STATS216-2015-Homework1

Darragh Hanley; SUID : 06001134

January 15, 2015

---

1) In this question we will consider some real-life applications of statistical learning:

(a) Describe three real-life applications (other than those described in this problem set) in which unsupervised learning might be useful.

1) A computer manufacturer, identifying closely competing products from competitors and its own product mix, based on features (RAM, Hard Drive, Price) in order to identify where there are gaps in the market and where competition is high.

2) A computer manufacturer may want to study the sentiment of product reviews and look for common terms for a particular product. Positive or negative sentiment may be derived from the start rating given in that review. For example, if one product has the term "battery" mentioned frequently in reviews with a poor star rating, it can be highlighted that there may be a battery problem for that product.

3) A retailer can segment its customer base (for eg. on buying habits, or age) to understand in which stores which product mix to use. For example, it could determine one region has a lot of young families, whereas another region has a more senior population. It could then market products in the associated stores accordingly. For example, baby products could be placed prominently in the stores where young families are common.

4) A search engine may want to cluster documents into similar groups, for example to find documents relating to an event, or a person.

(b) Describe three real-life applications (other than those described in this problem set) in which regression might be useful. Describe the response as well as the predictors. Is the goal of each application inference, prediction, or both? Explain your answer.

1) A consumer product company may want to forecast how many products it will need to ship in the next six months, based on historical information. In this case the historical information such as historical shipments, stage in product lifecycle, region of sale, month of sale serve as predictors, while the future shipments over the next six months would be the response. This is an example of prediction as the company is not interested in knowing the relationships between the predictors and the response; it simply want an accurate model to forecast future shipments.

2) A consumer product company may want to study it's historical marketing promotions, such as rebates or discounts, in order to understand which factors lead to the greatest increase in profit margin. In this case the price point, profit margin per unit, sales volume, product type, seasonality and ad size may serve as some of the predictors, while the total profit margin for the discounted period would serve as the response. This is an example of inference, as the

company is interested to understand how the input variables, or combination of variables, influence profit margin, and how to maximize the margin.

3) Regression may be used understand how many viewers a TV episode would receive in order to price advertising slots. By studying previous episodes and series of the same show and related shows, predictors may be the time of year, day of week, time of day, rating of the show, previous episodes viewers and the response would be the number of viewers for future shows.This is an example of prediction as we are not interested in knowing the relationships between the predictors and the response; we simply want an accurate model to forecast viewers of future episodes.

(c) Describe three real-life applications (other than those described in this problem set) in which classification might be useful. Describe the response as well as the predictors. Is the goal of each application inference, prediction, or both? Explain your answer.

1) A search engine owner may be interested to know relevant links for searches. In this case the model may be trained with historical information of which links were regularly clicked for previous searches from that user and others. There may be many predictors in this case such as the individual search terms, the users country, the number of terms entered, whether the page is a homepage or not, historical click-through links for the user etc. The predictor is a binary value indication the relevance of a link for the users current or future searches. The goal of the application is prediction as the search engine owner is interested only to provide the most relevant links, and not why these links are most relevant.

2) Another case of classification may be credit scoring to measure the risk of an individual paying back a loan. The predictor would be individual's age, employment status, salary, dependents, address, among others. While the response would be the risk of repayment. The goal could be both inference and prediction. Prediction could be used in order to decide whether to loan to a specific individual - in this case we are not interested in why he was approved. Whereas inference could be used to understand the features of the population that the company would like to target in advertising campaigns (eg. company strategy may be to charge higher interest rates for more risky customers in affluent areas.).In this case the attributes that feed into credit worthiness are important.

3) Classification may be used to predict fraudulent credit card transactions. In this case the predictors could include address of card owners, location, times and value of previous transactions. While the response would be whether the transaction is fraudulent or not. As the features of the transaction that highlighted the risk are not important, this would be predictive classification.

2. Explain whether each scenario below is a regression, classification or unsupervised learning problem, and indicate for each supervised learning scenario whether we are more interested in inference or prediction. Finally, provide n and p.

(a) Stanford received 42,000 undergraduate applications in the year 2014. The application includes the following data for each applicant: age, high school GPA, scores in the SAT Critical Reading, SAT Math and SAT Writing exams, whether they are domestic or international and whether they are transferring students or not. The university also knows which of these applicants have been admitted and wishes to understand how the different factors affect admission chances.

This is a classification problem since the response is qualitative (student is admitted or not). This is an inference problem as the university wishes to understand how the different factors affect admission chances, as opposed to predicting admission chances.

n=42000 (Undergraduate applications) p=8

Predictors : (1) age,

(2) high school GPA,

(3) scores in the SAT,

(4) Critical Reading,

(5) SAT Math and

(6) SAT Writing exams,

(7) whether they are domestic or international,

(8) whether they are transferring students or not.

Response : whether applicants have been admitted

(b) An online retailer wants to launch several different targeted ad campaigns and is interested in identifying distinct customer subtypes based on 1.5 million customers' past purchase histories. For each of 500,000 products sold, the company maintains a count of how often that product has been purchased by each customer in the past.

This is unsupervised learning.

Taking the course books definition " unsupervised learning describes the somewhat more challenging situation in which for every observation i = 1,...,n, we observe a vector of measurements xi but no associated response yi.". We have no response as we only want to break the customers into undefined subtypes. We also have a vector of measurements 500K Products * 1.5 million customers - and for each combination how often the product has been purchased.

n = 7.5e+11 observations

```
(500K products * 1.5 million customers)
```

p = 2

```
(I am assuming in this answer only the predictors mentioned in the question are used,
and other predictors such as date of purchase are not considered.)
```

Predictors are : (1) Customers (2) Products

(c) A book publisher would like to infer which factors turn a book into a best-seller. For each of the 4,000 books it has published, it knows whether the book published was considered a best-seller, a good sell, a bad sell, or a horrible sell. Also, for each book, it knows the number of books by the same author sold in the past, the book's theme, the book's length, whether it got mostly good or bad reviews, and the target audience.

This is obviously an inference problem, and not prediction, as the "book publisher would like to infer which factors turn a book into a best-seller". It is also a classification problem, as the response is whether the book was considered to sell ("best-seller, a good sell, a bad sell, or a horrible sell"). It does not look at the actual volumes sold.

n = 4000 p = 5

Predictors : For each book, it knows the,

(1) number of books by the same author sold in the past,

(2) the book's theme,

(3) the book's length,

(4) whether it got mostly good or bad reviews,

(5) the target audience.

Response, Sell Rate (a best-seller, a good sell, a bad sell, or a horrible sell)

(d) Scientists are very concerned with determining by how much global temperatures will rise in the coming years. They have annual data since 1900 on the world's temperature, the carbon, nitrous oxide and methane levels in the atmosphere, the world's GDP and a size estimate for Earth's polar ice caps.

This is a Regression problem. The response is rise in world temperature is quantitative. They are also interested in the rise in the temperature and not the factors behind this - therefore it is a prediction problem, and not inference.

n = 115 (Annual data since 1900 inclusive, and presumably up until 2014. I am assuming 2015 is not yet available.)

p = 6

Predictors,

(1) historical world's temperature,

(2) the carbon,

(3) nitrous oxide and

(4) methane levels in the atmosphere,

(5) the world's GDP and

(6) a size estimate for Earth's polar ice caps.

Response, rise in world temperature.

3. (a) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification?

Advantages of a Very Flexible Model

A flexible model can take full advantage of a large sample size.

A flexible model allows to find nonlinear effects.

A flexible model tends to work better when the variance of the data points is small.

Disadvantages of a Very Flexible Model

A flexible model can be prone to overfitting of the predictors in a high dimensional space.

A flexible model can be prone to overfitting of the known data points especially when the variance (and associated error or noise) is high.

A flexible model will not work as well with a small number of data points - an inflexible model would also not work well, however would perform better as it would not overfit on the limited data points.

(b) Describe two different circumstances (in terms of the number of samples and predictors, the true relationship between response and predictors, and the amount of noise in the problem) under which a more flexible approach would be preferred.

Explain your answer.

A more flexible approach would be preferred when, the number of samples, n, is extremely large, with a low variance and the number of predictors p is small In this case as there is low variance in the data points and a high number of data points, the given data points would represent the model well. In addition as there is a low number of predictors we will not be very prone to overfitting of the data points.

A more flexible approach would be preferred when the samples have a low variance and the relationship between the response and predictors in highly non-linear. In this case only a very flexible model can represent the very high linearity and as there is low variance in the given data points a flexible model should represent the true model well.

(c) Describe two different circumstances (in terms of the number of samples and predictors, the true relationship between response and predictors, and the amount of noise in the problem) under which a less flexible approach would be preferred.

Explain your answer.

A more inflexible approach would be preferred when the number of samples is small and variance is very high. The inflexible model will fit the general trend of the data as opposed to trying to map every point.

A more inflexible approach would be preferred when the number of predictors is extremely high and the relationship between the response and the predictors is linear. The inflexible model will

again, fit the general trend of the data and predictors - to give an inflexible linear model. Whereas a flexible model would not fit the model linearly due to the many predictors.

4. This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

first we load the package containing the dataset.

```
library(MASS)
```

Now the data set is contained in the object Boston.

```
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

How many rows are in this data set? How many columns? What do the rows and columns represent?

Lets look at the number of rows and columns.

```
dim(Boston)
```

```
## [1] 506  14
```

The Boston data frame has 506 rows and 14 columns. The rows represent suburbs of Boston. The columns represent attributes of suburb which can be used to predict the housing values. This data frame contains the following columns:

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.

black 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

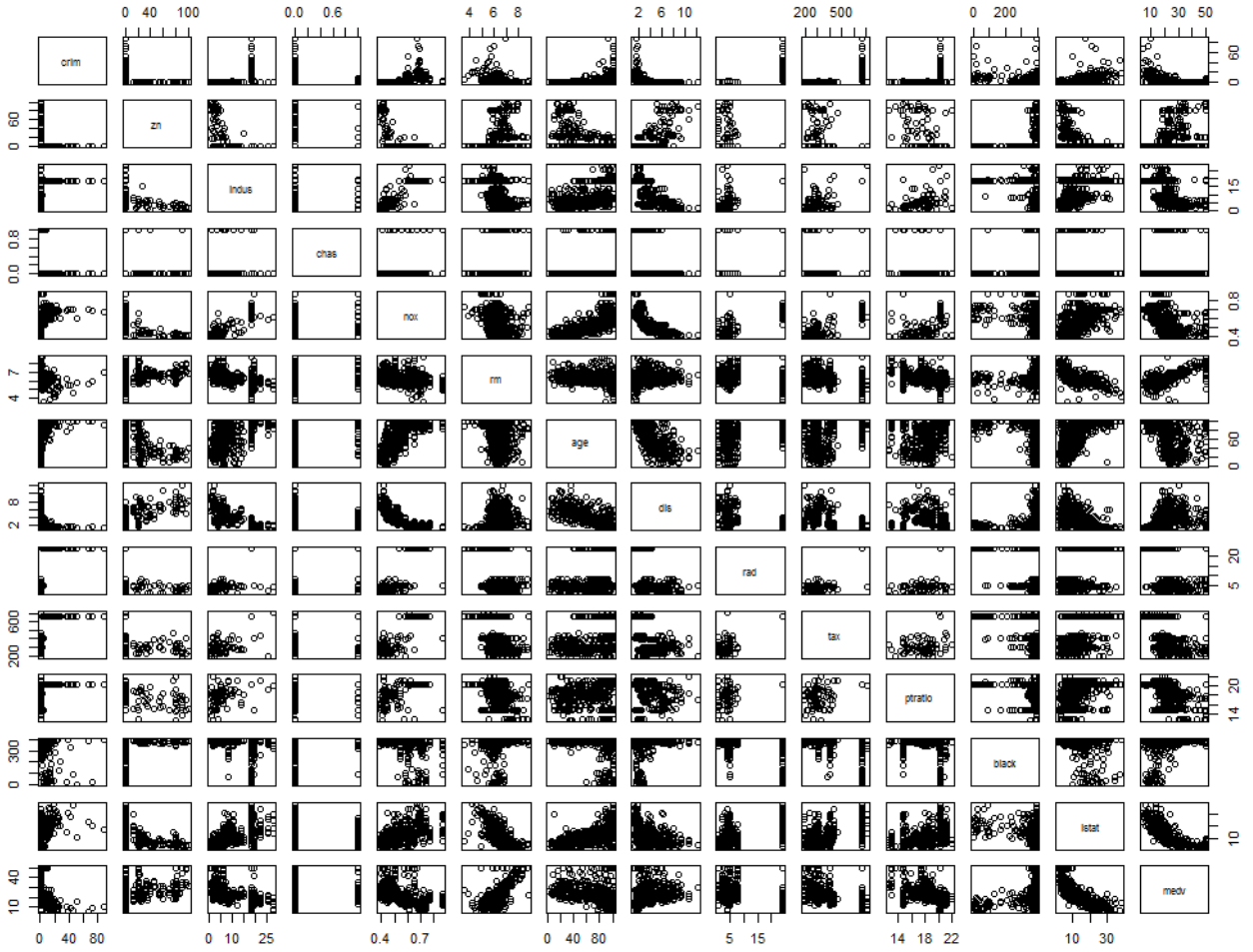
lstat lower status of the population (percent).

medv median value of owner-occupied homes in $1000s.

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your fndings.

In studying all features in a pairwise scatter plot (as seen below) it is is difficult to see where correlation seems high due to the amount of data on the page, however some pairs seem highly correlated even from this high level. For example, tax (property tax per $10k) and rad (index of accessibility to highways).

```
pairs(Boston)
```

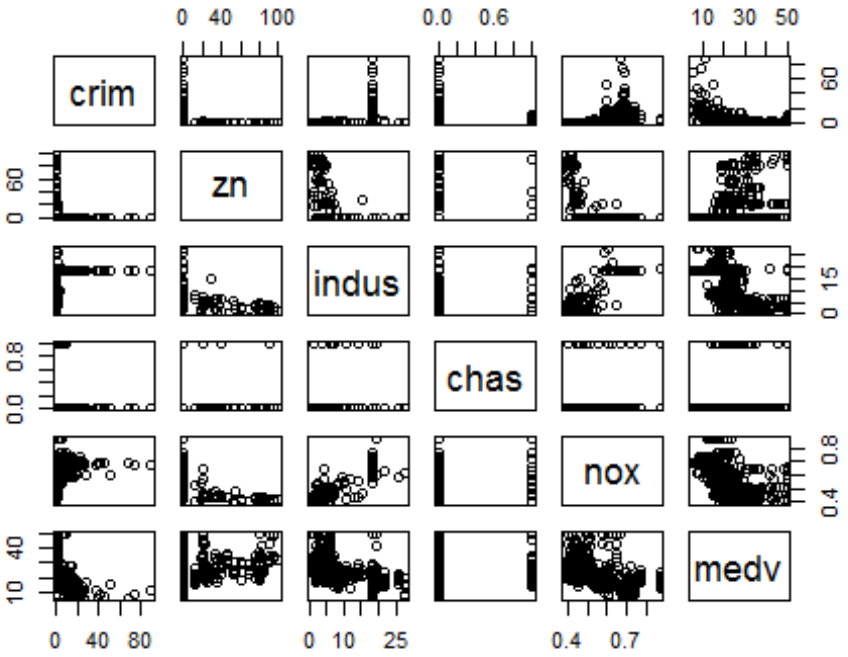


Drilling into some scatter plots, which compare the median value (medv) against other features in the data frame, we can see high correlations between medv & lstat (lower status of the population) and also medv & rm (average number of rooms per dwelling). lstat is negatively correlated to medv,

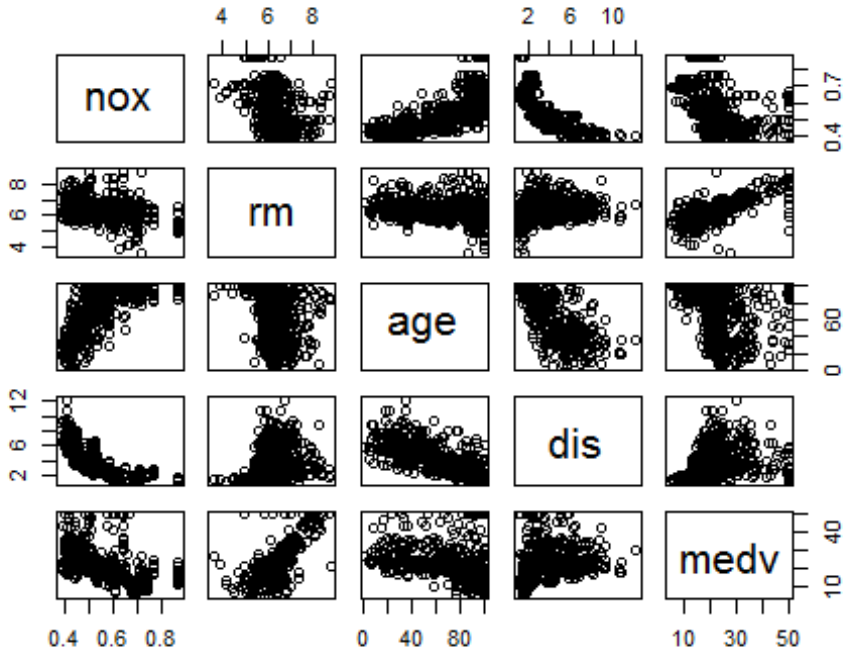
whereas rm is positively correlated which would be expected as the higher value areas would typically have larger dwellings.

Lower correlation of median value is seen with chas (Charles River Dummy Variable) and dis (distance to employment centres).

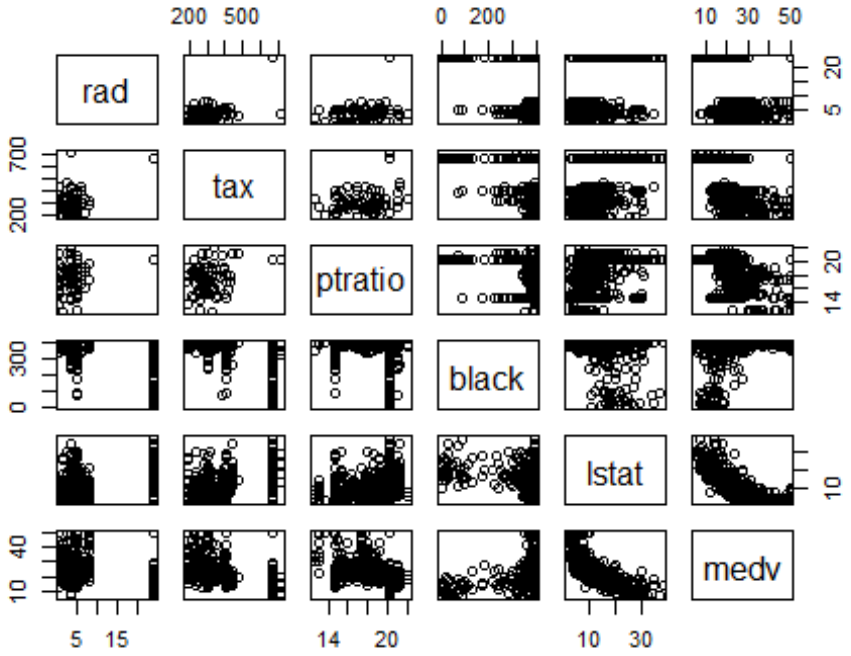
```
pairs(Boston[,c(1:5,14)])
```




```
pairs(Boston[,c(5:8,14)])
```




```
pairs(Boston[,c(9:14)])
```



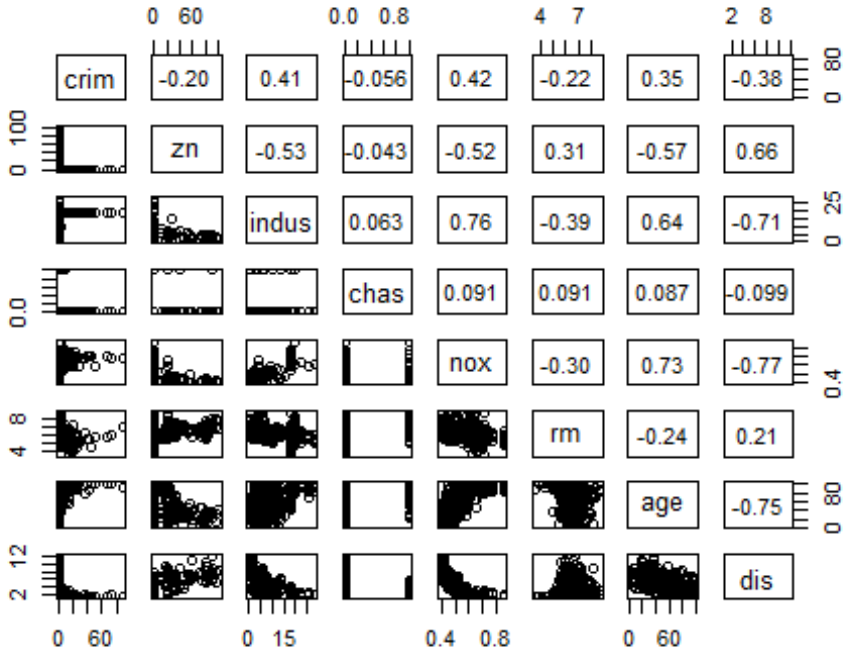
(c) Which predictors appear to associated with per capita crime rate? What are the apparent relationships?

From the below graphs of pairwise plots showing the correlation with per capita crime rate, column crim, the highest correlation appear to be the following features in order of the highest correlation to crim :

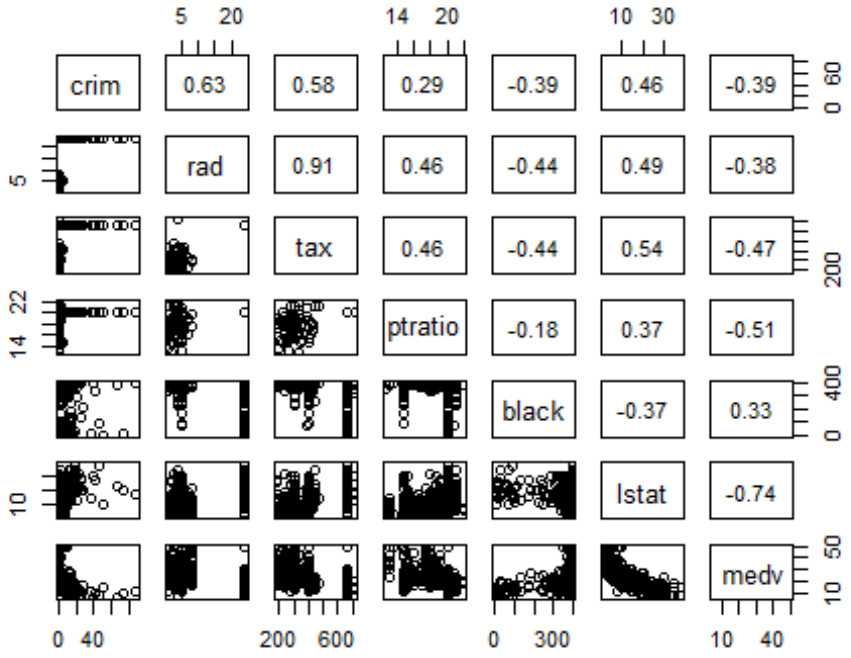
1. rad index of accessibility to radial highways. There appears to be a positive correlation to crime rate
2. tax full-value property-tax rate per $10,000. There appears to be a positive correlation to crime rate.
3. lstat lower status of the population (percent). There appears to be a positive correlation to crime rate.

In all cases, many areas have low crime (per capita crime rate <20), regardless of the accessibility to highways, tax rate or lower status of the population. However there are a few outlying suburbs with very high crime rates.

```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
     usr <- par("usr"); on.exit(par(usr))
     par(usr = c(0, 1, 0, 1))
     r <- cor(x, y)
     txt <- format(c(r, 0.123456789), digits = digits)[1]
     txt <- paste0(prefix, txt)
     if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
     text(0.5, 0.5, txt, cex = 1)
 }
pairs(Boston[,c(1:8)], upper.panel = panel.cor)
```

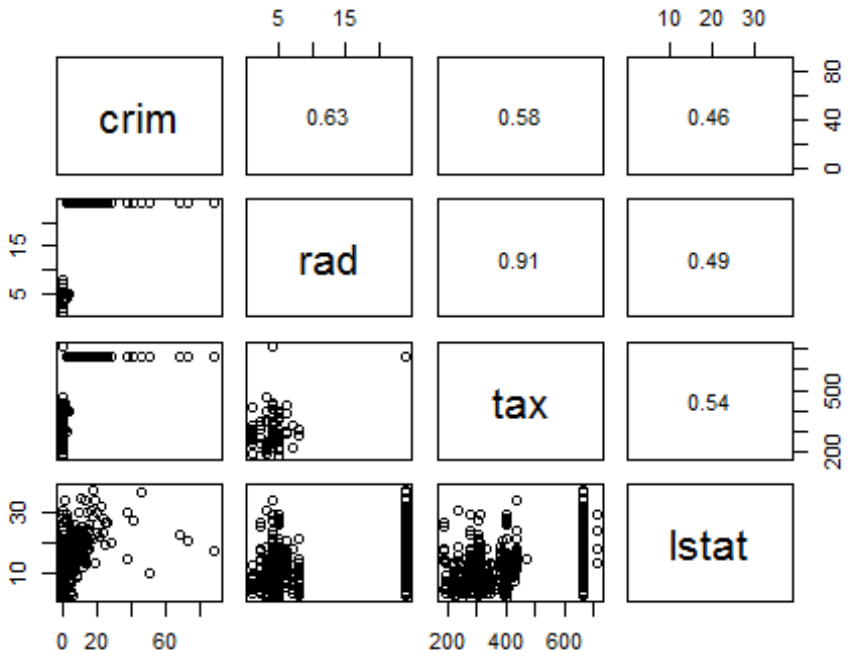



```
pairs(Boston[,c(1,9:14)], upper.panel = panel.cor)
```



Below is summarized the predictors with the highest correlation.

```
pairs(Boston[,c(1,9:10,13)], upper.panel = panel.cor)
```



(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Below can be seen a box plot for each of the variables in the data frame. The points lying outside the whiskers (>1.5 * Interquartile Range) are the outliers.

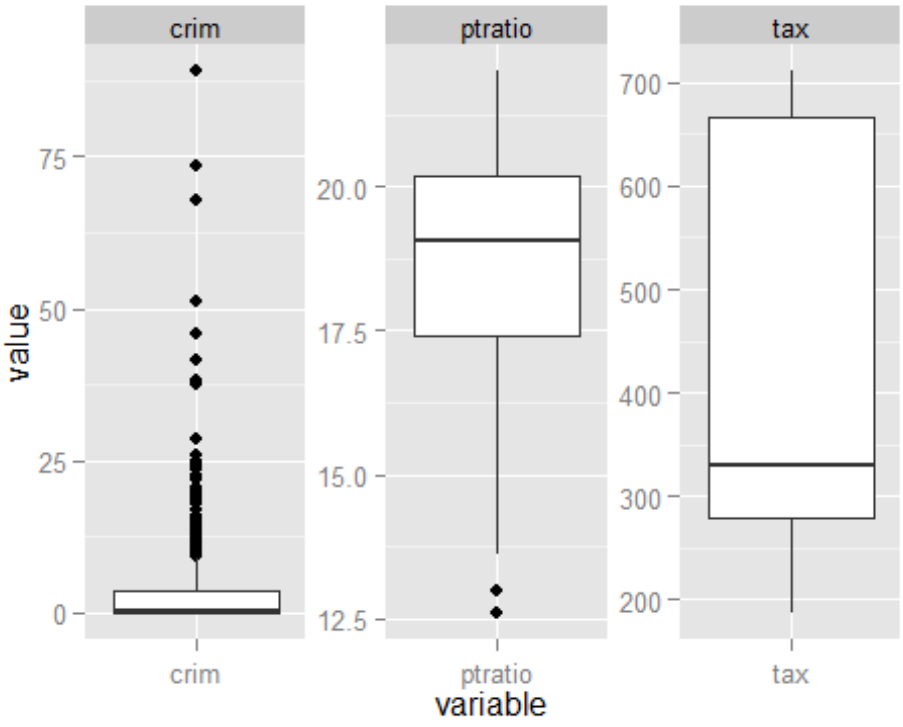
crim
The per capita crime rate by town. This appears to have many outliers as seen in the boxplot below. The majority of towns have close to zero crime rate (of under 5), while some suburbs have a rate as high as 70 or 80. There are many outlier suburbs marked as outliers indicating most suburbs have a very low crime rate, although there are a significant number of towns in the minority with an outlying crime rate on the higher end (eg. 54 towns have crim > 10).

tax
The full-value property-tax rate per $10,000. The property tax ranges from approximately 200 to 700, with no extreme outliers. The median lies around 320 dollars.

ptratio
The pupil-teacher ratio by town. the pupil teach ratio ranges from 12.5 to approximately 22. The Interquartile range lies between 17.5 and 20, and there are a couple of non extreme outliers around the 12.5 range.

```
#load two packages two reshape the data and create the boxplot
library(ggplot2)
library(reshape2)
#create a new data frame with two columns only (variable, value) for all three predic
tors
```

```
mdata <- melt(Boston[,c(1,11,10)])
# Output the boxplot
p <- ggplot(data = mdata, aes(x=variable, y=value)) +
             geom_boxplot()
p + facet_wrap( ~ variable, scales="free", ncol=3)
```



(e) How many of the suburbs in this data set bound the Charles river?

The variable chas is the Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). In this case 35 suburbs are bound by the Charles river.

```
# Attach the database, so variables can be accessed by simply giving their names
attach(Boston)
# Explore how many of each types of chas exist
table(chas)
```

```
## chas
##   0   1
## 471  35
```

(f) What is the median pupil-teacher ratio among the towns in this data set?

As can be seen below and in the boxplot above the pupil teacher ration is approximately 19 pupils per teacher.

```
summary(ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

(g) Which suburb of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your fndings.

There are two suburbs which have the lowest median value of owner-occupied homes. These can be seen below. Also can be seen the values of the other predictors for those two suburbs.

```
Boston[medv == min(medv),]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio  black
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97
##     lstat medv
## 399 30.59    5
## 406 22.98    5
```

The table below compares the suburbs with the lowest median value of owner-occupied homes, to some summary statistics of the wider population of all towns. The crime rates are obviously on the very upper end of the city, although not the suburbs with the highest crime.

The proportion of residential land zoned for lots over 25,000 sq.ft. is 0, which is the minimum of the city and indicating that there is little investment in these suburbs.

The proportion of non-retail business acres per town, lies within the 3rd quartiles for both suburbs. this indicates there are viable businesses and potentially employment in the area.

The Charles River dummy variable only indicates the suburbs do not lie on the river side.

The nitrogen oxides concentration (parts per 10 million) are in the upper quartile of the city, perhaps since the suburbs are so close to the highways.

The average number of rooms per dwelling is in the lower quartile indicating smaller apartments or houses, however it is not at a minimimum. Perhaps the areas closer to the city have smaller apartments instead of houses.

The age proportion of owner-occupied units built prior to 1940, is at the maximum indicating old housing units and no new builds.

The weighted mean of distances to five Boston employment centres is well into the lower quartile and close to the minimum. this indicates an area of high unemployment.

The index of accessibility to radial highways is at the maximum indicating that the areas lit on or very near a highway.

The full-value property-tax rate per $10,000 is quite high and in the upper quartile. Perhaps this is due to the small apartments/houses, in a city where the tax and unit area is not linearly correlated - so larger units actually pay less per square feet.

The pupil-teacher ratio in these areas is in the upper quartile suggesting some relative under investment in schooling. H

The variable black is around the median of all suburbs, indicating not a predominantly black population compared to the rest of the city.

The lower status of the population (percent) of population is close the maximum and high in the upper quartile.

```
# load a package which gives a flexible view of summary statistics in a data frame
library('fBasics')
# Find the records where medv is a minimum in the dataset
Minmedv <- Boston[medv == min(medv),]
# Indicate in the rownames that these rows are from the medv minmumu set
rownames(Minmedv) <- paste("Min Medv row", rownames(Minmedv))
# join these the two minmedv records with the summary statistics of the larger data s
et
rbind(Minmedv, as.data.frame(basicStats(Boston))[3:8,])
```

```
##                        crim        zn    indus    chas      nox       rm
## Min Medv row 399 38.351800   0.00000 18.10000 0.00000 0.693000 5.453000
## Min Medv row 406 67.920800   0.00000 18.10000 0.00000 0.693000 5.683000
## Minimum           0.006320   0.00000  0.46000 0.00000 0.385000 3.561000
## Maximum          88.976200 100.00000 27.74000 1.00000 0.871000 8.780000
## 1. Quartile       0.082045   0.00000  5.19000 0.00000 0.449000 5.885500
## 3. Quartile       3.677082  12.50000 18.10000 0.00000 0.624000 6.623500
## Mean              3.613524  11.36364 11.13678 0.06917 0.554695 6.284634
## Median            0.256510   0.00000  9.69000 0.00000 0.538000 6.208500
##                       age       dis       rad      tax  ptratio    black
## Min Medv row 399 100.0000  1.489600 24.000000 666.0000 20.20000 396.9000
## Min Medv row 406 100.0000  1.425400 24.000000 666.0000 20.20000 384.9700
## Minimum            2.9000  1.129600  1.000000 187.0000 12.60000   0.3200
## Maximum          100.0000 12.126500 24.000000 711.0000 22.00000 396.9000
## 1. Quartile       45.0250  2.100175  4.000000 279.0000 17.40000 375.3775
## 3. Quartile       94.0750  5.188425 24.000000 666.0000 20.20000 396.2250
## Mean              68.5749  3.795043  9.549407 408.2372 18.45553 356.6740
## Median            77.5000  3.207450  5.000000 330.0000 19.05000 391.4400
##                     lstat     medv
## Min Medv row 399 30.59000  5.00000
## Min Medv row 406 22.98000  5.00000
## Minimum           1.73000  5.00000
## Maximum          37.97000 50.00000
## 1. Quartile       6.95000 17.02500
## 3. Quartile      16.95500 25.00000
## Mean             12.65306 22.53281
## Median           11.36000 21.20000
```

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

64 suburbs average more than seven rooms per dwelling.

```
nrow(Boston[rm>7,])
```

```
## [1] 64
```

13 suburbs average more than eight rooms per dwelling.

```
nrow(Boston[rm>8,])
```

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

As can be seen in the two tables below comparing the 13 suburbs with on average over eight rooms per dwellings have a low crime rate, generally a high median value of the dwellings (with one exception under average) and generally a low pupil teacher ratio. The full-value property-tax rate per $10,000 is low, indicating that property tax is not judged off house value. In general there are not many non retail busiensses indicating they are purely residential areas. There are generally older houses with some exceptions with very low proportion of old buildings. They seem to be further away from the highways than average, again indicating purely residential areas perhaps on the outskirts. Other variables not mentioned above seem to be broadly consistent with the wider population.

Summary Statistics of whole population

```
Highrm <- Boston[rm>8,]
rownames(Highrm) <- paste("(RM > 8)", rownames(Highrm))
as.data.frame(basicStats(Highrm))[3:8,]
```

```
##                  crim       zn     indus     chas      nox       rm
## Minimum     0.020090  0.00000  2.680000 0.000000 0.416100 8.034000
## Maximum     3.474280 95.00000 19.580000 1.000000 0.718000 8.780000
## 1. Quartile 0.331470  0.00000  3.970000 0.000000 0.504000 8.247000
## 3. Quartile 0.578340 20.00000  6.200000 0.000000 0.605000 8.398000
## Mean        0.718795 13.61538  7.078462 0.153846 0.539238 8.348538
## Median      0.520140  0.00000  6.200000 0.000000 0.507000 8.297000
##                  age      dis       rad      tax  ptratio    black lstat
## Minimum      8.40000 1.801000  2.000000 224.0000 13.00000 354.5500  2.47
## Maximum     93.90000 8.906700 24.000000 666.0000 20.20000 396.9000  7.44
## 1. Quartile 70.40000 2.288500  5.000000 264.0000 14.70000 384.5400  3.32
## 3. Quartile 86.50000 3.651900  8.000000 307.0000 17.40000 389.7000  5.12
## Mean        71.53846 3.430192  7.461538 325.0769 16.36154 385.2108  4.31
## Median      78.30000 2.894400  7.000000 307.0000 17.40000 386.8600  4.14
##             medv
## Minimum     21.9
## Maximum     50.0
## 1. Quartile 41.7
## 3. Quartile 50.0
## Mean        44.2
## Median      48.3
```

Summary Statistics of dwellings with large number of rooms (average >8 rooms / dwelling)

```
as.data.frame(basicStats(Boston))[3:8,]
```

```
##                  crim        zn    indus    chas      nox       rm
## Minimum      0.006320   0.00000  0.46000 0.00000 0.385000 3.561000
## Maximum     88.976200 100.00000 27.74000 1.00000 0.871000 8.780000
## 1. Quartile  0.082045   0.00000  5.19000 0.00000 0.449000 5.885500
## 3. Quartile  3.677082  12.50000 18.10000 0.00000 0.624000 6.623500
## Mean         3.613524  11.36364 11.13678 0.06917 0.554695 6.284634
## Median       0.256510   0.00000  9.69000 0.00000 0.538000 6.208500
##                  age       dis       rad      tax  ptratio    black
## Minimum       2.9000  1.129600  1.000000 187.0000 12.60000   0.3200
## Maximum     100.0000 12.126500 24.000000 711.0000 22.00000 396.9000
## 1. Quartile  45.0250  2.100175  4.000000 279.0000 17.40000 375.3775
## 3. Quartile  94.0750  5.188425 24.000000 666.0000 20.20000 396.2250
## Mean         68.5749  3.795043  9.549407 408.2372 18.45553 356.6740
```

```
## Median       77.5000  3.207450  5.000000 330.0000 19.05000 391.4400
##                lstat     medv
## Minimum      1.73000  5.00000
## Maximum     37.97000 50.00000
## 1. Quartile  6.95000 17.02500
## 3. Quartile 16.95500 25.00000
## Mean        12.65306 22.53281
## Median      11.36000 21.20000
```

5. In this exercise, we will predict per capita crime rate by town using the other variables in the Boston data set.

(a) Split the data set into a training set and a test set of approximately equal size. Include your R code.

Here we will use the caret package to randomly split the data set into two folds of roughly equal size. We set the seed so we always get the same split when running it. The first fold will be our training set, and the second fold our test set.

```
library(caret)
set.seed(32343)
folds <- createFolds(y=Boston$crim[], k=2, list=TRUE, returnTrain=TRUE)
```

Lets take a look at the number of rows in both the training and test set.

```
c(nrow(Boston[folds[[1]],]), nrow(Boston[folds[[2]],]))
```

```
## [1] 254 252
```

Now, lets look at the first 5 rows of the training set.

```
head(Boston[folds[[1]],], n=5)
```

```
##       crim   zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 2  0.02731  0.0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3  0.02729  0.0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 5  0.06905  0.0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6  0.02985  0.0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
## 11 0.22489 12.5  7.87    0 0.524 6.377 94.3 6.3467   5 311    15.2 392.52
##    lstat medv
## 2   9.14 21.6
## 3   4.03 34.7
## 5   5.33 36.2
## 6   5.21 28.7
## 11 20.45 15.0
```

(b) Fit a linear model using least squares on the training set, and report the mean training and mean test error obtained.

Fit a linear model as seen below on the training set.

```
fit <- lm(crim ~., data=Boston[folds[[1]],])
```

Calculate out the mean training squared error using the formula : mean((actual.values - predicted.values)^2). Average RSS (as given by equation 3.22 in the book)

```
mean((Boston[folds[[1]],]$crim - (predict(fit,Boston[folds[[1]],])))^2)
```

```
## [1] 33.62213
```

Calculate out the mean test squared error in a similar fashion :

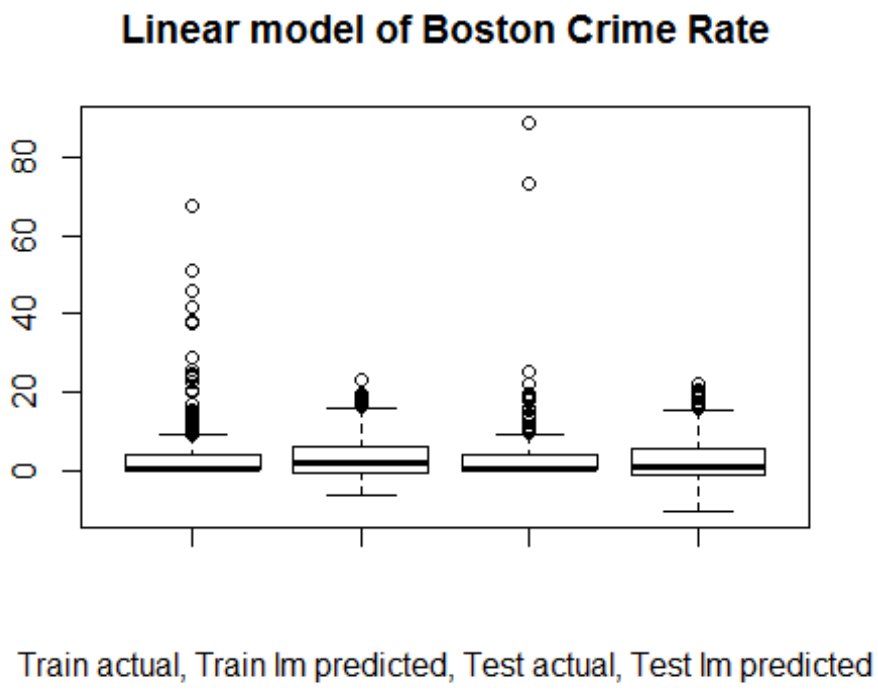
```
mean((Boston[folds[[2]],]$crim - (predict(fit,Boston[folds[[2]],])))^2)
```

```
## [1] 50.92324
```

(c) Comment on the results obtained. How accurately can we predict per capita crime rate by town? What are the most important predictors?

Below can be seen a boxplot of the training and test data set with actual values and predicted values. An interesting point when comparing the predicted and actual values is that we get a number of negative values predicted. This is most likely due to the inflexibility of the linear model and would not occur in a more flexible model. Also we see in the actual values we have most data points tightly bunched together with low crime rates then a large minority of outliers with high crime rates. The predicted values fail to pick up fully on this trend which again may be picked up with a more flexible model. this is perhaps the reason we get a high mean squared error on both the train and test data set. To sum up the model predicts reasonably well, however fails to pick up on the larger variance and most importantly incorrectly predicts values as negative.

```
boxplot(Boston$crim[folds[[1]]], predict(fit,Boston[folds[[1]],]), Boston$crim[folds[
[2]]], (predict(fit,Boston[folds[[2]],])), main="Linear model of Boston Crime Rate",
xlab="Train actual, Train lm predicted, Test actual, Test lm predicted")
```



The important predictors (very high significance) are : rad index of accessibility to radial highways. dis weighted mean of distances to five Boston employment centres.

Some less important predictors, however with p-values <0.05 are zn, nox, black, lstat and medv.

```
summary(fit)
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston[folds[[1]], ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.324  -2.420  -0.682   1.402  51.407
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.272251   9.584394   3.472 0.000614 ***
## zn            0.053297   0.024766   2.152 0.032389 *
## indus        -0.123539   0.111692  -1.106 0.269804
## chas         -1.818952   1.408245  -1.292 0.197723
## nox         -16.598468   6.854982  -2.421 0.016205 *
## rm           -0.818282   0.777452  -1.053 0.293621
## age          -0.002726   0.025796  -0.106 0.915930
## dis          -1.188060   0.389571  -3.050 0.002548 **
## rad           0.748279   0.119718   6.250 1.85e-09 ***
## tax          -0.008953   0.006914  -1.295 0.196573
## ptratio      -0.378939   0.249013  -1.522 0.129384
## black        -0.011108   0.004801  -2.314 0.021522 *
## lstat         0.200481   0.099688   2.011 0.045434 *
## medv         -0.215124   0.098226  -2.190 0.029480 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.965 on 240 degrees of freedom
## Multiple R-squared:  0.5608, Adjusted R-squared:  0.537
## F-statistic: 23.57 on 13 and 240 DF,  p-value: < 2.2e-16
```

(6) Using the same setup as in the previous question, form a new outcome variable Y which equals one if per capita crime rate by town (crim) is greater than or equal to the overall median and zero otherwise. Fit a logistic regression model to Y , and report the training and test misclassifcation rates and the most important predictors. Compare the results of this analysis to that of the linear regression approach in the previous question.

Create variable Y as described in the question above.

```
Y <- ifelse(Boston$crim>=median(Boston$crim), 1, 0)
```

Fit the model using only the first fold. I am assuming we do not use crim as a predictor, as this was used to create the response Y.

```
glm.fit <- glm(Y[folds[[1]]] ~ . -crim, data=Boston[folds[[1]],], family=binomial)
```

Make a prediction, using logistic regression for the whole dataset

```
pred <- ifelse(predict(glm.fit,Boston)>=median(Boston$crim), 1, 0)
```

Produce a confusion matrix of the training misclassifcation rates :

```
table(Y[folds[[1]]],pred[folds[[1]]])
```

```
##
##       0   1
##   0 118   9
##   1  14 113
```

Misclassification error rate on the training set

```
sum(Y[folds[[1]]]!=pred[folds[[1]]])/length(folds[[1]])
```

```
## [1] 0.09055118
```

Produce a confusion matrix of the test misclassifcation rates :

```
table(Y[folds[[2]]],pred[folds[[2]]])
```

```
##
##       0   1
##   0 117   9
##   1  13 113
```

Misclassification error rate on the test set :

```
sum(Y[folds[[2]]]!=pred[folds[[2]]])/length(folds[[2]])
```

```
## [1] 0.08730159
```

The important predictors (very high significance) are : nox (very high significance), age and rad. Less important predictors, however with p-values <0.05 are zn, age and ptratio.

```
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Y[folds[[1]]] ~ . - crim, family = binomial, data = Boston[folds[[1]
],
##     ])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.83201  -0.10166  -0.00001   0.00134   2.33672
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -51.525857  11.547526  -4.462 8.12e-06 ***
## zn           -0.147784   0.061408  -2.407  0.01610 *
## indus        -0.110173   0.067300  -1.637  0.10162
## chas          0.896842   0.935082   0.959  0.33751
## nox          69.375067  13.654635   5.081 3.76e-07 ***
## rm           -0.696391   1.070615  -0.650  0.51540
## age           0.052940   0.024127   2.194  0.02822 *
## dis           1.339633   0.425032   3.152  0.00162 **
## rad           0.700511   0.236516   2.962  0.00306 **
## tax          -0.008526   0.004541  -1.878  0.06044 .
## ptratio       0.393910   0.189612   2.077  0.03776 *
## black        -0.009971   0.005684  -1.754  0.07941 .
## lstat         0.095849   0.078249   1.225  0.22060
## medv          0.246808   0.113969   2.166  0.03034 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 352.119  on 253  degrees of freedom
## Residual deviance:  97.248  on 240  degrees of freedom
## AIC: 125.25
##
## Number of Fisher Scoring iterations: 9
```

Interestingly both models showed different predictors and significance of predictors. zn, dis and medv had equal significance in both models nox had very high significance in the lm model but not very high significance in lm. The opposite was true for rad(although the significance was quite high in the lm model also). black and lstat only showed significance in the linear model. For the difference in the predictors found, I would see two reasons for this difference, 1) The logistic regression only had a binary response to predict (greater than median crim or not), while the linear model needed to predict the rate of crime. Predicting the rate of crime need to differentiate between high levels of crim where factors different factors may come in. 2) Correlation between predictors.

Now we compare in how many cases the modified output from the linear regression (in question 5) agrees with the output from the logistic regression. We transform the linear regression output for the test data set into a qualitative response in a similar fashion to what was done above for logistic regression - take it to be 1 if it is above median, and 0 otherwise. To compare the linear regression and the logistic regression We start be producing a confusion matrix. Here we see that the logistic regression has a lower error rate of both results - above and below median crime rate.

```
# Make a prediction, using logistic regression for the whole dataset
pred_lm <- ifelse(predict(fit,Boston)>=median(Boston$crim), 1, 0)
# Produce a confusion matrix of the linear model with the training misclassifcation r
ates :
table(Y[folds[[2]]],pred_lm[folds[[2]]])
```

```
##
##      0  1
##   0 64 62
##   1 35 91
```

Here we see that the misclassification error rate on the training set for the linear regression is 0.38, while the misclassification for logistic regression was 0.09. This would support the view that for classification problems logistic regression is often more suited than linear regression.

```
# Misclassification error rate of the linar model with the on training set
sum(Y[folds[[2]]]!=pred_lm[folds[[2]]])/length(folds[[2]])
```

```
## [1] 0.3849206
```