Homework I

STA 629, Fall 2024

Due: September 13, Friday

1. (20 pts) Textbook Problem, ESL, Exercise 3.6

2. (80 pts) Gene Expression Data:
   Download the SRBCT microarray data from Canvas. This is a gene expression data set from a childhood cancer study with $n = 83$ patients and $p = 2308$ genes. Your response is the expression profile of p53, a major oncogene that acts as a tumor suppressor. Your goal is to select other genes whose expression profiles are associated with p53. The downloaded dataset includes two files: "X_SRBCT.csv" for the 2307 predictor genes and a separate file "Y_SRBCT.csv" for p53 as outcomes.

   (a) Fit the models and visualize regularization paths for the following methods:
      i. Elastic net
      ii. Lasso
      iii. SCAD
      iv. MC+

   (b) Reflection. Interpret the results. What are the top genes selected by each method? Are they different? If so, why? Which regularization paths look most variable? Why is this the case? If you had to report to a scientist the top 10 genes associated with p53, which ones would you report? Why?

   Note: You can use Python or R. Provided below is the code for loading the data in both Python and R.

```
### Python
import pandas as pd
import numpy as np
X = pd.read_csv("X_SRBCT.csv", header=None, sep=',').to_numpy()
Y = np.squeeze(pd.read_csv("Y_SRBCT.csv", header=None,
    sep=',').to_numpy())
```

```
### R
X <- read.csv("X_SRBCT.csv", header = F)
```

```
Y <- read.csv("Y_SRBCT.csv", header = F)
X <-as.matrix(X)
Y <- as.numeric(Y[,1])
```

- Bonus question (20 pts): Textbook Problem, ESL, Exercise 3.3