Data Analysis and Exploration
A.A. 2010/11 Second block of exam exercises.

Every student has to solve (through a program in R that answers the questions of the exercise and include comments if useful) the practical exercise using the dataset individually assigned to each student. The R program has to be sent, through e-mail to Andrea Pugliese <pugliese@science.unitn.it> at least 3 days before the day of the oral exam.

1. A file "pinecones_n.dat" where n is the number assigned to each student, contains (fake, but similar to real experimental data) data on experiments on the effect of CO_2 concentration on production of cones by pine trees.

   The column "tmt" describes the experimental treatment: the symbol AMB indicates that the tree has been grown under normal atmospheric concentrations of CO_2; the symbol CO2 indicates that the tree has been grown under increased concentrations of CO_2. The column named "diam" contains meaures of the tree diameter; the columns c98, c99 and c00 contain the number of cones produced by each tree in 1998, 1999 e 2000.

   It is required to study the effect of CO2 on the probability of producing cones. Primarily, the response to be considered is whether or not cones have been produced in 2000 ($c00 > 0$); one can also try with the same quantity for 1998 or 1999 ($c98 > 0$ or $c99 > 0$) or some combination of these. One should try models including or not (beyond tmt, the variable of interest) also tree diameter as predictor variable. Choose the model that appears most adequate, and show the results in graphical form, compared to data. Discuss whether it can be concluded that the concentration of CO_2 influences the probability of producing cones. If the conclusion depends on the model chosen, try to argue which can be the reasons for that.

2. Consider the model $y_i = \hat{y}_i + \varepsilon_i$, where $\hat{y}_i = 80 + 1.25 z_i$ and ε_i are independent normal random variables with mean 0 and variance 3. z_i are not observable; instead $x_i = z_i + \eta_i$ is observed, where η_i are independent normal random variables with mean 0 and variance 3/50.

   Choose 50 values for z_i from a uniform distribution between 0 and 30. Then compute x_i and y_i according to the model described above. On the resulting sample $\{(x_i, y_i),\ i = 1 \dots 50\}$ estimate the coefficient of (simple) linear regression.

   Repeat 100 times the procedure; then, find the average and the variance of regression coefficient (the slope of the regression line); finally, plot its empirical distribution from the sample. Compare with the theoretical value of 1.25.

   Repeat the procedure when η_i have variance 3/10 or 3/5. Comment the results.

   [The R script for this exercise should start setting the seed equal to the number assigned to each student]