Statistical Data Analysis
Exercise Session 6

Exercise 1: blood coagulation dataset.
Load the coagulation dataset from the faraway library (you can use ?coagulation for more information about this dataset).

1. Make the boxplots of coag for all 4 diets. What do they show?
2. Use the lm function to compute the mean of all 4 diets. Test whether the mean of diet B is different from the mean of diet C, without computing a confidence interval.
3. Test whether at least one of the diets has a different mean than the others.
4. Check the model assumptions (Levene test and QQ plot of the residuals).
5. Test whether the mean of diet B is different from the mean of diet C, without computing a confidence interval.
6. Compute 95% C.I. for all pairwise differences between the means of two diets. Also compute the 95% Bonferroni, Scheffé and Tukey HSD intervals. Which mean differences are significantly different from zero according to these intervals?
7. Use the wald.test in the aod package to test whether the average of the means of diets A and D is significantly different from the average of the means of the diets B and C.

Exercise 2 : car dataset.
Load the car2.txt dataset. The objective is to construct a model to forecast the variable "acceleration" based on the other variables.

1. Redefine the variable origin as a factor.
2. Use the appropriate diagnostic tools to check whether it seems useful to transform some of the predictor variables and/or to remove potentially influential observations.
3. Construct and fit a linear model to predict acceleration. Does the linear model seem appropriate?
4. Compute a Box-Cox transformation of the response variable that gives a better fit. Compare the fit and the diagnostic plots with the results you obtained using the original response variable.

Exercise 3 : election dataset.
Load the fpe dataset from the faraway library. The variables A to K represent the votes (in thousands) for each of the 10 candidates at the first round of the 1981 French presidential elections. The rows represent the French departments. A2 represents the votes (at the second round) for the final winner of the election. In each department the number of voters in the second round was larger than in the first and N represents this difference. We will consider N as if it were also a candidate in the first round. Finally EI represents the number of registered voters. Consider the vote transfer function:

$$A2_i = \beta_A A_i + \beta_B B_i + \beta_C C_i + \beta_D D_i + \beta_E E_i + \beta_F F_i + \beta_H H_i + \beta_J J_i + \beta_K K_i + \beta_N N_i + \epsilon_i$$

for all $1 \leq i \leq 24$. We are interested in estimating the different $\beta' s$.

1. Use lm to perform an OLS estimation of the β's

2. Explain the large value of the coefficient of determination.

3. Are all model assumptions satisfied?

4. Notice that the size of the different departments varies widely. In light of this, how would you modify the model you just fitted? Check whether it now satisfies the model assumptions. Compare the parameter estimates and their standard error.

Some useful functions:

- boxcox in package MASS.