City University of Hong Kong

2020 – 2021 Sem B

EE3211 Modelling Techniques

**Project Report**

**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. The aim of this task is to assess

the relationship between diabetics’ data and data on physical characteristics, blood pressure and

drinking habits.

**Objective:**

The objection of the project is to finding out the which risk factors are more related to having diabetes and their significance of casing having diabetes. In the following, logistic regression model is built to obverse and predict the relations between the factors and diabetes.

The selected data are the following: “Doctor told you have diabetes”, Weight, Height, Waist Circumference(cm), Hip Circumference(cm), and BMI, “Ever told you had high blood pressure”, “Doctor told you ‐ high cholesterol level”, “Had at least 12 alcohol drinks/1 year”, and “Average # alcohol drinks/day ‐ past 12 month”.

**Method:**

In problem 1, data will be selected from Diabetes Questionnaire Data, Body Measures Data, Blood Pressure & Cholesterol Questionnaire Data, Alcohol Use Questionnaire Data and will be merge into one data set. Summary is used to show the minimum, median, maximum and the number of NA value etc.

In problem 2, data with useless information like answer as “Refused”, “Don’t know” or NA value will be dropped. An overweight column will be added with BMI larger than 30 as will mark as 1 and vice versa. Correlation and Regression model will be built in R to measure the relativity of the factors.

In problem 3, logistic regression model and multiple logistic regression model will be used to find the relations between the risk factor and diabetes. A regression graph, Boxplot, distribution graph and histogram will be plotted to have a better vision to the situation.

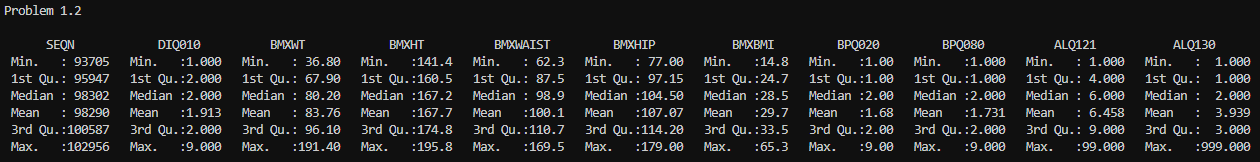
**Data:**

There are 8897 data imported from DIQ\_J.XPT, 8218 data imported from BMX\_J.XPT, 6161 data imported from BPQ\_J.XPT, 5533 data imported from ALQ\_J.XPT.

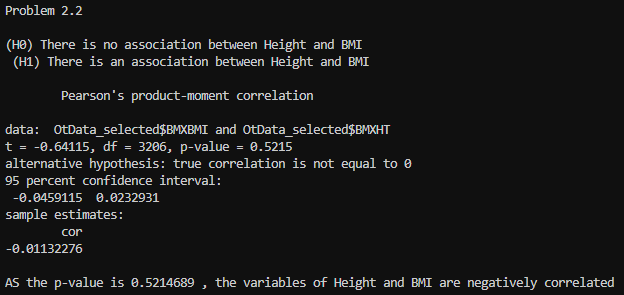
After merging and dropping unrelated data, there are 3208 data in the data set.

**Result:**

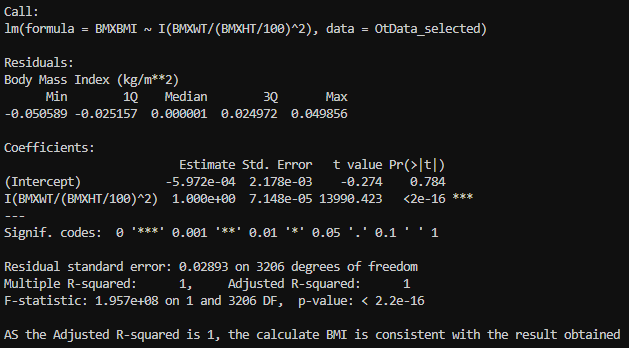
**In problem 1.2, the outputs are shown below:**



**In problem 2.2, the outputs are shown below:**



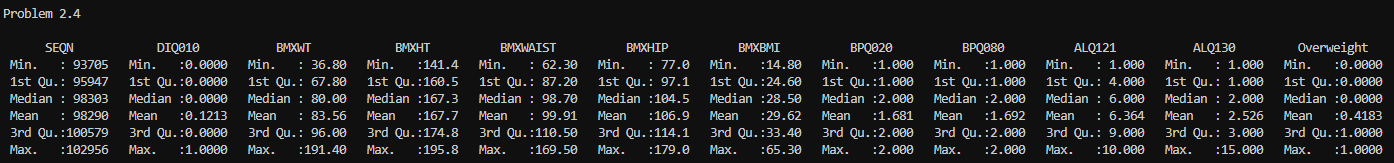
As the correlation is -0.01132276 which is smaller than 0, the height and BMI are negatively correlated. Thus, when height increase, BMI will decrease.



By calculating BMI manually with the following equation:

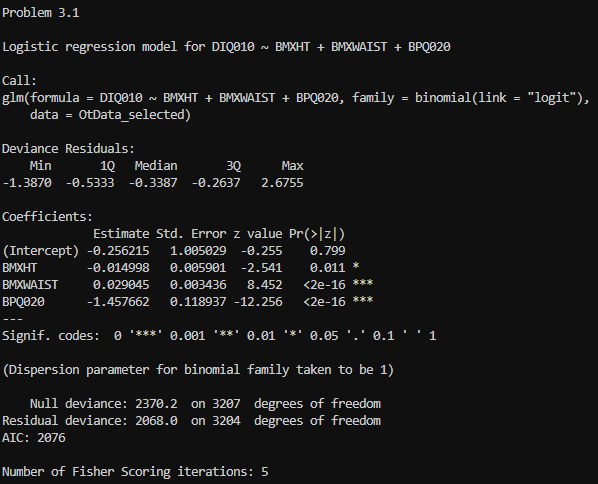
Then compare this value with the BMI given in the dataset, the Adjusted R-squared is found out to be 1 and the p-value is less than 0.05, the calculated BMI is consistent with the result obtained.

**In problem 2.4, the outputs are shown below:**



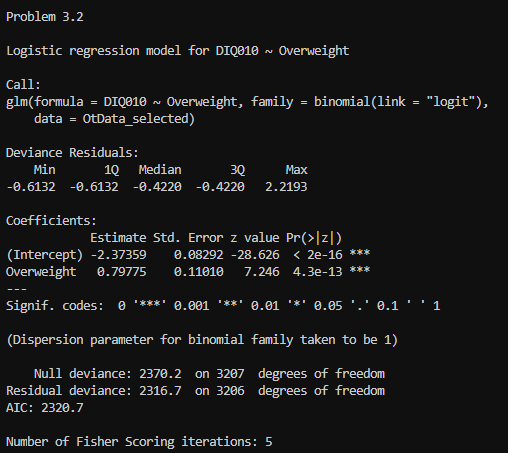
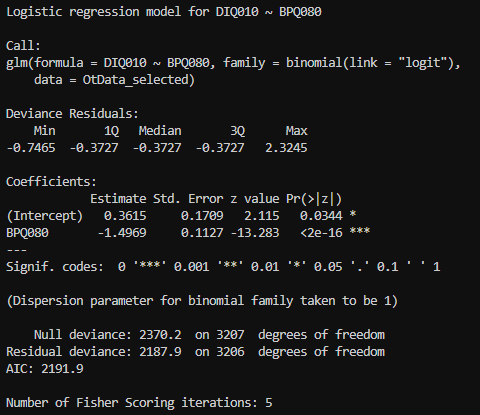
As uncertainty, refusal answer such as “Refused to answer” or “Don’t know”, and outlined answer are removed from the observation dataset, the value in the summary such as minimum, median or maximum will changed and updated.

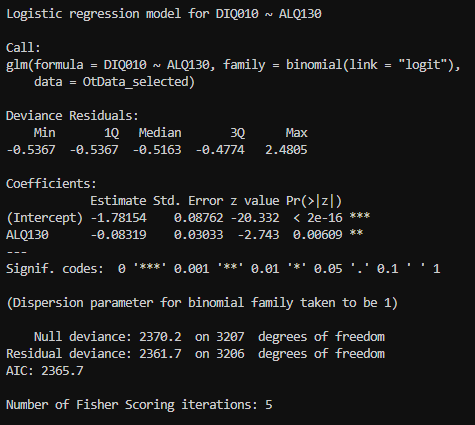
**In problem 3.1, the outputs are shown below:**



From the output, it shows that the height, waist circumference, blood pressure, they were all related to having diabetes with p-value 0.011, smaller than 2e-16 and smaller than 2e-16 respectively which all small than 0.05 after controlled all other risk factors in the dataset.

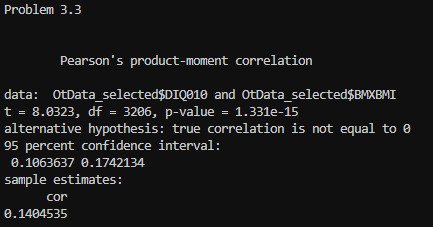
**In problem 3.2, the outputs are shown below:**

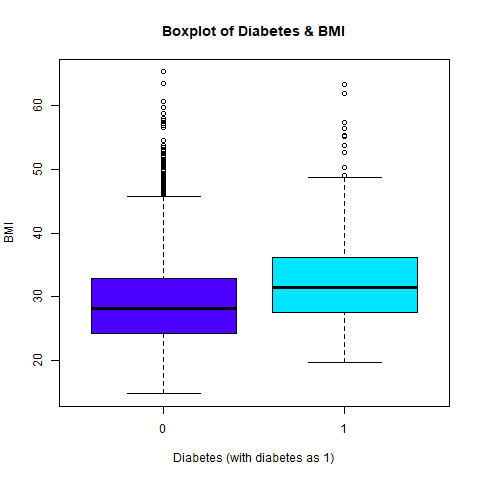
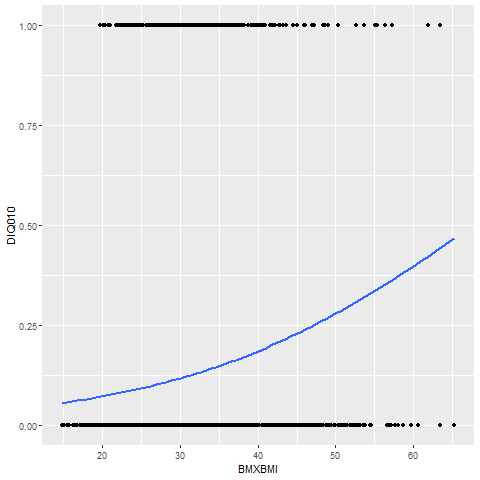
 ****



From the output, it shows that overweight, is related to having diabetes with p-value 4.3e-13, which is smaller than 0.05 after controlled all other risk factors in the dataset. For cholesterol level, it is related to having diabetes with p-value smaller than 2e-16, which is smaller than 0.05 after controlled all other risk factors in the dataset. For alcohol taking, it is related to having diabetes with p-value 0.00609, which is smaller than 0.05 after controlled all other risk factors in the dataset.

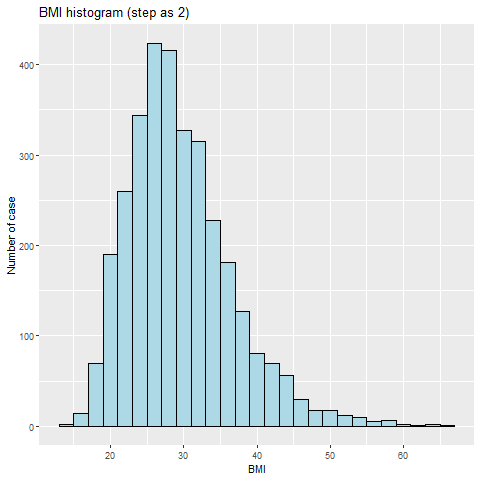
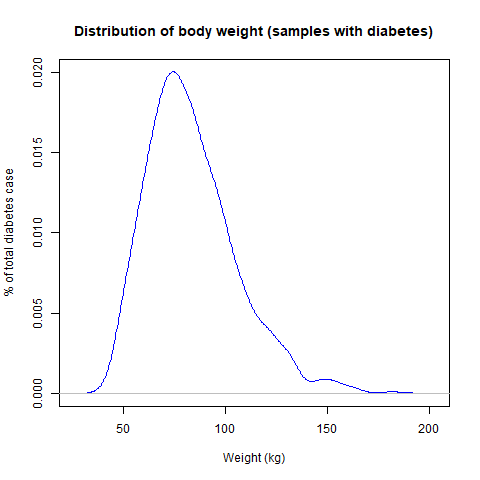
**In problem 3.3, the outputs are shown below:**





From above output and plots, they showed that, there is am relation between BMI and having diabetes. From the left graph, it shows that, then BMI increase, the chance of having diabetes will increase as they are positively related. From the right graph, it shows that higher BMI have more cases of having diabetes than lower BMI.

**In problem 3.4, the outputs are shown below:**



From the distribution of body weight plotting, it shows that, the distribution is positively skewed, the mass od the distribution is on the left of the diagram. From the histogram for BMI value, it shows that, most of the people who have diabetes has a BMI around 28.

**Conclusion:**

In conclude, there are significant relation between diabetes and blood pressure, BMI value, cholesterol level, and alcohol taking per day in the past 12 months. People have high blood pressure, high cholesterol level, BMI value or heavy alcohol taking are tended to have higher chance of having diabetes.

**References:**

U.S. Department of Health & Human Services. (2020, December). 2017-2018 Data Documentation, Codebook, and Frequencies Alcohol Use (ALQ\_J). Centers for Disease Control and Prevention. <https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/ALQ_J.htm#ALQ130>

U.S. Department of Health & Human Services. (2020, February). National Health and Nutrition Examination Survey 2017–2018 Data Documentation, Codebook, and Frequencies Blood Pressure & Cholesterol (BPQ\_J). Centers for Disease Control and Prevention. <https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BPQ_J.htm#BPQ080>

U.S. Department of Health & Human Services. (2020, February). 2017-2018 Data Documentation, Codebook, and Frequencies Body Measures (BMX\_J). Centers for Disease Control and Prevention. <https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BMX_J.htm#BMXWT>

U.S. Department of Health & Human Services. (2020, February). 2017-2018 Data Documentation, Codebook, and Frequencies Diabetes (DIQ\_J). Centers for Disease Control and Prevention. <https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DIQ_J.htm#DIQ010>

**Remarks:**

ALQ121 is used in the Alcohol Use Questionnaire Data set.