**Best Equation to Predict Y(Glucose)**

**DS 853\_01**

**Arpan Avvari**

**SFSU**

**12/14/2022**

**Table

Description automatically generatedTable

Description automatically generated**

**Data Explanation:**

My original dataset consists of 9 columns and 768 rows data, so I felt it better to drop the outcome and Pregnancies columns which doesn’t give lot of information to what I want to find. I wanted to find out that, do six coloums such as BloodPressure, SkinThinkness, Insulin, BMI, Age, and DiabetesPedigreeFunction have influence on Glucose level.

The Y column Glucose express the Glucose level in blood, And X dataset attributes represent:

* BloodPressure: To express the Blood pressure measurement
* SkinThickness: To express the thickness of the skin
* Insulin: To express the Insulin level in blood
* BMI: To express the Body mass index
* DiabetesPedigreeFunction: To express the Diabetes percentage
* Age: To express the age

**Multi regression Equation:**

I chose Forward method to find the best equation to predict my Y variable. I wanted to find:

***Which of 6 X variables might be used to predict Y which is Glucose?***

**Y= 75.796\* 0.093 (Insulin level) + 75.796\*1.022(Age)**

**Y = mX1 + mX2**

I would be able to predict the Glucose Y using the above equation. I would interpret the equations as Y is the dependent variable of the regression which is to predict Glucose level. The M is the slope of the regression, x1 is the first independent variable of the regression, and x2 is the second independent variable of the regression.

**Explanation of independent variables:**

To explain the two independent variable that’s helps to predict the Y in equations are the Insulin and the age.

Glucose Level

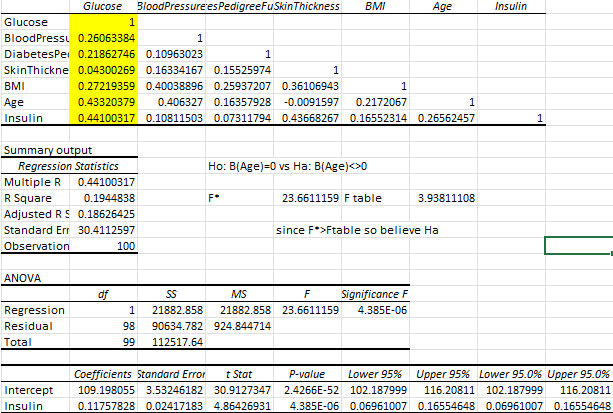
High Glucose level is dues to lack of enough insulin, higher the glucose is associated with diabetes, rapid heartbeat, excessive hunger and thirst, and vision problems.

Insulin: The insulin is measured within 30 minutes after food, Hyperglycemia (high blood glucose) means there is too much sugar in the blood because the body lacks enough insulin. The normal range of insulin should be between 30-230 mIU/L. That lack of insulin is called as hyperglycemia, which can lead to serious health problems. If there's not enough insulin, glucose can't get into your cells. It stays in the bloodstream instead which harmful for the health.

Age: Age is the fact the fact the production of the Insulin, as the age increase, people should exercise enough and maintain diet, if that’s not maintained, that would reduce the production of insulin, further it would influence the Glucose.

**Explanation of Forward multi–Regression Method through Excel**

I have taken the highest number in the correlation table, which is insulin, then ran the regression to find the coefficients, then I have done the linearity test using F\* and F Table; as the F\* is higher than the F table, that means that the insulin is part of equation to predict Y. To point out, I used 5 % significance while doing the F table.



I have created the residual variable 1 for glucose and insulin variable:

Glucose – (b0 Glucose coeff + b1 insulin coeff \* Insulin)

Then ran the correlation including the residual variable 1, to find the second highest variable that could best used to predict Y.

Table

Description automatically generated

The results showed that the Age as the second highest variable in the residual 1to predict Y. I ran the regression in Glucose, Age, and Insulin. I did the linearity test, the results shows that the F\* is highest then the F table, which means that, Age and Insulin are helpful in predicting Y.

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, table, Excel

Description automatically generated

Then, created Residual 2 using the age and Insulin:

Glucose – (b0 Glucose coeff + b1 insulin coeff \* Insulin variable + b2 Age coeff \* age variable)

To check the linearity test, I ran the regression, did the F\* and F table. The F\* is smaller then the F table, which represents that, BMI is not need in the equation to predict Y.

That explains that Age and insulin helps to explain the Glucose when they are in the equation.

**Normality**

Normality can be tested by checking weather the standard residuals ei/se is grouped under 1st, 2ns, 3rd standard derivation. If +/- 1st =~ 68% of the points, +/- 2nd =~ 95% of the points, +/- 3rd =~ 99% of the points are found, then we conclude that the data is normality distributed.

Table

Description automatically generated with low confidence Chart

Description automatically generated

The analysis in statistics show that the Glucose & Age, Glucose & Insulin condition that dataset is normality distributed.

**Homoscedasticity**

The Ratio od the standard deviation of error between S1-50 & S 50-100 should be in the range of 0.5 to 2.

Table

Description automatically generated

Calculating the ration between the Glucose & Age is 0.5<= 1.01 <= 2 which is true.

Calculating the ration between the Glucose & Insulin is 0.5<= 1.79 <= 2 which is true.

**R(value)**

R is a correlation coefficient that assesses both the direction and intensity of a relationship between two variables on a scatterplot. Always between a negative and a positive number, r's value. Values between the ranges make its true that the data is strong enough to predict the because the correlation between age & glucose, insulin & glucose.

age & glucose r vale is 0.433204, and insulin & glucose r value is 0.441003; to see this is other way the insulin & glucose has higher vale and it’s the better predictor.

**Jackknife**

The jackknife is a method used to estimate the variance and bias of a large population.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | B1 | | B0 | |
| Average | 1.302523 | | 73.10749 | |
| Stdv | 0.028282 | | 1.013432 | |
|  |  | |  | |
| Insulin | B1 | B0 | |
| Average | 0.29239 | 139.6553 | |
| Stdv | 0.02437 | 1.33033 | |

**Cross Validation**

A statistical technique called cross-validation is used to gauge the effectiveness (or accuracy) of machine learning models. It serves as a safeguard against overfitting in predictive models, especially when the available data may be scarce.

**Table

Description automatically generated**

**Table

Description automatically generated**