DATA 303/473 Assignment 3

NAME (ID): Write your name and student ID here

Boston Data Set

In this assignment, we use `Boston` data set. This is a data set containing housing values in 506 suburbs of Boston (a data frame with 506 rows and 13 variables).

- 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: nitrogen 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 mean of 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: percent of households with low socioeconomic status.
- medv: median value of owner-occupied homes in $1000s.

```
set.seed(1)
library(ISLR2)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
```

```
dim(Boston)
```

```
## [1] 506  13
```

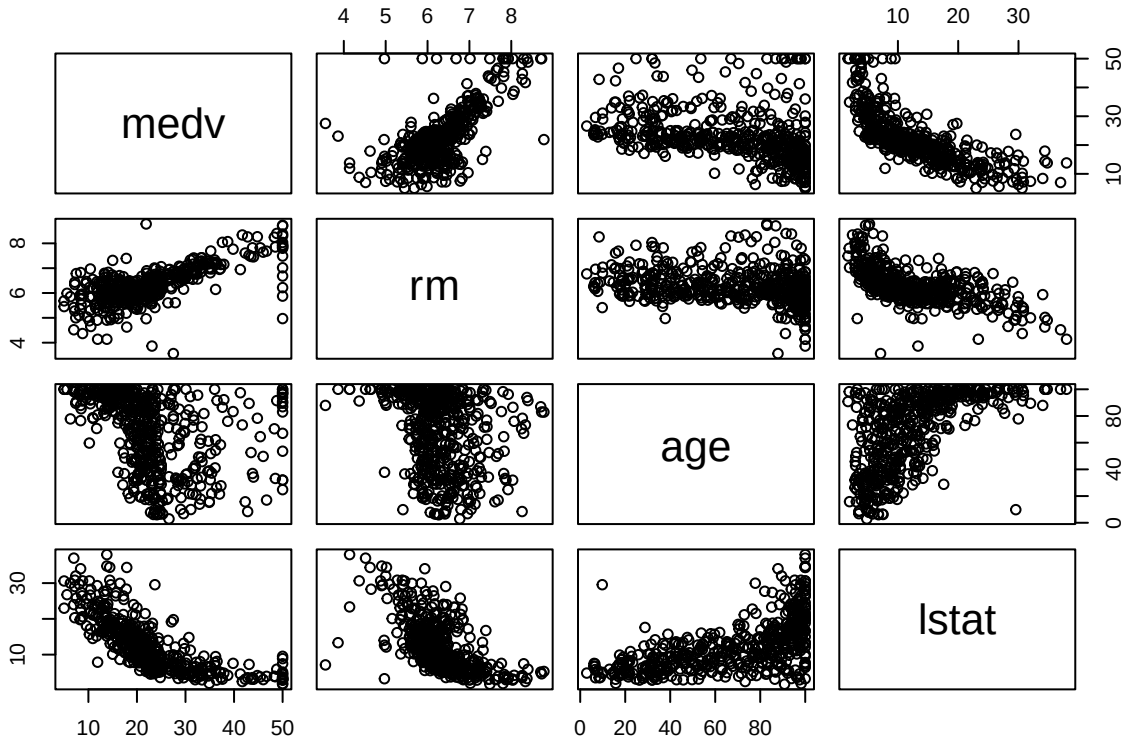
Q1 (Deviance test, AIC, test MSE)

Our interest is to predict `medv` (median house value) using predictors

- `rm`(average number of rooms per house),

- `age`(proportion of owner-occupied units built prior to 1940) and
- `lstat`(percent of households with low socioeconomic status).

```
pairs(~ medv + rm + age + lstat,  data = Boston)
```



We fit the following models:

```
m1 : medv ~ rm  + lstat,
m2 : medv ~ rm  + poly(lstat, df=2)
m3 : medv ~ rm + age + lstat
m4 : medv ~ rm + age + poly(lstat, df=2)
```

(a) **(10 marks)** Fit the model and use `anova()` function to do the deviance test to compare the models. Choose the best model.

(b) **(5 marks)** Calculate `AIC` for each model fitted in (a). Choose the best model using the value of `AIC`.

(c) **(10 marks)** Split the data set (100%) into a training set (80%) and a test set (20%). Then fit model1–model5 on the training set, and calculate the test MSE for each model. Choose the best model.

(d) **(10 marks)** By combining the result from (a), (b) and (c), decide the best model. Refit the chosen model using all of the `Boston` data set. Make a prediction of `medv` for a suburb with values `rm=10`, `age=50` and `lstat=10`. Interpret the predicted value.

Q2 (LASSO, best subset selection)

We continue to work on `Boston` data set. The aim in Q2 is to predict `medv` (median house value) using all predictors in `Boston` data set. In the following questions, we apply LASSO and the best subset selection methods.

(a) **(10 marks)** (LASSO) Fit a lasso model on the training set, with λ chosen by cross-validation with the `1 se rule` . Report the test error obtained, along with the values of non-zero coefficient estimates. We use the training set and the test set created in Q1 (c).

(b) **(10 marks)** (Best subset selection) Do the best subset selection with `BIC` and choose the best model. Report the values of coefficient estimates in the best model.

(c) **(10 marks)** Comparing the LASSO chosen model and the best subset selected model, which is the better model? Explain why?

(d) **(10 marks)** How can you improve the fit of the best subset selected model?

[**Total: 75 marks**]