SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

BIG DATA AND DATA ANALYTICS
LAB PROJECT 4

This lab project is based on a housing dataset of suburbs in Boston. The dataset is available from the UCI Machine Learning Repository (Lichman, 2013):

http://archive.ics.uci.edu/ml/datasets/Housing

EXERCISE 1 (2 MARKS) *[R-CODE]*

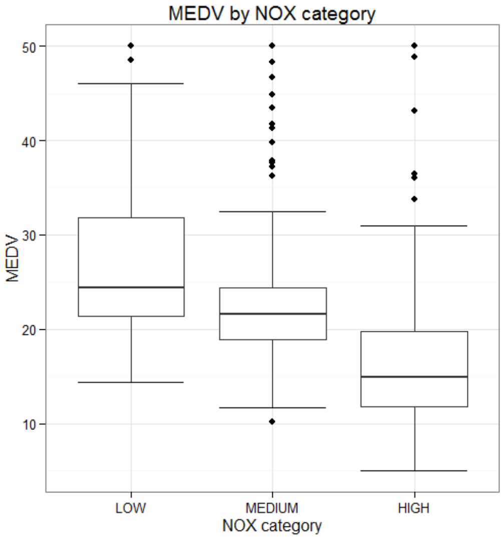
Use R to perform a multiple linear regression that regresses MEDV on CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), NOX (nitric oxides concentration; parts per 10 million), DIS (weighted distances to five Boston employment centres), and AGE (proportion of owner-occupied units built prior to 1940). Interpret the coefficients and report the results of the regression in APA style (including a regression table and reporting of F-values).

EXERCISE 2 (1 MARK) *[R-CODE]*

Use R to create a new factor variable called NOXCAT that categorizes the suburbs into towns with LOW, MEDIUM, and HIGH nitric oxides concentration (based on the variable NOX). The categorization should be as follows:

- LOW (<= 30% Quantile)
- MEDIUM (> 30% Quantile & <= 70% Quantile)
- HIGH (> 70% Quantile)

Then, use ggplot to create a boxplot that shows MEDV for the different values of NOXCAT (LOW, MEDIUM, HIGH).

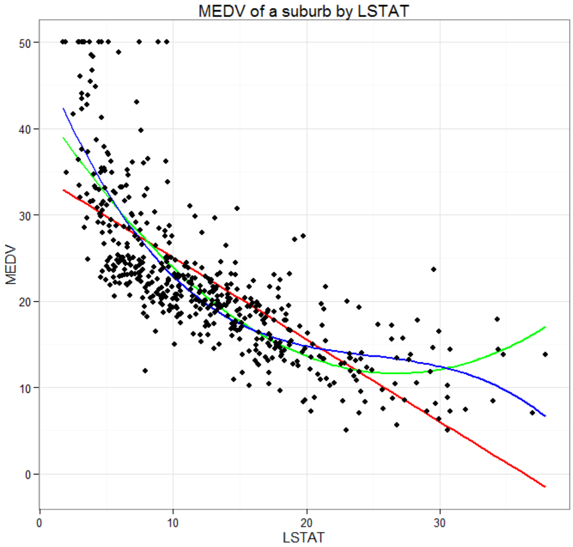


EXERCISE 3 (2 MARKS) *[R-CODE]*

The newly created variable NOXCAT is a categorical variable with three possible values (LOW, MEDIUM, and HIGH). Use R to manually create a set of dummy variables (for different values of NOXCAT) and then regress MEDV on the different NOX categories. The coding of the dummy variables in the regression should be such that the intercept reflects the MEDV value of suburbs in the MEDIUM category. Interpret the coefficients.

EXERCISE 4 (1 MARKS) *[R-CODE]*

Use ggplot() to create a scatterplot of MEDV by LSTAT. Add a linear fit (red), a quadratic fit (green), and a cubic fit (blue) to the plot.

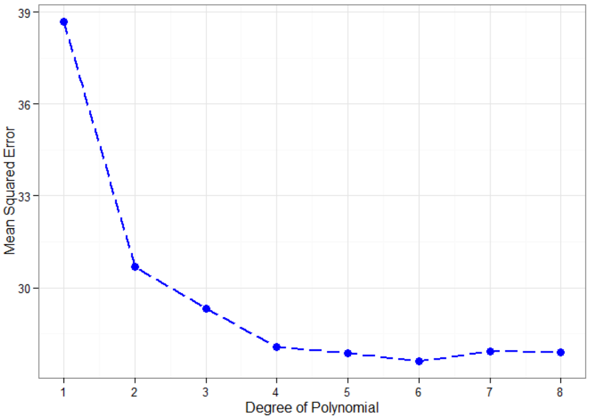


EXERCISE 5 (2 MARKS) *[R-CODE]*

Use Leave-One-Out Cross-Validation (LOOCV) to compare a linear model, a quadratic model, a cubic model, and a quartic model to regress MEDV on LSTAT. Interpret the results based on the mean-squared error (MSE).

EXERCISE 6 (2 MARKS) *[R-CODE]*

Use 11-fold cross-validation to compare 8 different degrees of polynomials to regress MEDV on LSTAT. Use ggplot() to plot the mean squared error (MSE) over the 8 different degrees of polynomials. Interpret the results based on the MSE. Why is 11-fold cross-validation in this particular case advantageous compared to 10-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.