# Statistical Data Analysis
## Exercise Session 8

Exercise 1 : PCR Regression.
Load the pollution dataset from the SMPracticals package. More information about this data can be found in ?SMPracticals::pollution. The variables hc, nox and so are right-skewed distributed. Work with the logarithm of these variables.

1. Investigate the presence of multicollinearity in this data set, based on the correlation matrix and the VIF values.
2. Perform a PCR regression on the first 50 observations.
   - Select the number of components based on the PCA analysis on the predictor variables.
   - Perform the PCR regression using the pcr function from the pls package.
   - Check that the coefficients obtained correspond with performing an LS analysis on the selected PCA scores.
   - Compared the obtained coefficients with an analysis of the complete data.
3. Now select the number of PCR components based on
   - The RMSEP of the training data set (observations 1 to 50)
   - The RMSEP based on leave-one-outcross validation (from the training data set)
   - The RMSEP calculated on the validation set (observations 51 to 60). Use the validation plot function for this.

Exercise 2: Ridge Regression.
Consider the same dataset as in exercise 1. Perform a ridge regression with the first 50 observations:

1. Determine a good optimal value for $\lambda$, the ridge parameter, based on the ridge trace and the VIF values.
2. Perform ridge regression with this chosen value for $\lambda$.
3. Calculate the RMSEP of the validation set (observations 51 to 60). Compare the results of the RMSEP values based on PCR.

Exercise 3: Robust Regression.
Load the hills data set from the MASS library.

1. Perform an LS analysis. Determine the observations with the largest studentized residual. Also make a residual plot and normal quantile plot.
2. Determine LS-based diagnostics to detect outliers: diagonal elements of the hat matrix, DFFITS, DFBETAS, Cook's distance. Always identify the points that are outliers according to the corresponding criteria.
3. Perform LTS regression with 50% breakpoint. Compare the parameter estimates with the LS solution.
4. Identify the outliers according to the LTS method. Make the diagnostic plot to divide them into good / bad leverage points and vertical outliers.

5. Make a scatter plot of the predictor variables. Add the tolerance ellipse to this, based on the classical average and covariance matrix. Also add the MCD-based tolerance ellipse. Compare the robust distances with the Mahalanobis distances.

6. Compare the LTS results when you lower the breakpoint.

Some useful functions:

1. lm.ridge() in de MASS library.

2. ltsReg() and covMcd() in the robustbase library.