

Multivariate Statistics: Exercise 3

October 29, 2014

Regression analysis:

Consider the data *world.R* from the web-page of our course. You can load these data into R with `source()`, resulting in the data frame *world*. The data contain for most countries on earth the following information:

PopGrowth	average population growth from 1986-2000
GiveBirth	percentage of women able to give birth (1986)
Women	proportion of women per 100 men in the population (1985)
LifeexpF	life expectation of females (1986)
LifeexpM	life expectation of males (1986)
InfMort	infant mortality (1986)
InhabDoc	number of inhabitants per medical doctor (1986)
Calorie	daily amount of available calories per person (1985)
BabyUnderw	Percentage of babies with underweight at birth (1984)

The first column includes a code for the continent in which the country is located, and the second column contains the full names of the countries.

The goal of this exercise is to model the response *PopGrowth* with the other explanatory variables. Start with fitting the full linear model:

```
lm.full=lm(PopGrowth ~ GiveBirth + Women + LifeexpF + LifeexpM + InfMort +  
           InhabDoc + Calorie + BabyUnderw, data=world)
```

Eliminate step-by-step variables to reach an “optimal” model (use of the function `summary()`). Are the model requirements fulfilled (use the function `plot()`)? Should certain variables be transformed first?

The above stepwise model selection is easier with the function `step()`. Try this function to arrive at a useful model.

An exhaustive model search can be done with the function `regsubsets()` from the `library(leaps)`. The results can be shown with `summary()` and `plot()`.

The use of these functions is explained in the document *regression.pdf*, available at our web-page.

Save your (successful) R code together with short documentations and interpretations of results in a text file, named as *Familyname3.R*. Send the file as an email attachment to *mehmet.mert@tuwien.ac.at*, at latest Monday (27.10).