Statistical Methods for Data Science
Mini Project 5

Consider the prostate cancer dataset available on eLearning. It consists of data on 97 men with advanced prostate cancer. Following is a description of the variables:

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)

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 and also to distinguish between quantitative and qualitative variables. Use the final model to predict the PSA level for a patient whose predictors are at the sample means of the variables.

<u>Instructions:</u>

- Due date: Thursday, April 20.
- Total points = 25.
- 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.
- Do a good job.
- You must use the following template for your report:

Mini 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.