Group Project

*Each group should submit **one** project to the relevant homework box (at the School of Mathematics, Statistics and Physics General Office, third floor, Herschel Building) no later than*

***3.00pm on Friday 27th March 2020**.*

Please note that:

- ***Every** member of the group needs to attach a **separate** NESS cover sheet to the front of the project before it is submitted.*
- *The report should not exceed 12 pages, written in Word or Latex. Project reports exceeding this limit will be penalised. Note that you may include an Appendix for supplementary R code, tabular and graphical output, and this does not count towards the page limit.*
- *At the end of the report, include a short statement which outlines each person's contribution to the project. All group members should sign this statement. It does not contribute to the page limit.*
- *You must also include the student login ID you used to create your random sample (see below) in order that your results are reproducible.*
- *The project is worth 15% of the overall mark for the module.*

1 Project brief

In this project, you will analyse the `Boston` dataset which concerns the value of housing in the Boston Standard Metropolitan Statistical Area in 1970. The dataset is part of the `nclSLR` package[1]. It can be loaded into R as follows:

```
## Load nclSLR package
library(nclSLR)
## Load the data
data(Boston, package="nclSLR")
```

Each group is required to take a random sample of size $n = 400$ from the full dataset, so that you are all working with your own dataset. This can be achieved in R as follows:

```
## Set the seed using one of your group members login ID
set.seed(1234567)
# Randomly sample 400 rows from the 506 rows in the full data
sampid = sample(dim(Boston)[1], 400)
BostonNew = Boston[sampid, ]
## Check that this has worked - dimensions should be 400 by 14
dim(BostonNew)


## [1] 400  14
```

[1] Note that there is also a dataset called `Boston` in the `MASS` package, but this is different and should not be used.

```
## Print first few rows
head(BostonNew, 3)

##         lcrim zn indus chas   nox    rm  age disf rad tax ptratio  black lstat medv
## 497 -1.239255  0  9.69    0 0.585 5.390 27.1    2   6 391    19.2 396.90 21.14 19.7
## 78  -2.441043  0 12.83    0 0.437 6.140 54.2    2   5 398    18.7 386.96 10.27 20.8
## 411  3.934485  0 18.10    0 0.597 5.757  0.0    1  24 666    20.2   2.60 10.11 15.0

## Save the new dataset
save(BostonNew, file = "BostonNew.RData")
```

You should then use this new dataset for all subsequent analyses. It can be loaded in a variety of ways. For example:

```
load("BostonNew.RData")
```

A description of each of the variables can be found by typing `?Boston` in the console.

The goal of this project is to build and interpret linear regression models for predicting the natural logarithm of the per capita crime rate (`lcrim`) in terms of (at least some of) the other variables. It should be stressed that this is a real dataset and there is no “correct” answer. Instead, what is required is evidence of an understanding the main statistical ideas, sound interpretation of results and demonstration of competence in the use of R as a tool for data analysis.

The project should be written up as a coherent report, giving consideration to the points detailed in Section 1.1 below. You do not need to include all R code, but you should include enough detail for the reader to be able to understand what you have done and why. You do not need to comprehensively describe everything you have done to explore and model the data. Instead, just describe the salient features of the modelling approaches you have used, then focus on reporting and interpreting your results.

1.1 Points to consider

- You should begin with some exploratory data analysis. For example, how might you summarise the data graphically, and what does this tell you about the relationships between the response variable and predictor variables? What about the relationships amongst the predictor variables?

- The variable `disf` is an ordered categorical variable with four levels. It could therefore be treated either as a factor (i.e. by introducing three indicator variables) or as a quantitative variable. Decide how (and if) you want to use it and include in your report a brief statistical justification for your decision. Note that you may like to do some experimentation to help inform your choice.

- You should identify a “best” model using each of the following:
  - At least one subset selection method;
  - At least one of ridge regression and the LASSO;
  - At least one of principal component regression and PLS.

In each case you should present the coefficients of the fitted model, and any other useful graphical or numerical summaries, and interpret your results. For example, which variables drop out of the model when you use subset selection or the LASSO? Can you interpret the transformed predictor variables obtained using principal component regression and PLS? What do the coefficients tell you about the relationships between the response and predictor variables?

- Formally compare the performance of your set of three or more “best” models (identified above) using cross-validation. Think about how you might do this in a way that makes the comparison fair. **Hint:** remember that for any given method, the MSE from k-fold cross-validation will be different for each partition of the data into k segments.
- Select a final “best” model, justifying your choice. Does it include all the predictor variables? Why or why not?
- Can you think of any possible use for the rows of data that were *not* included in your random sample?