MSG500/MVE190
Linear Models - Lab 3

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

November 28, 2011

1 Introduction

In this lab you will continue to work with your own random subset of the birthweight data. The goals of the lab are

1. Expand the models from lab 2 to include interactions if they are supported by the data.
2. Compare different model selection criteria.

2 Interactions

Explore the birthweight data. Are there any interactions in the data, e.g. does the fact that the mother smokes alter the relationship between the birthweight and the other variables? What about other variable interactions? Make sure to explain the meaning of the interactions you consider? Are they significant? What does automated model selection do?

3 Model selection

Use backward selection, cross-validation and Cp, AIC and BIC to select a model for the birthweight data (including interactions). Do selected models differ when you use the various criteria? Summarize your findings.

Repeat the model selection several times on different random splits of the data. What do you see?

4 Model selection function

Here is a function I wrote to help you out. Save the `randomsplits.R` function in your lab directory. The source it into R `source('randomsplits.R')`. Now you can use this as a function.
You call the function as follows:
`results<-randomsplits(baby,1,.75,100)`
if you want to apply the random splits with 3/4 for training, using `bwt` as the response (since it is column 1) and performing 100 random splits.

```
> randomsplits <- function(dataset, yind, frac = 0.5, K = 100) {
+     modsizecp <- rep(0, K)
+     modsizeaic <- rep(0, K)
+     modsizebic <- rep(0, K)
+     PEcpK <- rep(0, K)
+     PEaicK <- rep(0, K)
+     PEbicK <- rep(0, K)
+     modselcp <- rep(0, dim(dataset)[2] - 1)
+     modselaic <- rep(0, dim(dataset)[2] - 1)
+     modselbic <- rep(0, dim(dataset)[2] - 1)
+     for (kk in (1:K)) {
+         if (frac < 1) {
+             ii <- sample(seq(1, dim(dataset)[1]), round(dim(dataset)[1] *
+                 frac))
+             yy <- dataset[ii, yind]
+             xx <- as.matrix(dataset[ii, -yind])
+             yyt <- dataset[-ii, yind]
+             xxt <- as.matrix(dataset[-ii, -yind])
+         }
+         if (frac == 1) {
+             yy <- dataset[, yind]
+             yyt <- yy
+             xx <- dataset[, yind]
+             xxt <- xx
+         }
+         library(leaps)
+         rleaps <- regsubsets(xx, yy, int = T, nbest = 1, nvmax = dim(dataset)[2],
+             really.big = T, method = c("ex"))
+         cleaps <- summary(rleaps, matrix = T)
+         tt <- apply(cleaps$which, 1, sum)
+         BIC <- length(yy) * log(cleaps$rss/length(yy)) + tt *
+             log(length(yy))
+         AIC <- length(yy) * log(cleaps$rss/length(yy)) + tt *
+             2
+         cpmod <- cleaps$which[cleaps$cp == min(cleaps$cp), ]
+         aicmod <- cleaps$which[AIC == min(AIC), ]
+         bicmod <- cleaps$which[BIC == min(BIC), ]
+         mmcp <- lm(yy ~ xx[, cpmod[2:length(cpmod)] == T])
+         mmaic <- lm(yy ~ xx[, aicmod[2:length(aicmod)] == T])
+         mmbic <- lm(yy ~ xx[, bicmod[2:length(bicmod)] == T])
+         PEcpK[kk] <- sum((yyt - cbind(rep(1, dim(xxt)[1]), xxt[,
+             cpmod[2:length(cpmod)] == T]) %*% mmcp$coef)^2)/length(yyt)
+         PEaicK[kk] <- sum((yyt - cbind(rep(1, dim(xxt)[1]), xxt[,
+             aicmod[2:length(aicmod)] == T]) %*% mmaic$coef)^2)/length(yyt)
+         PEbicK[kk] <- sum((yyt - cbind(rep(1, dim(xxt)[1]), xxt[,
+             bicmod[2:length(bicmod)] == T]) %*% mmbic$coef)^2)/length(yyt)
+         modsizecp[kk] <- sum(cpmod)
+         modsizeaic[kk] <- sum(aicmod)
+         modsizebic[kk] <- sum(bicmod)
+         modselcp[cpmod[2:length(cpmod)] == T] <- modselcp[cpmod[2:length(cpmod)] ==
+             T] + 1
+         modselaic[aicmod[2:length(cpmod)] == T] <- modselaic[aicmod[2:length(cpmod)] ==
+             T] + 1
+         modselbic[bicmod[2:length(cpmod)] == T] <- modselbic[bicmod[2:length(cpmod)] ==
+             T] + 1
+     }
+     modseltabs <- cbind(names(dataset)[-12], modselcp, modselaic,
+         modselbic)
+     return(list(PEcpK = PEcpK, PEaicK = PEaicK, PEbicK = PEbicK,
+         modsizecp = modsizecp, modsizeaic = modsizeaic, modsizebic = modsizebic,
+         modseltabs = modseltabs))
+ }
```