.

[7] Fletcher B, Gulanick M, Lamendola C. Risk factors for type 2 diabetes mellitus. J Cardiovasc Nurs. 2002 Jan; 16(2):17-23. doi: 10.1097/00005082-200201000-00003. PMID: 11800065.

[8] Rice Bradley BH. Dietary Fat and Risk for Type 2 Diabetes: a Review of Recent Research. Curr Nutr Rep. 2018 Dec; 7(4):214-226. doi: 10.1007/s13668-018-0244-z. PMID: 30242725; PMCID: PMC6244743.

[9] Chakraborty S, Bhattacharyya R, Banerjee D. Infections: A Possible Risk Factor for Type 2 Diabetes. Adv Clin Chem. 2017; 80: 227-251. doi: 10.1016/bs.acc.2016.11.004. Epub 2017 Jan 3. PMID: 28431641.

[10] Dendup T, Feng X, Clingan S, Astell-Burt T. Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. Int J Environ Res Public Health. 2018 Jan 5; 15(1): 78. doi: 10.3390/ijerph15010078. PMID: 29304014; PMCID: PMC5800177.