Exercises 4

1. Have a look at block no. 6 on the Moodle page of this course. There are diagnostic plots for 13 artificially generated datasets. The usual model assumptions of (non-robust) multiple linear regression are fulfilled for 10 of them. For which three datasets do you think one or more model assumptions are violated? Why?

2. Explain in your own words and understandable to a non-statistician (with some solid school background in maths), what the model assumption of i.i.d. error terms in the general linear model means. (Don't assume that the reader knows what an "error term" is.)

3. .

   The dataset `auto.dat` can be found on the Moodle course page[1].
   Load with `read.table("auto.dat",header=TRUE)`.

   The dataset gives information about 390 cars on 7 variables. The cars have been randomly sampled in Chicago in 1983[2]. The variables are (in order of appearance as columns in the data file):

   mpg City-cycle fuel consumption in miles per gallon.

   cylinders

   displacement in cc

   horsepower

   weight in lbs

   acceleration in sec. 0-60 mph

   model year

   Imagine that you work for a public US traffic institute in 1983. The data have been collected in order to give car companies and the general public some information on how the fuel consumption (which is the response variable) is connected to the explanatory variables cylinders, displacement, horsepower, weight, acceleration and model year. A central goal is to support a reduction of fuel consumption in the future by giving reliable information about the impact of the other variables even though it is understood that customers have other desires as well (so don't just comment on the variables in the style "lower is better" etc.).

   You are asked to analyse the data using multiple linear regression. You can try to fit different models (generated for example by transforming or omitting variables; I don't expect you to use robust methods at this point) and are asked to come up with a finally recommended model.

   You are asked to

   - provide all relevant computer output including the R-code that produced it,
   - write comments on the fitted models that you don't recommend, stating why you don't recommend them and what else you have learnt from them that was relevant for the design or interpretation of your recommended model, making clear reference to R-output and graphics,
   - write comments on your recommended fitted model including diagnostic plots, making clear reference to R-output and graphics,
   - write a report for the institute, explaining the relevant information (including reasons for doubt about its reliability if necessary) in a clear and understandable way (interpreting regression coefficients is relevant information). This should be confined to the finally recommended model.

[1] In case of difficulties with the access to the dataset, please contact the Lecturer, `giampiero@stats.ucl.ac.uk`

[2] The dataset is based on a real dataset that has been manipulated.