Department of Statistics
2020/21 – Semester II

STA272 – STATISTICAL COMPUTING

PRACTICAL TEST I

MARCH 12, 2021

Time: 19h00 – 07h00 **Marks: 100**

Instructions:

- Answer **all** questions.
- All of your work must be typeset using Rmarkdown and submit the printed output showing all your `R` functions.
- Any work submitted late would be penalized as follows:
  - 0 – 10 minutes late submission attracts a penalty of 10%
  - 10 – 20 minutes late submission attracts a penalty of 25%
  - 20 – 30 minutes late submission attracts a penalty of 50%
  - Otherwise you'll be awarded a **zero mark**.
- Students are allowed to use the in-built help files of `R`, Moodle learning materials and the Internet.
- Any form cheating, including any form of collusion is **strictly prohibited** and would be punished harshly.
- Any plagiarized work, especially copied directly from the internet, will be awarded a **zero mark**.
- **Note that there is no one correct answer with programming and this makes it easy to identify copied work.**

Q1. (a) Explain in words how the following operations will be carried out in `R`.

```
exp(c(1,1) %*% matrix(c(4,2,2,9), 2)^(1/2) %*% c(1,1))
```

(b) A student was asked to approximate the value of e^x, $x = \sqrt{\pi}$ using the first 10 terms of its Taylor series expansion given by

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

where $n! = n \cdot n-1 \cdot \ldots 2 \cdot 1$ and $0! = 1$. In her attempt, she provided the following code which unfortunately couldn't run but also not quite correct.

```
x = sqroot(pi)
n = 0:10
approxExpVal = sum(x^n/comprod(n))
```

i. Critically analyse the code and correct the student's solution without re-writing the entire code. That is, edit the code until it works but without redoing everything from scratch.

ii. Compare your solution with the one you would get with the `exp()` function

[10+(10+5) = 25 marks]

Q2. Simple linear regression model for (x_i, y_i), for $i = 1, 2, \ldots, n$,

$$y_i = \alpha + \beta x_i + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

can be expressed in matrix form as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$$

where $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ is a vector of the response observations and $\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ is matrix of the explanatory observations. It can be shown that the least square estimator of $\boldsymbol{\beta} = (\alpha, \beta)^T$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

(a) Use the above formula to estimate the parameters of the following linear regression model $Y = \alpha + \beta X + \epsilon$ for the following simulated data.

```
X = rpois(100, 8)
Y = 2.5 + 0.9*X + rnorm(100, 0, 2)
```

(b) Comment on your parameter estimates compared to the ones used for simulation.

[20+5 = 25 marks]

Q3. Suppose you are tasked with creating email addresses and default passwords for the following students.

Names	Surnames	Student ID
KABELO	MALOME	202001428
DAVID	APRIL	202000090
LERATO	SEBAGA	202000178
BATHO	KGOSI	202005533

(a) Input the following names, surnames and student IDs as individual vectors in R.

(b) Use the vectors created above with `paste()/paste0()` functions to create the students email addresses of the form `studentid@ub.ac.bw`

(c) A package named `stringr` is used to handle character vectors in `R`. Use its `str_remove_all` and `str_sub` functions to create default passwords for the students. The default passwords are made up by concatenating the first three (3) consonants (B, C, D, F, ...) of the surnames with the first 3 letters of the names and the last four numbers of the student numbers. For example, `MLMKAB1428` will be the password for the first student.

[6+7+12 = 25 marks]

Q4. In `R`, there is an inbuilt data set called `mtcars`. This data set comprises of fuel consumption (`mpg`) and other 10 aspects of design and performance for 32 automobiles (1973/74 models).

(a) Use this data set to create a new data frame comprising of fuel consumption, number of cylinders and weights for automobiles with a V-shaped engine and manual transmission.
`HINT:` check out the help page of `mtcars` data set to identify cars with a V-shaped engine.

(b) Evaluate the mean and standard deviation of the fuel consumption of the automobiles in your new data set.

[15+10 = 25 marks]

ALL THE BEST