Financial Time Series I and Methods of Statistical Prediction

Project 2: Logistic Regression and Model Selection
Due Date: January 9th, 2003

In this project, you are asking to use a logistic regression to do prediction and get hold of the idea of model selection. In addition, you are also asked to learn additional *R*-commands to manipulate the data set. In particular, how do we convert a quantitative variable to a qualitative variable. All commands have been given but you should read the corresponding help files for your own benefit.

Hosmer and Lemeshow (*Applied Logistic Regression*, 1989) give a dataset on 189 births at the Baystate Medical Center, Springfield, Massachusetts during 1986 to attempt to identify which factors contributed to an increased risk of low birth weight infants. The objective of this study is to build a probability model to predict the probability of low birth weight before the birth. (This model can be used in prevention and policy making.) Information was recorded for 189 women of whom 59 had low birth weight infants. This data set is named *birthwt* in the package of *boot.* Read help file of the data frame *birthwt* on the explanation of all variables. In this original data, most variables are being treated as continuous variable. In the analysis, we will convert some of them to categorical variables. In particular, low, race, smoke, ptl, ht, ui, and ftv will be treated as categorical variable. It will be done according to the following steps.

1. Reduce ftv to three levels.
   - attach(birthwt)
   - race < − factor(race,labels=c("white", "black", "other"))
   - table(ftv); ftv < − factor(ftv); ftv
   - levels(ftv)[-(1:2)] < − "2+"; table(ftv)
2. Convert ptl to two levels and name the new variable as ptd.
   - table(ptl); ptd < − factor(ptl > 0)
3. Creat a new data frame *bwt.*
   - bwt < − data.frame(low=factor(low), age, lwt, race, smoke=(smoke > 0), ptd, ht=(ht > 0), ui=(ui > 0),ftv)
   - bwt
4. Clean up data.
   - detach("birthwt")
   - rm(race, ptd, ftv)

Question 1: Do the above commands and give a brief explanation on the meaning of those commands.

Question 2: Use the following *R* commands and again give a brief explanation on the meaning of those commands. In particular, explain the specification of the model.

- model fitting
  - birthwt.glm < − glm(low ∼ ., family=binomial, data=bwt)
  - summary(birthwt.glm, correlation=F)
- model selection with AIC
  - birthwt.step < − step(birthwt.glm, trace=F); birthwt.step$anova

Question 3: Repeat the steps in Question 2 and consider all models include pairwise interactions.

Question 4: Use the probability model obtained in Questions 2 and 3 to predict whether the actual live birth weight is below the threshold at 2.5 kilograms. Find out the prediction error.

Question 5: Repeat the above procedure (with the two model being chosen) with cross-validation to give prediction error again.