



**To create a Diabetes Risk Score: Making a multi factorial model to
assess the risk of an individual.**

Dr. Shami Nimgulkar Kamble

Assist Professor, Dept of Economics

Prahladrai Dalmia Lions College of Commerce & Economics

Abstract:

There has been a paradox prevailing in India where the awareness about diabetes among women has been quite high, but the behaviour to prevent it has not shown a significant change (Saliha,2018). The current paper attempts to create a multi factorial model to assess the risk of having diabetes using the secondary data published by Kaggle consisting of responses from 768 women respondents aged 20 years. R software is used to statistically analyse the data. First, the paper imputes the missing values in the explanatory variables and then checks each parameter to find out whether the data is normally distributed or not. Of the 8 explanatory variables, pregnancy is a parameter that is found to be not normally distributed. The paper therefore, applies two more tests- Kolmogorov Smirnov's test and Wilcoxon's test for normal distribution. Finally, the linear model fit is applied to identify the parameters that are responsible for increasing the risk of diabetes among respondents. Of the eight parameters, two parameters, namely- Diabetes Pedigree Function and Blood Pressure are found to be statistically significant in determining the risk of diabetes.

Keywords: Diabetes, Blood Pressure, Diabetes Pedigree Function.

Introduction

Diabetes has reached epidemic proportions in Asia - led by India and China -- and has dramatically increased the risk of premature death especially among women and middle-aged people, a significant study has found. India and China today have the highest diabetes burdens in the world (NDTV, 2019)

According to the World Health Organization (WHO), India has close to 62 million people living with the diseases and is projected to have close to 70 million diabetics by 2025 (ibid).

Women are anchors of the family. More often than less, this has led to them prioritising their families' health over their own. Traditionally, diabetes has not been spoken in the context of women. But secondary studies have shown that there has been a rise in incidence of diabetes among women.(Saliha, 2018).

Many Indians, a new research published in Diabetologia (journal of the European Association for the Study of Diabetes) shows that more than half of men (55 per cent) and some two-thirds (65 per cent) of women aged 20 years in

India will likely develop diabetes, with around 95 per cent of those cases likely be type 2 diabetes (T2D) — a lifelong disease that keeps your body from using insulin the way it should (Business Line, 2020).

The present papers attempts to examine 8 factors that affect the prevalence of diabetes among women in India. The definitions of 8 parameters (Ali, 2013) is mentioned below:

1. **Pregnancies** : Number of times pregnant
2. **Glucose**: Oral Glucose Tolerance Test result
3. **BloodPressure**: Diastolic Blood Pressure values in (mm Hg)
4. **SkinThickness**: Triceps skin fold thickness in (mm)
5. **Insulin**: 2-Hour serum insulin (μ U/ml)
6. **BMI**: The Body Mass Index (BMI) provides a simple, yet accurate method of assessing whether a patient is at risk from either over-or-underweight.
7. **DiabetesPedigreeFunction**: It provided some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence gave us an idea of the hereditary risk one might have with the onset of diabetes mellitus.
8. **Age**: Age in years
9. **Outcome**: Class 1 indicates person having diabetes and 0 indicates other.

Data and Methodology

2.1 Data Source:

The idea is to build diabetes test score for India and the data source has been secondary in nature. PIMA Indians diabetes data on Kaggle is used in order to arrive at conclusions. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. In particular, all patients here are females at least 21 years old of Pima Indian heritage. R software is used for the purpose of analysis.

2.2 Sample Size: The sample size is 768 respondents.

2.3 Parameters used in data:

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

2.4 Data Cleaning:

```
diabetes<-read.csv (file.choose())
attach (diabetes)
colnames (diabetes)
summary (diabetes)
```

```

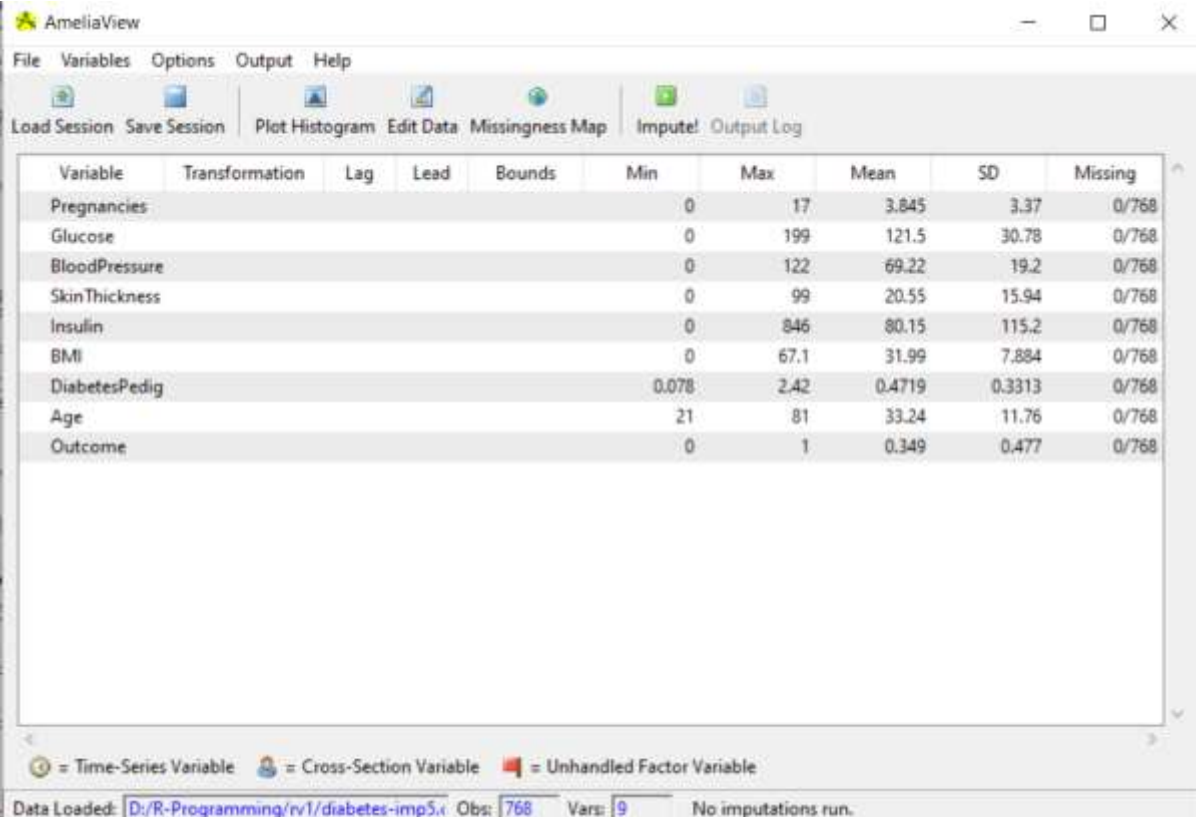
> summary(diabetes)
Pregnancies      Glucose      BloodPressure      SkinThickness      Insulin      BMI      DiabetesPedigreeFunction
Min. : 0.000  Min. : 44.0  Min. : 0.00  Min. : 7.00  Min. : 0.0  Min. : 0.00  Min. : 0.0780
1st Qu.: 1.000  1st Qu.: 99.0  1st Qu.: 64.00  1st Qu.: 22.00  1st Qu.: 76.0  1st Qu.: 27.40  1st Qu.: 0.2437
Median : 3.000  Median :117.0  Median : 72.00  Median : 29.00  Median :125.0  Median : 32.20  Median : 0.3725
Mean   : 3.845  Mean  :121.7  Mean   : 71.82  Mean   : 29.15  Mean   :155.2  Mean   : 32.29  Mean   : 0.4719
3rd Qu.: 6.000  3rd Qu.:141.0  3rd Qu.: 80.00  3rd Qu.: 36.00  3rd Qu.:190.0  3rd Qu.: 36.60  3rd Qu.: 0.6262
Max.   :17.000  Max.   :199.0  Max.   :122.00  Max.   : 99.00  Max.   :846.0  Max.   : 67.10  Max.   : 2.4200
      NA's : 5      NA's : 29      NA's : 227      NA's : 374      NA's : 7

Age      Outcome
Min. : 21.00  Min. : 0.000
1st Qu.: 24.00  1st Qu.: 0.000
Median : 29.00  Median : 0.000
Mean   : 33.24  Mean   : 0.349
3rd Qu.: 41.00  3rd Qu.: 1.000
Max.   : 81.00  Max.   : 1.000

```

The summary shows that the BloodPressure, Insulin and BMI all contain zeros. However, in reality, these parameters cannot take the value zero. Additionally, there appear some missing values for BloodPressure, SkinThickness, Insulin and BMI. I replaced all these zeros and missing values with the help of the package Amelia. Amelia is a tool that conducts multiple imputations for filling the missing cross section data.

The code for the same is:
 install.packages("Amelia")
 library(Amelia)
 AmeliaView()
 The output is as follows:



The screenshot shows the AmeliaView application window. The main area displays a summary table for the variables in the dataset. The table includes columns for Variable, Transformation, Lag, Lead, Bounds, Min, Max, Mean, SD, and Missing. The variables listed are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedig, Age, and Outcome. The 'Missing' column shows 0/768 for all variables, indicating no missing values. The status bar at the bottom indicates 'Data Loaded: D:/R-Programming/rv1/diabetes-imp5.c' with 768 observations and 9 variables, and 'No imputations run.'

Variable	Transformation	Lag	Lead	Bounds	Min	Max	Mean	SD	Missing
Pregnancies					0	17	3.845	3.37	0/768
Glucose					0	199	121.5	30.78	0/768
BloodPressure					0	122	69.22	19.2	0/768
SkinThickness					0	99	20.55	15.94	0/768
Insulin					0	846	80.15	115.2	0/768
BMI					0	67.1	31.99	7.884	0/768
DiabetesPedig					0.078	2.42	0.4719	0.3313	0/768
Age					21	81	33.24	11.76	0/768
Outcome					0	1	0.349	0.477	0/768

After clicking on the Impute tab, the imputations are done and the all the five imputations are stored in the directory.

The best imputation is the fifth imputation and it is now called in R with the following codes:

```
newdiabetes<- read.csv (file.choose())
```

```
attach (newdiabetes)
```

```
colnames (newdiabetes)
```

2.5 To check for Normal Distribution:

2.5.1 Age :

```
sddage<- sd(newdiabetes$Age)
```

```
sddage
```

```
#To test H0=Average Age =33 #(33 is the mean age)
```

```
      H1= Average Age <33
```

```
Age<- t.test(x = newdiabetes$Age, mu = 33, alternative =  
            "less")
```

```
Age
```

```
#Conclusion : p-value=0.71>0.05. Do not reject HO
```

```
#To test H0: Average age = 33
```

```
      versus H1: Average age ≠ 33
```

```
Age1<- t.test(x = newdiabetes$Age, mu = 33, alternative =  
            "two.sided")
```

```
Age1
```

```
#Conclusion: p-value=0.57>0.05. Do not reject HO
```

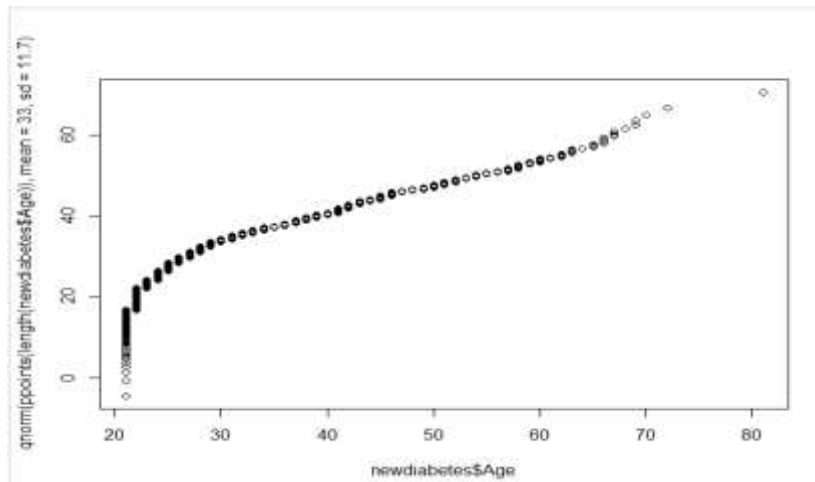
```
#To test:
```

```
      H0: Normal distribution is a good fit to the data on Age
```

```
      versus H1: Normal distribution is not a good fit
```

```
test1<- qqplot(newdiabetes$Age,  
              qnorm(ppoints(length(newdiabetes$Age))  
                    , mean = 33, sd = 11.7))
```

```
#Conclusion : normal distribution of Age with mean = 33  
and sd = 11.7 is a good fit.
```

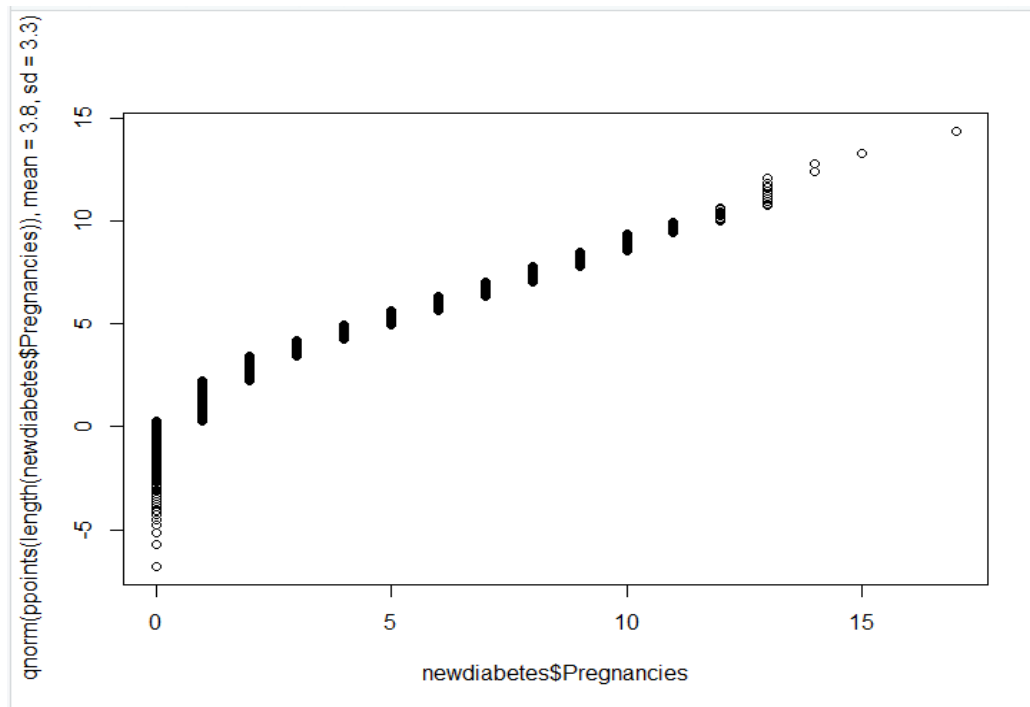


2.5.2 Pregnancies

```

sdpregnancies<- sd(newdiabetes$Pregnancies)
#To test H0=Average Age =3.8 (#3.8 is the mean pregnancy)
H1= Average Age <3.8
Pregnancies<- t.test(x = newdiabetes$Pregnancies, mu = 3.8, alternative =
"less")
Pregnancies
#Conclusion: p-value=0.64>0, Do not reject H0
#To test H0: Average Pregnancies = 3.8
versus H1: Average Pregnancies≠ 3.8
Pregnancies1<- t.test(x = newdiabetes$Pregnancies, mu = 3.8, alternative =
"two.sided")
Pregnancies1
#Conclusion: p-value=0.71>0.05. Do not reject HO
# To test:
H0: Normal distribution is a good fit to the data on Pregnancies
versus H1: Normal distribution is not a good fit
test2<- qqplot(newdiabetes$Pregnancies,
               qnorm(ppoints(length(newdiabetes$Pregnancies))
                    , mean = 3.8, sd = 3.3))
#Conclusion : normal distribution of pregnancies with mean = 3.8
and sd = 3.3 is not a good fit and the output is as follows:

```



Hence we use another test called as:

#Use Kolmogorov Smirnov's test

`ks.test(newdiabetes$Pregnancies, pnorm, mean=3.8, sd=3.3)`. The output

```

One-sample Kolmogorov-Smirnov test

data:  newdiabetes$Pregnancies
D = 0.16171, p-value < 2.2e-16
alternative hypothesis: two-sided
is:

```

#Conclusion: normal distribution with mean = 3.8 and sd = 3.3 is not a good fit.

Hence we conduct another test called as :

#Wilcoxon's test

H0: Average Pregnancies = 3.8

H1: Average Pregnancies \neq 3.8

```
wilcox.test(x = newdiabetes$Pregnancies, mu = 3.8,
            alternative = "two.sided")
```

The output is :

```

wilcoxon signed rank test with continuity correction

data:  newdiabetes$Pregnancies
V = 139176, p-value = 0.1676
alternative hypothesis: true location is not equal to 3.8

```

#Conclusion: p-value = 0.16 > 0.05. Do not reject Ho.

2.5.3 Glucose

```
sdglucose<- sd(newdiabetes$Glucose)
```

```
#To test H0=Average Glucose =122 (#122 is the mean glucose)
```

Dr. Shami Nimgulkar Kamble

H1= Average Glucose <122

```
Glucose<- t.test(x = newdiabetes$Glucose, mu = 122, alternative =
  "less")
```

Glucose

```
#Conclusion : p-value=0.42>0.05. Do not reject HO
```

```
#To test H0: Average Glucose = 122
```

```
versus H1: Average Glucose ≠ 122
```

```
Glucose1<- t.test(x = newdiabetes$Glucose, mu = 122, alternative =
  "two.sided")
```

Glucose1

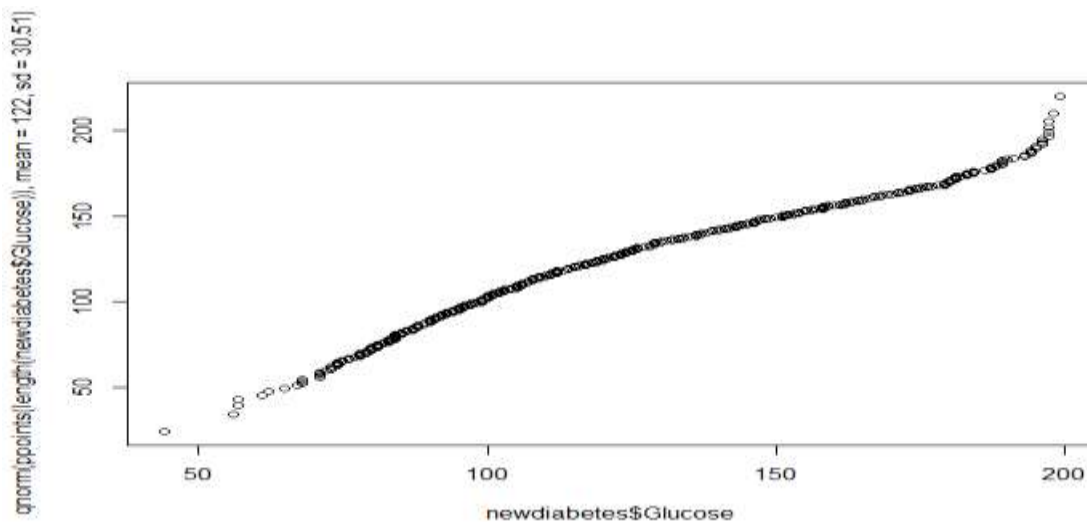
```
#Conclusion: p-value=0.84>0.05. Do not reject HO
```

```
#To test:
```

H0: Normal distribution is a good fit to the data on Glucose

versus H1: Normal distribution is not a good fit

```
test2<- qqplot(newdiabetes$Glucose,
  qnorm(ppoints(length(newdiabetes$Glucose))
    , mean = 122, sd = 30.51))
```



#Conclusion : normal distribution of Glucose with mean =122 and sd = 30.51 is a good fit.

2.5.4 Blood Pressure

```
sdBP<- sd(newdiabetes$BloodPressure)
```

```
#To test H0=Average Blood pressure =71.72 (#71.72 is the mean Blood Pressure)
```

```
H1= Average Blood Pressure < 71.72
```

```
BP<- t.test(x = newdiabetes$BloodPressure, mu = 71.72, alternative =
  "less")
```

BP

```
#Conclusion : p-value=0.49>0.05. Do not reject HO
```

```
#To test H0: Average Blood Pressure = 71.72
```

Dr. Shami Nimgulkar Kamble

versus H1: Average Blood Pressure \neq 71.72

```
BP1<- t.test(x = newdiabetes$BloodPressure, mu = 71.72, alternative =
"two.sided")
```

BP1

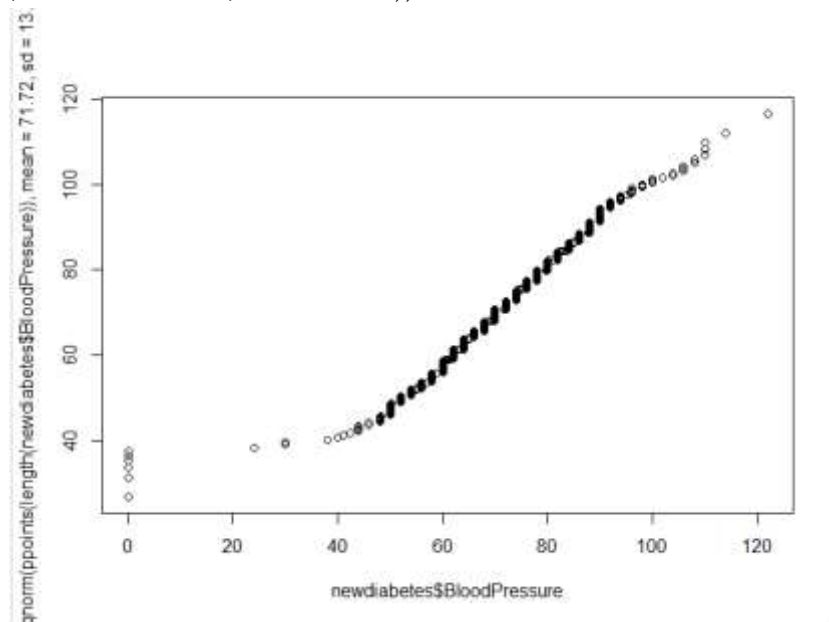
#Conclusion: p-value=0.99>0.05. Do not reject H0

#To test:

H0: Normal distribution is a good fit to the data on Blood Pressure

versus H1: Normal distribution is not a good fit

```
test3<- qqplot(newdiabetes$BloodPressure,
qnorm(ppoints(length(newdiabetes$BloodPressure))
, mean = 71.72, sd = 13.96))
```



#Conclusion : normal distribution of Blood Pressure with mean =71.72 and sd = 13.96 is a good fit.

2.5.5 Skin Thickness

```
sdskinthickness<- sd(newdiabetes$SkinThickness)
```

#To test H0=Average Skin Thickness =28.64 (#26.64 is the mean skin thickness)

H1= Average Skin Thickness < 28.64

```
Skinthickness<- t.test(x = newdiabetes$SkinThickness, mu = 28.64, alternative =
"less")
```

Skinthickness

#Conclusion : p-value=0.50>0.05. Do not reject H0

#To test H0: Average Skin Thickness = 28.64

versus H1: Average Skin Thickness \neq 28.64

```
ST1<- t.test(x = newdiabetes$SkinThickness, mu = 28.64, alternative =
"two.sided")
```

ST1

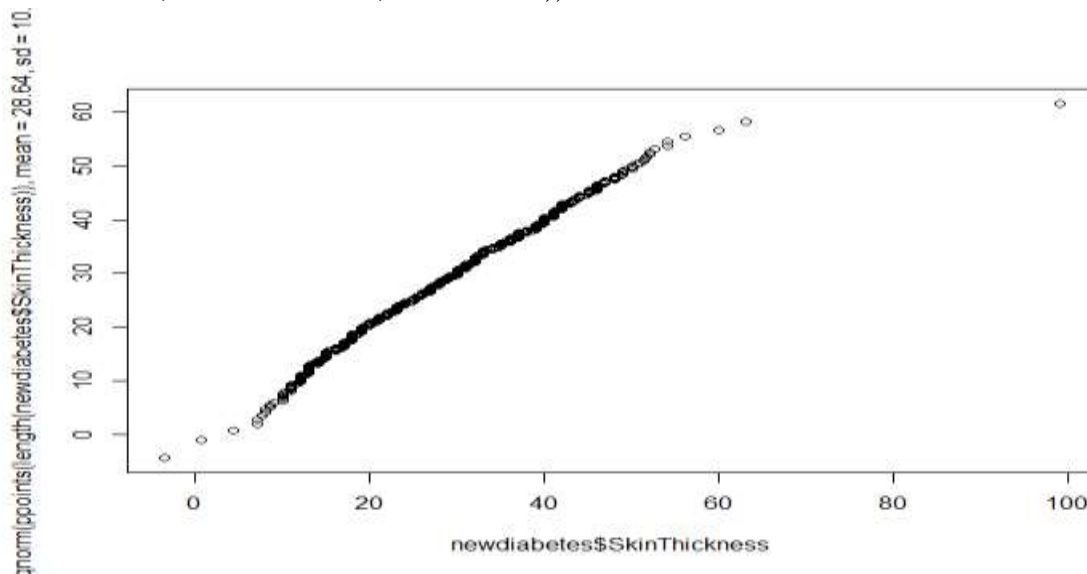
Dr. Shami Nimgulkar Kamble

#Conclusion: p-value=0.98>0.05. Do not reject HO

#To test:

H0: Normal distribution is a good fit to the data on Skin Thickness
versus H1: Normal distribution is not a good fit

```
test4<- qqplot(newdiabetes$SkinThickness,
               qnorm(ppoints(length(newdiabetes$SkinThickness))
                   , mean = 28.64, sd = 10.28))
```



#Conclusion : normal distribution of Skin Thickness with mean =28.64 and sd = 10.28 is a good fit.

2.5.6 Insulin

```
sdinsulin<- sd(newdiabetes$Insulin)
```

#To test H0=Average Insulin =155 (#155 is the mean Insulin)

H1= Average Insulin < 155

```
Insulin<- t.test(x = newdiabetes$Insulin, mu = 155, alternative =
                "less")
```

Insulin

#Conclusion : p-value=0.50>0.05. Do not reject HO

#To test H0: Average Insulin = 155

versus H1: Average Insulin \neq 155

```
Insulin1<- t.test(x = newdiabetes$Insulin, mu = 155, alternative =
                 "two.sided")
```

Insulin1

#Conclusion: p-value=0.99>0.05. Do not reject HO

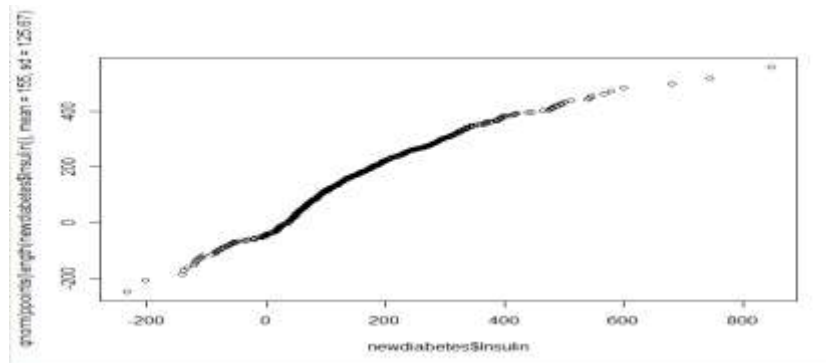
#To test:

H0: Normal distribution is a good fit to the data on Insulin
versus H1: Normal distribution is not a good fit

```
test5<- qqplot(newdiabetes$Insulin,
```

Dr. Shami Nimgulkar Kamble

```
qnorm(ppoints(length(newdiabetes$Insulin))
, mean = 155, sd = 125.67))
```



#Conclusion : normal distribution of Insulin with mean =155 and sd = 125.67 is a good fit.

2.5.7 BMI

```
sdbmi<- sd(newdiabetes$BMI)
```

```
#To test H0=Average BMI =32.28 (#32.28 is the mean BMI)
```

```
H1= Average BMI < 32.28
```

```
BMI<- t.test(x = newdiabetes$BMI, mu = 32.28, alternative =
"less")
```

BMI

```
#Conclusion : p-value=0.50>0.05. Do not reject HO
```

```
#To test H0: Average BMI = 32.28
```

```
versus H1: Average BMI ≠ 32.28
```

```
BMI1<- t.test(x = newdiabetes$BMI, mu = 32.28, alternative =
"two.sided")
```

BMI1

```
#Conclusion: p-value=0.99>0.05. Do not reject HO
```

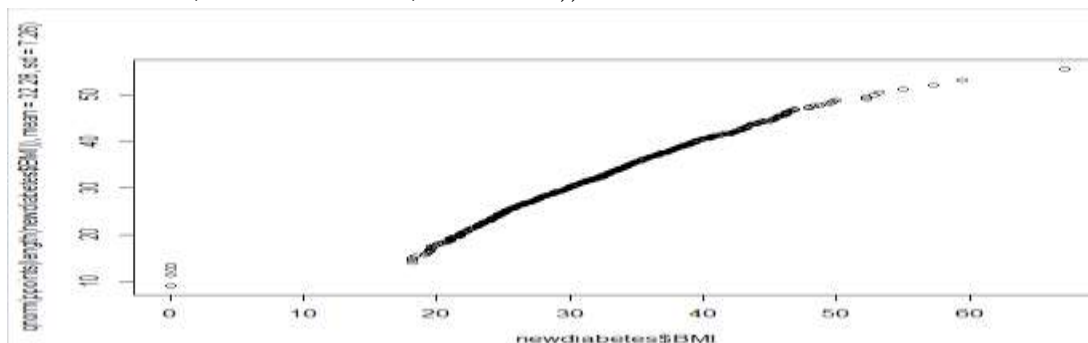
```
#To test:
```

```
H0: Normal distribution is a good fit to the data on BMI
```

```
versus H1: Normal distribution is not a good fit
```

```
test6<- qqplot(newdiabetes$BMI,
```

```
qnorm(ppoints(length(newdiabetes$BMI))
, mean = 32.28, sd = 7.26))
```



#Conclusion : normal distribution of BMI with mean =32.28 and sd = 7.26 is a good fit.

2.5.8 Diabetes Pedigree Function

```
sdpedigree<- sd(newdiabetes$DiabetesPedigreeFunction)
#To test H0=Average DPF =0.47 (#0.47 is the mean diabetespedigreefunction)
H1= Average DPF < 0.47
DPF<- t.test(x = newdiabetes$DiabetesPedigreeFunction, mu = 0.47, alternative =
"less")
```

DPF

#Conclusion : p-value=0.56>0.05. Do not reject H0

#To test H0: Average DPF = 0.47

versus H1: Average DPF \neq 0.47

```
DPF1<- t.test(x = newdiabetes$DiabetesPedigreeFunction, mu = 0.47,
alternative =
"two.sided")
```

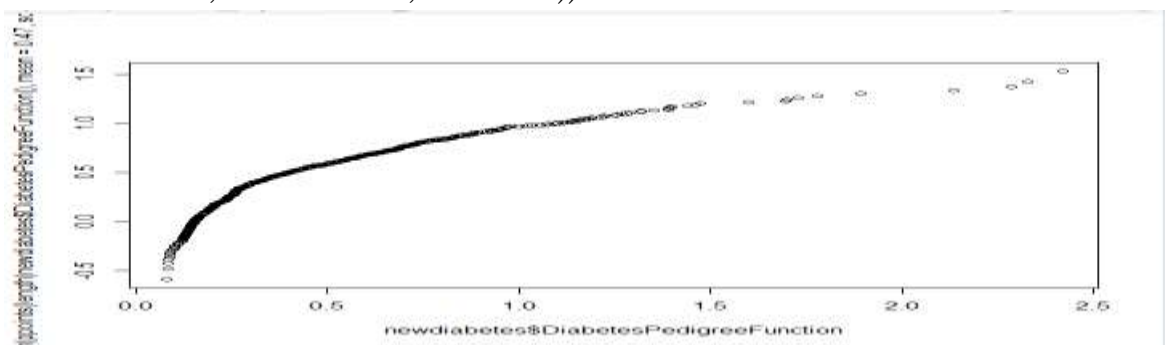
DPF1

[#Conclusion: p-value=0.87>0.05. Do not reject H0

To test:

H0: Normal distribution is a good fit to the data on Diabetespedigreefunction versus H1: Normal distribution is not a good fit

```
test7<- qqplot(newdiabetes$DiabetesPedigreeFunction,
qnorm(ppoints(length(newdiabetes$DiabetesPedigreeFunction))
, mean = 0.47, sd = 0.33))
```



#Conclusion : normal distribution of DiabetesPedigreeFunction with mean =0.47and sd = 0.33 is a good fit.

3. Findings

Since the data set follows a normal distribution we will apply the linear model to estimate the scores:

```
install.packages("lmtest")
```

```
library(lmtest)
```

Dr. Shami Nimgulkar Kamble

#THIS IS OUR LINEAR MODEL FIT TO FIND OUT THE STATISTICAL SIGNIFICANCE OF INDIVIDUAL PARAMETERS

```
colnames(newdiabetes)
```

```
mod=lm(Outcome~newdiabetes$Pregnancies+newdiabetes$Glucose+newdiabetes$BloodPressure+newdiabetes$SkinThickness+newdiabetes$Insulin+newdiabetes$BMI+newdiabetes$DiabetesPedigreeFunction+newdiabetes$Age,
```

```
data=newdiabetes)
```

```
mod
```

```
summary(mod)
```

```
Call:
```

```
lm(formula = Outcome ~ newdiabetes$Pregnancies + newdiabetes$Glucose +
    newdiabetes$BloodPressure + newdiabetes$SkinThickness + newdiabetes$Insulin +
    newdiabetes$BMI + newdiabetes$DiabetesPedigreeFunction +
    newdiabetes$Age, data = newdiabetes)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-1.13830 -0.28776 -0.07953  0.29521  0.96882
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-9.635e-01	9.887e-02	-9.745	< 2e-16	***
newdiabetes\$Pregnancies	2.038e-02	5.058e-03	4.030	6.15e-05	***
newdiabetes\$Glucose	6.552e-03	6.008e-04	10.905	< 2e-16	***
newdiabetes\$BloodPressure	-2.172e-03	1.120e-03	-1.940	0.05276	.
newdiabetes\$SkinThickness	4.962e-04	1.772e-03	0.280	0.77951	
newdiabetes\$Insulin	-4.215e-05	1.426e-04	-0.296	0.76761	
newdiabetes\$BMI	1.366e-02	2.608e-03	5.238	2.10e-07	***
newdiabetes\$DiabetesPedigreeFunction	1.370e-01	4.360e-02	3.141	0.00175	**
newdiabetes\$Age	2.362e-03	1.565e-03	1.509	0.13162	

```
---
```

```
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3933 on 759 degrees of freedom
```

```
Multiple R-squared:  0.327,    Adjusted R-squared:  0.3199
```

```
F-statistic: 46.09 on 8 and 759 DF,  p-value: < 2.2e-16
```

The tabulated t value for 759 degrees of freedom at 0.1, 0.05 and 0.01 confidence interval is -1.28267, -1.64686 and -2.33127. So, we can infer that, the intercept (outcome) is not statistically significant. All the estimate terms except Insulin are statistically significant in increasing the risk of diabetes in women considering the given sample.

However, going by the p value, since the p value of blood pressure 5.27 % < 10% and diabetes pedigree function 0.175% < 1%, these are the two estimates that are statistically significant in increasing the risk of diabetes in women in the given sample. All other factors are statistically insignificant.

Conclusion

Minor changes in lifestyle can greatly reduce the chances of getting diabetes. Proper testing, treatment and lifestyle changes, healthy eating as a strategy, promoting walking, exercise, and other physical activities will have beneficial effects on prevention or treatment of diabetes.

Dr. Shami Nimgulkar Kamble

References:

1. Ali, A. (2013), Analyzing Pima-Indian-Diabetes dataset, Available on <https://medium.com/analytics-vidhya/analyzing-pima-indian-diabetes-dataset-36d02a8a10e> Accessed on: September 16, 2021
2. Business Line (2020) Over half of men and two-thirds of women in India below 20 years likely to develop diabetes:Study, December 02, 2020, Available on: <https://www.thehindubusinessline.com/news/science/over-half-of-men-and-two-thirds-of-women-in-india-below-20-years-likely-to-develop-diabetes-study/article33227262.ece>, Accessed on: September 16, 2021
3. NDTV (2019) Indian Women At High Death Risk From Diabetes, Finds Study, April 23, 2019, Available at: <https://www.ndtv.com/health/diabetes-indian-women-at-high-death-risk-from-diabetes-finds-study-2027180>, Accessed on: September 16, 2021
4. Pima Indians Diabetes Database, Available on: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>.
5. Saliha,N. (2018) Women's diabetes in numbers: Why female population in India ignores the risks? *The Economic Times | Panache*, May 18, 2018, Available at: <https://economictimes.indiatimes.com/magazines/panache/womens-diabetes-in-numbers-why-indias-femalepopulationignoretherisks/articleshow/63590014.cms?from=mdr>, Accessed on :September 16, 2021