Assignment

For this assignment you will use the dataset “prostate” (prostate2.Rdata). The dataset contains data about prostate cancer patients with information on the size of the prostate, the age of the patient, a blood marker (lpsa) and so on. The response variable is a score (Cscore) on the progression of the cancer after detailed study of the tumor pathology.

1. Study and describe the predictor variables. Do you see any issues that are relevant for making predictions?
2. Generate your best linear regression model using only linear effects. Are there any indications that assumptions underlying inferences with the model are violated? Evaluate the effect of any influential point, or outlier.
3. Make an appropriate LASSO model, with the appropriate link and error function, and evaluate the prediction performance. Do you see evidence that over-learning is an issue?
4. Look at the coefficient for “lcavol” in your LASSO model. Does this coefficient correspond to how well it can predict Cscore? Explain your observation.
5. Fit your best model with *appropriate* non-linear effects. Report a comparison of performance to LASSO and your model reported under question 2. Explain what you find, and indicate relevant issues or limitations of your analysis.