STAT 534 Statistical Data Science I

Fall 2021

Exam I (Take-Home Portion)

Due Date: October 17 (Sunday)

Instructions: There are 50 points plus 10 extra credits on this portion of the exam. You may use any reference material for your exam. However, you are not to discuss any questions about the exam with anyone other than yourself. Please contact me with any questions and I will respond as soon as possible within the same day. Be sure to write/type your answers in complete sentences and show your work. Any items including R output not commented on will receive no credit. Attach your R code as an appendix or as a separate file. (If you use R markdown, you may keep the code inserted inline.)

1. (*20+10 pts*) In this problem, we will generate simulated data, and will then apply best subset selection to this data using the validation method. The goal is to explore that as model size increases, the training error will necessarily decrease, but the testing error may not.

   (a) (*5 pts*) Generate a data set with $p = 20$ predictors and $n = 1000$ observations according to the model
   $$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$
   where $X_j \sim N(0, 1)$ and $\epsilon \sim N(0, 1)$ independently. Randomly select your β values but let some elements to be exactly zero. Then split your data set into a training set containing 100 observations and a testing set containing 900 observations.

   (b) (*3 pts*) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size. (Hint: regsubsets() returns error (or residual) sum of squares (rss) for each model and MSE = RSS/n.)

   (c) (*6 pts*) Plot the testing set MSE associated with the best model of each size. For which model size does the testing set MSE take on its minimum value? (Hint: For each model, obtain the predicted values for the testing set and then compute associated MSE.)

   (d) (*3 pts*) What do you observe about the changes in training MSE and testing MSE as model size increases?

   (e) (*3 pts*) How does the model at which the testing MSE is minimized compare to the true model used to generate the data? (Hint: You want to refit the regression model to the entire data set using the selected predictors.)

   (f) (*+10 pts*) Create a plot displaying $\sqrt{\sum_{j=0}^{p}(\beta_j - \hat{\beta}_j^s)^2}$, where $\hat{\beta}_j^s$ is the jth coefficient estimate for the best model of size s using the entire data set. Comment on what you observe. How does this compare to the testing MSE plot from part (c)?

2. (*30 pts*) Use the *UScrime* data set in the *MASS* library to study the effect of punishment regimes on crime rates.

   (a) (*5 pts*) Explore the variables using appropriate graphics and summary statistics. Comment on your observations.

   (b) (*12 pts*) Split the data into 75% training and 25% testing and build the following models using the training set:

   - multiple linear regression
   - ridge regression
   - lasso regression
   - principal components regression (justify how many principal components should be used)
   - partial least squares (justify how many directions should be used)

   (c) (*5 pts*) Compare the effectiveness of each model on training vs. testing data.

   (d) (*8 pts*) Select the best two models from above. Interpret and compare their respective final fitted models.