P8130 Fall 2019: Biostatistical Methods I

Homework 5

Due, Dec 10 @ 5:00pm

P8130 Guidelines for Submitting Homework

Your homework should be submitted only through CourseWorks. No email submissions!

All derivations, graphs, output and interpretations to each section of the problem(s) must be included in the PDF (not the code), otherwise it will not be graded.

Only 1 PDF file should be submitted. When derivations were required and handwriting was allowed, scan the derivations and merge ALL PDF files (http://www.pdfmerge.com/) into a single one.

We are encouraged to use R for calculations, but you still have to show the mathematical formulae. Also, make sure to also submit your commented code as a separate R/RMD file.

DO NOT FORGET:

You are encouraged to collectively look for answers, explain things to each other, and use questions to test each other 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 (40p)

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. (3p)
b) Examine exploratory plots, e.g., scatter plots, histograms, box-plots to get a sense of the data and possible variable transformations. (2p)
(Be selective! Even if you create 20 plots, you don't want to show them all).
c) Use automatic procedures to find a 'best subset' of the full model. Present the results and comment on the following (10p):
   - Do the procedures generate the same model?
   - Is there any variable 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) (10p).
e) Compare the two 'subsets' from parts c) and d) and recommend a 'final' model. Using this 'final' model do the following (10p):
   - Identify any leverage and/or influential points and take appropriate measures.
   - Check the model assumptions.
   - Test the model predictive ability using a 10-fold cross-validation (10 repeats).
f) In a paragraph, summarize your findings to address the primary question posed by the investigator (that has limited statistical knowledge) (5p).

Problem 2 (25p)

Dataset 'CommercialProperties.csv' are taken from 81 suburban commercial properties that are the best located and most attractive for five geographic areas. The variables of interest are: age of the property, taxes, vacancy rates, total square footage and rental rates (monthly).

a) Fit a model regressing rental rates (monthly) on all the other variables. Summarize your findings focusing on the significance of the predictors and the overall fit. (5p)
b) From this point forward, let us consider only the significant predictors. Create scatter plots for each of the predictors and the outcome 'rental rates'. Comment on the form of these relationships. (3p)
c) Fit a multiple linear regression containing only the significant predictors. (2p)
d) Use the model from part 3, keep all the significant predictors in their initial form, but let us *play* with the 'age of the property'. (10p)
    - Given the relationship between 'age' and 'rental rates' (part 2), would you use higher order terms to fit 'age'? Should you use centered 'age'?
    - Or maybe you want to keep it 'linear', and try 'segments and knots' instead? If using piecewise linear, comment on your choice of knot(s).
    - Try one or both approaches and recommend a model.
e) Test whether the recommended model from part 4 is 'superior' to the model in part 3. Write and comment on the 'final' model that you chose. (5p)