**Report on Analysis of Pima Indian Women Diabetes Database**

TBANLT 540: Applied Regression Models

Group 1 Final Project

Contents

[Abstract 3](#_Toc4180527)

[Introduction 3](#_Toc4180528)

[Overview of the Data Set 4](#_Toc4180529)

[Analysis Plan 4](#_Toc4180530)

[Data Preprocessing 4](#_Toc4180531)

[Descriptive Statistics 5](#_Toc4180532)

[Exploratory Data Analysis 5](#_Toc4180533)

[Regression Output and Analysis 7](#_Toc4180534)

[Residual Plots 14](#_Toc4180535)

[Results 16](#_Toc4180536)

[Limitations of the study 17](#_Toc4180537)

[Conclusion 17](#_Toc4180538)

# Abstract

Diabetes mellitus is a very common and serious disease in many American Indian tribes, Indians, and many other populations in the world. Several well-known risk factors such as parental diabetes, genetic markers, obesity, diet are considered as the main risk factors for diabetes mellitus, while the precise nature of the gene or genes remains unknown.

Objectives: The Pima Indians, and many other populations in the world now suffer from the high prevalence rates of diabetes. The objective of this study is to identify a causal relationship between the risk factors and the disease. In the present report, data from Pima Indian women is analyzed to determine the causal factors of diabetes mellitus. This article aims to identify the determinants of diabetes mellitus in the Pima Indian women.

Results: The causal factors for diabetes mellitus of the Pima Indian women are identified. Statistically significant causal factors, namely, number of pregnancies, Glucose, BMI, Diabetes Pedigree Function indicating hereditary risk are identified as the determinants of diabetes mellitus. The effects of different causal factors on diabetes mellitus are explained based on probabilistic models.

Conclusions: Impacts of biochemical parameters, personal characteristics, family history, and dietary factors on risk of diabetes mellitus are explained based on mathematical relationships. The results of the present analyses support many earlier research findings. However, this research also captures a predictive model that can assess the risk of getting diabetes for Pima Indian women.

# Introduction

Diabetes mellitus is a group of metabolic disorders where the blood sugar levels are higher than normal for prolonged periods of time [[1]](#footnote-1). Diabetes is caused either due to the insufficient production of insulin in the body or due to improper response of the body’s cells to Insulin. The former cause of diabetes is also called Type 1 DM or insulin-dependent diabetes mellitus and the latter is known as Type 2 DM or non-insulin dependent DM. Gestational diabetes is a third type of diabetes where women not suffering from DM develop high sugar levels during pregnancy. In the United States, 30.3 million Americans were recorded as suffering from diabetes with 1.5 million being diagnosed with diabetes every year. Total cost of diagnosed Diabetes in the U.S. in 2017 was $327 billion [[2]](#footnote-2). Diabetes is especially hard on women as it can affect both the mother and their unborn children during pregnancy. Women with diabetes have a higher likelihood of having a heart attack, miscarriages or babies born with birth defects [[3]](#footnote-3).

Due to the increasing incidence rate of diabetes and prediabetes, it is a pressing issue in the healthcare industry to rightly identify the factors that contribute to the occurrence of diabetes in people, more so, in women. From secondary research, factors such as BMI, blood pressure, cholesterol and glucose levels are important factors that cause diabetes. In women, pregnancy seems to be an additional factor. To validate the above hypotheses, identify additional risk factors and build tools that can predict the occurrence of diabetes, particularly in women, the Pima Indians’ Diabetes dataset was chosen.

# Overview of the Data Set

This diabetes data contains information about Pima Indian females from 1993, originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The patients in this dataset are all at least 21 years old, of Pima Indian heritage, and had been studied continuously since 1965. It contains information of 768 females, of which 268 females were diagnosed with diabetes. Information available includes 8 variables, such as age, number of pregnancies, glucose, insulin, etc. More detailed description about the variables is listed in the table below. The response variable in the dataset is a binary classifier, Outcome, that indicates if the person was diagnosed with diabetes or not. Here is a brief description of the variables:

**Independent variables:**

* Pregnancies - Number of times patient was pregnant
* Glucose - Plasma glucose concentration 2 hours in an oral glucose tolerance test
* BloodPressure - Diastolic blood pressure (mm Hg)
* SkinThickness - Triceps skin-fold thickness (mm)
* Insulin - 2-Hour serum insulin (mu U/ml)
* BMI - Body mass index (weight in kg/ (height in m)^2)
* DiabetesPedigreeFunction - A calculation that provides data on diabetes history in relatives (hereditary risk)
* Age - (years)

**Dependent variable:**

* Outcome - Class variable: 0 (not having diabetes) and 1 (has diabetes)

The dataset can be found here: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

# Analysis Plan

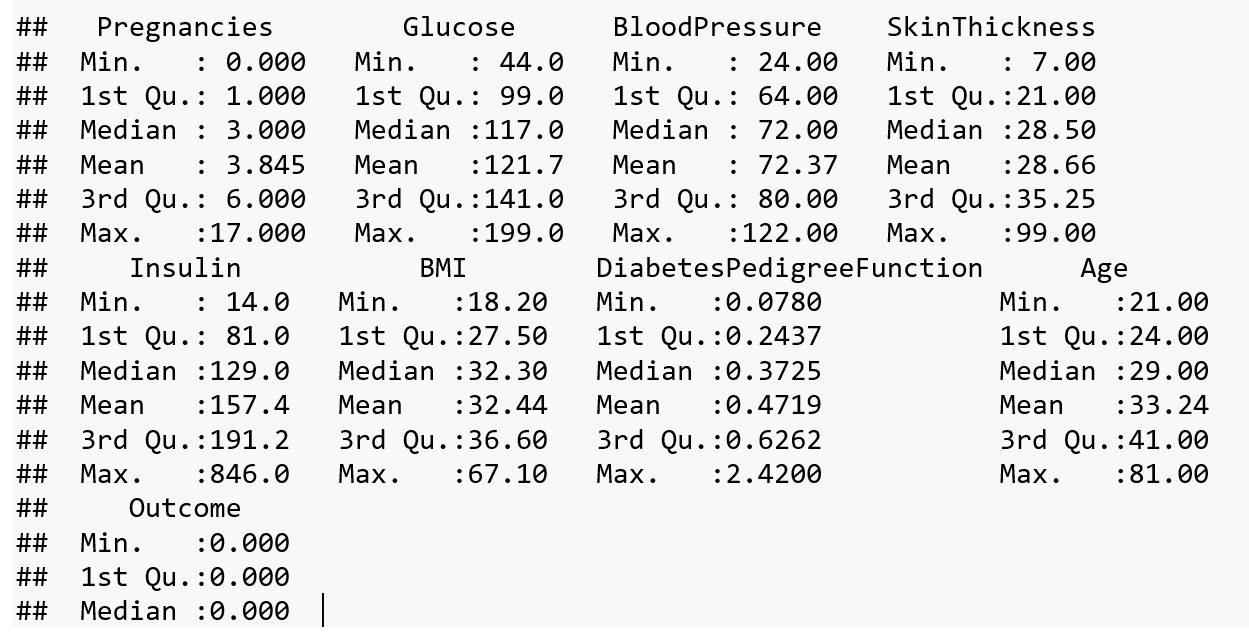
To build a predictive model that classifies women with/ without Diabetes, the following modeling approach was used.

1. Data Preprocessing
2. Exploratory Data Analysis
3. Modeling
   * Model Building - The modeling techniques performed on this data was Logistic Regression. The data set was split into an 80-20 train-test data set.
   * Arriving at parsimonious model based on AIC.
4. Model Performance Evaluation - Model performance was evaluated for the parsimonious model using sensitivity as the criterion. Sensitivity is the ability of the model to determine the true cases correctly, in this case, identify the patients with diabetes correctly.

# Data Preprocessing

At first glance, the dataset appeared to be clean. On deeper analysis, the dataset revealed many abnormal values for biological measures. Variables such as Skin Thickness and Glucose had 227 and 374 zero-values respectively. The fact that both measures cannot hold zero values indicated that the missing values in the dataset and they were represented as zero values in the dataset. The missing values in the dataset constituted to about 30% of the total observations. As removing these values would result in significant information loss, regression-based imputation using MICE library in R was performed to impute the missing values in the data set.

# Descriptive Statistics



# Exploratory Data Analysis

|  |  |
| --- | --- |
|  |  |
| Plot1 shows difference in the number of pregnancies between women who had diabetes and women who didn’t (in this dataset). However, in the histogram, no clear relationship between the number of pregnancies and the presence of Diabetes can be seen. We see skewness in the number of pregnancies as the box is towards the lower tail, therefore need standardization. | Plot1 shows a clear difference in the amount of Glucose present in the women who have been diagnosed with Diabetes and those who haven’t. While the density plot indicates slight overlap in the levels of glucose in both categories of women, we suspect that Glucose could be a good indicator of the response. |
|  |  |
| No clear difference is seen in the two categories of women who have and don’t have Diabetes. This shows that Blood Pressure might not be a good predictor of the response variable. | No clear difference can be seen in the two categories of women who have and don’t have Diabetes. This shows that Skin Thickness might not be a good predictor of the response variable. |
|  |  |
| It is observed that the median insulin level is higher for those who have diabetes, indicating Insulin may be a good predictor of the response variable. We see skewness in the insulin amount as the box is towards the lower tail, therefore need standardization. | In Plot1, all the women who had Diabetes had a median BMI around 35, which is above the normal levels. On the other hand, women who did not have Diabetes had a BMI (not very different from the other) but still slightly lower around 30. We’ll investigate further to see if this variable is a good predictor during modeling. We see skewness in BMI as the box is towards the lower tail, therefore need standardization. |
|  |  |
| No clear difference can be seen in the two categories of women who have and don’t have Diabetes. This shows that DPF might not be a good predictor of the response variable. We see skewness in DPF as the box is towards the lower tail, therefore need standardization. | It is seen that the median age of the women that have diabetes was higher than those who didn’t. This shows that Age might be a good predictor for the response variable. We see skewness in Age as the box is towards the lower tail, therefore need standardization. |

**Data Standardization:** From the box plots, in the exploratory data analysis, we identified skewness in Pregnancies, Insulin, Diabetes Pedigree Function, BMI and Age. We have standardized these variables using the scale function.

# Regression Output and Analysis

As the response variable is a binary categorical variable and since we would like to understand the presence of what explanatory variables causes a person to have diabetes meaning being able to effectively classify women as having diabetes or not, the modeling technique we’d be using is the logistic regression analysis.

Below are the steps by which we found the parsimonious model.

Adding explanatory variables to our model:

1. **Pregnancies**

As a first step, we build a model with outcome (the response variable) and the number of pregnancies (standardized) and below is the regression output.

## Null deviance: 804.39 on 614 degrees of freedom   
## Residual deviance: 773.15 on 613 degrees of freedom   
## AIC: 777.15 

The AIC tells the information loss and the **AIC** for this model is **777.15**. Akaike Information Criterion (AIC) is an estimator of the relative quality of the statistical models: it estimates the quality of each model relative to other models**.** This value is only useful in comparison. So, we move on to building our next model in the process of stepwise selection.

1. **Glucose**

Now, we introduce our second predictor variable, Glucose, into our model and the regression output showing the AIC value can be seen below:

Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.389888 0.481682 -11.190 < 2e-16 \*\*\*  
## Pregnancies.Z 0.415873 0.096460 4.311 1.62e-05 \*\*\*  
## Glucose 0.038249 0.003716 10.292 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 631.33 on 612 degrees of freedom  
## AIC: 637.33

The AIC tells the information loss and the **AIC** for this model is **637.33.** This means that our information loss has gone down to 637.33 from 777.15 with the addition of our second explanatory variable, Glucose.

1. **Blood Pressure**

We bring in the 3rd explanatory variable blood pressure into our model to check for the AIC value.

## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.364450 0.742028 -8.577 < 2e-16 \*\*\*  
## Pregnancies.Z 0.386710 0.097661 3.960 7.5e-05 \*\*\*  
## Glucose 0.037293 0.003727 10.005 < 2e-16 \*\*\*  
## BloodPressure 0.014993 0.008404 1.784 0.0744 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 628.13 on 611 degrees of freedom  
## AIC: 636.13

We can see that the AIC value has gone down to **636.13** from the previous 637.33.

1. **Skin Thickness:**

We add the predictor variable skin thickness to the model and check for the AIC.

## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.835597 0.755725 -7.722 1.15e-14 \*\*\*  
## Pregnancies.Z 0.377593 0.099838 3.782 0.000156 \*\*\*  
## Glucose 0.035694 0.003760 9.492 < 2e-16 \*\*\*  
## BloodPressure 0.009753 0.008634 1.130 0.258599   
## SkinThickness.Z 0.501140 0.109991 4.556 5.21e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 606.21 on 610 degrees of freedom  
## AIC: 616.21

We observe that the AIC value has dropped down to **616.21** from 636.13.

1. **Insulin:**

We add the predictor variable insulin to our model and below is the regression output for the same:

## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.841248 0.801230 -7.290 3.09e-13 \*\*\*  
## Pregnancies.Z 0.377510 0.099917 3.778 0.000158 \*\*\*  
## Glucose 0.035746 0.004479 7.982 1.44e-15 \*\*\*  
## BloodPressure 0.009744 0.008645 1.127 0.259728   
## SkinThickness.Z 0.501309 0.110284 4.546 5.48e-06 \*\*\*  
## Insulin.Z -0.002650 0.124691 -0.021 0.983045   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 606.21 on 609 degrees of freedom  
## AIC: 618.21

We see that our AIC value has gone up from 616.21 to 618.21 with the addition of the predictor variable insulin. We check to see how the model behaves with the addition of the rest of the predictor variables.

1. **BMI**

As a next step, we add the predictor variable BMI into our model and the regression report is as follows:

## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.286843 0.822244 -6.430 1.28e-10 \*\*\*  
## Pregnancies.Z 0.445519 0.105233 4.234 2.30e-05 \*\*\*  
## Glucose 0.036388 0.004641 7.841 4.47e-15 \*\*\*  
## BloodPressure 0.000807 0.008956 0.090 0.928   
## SkinThickness.Z 0.115731 0.136586 0.847 0.397   
## Insulin.Z -0.052094 0.126072 -0.413 0.679   
## BMI.Z 0.654091 0.144440 4.528 5.94e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 584.48 on 608 degrees of freedom  
## AIC: 598.48

Now, we see that our model is back with the trend of the AIC value dropping telling us the information loss has gone down to 598.48 with the addition of BMI.

1. **Diabetes Pedigree Function:**

We now add the diabetes pedigree function to our model.

## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.5568890 0.8389443 -6.624 3.50e-11 \*\*\*  
## Pregnancies.Z 0.4543594 0.1057771 4.295 1.74e-05 \*\*\*  
## Glucose 0.0363272 0.0046534 7.807 5.87e-15 \*\*\*  
## BloodPressure 0.0005276 0.0089673 0.059 0.9531   
## SkinThickness.Z 0.0976371 0.1370474 0.712 0.4762   
## Insulin.Z -0.0656859 0.1276605 -0.515 0.6069   
## BMI.Z 0.6531410 0.1447175 4.513 6.39e-06 \*\*\*  
## DiabetesPedigreeFunction 0.6352972 0.3300487 1.925 0.0542 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 580.75 on 607 degrees of freedom  
## AIC: 596.75

We see the information loss going further down through the AIC value.

1. **Age:**

We bring in the predictor variable Age into our model and following is the regression report.

## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.412631 0.861138 -6.285 3.27e-10 \*\*\*  
## Pregnancies.Z 0.410501 0.121967 3.366 0.000764 \*\*\*  
## Glucose 0.036003 0.004666 7.717 1.19e-14 \*\*\*  
## BloodPressure -0.000866 0.009173 -0.094 0.924783   
## SkinThickness.Z 0.090758 0.138022 0.658 0.510820   
## Insulin.Z -0.071966 0.127487 -0.564 0.572417   
## BMI.Z 0.666798 0.146235 4.560 5.12e-06 \*\*\*  
## DiabetesPedigreeFunction 0.625929 0.330337 1.895 0.058116 .   
## Age.Z 0.090951 0.127508 0.713 0.475664   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 580.24 on 606 degrees of freedom  
## AIC: 598.24

We see that our AIC value has gone up from **596.74 to 598.24** telling us the information loss now with the addition of the variable age has gone up in the model. Also, it should be noted that age isn’t significant at all (as seen in the above report) in predicting diabetes among Pima Indian women.

1. **Taking Age out of the model:**

Although in the explanatory analysis, we found that the median age of pregnant women who had diabetes was around 38 and the median age of pregnant women who didn’t was around 25 meaning that age can be one of the reasons pregnant women get diabetes, the AIC values of our model says otherwise. So, we take out the Age variable to check if the AIC drops down. Below is the regression output for the same.

## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.5568890 0.8389443 -6.624 3.50e-11 \*\*\*  
## Pregnancies.Z 0.4543594 0.1057771 4.295 1.74e-05 \*\*\*  
## Glucose 0.0363272 0.0046534 7.807 5.87e-15 \*\*\*  
## BloodPressure 0.0005276 0.0089673 0.059 0.9531   
## SkinThickness.Z 0.0976371 0.1370474 0.712 0.4762   
## Insulin.Z -0.0656859 0.1276605 -0.515 0.6069   
## BMI.Z 0.6531410 0.1447175 4.513 6.39e-06 \*\*\*  
## DiabetesPedigreeFunction 0.6352972 0.3300487 1.925 0.0542 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 580.75 on 607 degrees of freedom  
## AIC: 596.75

No Age Variable

We’re now back at the AIC (596.74) that we had before this last model.

1. **Taking Blood Pressure out of the model:**

Through our exploratory data analysis (EDA), we found that there was no clear difference in the blood pressure of women who had diabetes and who didn’t. So, we suspect that this variable may not be a good candidate for explaining our response variable. To test this, we remove this variable from our model to check how the model behaves.

Below is the regression output report:

## Call:  
## glm(formula = Outcome ~ Pregnancies.Z + Glucose + SkinThickness.Z +   
## Insulin.Z + BMI.Z + DiabetesPedigreeFunction, family = binomial,   
## data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5896 -0.7359 -0.3988 0.7298 2.2361   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.523176 0.612281 -9.021 < 2e-16 \*\*\*  
## Pregnancies.Z 0.455572 0.103772 4.390 1.13e-05 \*\*\*  
## Glucose 0.036363 0.004615 7.879 3.30e-15 \*\*\*  
## SkinThickness.Z 0.097493 0.137019 0.712 0.4768   
## Insulin.Z -0.065989 0.127548 -0.517 0.6049   
## BMI.Z 0.654737 0.142125 4.607 4.09e-06 \*\*\*  
## DiabetesPedigreeFunction 0.635578 0.329997 1.926 0.0541 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 580.75 on 608 degrees of freedom  
## AIC: 594.75

We see here that the AIC has gone down from 596.74 to **594.75.**

1. **Taking Insulin out:**

Now, we do the same with the variable insulin as it was also suspected in the EDA that it may not be a good predictor of diabetes. Below is the regression report:

## Call:  
## glm(formula = Outcome ~ Pregnancies.Z + Glucose + SkinThickness.Z +   
## BMI.Z + DiabetesPedigreeFunction, family = binomial, data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6671 -0.7341 -0.4002 0.7363 2.2168   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.358057 0.519007 -10.324 < 2e-16 \*\*\*  
## Pregnancies.Z 0.457595 0.103583 4.418 9.98e-06 \*\*\*  
## Glucose 0.035042 0.003813 9.189 < 2e-16 \*\*\*  
## SkinThickness.Z 0.097225 0.136953 0.710 0.4778   
## BMI.Z 0.648673 0.141311 4.590 4.42e-06 \*\*\*  
## DiabetesPedigreeFunction 0.628232 0.330291 1.902 0.0572 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 581.02 on 609 degrees of freedom  
## AIC: 593.02

We can see that our AIC value has gone further down.

1. **Taking skin thickness out:**

Similarly, as suspected in the EDA, we check for the AIC value by taking the skin thickness variable out of the model.

The regression report for the same is as follows:

## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.376461 0.518263 -10.374 < 2e-16 \*\*\*  
## Pregnancies.Z 0.464149 0.103281 4.494 6.99e-06 \*\*\*  
## Glucose 0.035185 0.003808 9.239 < 2e-16 \*\*\*  
## BMI.Z 0.708513 0.114312 6.198 5.72e-10 \*\*\*  
## DiabetesPedigreeFunction 0.643708 0.329238 1.955 0.0506 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 581.53 on 610 degrees of freedom  
## AIC: 591.53

We see that our AIC has dropped further down. The odds ratio for the model is below:

## Odds\_Ratio\_DiabetesVNODiabetes 2.5 %  
## (Intercept) 0.004624157 0.001616277  
## Pregnancies.Z 1.590660510 1.302360913  
## Glucose 1.035811363 1.028332272  
## BMI.Z 2.030969427 1.631696108  
## DiabetesPedigreeFunction 1.903526664 1.001196814  
## 97.5 %  
## (Intercept) 0.01236604  
## Pregnancies.Z 1.95380559  
## Glucose 1.04382711  
## BMI.Z 2.55601099  
## DiabetesPedigreeFunction 3.64410473

We understand that this is our parsimonious model with the least AIC and the predictor variables are:

1. Number of Pregnancies
2. Glucose
3. BMI
4. Diabetes Pedigree Function

Next, we created residual plots on our predictor variables to gain some insight into the residuals (observed – estimated).

# Residual Plots

|  |  |
| --- | --- |
|  |  |
| Residuals for this variable shows homoscedasticity meaning that the variance of errors is equal from observation to observation. The same goes with all our residual plots as well in this study. | Residuals for this variable shows homoscedasticity. |
|  |  |
| Residuals for this variable shows homoscedasticity. | Residuals for this variable shows homoscedasticity. |
|  |  |
| Residuals for this variable shows homoscedasticity. | Residuals for this variable shows homoscedasticity. |
|  |  |
| Residuals for this variable shows homoscedasticity. | Residuals for this variable shows homoscedasticity. |

# Results

1. **Parsimonious model:**

A full model was built with Outcome as the response variable with the rest of the 8 predictor variables. Stepwise selection method was used to identify the most important variables. The final model chosen with AIC as the criterion for selection generated a logistic regression model with the lowest AIC value of 591.53.

Call:  
## glm(formula = Outcome ~ Pregnancies.Z + Glucose + BMI.Z + DiabetesPedigreeFunction,   
## family = binomial, data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6615 -0.7267 -0.4055 0.7259 2.1765   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.376461 0.518263 -10.374 < 2e-16 \*\*\*  
## Pregnancies.Z 0.464149 0.103281 4.494 6.99e-06 \*\*\*  
## Glucose 0.035185 0.003808 9.239 < 2e-16 \*\*\*  
## BMI.Z 0.708513 0.114312 6.198 5.72e-10 \*\*\*  
## DiabetesPedigreeFunction 0.643708 0.329238 1.955 0.0506 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 804.39 on 614 degrees of freedom  
## Residual deviance: 581.53 on 610 degrees of freedom  
## AIC: 591.53  
## Odds\_Ratio\_DiabetesVNODiabetes   
## (Intercept) 0.004624157   
## Pregnancies.Z 1.590660510   
## Glucose 1.035811363   
## BMI.Z 2.030969427   
## DiabetesPedigreeFunction 1.903526664

From the regression output above, we identified that blood pressure, skin thickness, insulin, and age had p-values greater than 0.05 and were not statistically significant in our modeling efforts, so we removed them from the model.

We identified that number of pregnancies, glucose, BMI, and Diabetes Pedigree Function (family history) have p-values of less than 0.05. This indicates they are statistically significant. This model provides the lowest AIC of 591.53, indicating that it is the most likely parsimonious model.

We interpret the results of the model as follows:

* When all other variables are held constant, for each additional pregnancy there is an increase of 59.07% in the probability of diabetes. [exp (0.464149) = 1.590660]
* When all other variables are held constant, for each unit increase in blood glucose there is an increase of 3.58% in the probability of diabetes. [exp (0.035185) = 1.035811]
* When all other variables are held constant, for each unit increase in body mass index there is an increase of 103.10% in the probability of diabetes. [exp (0.708513) = 2.0309694]
* When all other variables are held constant, for each unit increase in the Diabetes Pedigree Function there is an increase of 90.35% in the probability of diabetes. [exp (0.643708) = 1.9035266]

**Model Evaluation:**

The logistic regression model for predicting the occurrence of diabetes in Pima women is evaluated based on sensitivity as the criterion. In our case, the model that we have built using the logistic regression data mining technique is able to identify the patients with diabetes **correctly at 66.36% of the time.** We can further evaluate the performance of models using various other techniques such as decision tree or support vector machine to evaluate the performance.

# Limitations of the study

We would like to have more variables put into consideration, such as the gestational week of the research object; if a woman got diabetes in her last pregnancy; and was there any viral or bacterial infection in the research period. These factors are probably significant in the predictive model, but they were not available in the current data set. The current model’s accuracy might be limited by lack of these variables. In future research, we will try to collect this data either from public resources or asking the record keeper to do so. In addition, due to the number of predictor variables in the model, a larger data set would be more helpful. With limited bandwidth we could not build and evaluate the performance of various other statistical techniques to predict the occurrence of diabetes in Pima women; this can be further explored. Also, the available dataset is old and there might have been changes in the variability of the data in recent times. Hence, we will try to get latest data to make our findings more relevant.

# Conclusion

We set out to understand what predictor variables can best predict diabetes in Pima Indian women. We used historical data collected on these women who were studied since the year 1965. The predictor variables of interest were:

* Number of Pregnancies
* Blood Glucose level
* BMI
* Diabetes Pedigree Function
* Age
* Skin Thickness
* Blood Pressure, and
* Insulin

Through the course of exploratory data analysis and logistic regression analysis we understand that the predictor variables number of pregnancies, glucose, and BMI are significant at the 99% confidence level in predicting diabetes. Although the other variable in our most likely parsimonious model, the Diabetes Pedigree Function, is not as significant as the other three above it still is found to be significant in predicting diabetes in Pima Indian women.

**One striking finding we had was:**

Our exploratory data analysis said that while the median age of pregnant women who had diabetes was around 38, the median age of pregnant women who didn’t were around 28, telling us that age could play a role in pregnant women getting diabetes. However, the Akaike Information Criterion and the p-values said otherwise in the modeling process. We found that age wasn’t significant at all in explaining diabetes among the Pima Indian women.

In conclusion, despite our limitations of not having additional information:

* on the gestational week?
* whether the diabetes was gestational?
* does the recorded outcome value of 1 mean that women with multiple pregnancies had diabetes in at least one of their pregnancies or all of them?

We have found our most likely parsimonious model that has a performance measure of sensitivity at 63.36% meaning that our model can **correctly predict the presence of diabetes** among Pima Indian women **63.36%** of the time.

1. <https://en.wikipedia.org/wiki/Diabetes_mellitus> [↑](#footnote-ref-1)
2. <http://www.diabetes.org/diabetes-basics/statistics/> [↑](#footnote-ref-2)
3. <http://www.diabetes.org/living-with-diabetes/treatment-and-care/women/> [↑](#footnote-ref-3)