DS 303: Practice Problems

Question 1: Concept Review

1. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq s$$

   for a particular value of s. Observe that this is just lasso but formulated differently. Provide a sketch of a typical training error and test error as we increase s from 0. The horizontal axis will represent s and the vertical axis will represent the MSE.

2. Draw a scatterplot of a dataset where there is a linear decision boundary but logistic regression would **not** perform well. Suppose for simplicity that the dataset contains only two groups (represented by circles and triangles) and $p = 2$ (two predictors X_1 and X_2) The horizontal axis of the scatterplot should be X_2 and the vertical axis should be X_1.

3. True or False? For a given data set, we can directly calculate the bias and variance of a regularized regression model to see whether or not the decrease in variance is enough to offset the increase in bias. Based on this, we can choose an optimal λ.

4. True or False? Since ridge regression always returns the full model (with all p predictors), its test MSE will always be smaller than that of Lasso.

5. True or False? QDA is equivalent to using Bayes Rule to approximate $P(Y = k|X)$, under the assumption that the predictors are normally distributed.

6. Suppose you implement QDA on a dataset with $n = 1000$ observations. There are $p = 10$ predictors and you observe that three of the predictors in your model are highly correlated ($VIF > 10$). Will the presence of multicollinearity affect the performance of QDA? State yes or no with a brief justification.

Question 2: Simulations

1. Design a simulation study to calculate the bias for a ridge regression model. Suppose we know that the true underlying population regression model is :

$$Y_i = 2 + 3 \times X_{i1} + 5 \times \log(X_{i2}) + \epsilon_i \quad (i = 1, \ldots, n), \quad \epsilon_i \sim \mathcal{N}(0, 1^2).$$

   You can generate your predictors using the following code:

```
n = 100
X1 = seq(0,10,length.out =100) #generates 100 equally spaced values from 0 to 10.
X2 = runif(100) #generates 100 uniform values.
```

Fix $\lambda = 2$. Calculate the bias for our estimates of β_0, β_1, and β_2. Report those values here.

2. Suppose we wish to invest a fixed sum of money into two financial assets that yield returns of X and Y, respectively. We will invest a fraction α of our money in X and invest the remaining $1-\alpha$ in Y. We want to find the value of α that minimizes the total risk of our investment. One can show that the value that minimizes the risk is given by:

$$\hat{\alpha} = \frac{\hat{\sigma^2}_Y - \hat{\sigma}_{XY}}{\hat{\sigma^2}_X + \hat{\sigma^2}_Y - 2\hat{\sigma}_{XY}},$$

which needs to be estimated from the data.

The `Portfolio` data set in the `ISLR2` package contains data for 100 pairs of stock returns. The `R` script `alpha_fn.R` will compute $\hat{\alpha}$ for you for this data set. Use bootstrap to quantify the accuracy of our estimate of $\hat{\alpha}$. in other words, estimate the standard error using bootstrap.