DS 303 Homework 6
Due: Oct. 18, 2021 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Concept Review

1. Explain in plain language (using limited statistics terminology) why lasso can set some of the regression coefficients to be 0 exactly, while ridge regression cannot. You may include a figure if that is helpful.
2. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

   a. As we increase λ from 0, the training MSE will:
      i. Increase initially, and then eventually start decreasing in an inverted U shape.
      ii. Decrease initially, and then eventually start decreasing in an inverted U shape.
      iii. Steadily increase.
      iv. Steadily decrease.
      v. Remain constant.
   b. Repeat (a) for test MSE.
   c. Repeat (a) for variance.
   d. Repeat (a) for (squared) bias.
   e. Repeat (a) for irreducible error.

Problem 2: Build a predictive model

We will work with the `Boston` housing data set; it is part of `library(ISLR2)`. Your goal here is to build a predictive model that can predict per capita crime rate. Split your data into a training set and test such that 90% of the observations go into the training set and the remaining 10% go into the test set. Your model building should include the following components:

1. A least-square model with the predictors chosen using a model selection technique of your choice. Explain and justify what technique you have chosen. Call the from this step Model1.
   - Do you think Model1 needs an interaction term between any predictors? Justify.
   - Do you think Model1 requires higher order terms to model any non-linearities? Justify.
   - Do you think Model1 can be improved using a regression spline? Justify.
2. A ridge regression with the optimal λ chosen using 10-fold cross-validation. Compare your models using the λ that gives the smallest CV error and the λ based on the one standard error rule. Call these models Model2a and Model2b, respectively. Report both models.
3. A lasso regression with the optimal λ chosen using 10-fold cross-validation. Compare your models using the λ that gives the smallest CV error and the λ based on the one standard error rule. Call these models Model3a and Model3b, respectively. Report both models.
4. Propose a model (or set of models) that seems to perform well on this dataset. Make sure you are evaluating your model performance using the test set and not using the training error. Report your chosen model(s) here. Does your chosen model involve all of the features in the data set? Why or why not?

Problem 3: Bootstrap

We will continue working with the `Boston` housing data set.

a. Based on this data set, provide an estimate for the population mean of `medv`. Call this estimate $\hat{\mu}$.

b. Provide an estimate of the standard error of $\hat{\mu}$ using an analytical formula. Interpret this result.

c. Now the estimate the standard error $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?

d. Using bootstrap, provide a 95% confidence interval for the mean of `medv`. Compare it to results using analytical formulas.

e. Based on this data set, provide an estimate $\hat{\mu}_{\text{med}}$ for the median value of `medv`.

f. We would like to estimate the standard error of $\hat{\mu}_{\text{med}}$. Since there is no simple formula for computing the standard error of the median, use bootstrap. Comment on your findings.

g. Based on this data set, provide an estimate $\hat{\mu}_{0.1}$, the 10th percentile of `medv`.

h. Use bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.

Problem 4: Properties of Bootstrap

a. What is the probability that the first bootstrap observation is the jth observation from the original sample? Justify your answer.

b. What is the probability that the first bootstrap observation is *not* the jth observation from the original sample? Justify your answer.

c. What is the probability that the jth observation from the original sample is *not* in the bootstrap sample?

d. When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

e. When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

f. When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample?

g. Create a plot (in `R`) that displays, for each integer value of n from 1 to 100,000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

h. Investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 5$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample. You may use the following code:

```
results <- rep(NA, 10000)
for(i in 1:10000){
      results[i] <- sum(sample(1:100, rep=TRUE) == 4) > 0
}
mean(results)
```

Comment on your findings.