CSCI E-106:Assignment 10

Due Date: December 7, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Problem 1

Refer to the Cement Composition Data. The variables collected were the amount of tricalcium aluminate (X_1), the amount of tricalcium silicate (X_2), the amount of tetracalcium alumino ferrite (X_3), the amount of dicalcium silicate (X_4), and the heat evolved in calories per gram of cement (Y). (25 points, 5 points each)

a -) Fit regression model for four predictor variables to the data. State the estimated regression function. (5 pt)

b-) Fit a ridge regression model and find the best λ.

c-) Fit a ridge regression model and find the best λ.

d-) Fir an elastic net model

e-) Compare all models built in part a, b, c, and d and choose the optimal model and explain your rationale

Problem 2

Refer to the Prostate cancer data set in the problem 3 in the Homework 9. Select a random sample of 65 observations to use as the model-building data set. (15 points, 5 each)

a-) Develop a regression tree for predicting PSA. Justify your choice of number of regions (tree size), and interpret your regression tree.

b-) Assess your model's ability to predict and discuss its usefulness to the oncologists.

c-) Compare the performance of your regression tree model with that of the best regression model obtained in the problem 3 in the Homework 9. Which model is more easily interpreted and why?

Problem 3

Refer to the Prostate cancer data set in the problem 3 in the Homework 9. Select a random sample of 65 observations to use as the model-building data set. (15 points, 5 each) a-) Develop a neural network model for predicting PSA. Justify your choice of number of hidden nodes and penalty function weight and interpret your model.

b-) Assess your model's ability to predict and discuss its usefulness to the oncologists.

c-) Compare the performance of your neural nerwork model with that of the best regression model obtained in the problem 3 in the Homework 9. Which model is more easily interpreted and why?

Problem 4

Refer to the Advertising Agency Data. Monthly data on amount of billings (Y, in thousands of constant dollars) and on number of hours of staff time (X, in thousand hours) for the 20 most recent months follow. A simple linear regression model is believed to be appropriate. but positively autocorrelated error terms may be present. (20 points 5 each)

a-) Fit a simple linear regression model by ordinary least squares and obtain the residuals. Conduct a formal test for positive autocorrelation using $\alpha = .01$.

b-) Use a Cochrane-Orcutt procedure to estimate the model and test if the autocorrelation remains after the first iteration

c-) Restate the estimated regression function obtained in part (b) in terms of the original variables. Also obtain $s(b_0)$ and $s(b_1)$. Compare the estimated regression coefficients obtained.

d-)Staff time in month 21 is expected to be 3.625 thousand hours. Predict the amount of billings in constant dollars for month 21, using a 99 percent prediction intervaL Interpret your interval.

Problem 5

Refer to the Advertising Agency Data and Problem 4. (25 points, 5 points each)

a-) Use the Hildreth-Lu procedure to obtain a point estimate of the autocorrelation parameter. Do a search at the values $\rho = .1, .2, \ldots\ , 1.0$ and select from these the value of ρ that minimizes SSE. Based on your model, obtain an estimate of the transformed regression function.

b-) Use the first difference procedure to obtain a point estimate of the autocorrelation parameter. Based on your model, obtain an estimate of the transformed regression function.

c-) Test whether any positive autocorrelation remains in the transformed regression model for both part a and b; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

d-) Which method would you choose? Explain your rationale

e-) For the selected model in part d. Staff time in month 21 is expected to be 3.625 thousand hours. Predict the amount of billings in constant dollars for month 21, using a 99 percent prediction interval. Interpret your interval.