STAT5P87 **Assignment I**

Brock University
STAT5P87 - Winter 2025
Assignment I

Due Date: Sunday January 26th

- **Instructions:**
  - Assignments should be submitted on Brightspace as a **single** R script (.R file)
    - Please be sure to use R version 3.6.1 or greater to ensure random number generation consistency.
    - Solutions should not use any additional packages.
    - Written parts of the solutions should be included in the script as comments.
    - I will make sure that R can find and load the required data sets, otherwise the code should run as submitted.
  - The weight of each question is indicated in bold [#].
  - Coding style and efficiency will account for [**4**] marks on the assignment.
  - Style Requirements:
    - appropriate use of white space (spaces, line breaks)
    - indentation of functions, for loops, etc.
    - minimal lines of unnecessary code

---

1) [**4**] This question uses the 'a1-q1.csv' spreadsheet available on Brightspace. Write an R script to load the csv into a dataframe and perform the following manipulations:
   - Rename the variable 'x' to 'x1'
   - Remove all rows corresponding to observation 2
   - Add rows to the data frame for a new observation 4 (x1 = 3, y = 2)
   - Add rows to the data frame for a new variable x2 (x2(observation = 1) = 3, x2(observation = 3) = 1, x2(observation = 4) = 5).
   - Reorder rows so observations are grouped together (if necessary)
   - Create a new column named 'value-squared' containing the squared y, $x1$ and $x2$ values for each observation

2) **[4]** This question uses the data available in 'a1-q2.csv' on Brightspace. There is data from 10 schools, each school has a variable number of students with scores recorded, and each student has a variable number of scores recorded. The problem is that each school numbered its students starting at 1, but 'student 1' from school A is not the same as 'student 1' from school B. Write an R script to load the data, and modify the `student` column so that each student is labeled with a unique number.

3) **[5]** This question uses the `iris` dataset within R. You can load the data using the command

```
> data(iris)
```

For this question, you should use the odd numbered observations as training data, and the even numbered observations as testing data. Ridge regression will be used to model `Sepal.Length` as a function of `Sepal.Width`, `Petal.Length`, and `Petal.Width`. Find the value of λ that minimizes testing error for the ridge regression model.

4) **[8]** Perform a simulation study to explore the bias-variance trade-off in the context of linear regression. In order to perform a simulation study, we need a probability model to simulate from (representing the 'ground truth'). The study has a $p = 15$ dimensional input. The probability model used to generate training data ($\boldsymbol{X}$) is that inputs are independent, and uniformly distributed (between -1 and $+1$). The output y is related to the inputs by

$$y = \sum_{j=1}^{15} \beta_j x_j + \epsilon,$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$ and $\beta_j = 1/j$ for all j (note, $\beta_0 = 0$). An R script with some preliminary work is posted to Brightspace as 'assignment1-question4.R'. The script includes a function to simulate input and output training data and the values of several simulation parameters.

The goal of the simulation study is to study the bias-variance trade-off in the context of linear regression. I've broken the simulation down into smaller steps below, but you only need to hand in a script that performs the simulation and generates the final plot.

a) Estimate $\hat{y}_{new}$ for linear regression models with different numbers of inputs k. Each of the following steps builds on the previous to accomplish the simulation
   i) Simulate a training dataset of size $n = 30$. Use the functions given in the R script to simulate input data ($\boldsymbol{X}$), and then use the input data to simulate corresponding output data ($\boldsymbol{y}$).
   ii) For a fixed value of k, use the training data to fit a linear regression model that includes only the first k inputs (don't count the intercept as one of the k).
   iii) Use the estimated model to predict y_{new} based on the input $x_{new} = (1, 1/2, \ldots, 1/2)$.
   iv) Repeat steps (ii-iii) for k from 1 to 15. Store the estimated values in first row of the matrix `hatY`.
   v) Repeat step (i-iv) 1000 times to get 1000 independent estimates of y_{new} for each value of k.

b) For each value k in 1 to 15, estimate the variance, squared bias and MSE of the linear regression model.

c) Plot the variance, squared-bias and MSE as a function of k (all on the same plot).