Assignment 1 (Due Date is 23:59pm 12th Aug 2022)

The data comes from a study of prostate cancer. The data information for the file "prostate.data.txt" is listed as follows.

- Predictors columns (2–9)
  - lcavol: log(cancer volume)
  - lweight: log(prostate weight)
  - age
  - lbph: the logarithm of the amount of benign prostatic hyperplasia
  - svi: seminal vesicle invasion
  - lcp: log(capsular penetration)
  - gleason: gleason score
  - pgg45: percentage Gleason score 4 or 5
- outcome (column 10)
  - lpsa: the logarithm of prostate-specific antigen
- train/test indicator (column 11)
  - This last column indicates which 67 observations were used as the "training set" and which 30 as the test set. "T" indicates training data while "F" means test data.

Question 1 [50 marks]: Apply the five estimation methods to this data set: OLS, ridge regression, the lasso, the naive elastic net and the elastic net. Please carry out model fitting and choose the tuning parameters (if necessary) by tenfold cross-validation (CV) on the training data. Then compute their prediction mean-squared error on the test data. After doing these work, please complete the following table and attach your R codes to the submission.

Method	Tuning Parameters	Test MSE	Variables Selected
OLS		?	?
Ridge Regression	$\lambda = ?$	?	?
Lasso	s=?	?	?
Naive elastic net	λ=?, s=?	?	?
Elastic net	$\lambda = ?, s = ?$	?	?

Question 2 [10 marks]: Based on the table results completed in Question 1, which method is the winner among all the competitors in terms of both prediction accuracy and sparsity? Which method is the worse method? Please interpret your conclusion.

Question 3 [10 marks]: Comparing the naive elastic net method and the ridge regression, which method is better in prediction and variable selection? Please interpret your conclusion.

Question 4 [10 marks]: Comparing the ridge regression and OLS, which method is better in prediction? Please interpret your conclusion.

Question 5 [20 marks]: Please explain why the elastic net method performs better than the lasso method on this data set.