Data Analysis and Exploration
A.A. 2010/11 First 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>. The theoretical parts (plus comments to the program, if useful) can be sent via e-mail (in some format) or handed in paper. Both parts have to arrive at least 3 days before the day of the oral exam.

1. A file "bulls_n.dat" where n is the number assigned to each student, contains (fake, but similar to an auction actually happened) data on selling price of bulls together to other variables. The variables are listed in the first line of the file, and explained here:

   ```
   # Col. 1: YrHgt = yearling height at shoulder (in)
   # Col. 2: FtFrBody = fat free body wt (lbs)
   # Col. 3: PrctFFB = percent fat-free body weight (%)
   # Col. 4: Frame = size scale (1 = small to 8 = large)
   # Col. 5: BkFat = back fat (in)
   # Col. 6: SaleHt = sale height at shoulder (in)
   # Col. 7: SaleWt = sale weight (lbs)
   # Col. 8: Breed (1 = Angus, 5 = Hereford, 8 = Simental)
   # Col. 9: SalePr = selling price ($)
   ```

   Make the regression of selling price (=SalePr) on all other variables *(remember to make qualitative the variables that clearly are such).* Present the results both in terms of an analysis of variance table, and of the estimates of the regression coefficients.

   Explain why the variables appear with a different significance level in the two cases (i.e., explain exactly to which tests the significance refer to).

   Make the regression of selling price first on the variable referring to the highest sum of squares in the analysis of variance table, then add Breed (as a qualitative variable) to the explanatory variable. Explain the difference between a model with or without interaction among the variables.

   Prepare some graphical output of the final regression, and comment whether this regression appears adequate.

   Use a systematic method to select the predictor variables to keep in the model. Compare the results to what previously obtained.

2. We organize an experimental treatment, in which a qualitative variable Z_1 can have values 1 and 2, and another qualitative variable Z_2 can have values A, B and C. For each combination of the two qualitative variables, 2 observations of the quantitative variable Y are taken.

We wish to analyse the resulting data through a linear model where A depends on Z_1 and Z_2; write down the assumptions for the individual data $Y_{i,j,k}$ ($i = 1, 2$ is the value of Z_1, $j = A, B, C$ is the value of Z_2, $k = 1, 2$ represents the observation) in an additive model or a model with interactions. Write down the corresponding model matrices X and the matrix $(X^t X)^{-1}$ in the additive case. Write down expected value and variance for the estimators in the resulting model[1]

[1] it is required just to write the ingredients for the computation, not an explicit formula that would be very cumbersome.