# SS4864A/SS9864A

## Assignment #1 due to September 29, 2021

**Instructions**

Assignments must be prepared by using RStudio Markdown/Notebook with appropriately annotated with comments, plots, and explanations; notice that neatness counts. Submit your pdf output as well as Rmd (R Markdown/Notebook) files to owl before deadline.

1. One of the most important tasks in using R is reading data from external sources into R objects, so that you can perform calculations and produce graphics.

   (a) Use **help(read.table)** to find the detailed usage of R function **read.table** and its variations: **read.csv, read.csv2, read.delim, read.delim2**. Find their similarities and differences.

   (b) Similarly do the same for R function **write.table** and its variations. Can you match corresponding functions in (a) (export vs import)?

   (c) Use **read.table** or one of its variations to import the file **bank.csv** (in Data Sets folder on owl) into R as a data.frame and explore some of its contents. This dataset contains the information about bank marketing (see the file bank.description.txt on owl). Find all variable names and how many observations. Does it contain any missing values?

2. Investigate R functions **apply, lapply**, and **sapply** functions (you can substitute **sapply** with **replicate**). Find their similarities and differences. Then use one of those functions to implement a simple Monte Carlo simulation. It is about CLT for sample mean. Let the population distribution be exp(rate = 1) (very skewed). Then the distribution of sample mean with sample size n will approximate to normal distribution when n is large. To test it, you need to simulate n such Exp samples and compute its sample mean. Next repeat it K times to get K sample means. With K sample means, a histogram can be produced. Now choose n=10, 50, 100 and K=10000 so that three histograms will be produced. You should try not to use any looping (for, while or repeat loop). You are allowed to use more than one of those functions. Try to see if you can write the whole computations in one line (not big long line). Comments out the results.

3. Investigate R functions **class** and **methods** for object-oriented programming. Use **help** to find what these two functions are used for. Then work on two different data objects **sunspots** and **cars**. Identify their classes. Use R functions **plot** and **summary** on both data sets. Explain why **plot** produces two different plots. What are the outcomes of **summary** and why? Create a new data vector as **x = rnorm(100)** and use **plot** to produce a plot. Then use **class** to change its class to "**ts**" (other equivalent function is **as.ts**) and apply **plot** again. Explain why a different plot is produced on the *same* object.

4. For the data.frame you imported in Question 1, it contains the inform about bank marketing. Do the follow calculations:

(a) Get the summaries of balance for y = "yes" as well as y = "no" respectively. Then compare their distributions. Can you conclude that balance can be used to predict y?

(b) Any association between marital and y? Similarly, check association between housing and y. Notice that when dealing with categorical variables, it might be easy to convert them into factors first.

5. In this question you will study a simple data mining procedure by fitting a simple linear regression model with a selected training data and cross-validating with multiple testing data. Please coding your steps as efficient as possible (by avoiding unnecessary looping).

(a) Generate a raw dataset by the following codes

```
set.seed(4864)  #all will have the same dataset
N = 10000
x = sample(seq(0, 20, by = 0.1), size = N, replace = TRUE)
y = 1 + 0.1 * x + rnorm(N, sd = 1 + 2 * rbinom(N, 1, prob = 0.1))
```

Put $x$ and $y$ into a data.frame with proper variable names and scatter plot $y$ against $x$. Comment your findings.

(b) Generate a training dateset by using the SRS procedure on the raw data generated in (a). The sample size is 500. Then build two regression models: one is $y$ against $x$ and another one is $y$ against$\sqrt{x}$. Run some model diagnostic tests to test model adequacy like checking normal errors. Find the RMSE (root mean square errors) for each model.

(c) Generate a testing dateset with size 500 by using the SRS procedure on the raw data generated in (a) after excluding the training dataset. Use this testing dataset to test/validate the models constructed in (b). You need to find predicted values and residuals. Comment your findings with some plots. For each model, find the RMSE and compare it to one found in (b).

(d) Repeat (c) 500 times and find their RMSE values for each model. Plot (histogram or density) those values to conclude if the models found in (b) are adequate or not. In each plot, you should add a vertical line with the corresponding value of RMSE found in (b). Comment your findings. Can this cross-validation method detect model misspecification? You may need to generate another training dataset in (b) to see if your conclusion is consistent.