Research on Influential Factors on Glucose

Yanyi Wang (1004901341)

INF2178H S LEC0101

Shion Guha

Mar.28th 2022

Abstract

There are two main types of diabetes, which are classified as type 1 diabetes and type 2 diabetes. Type 1 diabetes have insulin deficiency due to destruction of insulin secreting cells by immune system, and most happened among teenagers. Type 2 diabetes, on the other hand, reduced insulin sensitivity/secretion, and mostly happened among middle aged people with unhealthy living habit. Based on the research, there are about 10% diabetes in the world belongs to type 1 diabetes. So, in this study, I will focus on the type 2 diabetes, which is an acquired illness that can be avoid by healthier living habit. Additionally, there is also a special type named Gestational diabetes, for the situation of high blood glucose during pregnancy. And the cause of Gestational diabetes is the placental hormones that cause insulin resistance. Without treatment, Gestational diabetes will increase the probability of mother getting type 2 diabetes, and the baby have high birthweight, getting type 2 diabetes and obesity.

This research explores some diabetes-related indicators using Diabetes Dataset. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes[1]. It has 768 observations and 9 variables. This research will focus on type 2 diabetes considering it happened more frequently today when the living habit tend to be high in sugar and calories. This made me interested in the role glucoses is playing in diabetes. I would like to apply EDA to variables ‘Insulin’, ‘Glucose’, ‘Age’, and ‘Outcome’ in the diabetes dataset. Insulin is the recording of 2-Hour serum insulin, Glucose represents Plasma glucose concentration 2 hours in an oral glucose tolerance test. Based on study of diabetes, I come up with the questions05 if glucose is the key factors leading to change in the insulin level and if glucose has significant

[1] (Ashfaq, Diabetes data set 2021)

differences between different age groups using different modeling techniques. Significance level in this research is set as 0.05.

Data Wrangling

To test the relationship between those three variables, variable 'Insulin', 'Glucose', and 'Age' of the dataset should be used for the future modeling. Considering the topic of this dataset, I found the median insulin level is different between observations with diabetes and without diabetes. Thus, I split the dataset into two subsets, one for diabetes patient and another one for non-diabetes people. To clean the data, I have a glimpse of the original dataset found 'Insulin' of some variables are zero, but the insulin level should never be zero for any person, so it is reasonable to consider those observations with 0 insulin levels as missing values. To decide replacing the missing values with median or mean, we have a visualization of the insulin values including the missing values. According to the boxplot for variable Insulin in figure 1 and 2, there are lots of outliers. So, I replace the zero Insulin values for diabetes patient with median value of the Insulin for observations with outcome = 1, at the same time replace the zero Insulin values for non-diabetes patient with median value of the Insulin for observations with outcome = 0. Then append those two subsets together forms the whole dataset.

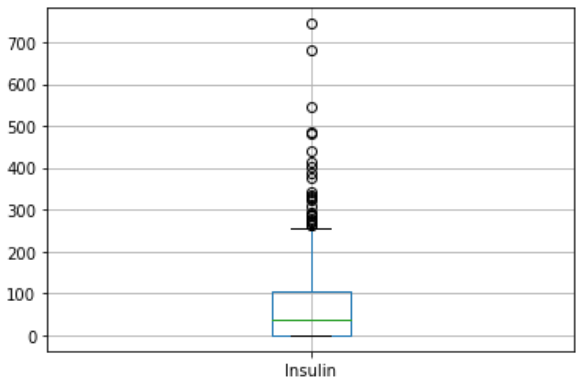

Figure 1. Boxplot of Insulin for non-diabetes observations

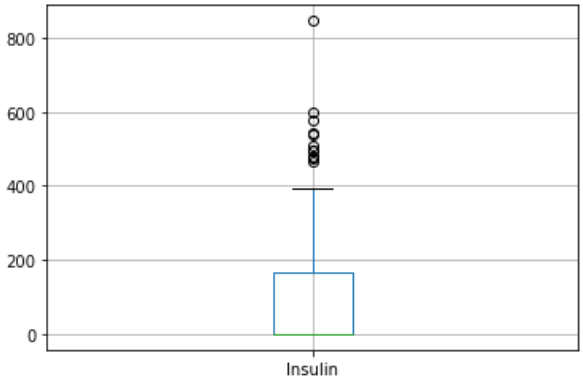

Figure 2. Boxplot of Insulin for diabetes observations

In order to do the statistical testing, I add two new variables to the diabetes dataset named 'AgeGroup' and 'InsulinLevel'. 'AgeGroup' groups observations into three different age groups. For observation with 'Age' between 0 and 35, they have the 'AgeGroup' value of 'YongeAdult'. For observation with 'Age' between 35 and 65, they have the 'AgeGroup' value of 'Adult'. For observation with 'Age' more than 65, they have the 'AgeGroup' value of 'Elderly. 'InsulinLevel' groups observations into three different groups. For observation with 'Insulin' between 0 and 300, they have the 'InsulinLevel' value of 'Low_Insulin'. For observation with 'Insulin' between 300 and 600, they have the 'InsulinLevel' value of 'Mid_Insulin'. For observation with 'Insulin' between 600 and 900, they have the 'InsulinLevel' value of 'High_Insulin'.

Data Visualization

To better understand the dataset and features of the three variables related to the research question, I applied both bar plot and scatterplot to the cleaned dataset. First for the scatterplot which displays a relationship between two sets of data. As shown in figure 3, I take Glucose as the response variable on the y-axis, and Insulin value as the explanatory value on x-axis. Additionally, to compare the difference between age groups, I choose to use colored scatters

representing the three different age groups. From the scatterplot, we can see there are some outliers. Most of the observations has glucose between 50-200, and there is an Inconspicuous positive correlation between Glucose and Insulin shown in the graph. Moreover, young adults tend to have a lower glucose level comparing to the adult age group.

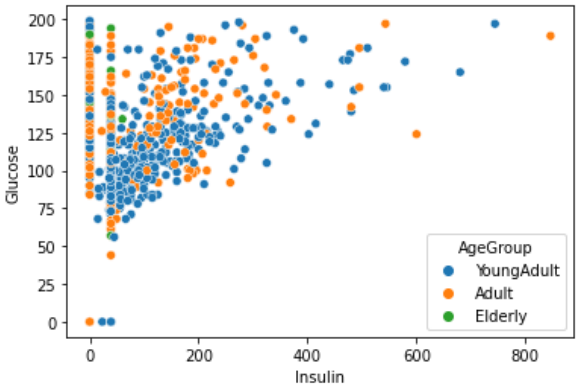

Figure 3. Scatterplot of Insulin vs. Glucose

In the boxplot shown in figure 4, I also take Glucose as the response variable on the y-axis, but changed to new variable 'InsulinLevel' as the explanatory value on x-axis. In the boxplot, first it is noticeable that there are some outliers for low insulin level of the young adult age group, there are some observations have glucose higher that the upper whisker. Second, we can see the Elderly age group only appears at the low insulin level, while only the young adult age group have observations with high insulin level. Third, we can again see the positive relationship between Insulin level and Glucose. By comparing the position of boxplot for different insulin levels, we can see that glucose is highest for the high insulin level group, and then followed by the middle insulin level group, low insulin groups have the lowest glucose level as well.

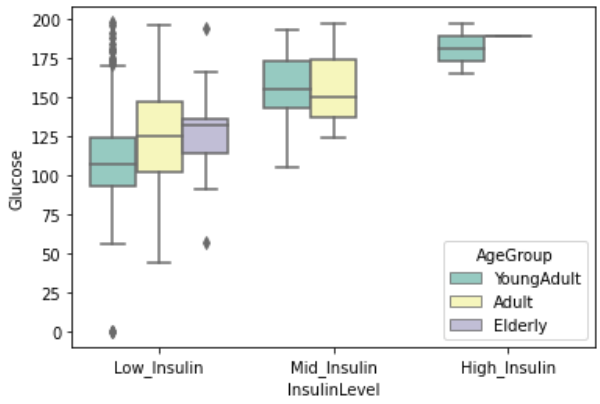

Figure 4. Boxplot of Glucose for different Insulin Levels

Statistical Testing

T-Test:

I applied the independent t test to the two subsets of the diabetes dataset, which is trying to identify if there is any difference in Glucose between the diabetes patient and non-diabetes observations in the dataset. The null hypothesis of t-test is that the difference between group means is zero. The result of the independent t test of the two groups is shown in table 1. From the result we can see p-value is about 2.64e-36, which is much less than the significance level 0.05. Thus, the independent t-test results are significant, therefore, we can reject the null hypothesis in support of the alternative hypothesis. From the t-test result, we can conclude that the mean of Glucose value is different for observations with diabete group and non-diabetes group. Then based on this conclusion, I will further explore the relationship between Insulin level, Glucose value, and age group.

P-Value	2.6441613495403223e-36
T-Statistic	-13.751537067396411

Table 1. Result of T-Test

Two-Way ANOVA:

ANOVA stands for analysis of variance and tests for differences in the effects of independent variables on a dependent variable. And the two-way ANOVA test is a statistical test used to determine the effect of two nominal predictor variables on a continuous outcome variable2. Followed by t-test, I am going to use the two categorical and one numerical variable chosen from the diabetes dataset modeling the two-way ANOVA. There are 2 independent variables, 'InsulinLevel' and 'AgeGroup', and both variables have 3 levels respectively. With those three variables, I am going to test the effect of age group on glucose, effect of insulin level on glucose, and effect of age group and insulin level interactions on glucose.

The result of two-way ANOVA is as shown in table 2. The P-value is getting from ANOVA analysis for 'AgeGroup' is 1.91e-08, for 'InsulinLevel' is 3.05e-16 and interaction is 3.31e-01. The P-value for both 'AgeGroup' and 'InsulinLevel' are both less than 0.05, which is statistically important. Thus, it is able to conclude that both age group and insulin level importantly affect the yield glucose. But the P-value for interaction is greater than significant level, so it is not able to conclude that interaction of both age group and insulin level importantly affects the yield outcome.

	df	sum_sq	mean_sq	F	PR(>F)
C(AgeGroup)	2.0	50856.204808	25428.102404	32.429142	1.905980e-08
C(InsulinLevel)	2.0	110615.980712	55307.990356	70.535765	3.054868e-16
C(AgeGroup):C(InsulinLevel)	4.0	3469.911911	867.477978	1.106318	3.314236e-01
Residual	623.0	488502.222742	784.112717	NaN	NaN

Table 2. Result of Two-Way ANOVA

It is possible to realize that age group and insulin level differences are statistically Important. But we cannot conclude which age group and insulin level are significantly different from each other from ANOVA. In order to get the pairs of significant different age group and

[2] (Hayes, Two-way ANOVA 2022)

insulin level, I will apply the perform multiple pairwise comparison analysis by Tukey HSD test. As well as Levene and Shapiro-Wilk test to validate the assumptions for homogeneity of variances and normal distribution of residuals.

Post-hoc Comparison:

Tukey's HSD test is conservative and increases the critical value to control the experiment wise type I error rate[3]. To know the pairs of significant different age group and insulin level, I will perform post hoc comparison analysis using Tukey's HSD test. The results of post hoc comparisons are as shown in table 3 and table 4. Table 3 shows the post-hoc test result of pairs of significant different age group, while table 3 shows the result of pairs of significant different insulin levels. From result shown in table 3, P-value for YoungAdult-Adult is less than 0.001, for YoungAdult-Elderly is 0.30, and for Adult-Elderly is 0.90. Thus, Tukey's HSD suggests that except YoungAdult-Adult, all other pair wise comparison for age groups have a p-value higher than the significance level, which fail to rejects the null hypothesis, indicates statistical significant differences between the young adult age group and adult age group. Then from result shown in table 4, P-value for Low_Insulin-Mid_Insulin is less than 0.001, for Low_Insulin-High_Insulin is also less than 0.01, and for Mid_Insulin-High_Insulin is 0.22. Thus, Tukey's HSD suggests that except Mid_Insulin-High_Insulin, other two pair wise comparison for insulin level have a p-value less than the significance level, which rejects the null hypothesis, indicates statistically significant differences between the low insulin level group and middle insulin level group, and also between the low insulin level and high insulin level group.

[3] (Bedre, ANOVA using python (with examples) 2022)

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	YoungAdult	Adult	14.324598	8.515279	20.133918	8.192718	0.001000
1	YoungAdult	Elderly	14.070455	-8.081939	36.222848	2.110369	0.295835
2	Adult	Elderly	0.254144	-22.213697	22.721984	0.037583	0.900000

Table 3. Post Hoc Test for AgeGroup

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Low_Insulin	Mid_Insulin	41.000992	29.399548	52.602436	11.742299	0.001000
1	Low_Insulin	High_Insulin	68.961776	30.883241	107.040312	6.017264	0.001000
2	Mid_Insulin	High_Insulin	27.960784	-11.662079	67.583648	2.344630	0.222687

Table 4. Post Hoc Test for InsulinLevel

ANOVA Assumption Testing

QQ-Plot:

ANOVA assumptions can be checked using visual approaches such as residual plots, the visual approaches perform better than the statistics test. So, I will first generate QQ-plot from standardized residuals and the result is show in figure 5. From the image we can see most of the standardized residuals lie around the 45-degree line, which suggests that the residuals are approximately normally distributed. But we can also see there are 3 outliers lying at the left bottom of the plot.

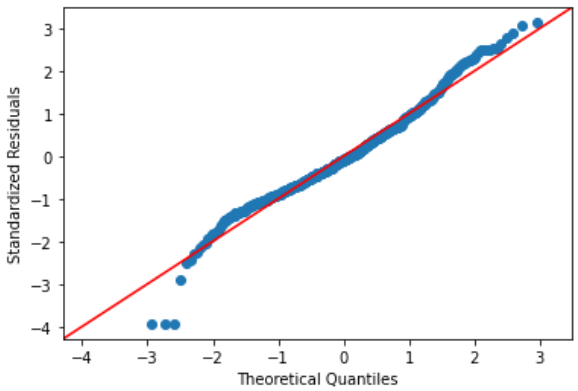

Figure 5. QQ plot

Histogram:

Followed by QQ-plot, I will generate histogram as another visual approaches to check the ANOVA assumptions. As show in figure 6, the distribution of the histogram is approximately normally distributed. And there are also some outliers with residuals less than -100.

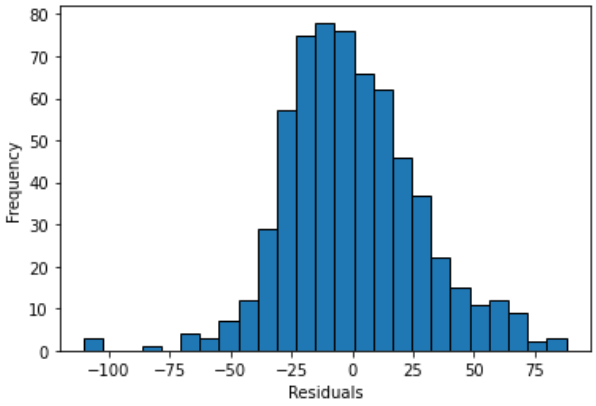

Figure 6. Histogram

Shapiro-Wilk Test:

Shapiro-Wilk test can also be used to check the normal distribution of residuals. The null hypothesis of Shapiro-Wilk test is that data is drawn from normal distribution. As shown in table 5, p-value is 2.96e-8, which is less than the significance level. Thus, we reject the null hypothesis and conclude that data is not drawn from normal distribution.

P-Value	2.962094924896519e-08
w	0.9775135517120361

Table 5. Shapiro-Wilk test

Levene's Test:

Levene's test can be used to check the homogeneity of variances when the data is not drawn from normal distribution. The results of Levene's test are shown in table 6 and table 7. As shown in table 6, the p-value is 0.04 which is significant. Thus, the p-value reject the null hypothesis that age groups do not have the same an equal variance. Then show in table 7, the p-

value is 0.298 which is non-significant. The p-value fail to reject the null hypothesis, which implies that insulin level groups have the same an equal variance.

	Parameter	Value
0	Test statistics (W)	5.4248
1	Degrees of freedom (Df)	2.0000
2	p value	0.0046

Table 6. Levene's Test for AgeGroup

	Parameter	Value
0	Test statistics (W)	1.2127
1	Degrees of freedom (Df)	2.0000
2	p value	0.2981

Table 7. Levene's Test for InsulinLevel

Multiple Linear Regression (MLR)

After testing the glucose of different age groups and insulin level groups with two-way ANOVA, I am going to test relationship between glucose and four variables in the diabetes dataset with MLR. In the MLR, we will have 'Glucose' as the dependent variable again. And for the independent variable, we will have 'Age', 'Insulin', 'Pregnancies' and 'DiabetesPedigreeFunction'.

In case of multivariable linear regression, the regression model must find the most optimal coefficients for all the attributes[4]. And table 8 shows the result of coefficients our regression model has chosen. Those coefficients means that for a unit increase in insulin, there will be a 0.09 unit decrease in the glucose level. Similarly, a unit increase in pregnancies will

[4] (Robinson, *Linear regression in python with scikit-learn* 2021)

result in an increase of 0.0367 unit decrease in glucose level. And a unit increase in age will result in an increase of 0.7188 unit decrease in glucose level. Then, a unit increase in 'DiabetesPedigreeFunction' will result in an increase of 5.0844 unit decrease in glucose level. Thus, we can conclude that 'DiabetesPedigreeFunction' have the biggest effect on the glucose level. And all four variables have a positive relationship to the glucose value.

	Coefficient
Insulin	0.093708
Pregnancies	0.036723
Age	0.718811
DiabetesPedigreeFunction	5.084430

Table 8. MLR Coefficient

Then I execute the predictions with the regression. And then the final step is to evaluate the performance of algorithm. The output in table 9 shows the values for MAE, MSE, and RMSE. The value of mean absolute error (MAE) is 22.99, which represents the difference between the actual and predicted values extracted by averaged the absolute difference. And mean square error (MSE) is 835.09, which represents the difference between the actual and predicted values extracted by squared the averaged difference. And we can see the value of root mean squared is 28.98, which is which is much lower than the mean of glucose value. This means that out algorithm was not accurate to make reasonably good predictions. This can either cause by small dataset, bad assumptions, or poor features which doesn't have high enough correlations.

```
MAE: 22.990080682333573
MSE: 835.0926609520016
RMSE: 28.897969841357394
```

Table 9. MAE&MSE&RMSE

Result

In this research, I explore the glucose level difference between different age groups and for different insulin levels, the result also implies the relationship between those three variables. According to the t-test, I confirm the glucose level is different between diabetes patient and non-diabetes people. And with two-way ANOVA, the result suggests both age group and insulin level importantly affect the yield glucose value. And combine the visualization of the data set, the two-way ANOVA implies that high insulin level can lead to a higher level of glucose, and with same level of insulin, elder people have a higher glucose level as well. The post hoc indicates the statistically significant differences between the young adult age group and adult age group. It also indicates statistically significant differences between the low insulin level group and middle insulin level group, and between the low insulin level and high insulin level group. Then to test the authenticity of ANOVA result, I applied test to the ANOVA assumptions. Both QQ-Plot and Histogram suggests the residuals are approximately normally distributed, but Shapiro-Wilk test result suggest the opposite way. And Levene's Test confirms the homogeneity of variances. Based on the test results, the used variables satisfy the assumptions of the two way ANOVA, and the two-way ANOVA result is reliable.

As other medical research shows, the type 2 diabetes embodied in reduction of insulin sensitivity/secretion, which more frequently happened among elderlies. And the mechanism for the reduced insulin secretion is started with glucose rapid increasing with no organ reactions to the insulin. Based on visualization of the diabetes dataset and statistical testing, we observe the interactive influences between those three factors, and the relationship of glucose with those two factors.

Discussion

In this research, I have found that high insulin level can lead to a higher level of glucose, and with same level of insulin, elder people have a higher glucose level. Thus, based on this result I would consider glucose as one of the key factors reflecting a people's health status. It not only reflects risk of individuals getting diabetes, but also an effective indicator of other related illness. This makes glucose level one of the most important physical indicators. From the research, we can see high glucose level has happening more often among elder people, that can also imply why type 2 diabetes also happened more often among elderly. And this is a good warning for people keeping a healthy lifestyle, especially for elder people.

Then, in the MLR part, it is noticeable that the RMSE for the MLR suggests that the algorithm was not accurate to make reasonably good predictions. As mentioned above, this inaccuracy can be caused by small dataset, bad assumptions, or poor. Considering those reasons, there are few ways to improve the MLR prediction. First is to get a larger dataset including more observations, which could help improving the accuracy. Second is to visualize the original dataset and determine a better assumption. Third is to choose features that has high correlation to the Glucose.

Reference

Ashfaq, H. (2021, September 18). Diabetes data set. Kaggle. Retrieved March 27, 2022, from https://www.kaggle.com/hassanashfaq2001/diabetes-data-set

Bedre, R. (2022, March 6). ANOVA using python (with examples). Data science blog. Retrieved March 27, 2022, from https://www.reneshbedre.com/blog/anova.html

Hayes, A. (2022, February 8). Two-way ANOVA. Investopedia. Retrieved March 27, 2022, from https://www.investopedia.com/terms/t/two-way-anova.asp

Robinson, S. (2021, June 7). Linear regression in python with scikit-learn. Stack Abuse. Retrieved March 27, 2022, from https://stackabuse.com/linear-regression-in-python-with-scikit-learn/