Prostate Case Study

Instructor: A. Chronopoulou

Prostate Data Set

Serum prostate-specific antigen (PSA) is a well-established screening test for prostate cancer. The oncologists wanted to examine the *correlation* between `level of PSA` and a number of clinical measures for 97 men with advanced prostate cancer who were about to undergo a radical prostatectomy. These measures are:

X_1: cancer volume
X_2: patient age
X_3: amount of bening prostatic hyperplasia
X_4: seminal vesicle invasion
X_5: capsular penetration
X_6: Gleason score

Our goal is to fit a model to describe the association between the `level of PSA` and the clinical measures (potential predictors X_1-X_6). The data can be found in the `prostate.txt` data set on Moodle.

Remark: The data set does not have labels. The columns $V1 - V9$ correspond (in order) to `ID`, `psalevel`, `cancervolume`, `prostateweight`, `age`, `hyperplasia`, `svi`, `capsular`, `gleason`.

Instructions

You should perform a *full* regression analysis using the tools we have already discussed in class. It should include evaluating the statistical significance of the variables in the model, removing variables that do not contribute to explaining the variation in the response, check unusual observations and model assumptions, perform remedial measures -if necessary.

Groups

The case study should be done in a group of 2–3 students. You are free to choose your own group. If you do not have a group in mind, please use this form to let me know and I will randomly assign you to a group.

Deliverables

The case study should be submitted on Gradescope as a group (only once case study per group). You should submit:

(1) a **PDF** file containing a 2-3 executive summary of the analysis. You need to make sure that your report is professionally and clearly written, addressed to someone who *knows statistics*. You should also include a concluding paragraph where you should state your conclusions in layman's terms. Any necessary plots or **R** output should be attached in an *appendix*. You should include no **R** code in the summary.

(2) an **R Markdown** and corresponding **HTML** file with comments with all the R code that you built to analyze the data set.

A rubric on which the grading of the case study will be based is posted on Moodle for your reference.

Deadline: Submit **one case study report per group** on Gradescope by **Tuesday, March 29** @ 11.59PM.

Learning Objectives

By the end of this case study, you will

1. enhance your skills in using R for the purpose of statistical analysis of a data set.
2. independently apply the regression in a real-world problem.
3. evaluate the applicability of the regression model.
4. draw reasonable conclusions, and make decisions about the initially stated research questions.
5. interpret your statistical outcomes using plain English.
6. demonstrate your team collaboration skills.