Exercise Series 8

Important remark:
In the Linux computer rooms (HG E 26.1 and HG D 13), **before** you can use the package **earth**, you have to update your environment. For this, you have to type in the **terminal** the following line:
source ∼sfs/adm/Source-me
and press **Enter**. After that, you have to log out and then log in once again.
To open the **terminal**, click on the fifth icon from the left in the tool bar.

1. In this exercise we are investigating the ozone dataset which you have already seen in the lecture. The dataset `ozone` is available in numerous R-packages, e.g. in the package `gss`. You can load it with `data(ozone, package = "gss")`. If you do not have access to the required R-package you can get it via
`read.table("http://stat.ethz.ch/Teaching/Datasets/ozone.dat", header = TRUE)`.

A short description of the variables is available at `help(ozone, package = "gss")` or at `http://stat.ethz.ch/Teaching/Datasets/ozone.txt`.

a) Get an overview of the data with `pairs()`. You should take the log of the response (`upo3`) and remove the outlier in the predictor `wdsp`.
R-Hints:
The transformation can be done efficiently using

```
d.ozone <- subset(transform(ozone,logupo3=log(upo3)),select=-upo3) .
```

b) We want to use MARS as regression method. To decide on the "best" interaction degree we use a leave-one-out cross validation (see 4.3.1). You should use the sum of squared errors as loss function. Try interaction degrees 1, 2, 3 and 4. Which one performs best?
R-Hints:
`earth()` (a function performing MARS regression method) is available in the package `earth`.
Usage: `earth(formula, data, degree)`, where `formula` is the model formula (syntax as in `lm()`), `data` is a dataframe containing the measurements for the different variables and `degree` is the maximum interaction degree. For more information see `help(earth)`.

c) Independantly from b) we work now with a model with maximal interaction degree equal 2. Fit the model (with `summary()` of the `earth()` output you can take a look at the fitted model) and check the model-assumptions with a Tukey-Anscombe plot. The residuals and the fitted values can be found in the output object of `earth()`. Use the help to find out how to extract them.
We now want to visualize the effects of the predictors: Therefore we vary the one variable of the main effects or the two of an interaction while leaving the others constant. This can be done with the command `plotmo()`, which is also in the `earth` package. Read the help (`help(plotmo)`) and try different options to create nice plots.

d) Now we want to compare MARS with an additive model (7.2 in the lecture notes). Therefore you have to use a MARS regression model with maximal interaction degree 1.

R-Hints:

`gam()` (to fit an additive model) can be found in the package `mgcv`.

Usage: `gam(formula, data)`. In the formula, it is possible to define smooth terms using `s(variable name)`. To create the model formula, you may want to use the following code:

```
vars <- colnames(d.ozone)[ -10]
RHS <- paste(paste("s(", vars,")", sep=''), collapse= " + ")
gamForm <- as.formula(paste("logupo3 ~ ", RHS)) .
```

Make use of `summary()` to get an overview of your `gam()` output. Remove the insignificant explanatory variable from the model and fit this reduced model once again. (You can create the model formula in a similar way as before.)

To create nice plots use the function `p.gam()` from Dr. Mächler. You can get it with `source("ftp://stat.ethz.ch/Teaching/maechler/CompStat/plotGAM.R")`.

Preliminary discussion: Friday, May 25, 2007. **Deadline:** Friday, June 1, 2007.

Advice (for this exercise): Contact Michael Amrein, `amrein@stat.math.ethz.ch`.