Some comments about the General Linear Model

The detailed study of the General Linear Model is almost endless because of its flexibility, the variety of behaviours the residuals can display and the many possibilities to improve their behaviour without resorting to much more complex or black-box models.

If the *default* tests don't show any important departure from the model assumptions, you are very lucky. Depending on your objectives and the overall variability explained by the covariates you can use the model for explanation, prediction or confirmatory analysis. It's very usual than one or more assumptions are not met. Besides the *default* tests, we are going to name other than can guide you through the difficult task of finding out what is going wrong with your model and why. The first remedy to try to ameliorate bad behavoiur of the residuals is transformation (of the dependent variable or the explanatory ones); but this is not magical and it doesn't always work. For every assumption not met there is a typical list of things to do. These options range from transformation to considering different or more complex models. Some of these methods are listed in the slides. Thus, this will give you the starting point to further study extensions of the Linear Model and other models closely related to it. Before jumping in a new type of model *googling* it, I strongly recommend to look for its basis, even theoretical ones, to learn a little bit more about its structure and assumptions. Some excellent (in my opinion) and free references are the following books (available on Moodle):

- Rawlings, J.O., Pantula, S.G., Dickey, D.A. "Applied regression Analysis, A Research Tool", (1998). Springer.
- "Practical regression and Anova using R" , by J.J. Faraway.

Exercise, business data

Paper Inc. is a manufacturer of paper products. It has data consisting on 100 observations on 18 separate variables based on a market segmentation study of its customers. The company sells paper products to two market segments: the newsprint industry and the magazine industry. Also, paper products are sold to these market segments either directly to the customer or indirectly through a broker.

Paper Inc. management has long been interested in more accurately predicting the satisfaction level of its customers. To this end, researchers at this Company have proposed that a multiple linear regression analysis should be attempted to predict the customer satisfaction based on their perceived values oy company's performance. In addition to finding a way to predict satisfaction, the researchers also were interested in identifying the factors that lead to increased satisfaction for use in differentiated marketing campaigns. To apply the regression procedure, they selected $X_{19} = customer\ satisfaction$ as the dependent variable to be predicted by independent variables representing perceptions of the company's performance.

This exercise is going to be a pretext to illustrate several aspects of model building, model comparison, checking assumption on the residuals and different aspects of their distribution.

Models are going to be simplified, compared, etc. but just for illustration purposes, not because they are actually important in that particular data set or problem. The idea is to show you different tools when dealing with multiple linear regression models.

We will start by fitting a Full model with all the cuantitative X variables available, from $X6$ to $X18$. The response variable is $X19 =$*overall customer satisfaction*. First result indicates that not all variables are needed in the model. There are several approaches for deciding which variables to include in the final model. For this example, one variable at a time is eliminated, the one whose elimination will cause the smallest increase in the residual sum of squares (the one with the highest p-value in the t-test). This process will stop when all the variables remaining in the model are significant.

We get a model containing variables $X6, X7, X9, X11$ and $X12$. For illustration purposes only, lets compare this model (lm.1) with one having $\beta_6 = \beta_9$ (reduced model).

Let's compute also confidence intervals for the parameters β of the model.

Diagnostic plots. R shows us 6 plots:

- Residuals against fitted values. Here is where we have to find those patterns showing lack of independence and non-linear relationships between outcome and predictor.
- Normal Q-Q plot: a graphical tool to help us if the residuals come from a Normal distribution.
- A Scale-location plot (also called Spread-Location plot) of $\sqrt{|e_i|}$ against fitted values. It shows if the residuals are spread equally along the range of the predictor. It is good to see a roughly horizontal line with equally (randomly) spread points.
- A plot of Cook's distances in the same order as the observations. It is useful for finding influential observation and we talk more about it on multilple linear regression.
- Residuals vs. Leverage.
- Cook's distance vs. Leverage.

Miscellanea of things we will cover:

- Function residualPlots(), package car.
- Influential observations, built-in function influence.measures()
- Other type of plots to study influential observations: infIndexPlot()
- Added-variable plots and Component+residual plots
- Basic analysis of the residuals.

- Function spl(): Spread-level plot. It suggests a power (variance-stabilizing) transformation (for the Y variable) if there is a problem with non constant variance.
- Function resettest(): general test for errors of specification in the model, it is a test about the functional form of the model. The basic asssumption is that under the alternative hypothesis the model can be written in terms of powers of either the fitted response, the regressor variables or the first principal components of x. Here we are interested in high p-values.
- Harvey-Collier test for linearity. It performs a t-test on the recursive residuals. If the true relationship is not linear but convex or concave the mean of the recursive residuals should differ from 0 significantly.
- Global Diagnostic measure, package gvlma.
- Variable Transformation.
- Predictions.

Besides all these tools, there is a package called lmtest (after linear model test) where you can find many more tests related to the Linear Model.

One-way ANOVA: comparisons among several means

Basics One quantitative variable and one categorical one, its categories are often refered as *treatments*. The primary question may be *Are there differences between any of the means?*. The ANOVA F-test provides evidence in answer to this question. The term *variance* should not mislead: this is most definetly a test about *means*. It assesses mean differences by comparing amounts of variability explained by different sources. Assumptions in this model are as usual: (1) The populations have normal distributions. (2) The population standard deviations are all the same. (3) Observations within each sample are independent. (4) Observations in any one sample are independent of observations in other samples.

Multiple comparisons arise when one wants to compare more than two things at the same time, simultaneously (for the same dataset, in the same explanatory framework). In the One-way ANOVA framework, one wants to isolate where (in which groups) the differences are between the means. Multiple tests that are not carried out in an independent way lead to an inflation of the overall type I error rate. For instance, when faced with, say, 10 comparisons on the same data, the probability of falsely detecting a significant difference becomes grater that the *classical* 5 % level.

Practice We will begin studying differences in variable X_{20} of the hbat dataset through the different types of customers according to the length of time they have been customers. Repeat the study with X_{22} as the response variable. Reporting conclusions from the first example:

A one-way Analysis of Variance showed a highly significant effect of type of customer on its likelihood of recommending this company to other firms (variable X_{20}) ($F_{2,97} = 25.41$, p-v < 0.001). Multiple Comparisons (adjusting p-values by Bonferroni method, overall level 0.05) indicated that the only significant differences were between customers of type I and II, and customers of type I and III.

- An example with 5 groups. Data *competition.txt*. An example of composite comparisons (and planned comparisons).
- Ancova (Analysis of Covariance) model: It involves a combination of regression and analysis of variance. The response variable is continuous and there is at least one continuous explanatory variable and at least one categorical explanatory variable. Example with dataset *perf*.