HOMEWORK 1
BIOSTATISTICS 835
DUE SEPTEMBER 7TH, 2023

The solution to your homework should be emailed to the instructor. Your solution can be the output of an `rmarkdown` file (containing code and output), or the combination of a word file with the solutions and an `R` script containing the code used to get the solutions.

1. **(15 points)** Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric $J \times J$ matricies. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ have eigen values $\{\lambda_j(\boldsymbol{A})\}$ and $\{\lambda_j(\boldsymbol{B})\}$. The Hoffman-Wielandt Theorem (Hoffman and Wielandt, 1953) states that

$$\sum_{j=1}^{J}\{\lambda_j(\boldsymbol{A}) - \lambda_j(\boldsymbol{B})\}^2 \leq tr\{(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}-\boldsymbol{B})^\top\}$$

   Prove this result. Hint: Use the spectral decomposition theorem on $\boldsymbol{A}$ and on $\boldsymbol{B}$; express $tr\{(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}-\boldsymbol{B})^\top\}$ in terms of the decomposition matrices of $\boldsymbol{A}$ and $\boldsymbol{B}$, and simplify; then, show that the result is minimized by $\sum_{j=1}^{J}\{\lambda_j(\boldsymbol{A}) - \lambda_j(\boldsymbol{B})\}^2$.

2. As discussed in class, the `RcppArmadillo` package in `R` can used to speed up calculations, specifically with matrix multiplication. The matrix multiplication we are interested in this problem is the following:

$$D = Z'(A'A)Z = Z\Sigma Z' \tag{1}$$

   where Z is an $n \times p$ matrix, A is an $n \times n$ matrix and $\Sigma = A'A$.

   The A and Z we'll perform (1) on will be as follows:

```
set.seed(1234)
n <- 2000
p <- 10

A <- matrix(runif(n^2)*2-1, ncol=n)
Z <- matrix(rnorm(n*p),n,p)
```

   Two similar, but different functions we could use to perform (1) are:

```
f1 <- function(A,Z){ t(Z) %*% (t(A)%*%A) %*% Z }

f2 <- function(A,Z){
  Sigma <- (t(A)%*%A)
  temp1 <- t(Z) %*% Sigma
  temp1 %*% Z
  }
```

Both of these will give the same answer, but are not the same. Which is faster? We can answer this question using the `mark` function in the package `bench` using the following code:

```
install.packages("bench")
library(bench)
mark( f1(A,Z), f2(A,Z) , min_iterations = 10)
```

We're going to use the above to answer the following questions.

(a) **(10 points)** Which function (f1 or f2) has a shorter median time of execution? Which uses a smaller amount of memory?

(b) **(10 points)** Write a function that will perform (1) using `RcppArmadillo`. For an example of how to write an `RcppArmadillo` function see here. To call the function in `R` you will need to install and load the `Rcpp` and `RcppArmadillo` packages.

```
install.packages("Rcpp")
install.packages("RcppArmadillo")
library(Rcpp)
library(RcppArmadillo)
```

(c) **(10 points)** Compare the performance of the `RcppArmadillo` function to f1 and f2 using the `mark` function. Which has the shortest run time? Which uses a smaller amount of memory?

(d) (BONUS: worth 5 points, optional) Repeat (b) and (c) using `RcppEigen` instead of `RcppArmadillo`.

3. This problem will use the *"baseball"* data from Journal of Statistics Education (JSE) data archive, which is available on the course website. The data consists of measurements salary information for 337 Major League Baseball (MLB) players who are not pitchers and played at least one game during both the 1991 and 1992 seasons. The purpose of the study is to determine whether a baseball player's salary is a reflection of his offensive performance.

(a) **(7 points)** Using the `lm` command regress 'salary' on all variables in the dataset (see the description of the data for the labels). Turn in the the table of estimates, standard errors, T-statistics and p-values (each variable should have the appropriate label). Give two estimates of the average squared error of the model, and comment on which is better.

(b) **(7 points)** Discuss whether you think the two estimates of the average squared error in the previous question are a good estimates of the *expected prediction error* (EPE). Why or why not?

(c) **(5 points)** Using the X variable that appears to be the most significantly related to the outcome, fit a K-Nearest Neighbor (NN) model for $K = 1, 5, 10,$ and 15. Estimate the MSE for each k

(d) **(5 points)** Repeat what you did in the previous question for the J-most significantly related X variables for $J = 2, 4,$ and 8. Give what criteria you used to determine which variables you found to be the J-most significant.

(e) **(6 points)** At this point, you know should have estimated the MSE for $J = 1, 2, 4, 8$ and $K = 1, 5, 10, 15$. Construct a plot with K on the x-axis and APE on the y-axis. For each J the MSE values should be connected by lines (using different line types, point types or colors for each J).

(f) **(5 points)** Comment on the pattern of the MSE by K and J. If we were to estimate the EPE (i.e., using test data), what do you think the pattern would be?

(g) **(5 points)** Suppose we were to estimate $bias^2$ instead or the error rate. What do you think the pattern of $bias^2$ by K and J would look like? Explain why you think this pattern exists.

4. **(All four parts of the question worth 5 points.)** Complete parts a), b), and c) from exercise 2.7 from the ESL textbook. Then comment on the differences for the expression for b) and c) (in general, you don't have to apply them to the linear regression or nearest neighbor situations).