PROJECT REPORT

Unveiling Patterns in Health: A Statistical Investigation of Diabetes Risk Factors in Pima Indian Women

Introduction

As a part of this project, we explore the comprehensive dataset sourced from the National Institute of Diabetes, focusing on Pima Indian women aged 21 and above.Through meticulous statistical analysis, we aim to unravel intricate connections, addressing crucial questions regarding pregnancies, BMI fluctuations, and familial susceptibility to diabetes. Employing diverse techniques, including two-sample t-tests, one-sample hypothesis testing, and both multiple and simple linear regression, our investigation is tailored to enhance comprehension of the intricate interplay between various factors influencing diabetes within this specific demographic.

Methods

The dataset used contains 767 observations with the following variables :

Key variables:

Continuous Variables: Glucose, BloodPressure, SkinThickness, Insulin, BMI,DiabetesPedigreeFunction(DPF), Age.
Categorical Variables: Outcome (response variable, binary - 0 or 1).
Discrete Variables: Pregnancies.

Response variable :

1. ***Outcome*** - This is a binary variable indicating whether a patient has diabetes (1) or does not have diabetes (0).

Thorough examination of the dataset revealed compelling questions addressable through statistical analysis.The following questions were raised and addressed:

Analysis Question 1: Do the number of pregnancies differ between diabetic and non-diabetic women?
Aim: The analysis will help determine whether female individuals with diabetes exhibit a significantly different number of pregnancies compared to those without diabetes
Statistical Technique 1: Employed a Two-Sample t-Test for unequal variances for comparing means using the variables Pregnancy and Outcome, facilitating the determination of proposed variations in the number of pregnancies among female individuals with and without diabetes.
Statistical Technique 2: Employed test Two-Sample t-Test for equal variances facilitating the determination of significant variations in the number of pregnancies among female individuals with and without diabetes.

Analysis Question 2: Does the BMI (Body Mass Index) of the diabetic patients differ from theNational Average?
Aim: The analysis will help determine if the average BMI (Body Mass Index) of diabetic patients is significantly different from the national average BMI of 28.
Statistical Technique 3: Utilized Hypothesis Testing for One-Sample, using the variable BMI , enabling a conclusive assessment of the average BMI of diabetic patients in relation to the established national benchmark.

Analysis Question 3: What is the association between a person's family history and the other features?
Aim: Explored the association between a person's family history and other features, seeking optimal predictors for DiabetesPedigreeFunction.
Statistical Technique 4: Applied Multiple Linear Regression to identify the key features influencing DiabetesPedigreeFunction, contributing to a nuanced understanding of familial predispositions to diabetes.

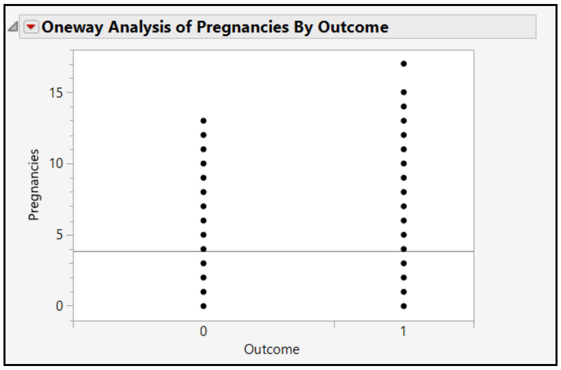
Results

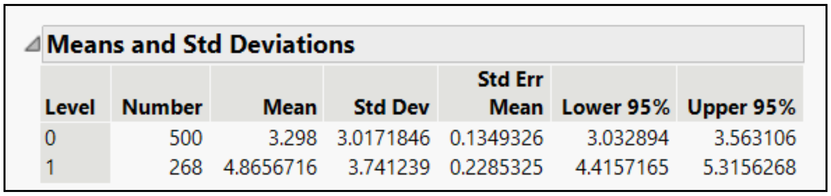
Analysis Question 1:

To determine whether pregnancies differ between diabetic and non-diabetic women, we perform a Two-Sample t-Test for comparing means.
In this scenario,
We **assumed** that the populations are **normal distribution** and α = **0.05**
μ_2 = Mean of pregnancies of non diabetic women

μ_1 = Mean of pregnancies of diabetic women

Null Hypothesis : $H_0 : \mu_2 = \mu_1$ **Alternate Hypothesis :** $H_1 : \mu_2 - \mu_1 \neq 0$

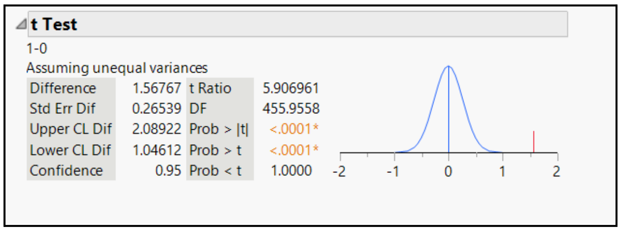



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
0	500	3.298	3.0171846	0.1349326	3.032894	3.563106
1	268	4.8656716	3.741239	0.2285325	4.4157165	5.3156268

We proceed with the **Two-Sample t-Test** for comparing means:

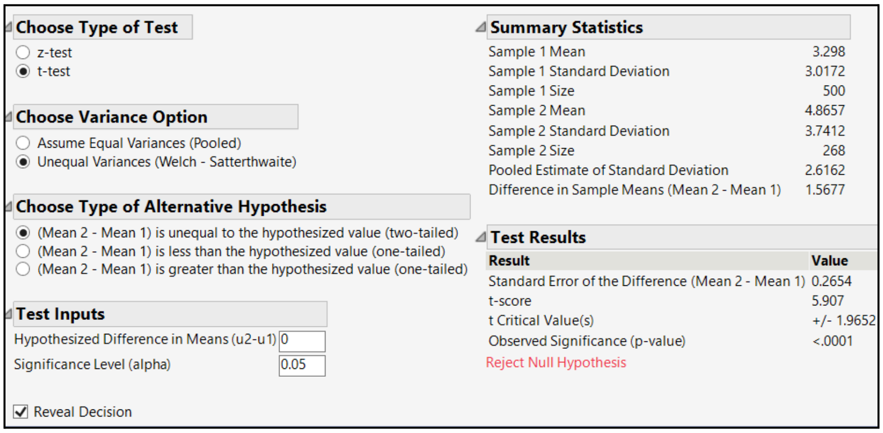
Statistical Technique 1 : From the Oneway Analysis the spread from these two groups are not similar, hence we employ the t test with the *Variance Option:* ***Unequal Variances.***


t Test

1-0

Assuming unequal variances

Difference	1.56767	t Ratio	5.906961
Std Err Dif	0.26539	DF	455.9558
Upper CL Dif	2.08922	Prob > |t|	<.0001*
Lower CL Dif	1.04612	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000


Choose Type of Test

○ z-test
● t-test

Choose Variance Option

○ Assume Equal Variances (Pooled)
● Unequal Variances (Welch - Satterthwaite)

Choose Type of Alternative Hypothesis

● (Mean 2 - Mean 1) is unequal to the hypothesized value (two-tailed)
○ (Mean 2 - Mean 1) is less than the hypothesized value (one-tailed)
○ (Mean 2 - Mean 1) is greater than the hypothesized value (one-tailed)

Test Inputs

Hypothesized Difference in Means (u2-u1) 0
Significance Level (alpha) 0.05

☑ Reveal Decision

Summary Statistics

Sample 1 Mean	3.298
Sample 1 Standard Deviation	3.0172
Sample 1 Size	500
Sample 2 Mean	4.8657
Sample 2 Standard Deviation	3.7412
Sample 2 Size	268
Pooled Estimate of Standard Deviation	2.6162
Difference in Sample Means (Mean 2 - Mean 1)	1.5677

Test Results

Result	Value
Standard Error of the Difference (Mean 2 - Mean 1)	0.2654
t-score	5.907
t Critical Value(s)	+/- 1.9652
Observed Significance (p-value)	<.0001

Reject Null Hypothesis

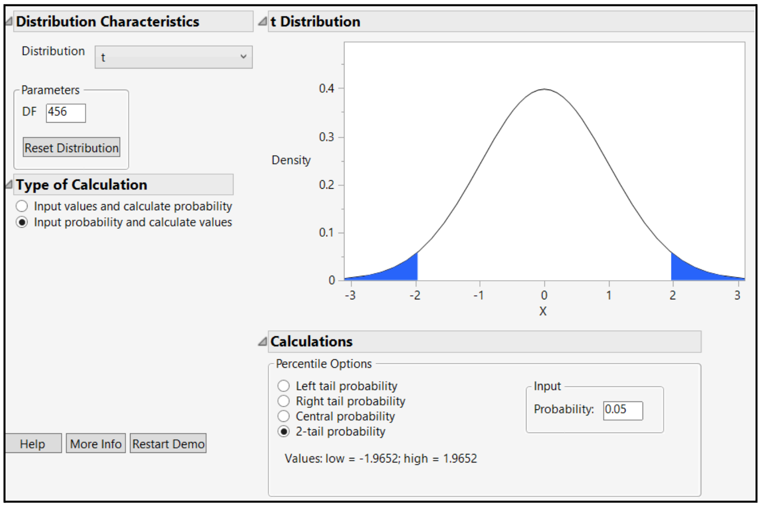
Utilizing this method: $t_0 = 5.907$

Based on our assumptions, **The Critical value = +/- 1.9652**

Rejection Region: anything smaller than -1.9652 or bigger than 1.9652

P-value = <0.0001 .

As the **p-value <** α , we reject the null hypothesis.



Distribution Characteristics

Distribution: t

Parameters

DF: 456

Reset Distribution

Type of Calculation

○ Input values and calculate probability

● Input probability and calculate values

t Distribution

Calculations

Percentile Options

○ Left tail probability

○ Right tail probability

○ Central probability

● 2-tail probability

Input

Probability: 0.05

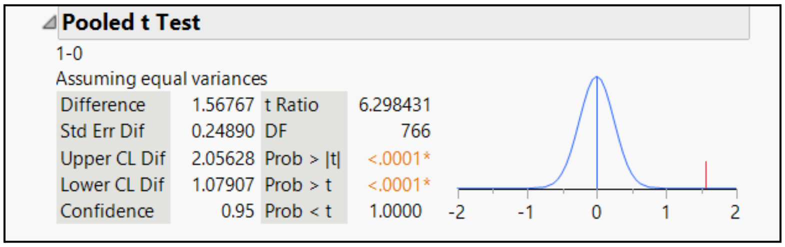
Values: low = -1.9652; high = 1.9652

Hence,

Decision: Reject Null hypothesis

Conclusion : At 0.05 level of significance, we have **sufficient evidence to conclude** that the mean value of pregnancies in diabetic women differ from non-diabetic women hence proving that the **diabetic condition influences the pregnancies in the women in the specified demographic .**

Statistical Technique 2: We employ the t test with the Variance Option as : Equal Variances.

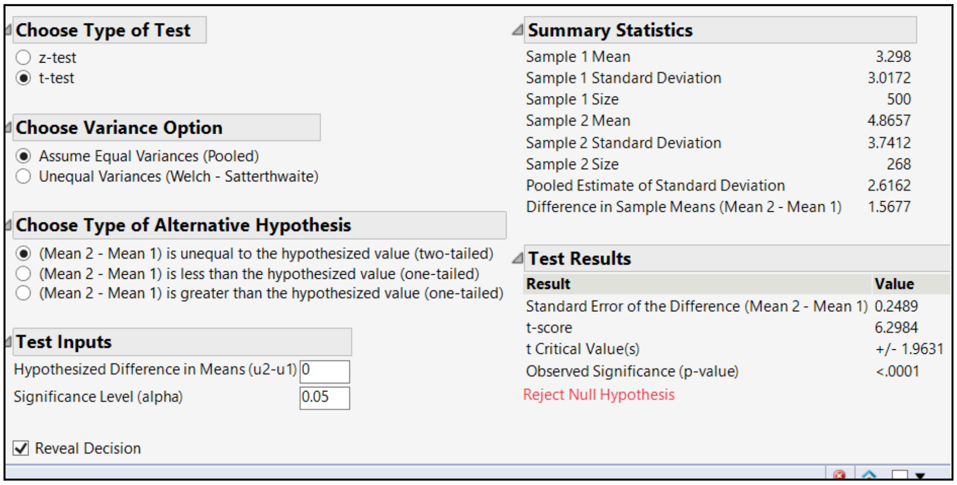


Pooled t Test

1-0

Assuming equal variances

Difference	1.56767	t Ratio	6.298431
Std Err Dif	0.24890	DF	766
Upper CL Dif	2.05628	Prob > |t|	<.0001*
Lower CL Dif	1.07907	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Choose Type of Test

○ z-test

● t-test

Choose Variance Option

● Assume Equal Variances (Pooled)

○ Unequal Variances (Welch - Satterthwaite)

Choose Type of Alternative Hypothesis

● (Mean 2 - Mean 1) is unequal to the hypothesized value (two-tailed)

○ (Mean 2 - Mean 1) is less than the hypothesized value (one-tailed)

○ (Mean 2 - Mean 1) is greater than the hypothesized value (one-tailed)

Test Inputs

Hypothesized Difference in Means (u2-u1): 0

Significance Level (alpha): 0.05

☑ Reveal Decision

Summary Statistics

Sample 1 Mean	3.298
Sample 1 Standard Deviation	3.0172
Sample 1 Size	500
Sample 2 Mean	4.8657
Sample 2 Standard Deviation	3.7412
Sample 2 Size	268
Pooled Estimate of Standard Deviation	2.6162
Difference in Sample Means (Mean 2 - Mean 1)	1.5677

Test Results

Result	Value
Standard Error of the Difference (Mean 2 - Mean 1)	0.2489
t-score	6.2984
t Critical Value(s)	+/- 1.9631
Observed Significance (p-value)	<.0001

Reject Null Hypothesis

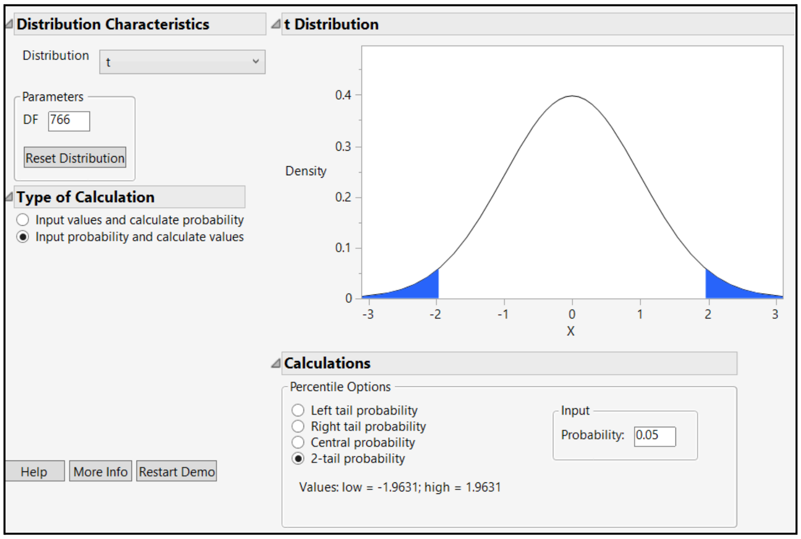
$t_0 \ = \ 6.2984$

Based on our assumptions, **the Critical value = +/- 1.9631**

Reject Region: anything smaller than - 1.9631 or bigger than 1.9631

P-value = <0.0001

As the p-value < α , we reject the null hypothesis.



Distribution Characteristics

Distribution: t

Parameters

DF: 766

Type of Calculation

○ Input values and calculate probability
● Input probability and calculate values

t Distribution

Calculations

Percentile Options

○ Left tail probability
○ Right tail probability
○ Central probability
● 2-tail probability

Input

Probability: 0.05

Values: low = -1.9631; high = 1.9631

Decision: Reject Null hypothesis

Conclusion : At 0.05 level of significance, we have sufficient evidence to conclude that the mean value of pregnancies in diabetic women differ from non-diabetic women hence proving that the diabetic condition influences the pregnancies in the women in the specified demographic.

Recommendations based on the results obtained for Analysis Question 1:

The number of pregnancies a woman experiences can be influenced by her diabetic condition, affecting the likelihood of carrying a pregnancy to term. Therefore, it is advisable for women planning a pregnancy to test for diabetes, ensuring the necessary hormonal levels are monitored. This approach also informs the healthcare planning required for potential diabetes-related conditions.

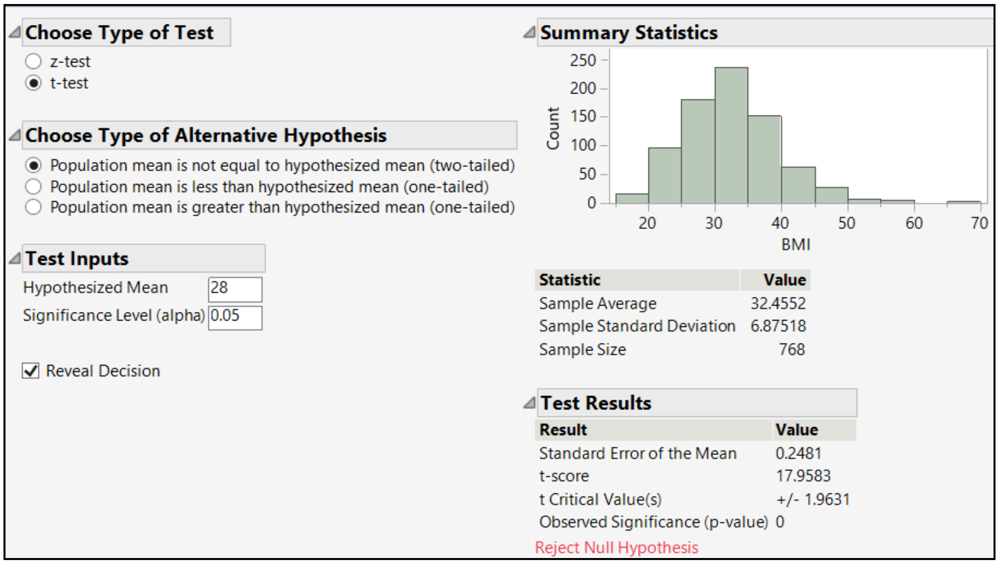
Statistical analyses show that diabetes significantly affects the average number of pregnancies. Non-diabetic women with fewer pregnancies and a high Diabetes Pedigree Function (DPF) value, indicating a higher risk of diabetes from family history, should be especially cautious, as future diabetes can present extra challenges in pregnancy.

Analysis Question 2:

To determine if the BMI (Body Mass Index) of the diabetic patients differ from the National Average of 28, we utilize Hypothesis Testing for One-Sample test.

Statistical Technique 3:

In this scenario,
We assume that the populations are normal and $\alpha = 0.05$
μ = Mean of BMI (Body Mass Index) of the diabetic patients , μ_0 = National Average for BMI in USA= 28



Choose Type of Test

○ z-test
● t-test

Choose Type of Alternative Hypothesis

● Population mean is not equal to hypothesized mean (two-tailed)
○ Population mean is less than hypothesized mean (one-tailed)
○ Population mean is greater than hypothesized mean (one-tailed)

Test Inputs

Hypothesized Mean: 28
Significance Level (alpha): 0.05

☑ Reveal Decision

Summary Statistics

Statistic	Value
Sample Average	32.4552
Sample Standard Deviation	6.87518
Sample Size	768

Test Results

Result	Value
Standard Error of the Mean	0.2481
t-score	17.9583
t Critical Value(s)	+/- 1.9631
Observed Significance (p-value)	0

Reject Null Hypothesis

Null Hypothesis : $H_0 : \mu = \mu_0$ **Alternate Hypothesis :** $H_1 : \mu \neq \mu_0$

As observed above, the Critical value = +/- 1.9631 , Standard Error of mean =0.2481
t score = 17.9583
Since test statistic is more extreme than critical value, our decision is to Reject H_0 .

Decision: Reject Null hypothesis
Conclusion : At 0.05 level of significance, we have sufficient evidence to conclude that the mean value of BMI in diabetic patients differs from the national average and hence we can reject the claim that the BMI of diabetic patients is the same as the national average BMI.

Recommendations based on the results obtained for Analysis Question 2:

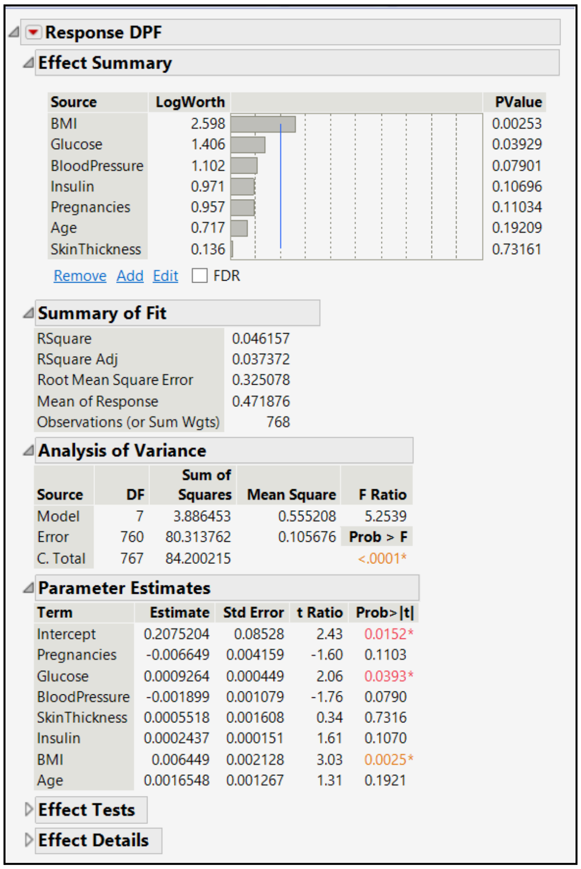
We can see that the sample average mean of this data is 32.4552 which falls right in the Obese category.
Since the BMI of the diabetic patients is different from the national average. As per the result obtained from our analysis, we found that the sample average mean is much greater than the National Average BMI hinting at higher BMI levels(Obese) in diabetic patients.

As BMI is a leading factor of Diabetes, it is recommended for these women to keep their health in check. A BMI of 32.45 is concerning in the Pima Indian community. It should be recommended that the BMI levels are kept in check by leading a much healthier lifestyle.

Analysis Question 3:

We want to determine if an association is present between a person's family history and the other features.
The goal is to find optimal predictors for **predicting** the ***DiabetesPedigreeFunction(DPF)*** by implementing **Multiple Linear Regression.**

Statistical Technique 4:



Response DPF

Effect Summary

Source	LogWorth	PValue
BMI	2.598	0.00253
Glucose	1.406	0.03929
BloodPressure	1.102	0.07901
Insulin	0.971	0.10696
Pregnancies	0.957	0.11034
Age	0.717	0.19209
SkinThickness	0.136	0.73161

Summary of Fit

RSquare	0.046157
RSquare Adj	0.037372
Root Mean Square Error	0.325078
Mean of Response	0.471876
Observations (or Sum Wgts)	768

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	3.886453	0.555208	5.2539
Error	760	80.313762	0.105676	Prob > F
C. Total	767	84.200215		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.2075204	0.08528	2.43	0.0152*
Pregnancies	-0.006649	0.004159	-1.60	0.1103
Glucose	0.0009264	0.000449	2.06	0.0393*
BloodPressure	-0.001899	0.001079	-1.76	0.0790
SkinThickness	0.0005518	0.001608	0.34	0.7316
Insulin	0.0002437	0.000151	1.61	0.1070
BMI	0.006449	0.002128	3.03	0.0025*
Age	0.0016548	0.001267	1.31	0.1921

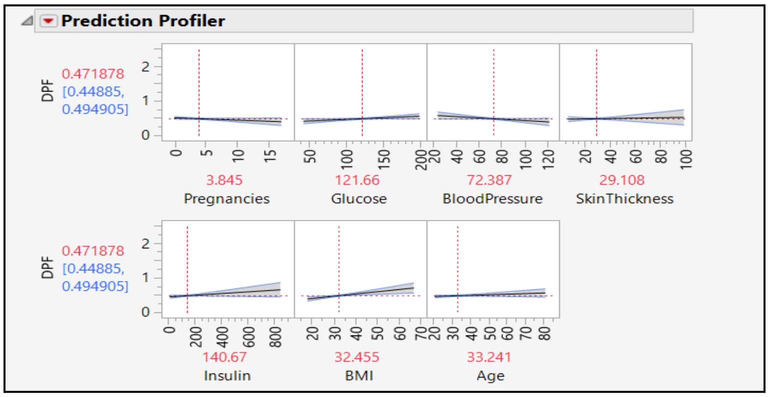
Effect Tests

Effect Details

In this scenario,
When we fit the model using all the variables,
The model can be represented as

$$y = 0.2075 - 0.0066 * x_1 + 0.0009 * x_2 - 0.0018 * x_3 + 0.0005 * x_4 + 0.0002 * x_5 + 0.0064 * x_6 + 0.0016 * x_7$$

We find that **BMI** has the **greatest** contribution to a person's diabetic condition followed by Glucose, Blood Pressure, Insulin levels and **Pregnancies**. **Age** and **Skin thickness** contributes **least** to a person's diabetic condition and its effects to be ignored.



Adjusted R-squared (RSquareAdj) is a modification of the R-squared statistic in regression analysis that takes into account the number of predictors in the model and adjusts the R-squared value to penalize the variables that do not significantly contribute in explaining the variations in the dependent variable. In multi-linear regression, R-square Adjusted is generally preferred over the simple R-squared. For our scenario, **R square Adj = 0.037372.** This however, **isn't that great**. Theoretically, the R-squared metric can be inflated by adding more data points. So for further analysis, we chose the Root Mean Square Error metric

The Root Mean Square Error provides a way to measure the spread of the residuals and indicates how well the model's predictions align with the actual data. Lower Root Mean Square Error values indicate a better fit of the model to the data, while higher values suggest that the model's predictions are farther from the observed values on average.The Root mean Square Error for our predicted model is **0.3250 which is fairly good.**

Conclusion & Recommendations based on the results obtained for Analysis Question 3:

As per the analysis above, we know BMI and Blood Pressure are leading factors for causing diabetes, which can be controlled by leading a healthier lifestyle. Glucose and Insulin can be controlled by taking proper medication.
Hence, if the Pima Indian women are successful in managing these factors, we can mitigate the symptoms of diabetes by a huge extent.

References

- Kaggle
- A Machine Learning Approach to Predicting the Onset of Type II Diabetes in a Sample of Pima Indian Women
- The Study of Pima Indian Diabetes
- Analysis of Pima Indian Diabetes Using KNN Classifier and Support Vector Machine Technique