ADA2: HW2 (Chaps 03 and 10)

Due Thursday 02/14/2019

Prostate-specific antigen (PSA)

A university medical center urology group (Stamey, *et al.*, 1989) was interested in the association between a prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies.

Following variable PSA (Serum prostate-specific antigen level (ng/ml)) are the measures: cancer volume (Estimate of prostate cancer volume (cc)), prostate weight (gm), patient age, the amount of benign prostatic hyperplasia (cm2) hyperplasia, seminal vesicle invasion with 1 if yes, 0 if no, capsular penetration (cm), and gleason score (Pathologically determined grade of disease (6,7,8), higher indicates worse prognosis).

Background: Until recently, PSA was commonly recommended as a screening mechanism for detecting prostate cancer. To be an efficient screening tool it is important that we understand how PSA levels relate to factors that may determine prognosis and outcome. The PSA test measures the blood level of prostate-specific antigen, an enzyme produced by the prostate. PSA levels under 4 ng/mL (nanograms per milliliter) are generally considered normal, while levels over 4 ng/mL are considered abnormal (although in men over 65 levels up to 6.5 ng/mL may be acceptable, depending upon each laboratorys reference ranges). PSA levels between 4 and 10 ng/mL indicate a risk of prostate cancer higher than normal, but the risk does not seem to rise within this six-point range. When the PSA level is above 10 ng/mL, the association with cancer becomes stronger. However, PSA is not a perfect test. Some men with prostate cancer do not have an elevated PSA, and most men with an elevated PSA do not have prostate cancer. PSA levels can change for many reasons other than cancer. Two common causes of high PSA levels are enlargement of the prostate (benign prostatic hypertrophy (BPH)) and infection in the prostate (prostatitis).

Goal

A goal here is to build a multiple regression model to predict PSA level `PSA` from a subset of the predictors: the cancer volume, Prostate weight, Age, Benign prostatic and Capsular penetration.

(1) Looking at the data. Plot scatter plot matrix, correlation matrix of Y and a subset of predictors, and briefly comment on them.

```r
fn.data <- "http://statacumen.com/teach/ADA2/homework/ADA2_HW_03_psa.txt"
psa <- read.table(fn.data, header=TRUE)

# remove the variables we don't want to use for this assignment
psa <- subset(psa, select = names(psa)[-c(1,7,9)])

names(psa)
```

```
[1] "PSA"                          "cancer_volume"
[3] "prostate_weight"              "patient_age"
[5] "benign_prostatic_hyperplasia" "capsular_penetration"
```

```r
# simplify column names
colnames(psa) <- c("PSA", "v", "wt", "age","benigh","capsular")
```

(2) Fit the full aditive model, check model assumptions.

(3) Check Y outliers, X outliers, and influential data points.

(4) Are there any interaction terms need to be included?

(5) Do transformation on Y or Xs or both if needed. After that, check model assumptions again. (Please only include transformations that make model assumptions met.)

(6) Fit an appropriate full model and perform backward selection.

(7) Fit an appropriate full model and perform best subset selection using adjusted R^2, Cp and BIC criteria.

(8) From the above model selection methods, which model seems appropriate? Check model assumptions for this model. Check VIF.

(9) For the model you derived in (8), check Y outliers, X outliers, and influential data points. If you identify some outliers, can you remove them?

(10) Suggest a final model for use.