Statistical Methods for Data Science
Mini Project 6

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)

Take PSA level as the response variable, and in this exercise, we will focus only on the predictors that are **quantitative**.

(a) Make scatterplots of PSA level against each quantitative variable. Comment on what you see. Based on your assessment, choose one **quantitative** variable that you think may be used most effectively to predict PSA level.

(b) Fit a simple linear regression model and carry out regression diagnostics. The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If an assumption is not met, attempt to remedy the situation. Comment on the fit of the final model using appropriate tests and statistics.

(c) Use the final model to predict the PSA level for a patient whose predictor variable value is at the sample median of the variable.

2 Bonus points will be given for good, neat work.

Instructions:

- Due date: Tuesday, Dec 6.
- Total points = 20
- Submit a typed report and include any relevant plots. Justify each step in the analysis. For example, plot/test indicate …. assumption may be violated. So, …. is attempted, which shows ….
- 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)
  Contributions of group members
  Answers and justifications for each exercise

  Provide the R codes in an appendix. 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.