STOR 565 Spring 2020 Homework 3

Due on 02/07/2020 in Class

YOUR NAME

Remark. This homework aims to help you further understand the model selection techniques in linear models. For the **Computational Part**, please complete your answers in the **RMarkdown** file and print your generated PDF file created by it.

Computational Part

Hint. Before starting your work, carefully read Textbook Chapter 6.5-6.7 (Labs 1-3). Mimic the related analyses you learn from it. Related packages have been loaded in the setup.

1. (Model Selection and Best Subset Prediction, *25 pt*) In this exercise, we will generate simulated data, and will then use the data to perform model selection.

   (a) Use the `rnorm` function to generate a predictor $\boldsymbol{X}$ of length $n = 200$, as well as a noise vector $\boldsymbol{\epsilon}$ of length $n = 200$.

   (b) Generate a response vector $\boldsymbol{Y}$ of length $n = 200$ according to the model

   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

   where $\beta_0 = 4$, $\beta_1 = 3$, $\beta_2 = -2$, $\beta_3 = 0.5$. Split the data set into a size-100 training set and a size-100 test set. (You can just let the first 100 samples be the training set)

   (c) Use the `regsubsets` function from `leaps` package to perform best subset selection in order to choose the best model containing the predictors $(X, X^2, \cdots, X^{10})$. What is the best model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Calculate the test errors of these coefficients.

   (d) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. Report the best coefficients and calculate their test errors. How does your answer compare to the results in (c)?

   (e) Now fit a LASSO model with `glmnet` function from `glmnet` package to the simulated data, again using $(X, X^2, \cdots, X^{10})$ as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

   (f) Now generate a response vector Y according to the model

   $$Y = \beta_0 + \beta_6 X^6 + \epsilon,$$

   where $\beta_6 = 6$, and perform best subset selection and the LASSO. Discuss the results carefully.

2. (Prediction, *25 pt*) In this exercise, we will predict the number of applications received using the other variables in the `Boston` data set from `MASS` package.

   (a) Randomly split the data set into a training set and a test set (2:1).

   (b) Fit a linear model using least squares on the training set, and report the test error obtained.

(c) Fit a ridge regression model on the training set, with λ chosen by 5-fold cross-validation. Report the test error obtained.

(d) Fit a LASSO model on the training set, with λ chosen by 5-fold cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

(e) Fit a PCR model on the training set, with M chosen by 5-fold cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

(f) Comment on the results obtained. How accurately can we predict the number of crimes? Is there much difference among the test errors resulting from these four approaches?