NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

DSA1101 Introduction to Data Science

(Semester 2 : AY 2022/2023)

Midterm Test

Time Allowed: 80 minutes

INSTRUCTIONS TO STUDENTS

1. Students are required to complete this test individually.
2. All submissions are done online via **Canvas**, folder “Midterm Test Submission”.
3. **Your submission should have only one file (.R) which includes the R code and your interpretation/analysis as comments**. Make sure that there is no error when the graders open and run your code.
4. Be sure to lay out systematically the various parts and steps in your code file.
5. Your submission file should be named **using your student number** such as A0123456B.R.

1. (*50 points*) A study of nesting horseshoe crabs (J. Rrockmann, *Ethology*, **102**: 1-21, 1996) collected data on 173 female horseshoe crabs.
Explanatory variables included the female crab's **color** (2,3,4 and 5 which increases when the darkness increases), spine condition, **weight** (kg), and carapace **width** (cm).

Variable "**satell**" indicates the number of male crab (called satellite) attached to the female crabs (for example, the first crab has 8 satellites attached to it).
In this problem, we are interested in the possible factors like **color**, **weight** and **width** that may affect the number of satellites attach to a female crab.

Data are given in the file `data1.txt`.
Import the given data into **R** and keep all the variable names as original.

Part I: Exploring the response variable - **satell**

1. Create a histogram of this variable with a normal density curve overlaying. Give your comment about this plot.

2. Create a box plot of this variable. Does it show any outliers? If yes, retrieve the full information (full row) of the crabs that are outliers in number of satellites. Copy the information and paste into R code file as comments.

3. Create a qq plot of this variable. Give your comment about this plot.

Part II: Variable color

4. Write the code to create a new categorical variable, **col**, which equals to "light" if the crab has color 2 or 3, and equals to "dark" if the crab has color 4 or 5.

5. Create a frequency table for variable **col** created above. How many crabs are of light color and how many crabs are of dark color?

6. Plot a scatter plot of **weight** and **satell**, classified by **col**. Add a legend box for this plot. Give your comment about this plot.

Part III: Modelling

7. Fit a linear regression model for **satell** using three features **color**, **weight** and **width**.

8. Report the value of R^2 of the model above. Give your comments on the goodness-of-fit of the model.

2. (*30 points*) A data set, `data2.csv`, contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. [1]

Variable	Description
age	age of patient at which they undergone surgery
year	year in which patient was undergone surgery(1958–1969)
node	number of lymph nodes that have cancer cells detected
status	1 = the patient survived 5 years or longer (negative); and 2 = the patient died within 5 years (positive)

A table of variable description is given below. **status** is the response in this study.

For this problem, we will not consider the year when patient was undergone the surgery, **year**, as a feature for classification.

In R, use `set.seed(999)`.

We would:

- Use KNN algorithm where K can be any positive integers, from 1 up to 50, to form classifiers to predict the outcome.
- For each classifier, evaluate it's performance by 3-fold cross validation.
- Use type I error rate and type II error rate as the measures to evaluate each classifiers.
- Select the best K.

1. Write the code for the purposes above. For each classifier, the average of Type I error rates from 3-fold cross validation is saved in a vector, (named **ave.type1**); and the average of Type II error rates is saved in another vector (named **ave.type2**).
2. Report the length of vector **ave.type1** and the length of vector **ave.type2**.
3. Write the code to produce a scatter plot where **ave.type1** is in X-axis and **ave.type2** is in Y-axis.
4. For this study, we assume that type I error can be tolerated while type II error is not. Which value of K would you choose among the three smallest type II error rate, yet the type I error rate is not larger than 15%? Report the type I and type II error rate for that value of K.

END OF QUESTIONS

[1] https://www.kaggle.com/datasets/gilsousa/habermans-survival-data-set