Statistical and Machine Learning (Spring 2019)

Mini Project 2

Instructions:

- Due date: Feb 13, 2019.
- Total points = 20.
- Submit a typed report.
- It is OK to discuss the project with other students in the class, but each student must write their own code and answers. If your 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.
- You must use the following template for your report:

  Mini Project #
  Name
  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.

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

.

(a) Perform an exploratory analysis of data.

(b) Is `psa` appropriate as a response variable or a transformation is necessary? In case a transformation of response is necessary, try the natural log transformation or some other transformation and use it for the rest of this problem.

(c) Do part (a) of Exercise 15 in Chapter 3 for these data.

(d) Do part (b) of Exercise 15 in Chapter 3 for these data.

(e) Build a "reasonably good" multiple regression model for these data. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions.

(f) Write the final model in equation form, being careful to handle qualitative predictors (if any) properly.

(g) 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 (if any) are at the most frequent category.