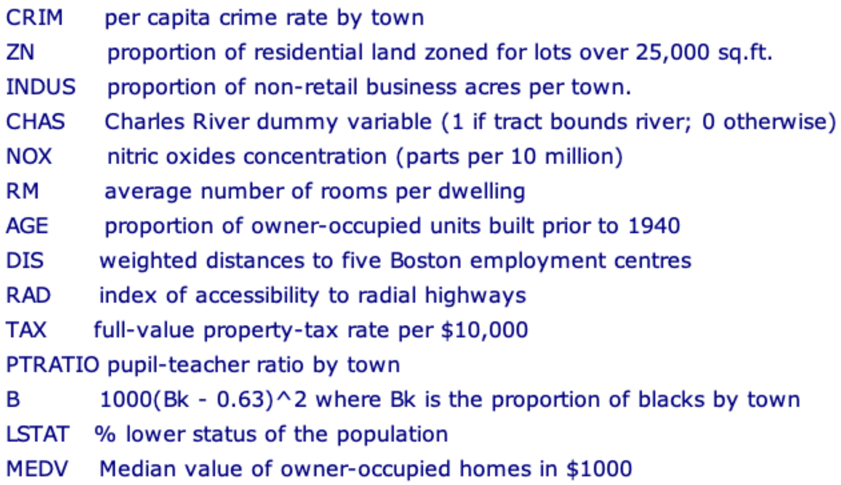
Multiple Linear Regression

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

Predicting MEDV housing price

This dataset contains information collected by the US Census Service concerning housing in the area of Boston Massachusetts. The dataset has 506 cases. The data was originally published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

The file Boston.xls contains the following 14 attributes:



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 oxides 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 centres
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 blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000

Answer the following questions

Predict New House using given predictors

a) Fit a multiple linear regression model to predict MEDV using CRIM, CHAS and RM as predictors.
    - Write the equation for predicting the median house price from the predictors in the model.
    - What is the median house price for (not bound Charles river, crime rate of 0.00632, avg number of rooms 6.575)?

Search for possible multicollinearity

b) There are several variables that measure levels of industrialization, which are expected to be positively correlated. These include INDUS, NOX (pollution), and TAX. **We expect a positive relationship between NOX (nitric oxides concentration, a pollutant), INDUS (proportion of non-retail business acres per town) and TAX (tax rate),** because areas that have a high proportion of non-retail businesses tend to have higher taxes and more pollution. These 3 predictors are likely to measure the same thing.
    - Compute the correlation table between them
    - Search for highly correlated pairs. These have potential redundancy and can cause multi-collinearity. Choose which ones to remove based on this table.

Reduce the number of predictors & Propose Your Best Model

c) Use a feature selection mechanism to reduce the remaining predictors (from previous step). Run each model separately using the training/testing datasets. Then, give the best model in terms of regression equation. What is RMSE ?