Homework 5

P8130 Fall 2022

Due: December 5, 2022 at midnight Eastern

P8130 Guidelines for Submitting Homework

- Your homework must be submitted through Courseworks. No email submissions!
- Only one PDF file should be submitted, including all derivations, graphs, output, and interpretations. When handwriting is allowed (this will be specified), scan the derivations and merge ALL PDF files (http: //www.pdfmerge.com/).
- You are encouraged to use R for calculations, but you must show all mathematical formulas and derivations. Please include the important parts of your R code in the PDF file but also submit your full, commented code as a separate R/RMD file.
- To best follow these guidelines, we suggest using Word (built in equation editor), R Markdown, Latex, or embedding a screenshot or scanned picture to compile your work.

DO NOT FORGET: You are encouraged to collaborate on homeworks, explain things to each other, and test each other's knowledge. But Do NOT hand out answers to someone who has not done any work. Everyone ought to have ideas about the possible answers or at least some thoughts about how to probe the problem further. Write your own solutions!

Problem 1 (38 points)

R dataset `state.x77` from `library(faraway)` contains information on 50 states from 1970s collected by US Census Bureau. The goal is to predict ‘life expectancy’ using a combination of remaining variables.

a) Provide descriptive statistics for all variables of interest (continuous and categorical) – no test required.

b) Examine exploratory plots, e.g., scatter plots, histograms, box-plots to get a sense of the data and possible variable transformations. (Be selective! Even if you create 20 plots, you don’t want to show them all). If you find a transformation to be necessary or recommended, perform the transformation and use it through the rest of the problem.

c) Use automatic procedures to find a ‘best subset’ of the full model. Present the results and comment on the following:

- Do the procedures generate the same model?
- Are any variables a close call? What was your decision: keep or discard? Provide arguments for your choice. (Note: this question might have more or less relevance depending on the ‘subset’ you choose).
- Is there any association between ‘Illiteracy’ and ‘HS graduation rate’? Does your ‘subset’ contain both?

d) Use criterion-based procedures to guide your selection of the ‘best subset’. Summarize your results (tabular or graphical).

e) Use the LASSO method to perform variable selection. Make sure you choose the “best lambda” to use and show how you determined this.

f) Compare the ‘subsets’ from parts c, d, and e and recommend a ‘final’ model. Using this ‘final’ model do the following:

- Check the model assumptions.
- Test the model predictive ability using a 10-fold cross-validation.

g) In a paragraph, summarize your findings to address the primary question posed by the investigator (that has limited statistical knowledge).