MSG500/MVE190
Linear Models - Lab 3

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

December 3, 2012

1 Introduction

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

1. expand the models from lab 2 to include interactions if they are supported by the data.
2. compare different model selection criteria.
3. If time permits, create a new variable called ”low birthweight”, where you divide the baby weight into a 0/1 variable with 1's indicating very low weight babies (you can pick the threshold for low weight). Now, try to model this new data with logistic regression. Can you classify low birthweight babies with this model?

In Part II you will try to model to model the cars data from class, using ”domestic” (0/1) as the outcome. Can you model american/non-american cars with logistic regression? Which variables are good predictors of the ”domestic” variable?

2 PART I

2.1 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?

2.2 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?

2.3 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))
+ }
```

2.4 Optional: Logistic regression

Choose a threshold for the birthweight data to create a 0/1 variable to indicate very low birthweight babies. This new variable is your outcome in a logistic regression model (do not include birthweight as a predictor).

Use the codes from lectures 11 and 12 to model the low birthweight data. Can the predictor variables be used to classify low and non-low birthweight babies? Which variables seem to be relevant. Compare with the linear model from above.

3 PART II

The low birthweight data is not an easy classification problem. Try a logistic regression analysis of the cars data set. The question I want you to investigate is the following: can you, based on data on car prices, mileage, various size measures and engine data etc, tell american cars from non-american cars (domestic=1 is american built cars)?

That is, use domestic as the outcome variable (0/1) and model this data using logistic regression. You can use the codes from lectures 11 and 12 adapted to this data set. Does the binomial model fit this data set? Are there any outliers? Do you need to transform any of the predictor variables?

Using stepwise model selection (if you want you can also try all-subset selection (see Demo 12 code)), which variables are predictive of american built vehicles? Can you interpret the selected model?

If you split the data into a training and test set and repeat the exercise, are the same variables selected? Try predicting the cars in the test set as american or not. Use the `pp<-predict(model, type="response")` function. This gives you the value on the dose-response curve, i.e. the probability that the car is american given the x-variables. Turn this into a classifier (0/1 variable) by setting all probabilities over 0.5 to 1 and all under 0.5 to 0. Compare with the domestic label of the test set (use the function `table(test$domestic, pp)`). Is the model able to classify cars as american built based on the x-variables? What is the accuracy of prediction (how many misclassifications did you get)?