STAT 362 W23 R for Data Science

Assignment 1

Due: 20 Jan (Friday) 11:59pm.

Instruction:

1. Submit a `.R` file in onQ under Assignment 1. Your file name should be `FirstName_LastName_studentID_Asg1.R`. For example, `Brian_Ling_12345678_Asg1.R`.
2. All the questions require you to write some R code. Clearly separate the R code for different questions using some appropriate comments:

```
# Q1

# your code

# Q2

# your code

# Q3

# your code

# and so on
```

3. You may receive **at most half** of the points for a question if the code cannot be executed.
4. You should have proper indentation in your code.
5. ***Posting any questions on Chegg or other platforms may result in a fail in this course***. You may discuss with your classmates with proper acknowledgment in your assignment. You can send me an email or come to my office hours to ask questions about the assignment.
6. There are 10 questions in this assignment. Full marks = 100. Each question is worth 10 points.

Additional remarks:

1. For questions that do not ask you to write a function, do not write a function.
2. Do not include any unnecessary code in your submission to save our TA work from looking for your answer.
3. Use meaningful names for your variables if you need them in your intermediate steps.

Q1(a): Write a function (using the operators and built-in functions in Section 1 of the lecture notes) called `my_LS` to compute the least squares estimate in linear regression given the design matrix X ($n \times p$) and response vector Y (length n). The least squares estimate is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

The function should take X and Y as two inputs and outputs a vector (not a matrix). In addition, you are not allowed to use `lm` in this question.

For example,

```r
X <- cbind(rep(1, 10), 1:10)
Y <- seq(1, 30, length = 10)
my_LS(X, Y)
```

```
## [1] -2.222222  3.222222
```

Your function should give the same output as above.

Hint: you may use `as.vector` to change your matrix into a vector. E.g.,

```r
A <- matrix(1:3, nrow = 3, ncol =1)
A
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
```

```r
as.vector(A)
```

```
## [1] 1 2 3
```

Common problems observed from previous years:

1. You cannot use `^{-1}` to find the inverse of a matrix in R.
2. It is incorrect to use `*` to do the matrix multiplication. Use `%*%`.
3. Naming of the arguments: you should write `my_LS <- function(X, Y)` not `my_LS <- function(x, y)`.

Q1(b): Write a function (using the operators and built-in function in Chapter 1 of the lecture notes) called `my_ridge` to compute the ridge estimate in ridge regression given the design matrix X ($n \times p$), response vector Y (length n) and penalty parameter (a scalar) λ. The function should take X, Y, and λ as inputs and outputs a vector (not a matrix). The ridge estimate is given by

$$\hat{\beta}_R = (X^T X + \lambda I)^{-1} X^T Y,$$

where I is an identity matrix of appropriate dimension.

For example,

```r
X <- cbind(rep(1, 10), 1:10)
Y <- seq(1, 30, length = 10)
lambda <- 1
my_ridge(X, Y, lambda)
```

```
## [1] -1.374556  3.093093
```

Your function should give the same output as above.

Remark: $\hat{\beta}_R$ minimize

$$\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2 + \lambda \sum_{i=1}^{p} \beta_i^2.$$

The goal of this question is for you to practise some basic operations in R (not to introduce ridge regression).

Hint: the dimension of the identity matrix is the same as `t(X)%*%X`.

Q2(a): Write a function called `my_sum` using two for loops to find $\sum_{x=1}^{n}\sum_{y=1}^{m}\frac{x^2y}{x+y}$.

For example, `my_sum(n = 5, m = 6) =` 145.5935065.

Q2(b): Write a function called `my_sum2` without any loops (for loops, while loops, repeat loops) to do the same task in Q2(a).

Hint: you may define two matrices A and B of suitable dimensions, obtain a new matrix C by performing vectorized operators on A and B such that the sum of all the elements in C gives you the required sum.

Q3: Write a one-line R code to sample an integer randomly from $\{1, 2, 3, 4, 5\}$ using `runif` (each integer has 0.2 probability of being selected).

Hint: you may use some rounding functions. In this question, you cannot use `sample` or other simulation functions other than `runif`.

Also, with probability one, you will not get b if you simulate from $Unif[a, b]$ because this is a continuous distribution.

Q4: You have a biased coin with $P(Head) = 0.6$. Write down a code for simulating the results from flipping this coin 10 times independently. You may use 1 to denote Head and 0 to denote Tail.

Q5: The l_p-norm of a vector $x = (x_1, \ldots, x_n)$ is defined as $\|x\|_p := (\sum_{i=1}^{n}|x_i|^p)^{1/p}$. Write a function called `l_p` to compute $\|x\|_p$. The inputs are `x` and `p`, and the output is $\|x\|_p$.

Q6: Let `v` be a vector of some positive integers. Write a one-line R code to compute the sum of all the odd integers in v. When you test your code, you may define some v.

E.g., if $v = (1, 5, 9, 4)$, your output is 15.

Q7: Write R code to create a plot of the probability mass function of the geometric distribution with parameter 0.3 from $x = 1$ to $x = 10$.

Q8: A simple model on the stock return assumes that (i)

$$r_{t+1} := \log\frac{P_{t+1}}{P_t} \sim N(\mu, \sigma^2),$$

where r_{t+1} is the log-return at Day $t + 1$, P_t is the stock price at the end of Day t; (ii) $r_1, r_2, \ldots$ are iid.

Today is Day 0. Suppose that the current price of a certain stock P_0 is 100, $\mu = 0.0002$ and $\sigma = 0.015$. Denote A to be the event that the price is below \$95 at the close of at least one of the next 20 trading days (i.e., Day 1 - Day 20) and B to be the event the price is above \$101 at the close of at least one of the next 40 trading days (i.e., Day 1 - Day 40). Using simulation, estimate $P(A \cap B)$.

Hint:

1. see Section 3.5 in the lecture notes.
2. The two events A and B are not independent. You were asked to find the joint probability of the two events. You cannot multiply the $P(A)$ and $P(B)$ together to get the joint probability $P(A \cap B)$ because they are not independent.

Q9: Suppose we have a biased coin with the probability of getting a head begin equal to 0.7. Consider a game where we flip the coin, win \$1 if the result is head and lose \$1.5 if the result is tail. You play the game 100 times. You are interested in the possible paths of the cumulative profit. Following Section 2.2 in the lecture notes, write R code to create a plot of one simulated path of the cumculative profit vs the number of games.

Q10: Go to https://support.rstudio.com/hc/en-us/articles/200711853-Keyboard-Shortcuts and then write down the shortcuts (Windows & Linus/ Mac) for performing the following actions in RStudio.

1. Clear console
2. Move cursor to Console
3. Interrupt currently executing command
4. Move cursor to Source Editor
5. Save active document
6. Run current line/selection
7. Undo
8. Cut
9. Copy
10. Paste