SP JAIN SCHOOL OF GLOBAL MANAGEMENT, MUMBAI

STATISTICAL ANALYSIS ON DIABETES PROGRESSION

Prepared By: Sooryajith M.Y.

Student ID: BS21DMU016

Guided By: Dr Suchismita Das

CONTENT

INTRODUCTION

Diabetes is a chronic condition that develops when blood glucose (also known as blood sugar) levels are too high. The main source of energy is blood glucose, which is obtained from food. Insulin, a hormone produced by the pancreas, aids glucose absorption into cells for energy usage. Sometimes the body does not produce enough insulin or does not utilise it properly. Glucose remains in the bloodstream and does not reach the cells. Having too much glucose in your blood might lead to health issues over time. Although there is no cure for diabetes, we may take steps to manage it and be healthy.

THE DIABETES DATASET

For each of 442 diabetic patients, 10 baseline characteristics, including age, sex, BMI, average blood pressure, and six blood serum measures, as well as the response of interest, a quantitative measure of disease progression one year after baseline, were gathered. This dataset was first published in Annals of Statistics in 2004 where it was used in "Least Angle Regression" by Efron et al. The data consists of the following attributes:

1. **Disease Progression one year after baseline(Y)**
2. Age
3. Gender
4. BMI
5. BP
6. Total Cholesterol
7. LDL
8. HDL
9. TCH
10. LTG
11. Glucose

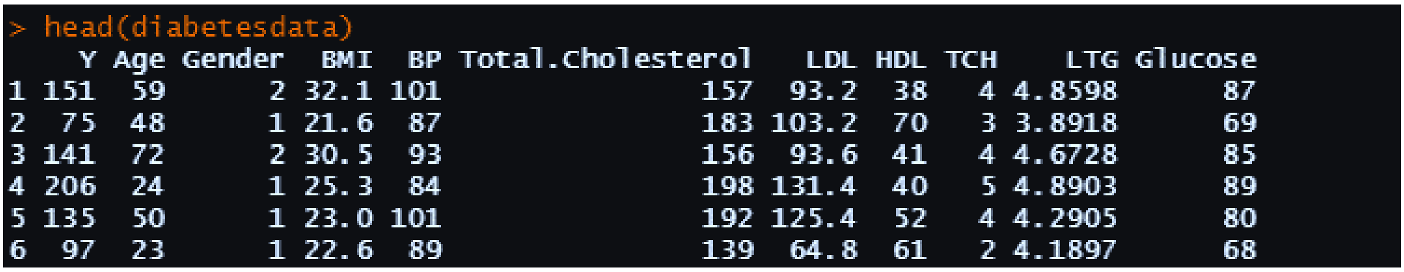
Y is the dependant variable (response variable) to be studied based on the input variables.

DATA UNDERSTANDING

Sample Data

The sample Data is as shown below:

Note: The figure henceforth shown in the report will be an output from R-Code and JMP. Box plot, Histogram, Correlation Matrix, Correlation Scatter plot were plotted using JMP.



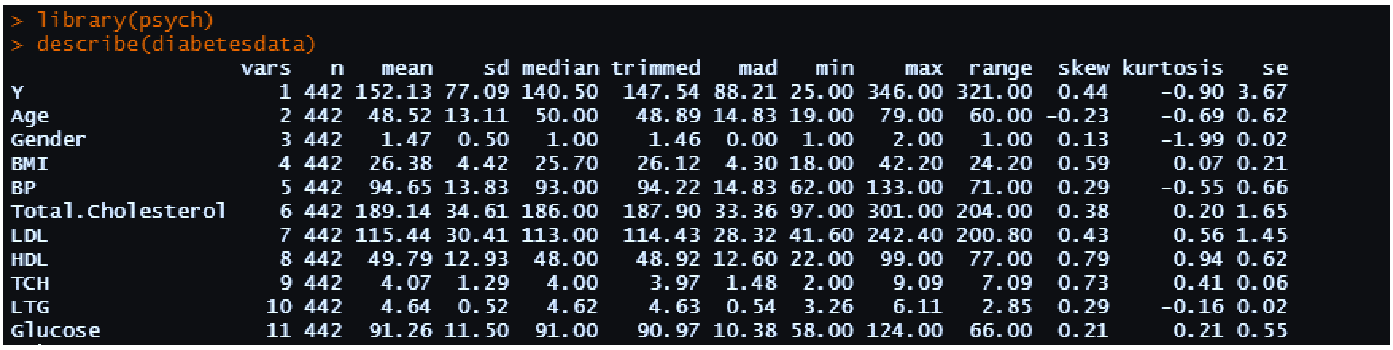
```
> head(diabetesdata)
    Y Age Gender  BMI  BP Total.Cholesterol   LDL HDL TCH    LTG Glucose
1 151  59      2 32.1 101               157  93.2  38   4 4.8598      87
2  75  48      1 21.6  87               183 103.2  70   3 3.8918      69
3 141  72      2 30.5  93               156  93.6  41   4 4.6728      85
4 206  24      1 25.3  84               198 131.4  40   5 4.8903      89
5 135  50      1 23.0 101               192 125.4  52   4 4.2905      80
6  97  23      1 22.6  89               139  64.8  61   2 4.1897      68
```

As per the NOIR classification (Nominal, Ordinal, Interval and Ratio classification) the data in dataset can be classified into Interval data of continuous type.

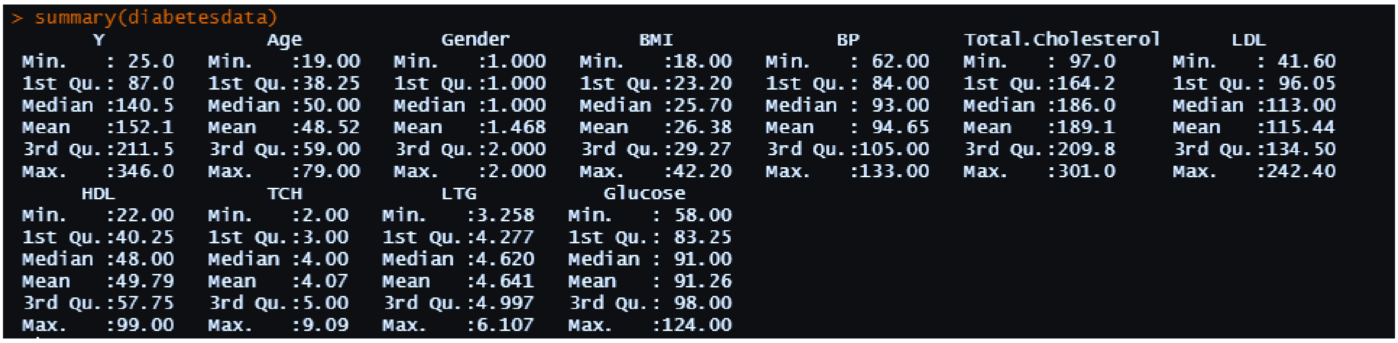
NULL value test was performed on the dataset. No NULL values where present in the dataset.

Key Statistics for Data

Before we proceed let us find the key parameters of the data attribute.



```
> library(psych)
> describe(diabetesdata)
                  vars   n   mean    sd median trimmed   mad   min    max  range  skew kurtosis   se
Y                    1 442 152.13 77.09 140.50  147.54 88.21 25.00 346.00 321.00  0.44    -0.90 3.67
Age                  2 442  48.52 13.11  50.00   48.89 14.83 19.00  79.00  60.00 -0.23    -0.69 0.62
Gender               3 442   1.47  0.50   1.00    1.46  0.00  1.00   2.00   1.00  0.13    -1.99 0.02
BMI                  4 442  26.38  4.42  25.70   26.12  4.30 18.00  42.20  24.20  0.59     0.07 0.21
BP                   5 442  94.65 13.83  93.00   94.22 14.83 62.00 133.00  71.00  0.29    -0.55 0.66
Total.Cholesterol    6 442 189.14 34.61 186.00  187.90 33.36 97.00 301.00 204.00  0.38     0.20 1.65
LDL                  7 442 115.44 30.41 113.00  114.43 28.32 41.60 242.40 200.80  0.43     0.56 1.45
HDL                  8 442  49.79 12.93  48.00   48.92 12.60 22.00  99.00  77.00  0.79     0.94 0.62
TCH                  9 442   4.07  1.29   4.00    3.97  1.48  2.00   9.09   7.09  0.73     0.41 0.06
LTG                 10 442   4.64  0.52   4.62    4.63  0.54  3.26   6.11   2.85  0.29    -0.16 0.02
Glucose             11 442  91.26 11.50  91.00   90.97 10.38 58.00 124.00  66.00  0.21     0.21 0.55
```



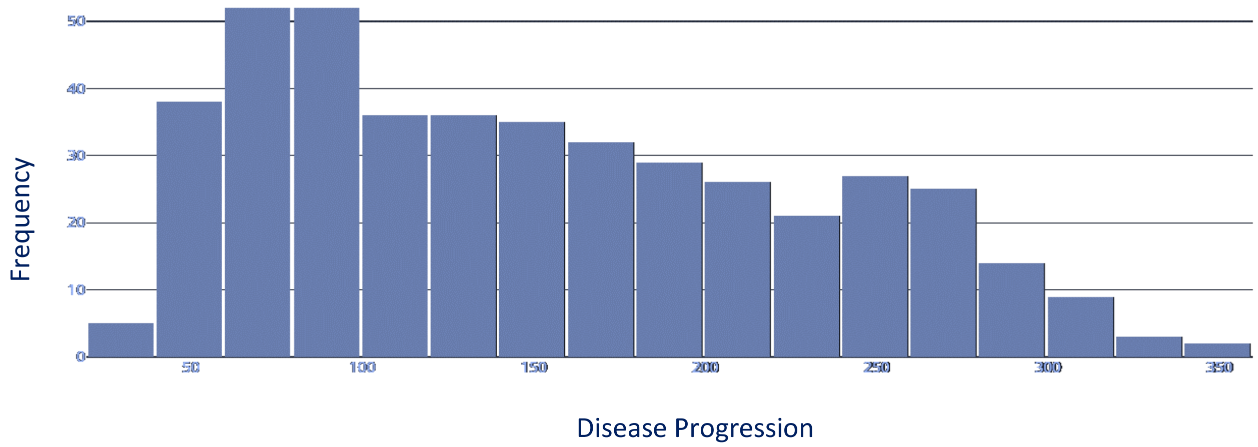
```
> summary(diabetesdata)
       Y              Age            Gender           BMI              BP         Total.Cholesterol       LDL
 Min.   : 25.0   Min.   :19.00   Min.   :1.000   Min.   :18.00   Min.   : 62.00   Min.   : 97.0     Min.   : 41.60
 1st Qu.: 87.0   1st Qu.:38.25   1st Qu.:1.000   1st Qu.:23.20   1st Qu.: 84.00   1st Qu.:164.2     1st Qu.: 96.05
 Median :140.5   Median :50.00   Median :1.000   Median :25.70   Median : 93.00   Median :186.0     Median :113.00
 Mean   :152.1   Mean   :48.52   Mean   :1.468   Mean   :26.38   Mean   : 94.65   Mean   :189.1     Mean   :115.44
 3rd Qu.:211.5   3rd Qu.:59.00   3rd Qu.:2.000   3rd Qu.:29.27   3rd Qu.:105.00   3rd Qu.:209.8     3rd Qu.:134.50
 Max.   :346.0   Max.   :79.00   Max.   :2.000   Max.   :42.20   Max.   :133.00   Max.   :301.0     Max.   :242.40
      HDL             TCH            LTG            Glucose
 Min.   :22.00   Min.   :2.00   Min.   :3.258   Min.   : 58.00
 1st Qu.:40.25   1st Qu.:3.00   1st Qu.:4.277   1st Qu.: 83.25
 Median :48.00   Median :4.00   Median :4.620   Median : 91.00
 Mean   :49.79   Mean   :4.07   Mean   :4.641   Mean   : 91.26
 3rd Qu.:57.75   3rd Qu.:5.00   3rd Qu.:4.997   3rd Qu.: 98.00
 Max.   :99.00   Max.   :9.09   Max.   :6.107   Max.   :124.00
```

From the above table output following key observations can be made:

1. Mean of Y= 152.13 is higher than median 140.50, indicating a positive skewness
    a. We will check for outliers and make the mean closer to median
2. The range of Y is [25,346].
3. HDL is having the highest kurtosis (4th derivative of moment generating function) that is 0.94.

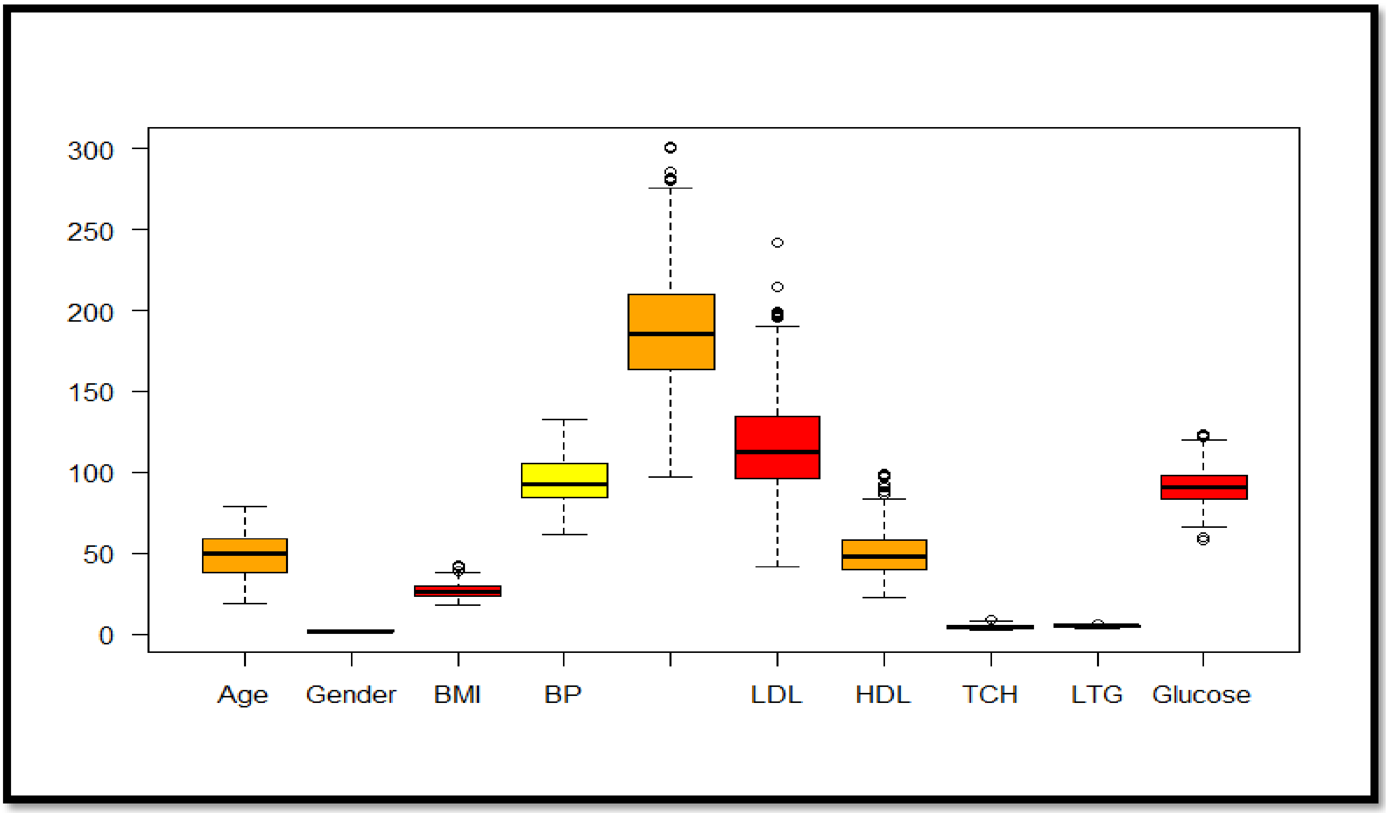
Histogram & Box Plots

Let us draw some important plots to understand the distribution of our dependent variable:



As the above figure is a normal approximation but slightly positively skewed.

The box plot below shows how each attribute is classified and how much outliers are present



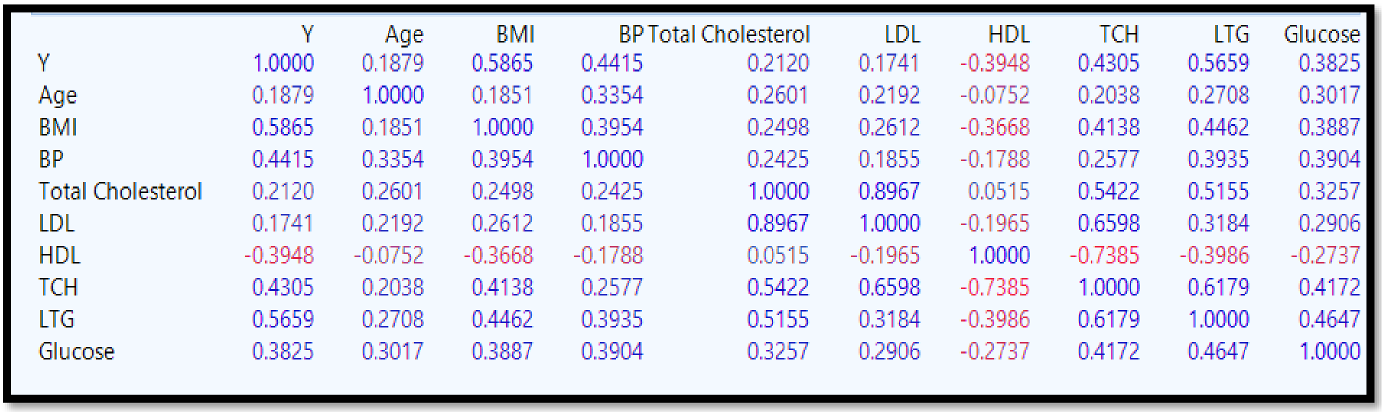
From the box plot, it can be inferred that there are outliers present in the data. However, after analysing the outliers, it was found that no parameters were able to be removed due to the correlation present between the input variables and response variable.

Correlation Matrix

Correlation coefficient between two random variables X and Y, usually denoted by $r\ (X, Y)\ or\ rXY$ is a numerical measure of linear relationship between them and is defined as:

$$R_{XY} = \frac{Cov\ (X, Y)}{\sigma X \sigma Y}$$

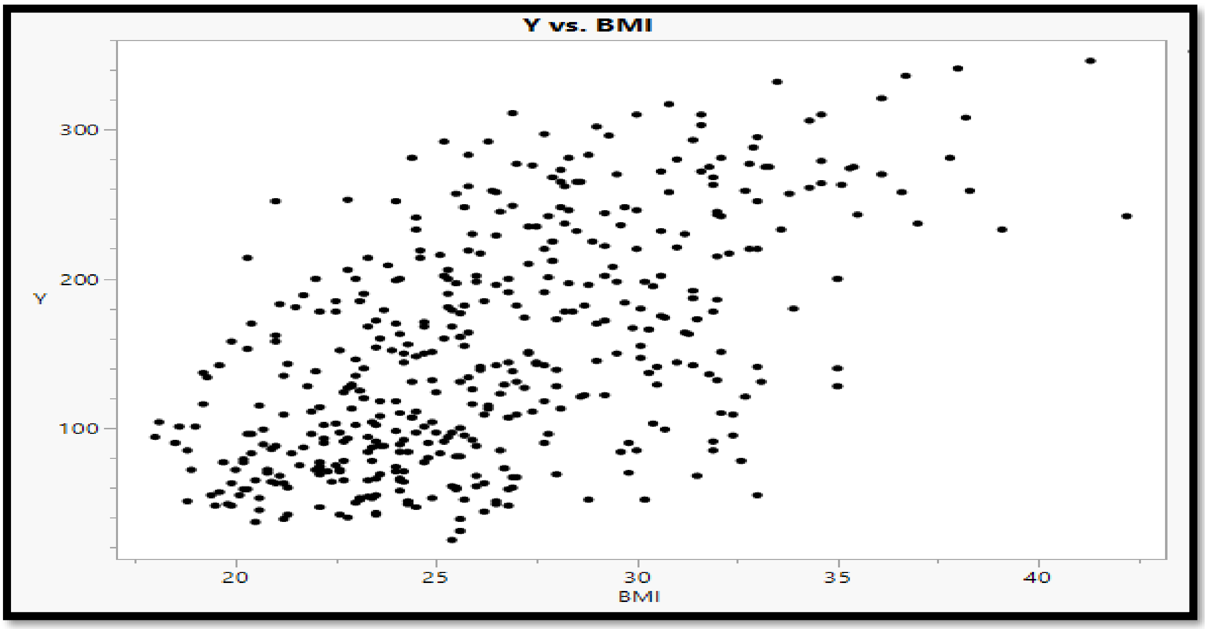
- rXY provided a measure of linear relationship between X and Y.
- It is a measure of degree of relationship.



	Y	Age	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose
Y	1.0000	0.1879	0.5865	0.4415	0.2120	0.1741	-0.3948	0.4305	0.5659	0.3825
Age	0.1879	1.0000	0.1851	0.3354	0.2601	0.2192	-0.0752	0.2038	0.2708	0.3017
BMI	0.5865	0.1851	1.0000	0.3954	0.2498	0.2612	-0.3668	0.4138	0.4462	0.3887
BP	0.4415	0.3354	0.3954	1.0000	0.2425	0.1855	-0.1788	0.2577	0.3935	0.3904
Total Cholesterol	0.2120	0.2601	0.2498	0.2425	1.0000	0.8967	0.0515	0.5422	0.5155	0.3257
LDL	0.1741	0.2192	0.2612	0.1855	0.8967	1.0000	-0.1965	0.6598	0.3184	0.2906
HDL	-0.3948	-0.0752	-0.3668	-0.1788	0.0515	-0.1965	1.0000	-0.7385	-0.3986	-0.2737
TCH	0.4305	0.2038	0.4138	0.2577	0.5422	0.6598	-0.7385	1.0000	0.6179	0.4172
LTG	0.5659	0.2708	0.4462	0.3935	0.5155	0.3184	-0.3986	0.6179	1.0000	0.4647
Glucose	0.3825	0.3017	0.3887	0.3904	0.3257	0.2906	-0.2737	0.4172	0.4647	1.0000

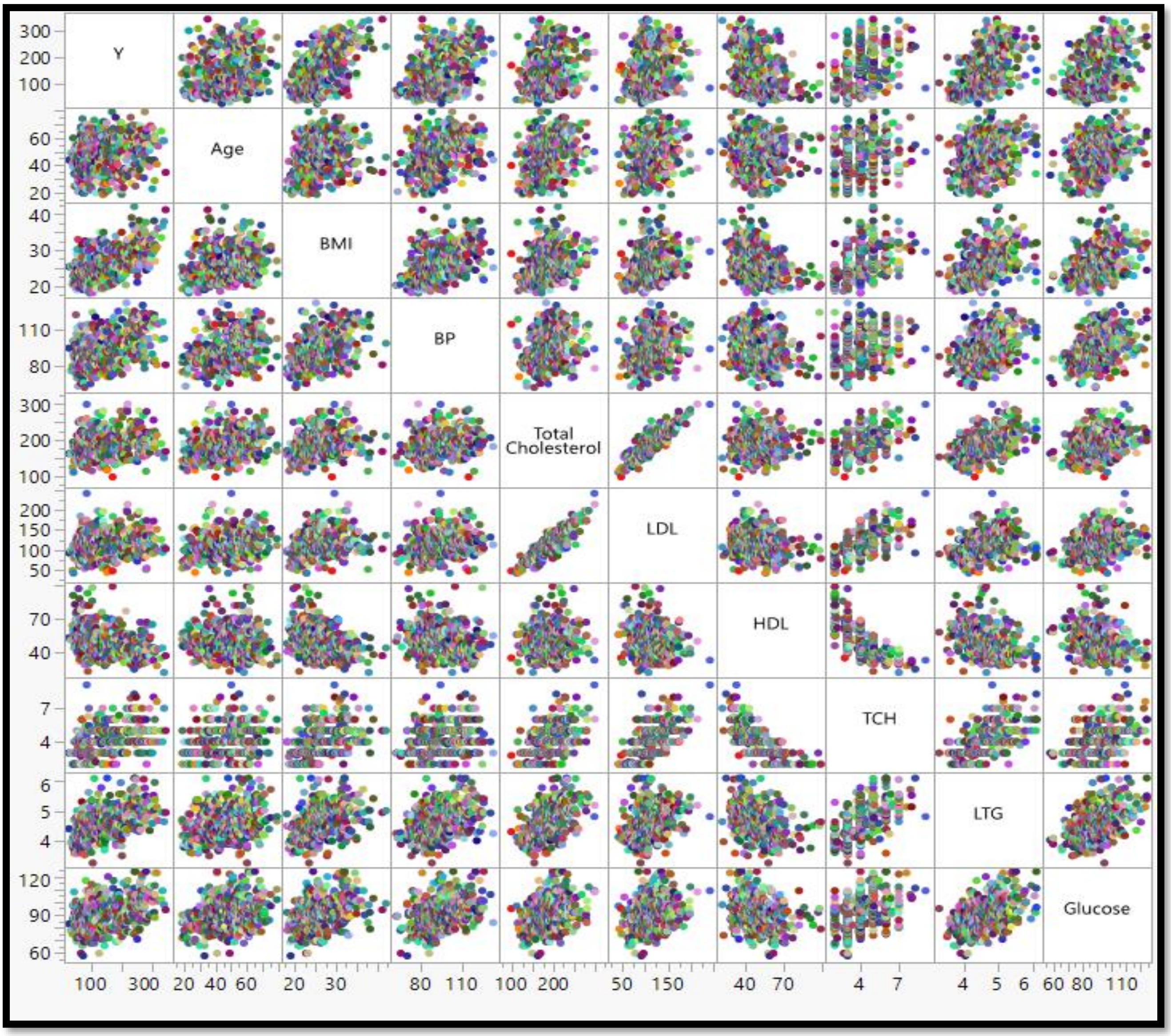
Some observations:

- BMI has the highest correlation with the response variable comparing to other input variables. From the scatterplot it can be easily observed that BMI has an average positive correlation with Y.
- LDL has the least correlation with the response variable.
- There is high correlation between the independent variables like LDL and Total Cholesterol indicating there is multicollinearity which have to be solved during regression analysis.



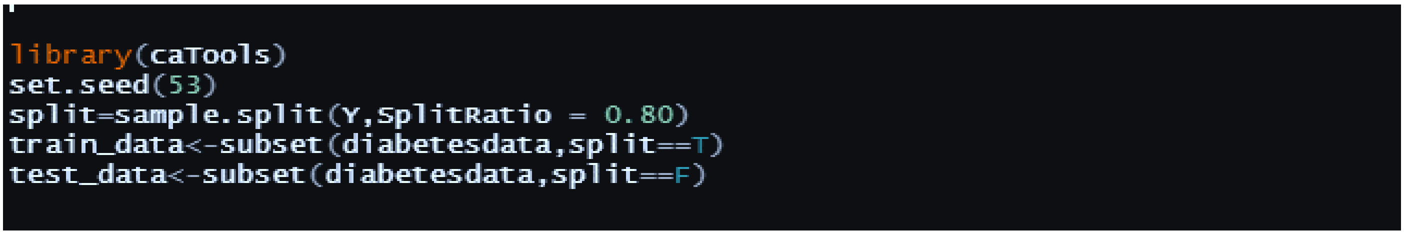
Correlation Scatterplot

Correlation Scatterplot is a multivariate descriptive plot which are designed to reveal the relationship among several variables simultaneously. It displays the strength, direction, and form of relationship between every variable.



TRAIN AND TEST SPLIT

The data was split into training and testing data. There was a total of 442 instances. After splitting with a ratio of 80:20, there was 380 observations in training data set and 62 observations in the testing data. The following lines of code were used for splitting the data.



```
library(caTools)
set.seed(53)
split=sample.split(Y,SplitRatio = 0.80)
train_data<-subset(diabetesdata,split==T)
test_data<-subset(diabetesdata,split==F)
```

The multiple linear regression was applied on the training data set only.

REGRESSION ANALYSIS

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. Assumptions of multiple linear regression

Multiple linear regression makes all of the same assumptions as simple linear regression:

Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.

Independence of observations: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.

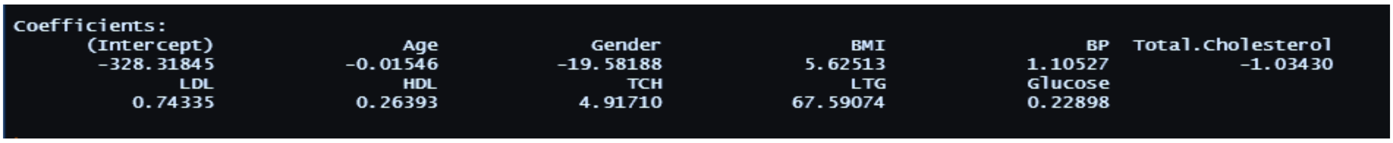
The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- B_1X_1= the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- B_nX_n = the regression coefficient of the last independent variable
- e = model error (a.k.a. how much variation there is in our estimate of y)

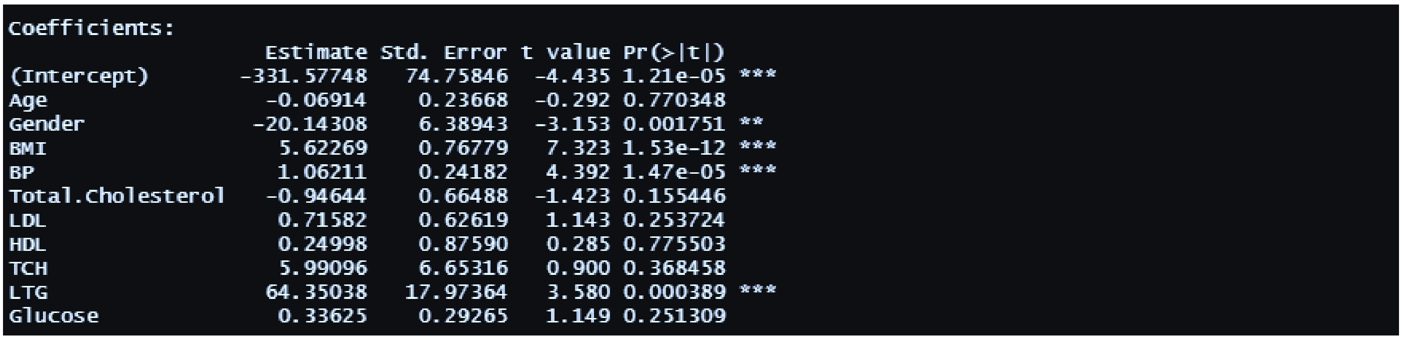
Model 1

The coefficients of x variables (β) are given below:



```
Coefficients:
      (Intercept)                Age             Gender                BMI                 BP  Total.Cholesterol
       -328.31845           -0.01546          -19.58188            5.62513            1.10527           -1.03430
              LDL                HDL                TCH                LTG            Glucose
          0.74335            0.26393            4.91710           67.59074            0.22898
```

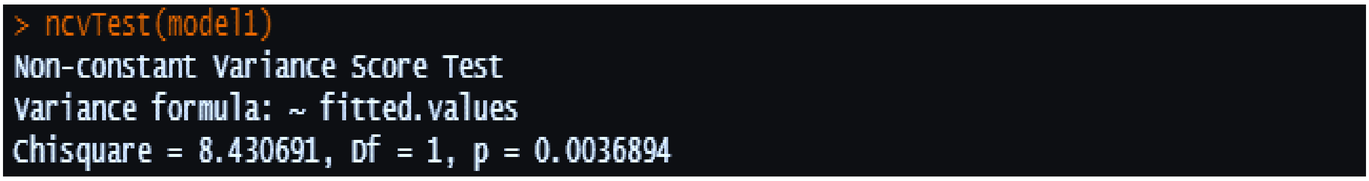
Output of Model 1 is given as follow:



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -331.57748   74.75846  -4.435 1.21e-05 ***
Age                -0.06914    0.23668  -0.292 0.770348    
Gender            -20.14308    6.38943  -3.153 0.001751 ** 
BMI                 5.62269    0.76779   7.323 1.53e-12 ***
BP                  1.06211    0.24182   4.392 1.47e-05 ***
Total.Cholesterol  -0.94644    0.66488  -1.423 0.155446    
LDL                 0.71582    0.62619   1.143 0.253724    
HDL                 0.24998    0.87590   0.285 0.775503    
TCH                 5.99096    6.65316   0.900 0.368458    
LTG                64.35038   17.97364   3.580 0.000389 ***
Glucose             0.33625    0.29265   1.149 0.251309    
```

STATISTIC	VALUE
Residual standard error	54.69
Multiple R- Squared	0.514
Adjusted R-Squared	0.5008
F Statistic	39.03
P value	< 2.2e-16

From the output it is clear that the P value is < 2.2e-16 which is < 0.05. Therefore, it indicates that the Model-1 holds good for predicting the output. However, the NCV Test was also conducted which computes a score test of the null hypothesis of constant error variance against the alternative that the error variance changes with the level of the response (fitted values), or with a linear combination of predictors.

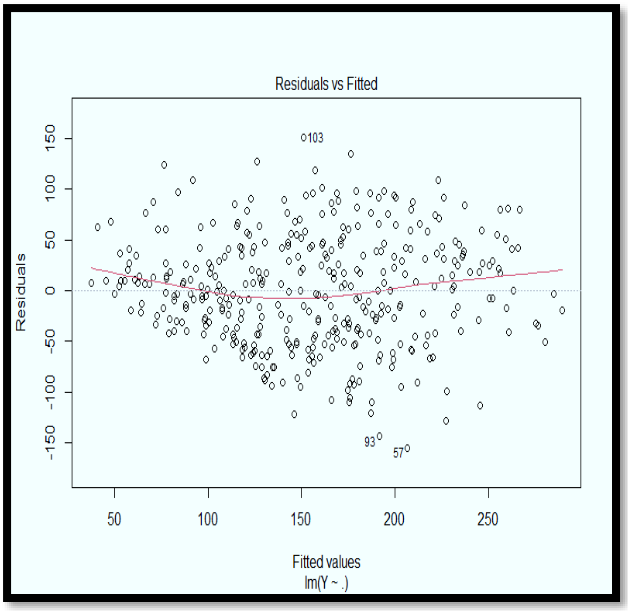


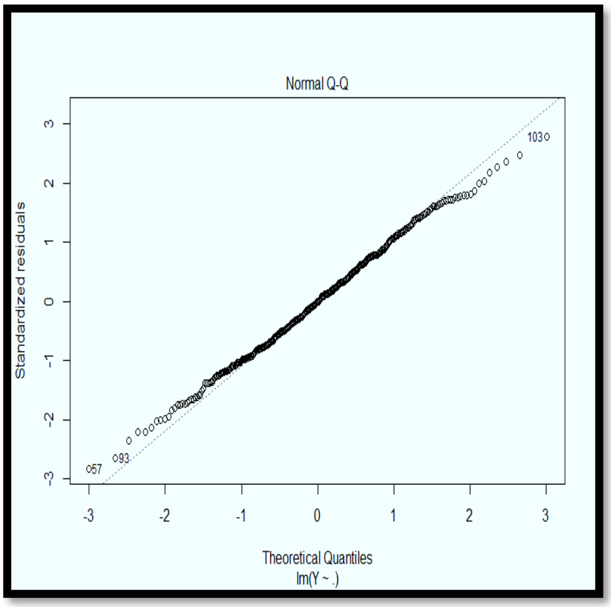
```
> ncvTest(model1)
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 8.430691, Df = 1, p = 0.0036894
```

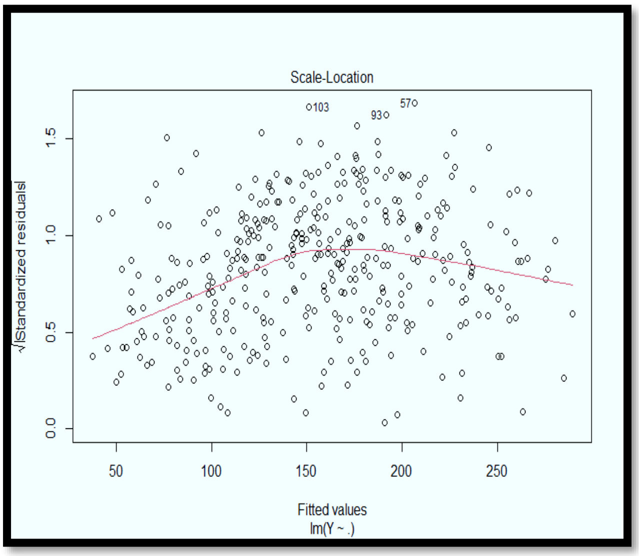
The p-value is 0.00368 which is very highly significant. But the respective p values for few variables such as HDL, TCH, LDL are not significant indicating they don’t have any impact in predicting.

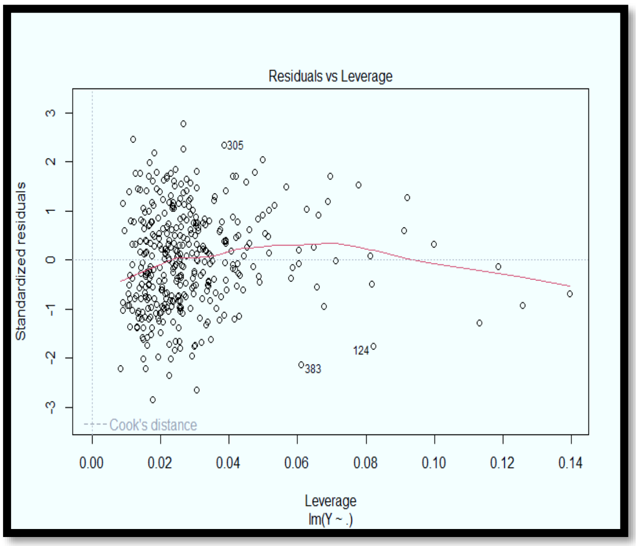
Validating Model 1

Plots related to the model for validating the model









- **Residuals versus fits plot**: When conducting a residual analysis, a " **Residuals versus fits plot** " is the most frequently created plot. It is a scatter plot of residuals on the *y* axis and fitted values (estimated responses) on the *x* axis. The plot is used to detect non-linearity, unequal error variances, and outliers.

- **Q-Q plot:** A straight line suggests that the residual errors are normally distributed. It can show the skewness and outliers for the residual. Here the presence of outliers can be observed.

- **Residuals vs. leverage plot**: It is a type of diagnostic plot that allows us to identify influential observations in a regression model. Leverage measures how far away the XX values of an observation are from those of the other observations. Usually 2(k+1)/N2(k+1)/N (kk is the number of X variables in the model, and NN is the sample size) is used as the threshold for leverage.

Dealing Multicollinearity

Multicollinearity is defined the correlation between several independent variables. In the begin of the discussion from the correlation matrix it was observed there are high correlation between few independent variables which is not good for our model. So, for checking the multicollinearity VIF was conducted.

VIF stands for Variance Inflation Factor which tells the measure of the amount of multicollinearity present. Usually, Variance Inflation Factor (VIF) value over 10, or a mean of the VIF values over 5 indicates potential multicollinearity problem.

VIF observed was:

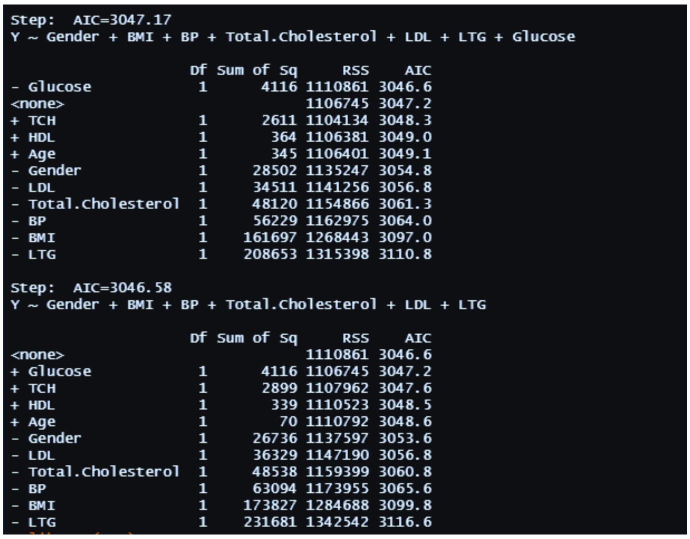
Age	**1.224297**
Gender	**1.267842**
BMI	**1.447970**
BP	**54.749114**
Total Cholesterol	**36.113844**

LDL	**36.113844**
HDL	**14.908731**
TCH	**9.278520**
LTG	**9.931825**
Glucose	**1.487262**

From the above output it is observed that there is multicollinearity in this data in variables Total Cholesterol, LDL, HDL, TCH and LTG.

One way to remove the multicollinearity problem is to remove one or more of highly correlated independent variable. Using stepwise regression by conducting step AIC. AIC stands for (Akaike information criterion).

Step AIC is performed on model 1 only considering all the input parameters:



```
Step:  AIC=3047.17
Y ~ Gender + BMI + BP + Total.Cholesterol + LDL + LTG + Glucose

                    Df Sum of Sq     RSS    AIC
- Glucose            1      4116 1110861 3046.6
<none>                           1106745 3047.2
+ TCH                1      2611 1104134 3048.3
+ HDL                1       364 1106381 3049.0
+ Age                1       345 1106401 3049.1
- Gender             1     28502 1135247 3054.8
- LDL                1     34511 1141256 3056.8
- Total.Cholesterol  1     48120 1154866 3061.3
- BP                 1     56229 1162975 3064.0
- BMI                1    161697 1268443 3097.0
- LTG                1    208653 1315398 3110.8

Step:  AIC=3046.58
Y ~ Gender + BMI + BP + Total.Cholesterol + LDL + LTG

                    Df Sum of Sq     RSS    AIC
<none>                           1110861 3046.6
+ Glucose            1      4116 1106745 3047.2
+ TCH                1      2899 1107962 3047.6
+ HDL                1       339 1110523 3048.5
+ Age                1        70 1110792 3048.6
- Gender             1     26736 1137597 3053.6
- LDL                1     36329 1147190 3056.8
- Total.Cholesterol  1     48538 1159399 3060.8
- BP                 1     63094 1173955 3065.6
- BMI                1    173827 1284688 3099.8
- LTG                1    231681 1342542 3116.6
```

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

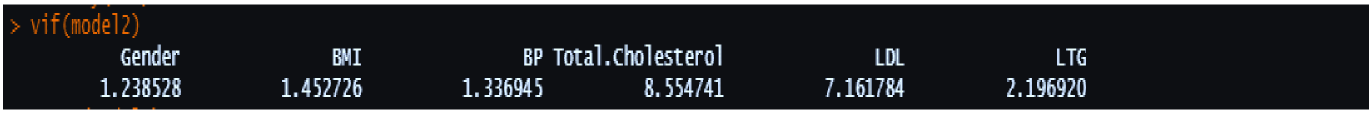
n: number of observations

$\hat{\sigma}^2$: estimate of error or residual variance

d: number of x variables included in the model

RSS: Residual sum of squares

Now to ensure that the multicollinearity is relieved, the VIF was conducted.

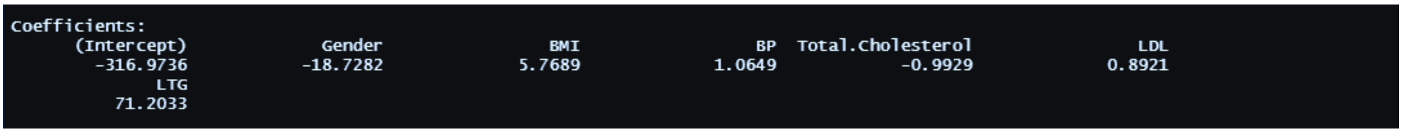


```
> vif(model2)
           Gender               BMI                BP Total.Cholesterol               LDL               LTG 
         1.238528          1.452726          1.336945          8.554741          7.161784          2.196920 
```

This shows that the multicollinearity problem is solved and the model-2 created after Step regression performs better than the earlier Model1.

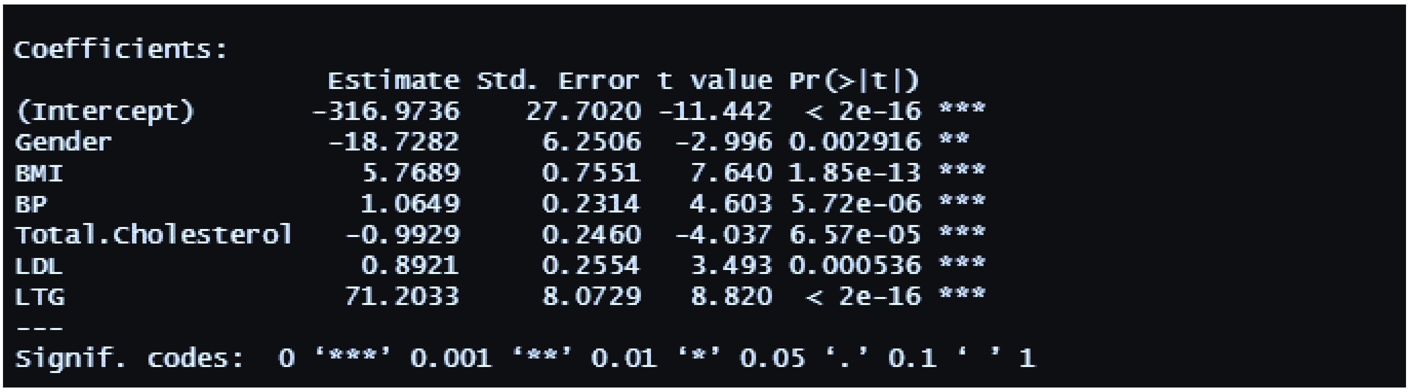
Model2

The coefficients of x variables (β) are given below:



```
Coefficients:
        (Intercept)             Gender                 BMI                  BP  Total.Cholesterol                 LDL
          -316.9736           -18.7282              5.7689              1.0649            -0.9929              0.8921
                LTG
            71.2033
```

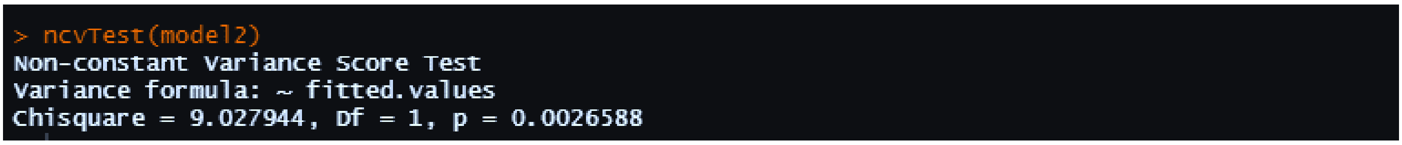
The summary of model2 is given below:



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -316.9736    27.7020 -11.442  < 2e-16 ***
Gender              -18.7282     6.2506  -2.996 0.002916 ** 
BMI                   5.7689     0.7551   7.640 1.85e-13 ***
BP                    1.0649     0.2314   4.603 5.72e-06 ***
Total.Cholesterol    -0.9929     0.2460  -4.037 6.57e-05 ***
LDL                   0.8921     0.2554   3.493 0.000536 ***
LTG                  71.2033     8.0729   8.820  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

STATISTIC	VALUE
Residual standard error	54.57
Multiple R- Squared	0.5108
Adjusted R-Squared	0.503
F Statistic	64.92
P value	< 2.2e-16

From the output it is clear that the P value is < 2.2e-16 which is < 0.05. Therefore, it indicates that the Model 2 holds good for predicting the output. However, the NCV Test was also conducted for this model as well.



```
> ncvTest(model2)
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 9.027944, Df = 1, p = 0.0026588
```

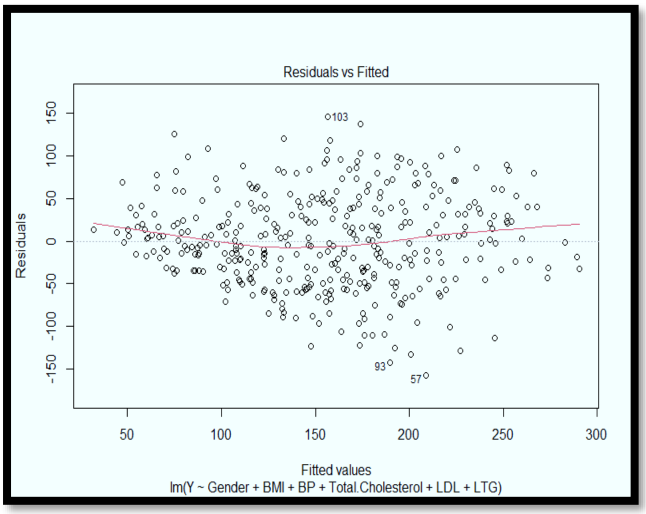
The p-value is 0.00265 which is very highly significant and is less than the p value of previous model. Therefore, we can say there is a slight improvement in our model2.

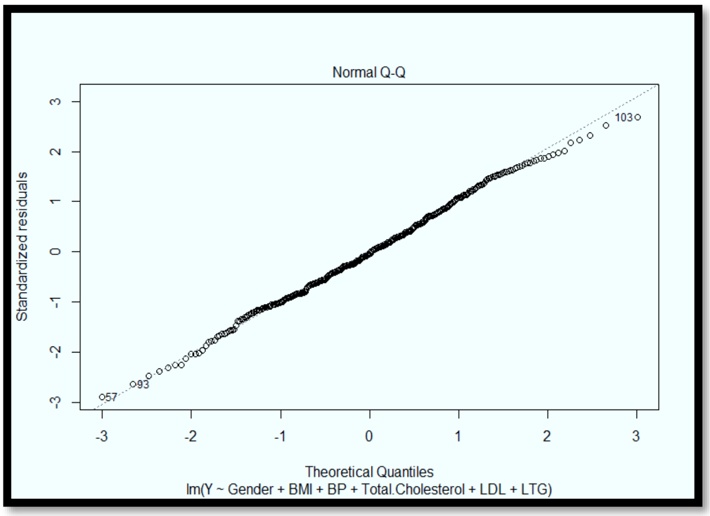
MODEL EQUATION:

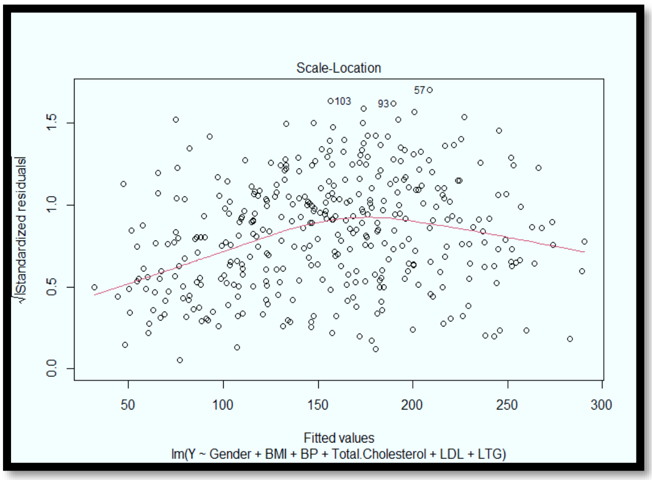
Y = – 316.97 + 71.2* LTG + 5.7689*BMI + 1.06*BP + 0.89*LDL – 0.99*Total Cholesterol – 18.72*Gender

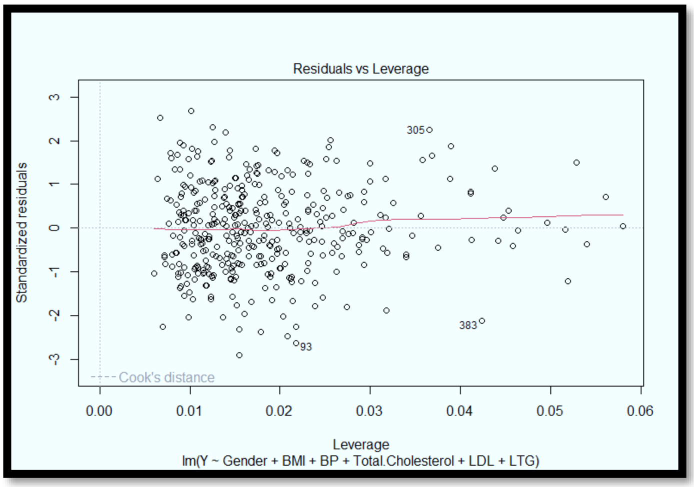
Validating Model-2

Plots related to Model-2









MSE

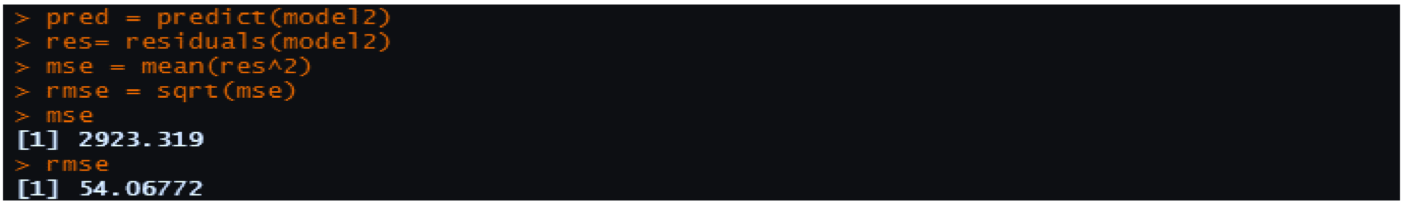
Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

For model2 the MSE Calculated is: 2923.319

RMSE

It is the square of Mean Squared error.

For Model2 the RMSE calculates is: 54.067



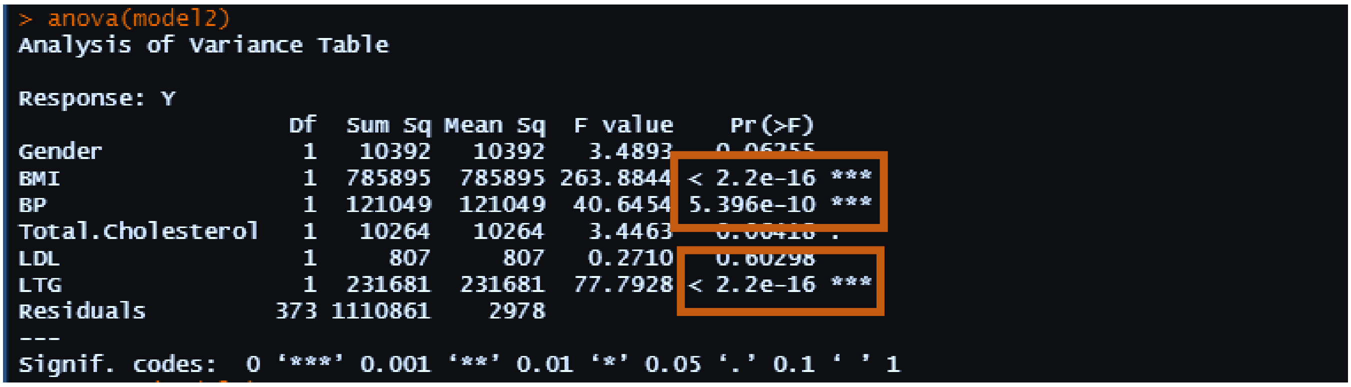
```
> pred = predict(model2)
> res= residuals(model2)
> mse = mean(res^2)
> rmse = sqrt(mse)
> mse
[1] 2923.319
> rmse
[1] 54.06772
```

Hypothesis Testing on Model-2 outcome:

Null Hypothesis, $\mathbf{H_0} : \boldsymbol{\beta 1 = \beta 2 = \cdots = \beta k{-}1 = 0}$

Alternate Hypothesis, $\mathbf{H_a} : \boldsymbol{\beta j \neq 0, for\ atleast\ one\ j.}$

ANOVA Output:



```
> anova(model2)
Analysis of Variance Table

Response: Y
                   Df  Sum Sq Mean Sq  F value    Pr(>F)    
Gender              1   10392   10392   3.4893   0.06255 .  
BMI                 1  785895  785895 263.8844 < 2.2e-16 ***
BP                  1  121049  121049  40.6454 5.396e-10 ***
Total.Cholesterol   1   10264   10264   3.4463   0.06418 .  
LDL                 1     807     807   0.2710   0.60298    
LTG                 1  231681  231681  77.7928 < 2.2e-16 ***
Residuals         373 1110861    2978                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the above results show there exists p values that are significant. So, we can reject the NULL hypothesis. So, the model can be used for prediction.

Prediction for Test data

The prediction for test data was conducted and the MSE and RMSE resulted were 2691.952 and 51.88402 respectively.

CONCLUSION

The model 2 can be considered for prediction. However, disease progression has various external factors as it depends mainly on the lifestyle of different individuals. Therefore, any model that is made for diabetes progression will not be highly accurate. But for the model 2 I have created when the predictions were done for test data, the RMSE was low than the RMSE for training data which shows that the model I not overfitted for the training data. As the model had a p value of 0.00265, we can say that it is highly significant.

ACKNOWLEDGEMENT

I would like to extend my gratitude to our Professor Dr Suchismita Das for her valuable guidance. I would also like to thank SP Jain School of Global Management, Mumbai for this wonderful opportunity.

REFERENCES

- Efron, B., Hastie, T., Johnstone, J., and Tibshirani, R. (2004). Least Angle Regression. Annals of Statistics (with discussion), 32, 407-499. Permission to distribute with JMP has been granted.
- https://www.tutorialspoint.com/r/r_boxplots.htm
- https://www.scribbr.com/statistics/multiple-linear-regression/
- https://www.geeksforgeeks.org/residual-leverage-plot-regression-diagnostic/
- https://www.scribbr.com/statistics/akaike-information-criterion/
- https://www.investopedia.com/terms/v/variance-inflation-factor.asp

Appendix

R-Script Submitted along with the report