STAT 534: Homework 4

Fall 2021

Due: Friday, October 8

1. Using the *Boston* dataset from the *MASS* package, the goal is to predict the crime rate by the other variables.
   (a) Create bivariate plots to explore relations between variables. Comment on your observations. (Hint: you may use ggpair() and remember to factorize any categorical variables.)
   (b) Log-transform the crime rate and repeat part (a).
   (c) Create a correlation matrix of all the continuous variables and make comments. (Hint: you may use ggcorr().)
   (d) Split the data into training and testing subsets.
   (e) Build the following models using the training set:
      - multiple linear regression
      - principal components regression (indicate how many principal components are selected)
      - partial least squares (indicate how many directions are selected)
      - lasso
   (f) Compare the effectiveness of each model on training vs. testing data. Which model is the best?
   (g) Refit the principal components regression model and the lasso model to the entire dataset. Comment on the differences between the two methods. (Hint: also pay attention to highly correlated variables that you found in part (c).)
   (h) Refit the partial least squares model to the entire dataset, and compare with the principal components regression model.