STAT 534: Homework 3

Fall 2021

Due: Wednesday, September 29, 12:30 pm

1. In this problem, we will generate simulated data, and will then use this data to perform variable selection.
   (a) Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of the same length. Then generate a response vector Y according to the model
   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$
   where $\beta_0, ..., \beta_3$ are constants of your choice.
   (b) Given the predictors $X, X^2, ..., X^{10}$, perform best subset selection in order to choose the best model. What is the best model obtained according to C_p, AIC, BIC and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained.
   (c) Repeat (b), using forward selection. How does your answer compare to the results in (b)?
   (d) Now fit a lasso model and use cross-validation to select the optimal values of λ. Create plots of the cross-validation error as a functionf of λ. Report the resulting coefficient estimates, and discuss the results obtained
2. (Excerpted from Problem 11.6 in textbook) The ability to get a good night's sleep is correlated with many positive health outcomes. Use the *NHANES* data set from the *NHANES* package to predict *SleepHrsNight*. Check the R help document for detailed information about the data set.
   (a) First separate the data set at random into 75% training and 25% testing sets.
   (b) Select your own predictors, and create plots or summary tables to explore the variables.
   (c) Build the following models using the training set with your predictors of choice:
   - Multiple linear regression
   - Ridge regression
   - LASSO regression

   (d) Compare the effectiveness of each model on training vs. testing data.
   (e) Choose one best model and interpret the results. What have you learned about people's sleeping quality?