Statistical and Machine Learning (Spring 2018)

Mini Project 4

Instructions:

- Due date: March 29, 2018.
- Total points = 40.
- Submit a typed report.
- Submit only one report per group, and include a description of the contribution of each member.
- It is OK to discuss the project with other students in the class (even those who are not in your group), but each group must write its own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- **Interpret all results and justify all conclusions**.
- You must use the following template for your report:

  Mini Project #
  Name
  Names of group members
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. Consider the Prostate Cancer data available in `ElemStatLearn` package as `prostate`. Look at the data description using `?prostate`. You are also encouraged to read about these data on pages 3-4 and 49-51 and in Section 3.4 of the book *The Elements of Statistical Learning*, 2nd edition, available from `https://web.stanford.edu/~hastie/ElemStatLearn/`. We will take `lpsa` as the response and the following 8 variables are predictors: `lcavol`, `lweight`, `age`, `lbph`, `svi`, `lcp`, `gleason`, and `pgg45`. Note that the book divides the dataset into training and test sets. However, we will take all the observations as training data. Further, for any fitting method (see below) that has a complexity parameter, choose the best value for the parameter using estimated test error rates based on 10-fold cross-validation.

   (a) Perform an exploratory analysis of the data.
   (b) Fit a linear model using the usual least squares.
   (c) Repeat (b) using best-subset selection.
   (d) Repeat (b) using ridge regression.
   (e) Repeat (b) using lasso.
   (f) Repeat (b) using PCR.
   (g) Repeat (b) using PLS.
   (h) Make a summary of the parameter estimates and test error rates as in Table 3.3 on page 63 of the aforementioned book. Compare the results. Which method would you recommend? Justify your conclusions.