Spring 2020

Quantitative Analysis
Chung-Ming Kuan

Problem Set 7

Due: 4/27

Part One: Hand-Written Exercise

1. For a data set $(y_i, \mathbf{x}_i)_{i=1}^n$, where y_i is a scalar and $\mathbf{x}_i$ a $p \times 1$ column vector. That is, the regression model is $y_i = \sum_{j=1}^p \beta_j x_{ij} + u_i$. Please show that the Ridge Regression estimator $\hat{\beta}_R$ is given by:

$$\hat{\beta}_R = \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' + \lambda \mathbf{I} \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i y_i \right),$$

   where λ is the tuning parameter and $\mathbf{I}$ the p-dimensional identity matrix.

2. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p+1$ models, containing 0, 1, 2, ..., p predictors. For $k = 1, .., p$, please answer the following questions and justify your answers:

   (a) Which of the three models with k predictors has the smallest training RSS?

   (b) Which of the three models with k predictors has the smallest testing MSE?

   (c) (True or False) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.

   (d) (True or False) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

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

$$\min_\beta \text{RSS} = \min_\beta \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^p |\beta_j| \leq s$$

   for some $s \in \mathbb{R}$. Please answer the following questions and justify your answers:

   (a) As s increases from 0 to ∞, what will happen to the training RSS?

   (b) As s increases from 0 to ∞, what will happen to the testing MSE?

   (c) As s increases from 0 to ∞, what will happen to the variance of our estimated coefficients?

   (d) As s increases from 0 to ∞, what will happen to the bias of our estimated coefficients?

Part Two: Computer Exercise

Please load the data set `Hitters` from the package `ISLR`. `Hitters` contains the Major League Baseball Data from the 1986 and 1987 seasons. Note that in your R answer sheets, the `na.omit()` function helps you to remove all of the data that have missing values in any variable. Please answer the following questions:

1. We wish to predict a baseball player's `Salary` on the basis of all the other 19 variables in the `Hitters` dataset. Please derive the best model based on the best subset selection approach. What are the chosen variables in `Hitters` and their corresponding coefficients?

2. Repeat (1) using backward stepwise procedure.

3. Repeat (1) using forward stepwise procedure.

4. For the three different models obtain in (1), (2) and (3), please use 10-fold CV to determine which one is better?

5. Please construct the ridge regression model with `Salary` as the response and all the other variables in `Hitters` dataset as the predictors.

   (a) What is the optimal lambda chosen by LOOCV? (Ignore the warning message if there is any.)

   (b) What is the estimated coefficient for `Hits` under the optimal lambda?

6. Repeat (5) using LASSO regression.

   (a) What is the optimal lambda chosen by LOOCV?

   (b) How many variables, under optimal lambda, are force to be zero?