# MATH70071 Applied Statistics – Assessed Coursework 1

## Due Monday 25 October 2021 – deadline specific to your group

Upload your final version only - once the report is uploaded there is no option for re-uploading. Avoid last minute uploads. Hand-in no more than **10 pages**. Keep your R code concise and present it inline as part of the report. Considerable emphasis will be put on clarity of expression and a clean presentation. Only detailed, well-written answers will score highly.

This coursework concerns a version of the Boston Housing data relating median house prices in different suburbs of Boston to some measured attributes. The data can be obtained from the R file bos.R posted on Blackboard, and contain the following fields:

| | |
|---|---|
| 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 |

a) Perform an exploratory analysis of the BostonHousing data and write a summary reporting any important features, including the data types of the variables and their ranges. Are there any unusual features in the house prices? Make your summary concise and informative.

The remainder of the coursework concerns fitting a normal linear regression model for nitric oxide concentration (NOX) with the following predictors: INDUS, RAD, TAX, AGE, along with an intercept term.

b) Write a formal mathematical expression for this linear model, stating clearly any assumptions which are made.
c) Which of these variables have statistically significant regression coefficients?
d) Make a Q-Q plot for the linear regression above, and comment on the model fit. What is the most prominent weakness of the linear model fit?
e) Perform two analysis of variance for the above linear model, firstly including the variable TAX last in the formula, secondly with TAX at the start of the formula. Comment on the comparison.
f) Make a plot of Cook's distances against leverages for this regression model to identify some outlying data points. Remove from the data those observations with leverage exceeding 0.05 or Cook's distance exceeding 0.02, and refit the linear model. How many points are discarded, and how do the parameter estimates and model fit compare?
g) Using the estimates from the revised model fit in the previous part, predict the nitric oxide concentration for another suburb with the four predictor variables equal to their median values from the full data set.
h) Construct a 99% confidence interval for prediction made in the previous part. Give a clear interpretation for this interval.
i) With the data still filtered, make an added variable plot for the variable TAX with respect to the other predictors INDUS, RAD, AGE. Interpret this plot.
j) Returning to the unfiltered data set, choose another variable to add to the normal linear model for NOX. Justify your choice, and comment on whether the additional variable improves the model.

---