EE3211 Exercise Topic 1

Due: 23:59, April 23, 2021

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.

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 Center for Health Statistics (NCHS).

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

Raw data & Data dictionary:
https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Questionnaire&CycleBeginYear=2017 (Questionnaire Datasets)
https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Examination&CycleBeginYear=2017 (Examination Data)

Problem1:

1.1 Load **Diabetes Questionnaire Data**, extract sample ID and "Doctor told you have diabetes"; load **Body Measures Data**, extract sample ID, Weight, Height, Waist Circumference(cm), Hip Circumference(cm), and BMI; load **Blood Pressure & Cholesterol Questionnaire Data**, extract sample ID, "Ever told you had high blood pressure", "Doctor told you - high cholesterol level"; load **Alcohol Use Questionnaire Data**, extract sample ID, "Had at least 12 alcohol drinks/1 yr", and "Avg # alcohol drinks/day - past 12 mos".

1.2 Merge all these data into one matrix and omit all the missing values (NA) from this matrix. Use the "summary()" function to print the results.

Problem2:

2.1 Adjust the information in the matrix. Record those with diabetes as 1, those without diabetes as 0, those with uncertainty or refusal to answer as NA and exclude. Do a similar deletion of the rest of the data sets (remove entries such as 'Refused', 'Don't know').

2.2 Add a column of "overweight" to the matrix, recording a BMI greater than 30 as overweight (marked 1) and the rest as 0. Please analyze whether there is a correlation between height and BMI. Please search the formula of BMI by yourself, and find if it is consistent with the result obtained.

2.3 Use "factor()" to set the classification variable type to factor, and use the "class()" function to check the type of each data in the table.

2.4 Use the "summary()" function again, and compare the result with the last summary().

Problem3:

3.1 Use logistic regression model, find the relation between (dependent variable: diabetes) vs. (independent variable: height + waist circumference + blood pressure), put these independent variables together in the regression model.

3.2 Use logistic regression model, find the relation between (dependent variable: diabetes) vs. (independent variable: overweight, cholesterol level, and alcohol taking), test each independent variable individually.

3.3 Find the relation between (dependent variable: diabetes) vs. (independent variable: BMI value), plot the result.

3.4 Select the samples with diabetes, plot the distribution of body weight, and plot the histogram for BMI value (number of samples with BMI between 20-22, 22-24, 24-26 …). Please add coordinate axis information, title and other necessary components for all the figures.

3.5 State the conclusions obtained from the figures in the report.

Hint:
load data: library("SASxport"), read.xport()
merge data: library("dplyr"), merge()
plot: library("ggplot2"), boxplot(), ggplot()