Open the file “Housing raw data.unsb”

The Boston Housing data set was analyzed by Harrison and Rubinfeld (1978) who wanted to find out whether ``clean air'' had an influence on house prices.

The objective with the analysis: From the available data, can this hypothesis be confirmed?

Already, the data set has been analyzed with MLR in a previous assignment. You will now use this data set for PCR and PLSR.

The data set is described by the following variables:

Independent variables (X):

CRIM: Per capita crime rate by town
ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS: Proportion of non-retail business acres per town
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX: Nitric oxide concentration (parts per 10 million)
RM: Average number of rooms per dwelling
AGE: Proportion of owner-occupied units built prior to 1940
DIS: Weighted distances to five Boston employment centers
RAD: Index of accessibility to radial highways
TAX: Full-value property tax rate per $10,000
PTRATIO: Pupil-teacher ratio by town
B: 1000(Bk — 0.63)^2, where Bk is the proportion of [people of African American descent] by town
LSTAT: Percentage of lower status of the population

Response variable:

MEDV Median value of owner-occupied homes in $1000's

Procedure:

- Divide the 506 samples into training (2/3) and test (1/3). Create row sets in the Define Range editor (Edit – Define range). Hint: Click on Special intervals. You might use the sets in the already stored file from the previous assignment. Name the sets “training” and “test”
- Make a PCR model on the samples set “training” with the column sets “X-variables” and MedianValue(Y) as Y (as was done last week). Interpret the various plots (scores, loadings, influence, predicted vs. reference, regression coefficients, residuals). Decide on the optimal number of PCs to use for prediction.
- Recalculate without some of the not-so-important variables and compare the models. Mark variables and do Tasks – Recalculate - Without marked - Variables. Compare the results, can you improve the model?
- Now predict the test set: Tasks-Predict-Regression. Is the RMSEP similar to the model on the training set?
- Now do the same with PLS regression. Compare the number of PCs needed for modelling Y. Why is there a difference? Are the final regression coefficients similar?
- Compare the regression coefficients from MLR with PCR and PLSR. Why are they different?

For interpretation of the various plots you’ll find a lot of information in the help system!