STA 501- Final Examination

Methodology of Applied Statistics

Due: Thursday, 05/13/2021, 11:30 PM

1 Introduction

This final exam will use a publicly available heart failure data set taken from *Kaggle.com*. You can download the data set and the variable description from the following links:

1. https://stat501.s3.amazonaws.com/w15-HeartFailure.csv.
2. https://stat501.s3.amazonaws.com/w15-heart-failure-description.txt.

The following code reads this data set from the website directly.

```r
heart.failure = read.csv("https://stat501.s3.amazonaws.com/w15-HeartFailure.csv")
```

If you want to pick a specific variable in the data set to study, for example, age, you can simply use the command **heart.failure$age**. You need also to run the following code in order to use R functions **Pearson.chisq()** and **table2x2()** to perform Pearson χ^2 test and calculate measures of association.

```r
source("https://stat501.s3.amazonaws.com/w12-table2x2Calculator.txt")
```

The specific questions will be listed in the last section.

2 Data Set Description

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyper-lipidaemia, or already established disease) need early detection and management wherein a statistical can be of great help.

3 Data Dictionary

The variable names and their corresponding definitions are listed below.

- age: patient age
- anaemia: Decrease of red blood cells or hemoglobin (boolean, 1 = yes, 0 = no)
- creatinine_phosphokinase: Level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (boolean, 1 = yes, 0 = no)
- ejection_fraction: Percentage of blood leaving the heart at each contraction (percentage)
- high_blood_pressure: If the patient has hypertension (boolean, 1 = yes, 0 = no)
- platelets: Platelets in the blood (kiloplatelets/mL)
- serum_creatinine: Level of serum creatinine in the blood (mg/dL)
- serum_sodium: Level of serum sodium in the blood (mEq/L)
- sex: Woman or man (binary, 1 = woman, 0 = man)
- smoking: patient smoking status (1=yes, 0 = no)
- death.event: 1 = yes, 0 = no

4 Problem Set

Problem 1.

In regression models, an implicit assumption about predictor variables is that all numerical predictor variables are **not** linearly correlated. Sometimes, people make pair-wise scatter plots to identify the correlation between two numerical variables. In this problem, we want to assess the relationship between two numerical variables *serum sodium* and *ejection fraction* by

- making a scatter plot to visualize the potential correlation between *serum sodium* and *ejection fraction.*
- compute the Pearson correlation between *serum sodium* and *ejection fraction.*
- Fit a simple linear regression model using *ejection fraction* as response and *serum sodium* as the predictor.

Write a paragraph or two to summarize what you observed from the output of the above three analyses.

Problem 2.

The key variable in the data set is the death status. Pretend that we want to know whether smoking affects death. The following R code creates a 2 × 2 contingency table.

```r
cont.table = table(heart.failure$death.event, heart.failure$smoking)
colnames(cont.table) = c("survival", "death")
rownames(cont.table) = c("non-smoker", "smoker")
kable(cont.table, caption = "Contingency table: death vs smoking", align="c")
```

Table 1: Contingency table: death vs smoking

	survival	death
non-smoker	137	66
smoker	66	30

Using the above 2 × 2 table to answer the following questions:

- Perform a Pearson χ^2 test to justify whether smoking affects the chance of dying from heart failure.
- Calculate the measures of association.
- Fit the logistic regression model using **death event** as response and **smoking** as the predictor variable.

Write one paragraph or two to summarize the results of the above analyses.

Problem 3.

The data was collected to identify potential risk factors that are associated with death due to heart failure. There are 11 potential predictor variables in the data. The common practice in the real-world application is to work with domain experts (clinicians in this case) to identify practically important predictor variables to be included in the model regardless of their statistical significance and then search statistically significant predictor variables to include in the final working model.

Let's assume that **diabetes**, **smoking**, **ejection fraction**, and **serum creatinine** are the variables identified by clinicians and **age** was identified by statistical approaches. Fit the logistic regression model with the aforementioned predictor variables to the data and summarize the results.