EE3211 Exercise Topic 3

Due: 23:59, April 23, 2021

Blood pressure, cholesterol levels, height, weight and BMI are frequently tested in physical examinations. Smoking status, drinking habits are the potential risk factors leading to multiple diseases. The aim of this exercise is to find the relation between these medical parameters and lifestyle factors.
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 & Cholesterol Questionnaire Data, Alcohol Use Questionnaire Data, Body Measures, Demographic Variables and Sample Weights, Smoking - Cigarette Use, Medical Conditions, Current Health Status (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 Download all needed datasets ("**Demographic Variables**", "**Body Measures**", " **Blood Pressure & Cholesterol** ", "**Smoking - Cigarette Use**", "**Medical Conditions**", and "**Alcohol Use**"). Extract sample ID, gender, age, Race(Hispanic origin w/ NH Asian), weight, height, BMI, "ever told you had high blood pressure", "high cholesterol level", "Smoked at least 100 cigarettes in life", "Ever smoked a cigar even 1 time?", "Ever used an e-cigarette?", "Ever used smokeless tobacco?", "Ever been told you have asthma", "Doctor ever said you were overweight", "Doctor ever said you had arthritis", " Ever told you had a stroke", "Ever told you had chronic bronchitis", "Avg # alcohol drinks/day - past 12 mos".

1.2 For the yes / no question data, set yes as "1" and no as "0", remove other answers. Delete all rows of data with NA. Use the summary() function to print the results.

1.3 Download Current Health Status data ("**Current Health Status**"), extract sample ID, "General health condition", "SP have head cold or chest cold", "SP have stomach or intestinal illness?", and "SP have flu, pneumonia, ear infection?". Remove the NA (missing values), "refuse" and "don't know".

1.4 Plot the histogram for "General health condition" (by using the score 1-5 from data dictionary).

Problem2:

2.1 Merge two datasets you got in Problem 1.

2.2 Use logistic regression model to evaluate the relation between blood pressure and all smoking data.

2.3 Use logistic regression model to assess the relation between (dependent variable: cholesterol level) and (independent variables: BMI + Alcohol Use data), test the independent variables separately.

2.4 Use linear regression method the find the relation between (dependent variable: Alcohol Use data) and (independent variable: age).

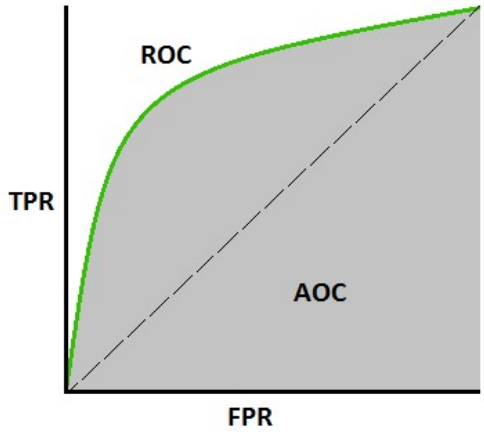
Problem 3:

3.1 Use ANOVA methods to test whether BMI levels is comparable for the five general health condition. Identify and test for any specific group differences (use Bonferroni for adjust p value).
3.2 Divide the "General health condition" into 2 classes: (class0: Value = 1,2,3; class1: Value = 4,5). Use logistic regression model to investigate the relation between (dependent variable: General health condition class) and (independent variables: group 1: Age, BMI; group 2: Alcohol use data), test the two groups of independent variables separately, and calculate the OR and CI.
3.3 Suppose that all samples in the data were re-tested for cholesterol, with 2000 class=0 samples, 1900 of them were also class=0 in the original test. Please use McNemar test method to calculate whether the detection rate of the two test methods is different.

Problem 4 (optional)

In Machine Learning, performance measurement is an essential task. So, when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi - class classification problem, we use AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model's performance.
AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. The ROC curve is plotted with true positive rate (TPR) against the false positive rate (FPR) where TPR is on y-axis and FPR is on the x-axis.

Here is an example of ROC-AUC curve:



Question: Use logistic regression model to assess the relation between the new score and all the remaining variablesand predict the score of health condition? (except ID). Draw the prediction table, the ROC curve and AUC line. Calculate the AUC value.

	0 (true)	1 (true)
0 (predict)	?	?
1 (predict)	?	?

Hint:
load data: library("SASxport"), read.xport()
merge data: library("dplyr"), merge()
plot: library("ggplot2")
ROC and AUC: library(pROC), roc(), predict (), table ()
install ("ggplot2"), set X axis into "1-specificities", Y axis into "sensitivities", AUC line is the line from (0, 0) to (1, 1).