Statistical Methods for Data Science (Fall 2019)

Project 6

Instructions:

- Due date: Dec 6, 2019.
- Total points = 20.
- Submit a typed report.
- You can work on the project either individually or in a group of no more than two students. In case of the latter, submit only one report for the 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.
- Start early, do a good job.
- You must use the following template for your report:

  Project #
  Name
  Names of group members (if applicable)
  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 dataset available on eLearning as `prostate_cancer.csv`. It consists of data on 97 men with advanced prostate cancer. A description of the variables is given in Figure 1. We would like to understand how PSA level is related to the other predictors in the dataset. Note that `vesinv` is a qualitative variable. You can treat `gleason` as a quantitative variable.

   Build a "reasonably good" linear model for these data by taking PSA level as the response variable. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions. In case a transformation of response is necessary, try the natural log transformation. Use the final model to predict the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors are at the most frequent category.

header	name	description
subject	ID	1 to 97
psa	PSA level	Serum prostate-specific antigen level (mg/ml)
cancervol	Cancer Volume	Estimate of prostate cancer volume (cc)
weight	Weight	prostate weight (gm)
age	Age	Age of patient (years)
benpros	Benign prostatic hyperplasia	Amount of benign prostatic hyperplasia (cm^2)
vesinv	Seminal vesicle invasion	Presence (1) or absence (0) of seminal vesicle invasion
capspen	Capsular penetration	Degree of capsular penetration (cm)
gleason	Gleason score	Pathologically determined grade of disease (6, 7 or 8)

Figure 1: List of variables in the prostate cancer data

.