Exercise 11

Multiple linear regression analysis

As cheese ages, various chemical processes take place that determine the taste of the final product. This dataset contains concentrations of various chemicals in 30 samples of mature cheddar cheese, and a subjective measure of taste for each sample. The predictor variables "Acetic" and "H2S" are the natural logarithm of the concentration of acetic asid and hydrogen sulfide respectively. The variable "Lactic" has not been transformed. To what extent do the predictor variables predict the subjective tast measurements?

(Reference: Moore, David S., and George P. McCabe (1989). *Introduction to the Practice of Statistics*. Download from *The Data and Story Library*, see http://lib.stat.cmu.edu/DASL/)

Download the table cheese.sav from: http://www.let.rug.nl/~heeringa/statistics/stat03_2013/ and load the table in SPSS.

1. Create for each pair of predictor variables a scatter plot and calculate the correlation coefficients between the predictor variables. Do you find multicollinearity?

2. Perform a multiple linear regression analyis (all-at-once), and create the residual plot (predicted values versus residuals). Save the residuals and calculate the Cook's distances.

3. Look at the residual plot. Are the assumptions of linearity and homoskedasticity met?

4. Create a bar graph on the basis of the Cook's distances. Do you expect that there are any influential points?

5. Test the normality of the residuals by creating a normal quantile plot and performing the Shapiro-Wilk test.

6. Look at the results of the linear regression analysis which are found in the table 'Coefficients' in the SPSS output.What do you conclude? What is the adjusted R square?

7. Perform also a stepwise linear regression analysis. Are all variables retained in the model? What is the adjusted R square?