Linear models, part II **lp2 2010**

Lab 3: Model Selection

A university medical center urology group 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. A random subset of these data (65 observations) are available on the course website. Each line of data set has an identification number and provides information on 8 other variables for each person. I have also posted a test data set with 32 observations, where PSA is unknown. Note, a higher Gleason score indicates worse prognosis.

	Variable name	Description
1	Identification number	1-97
2	PSA level	Serum prostate-specific antigen level (mg/ml)
3	Cancer volume	Estimate of prostate cancer volume (cc)
4	Weight	Prostate weight (gm)
5	Age	Age of patient (years)
6	Benign prostatic	Amount of benign prostatic hyperplasia (cm2) hyperplasia
7	Seminal vesicle invasion	Presence or absence of seminal vesicle invasion: 1 if yes; 0 o.w.
8	Capsular penetration	Degree of capsular penetration (cm)
9	Gleason score	Pathologically determined grade of disease (6,7,8)

1 Modeling the Prostate data

PSA is commonly used as a screening mechanism for detecting prostate cancer. However, 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).

Some of the variable names may look unfamiliar to you - please use resources on the web if you feel unsure as to what these variables measure. The section above is based on excerpts from Wikipedia.org, and you can also find variable definitions at http://www.prostate-cancer.org/resource/glossary.html. For example, a large tumor may invade surrounding tissue and penetrate the wall of the prostate (variable 7 and 8). Also, benign hyperplasia is associated with higher PSA levels, but is non-cancerous (variable 6).

1.1 Exploring the data

Use `read.table` to upload the data into R. The goal of this lab is to develop a predictive model for PSA. The data set is rich, but also complex. You need to pay attention to the variable types in the data. There are clearly indicator variables or qualitative variables such as variable 7. However, variable 8 contains a hidden indicator - either capsular penetration has been detected and we have a positive measurement, or not - in which case our measurement is 0 exactly. You should look for the main variables that relates to PSA in an additive model, and also consider the inclusion of interactions. That being said - remember that the overall goal is prediction so its often better to keep things simple.

I expect you to explore the data in full. Come up with an initial model for which all basic assumptions have been verified. Remember to check the residual plots, the normal-normal plots, and leverage and Cooks D plots for outliers.

1.2 Model Selection

Once you are satisfied that you have a working model, consider model selection via F - test, AIC, BIC or Cp. You can use all-subset or backward selection. Compare at least two selection criteria in your report. Do you identify the same model using both criteria? Please provide a table of models selected with different criteria. It is also a good idea to report the trace of the selection (i.e. the order in which the variables were dropped or added). While our ultimate goal is prediction, try to come up with a reasonably interpretable model for the data. I also want you to comment on the following; PSA is used as a screening tool to detect prostate cancer, and especially detect cases with poor prognosis. Does your model assist in increasing our understanding of how PSA relates to clinical measures of prognosis? Perhaps certain variable relationships are more important than others?...

1.3 Selection Stability

Create a new data set by selecting observations at random. Perform the model selection task (choose one criterion) on this new data set. Repeat this exercise several times and examine the variability of model selection. Is the same model always selected? Is there a set of variables that are always selected. Summarize the (in)stability of model selection and discuss. You can either choose to randomly select from all observations to create a new data set of equal size to the original (meaning some observations will appear more than once), or by choosing a subset of observations. Please make sure to specify what you did.

1.4 Class competition

I actually have the PSA values for the 32 observations in the test data. As part of the lab, each of you will produce a vector of 32 predictions and submit to me. The winning model will be announced in the class following the lab due date.

2 Writing the report

Write a data analysis report with an introduction, a results section (selected models, validation), and a summary/conclusions. Tables and figures should go in the main body of the text, and captions should be informative and complete. Dont intermix code and text in your report.

Focus on answering the following questions; (1) Can you interpret your selected model - what have you learnt about the relationship between PSA and clinical measures of prognosis?l (2) Which clinical measures of prognosis are related to PSA? Does your analysis support the use of PSA as an early screening mechanism? Does your analysis support the use of PSA as a screening mechanism for patients with poor prognosis? Does your analysis support the use of PSA as a screening mechanism at all?;