Stat 6242
Assignment 5, Due Nov. 30

November 22, 2016

1. (30 pts) The *bodyfat* data set contains estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for men. Accurate measurement of body fat is inconvenient and costly (see preamble of the data set) and it is desirable to have easy methods of estimating body fat that are not inconvenient or costly.

   A variety of popular health books suggest that the readers assess their health

   at least in part

   by estimating their percentage of body fat In Bailey (1994), for instance,

   the reader can estimate body fat from tables using their age and various skin-fold measurements obtained by using a caliper. Other texts give predictive equations for body fat using body circumference measurements; e.g. abdominal circumference and/or skin-fold measurements

   The file *bodyfat* contains introductory information that you should delete in order to import the data in R. The variables provided in the dataset

   from left to right are:

   - Density determined from underwater weighing
   - Percent body fat from Sir's (1956) equation
   - Age (years)
   - Weight (lbs)
   - Height (inches)
   - Neck circumference (cm)
   - Chest circumference (cm)

- Abdomen circumference (cm)
- Hip circumference (cm)
- Thigh circumference (cm)
- Knee circumference (cm)
- Ankle circumference (cm)
- Biceps extended circumference (cm)
- Forearm circumference (cm)
- Wrist circumference (cm)

Analyze these data to produce predictive equations for lean body weight using multiple linear regression (model selection), regression trees and additive models (model selection). Build your model using the first 143 of the 252 cases and the rest to assess the predictive ability of your models. Do the three methods use the same variables to make the prediction?

2. (30 pts) **Classification:** When a bank receives a loan application, based on the applicants profile the bank has to make a decision regarding whether to go ahead with the loan approval or not. Two types of risks are associated with the banks decision:

   - If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank
   - If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank

   Objective of Analysis: Minimization of risk and maximization of profit on behalf of the bank.

   To minimize loss from the banks perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicants demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

The German Credit Data contain data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. The response is binary (Good credit risk or Bad, $Creditability = 1$ if credit worthy and 0 otherwise). A predictive model developed on these data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.

Build your classification models using the training data (Training50.csv), all the other variables as predictors and

1. Logistic regression
2. Discriminant Analysis
3. Classification Trees

Assess the predictive accuracy of your models using the test data (Test.csv). Which method results in the model with best predictive ability?

Use the following predictors in your analysis:

1. Account Balance: No account (1), None (No balance) (2), Some Balance (3)
2. Payment Status: Some Problems (1), Paid Up (2), No Problems (in this bank) (3)
3. Savings/Stock Value: None, Below 100 DM, [100, 1000] DM, Above 1000 DM
4. Employment Length: Below 1 year (including unemployed), [1, 4), [4, 7), Above 7
5. Sex/Marital Status: Male Divorced/Single, Male Married/Widowed, Female
6. No of Credits at this bank: 1, More than 1
7. Guarantor: None, Yes
8. Concurrent Credits: Other Banks or Dept Stores, None
9. Purpose of Credit: New car, Used car, Home Related, Other