EE3211 Modelling Techniques Report

PATRA Yuvraj (SID: 55907774)

TOPIC 1

Background

Diabetes mellitus (DM), commonly known as diabetes, is a group of metabolic disorders characterized by a high blood sugar level over a prolonged period. We are trying to find a relation between diabetes and a person's physical characteristics such as blood pressure, blood cholesterol and alcohol drinking habits.

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. NHANES is a major program of the National Centre for Health Statistics (NCHS). NCHS is part of the Centres for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.

Objective

The objective of this project is to identify a number of body-factors and life habits and to determine whether they have an association with diabetes. Data for this project will be collected from the NHANES website. The datasets used for this project are as follows:

Datasets: Diabetes Questionnaire Data, Body Measures Data, Blood Pressure & Cholesterol Questionnaire Data, Alcohol Use Questionnaire Data (2017-2018).

Methods

Data Pre-processing: The required data was extracted from the given data sets from the NHANES website and merged into a single matrix.

Data Cleansing: A data frame was created and used to omit the NA values from the extracted data. Binary classification was employed to turn the DIQ010(diabetes) attribute dichotomous.

Correlation: The statistical method of Pearson's correlation is used for analysing whether there is a correlation between the attributes Body Mass Index (BMI) and Height.

Logistic Regression (Simple and Multiple): The statistical method of multiple logistic regression is employed to determine the body-factors that are associated with diabetes. It is used to examine the association between the independent variables (the body-factors) with the dichotomous dependent variable (diabetes (DIQ010)).

Observations and Graph Plots: Boxplot and is used to represent the Body Weight Distribution of the people who have diabetes, and a Histogram is plotted for the Body Mass index (BMI).

Results

1. Correlation between BMI and Height

```
        Pearson's product-moment correlation

data:  data_final$BMXBMI and data_final$BMXHT
t = -0.87763, df = 482, p-value = 0.3806
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.12861462  0.04936227
sample estimates:
        cor
-0.03994298
```

The Pearson's correlation between BMI and Height was obtained as -0.0399428 (negative correlation) and a p-value of 0.3806 was obtained which means that it is statistically insignificant since this value is greater than the level of significance ($\alpha = 0.05$).
The formula for BMI is BMI = kg/m2 where kg is a person's weight in kilograms and m^2 is their height in metres squared. Since the correlation between the BMI and height attributes of the dataset is negative it is consistent with the formula since BMI of the diabetic person is inversely proportional to the square of the height of the diabetic person.

2. Multiple Logistic Regression (Height, Waist Circumference and Blood Pressure)

H_0: All features are not related to diabetes
H_1: At least one of the features are related to diabetes.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.03069    3.65763  -0.282  0.77810
BMXHT       -0.01997    0.02045  -0.976  0.32897
BMXWAIST     0.03434    0.01150   2.986  0.00283 **
BPQ020      -1.11467    0.38570  -2.890  0.00385 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Since the p-values obtained from the multiple logistic regression model are less than the significance level ($\alpha = 0.05$) for features BMXWAIST and BPQ020, they are statistically significant.
- Since the p-value obtained for BMXHT is greater than the significance level ($\alpha = 0.05$), it is statistically insignificant
- Null Hypothesis is rejected
- A one-unit increase in the Height is associated with the decrease in the log-odds of being diabetic by 0.01997.
- A one-unit increase in the Waist Circumference is associated with the increase in the log-odds of being diabetic by 0.03434.
- A one-unit increase in the Blood Pressure is associated with the decrease in the log-odds of being diabetic by 1.11467.

3. Simple Logistic Regression (feature: Overweight)

H_0: Attribute Overweight is not related to diabetes.
H_1: Attribute Overweight is related to diabetes.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3032     0.3220 -10.259  < 2e-16 ***
Overweight    1.2013     0.3933   3.054  0.00226 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Since the p-value obtained from the logistic regression model (0.00226) is less than the significance level ($\alpha = 0.05$), the attribute is Overweight is statistically significant.
- Null Hypothesis is rejected. Attribute Overweight is related to diabetes.

4. Simple Logistic Regression (feature: Cholesterol Level)

H_0: Attribute Cholesterol Level is not related to diabetes.
H_1: Attribute Cholesterol Level is related to diabetes.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5658     0.6258  -0.904  0.36591
BPQ080       -1.2260     0.3763  -3.258  0.00112 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Since the p-value obtained from the logistic regression model (0.00112) is less than the significance level ($\alpha = 0.05$), the attribute Cholesterol is statistically significant.
- Null Hypothesis is rejected. Attribute Cholesterol is related to diabetes.

5. Simple Logistic Regression (feature: Alcohol Intake)

a. ALQ130 (Avg # alcohol drinks/day - past 12 mos)

H_0: Attribute Alcohol Intake (ALQ130) is not related to diabetes.
H_1: Attribute Alcohol Intake (ALQ130) is related to diabetes.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.81153    0.35058  -8.020 1.06e-15 ***
ALQ130       0.02935    0.05224   0.562    0.574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Since the p-value obtained from the logistic regression model (0.574) is greater than the significance level ($\alpha = 0.05$), the attribute Alcohol Intake (ALQ130) is statistically insignificant.
- We fail to reject the Null Hypothesis. Attribute Alcohol Intake (ALQ130) is not related to diabetes.

b. ALQ290 (# times 8+ drinks in 1 day/past 12 mos)

H_0: Attribute Alcohol Intake (ALQ290) is not related to diabetes.
H_1: Attribute Alcohol Intake (ALQ290) is related to diabetes.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.83259    0.26110 -10.849   <2e-16 ***
ALQ290       0.04663    0.04311   1.082    0.279
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Since the p-value obtained from the logistic regression model (0.279) is greater than the significance level ($\alpha = 0.05$), the attribute Alcohol Intake (ALQ290) is statistically insignificant.
- We fail to reject the Null Hypothesis. Attribute Alcohol Intake (ALQ290) is not related to diabetes.

6. Simple Logistic Regression (feature: Body Mass Index (BMI))

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.01404    0.82750  -6.059 1.37e-09 ***
BMXBMI       0.07665    0.02491   3.077  0.00209 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Since the p-value obtained from the logistic regression model (0.00209) is less than the significance level ($\alpha = 0.05$), the attribute BMI is statistically insignificant.
- Null Hypothesis is rejected. Attribute BMI is related to diabetes.

7. Observations from Graphs and Figures

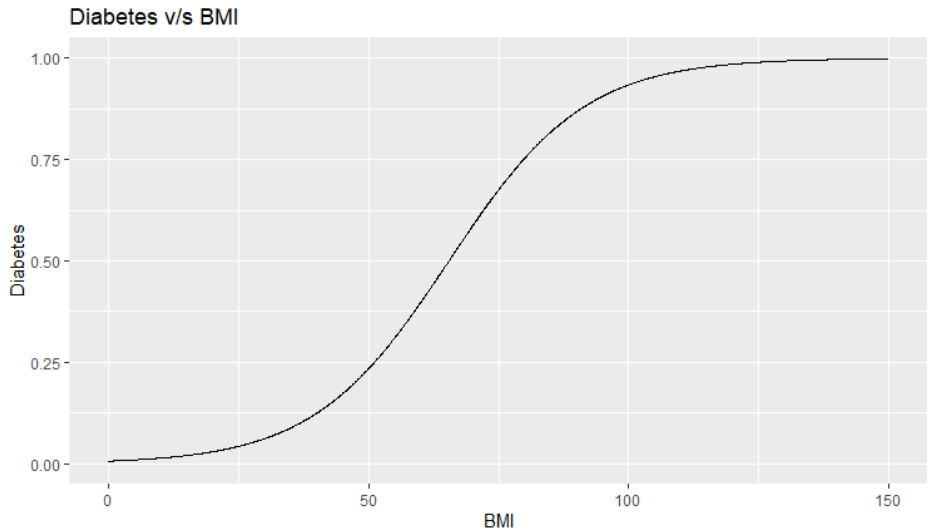

Fig 1: Logistic Regression Graph Diabetes v/s BMI

BMI is related to Diabetes and the figure above is the logistic regression graph between the dichotomous dependent variable Diabetes and the independent variable BMI.

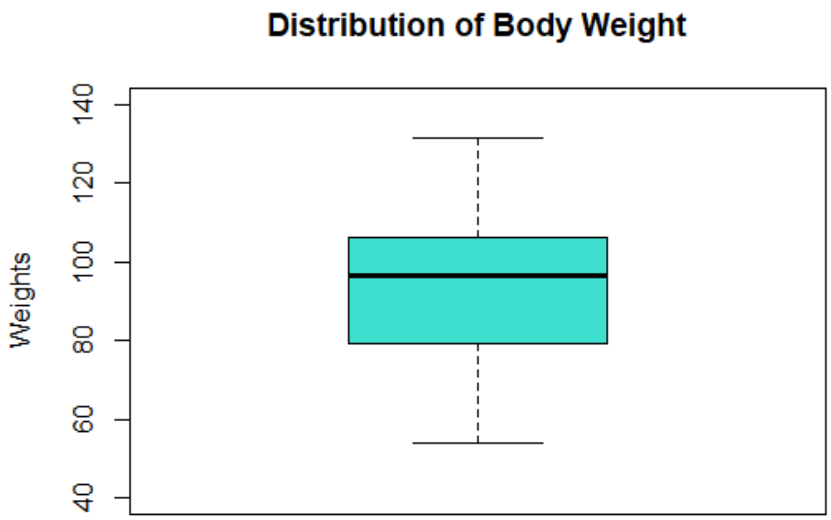

Fig 2: Boxplot Distribution of Bodyweight

Fig 2 shows the distribution of bodyweight in the form of a boxplot. From the boxplot we can observe the following about the bodyweight of persons who are diabetic:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 53.90   79.42   96.40   97.40  106.12  191.40
```

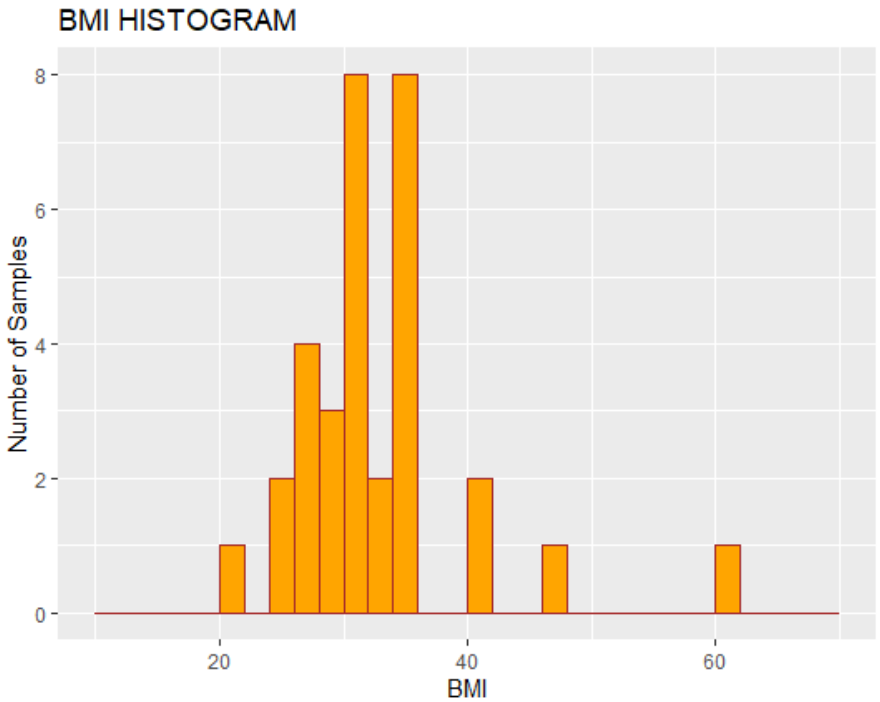

Fig 3: Histogram for Body Mass Index (BMI)

Fig 3 is a histogram of Body Mass Index of persons who have diabetes. It shows number of samples with BMI between 20-22, 22-24, 24-26 and so on.

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 20.10   29.18   31.65   32.87   34.80   61.90
```

Conclusion

Using R programming in R Studio software, statistical tests were conducted on datasets obtained from NHANES to determine which body-features and life habits have an association with Diabetes mellitus (DM). From the results obtained, it can be concluded that BMI and Height have a negative correlation (based on Karl Pearson's correlation coefficient (r)) which is consistent with the mathematical formula of BMI. Multiple Logistic Regression method was used to determine that Waist Circumference and Blood Pressure are associated with Diabetes while Height of the person is not associated with it. Simple Logistic Regression method was employed for each of the attributes Overweight, Cholesterol and Alcohol Intake individually as the independent variable in the model. Based on these models, it was inferred that the attributes Overweight and Cholesterol are associated with Diabetes, however Alcohol Intake of the individual has no association with Diabetes. Logistic Regression model with BMI as the independent variable helped us conclude that BMI is associated with Diabetes. Furthermore, the logistic regression graph plot was useful in observing how the predicted probabilities changed as the independent variable BMI was varied. The box and whisker plot distribution helped visualize the measures of central tendency as well as the quartiles of bodyweight of individuals who have diabetes. The histogram graph plot was useful in visually representing the frequency distribution of Body Mass Indices of diabetic people, showing the number of datapoints falling withing specified range of values.

In conclusion, Diabetes mellitus (DM) can be prevented by reducing the risk of having a high cholesterol level, high blood pressure and being overweight. This can be achieved with a balanced diet and following a healthy lifestyle.