MSBA7002 Business Statistics - HW 3

Name: ________________________ Student ID: ____________________________

20 Nov 2023

Overview / Instructions

This homework will be *due on 1 December 2023 by 11:55 PM* via Moodle.

You are required to submit 1) original R Markdown file and 2) a knitted HTML or PDF file. Please provide comments for R code wherever you see appropriate. Nice formatting of the assignment will have extra points.

In general, be as concise as possible while giving a fully complete answer. All necessary data are available in Moodle.

Remember that the Class Policy strictly applies to homework. You are encouraged to work in groups and discuss with fellow students. However, each student has to know how to answer the questions on her/his own. Note that the final exam is individual based.

Question 0

Review the lectures.

Question 1: Awards Data

The data **awards.csv** contains the number of awards earned by students at one high school. We consider two predictors:

□ **prog**: the type of program in which the student is enrolled (3 categories of general, academic or vocational).

□ **math**: a continuous variable representing students' scores on their math final exam.

Use the following to load data and address the questions using Poisson regression.

```
awards <- read.csv('awards.csv')
awards$prog <- factor(awards$prog, levels=c("General","Academic","Vocational"))
```

1. Is mean roughly equal to variance for each type of program?
2. Consider model_1 "num_awards ~ math + prog" and model_2 "num_awards ~ math". Use "anova(model_2, model_1, test='Chisq')" to test significance of "prog".
3. How many more wards do we expect one student to get in the "Academic" program compared to the "General" program with a math score of 90?

Question 2: Lung Cancer Data

The data.txt file is a 12625 x 56 matrix; each column (row) of the matrix corresponds to the individual case (gene).

Among the 56 cases,

Columns 1~20: pulmonary carcinoid samples (Carcinoid);

Columns 21~33: colon cancer metastasis samples (Colon);

Columns 34~50: normal lung samples (Normal);

Columns 51~56: small cell carcinoma samples (SmallCell).

Before the following analyses, please first center each row of the data, i.e. remove the mean of each row and transpose the matrix.

Q2.1 Principal Component Analysis

- Apply PCA to the transposed matrix to extract the first three principal components. Be careful with outliers.
- Use the scree plot to explain whether these three components are sufficient.
- Plot pair-wise scatterplots and use the sample labels (Carcinoid, Colon, Normal, SmallCell) to explain the plots.

Q2.2 Nominal Logistic Regression, LDA and SVM

Consider the first three principal components in Q2.1 as predictors, and the sample labels (Carcinoid, Colon, Normal, SmallCell) as the categorical response.

- Perform nominal logistics regression, LDA and SVM.
- Present, compare and discuss the results.

Q2.3 Clustering

Consider the first three principal components in Q2.1 as predictors.

- Perform K-means and hierarchical clustering analyses using the first three principal components.
- Use the sample labels to discuss whether the two analyses are reasonable.