## Question 1 (10 marks)

This question will require you to analyse a regression dataset. The file housing.ass3.2020.csv contains the data that we will use for this question. This dataset is a modified version of the Boston housing data which was collected to study house prices in the metropolitan region of Boston. In this data set, each observation represents a particular suburb from the Boston region. The outcome, medv, is the median value of owner-occupied homes in 1,000 in the suburb. The variables are summarised in Table 1. The data consists of $p = 12$ variables measured on $n = 250$ suburbs. We are interested in discovering which predictors are good determinants of housing price, and how these variables effect the median house price.

1. Fit a multiple linear model to the housing data using R. Using the results of fitting the linear model, which predictors do you think are possibly associated with median house value, and why? Which three variables appear to be the strongest predictors of housing price, and why? **[3 marks]**

2. How would your assessment of which predictors are associated change if you used the Bonferroni procedure with $\alpha = 0.05$? **[1 mark]**

3. Describe what effect the per-capita crime rate (crim) appears to have on the median house price. Describe what effect a suburb having frontage on the Charles River has on the median house price for that suburb. **[2 marks]**

4. Use the stepwise selection procedure, with the BIC criterion, to prune out potentially unimportant variables. Write down the final regression equation obtained after pruning. **[1 mark]**

5. If a council wanted to try and improve the median house value in their suburb, what does the model that we found in Question 1.4 suggest they could try and do? **[2 marks]**

6. Table 2 gives the values of predictors for a new suburb. Use the model found in Question 1.4 to predict the median house price for this suburb. Provide a 95% confidence interval for this prediction. **[1 mark]**

| Variable name | Description | Values |
|---|---|---|
| crim | Per-capita crime rate | > 0 |
| zn | Proportion of residential land zoned for lots over 25,000 sq. ft. | 0 − 100 |
| indus | Proportion of non-retail business acres per town | 0 − 100 |
| chas | Does the suburb front the Charles River? | 0 = No, 1 = Yes |
| nox | Nitric oxides concentration (parts per 10 million) | > 0 |
| rm | Average number of rooms per dwelling | ≥ 1 |
| age | Proportion of owner-occupied units built prior to 1940 | 0 − 100 |
| dis | Weighted distances to five Boston employment centres | > 0 |
| rad | Index of accessibility to radial highways | > 0 |
| tax | Full-value property-tax rate per $10,000 | 187 − 711 |
| ptratio | Pupil-teacher ratio | > 0 |
| lstat | Percentage of “lower status” of the population | 0 − 100 |

| | | |
|---|---|---|
| medv | Median value of owner-occupied homes in $1,000s | > 0 |

Table 1: Boston Housing Data Dictionary.

| |
|---|
| **Variable** crim zn indus chas nox rm age dis rad tax ptratio lstat **Value** 0.04741 0 11.93 0 0.573 6.03 80.8 2.505 1 273 21 7.88 |

Table 2: Boston Housing Data Dictionary.