Name:

SID:

Submission Date: 24/04/2020

Hypothesis:

Generally, this exercise is to test for whether the significant relationship between diabetic's data and data on physical characteristics, blood pressure, and drinking habits. I have been settled an assumption that those factors are pertinent to diabetes.

In problem 2, I tested for the significance of the regression line derived for the BMXBMI- BMXHT data to determine a correlation between height and BMI. The null hypothesis (H0) is that there is no association between them. The alternative hypothesis (H1) is that there is an association between them.

In problem 3, I assessed the statistical significance of the individual risk factors to find the relation between diabetes and height, overweight, blood pressure, cholesterol level, and alcohol taking. Under the null hypothesis that a specific factor does not affect diabetes after controlling for all other risk factors. The alternative hypothesis (H1) is that a specific factor would be a risk factor.

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. Symptoms usually include increased thirst and increased hunger. Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, damage to the eyes, and cognitive impairment.

The National Health and Nutrition Examination Survey organization conducted the questionnaire data and body measures data by focusing on the adults and children in the United States. The survey is valuable and unique which combines interviews and physical examination. It is beneficial and accurate for us to assess their health status.

In Diabetes questionnaire data, DIQ010 is used, the respondents were asked: “Have you ever been told by a doctor or health professional that you have diabetes?”. It has been treated as a case-control study: the diabetes group and non-diabetes group are initially identified. Also, the body measures data, blood pressure & cholesterol data and alcohol data are adopted as the factors, multiple risk factors may be involved, several risk factors must be controlled for simultaneously in analyzing variables associated with diabetes. I have been merged all these data into one matrix, the data set Matrix contains for the following variables: BMXHT-Standing Height (cm), BMXBMI-Body Mass Index(kg/m^2), BMXWT-Weight(kg), BPQ020-”Ever told you had high blood pressure”, BPQ080-”Doctor told you-high cholesterol level”, DIQ010-"Doctor told you to have diabetes”, ALQ101-”Had at least 12 alcohol drinks/1yr”, and ALQ130-“alcoholic drinks/day-past 12 mos.”

Objective:

Since diabetes would frequently lead to a variety of complications and symptoms which is detrimental to our health. It raises a question about what dicey factors may cause these health issues. Therefore, this exercise aims to assess the relationship or potential determinants between diabetes and those factors.

Methods:

In problem 1.2, I would like to use the descriptive statistics in R, the data set would be described numerically in terms of a measure of location and a measure of spread. (arithmetic mean, median, quantiles, etc.)

In problem 2.2, regression, and correlation method are chosen in R, I would use a correlation coefficient to test for the variables that are positively correlated or negatively correlated or uncorrelated. Also, I would like to use the regression line for the BMI, weight, and height data.

In problem 3.1, multiple logistic regression is adopted in R for finding the relation between diabetes and factors the relation between diabetes and BMI.

In problem 3.2, the sample with diabetes is picked out to form the distribution of body weight for testing the relationship between body weight and diabetes. Also, the data sets of BMI are categorized into groups that differ in 2 units. The plot of a histogram is used to visualize the difference in high BMI and low BMI.

Results:

In problem 1.2, the outputs are shown below:

```
> summary(Matrix)
      SEQN           DIQ010          BMXWT           BMXHT           BMXBMI          BPQ020          BPQ080          ALQ101          ALQ130      
 Min.   :83732   Min.   :1.000   Min.   : 32.40   Min.   :129.7   Min.   :14.50   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :  1.000  
 1st Qu.:86164   1st Qu.:2.000   1st Qu.: 65.90   1st Qu.:158.7   1st Qu.:24.30   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:  1.000  
 Median :88668   Median :2.000   Median : 78.20   Median :166.0   Median :28.30   Median :2.000   Median :2.000   Median :1.000   Median :  2.000  
 Mean   :88679   Mean   :1.886   Mean   : 81.34   Mean   :166.1   Mean   :29.38   Mean   :1.657   Mean   :1.723   Mean   :1.336   Mean   :  3.912  
 3rd Qu.:91179   3rd Qu.:2.000   3rd Qu.: 92.70   3rd Qu.:173.5   3rd Qu.:33.00   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:  3.000  
 Max.   :93702   Max.   :9.000   Max.   :198.90   Max.   :202.7   Max.   :67.30   Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :999.000  
                                 NA's   :69       NA's   :62      NA's   :73                                      NA's   :527     NA's   :2356     
```

In problem 2.2, the outputs are shown below:

```
> model = cor.test(Matrix$BMXBMI,Matrix$BMXHT)
> model

        Pearson's product-moment correlation

data:  Matrix$BMXBMI and Matrix$BMXHT
t = -1.783, df = 3258, p-value = 0.07468
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.065481550  0.003110763
sample estimates:
        cor
-0.03122215
```

Since the correlation = -0.03122215 (<0), so the variables of height and BMI are negatively correlated, BMI decreases as height increases.

```
> model = lm(Matrix$BMXBMI~Matrix$BMXWT/(Matrix$BMXHT^2))
> summary(model)

Call:
lm(formula = Matrix$BMXBMI ~ Matrix$BMXWT/(Matrix$BMXHT^2))

Residuals:
Body Mass Index (kg/m**2)
    Min      1Q  Median      3Q     Max
-0.5728 -0.2349 -0.1249  0.1104  3.5732

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 0.0189313  0.0281331    0.673    0.501
Matrix$BMXWT                1.0532734  0.0016242  648.480   <2e-16 ***
Matrix$BMXWT:Matrix$BMXHT -0.4141619  0.0008652 -478.716   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3756 on 3257 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9972
F-statistic: 5.738e+05 on 2 and 3257 DF,  p-value: < 2.2e-16
```

By formula of BMI, BMI = weight / height^2 (kg/m^2), we could have a prediction that weight is positively correlated with BMI, and height is negatively correlated with BMI. From the actual output, the coefficient of height = -0.4141619 (<0), and the p-value is less than 0.05, so we can conclude that there is a significant association between height and BMI.

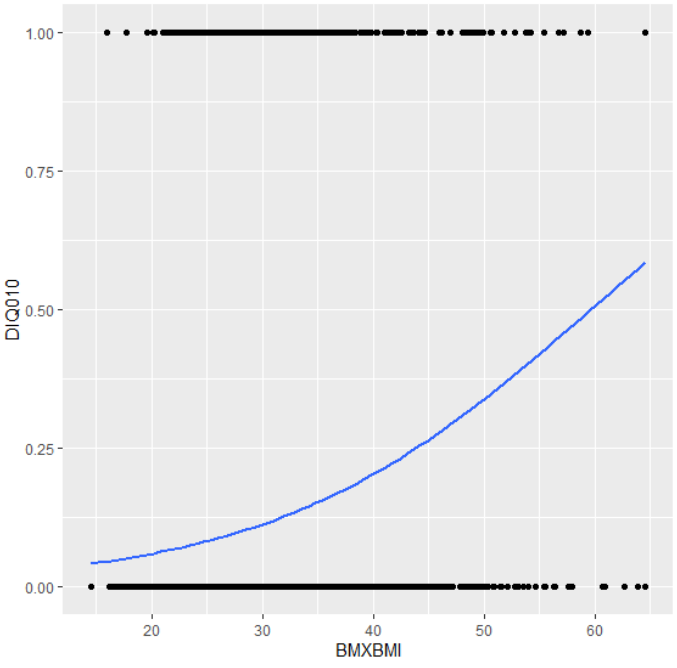
In problem 3.1, the outputs are shown below:

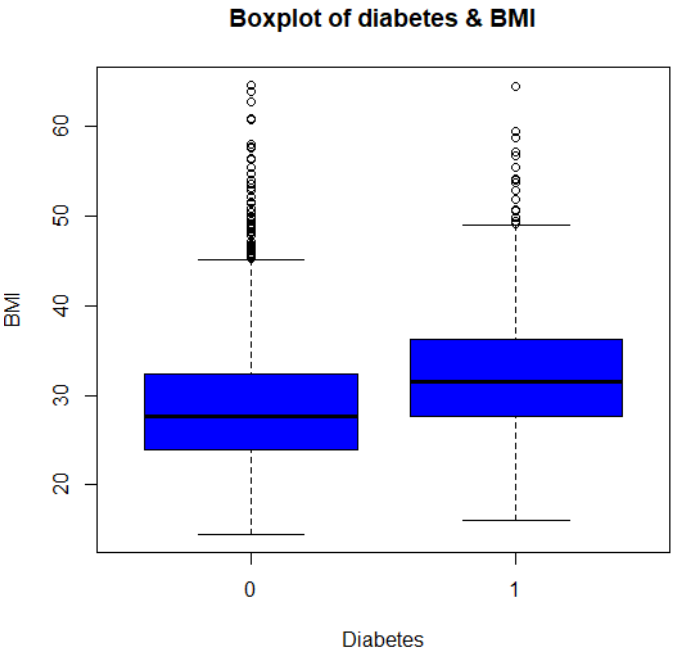
```
Call:
glm(formula = DIQ010 ~ BMXHT + overweight + BPQ020 + BPQ080 +
    ALQ101 + ALQ130, family = binomial(link = logit), data = Matrix)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0991  -0.4743  -0.3127  -0.2403   2.7588

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.82048    1.07983   1.686   0.0918 .
BMXHT       -0.07778    0.60955  -0.128   0.8985
overweight1  0.72228    0.11894   6.073 1.26e-09 ***
BPQ020      -1.35064    0.12625 -10.698  < 2e-16 ***
BPQ080      -1.13384    0.12339  -9.189  < 2e-16 ***
ALQ101      -0.06890    0.13912  -0.495   0.6204
ALQ130      -0.06079    0.02838  -2.142   0.0322 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2355.4  on 3259  degrees of freedom
Residual deviance: 1969.2  on 3253  degrees of freedom
AIC: 1983.2

Number of Fisher Scoring iterations: 7
```

By the outputs' result, we note that overweight, high blood pressure, high cholesterol level, and over-drink were significantly related (p-value<0.05) to the incidence of diabetes after simultaneously controlling for the effects of all other risk factors in the model. Interestingly, height and low level of drinking alcohol had no significant effect on the incidence of diabetes after controlling for the other risk factors.

```
Call:
glm(formula = DIQ010 ~ BMXBMI, family = binomial(link = logit),
    data = Matrix)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3266  -0.5131  -0.4366  -0.3782   2.4849

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.154895   0.226233  -18.37   <2e-16 ***
BMXBMI       0.069645   0.006856   10.16   <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: 2355.4  on 3259  degrees of freedom
Residual deviance: 2256.0  on 3258  degrees of freedom
AIC: 2260

Number of Fisher Scoring iterations: 5
```

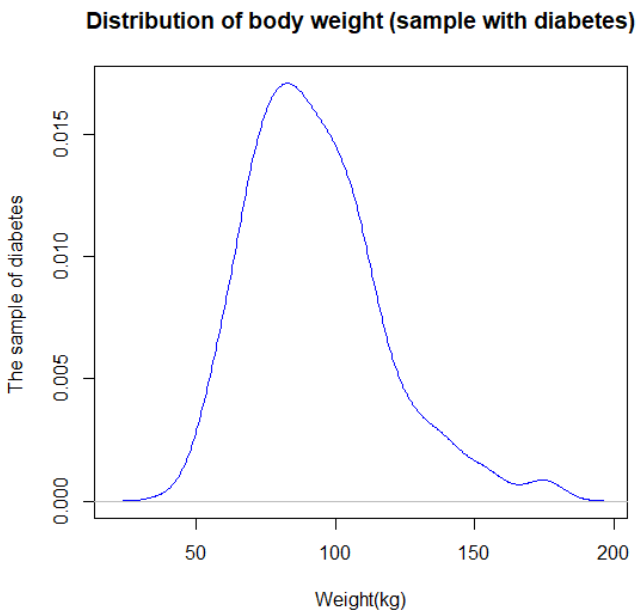
By the output's result, we note that BMI was significantly related (p-value<0.05), to the incidence of diabetes after simultaneously controlling for the effects of all other risk factors in the model which implied that low index of BMI had a low chance of suffering from diabetes.

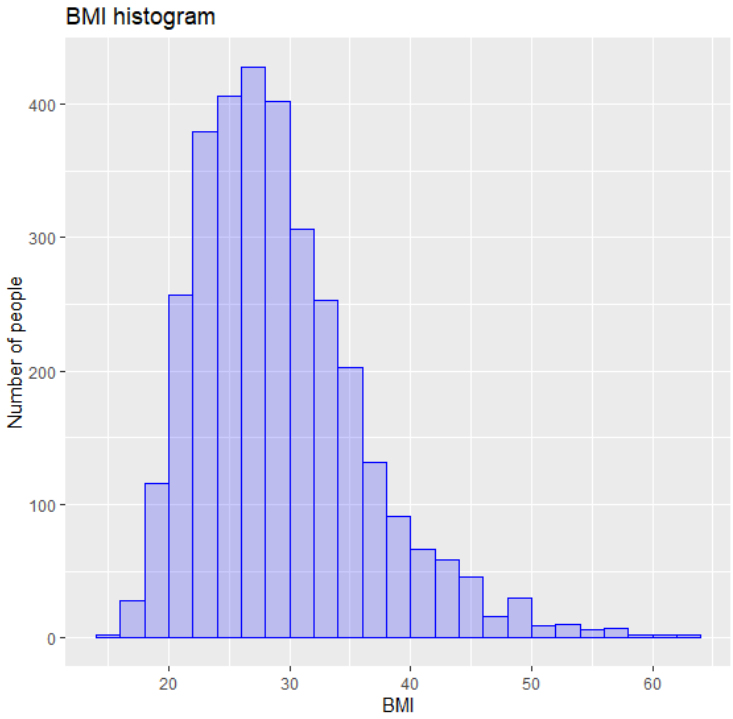




In both diagrams, we could get the relation between BMI and diabetes. On the left diagram, BMI increases as the incidence of diabetes increases (positively related), vice versa. On the right diagram, a non-diabetes person had low BMI (mean <30), on the contrary, diabetes person had a high BMI (mean>30). It had shown that diabetes had a positive correlation with BMI.

In problem 3.2, the outputs are shown below:





In both distributions, we found that they were positively skewed unimodal distribution, the mass of the distribution is concentrated on the left of the figure. Also, it implied that the mean being skewed to the right of a typical center of the data, the person had diabetes which tended to have high weight and the average BMI tended to be larger generally.

Conclusion:

To conclude, there are significant relations between diabetes and BMI, blood pressure, cholesterol level, and alcoholic drinks per day, in which the potential determinants are blood pressure and cholesterol. People with high blood pressure, high cholesterol level, and larger weight have a higher probability that he or she may experience diabetes. However, the results found that there is an insignificant association between diabetes and the low level of alcoholic drinks per year.