Coding Assignment 2

Due Date: *Monday, September 22 (11.59PM)*

Submission: on Gradescope

Instructions:

The coding assignments may be completed in groups up to 4 students. If you do so, please make sure that you include everyone's full name, and that you *also select everyone's name when submitting the assignment on Gradescope*. This ensures that each group member will get a grade assigned and have access to the comments from the graders.

For this assignment, the following items need to be submitted:
(1) the **code** in `R` or *Python* (a Markdown file is ok)
(2) a **pdf** file with your code and results (Markdown-style) with all necessary plots and comments.

***Problem 1**: [50 points]*

In this problem, we are going to use a simulated example to illustrate the behavior of *testing* and *training* errors *as the number of predictor* ***increases***.

(a) *[5 points]* Generate a data set with $p = 20$ (independent) covariates, $n = 1,000$ observations and an associated quantitative response vector generated according to the model

$$Y = X\beta + \varepsilon$$

where β is the following vector:

$$\beta = (1,\ 0.5,\ 0,\ -0.5,\ -1,\ 1,\ 0.5,\ 2,\ 0,\ 0,\ 0.1,\ 0.2,\ 2,\ 0,\ 0,\ 0,\ -2,\ 1,\ 0,\ 0)$$

The covariates and the ε can be generated using standard normal random variables. (This is similar to what you did in Coding Assignment # 1).

(b) *[5 points]* Randomly split your data set into a *training* set containing 200 observations and a *test* set containing 800 observations.

(c) *[10 points]* Perform **best subset selection** on the training set, and **plot** the *training* set MSE associated with the best model of each size. Recall that

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(d) *10 points]* **Plot** the *test* set MSE associated with the best model of each size.

(e) *[5 points]* For which model size does the test set MSE take on its minimum value? Comment on your results.
Remark: If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.

(f) *[5 points]* How does the model at which the test set MSE is minimized compare to the true model used to generate the data? *Comment on the coefficient values.*

(g) *[10 points]* Create a plot displaying $\sqrt{\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^r)^2}$ for a range of values of r, where $\hat{\beta}_j^r$ is the jth coefficient estimate for the best model containing r coefficients. Comment on what you observe. How does this compare to the test MSE plot from (d)?

Problem 2: *[50 points - 10 points each]*

Download the soccer data provided on Coursera: there are two files that you need `mls_train.csv` and *mls_test.csv*. These data sets contain publicly available data of soccer players in Major League Soccer. Our goal in this case study is to do *subset selection* and *penalized regression* to predict a player's **salary**.

The covariates that we have available are the following:

Variable	Description
`salary`	A player's salary
`height, cm`	A player's height in cm
`weight, kg`	A player's weight in kg
`game_started`	Total # of games the player was a starter
`mins`	Total # of minutes played in a game
`sub_on`	Total # of games player was a sub
`total_wins`	Total wins a player has
`goals`	Total goals a player scored
`duel_won`	Duel over the possession of the ball where a player wins the ball
`accurate_cross`	Total # of accurate crosses
`assist`	Total # of assists
`yellow_card`	Total # of yellow cards
`won_tackle`	Total # of tackels won
`aerial_won`	Total # of aerials won
`ontarget_scoring_att`	Total on target scoring attempts
`successful_short_pass`	Total # of successful short passes
`won_corners`	Total # of corners won
`ball_recovery`	Total # of loose balls a player takes possession of
`total_offside`	Total # of offsides

(a) Fit a **best subset selection** algorithm to the data set and *report* the best model of each model size (up to 8 variables, excluding the intercept) and their prediction errors. Make sure that you simplify your output to only present the essential information.

(b) Using the models reported in part (a), which is the best model acording to: (i) AIC, (ii) BIC, (iii) C_p-Mallows, and (iv) R_a^2? For each criterion, report the MSE for both training and testing data.

(c) Fit a **ridge** regression model to predict a player's salary. Use cross-validation to select the best regularization parameter λ.

(d) Fit a **lasso** regression model on the same data set. Identify which features are shrunk to zero.

(e) Compare the performance of the models in (b) vs. ridge vs. lasso using the MSE. Which model would you recommend for predicting MLS player salaries and why?