P8130 Fall 2018: Biostatistical Methods I

Homework 5

Due, Dec 3rd @ 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.

You are encouraged to use R for calculations, but you still have to show the mathematical formulae. For some problems, you can use R ONLY and the problem will specifically state that.

Also, make sure you include your commented code at the end of the document (PDF) or attach a separate R 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

You are NOT supposed to 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!

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.

1. Explore the dataset and generate appropriate descriptive statistics and relevant graphs for all variables of interest (continuous and categorical) – no test required. Be selective! Even if you create 20 plots, you don't want to show them all.(5p)

2. Use automatic procedures to find a 'best subset' of the full model. Present the results and comment on the following (10p):
   a) Do the procedures generate the same model?
   b) 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).
   c) Is there any association between 'Illiteracy' and 'HS graduation rate'? Does your 'subset' contain both?

3. Use criterion-based procedures studied in class to guide your selection of the 'best subset'. Summarize your results (tabular or graphical) (10p).

4. Compare the two 'subsets' from parts 2 and 3 and recommend a 'final' model. Using this 'final' model do the following (10p):
   a) Identify any leverage and/or influential points and take appropriate measures.
   b) Check the model assumptions.

5. Using the 'final' model chosen in part 4, focus on MSE to test the model predictive ability (25p):
   a) Use a 10-fold cross-validation (10 repeats).
   b) Experiment a new, but simple bootstrap technique called "residual sampling". Several references can be found on this topic, but you can just follow these steps:
      i) Perform a regression model with the original sample; calculate predicted values ($\widehat{Y}_i$) and residuals (e_i)
      ii) Randomly resample the residuals (with replacement), but leave the X values and ($\widehat{Y}_i$) unchanged.
      iii) Construct new Y_i^* values by adding the original predicted values to the bootstrap residuals, $Y_i^* = \widehat{Y}_i + e_i^*$, e_i^* are re-sampled residuals from ii)
      iv) Regress Y_i^* on the original X variable(s).
      v) Repeat steps (ii) – (iv) 10 times and 1,000 times
      vi) Summarize the MSE for all repetitions.
   c) In a paragraph, compare the MSE values generated by the two methods a) and b). Briefly comment on the differences and your recommendation for assessing model performance.

Ref: Efron, B., Tibshirani. R.,(1993). Introduction to the Bootstrap. Chapman and Hall, London.