EE3211 Exercise Topic 2

Due: 23:59, April 23, 2021

Blood pressure is related to heart disease, kidney disease and other diseases, so it is an important risk factor for metabolic diseases. The goal of this exercise is to find the relation between demographics data, body characteristic data and blood pressure.

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: Blood Pressure, Body Measures, Demographic Variables and Sample Weights (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=Demographics&CycleBeginYear=2017 (Demographic Data)

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

Problem1:

1.1 Load **Demographic Variables and Sample Weights data**, extract Sample ID, Gender, Age, Race/Hispanic origin w/ NH Asian, and create a new matrix for them. Use "factor()" to set the classification variable type to factor, and use the "class()" function to check the type of each data in the matrix.

1.2 Please find if there is any missing data in this matrix.

1.3 Please analyze whether there is a correlation between Age and Gender.

Problem2:

2.1 Load **Body Measures data**, extract sample ID, Weight, Height, BMI, Waist circumference and Hip circumference. Merge this dataset with the matrix you got from Problem 1 (based on sample ID).

2.2 Omit all the missing values (NA) from this new matrix.

2.3 Please find the relation between (dependent variable: BMI) and (independent variables: sample ID + Age + Gender + Race + Height + Waist circumference), put these independent variables together in the regression model, and summarize your result.

Problem3:

3.1 Load **Blood Pressure data**, extract sample ID, Pulse (60 sec.), Pulse (regular/irregular), Maximum inflation levels, Systolic BP (1, 2, 3 times testing), Diastolic BP (1, 2, 3 times testing). Merge this dataset with the matrix you got from Problem 2 (based on sample ID). Omit all the missing values (NA) from this new matrix.

3.2 Make new columns for: mean Systolic BP, mean Diastolic BP (both as integer). Summarize the result.

3.3 Make new column for: BP score. The blood pressure levels were divided into four categories:
Score = 0: normal blood pressure (≤120/80 mmHg);
Score = 1: prehypertension (120-139/80-89 mmHg);
Score = 2: stage 1 hypertension (140-159/90-99 mmHg);

Score = 3: stage 2 hypertension (≥160/100 mmHg).
Plot the box plot of BMI value for these 4 group.

3.4 Please calculate the odds ratio for Age, BMI, Waist circumference, Hip circumference, and Pulse.

3.5 Is there any relationship between these variables and the occurrence of hypertension?

Hint:
load data: library("SASxport"), read.xport()
merge data: library("dplyr"), merge()
box chart: data.frame(group = , value =), plot.data <- rbind(),
ggplot(plot.data, aes() + geom_boxplot()