Mercer University

Stetson-Hatcher School of Business

RESEARCH PROJECT

Which factors are most closely associated with diabetes diagnosis?

Course: BDA 602 - Statistics for Business Analytics

Name: Nhan Nguyen

Student ID: 11042660

1. Introdutction

Diabetes is a long-term medical illness that affects how our body converts food into energy. Most of our food consumed is converted by the body into sugar, which is then released into the circulation system. The hormone insulin is essential for allowing blood sugar into our cells as fuel. When the high level of blood sugar stays in the bloodstream, the body will not produce enough insulin, which can lead to serious health issues such as eyesight loss, and heart disease (Centers for Disease Control and Prevention, n.d.). More seriously, according to Centers for Disease Control and Prevention, the number of patients who are diagnosed with diabetes has doubled to 37 million patients, but one out of five of these people does not perceive they contract it. Therefore, the main purpose of this project is to identify which risk factors are most closely correlated with diabetes. These factors under the study include Body Mass Index (BMI), weight, glucose, and cholesterol,.. By identifying the risk factors that contribute to the development of diabetes, and healthcare professionals can develop targeted intervention strategies to prevent the disease more effectively.

2. Descriptive analysis

(i). The data sources used in this study included a dataset obtained from the online community of data scientists and machine learning practitioners. The dataset used is titled “Predict diabetes based on diagnostic measures” (Kaggle, n.d.) and it contains 16 variables related to diabetes disease, including glucose, cholesterol, BMI, height, weight, and diabetes status,..

(ii). Limitations: In the dataset, the used variables in this study may not include enough of all relevant variables that could impact the risk of diabetes. Secondly, the dataset used in this study was limited to certain geographic locations, which could affect the generalizability of findings.

(iii). Summary statistics table

- By applying **`st()`** from package **`vtable`** in R, it provides a summary of the statistical properties of a dataset, including the minimum and maximum, mean and median, the first and third quartiles. Figure 1 displays a table with the summary statistics of each variable in the dataset.

(i.) Data visualization

+ Histogram: The distribution of age of individuals with diabetes (Figure 2)

+ Box plot: The graph shows differences in glucose levels in individuals with Diabetes and Non-Diabetes (Figure 3)

3. Hypothesis Test

(i). - Null Hypothesis: The true population mean age of individuals who have diabetes is equal to 45 years old.

$$\mu - \mu 0 \geq 45$$

- **Alternative Hypothesis :** The true population mean age of individuals who have diabetes is different from 45 years old.

$$\mu - \mu 0 < 45$$

- The choice of 45 years old as the hypothesized value for the true population mean age in this study is based on epidemiological data indicating that a majority of individuals with diabetes in the US are over the age of 45, as reported by the Center for Diease Control and Prevention (CDC). Therefore, it is important to examine the validity of this assumption in order to better understand the range of ages having diabetes.

(ii). Calculate the test statistic and critical value

- Applying the **`t.test()`** function in R, the result for t-value is 7.9103. To calculate the critical value, the study uses **`qt()`** function in R with a significance level is 0.05 and degree of freedom (= 69), the result for critical value is 1.667

- It can be observed from the graph above (figure 4) that the critical value does not fall into the critical region, so it means that the test statistics fall within the acceptance region and the null hypothesis can not be rejected.

(iv). Confidence Interval

- Applying the t.test() function in R, the result of confidence interval is (- Inf, 61.23). This suggests that with 95% confidence, the true population mean of age is likely to fall into this interval. Moreover, the confidence interval contains the hypothesized value of 45. This means that we can not reject the null hypothesis at the 5% significance level. As a result, we can conclude that the null hypothesis is true which means that the true population mean of age who have diabetes is greater than 45.

4. Analysis of Variance (ANOVA)

(i). In this study, I want to compare the mean of glucose levels between individuals with diabetes and non-diabetes. The reason why I want to testify the glucose levels is because it is important to know that diabetes is a chronic condition affecting higher blood glucose in our body, which can lead to other serious health complications (American Diabetes Association, n.d).

(ii). Creating the null hypothesis and alternative hypothesis

+ Null hypothesis: There is no significant difference in mean glucose levels between individuals with diabetes and non-diabetes.

+ Alternative hypothesis: There is a significant difference in mean glucose levels between individuals with diabetes and non-diabetes.

- This study choose these hypotheses because it wants to determine if there is evidence to support the idea having diabetes affects glucose levels.

(iii). F-statistics and critical value

- In order to perform the ANOVA test, this study applies **`aov()`** function to extract the F-statistics and p-value. Then, the result for F-statistics is 350.8 and p-value <2e-16. By looking at the p-value, it is significantly smaller than alpha level 0.05 . So, we can conclude that there is strong evidence to reject the null hypothesis.

- Applying the **`qf()`** function in R, we can extract the critical value equal to 3.86. From this result, we can confidently reject the null hypothesis because the F-statistics are much larger than the critical value. Therefore, it can be concluded that there is a significant difference in glucose levels between individuals with diabetes and non-diabetes.

(iv). The broader implications of the finding would be that diabetes significantly affects glucose levels and therefore, it is important to monitor glucose levels in individuals with diabetes and manage it accordingly. Additionally, the finding can contribute to our understanding of how diabetes affects glucose levels, which can inform future research and interventions such as adjusting the eating diets, and more regularly checking on blood glucose levels.

5. Regression Analysis

(i). Research question: Which factors are significantly associated with diabetes diagnosis?

- **Purpose**: By applying regression analysis, this study can investigate the relationship between predictor variables and determine which factors are the most significant predictors of diabetes. This information is important to help not only patients but also healthcare professionals develop effective strategies for prevention and treatment.

(ii). Dependent variable: Diabetes (1 = Diabetes ; 0 = No Diabetes) .

It is of interest to understand its variation because diabetes is a serious health problem affecting millions of people over the world. If we don’t prevent or have a proper treatment, it could lead to other serous health complications such as cardiovascular disease, kidney damage, eysesight loss, and hearing impairment (Diabetes, 2023),

(iii). Independent variables (Predictors variables) : Cholesterol, Glucose, Hdl_chol (High-density lipoprotein cholesterol), height, weight, BMI (Body Mass Index), Systolic_bp (Systolic blood pressure), diastolic_bp (Diastolic blood pressure), Waist, Hip. These variables were included in the regression analysis because they have been previously identified as potential risk factors for diabetes. Understanding the relationship between these variables and development of diabetes can help to identify individuals who may be at increased risk and inform targeted prevention and intervention strategies.

(iv). Conducting regression analysis: This study using function **`lm`** in R to conduct the regression analysis, and applying function **`summary`** to include the result of this regression model, and the result table is displayed in figure 5 of appendix.

(v). The result shows that only **glucose** and **cholesterol** are both significant values with diabetes because the p-value is smaller than significant level 5%. Other variables are not significant values because all of them has p-value is greater than 0.05. The coeffcient for glucose is estimated to be 0.0044, which means that a one-unit increase in glucose is associated with a 0.0044 unit increase in risk of diabetes, holding all other variables constant. Similarity, a one-unit increase in cholesterol is associated with 0.0007 unit increase in risk of diabetes, holding all other variables constant.

(vi). One surprising result that weight, and BMI do not have significant coefficients in predicting diabetes. This may seem counterintuitive since obesity is a known risk factor for diabetes. However, the lack of significance may be due to the relatively small sample size or the inclusion of other variables that account for the variance in diabetes risk.

(vii). Strength of model: The R-squared value of the model is 0.4955, which means that about 49.55% of the variance in the dependent variable (Diabetes) is explained by the independent variables in the model. This suggests that the model has moderate predictive power. However, the standard error of the estimate is 0.26, which means that the predicted values of the dependent variable may deviate from the actual values by an average of 0.26 units. This suggest that the model has some degree of error in its predictions.

(viii). To strengthen the model and its accuracy, there are some recommended variables should be included:

+ **Family history:** A family history is known as a risk factor of diabete (CDC, 2022), this could be a factor to diagnose diabetes.

+ **Gestational diabetes:** This factors could affect to newborn baby if a mother contracted to diabetes while pregnancy (CDC, 2022)

+ **Physical activities:** If a people who are inactive less 3 times a week could lead to a risk of being diabetes (CDC, 2022)

(ix). Based on the result of regression model, it can be concluded that glucose and cholesterol are important predictors of diabetes. However, other factors such as, blood pressure, BMI, height, and weight do not appear to be significantly associated with diabetes in this model. These findings can help healthcare professionals to better understand the risk factors for diabetes and develop targeted strategies for better prevention and treatment for patients with this condition.

6. Conclusions

(i). Based on the hypothesis testing of this study, it can be concluded that the true population mean of individuals who have diabetes is greater than 45 years old. This finding can help both patients and healthcare professionals have a better understanding of the range of ages getting risk of diabetes, so they can have an immediate intervention for this disease.

Secondly, from the ANOVA testing, the finding shows that the glucose levels are significantly affect diabetes. Therefore, in order to minimize the risks of diabetes, we should adjust glucose levels by having a healthy diet, involving the physical activities and more regularly check it.

From the regression analysis of this study, it can be concluded that glucose and cholesterol are imporant predictors of diabetes. This finding can help healthcare professionals to better understand the risk factors for diabetes and develop targeted strategies for better prevention and treatment for patients with this condition.

(ii). Limitations : This study has a data with a relatively small size with only 390 observations may not be representative of the population. A larger sample size may improve the accuracy and statistical power of the analysis. Furthermore, this study included limited variables, there are many important predictors for diabetes such as family history, gestational diabetes,… that are not included in this study.

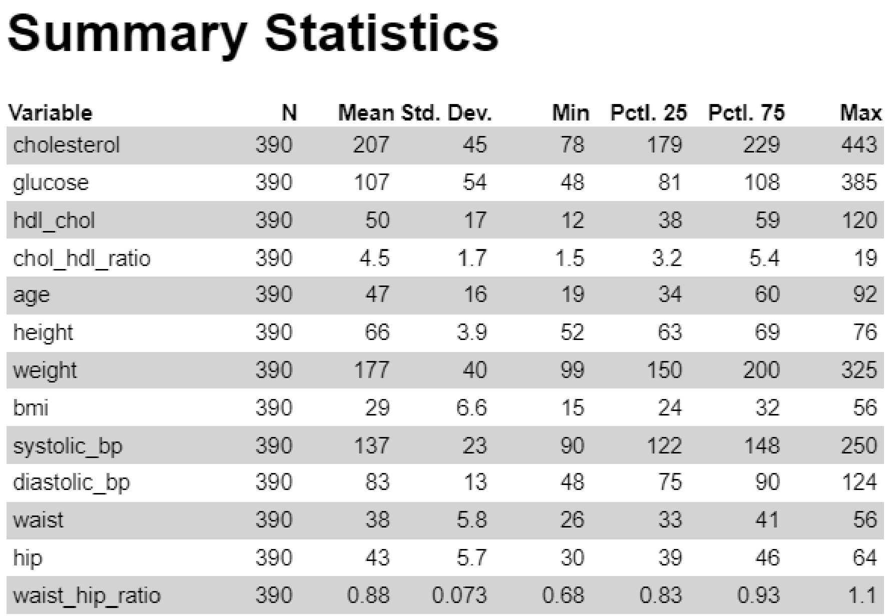
Possible areas for future research: Examining the effects of other potential risk factors such as genetics or environmental factors. Or, investigating the impact of lifestyle factors on the development of diabetes such as diet and exercise.

(iii). There are some suggested research questions such as:

+ Can genetic factors be incorporated into the analysis to better predict an individual's risk of diabetes?

+ How do lifestyle factors, such as diet and exercise, affect the risk of developing diabetes?

Appendix



Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
cholesterol	390	207	45	78	179	229	443
glucose	390	107	54	48	81	108	385
hdl_chol	390	50	17	12	38	59	120
chol_hdl_ratio	390	4.5	1.7	1.5	3.2	5.4	19
age	390	47	16	19	34	60	92
height	390	66	3.9	52	63	69	76
weight	390	177	40	99	150	200	325
bmi	390	29	6.6	15	24	32	56
systolic_bp	390	137	23	90	122	148	250
diastolic_bp	390	83	13	48	75	90	124
waist	390	38	5.8	26	33	41	56
hip	390	43	5.7	30	39	46	64
waist_hip_ratio	390	0.88	0.073	0.68	0.83	0.93	1.1

Figure 1: Summary statistics table

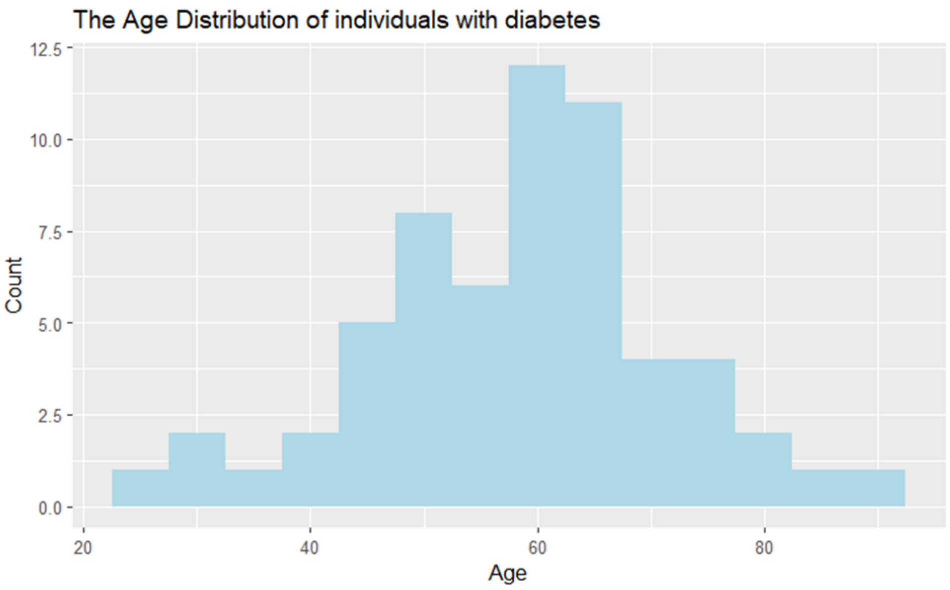


Figure 2. The distribution of age of individuals with diabetes

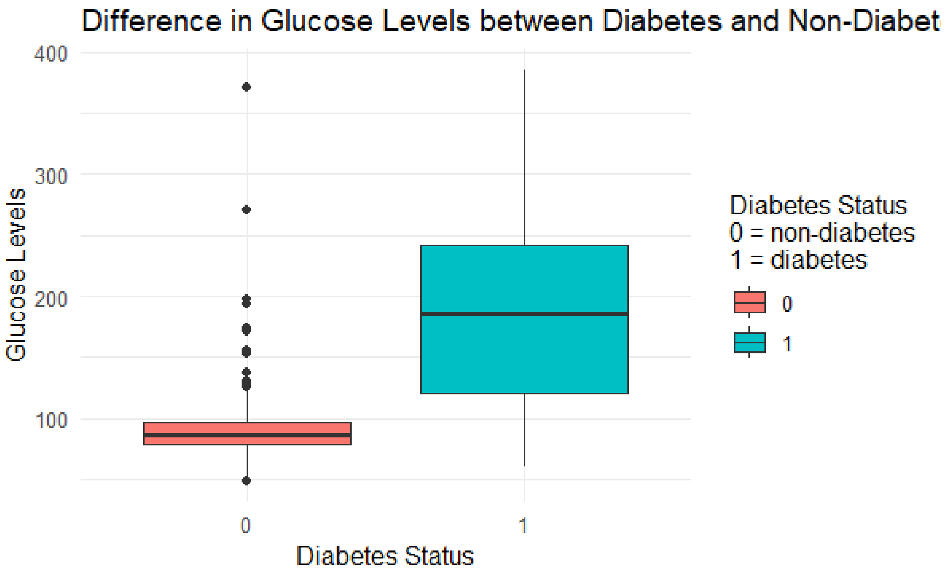


Figure 3: The graph shows the difference in glucose levels between Diabetes and Non-Diabetes

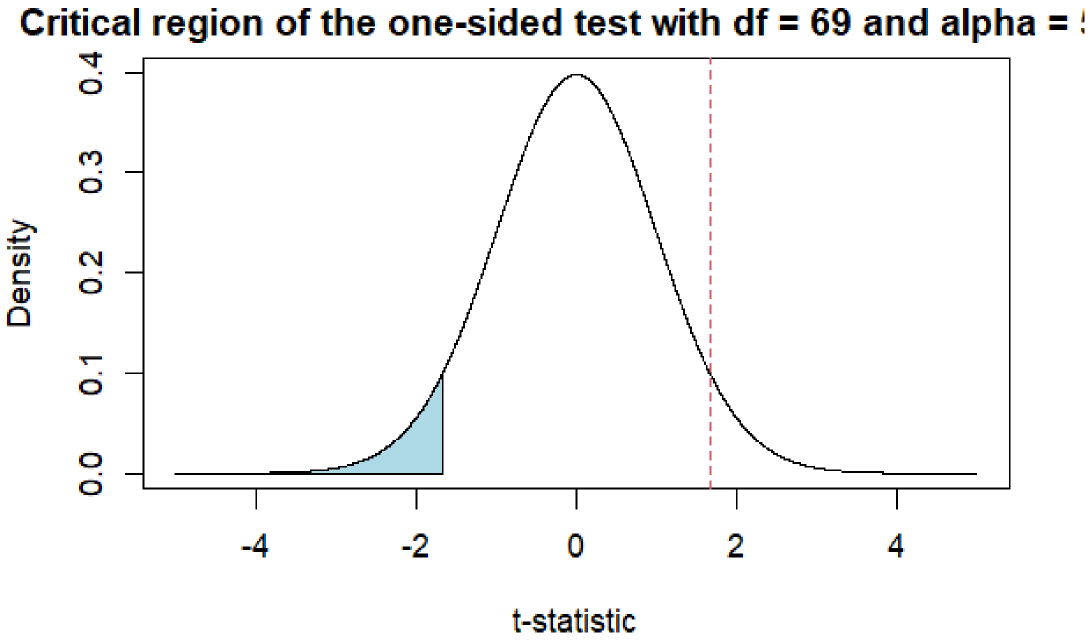


Figure 4: The graph shows the critical region of one-sided test

Coefficients				
	Estimate	Std.Error	t-value	Pr(> |t|)
(Intercept)	-0.8981008	0.9897599	-0.907	0.3648
Cholesterol	0.000685	0.000316	2.168	0.0308
glucose	0.0043724	0.0002616	16.716	<2e-16
hdl_chol	-0.0005633	0.0008354	-0.674	0.5005
Height	0.0041846	0.0149255	0.28	0.7794
Weight	-0.00159	0.0027989	-0.568	0.5703
Bmi	0.0091308	0.0172091	0.531	0.596
systolic_bp	0.0012488	0.000766	1.63	0.1039
diastolic_bp	-0.0008025	0.0012649	-0.634	0.5262
waist	0.0063786	0.005067	1.259	0.2089
Hip	-0.003108	0.0055478	-0.56	0.5757
Residual standard error: 0.26 on 379 degrees of freedom				
(1 observation deleted due to missingness)				
Multiple R-squared: 0.4955, Adjusted R-squared: 0.4822				

F-statistic: 37.23 on 10 and 379 DF, p-value: < 2.2e-16				

Figure 5: The table of result of regression analysis

References:

1. Centers for Disease Control and Prevention. (n.d.). Diabetes basics. Retrieved from https://www.cdc.gov/diabetes/basics/diabetes.html
2. Kaggle. (n.d.). Predict diabetes based on diagnostic measures. Retrieved from https://www.kaggle.com/datasets/houcembenmansour/predict-diabetes-based-on-diagnostic-measures

3. Centers for Disease Control and Prevention. (2022, December 30). *Type 2 diabetes*. Centers for Disease Control and Prevention. Retrieved April 26, 2023, from https://www.cdc.gov/diabetes/basics/type2.
4. *Blood glucose and insulin at work*. Blood Glucose and Insulin at Work | ADA. (n.d.). Retrieved April 27, 2023, from https://diabetes.org/tools-support/diabetes-prevention/high-blood
5. Mayo Foundation for Medical Education and Research. (2023, January 20). *Diabetes*. Mayo Clinic. Retrieved May 1, 2023, from https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444
6. Centers for Disease Control and Prevention. (2022, April 5). *Diabetes risk factors*. Centers for Disease Control and Prevention. Retrieved May 3, 2023, from https://www.cdc.gov/diabetes/basics/risk-factors.html