SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

BIG DATA AND DATA ANALYTICS
LAB PROJECT 4
(DATASET DESCRIPTION & PREPARATION)

This lab project is based on a housing dataset of suburbs in Boston.

PREPARATION

In preparation for Lab Project 4, load the Housing.csv dataset. Use the dataset to practice the following topics:

- Multiple linear regressions: coefficient estimates, predicted values, residuals, standard errors, confidence intervals, t-values, residual standard error, R^2
- Reporting of results in APA style
- Analysing and interpreting the results of multiple linear regressions
- Parallel computing using “foreach” and “doParallel”, determining the number of cores of the current machine, setting up a parallel environment in R
- Training regression models for prediction
- Validation set approach, Leave-one-out cross-validation (LOOCV), k-fold cross-validation

REFERENCES

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

DATASET

Housing	*Housing Dataset*

Description

A dataset about housing values in suburbs of Boston at the end of the 1970s.

Usage

Housing

Format

A data frame with 506 observations on the following 14 variables.

`ID`	Town identifier
`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
`LSTAT`	% lower status of the population
`MEDV`	Median value of owner-occupied homes in $1000's

Source

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Harrison, D., & Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics & Management*, 5, 81-102.