Your data set contains 20 variables and 96 objects. The first 19 variables are X-variables (absorbances measured on a spectrophotometer) and the last variable is the Y-variable (concentration of a chemical constituent (protein)). The data set is organised the normal way as a 96*20 matrix.

1. Check for collinearity among the X-variables by the use of the condition number/index and the correlation matrix.

2. Build a regression model of Y vs. all the X-variables by the use of principal components regression (PCR) based on the first 48 objects. Use the last 48 objects as a test set. Determine the number of components by the use of full (leave-one-out) cross-validation on the first 48 objects and determine the prediction ability of the best solution (on the test set, the last 48 objects).

3. Standardise the X-variables (divide the variables by their standard deviation) and do a new PCR on the standardised data. Are there any differences between this and the above solution?

4. Are there any outliers in the training data? (Use the original, unstandardised, data for this and the following exercises.)

5. Is there any pattern of interest in the PCA scores plot of the training data.

6. Use only the first 10 X-variables and split the calibration/training data set (first 48 objects) in two equal parts (24 objects in each). Compare the two groups by the use of the Hotelling's T square method (page 77 in the book). Is there any significant difference in the mean value of the two groups?

Present and comment all relevant results and solutions and give a brief description of what the methods you use do with the data. Enclose computer code used for the calculations. If a menu based computer system is applied, describe the procedure used. Emphasise assumptions made.