In-Class Practice (HW) 9

Due: April 13, 2018

- Asking questions to TAs and collaborating with classmates are encouraged, but copying, sharing, or distributing any material is strictly prohibited. Homework should be students' original work.
- Please submit
  1) SAS code (.SAS) with detailed comments
  2) PDF document with relevant output and interpretations
- Late homework will not be accepted.

<u>**Coronary Heart Disease (CHD)**</u>

Dataset 'chd.xlsx' contains a subset of a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD (302 controls / 160 CHD cases). Many of the CHD positive mean have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. Following is the list of variables included in the dataset:

Variable	Description
ID	ID
SBP	Systolic blood pressure
Tobacco	Cumulative tobacco (kg)
LDL	Low density lipoprotein cholesterol
BAI	Body adiposity index
Famhist	Family history of heart disease (Present, Absent)
TypeA	Type-A personality
BMI	Body Mass Index
Alcohol	Current alcohol consumption
Age	Age at onset (year)
CHD	Coronary heart disease (1 = Case / 0 = Control)

The data have been taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal.

a) Import the dataset, name it 'CHD', and apply labels (SBP, LDL, BAI, Famhist, BMI, and CHD) and formats (CHD) <u>in DATA step</u>.

Print the first 5 observations of dataset with the labels and formats.

b) Descriptive statistics: Provide the following tables and plots and describe the distribution (e.g. missing values, symmetry, skewness, association between variables, location (mean, median), dispersion (range, standard deviation), outliers) of variables displayed in those tables and plots.

i. Cross-tabular frequency family history (rows) and CHD status (columns)

ii. Distribution of systolic pressure

1) Descriptive statistics (n, mean, median, standard deviation, min, max) of systolic blood pressure for each level of CHD status. Use two decimal points.

2) Boxplots of systolic pressure for each level of CHD status

3) Scatterplot and Pearson's correlation coefficient of systolic blood pressure and tobacco consumption for each level of CHD status

iii. Histograms of body adiposity index for those with and without family history, separately. Overlay each histogram with normal density curve. (Hint: Panel)

c) Macro: Create a macro program named 'table' that takes two numeric variables as inputs and produces a table with CHD status and family history. Following is an example with numeric variables 'alcohol' and 'tobacco'.

		Alcohol			**Tobacco**		
		Freq	**Mean**	**Std Dev**	**Freq**	**Mean**	**Std Dev**
Control	**Absent**	206	15.1	22.19	206	2.5	3.74
	Present	96	17.7	26.12	96	3.0	3.32
	Total	302	15.9	23.50	302	2.6	3.61
Case	**Absent**	64	16.3	19.81	64	5.9	6.67
	Present	96	21.1	29.63	96	5.3	4.70
	Total	160	19.1	26.18	160	5.5	5.57
Total	**Absent**	270	15.4	21.62	270	3.3	4.82
	Present	192	19.4	27.91	192	4.1	4.22
	Total	462	17.0	24.48	462	3.6	4.59

d) Hypothesis testing: The researchers aim to answer the following questions by investing the dataset 'CHD'. For each question,

1) Clarify the null and alternative hypotheses.
2) Determine an appropriate statistical test.
3) Check the assumptions.
4) Report your conclusion based on the test result. Test at the significance level of 0.05.

i. Is the CHD status independent of family history?
ii. Is there a difference in mean type-A personality score depending on family history?
iii. Is the Pearson's correlation coefficient of alcohol and tobacco consumption equal to 0?
iv. Is the proportion of having family history greater than 40%?

e) Fitting a model: Participants were queried about their medical status and personal habits with the ultimate goal of testing whether alcohol and tobacco are related to heart disease controlling for potential confounders. Fit an appropriate model with CHD status as a response variable. Interpret the final model you choose after model selection including, but not limited to

i. Overall significance
ii. ROC curve (If applicable)
iii. Goodness-of-fit (If applicable)
iv. Estimated coefficients (Interpretation, significance)