STAT 435: Quiz2

All work should be done in Rmarkdown.

Q1 This problem is regarding writing your own code for choosing the regularization parameter λ via cross validation for a lasso estimator.

a. Simulate a data set as follows

```
set.seed(1)
n=300;p=200;s=5
x=matrix(rnorm(n*p),n,p)
b=c(rep(1,s),rep(0,p-s))
y=1+x%*%b+rnorm(n)
```

In this synthetic dataset there are 300 observations, 200 regression coefficients, of which only 5 are non-zero. The objective of the following steps is to recover the vector b from the data in x, and y

b. Define a grid $L = \{0, ..., 2\}$ of 100 numbers between 0 and 2, this grid shall serve as potential values for the regularizer λ. *Hint: you may want to use the function* `seq()`*, with the argument* `length.out=100`

c. Use the function `glmnet()` to obtain lasso estimates for each value in the grid that you define in part (b). Using the function `coef()`, extract the estimated coefficient vector when $\lambda = L[10]$, i.e., the 10^{th} value in the grid L.

d. For each value of λ in the grid L of part (b), compute the mean squared error on the entire dataset. (Use a `for()` loop for this purpose). This will provide a vector of 100 values of mse, one for each value of λ. Plot λ vs mse. What do you observe?

e. Using a `for()` loop. compute the cross vlaidation error, for each value of λ in the grid L, under a $k = 5$ fold cross validation setup. (you will need to repeatedly divide the data into testing and training, this will require another `for()` loop).

f. Compile all your code into a custom function with input arguments x, y, k and a grid L. This function should output the following results, (i) the best fit model with a k-fold cross validation, (ii) the vector of cross validation errors (one for each value of λ), (iii) the grid L used for cross validation, (iv) the value of lambda at which the best fit model is obtained.

g. Finally, use the function you make in Part (f) with $k = 5$, then extract the vector of cross validation errors (say, CVV) and the grid L that is used. Make a plot of L vs. CVV.