Exam: Predictive Modeling

Stuti Madaan (sm63332)

August 04, 2016

Exam Questions

Exam Question 1

Part -1 Using the data, estimate the effect of "beauty" into course ratings. Make sure to think about the potential many "other determinants". Describe your analysis and your conclusions.

(A)Linear model Score vs beauty:

```
beauty = read.csv("E:/R Directory/csv_files/BeautyData.csv")
beauty.fit<- glm(CourseEvals~BeautyScore, data =beauty)
summary(beauty.fit)
```

```
##
## Call:
## glm(formula = CourseEvals ~ BeautyScore, data = beauty)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5936  -0.3346   0.0097   0.3702   1.2321
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.71340    0.02249 165.119   <2e-16 ***
## BeautyScore  0.27148    0.02837   9.569   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2312617)
##
##     Null deviance: 127.79  on 462  degrees of freedom
## Residual deviance: 106.61  on 461  degrees of freedom
## AIC: 640.01
##
## Number of Fisher Scoring iterations: 2
```

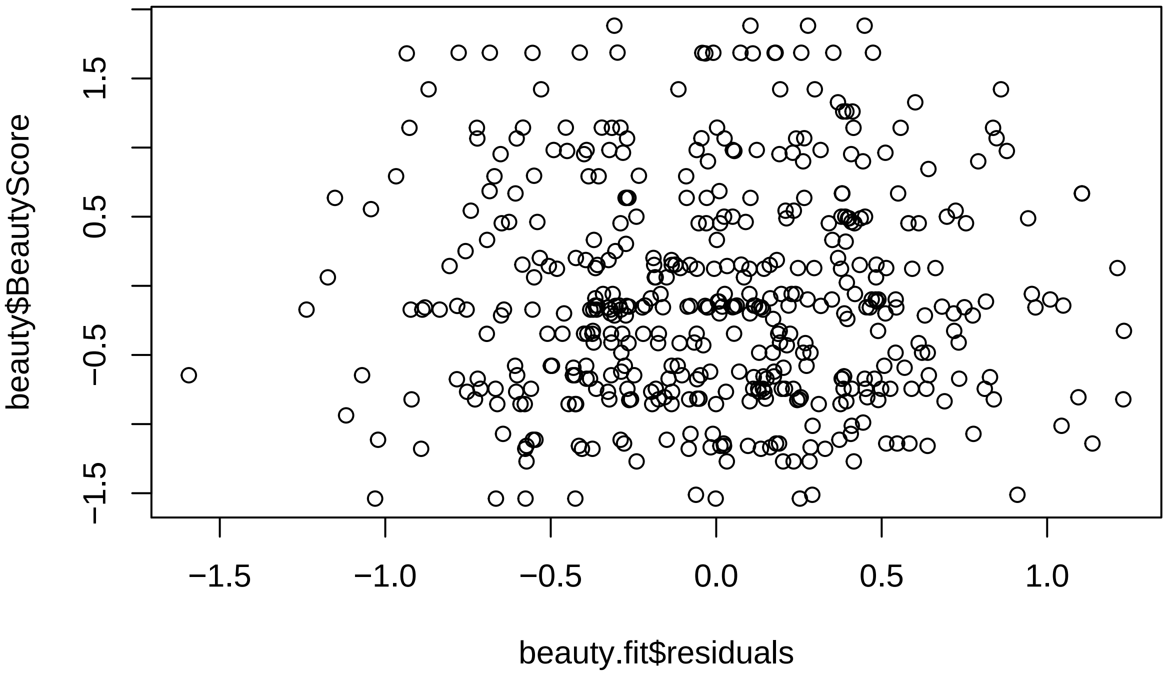
```
confint(beauty.fit)
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %    97.5 %
## (Intercept) 3.6693213 3.7574776
## BeautyScore 0.2158755 0.3270809
```

```
beauty$pred = predict(beauty.fit, beauty)
```

It can be seen below that all the variance has been explained by beauty variable through a linear relationship.



```
## RMSE from CourseEvals vs BeautyScore: 0.479857
```

RMSE of 0.4798 with a confidence interval not containing zero shows that there is a correlation between BeautyScore and CourseEvals. The correlation is positive as an be judged by the positive slope. The plot above shows that the residuals are random when plotted against Beauty Score. This implies that all the variation in CourseEvals that is dependent on BeautyScore has been explained by the linear relationship between the two. Even though Beauty seems to affect the course ratings, in reality, there can be several other factors that will impact the CourseEvals. For example, it can be that the people, who are physicaly fit(therefore, higher beauty score) are more organized and effective in their teaching skills. It is also possible that the faculty with a higher beauty score is younger and employs more creative methods for educating the students. This will again lead to a higher CourseEvals for a faculty with higher beauty score.

Thus, it can be concluded that a higher beauty score may not be directly impacting the CourseEvals. The relationship between Beauty score and CourseEvals could be more of a correlation rather than causal effect.That’s why we might get a positive relationship. The ‘true’ factors affecting CourseEvals might be closely related to ‘beauty score’.

Trying out the combination of beauty and other variables:

(B)Linear model Score vs beauty and Tenure Track

If we look at the Confidence interval for tenuretrack, it contains zero. Thus, we can't be certain of the relationship between Tenure track and CourseEvals. Also, the p-value is very low for tenuretrack which shows that it does not predict CourseEvals all that well.

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)  3.74284902 0.04767262 78.5115011 1.451005e-268
## BeautyScore  0.27109757 0.02839013  9.5490093  7.808259e-20
## tenuretrack -0.03781361 0.05396450 -0.7007127  4.838366e-01
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept)  3.6494124 3.83628564
## BeautyScore  0.2154540 0.32674120
## tenuretrack -0.1435821 0.06795486
```

```
## RMSE from CourseEvals vs BeautyScore & tenuretrack: 0.4796011
```

(C)linear model score vs Beauty and nonenglish

We can see that both BeautyScore and nonenglish have positive and negative correlation,respectively, with CourseEvals. Also, the confidence intervals do not contain zero thus, the null hypothesis is rejected for both these variables. However, the marginal reduction in the RMSE because of nonenglish is very less.

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  3.7246587 0.02312604 161.059086 0.000000e+00
## BeautyScore  0.2720555 0.02828097   9.619735 4.411227e-20
## nonenglish  -0.1853362 0.09346848  -1.982874 4.797591e-02
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %       97.5 %
## (Intercept)  3.6793325  3.769984902
## BeautyScore  0.2166258  0.327485138
## nonenglish  -0.3685310 -0.002141331
```

```
## RMSE from CourseEvals vs BeautyScore & nonenglish: 0.4778193
```

(D) linear model score vs Beauty + non-English + female

It can be seen that the female variable is very predictive in determining the CourseEvals. i.e. RMSE has reduced directly to 0.21. Also, the confidence interval does not contain zero.Thus, we can be certain of the negative correlation between female and CourseEvals.

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  3.8554573 0.02873129 134.190176 0.000000e+00
## BeautyScore  0.2961408 0.02710015  10.927646 7.280214e-25
## nonenglish  -0.1837458 0.08885588  -2.067908 3.920839e-02
## female      -0.3057388 0.04323760  -7.071131 5.761231e-12
## Waiting for profiling to be done...

##                  2.5 %       97.5 %
## (Intercept)  3.7991450  3.911769615
## BeautyScore  0.2430255  0.349256166
## nonenglish  -0.3579001 -0.009591451
## female      -0.3904829 -0.220994636

## RMSE from CourseEvals vs BeautyScore, nonenglish and female: 0.2058835
```

(E) linear model score vs Beauty + non English + Female + Lower

> Lastly,adding the variable ‘lower’ further reduces the error to 0.18 with a negative correlation between lower and CourseEvals.

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  3.9826158 0.03158804 126.079845 0.000000e+00
## BeautyScore  0.3044656 0.02551939  11.930757 9.237229e-29
## nonenglish  -0.2783854 0.08447957  -3.295299 1.059507e-03
## female      -0.3252310 0.04075680  -7.979796 1.187452e-14
## lower       -0.3317386 0.04263690  -7.780551 4.835206e-14

## Waiting for profiling to be done...

##                  2.5 %     97.5 %
## (Intercept)  3.9207043  4.0445272
## BeautyScore  0.2544485  0.3544827
## nonenglish  -0.4439623 -0.1128085
## female      -0.4051128 -0.2453491
## lower       -0.4153053 -0.2481718

## RMSE from CourseEvals vs BeautyScore, nonenglish, female and lower: 0.1818475
```

Anlaysis:

> I linearly regressed the Evaluation Score against beauty score only to establish if we see a relationship significant enough to keep this variable as one of the determinants. A positive slope, lower p-value and 95% confidence interval greater than 0 establishes that there is a positive correlation between Evaluation Score and beauty score. i.e. For every one unit increase in beauty score, there is 0.27 points increase in Evaluation score.
>
> Further, I tried to add the variable tenuretrack to establish any relationship. However, adding tenuretrack to the equation only marginally improves the standard error. Also, the confidence interval for tenuretrack contains zero which establishes that it is not a statistically significant variable in this context. However, the significance of beauty score is intact.
>
> Sequentially adding the variables female, lower and non-English further reduce the error with negative correlations to the Evaluation Score. i.e. with increase in one unit of any of these variables (keeping others constant) will decrease the Evaluation Score by abs(coefficient) amount. From the analysis, it can be derived that ‘beauty score’ is one of the drivers of the Evaluation Score i.e. Keeping all other variables constant, one unit increase in beauty score will increase Evaluation score by 0.25 units. However, this is not the only major driver for Evaluation Score. Variables like male/Female, non-English, lower better determine the overall Evaluation Score. So, from the data at hand, it can be concluded that an English speaking woman not of lower division tends to have a higher Evaluation Score.

Part -2: In his paper, Dr. Hamermesh has the following sentence: "Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible". Using the concepts, we have talked about so far, what does he mean by that?

> The productivity of an instructor should have been the predictor of the Evaluation Score. However, productivity is something that cannot be measured directly and thus, we use proxy for productivity like non English, female, lower division to try and capture what all factors might be impacting students while evaluating the instructor. Now, the ratings by students are biased both by beauty and productivity of the instructor. For this scenario, beauty and productivity are probably highly correlated whihc makes it difficult to isolate the effects of their individual effects. Thus, it is not clear whether the final Evaluation Score is due to productivity or discrimination based on beauty. This is exactly what the author means by "Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible".

Exam Question 2

Use regression models to estimate the pricing structure of houses in this town and answer the following questions:

```r
#model fitting:
price.fit<- lm(Price~Nbhd+Offers+SqFt+Brick+Bedrooms+Bathrooms+ Nbhd:Brick ,data=final_data)
summary(price.fit)
```

```
##
## Call:
## lm(formula = Price ~ Nbhd + Offers + SqFt + Brick + Bedrooms +
##     Bathrooms + Nbhd:Brick, data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27225.1  -5219.0   -273.7   4297.4  27507.2
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3695.511   8829.382   0.419  0.67631
## Nbhd2           -1317.656   2679.849  -0.492  0.62385
## Nbhd3           16980.797   3437.529   4.940 2.60e-06 ***
## Offers          -8381.770   1068.248  -7.846 2.15e-12 ***
## SqFt               53.745      5.686   9.453 3.96e-16 ***
## BrickYes        12093.056   4082.168   2.962  0.00369 **
## Bedrooms         4777.216   1586.397   3.011  0.00318 **
## Bathrooms        6457.287   2160.867   2.988  0.00341 **
## Nbhd2:BrickYes   2668.449   5068.893   0.526  0.59957
## Nbhd3:BrickYes  11933.197   5341.027   2.234  0.02735 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9847 on 118 degrees of freedom
## Multiple R-squared:  0.8752, Adjusted R-squared:  0.8657
## F-statistic: 91.94 on 9 and 118 DF,  p-value: < 2.2e-16
```

```r
final_data$pred1 <- predict (price.fit, final_data)
```

1. **Is there a premium for brick houses everything else being equal?**

   Yes. The slope for brick houses is 12093 and intercept is 3695. Net quantity, being positive, implies that if it's a brick house, the price will be 12093 dollars more expensive as compared to non-brick houses, given all other variables are constant. Thus, there is a premium for the brick houses.

2. **Is there a premium for houses in neighborhood 3?**

   Yes. The slope for Nbhd3 is 16980 and intercept is 3695. Net quantity being positive implies that if it's in Neighborhood 3, the price will be 16980 dollars higher for it as compared to neighborhood 1, given all other factors do not change. However, for Neighborhood 2, the price will be 1317 dollars less than that of Neighborhood 1 given all other variables are same. Also, the price will be 15663 dollars higher than neighbor 2.Thus, there is a premium for houses in neighborhood 3.

3. **Is there an extra premium for brick houses in neighborhood 3?**

   Yes. We find this out by using the interaction term between Neighborhod variable and Brick variable. It can be observed below that the interaction between Neighborhood 3 and Brick yes is significant. i.e. A house in neighborhood 3 with bricks will be 11933 dollars more than the ones without the brick. So, there is a premium for houses in Neighborhood 3 with bricks or without bricks

```
##
## Call:
## lm(formula = Price ~ Nbhd + Offers + SqFt + Brick + Bedrooms +
##     Bathrooms + Nbhd:Brick, data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27225.1  -5219.0   -273.7   4297.4  27507.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3695.511   8829.382   0.419  0.67631
## Nbhd2          -1317.656   2679.849  -0.492  0.62385
## Nbhd3          16980.797   3437.529   4.940 2.60e-06 ***
## Offers         -8381.770   1068.248  -7.846 2.15e-12 ***
## SqFt              53.745      5.686   9.453 3.96e-16 ***
## BrickYes       12093.056   4082.168   2.962  0.00369 **
## Bedrooms        4777.216   1586.397   3.011  0.00318 **
## Bathrooms       6457.287   2160.867   2.988  0.00341 **
## Nbhd2:BrickYes  2668.449   5068.893   0.526  0.59957
## Nbhd3:BrickYes 11933.197   5341.027   2.234  0.02735 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9847 on 118 degrees of freedom
## Multiple R-squared:  0.8752, Adjusted R-squared:  0.8657
## F-statistic: 91.94 on 9 and 118 DF,  p-value: < 2.2e-16
```

4. For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

> Although there is no reduction in the standard error by making this change. The model is better. As can be seen in previous example, the Neighborhood 1 and 2 segregations is insignificant with a high p-value for nighborhood2 dummy variable. Thus, combining them makes all the neighborhood variables significant.

```
##
## Call:
## lm(formula = Price ~ Nb_new + Offers + SqFt + Brick + Bedrooms +
##     Bathrooms + Nb_new:Brick, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26710.2  -5797.0   -277.9   4337.6  26483.2
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         21117.266   9574.808   2.206  0.02932 *
## Nb_newold          -17709.779   2949.399  -6.005 2.10e-08 ***
## Offers              -8298.791    997.359  -8.321 1.60e-13 ***
## SqFt                   53.728      5.488   9.790  < 2e-16 ***
## BrickYes            24026.344   3349.112   7.174 6.53e-11 ***
## Bedrooms             4652.044   1554.250   2.993  0.00335 **
## Bathrooms            6407.659   2137.027   2.998  0.00330 **
## Nb_newold:BrickYes -10361.809   4100.564  -2.527  0.01281 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9780 on 120 degrees of freedom
## Multiple R-squared:  0.8748, Adjusted R-squared:  0.8675
## F-statistic: 119.8 on 7 and 120 DF,  p-value: < 2.2e-16
```

```
##                          2.5 %       97.5 %
## (Intercept)          2159.8124  40074.71994
## Nb_newold          -23549.3844 -11870.17306
## Offers             -10273.4916  -6324.08954
## SqFt                   42.8622     64.59387
## BrickYes            17395.3350  30657.35292
## Bedrooms             1574.7364   7729.35129
## Bathrooms            2176.4938  10638.82461
## Nb_newold:BrickYes -18480.6401  -2242.97815
```

Exam Question 3

1. Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city)

> Because presence of more police in a city can be due to several other reasons such as to prevent a terrorist attack or to carry out a political rally. In such cases, it will not be correct to directly

attribute the presence of more police to more 'crime'. Although, it has been observed that if a lot of police is present, the crime rate decreases, which represents the causal effect of more police on crime. Also, the data from different cities would vary a lot. Reasons for more police can vary from city to city. If there are more police in one city and crime has reduced, this does not establish that the cities with low police will have an increased crime. Some cities inherently carry a crime rate which is independent of the police force available.

Also, the presence of police and crime rates in different cities are independent of each other and combining the data from different cities to establish this relationship can be misleading.

2. How were the researchers from UPENN able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below.

Researchers checked the midday ridership on alert days vs normal days. They found out that the midday ridership hadn't changed even on the alert days. This negates the theory that "less number of crimes on alert days can be attributed to people staying indoors due to fear of terror". The researchers found that there was no change in midday ridership but the crime had actually reduced because of the presence of more police on terror alert days.

Initially, the researchers tried to regress the Daily total number of crimes in DC with High Alert. The negative coefficient of -7 says that on high alert days, there are seven less crimes per day. The second model considers the log of midday ridership along with the High alert categorical variable. The negative coefficient of -6 says that given the ridership does not change, the number of crimes go down by 6 on high alert days. Also, it says that on any given day (high alert or not is fixed), the crime goes up by 17 crimes if the log of ridership increases by one unit.

Since in the study by researchers, midday ridership was observed to be the same as no alert day, the crime should decrease by 6 units. Thus, there is a causal relationship between 'Police' and 'Crime'.

3. Why did they have to control for METRO ridership? What was that trying to capture?

The researchers proposed that the presence of increased police in DC led to a decrease in crimes. However, a possible counter argument was that people are staying indoors due to terror threat. Thus, Criminals are also staying indoors due to terror threat, and thus, lower crime rates. To isolate the effect of high alert on crimes, they controlled the METRO ridership. They found that the metro ridership was in fact the same and the reduction in crime rate was due to increased police. The researchers were trying to capture this isolated impact of high alert on crime rates through this experiment.

4. In the next page, I am showing you "Table 4" from the research paper. Just focuson the first column of the table. Can you describe the model being estimated here? What is the conclusion?

The table shows that the effect of High Alert can be segregated for different cities. For example, for district 1, a high alert would decrease the crimes by 2-3 crimes per day given the midday ridership doesn't change. The standard error of 0.044 shows that the interval of variation of coefficient is (-2.533, -2.709). Since zero does not lie in this range, we can be 95% confident about the decrease in crime of District 1 on high alert day.

However, for other districts, the decrease in crimes would be 0.5 on high alert days given midday ridership stays the same. The standard error is 0.455 that gives the confidence interval of (-1.471, 0.329). Since zero lies in the confidence interval. We cannot be sure whether there is a relationship between Other districts' crimes and high alert days. Thus, there is a negligible effect on crimes in other districts even on high alert days. One possible reason is that if the high alert is for district 1, it would not increase the police in other districts and the crime would continue in the normal way. Now looking at the midday ridership variable, given any fixed day (either high alert or not), it seems that an increase by one unit in log(midday ridership), it will lead to an increase in the crimes by 2 per day. Also the confidence interval does not contain zero. Thus, we can be 95% confident of the positive correlation between ridership and crimes.

It can be concluded that the impact of high alert is different for different districts.

Book Questions: Chapter 2

Question10

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

Reading the first 6 rows:

```
##      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?

```r
nrow(Boston)
```

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

How many columns?

```r
ncol(Boston)
```

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

What do the rows and columns represent?

> The 506 rows are the observations for housing values in Suburbs of Boston and the columns represent the features for them.
>
> 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.

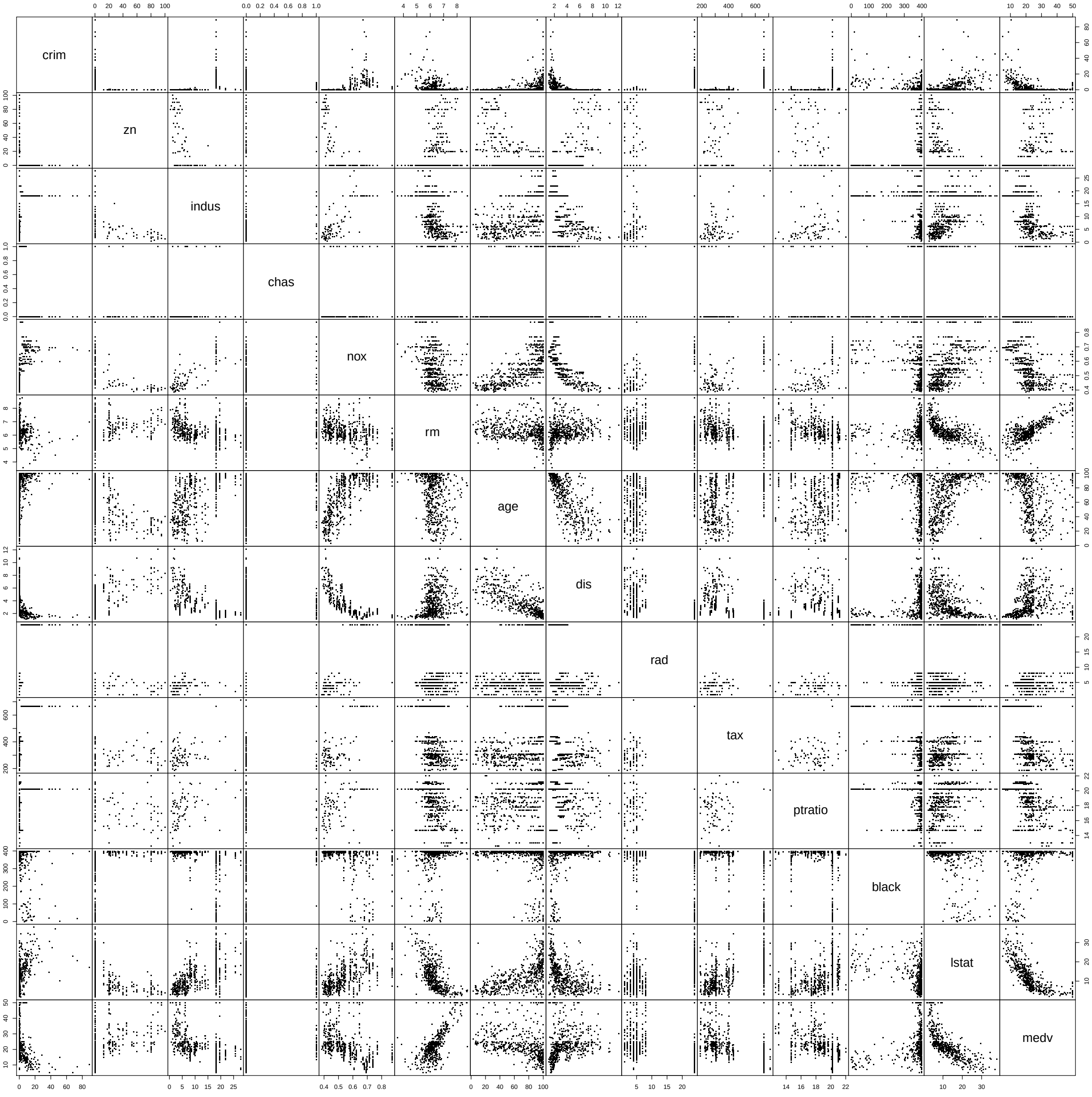
ptratio:pupil-teacher ratio by town.

black:1000(Bk - 0.63)ˆ2 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 findings.



Description:

- There Is a negative correlation between crime rate and median value of house. Most of the houses are in the zero crime rate region. Crime also correlates with age, dis, rad, tax and ptratio

- As the weighted mean of distances to five Boston employment centres increases, median value seems to increase. Possible explanation people want to buy houses not very close to work place/industrial areas. The employment centres might be very noisy

- There is a correlation between zn and nox. Other variables correlated to zn are age, lstat and nox.

- indus is correlated with age and dis
- nox is also correlated with age and dis. dis is additionally correlated with lstat and lstat is correlated with medv as well.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

By looking at the correlation chart above, the crime rate per capita is related to the following variables:

- Age: the crime rate is higher in regions where houses are old.
- Dis : As the distance from the employment companies increase, crime rate decreases
- Black: After a certain population for blacks, the crime rate seems to be increasing with increasing number of blacks. Its also possible that since all the blacks live together, any crime rate in any other proportion of blacks is not captured
- Lstat: As the %of lower status is low, the crime rates are not so high. They increase with the increase in % of people with lstat

(d) Do any of the suburbs of Boston appear to have particularly high crime rates?

Overall values of All variables:

Though overall summary, we can find the variables that vary from their mean/median values:

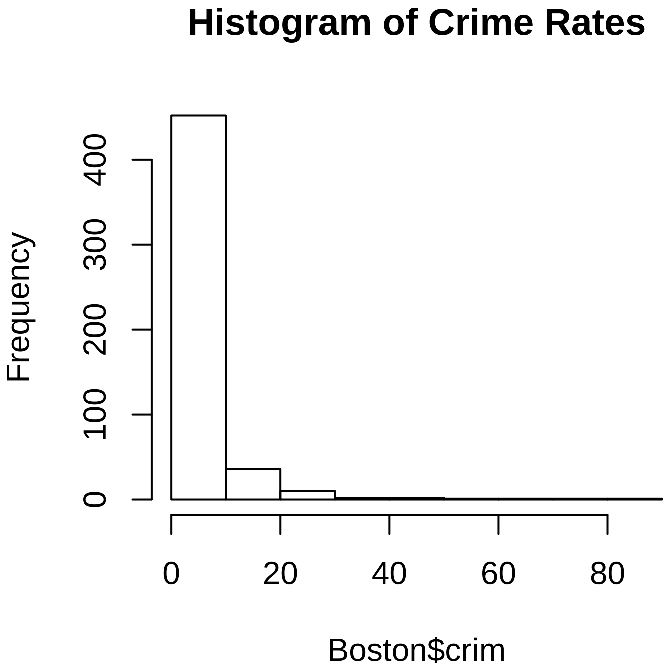
```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Checking crime rates

Yes. Around 2 percent suburbs have high crime rates. Crime rate for suburbs which are 99 perentile and above is 41%.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25650  3.61400  3.67700 88.98000
```

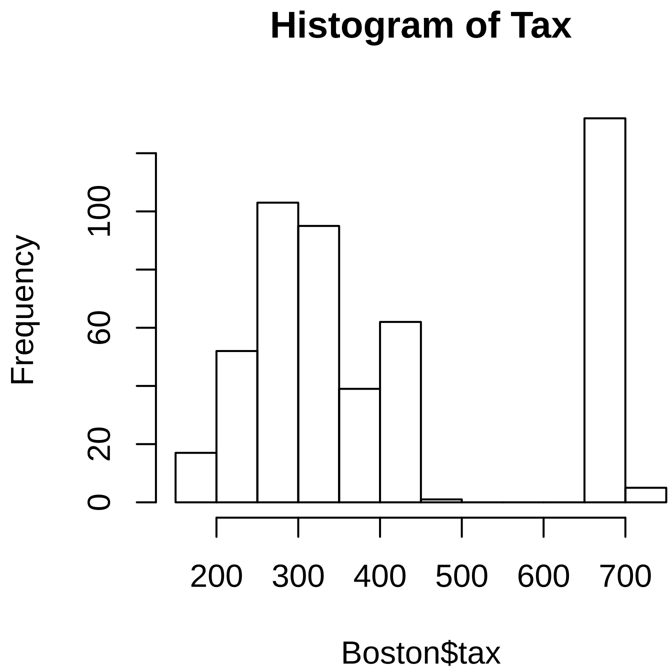
```
##      99%
## 41.37033
```



Checking Tax rates:

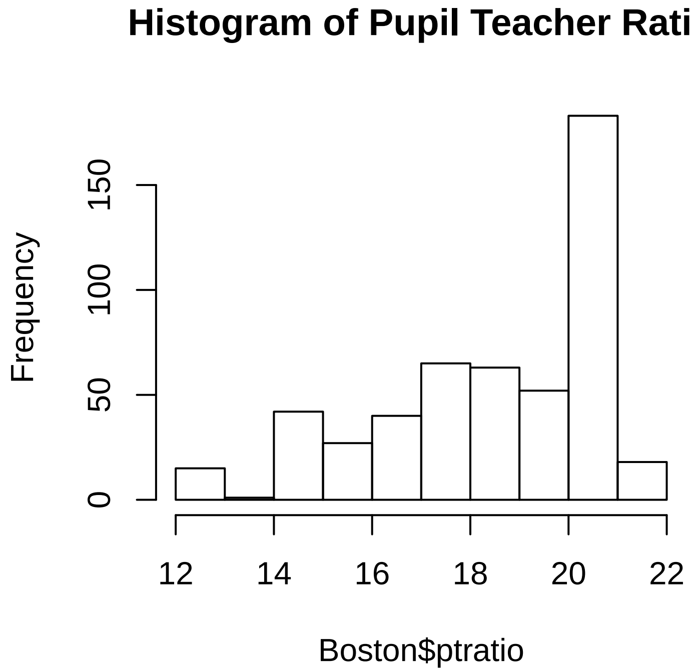
Yes. 20% of the suburbs have more than 666 as full-value property-tax rate per $10,000

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```



Checking Pupil-teacher ratios?

No suburbs have a particularly high Pupil teacher ratio as compared to the mean, median or 99th percentile of the data



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

Comment on the range of each predictor.

```
## [1] "dis"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.130   2.100   3.207   3.795   5.188  12.130
```

```
## [1] "crim"
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25650  3.61400  3.67700 88.98000
```

```
## [1] "lstat"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.73    6.95   11.36   12.65   16.96   37.97
```

- Crime Rate has a wide range from 0% to upto 89% with 2% of the suburbs having more than 40% crime rate

- Proportion for zone allotment is higher for the areas: low crime rate, less retail services, most of them do bound the river , moderate age, moderate other factors as well

- Only 25% top percentile of the suburbs have a displacement >6

- There are around 25% suburbs with percent of lower status of population >17%. Also, the maximum percentage is 38%

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

471 suburbs

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

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

```
## [1] 19.05
```

(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 findings.

```
## Data:
```

```
##        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
```

```
## Summary of suburbs with lowest median value of owner occupied homes:
```

```
##       crim             zn        indus           chas        nox       
##  Min.   :38.35   Min.   :0   Min.   :18.1   Min.   :0   Min.   :0.693  
##  1st Qu.:45.74   1st Qu.:0   1st Qu.:18.1   1st Qu.:0   1st Qu.:0.693  
##  Median :53.14   Median :0   Median :18.1   Median :0   Median :0.693  
##  Mean   :53.14   Mean   :0   Mean   :18.1   Mean   :0   Mean   :0.693  
##  3rd Qu.:60.53   3rd Qu.:0   3rd Qu.:18.1   3rd Qu.:0   3rd Qu.:0.693  
##  Max.   :67.92   Max.   :0   Max.   :18.1   Max.   :0   Max.   :0.693  
##        rm             age           dis              rad          tax     
##  Min.   :5.453   Min.   :100   Min.   :1.425   Min.   :24   Min.   :666  
##  1st Qu.:5.511   1st Qu.:100   1st Qu.:1.441   1st Qu.:24   1st Qu.:666  
##  Median :5.568   Median :100   Median :1.458   Median :24   Median :666  
##  Mean   :5.568   Mean   :100   Mean   :1.458   Mean   :24   Mean   :666  
##  3rd Qu.:5.625   3rd Qu.:100   3rd Qu.:1.474   3rd Qu.:24   3rd Qu.:666  
##  Max.   :5.683   Max.   :100   Max.   :1.490   Max.   :24   Max.   :666  
##     ptratio         black           lstat            medv  
##  Min.   :20.2   Min.   :385.0   Min.   :22.98   Min.   :5  
##  1st Qu.:20.2   1st Qu.:388.0   1st Qu.:24.88   1st Qu.:5  
##  Median :20.2   Median :390.9   Median :26.79   Median :5  
##  Mean   :20.2   Mean   :390.9   Mean   :26.79   Mean   :5  
##  3rd Qu.:20.2   3rd Qu.:393.9   3rd Qu.:28.69   3rd Qu.:5  
##  Max.   :20.2   Max.   :396.9   Max.   :30.59   Max.   :5  
```

Features of Suburbs with lowest median value of owner occupied homes:

- Crime rate lies between 75 percentile and maximum value i.e. High Crime Rate
- Zero land proportion of residential land zoned for lots over 25,000 sq.ft
- Proportion of non-retail business acres per town is relatively more: 75 percentile
- These do not bound river
- Nitrogen oxides are also on the higher sides
- Average number of rooms per dwelling <6 (within 25 percentile)
- These suburbs are very old
- weighted mean of distances to five Boston employment centres is within 1st quartile. i.e . close to employment centers
- high Rad: index of accessibility to radial highways

- High taxes: full-value property-tax rate per $10,000.

- High: pupil-teacher ratio by town (less teachers available)

- 1000(Bk - 0.63) ˆ2 where Bk is the proportion of blacks by town : on the higher end

- Lstat: lower status of population: on the higher end

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling?

> *64 suburbs*

```
##
## FALSE  TRUE
##   442    64
```

Check for more than eight rooms per dwelling?

> *13 suburbs*

```
##
## FALSE  TRUE
##   493    13
```

Comment on the suburbs that average more than eight rooms per dwelling.

```
summary(Boston[which(Boston$rm>8),])
```

```
##       crim               zn            indus             chas       
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm             age             dis       
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black      
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat           medv     
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7
##  Median :4.14   Median :48.3
##  Mean   :4.31   Mean   :44.2
##  3rd Qu.:5.12   3rd Qu.:50.0
##  Max.   :7.44   Max.   :50.0
```

- The crime rate is not so high
- Proportion of Area zoned for lots over 25,000 sq.ft is 13% on average
- proportion of non-retail business acres per town is not so high
- Most of the areas do not bound river
- Nitrogen oxide is on the lower side
- The suburbs are very old on average
- Distance from employment places is less
- Radial expressway is close except for a few exceptional cases
- low taxes (less than average)
- High teacher pupil ratio
- Good number of blacks live there
- Usually lower status of living

Book Questions: Chapter 3

Question15

```
##      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
```

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

i) crim ~ zn:

Significant relationship exists

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  4.45369376  0.4172178 10.674746 4.037668e-24
## zn          -0.07393498  0.0160946 -4.593776 5.506472e-06
```

ii)crim~indus

Significant relationship exists

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.0637426 0.66722830 -3.093008 2.091266e-03
## indus        0.5097763 0.05102433  9.990848 1.450349e-21
```

iii)crim~chas

Significant relationship does not exist

There is no relationship between chase and crim rates on the plot

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  3.744447  0.3961111  9.453021 1.239505e-19
## chas        -1.892777  1.5061155 -1.256727 2.094345e-01
```

iv) crim~nox

Significant relationship exists

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -13.71988   1.699479 -8.072992 5.076814e-15
## nox          31.24853   2.999190 10.418989 3.751739e-23
```

v)crim~rm

Significant relationship exists

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 20.481804  3.3644742  6.087669 2.272000e-09
## rm          -2.684051  0.5320411 -5.044819 6.346703e-07
```

vi)crim~age

Significant relationship exists

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -3.7779063 0.94398472 -4.002084 7.221718e-05
## age          0.1077862 0.01273644  8.462825 2.854869e-16
```

vii)crim~dis

Significant relationship exists

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.499262  0.7303972 13.005611 1.502748e-33
## dis         -1.550902  0.1683300 -9.213458 8.519949e-19
```

viii)crim~red

Significant relationship exists

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.2871594 0.44347583 -5.157349 3.605846e-07
## rad          0.6179109 0.03433182 17.998199 2.693844e-56
```

ix)crim~gtax

Significant relationship exists

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -8.52836909 0.815809392 -10.45387 2.773600e-23
## tax          0.02974225 0.001847415  16.09939 2.357127e-47
```

x)crim~ptratio

Significant relationship exists

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -17.646933  3.1472718 -5.607057 3.395255e-08
## ptratio       1.151983  0.1693736  6.801430 2.942922e-11
```

xi)crim~black

Significant relationship exists

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 16.55352922 1.425902755 11.609157 8.922239e-28
## black       -0.03627964 0.003873154 -9.366951 2.487274e-19
```

xii)crim~lstat

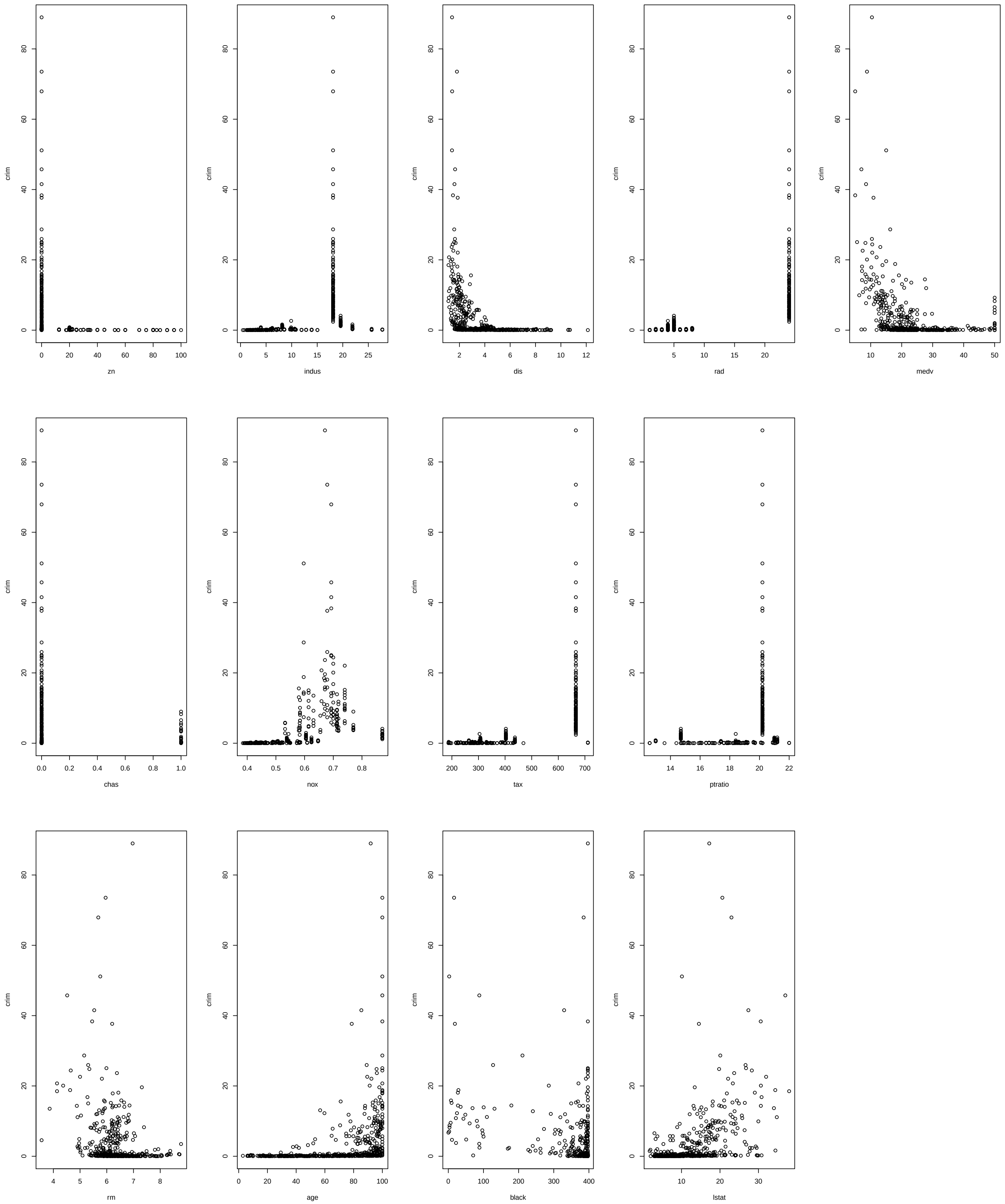
Significant relationship exists

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -3.3305381 0.69375829 -4.800718 2.087022e-06
## lstat        0.5488048 0.04776097 11.490654 2.654277e-27
```

xiii)crim~medv

Significant relationship exists

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 11.7965358 0.93418916 12.62757 5.934119e-32
## medv        -0.3631599 0.03839017 -9.45971 1.173987e-19
```


crim
0
20
40
60
80
zn
0 20 40 60 80 100
indus
0 5 10 15 20 25
dis
2 4 6 8 10 12
rad
5 10 15 20
medv
10 20 30 40 50
chas
0.0 0.2 0.4 0.6 0.8 1.0
nox
0.4 0.5 0.6 0.7 0.8
tax
200 300 400 500 600 700
ptratio
14 16 18 20 22
rm
4 5 6 7 8
age
0 20 40 60 80 100
black
0 100 200 300 400
lstat
10 20 30

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : Bj = 0?

Using all predictors, I built a multivariable linear regression model

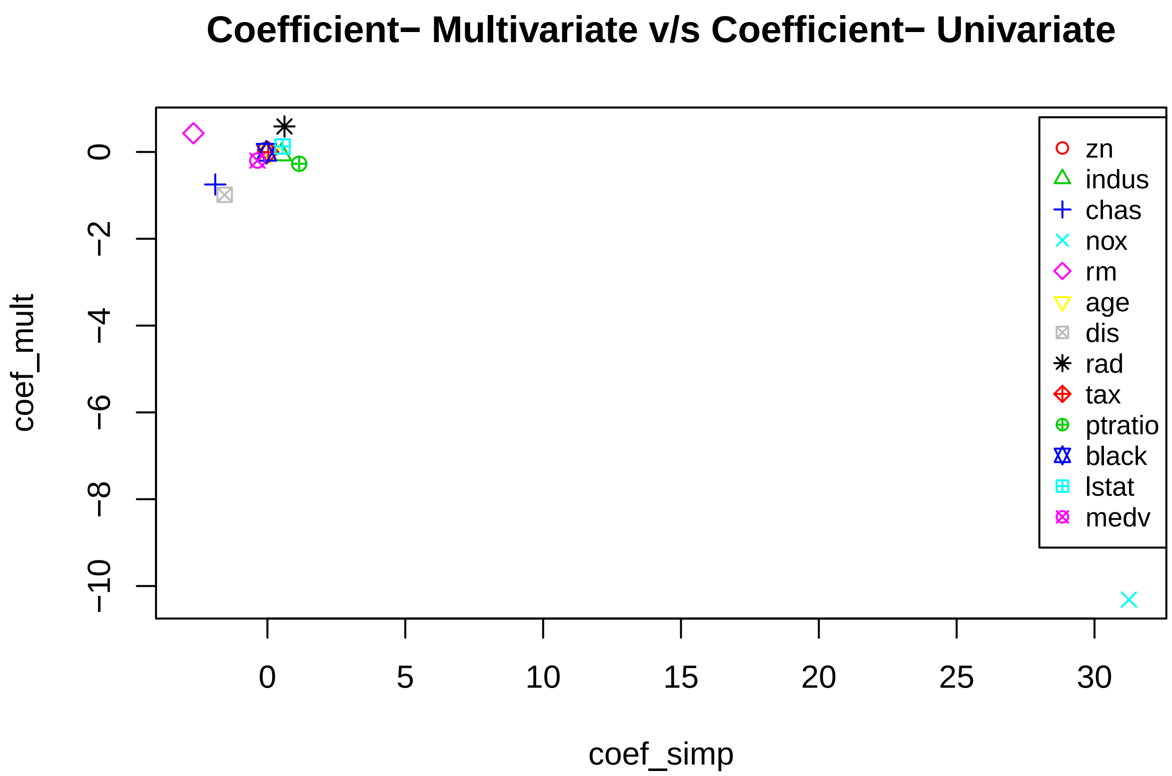
For zn,dis,rad,black and medv, Null Hypothesis H0 : Bj=0 will be rejected

```
##                   Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept)  17.033227523 7.234903031  2.35431317 1.894909e-02
## zn            0.044855215 0.018734071  2.39431224 1.702489e-02
## indus        -0.063854824 0.083407241 -0.76557890 4.442940e-01
## chas         -0.749133611 1.180146772 -0.63478004 5.258670e-01
## nox         -10.313534912 5.275536315 -1.95497373 5.115200e-02
## rm            0.430130506 0.612830309  0.70187538 4.830888e-01
## age           0.001451643 0.017925128  0.08098372 9.354878e-01
## dis          -0.987175726 0.281817266 -3.50289299 5.022039e-04
## rad           0.588208591 0.088049274  6.68044796 6.460451e-11
## tax          -0.003780016 0.005155587 -0.73318838 4.637927e-01
## ptratio      -0.271080558 0.186450494 -1.45390099 1.466113e-01
## black        -0.007537505 0.003673322 -2.05195893 4.070233e-02
## lstat         0.126211376 0.075724837  1.66671043 9.620842e-02
## medv         -0.198886821 0.060515990 -3.28651687 1.086810e-03
```

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

In Simple linear regression, all variables except 'chas' were significant. However, in Multiple regression, the significant variables are:

zn, dis, rad, black and medv only



```
##          coef_simp     coef_mult
## zn      -0.07393498   0.044855215
## indus    0.50977633  -0.063854824
## chas    -1.89277655  -0.749133611
## nox     31.24853120 -10.313534912
## rm      -2.68405122   0.430130506
## age      0.10778623   0.001451643
## dis     -1.55090168  -0.987175726
## rad      0.61791093   0.588208591
## tax      0.02974225  -0.003780016
## ptratio  1.15198279  -0.271080558
## black   -0.03627964  -0.007537505
## lstat    0.54880478   0.126211376
## medv    -0.36315992  -0.198886821
```

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form Y = B0 + B1X + B2X2 + B3X3 + e

i)crim∼ zn+ znˆ2 + znˆ3

There is Association of non linear parameters till zn squared

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      3.613524   0.372190  9.708814 1.547150e-20
## poly(zn, 3)1   -38.749835   8.372207 -4.628389 4.697806e-06
## poly(zn, 3)2  23.939832   8.372207  2.859441 4.420507e-03
## poly(zn, 3)3 -10.071868   8.372207 -1.203012 2.295386e-01
```

ii)crim~ indus+ indus^2 + indus^3

Yes, Non linear terms are predictive

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        3.613524   0.329998 10.950138 3.606468e-25
## poly(indus, 3)1  78.590819   7.423121 10.587301 8.854243e-24
## poly(indus, 3)2 -24.394796   7.423121 -3.286326 1.086057e-03
## poly(indus, 3)3 -54.129763   7.423121 -7.292049 1.196405e-12
```

iii)crim~nox+ nox^2+ nox^3

Yes, they are associated

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      3.613524   0.321573 11.237025 2.742908e-26
## poly(nox, 3)1  81.372015   7.233605 11.249165 2.457491e-26
## poly(nox, 3)2 -28.828594   7.233605 -3.985370 7.736755e-05
## poly(nox, 3)3 -60.361894   7.233605 -8.344649 6.961110e-16
```

iv)crim~rm+rm^2 + rm^3

yes there is an association with non linear terms upto rm squared

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     3.613524  0.3702993  9.7583873 1.026665e-20
## poly(rm, 3)1 -42.379442  8.3296758 -5.0877661 5.128048e-07
## poly(rm, 3)2  26.576770  8.3296758  3.1906128 1.508545e-03
## poly(rm, 3)3  -5.510342  8.3296758 -0.6615314 5.085751e-01
```

v)crim~age +age^2 + age^3

Yes, there is an association with the non linear terms

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     3.613524  0.3485173 10.368276 5.918933e-23
## poly(age, 3)1 68.182009  7.8397027  8.697015 4.878803e-17
## poly(age, 3)2 37.484470  7.8397027  4.781364 2.291156e-06
## poly(age, 3)3 21.353207  7.8397027  2.723727 6.679915e-03
```

vi)crim~ dis+ dis^2 + dis^3

Yes, there is an association with the non linear terms

```
##                  Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)      3.613524   0.325924  11.087013 1.060226e-25
## poly(dis, 3)1 -73.388590   7.331479 -10.010066 1.253249e-21
## poly(dis, 3)2  56.373036   7.331479   7.689176 7.869767e-14
## poly(dis, 3)3 -42.621877   7.331479  -5.813544 1.088832e-08
```

vii)crim~rad+rad^2 + rad^3

Yes there is an association with non linear terms upto rad squared

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     3.613524   0.297069 12.163920 5.149845e-30
## poly(rad, 3)1 120.907446   6.682402 18.093412 1.053211e-56
## poly(rad, 3)2  17.492299   6.682402  2.617666 9.120558e-03
## poly(rad, 3)3   4.698457   6.682402  0.703109 4.823138e-01
```

viii)crim~tax +tax^2 + tax^3

Yes there is an association with non linear terms upto tax squared

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     3.613524  0.3046845 11.859888 8.955923e-29
## poly(tax, 3)1 112.645827  6.8537074 16.435751 6.976314e-49
## poly(tax, 3)2  32.087251  6.8537074  4.681736 3.665348e-06
## poly(tax, 3)3  -7.996811  6.8537074 -1.166786 2.438507e-01
```

ix)crim~ptratio+ ptratio^2 + ptratio^3

Yes, there is an association with the non linear terms

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       3.613524   0.329998 10.950138 3.606468e-25
## poly(indus, 3)1  78.590819   7.423121 10.587301 8.854243e-24
## poly(indus, 3)2 -24.394796   7.423121 -3.286326 1.086057e-03
## poly(indus, 3)3 -54.129763   7.423121 -7.292049 1.196405e-12
```

x)crim~ black+ black^2 + black^3

No association with non linear terms

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       3.613524   0.353627 10.2184605 2.139710e-22
## poly(black, 3)1 -74.431199   7.954643 -9.3569505 2.730082e-19
## poly(black, 3)2   5.926419   7.954643  0.7450264 4.566044e-01
## poly(black, 3)3  -4.834565   7.954643 -0.6077665 5.436172e-01
```

xi)crim~lstat+ lstat^2 + lstat^3

Yes there is an association with non linear term till lstat squared

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       3.613524  0.3391698 10.654025 4.939398e-24
## poly(lstat, 3)1  88.069666  7.6294361 11.543404 1.678072e-27
## poly(lstat, 3)2  15.888164  7.6294361  2.082482 3.780418e-02
## poly(lstat, 3)3 -11.574022  7.6294361 -1.517022 1.298906e-01
```

xii)crim~medv+ medv^2 + medv^3

Yes, the non linear term is predictive

```
##                  Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)      3.613524  0.2920344  12.373622 7.024110e-31
## poly(medv, 3)1 -75.057605  6.5691520 -11.425768 4.930818e-27
## poly(medv, 3)2  88.086211  6.5691520  13.409069 2.928577e-35
## poly(medv, 3)3 -48.033435  6.5691520  -7.311969 1.046510e-12
```

Book Questions: Chapter 4

Question10

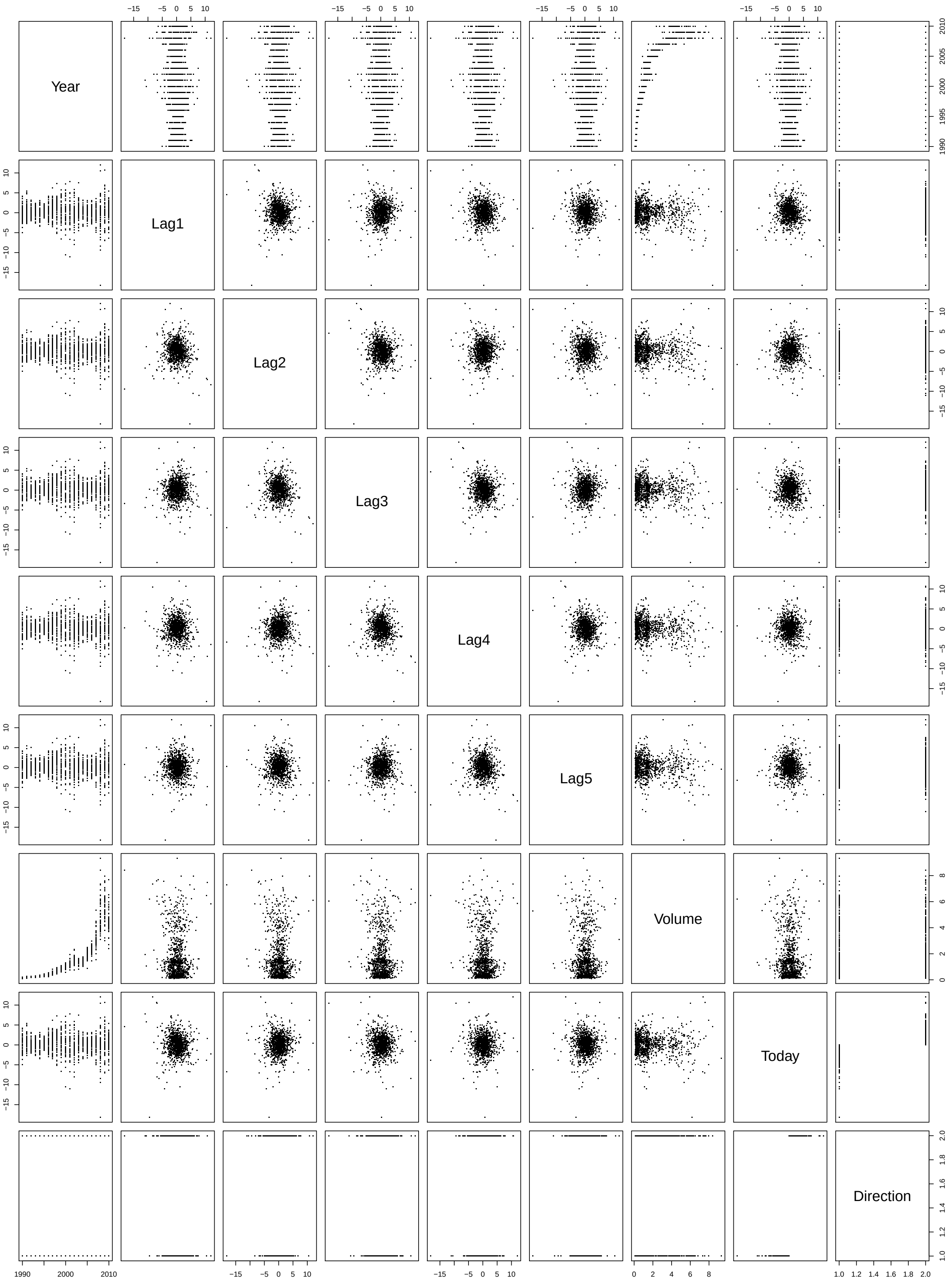
(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

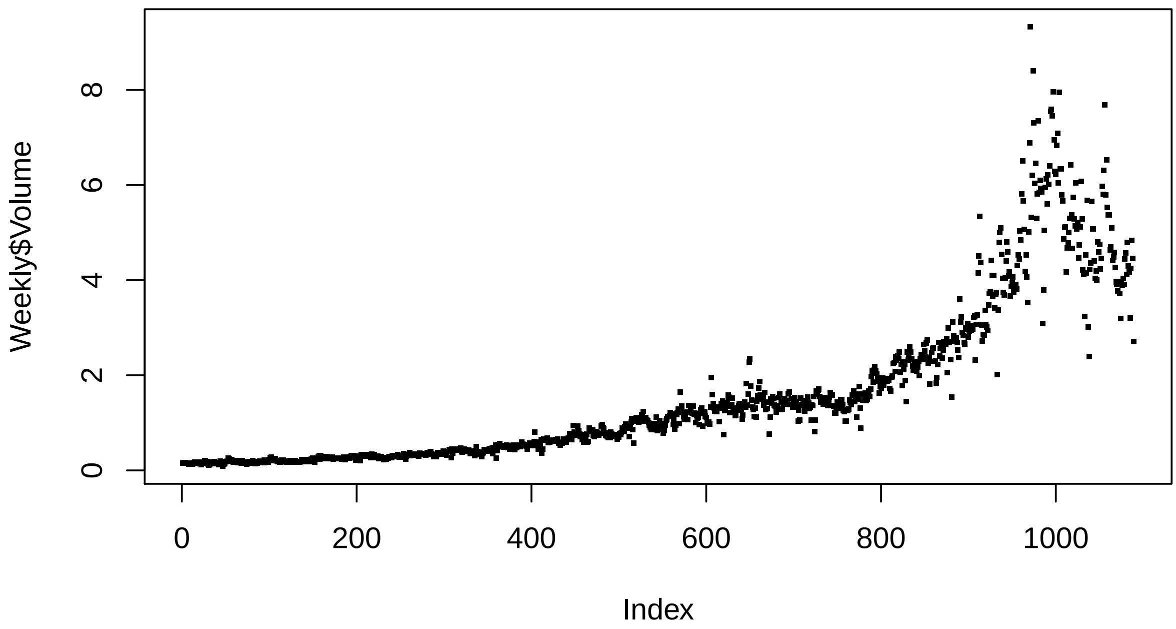
- Only Volume variable has a visible pattern over time.
- The same can be seen in correlation matrix as well.
- Lag variables do not have any identifiable pattern.

```r
#Numerical Summary
summary(Weekly)
```

```
##       Year           Lag1               Lag2               Lag3
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
##       Lag4               Lag5              Volume
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202
##  Median :  0.2380   Median :  0.2340   Median :1.00268
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821
##      Today          Direction
##  Min.   :-18.1950   Down:484
##  1st Qu.: -1.1540   Up  :605
##  Median :  0.2410
##  Mean   :  0.1499
##  3rd Qu.:  1.4050
##  Max.   : 12.0260
```

```r
#Graphical Summary
pairs(Weekly,pch='.',cex=2.5)
```


Year
Lag1
Lag2
Lag3
Lag4
Lag5
Volume
Today
Direction
−15
−5
0
5
10
1990
1995
2000
2005
2010
0
2
4
6
8
1.0
1.2
1.4
1.6
1.8
2.0



```
## Correlation check:

##                Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```

(b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

Only Lag2 came out to be statistically significant

```r
#model fitting
glm.fit<- glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,data= Weekly,family='binomial')
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = "binomial", data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

> According to confusion matrix, we are correctly predicting 557 times when the market goes ‘UP’ and 54 times when the market goes ‘DOWN’. However, we are mispredicting 430 ‘DOWN’s as ’UP’s and 48 ’UPs’ as ’DOWN’s.
>
> - *Overall correct guess = 56%*
> - *False Down Rate=47%*
> - *False Up Rate=43%*

```r
glm.pred<- predict(glm.fit, type='response')
glm.prob=rep('Down',nrow(Weekly))
glm.prob[glm.pred>0.5]<-'Up'#56% correct : false 'UP's
cat("Confusion matrix: ")
```

```
## Confusion matrix:
```

```r
table(glm.prob , Weekly$Direction)
```

```
##
## glm.prob Down  Up
##     Down   54  48
##     Up    430 557
```

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

> Overall test correct rate : 62.5%

```r
train = Weekly[which(Weekly$Year<=2008),]
test = Weekly[which(Weekly$Year>2008),]

glm.pred.new<- glm(Direction~Lag2, data = train, family='binomial')
glm.test.pred<- predict(glm.pred.new, newdata=test, type='response')
glm.test.prob=rep('Down',104)
glm.test.prob[glm.test.pred>0.5]<- 'Up'

cat("Logistic Regression Confusion Matrix: ")
```

```
## Logistic Regression Confusion Matrix:
```

```r
table(test$Direction, glm.test.prob)
```

```
##       glm.test.prob
##        Down Up
##   Down    9 34
##   Up      5 56
```

(g) Repeat (d) using KNN with K = 1. >Overall test correct rate :

- *50% with k =1*
- *55% with k = 3*
- *53% with k=5*

```
## Confusion Matrix using KNN with K=1:
```

```
##       knn.pred
##        Down Up
##   Down   21 22
##   Up     30 31
```

(h) Which of these methods appears to provide the best results on this data?

> Logistic Regression is performing the best amongst all the above methods with an overall correct rate of 62.5%.

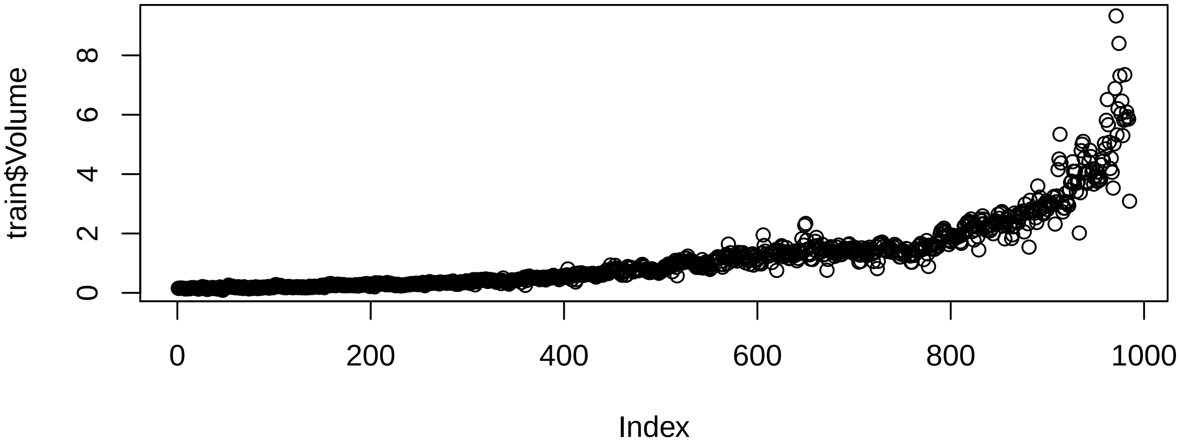
(i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

> For logistic regression, I checked the interaction of Lag2 with Lag4. This gave a correct rate of 55.7%. ALso, I checked the pattern between Volume and Lag2 variables. Upon checking the trend between Lag2 and log(Volume) variables, the relation between the two seemes significant. However, upon using log(Volume) as one of the predictors, both Lag2 and Volume became insignificant. The best results are being produced by logistic regression with only Lag2 as the predictor. The overall correct rate for logistic regression is 62.5%. I also tried the classification using KNN which gives a correctness rate of 50.9%.

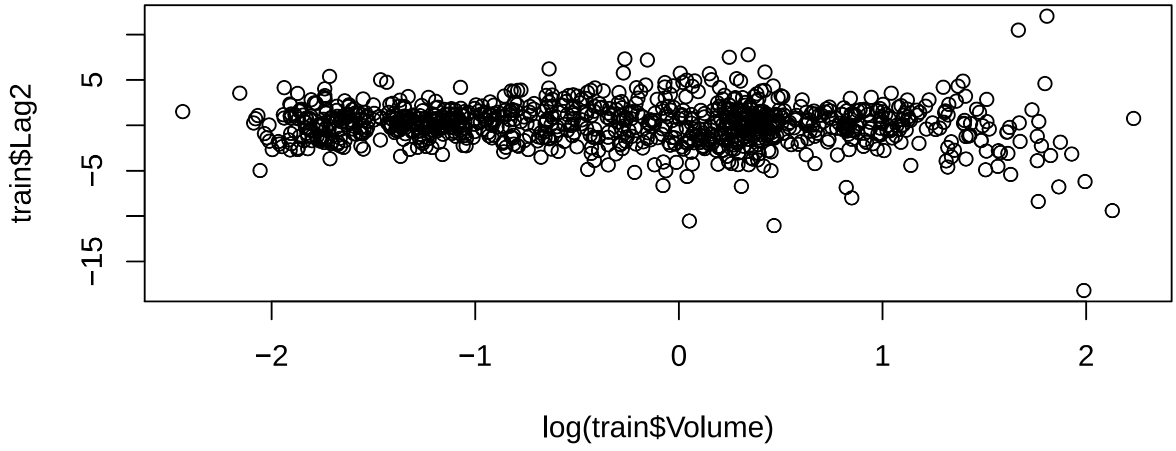
```
## Confusion Matrix(Logistic with Lag2:Lag4):
```

```
##       glm.test.prob2
##        Down Up
##   Down    1 42
##   Up      4 57
```

```
## Plot between Lag2 and Volume:
```



```
## Plot between Lag2 and log(Volume):
```



```
## Confusion Matrix(Logistic):

##       glm.test.prob2
##         Down Up
##   Down     9 34
##   Up       5 56

## Confusion Matrix(KNN):

##       knn.pred
##         Down Up
##   Down    11 32
##   Up       8 53
```

Book Questions:Chapter 6

Question 9

In this exercise, we will predict the number of applications received using the other variables in the College data set.

(a) Split the data set into a training set and a test set.

```
## dim(train) 427 18
```

```
## dim(test) 350 18
```

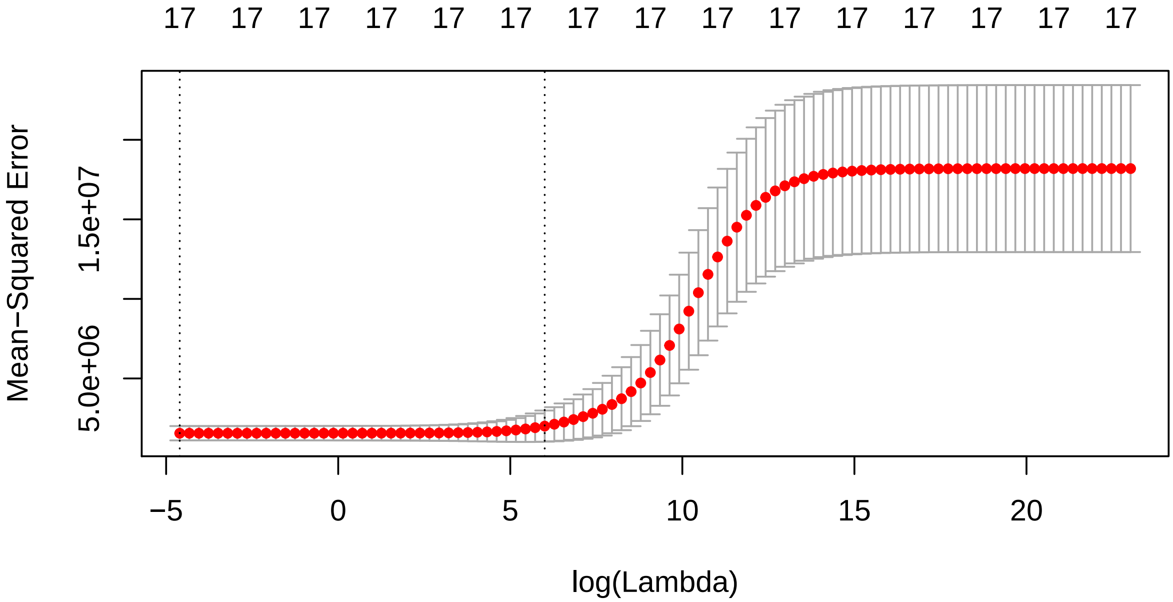
(b) Fit a linear model using least squares on the training set, and report the test error obtained.

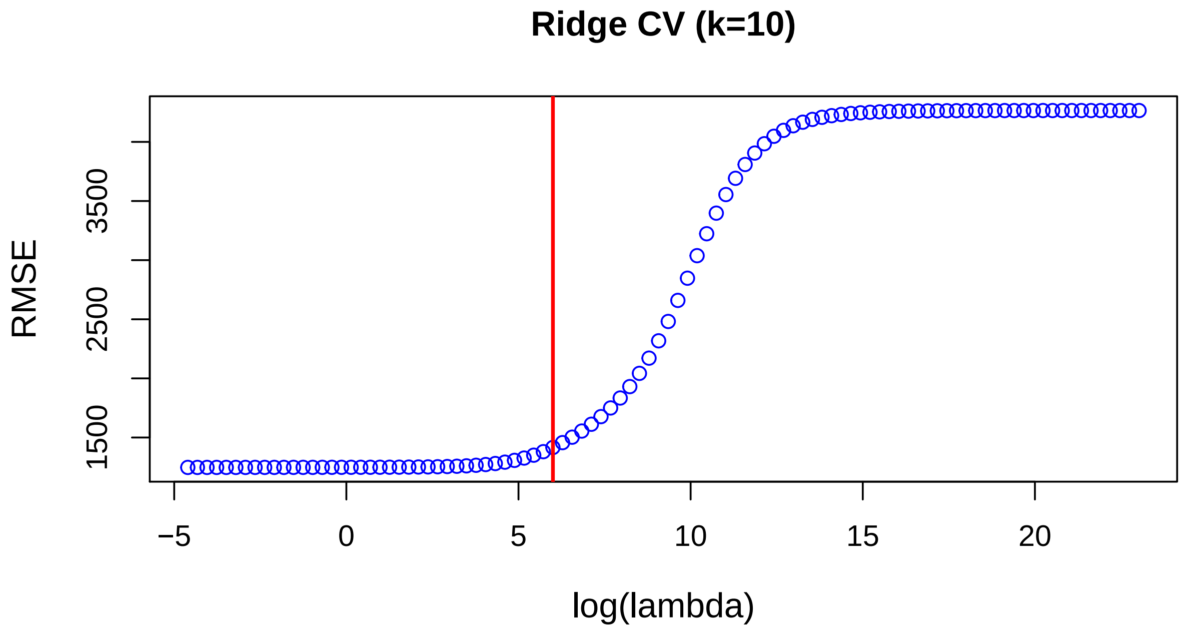
Fit an MLR(Multivariate Linear Model) where I chose the variables manually in the forward selection modeling technique. The test error obtained using this method is 1211 Applications with variables Accept, Top10perc,Expend and F.Undergrad as significant variables.

```
## 
## Call:
## lm(formula = college1.Apps ~ ., data = train)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3307.0  -367.9    14.0   286.1  7344.2 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.102e+03  5.096e+02  -2.163 0.031129 *  
## PrivateYes  -4.059e+02  1.674e+02  -2.425 0.015731 *  
## Accept       1.406e+00  6.442e-02  21.832  < 2e-16 ***
## Enroll      -8.435e-01  2.466e-01  -3.420 0.000689 ***
## Top10perc    3.070e+01  6.894e+00   4.454 1.09e-05 ***
## Top25perc   -6.800e+00  5.528e+00  -1.230 0.219343    
## F.Undergrad  1.312e-01  3.859e-02   3.399 0.000744 ***
## P.Undergrad  4.316e-02  3.365e-02   1.283 0.200330    
## Outstate    -6.038e-02  2.356e-02  -2.563 0.010724 *  
## Room.Board   8.347e-02  5.791e-02   1.441 0.150228    
## Books        3.624e-02  2.898e-01   0.125 0.900544    
## Personal     1.537e-02  7.870e-02   0.195 0.845301    
## PhD         -8.733e+00  5.527e+00  -1.580 0.114839    
## Terminal    -2.434e+00  5.894e+00  -0.413 0.679855    
## S.F.Ratio    3.289e+01  1.516e+01   2.170 0.030592 *  
## perc.alumni -8.873e-01  5.000e+00  -0.177 0.859238    
## Expend       1.123e-01  1.483e-02   7.570 2.51e-13 ***
## Grad.Rate    1.075e+01  3.542e+00   3.036 0.002552 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 937.3 on 409 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9174 
## F-statistic: 279.3 on 17 and 409 DF,  p-value: < 2.2e-16
## Test Error(RMSE) using Linear regression is : 1211.41
```

(c) Fit a ridge regression model on the training set, with ?? chosen by cross-validation. Report the test error obtained.

On fitting the Ridge regression and using cross-validation, the lambda with the lowest error is 0.01 which has a test error of 1062 Applications. However, if for the sake of simplicity of the model, we choose lambda.1se(largest value of lambda with error within 1 standard error of minimum error)i.e. 403 instead of lambda.min(lambda with minimum cross validation error), we get a test error of 1058 Applications.

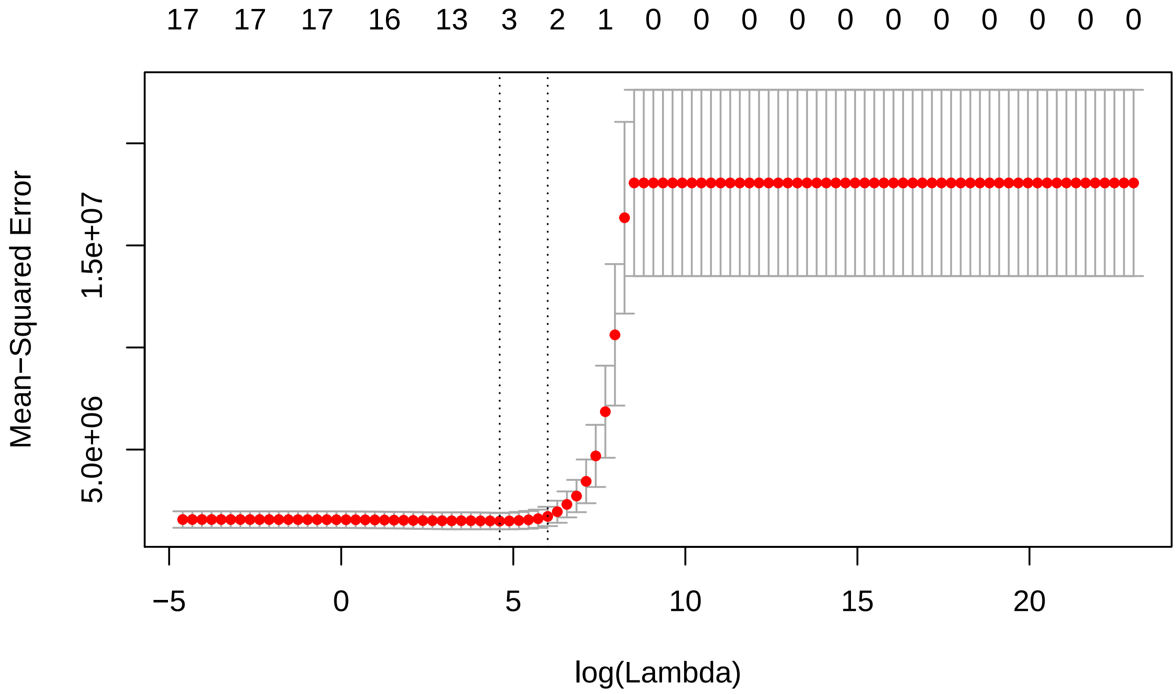


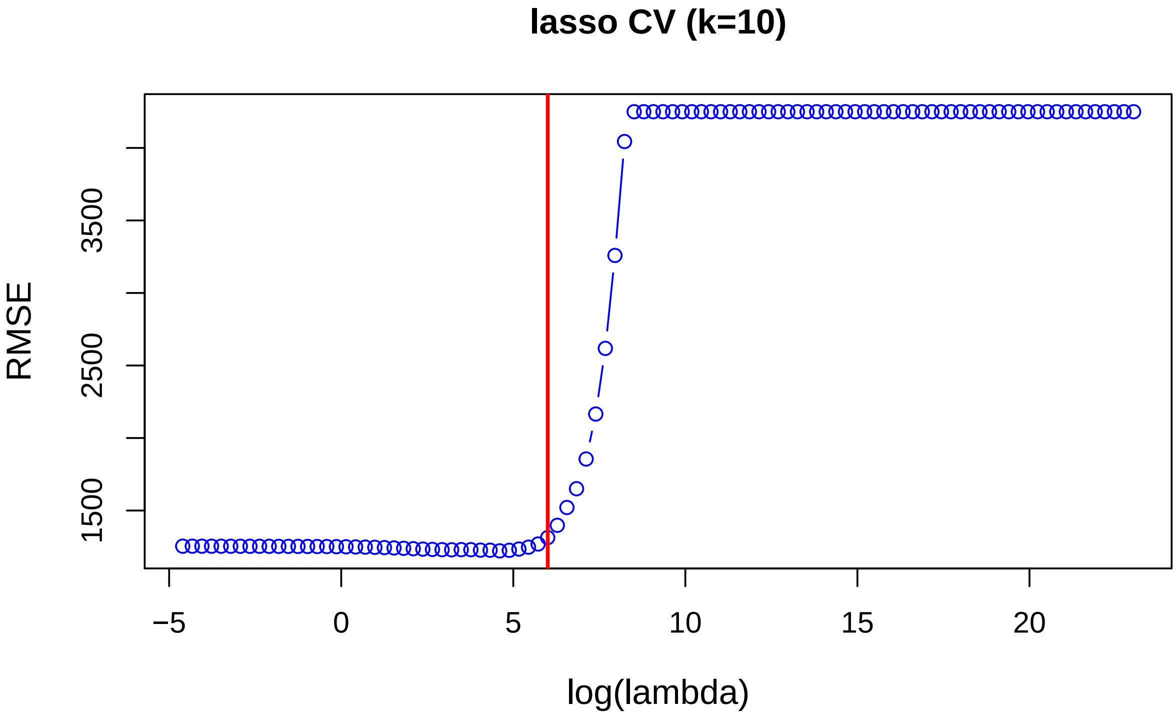


```
## Test Error(RMSE) from Ridge Regression is:  1058.549
```

(d) Fit a lasso model on the training set, with ?? chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

On fitting the Lasso regression and using cross-validation, the lambda with the lowest error is 100 which has a test error of 1109 Applications. However, if for the sake of simplicity of the model, we choose lambda.1se(largest value of lambda with error within 1 standard error of minimum error)i.e. 403 instead of lambda.min(lambda with minimum cross validation error), we get an error of 1155 applications.



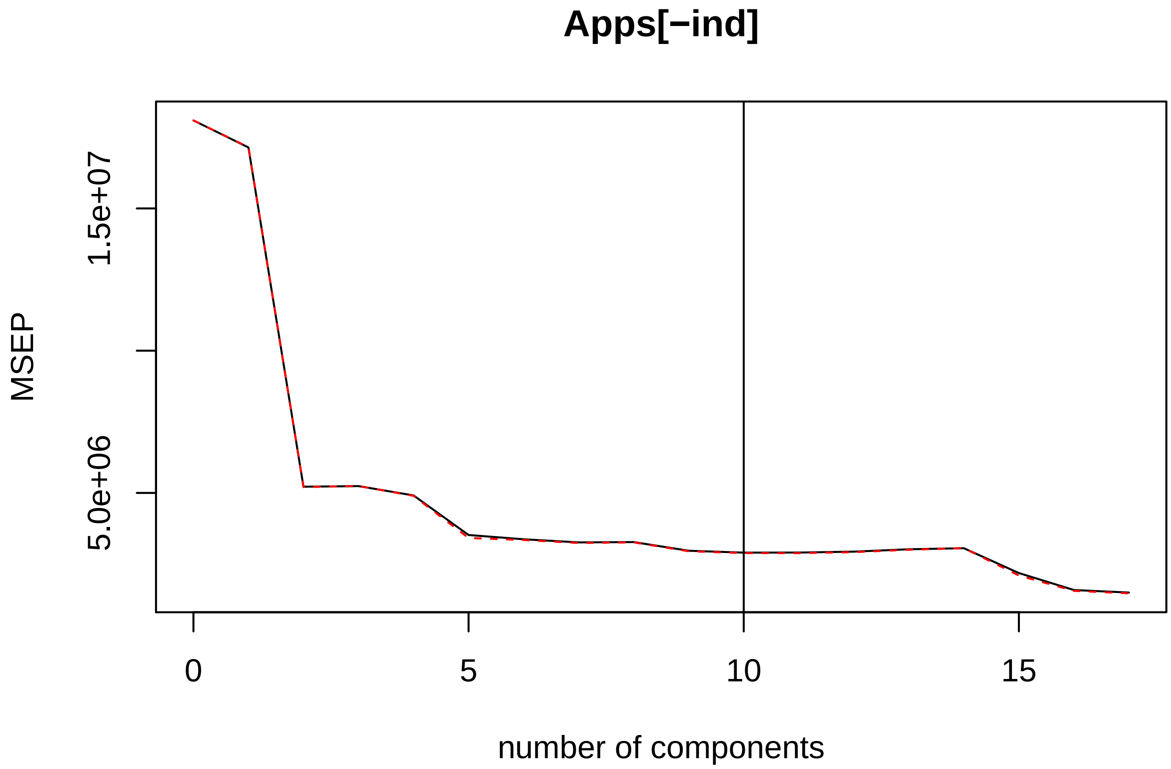


```
## Test Error(RMSE) from Lasso regression is: 1155.69

## Total non-zero coefficients are: 3
```

(e) Fit a PCR model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selected by cross-validation.

Fitting a model using cross validation on PCR gives an error of 1280 Applications with 10 Principal components. Error obtained using just two PCs is 1843. Since, the number of variables p<<n (number of observations), it uses all the variables and produces for lower error.

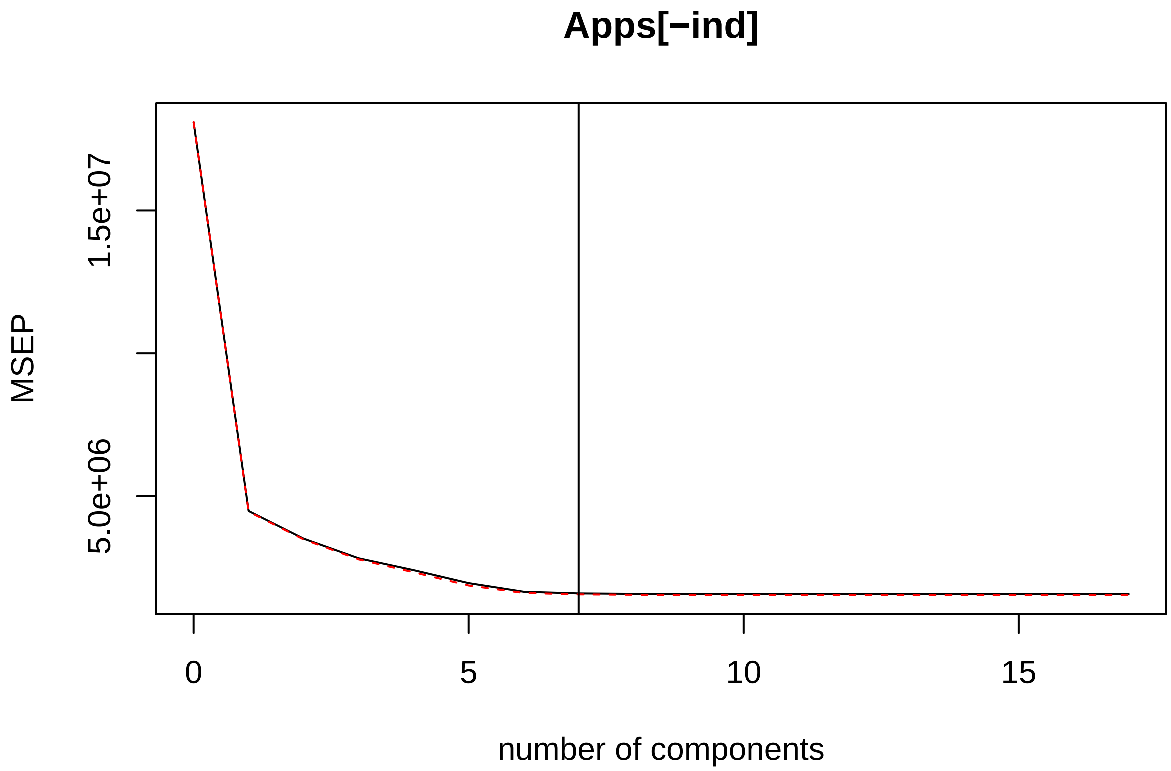
```
##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##     loadings
```



```
## Test Error(RMSE):  1280.811

## M selected by cross validation: 10
```

(f) Fit a PLS model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selected by cross-validation.

I fitted a PLS model by using 7 principal components and a test error of 1080 Applications



```
## Data:    X dimension: 427 17
##  Y dimension: 427 1
## Fit method: kernelpls
## Number of components considered: 17
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            4254     2117     1875     1681     1553     1398     1286
## adjCV         4254     2112     1867     1670     1529     1369     1270
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1264     1258     1256      1258      1258      1258      1256
## adjCV     1250     1245     1243      1245      1245      1245      1242
##        14 comps  15 comps  16 comps  17 comps
## CV         1255      1255      1255      1255
## adjCV      1241      1242      1242      1242
##
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X             26.38    43.81    62.60    65.06    67.89     72.4    76.18
## Apps[-ind]    77.07    84.08    87.87    91.79    93.56     93.9    93.96
##             8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X             80.84    83.17     85.66     89.47     91.37     92.53
## Apps[-ind]    93.99    94.04     94.07     94.07     94.08     94.09
##             14 comps  15 comps  16 comps  17 comps
## X              94.09     96.79     98.45    100.00
## Apps[-ind]     94.09     94.09     94.09     94.09
```

```
## Test Error(RMSE):  1080.331
```

```
## M selected by cross validation: 7
```

(g) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

From the above analysis, following are the test errors obtained from different approaches:

- linear:1211
- ridge: 1058
- lasso: 1155
- PLS:1080
- PCR:1280

On an average, we can predict the number of Applications within +/- of 2000 aplications(+/- two sigma) with a 95% confidence interval. Also, the error from linear regression is the maximum at +/- 2400 applications with 95% confidence interval. For Ridge, PCR and PLS, the error is almost similar with slight deviations at ~1000 applications. Since these four techniques improve upon the Linear Regression model, no one technique is overpowering any other while improving the model. They all produce more or less the same test error.

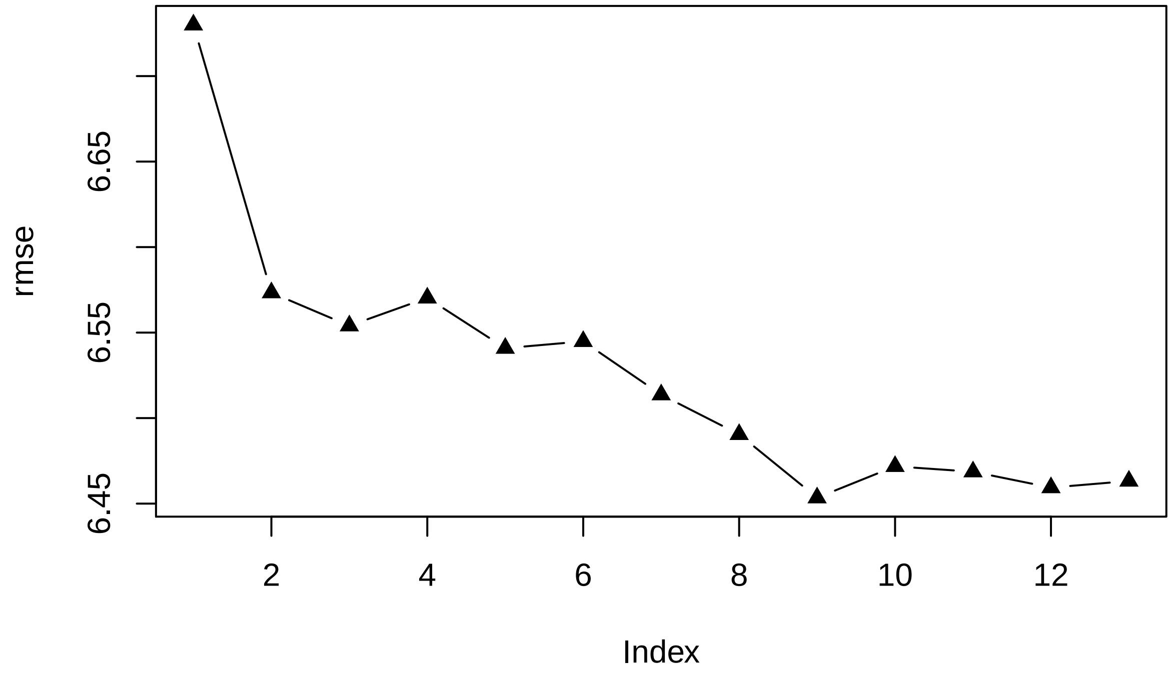
Question 11

We will now try to predict per capita crime rate in the Boston data set.

(a) Try out some of the regression methods explored in this chapter, such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

Subset Selection

I used the regsubsets function to perform the subset selection along with 10 fold cross validation. I checked the graph for optimum value of predictors for minimum error.



```
which.min(rmse)
```

```
## [1] 9
```

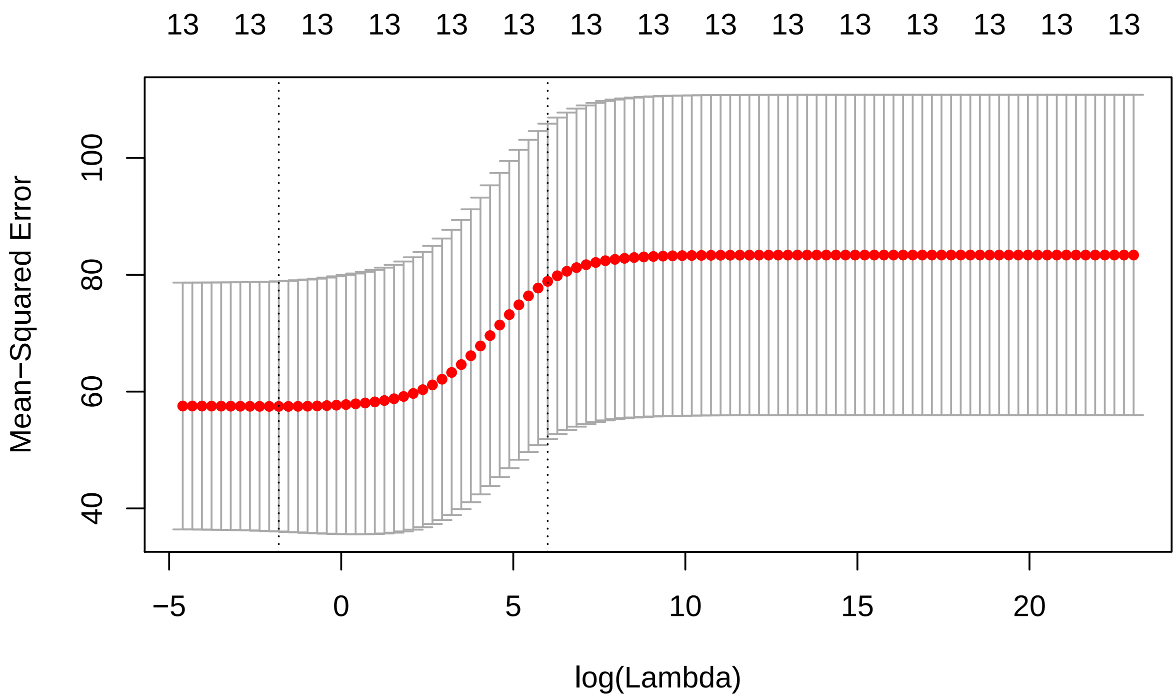
```
rmse[which.min(rmse)]
```

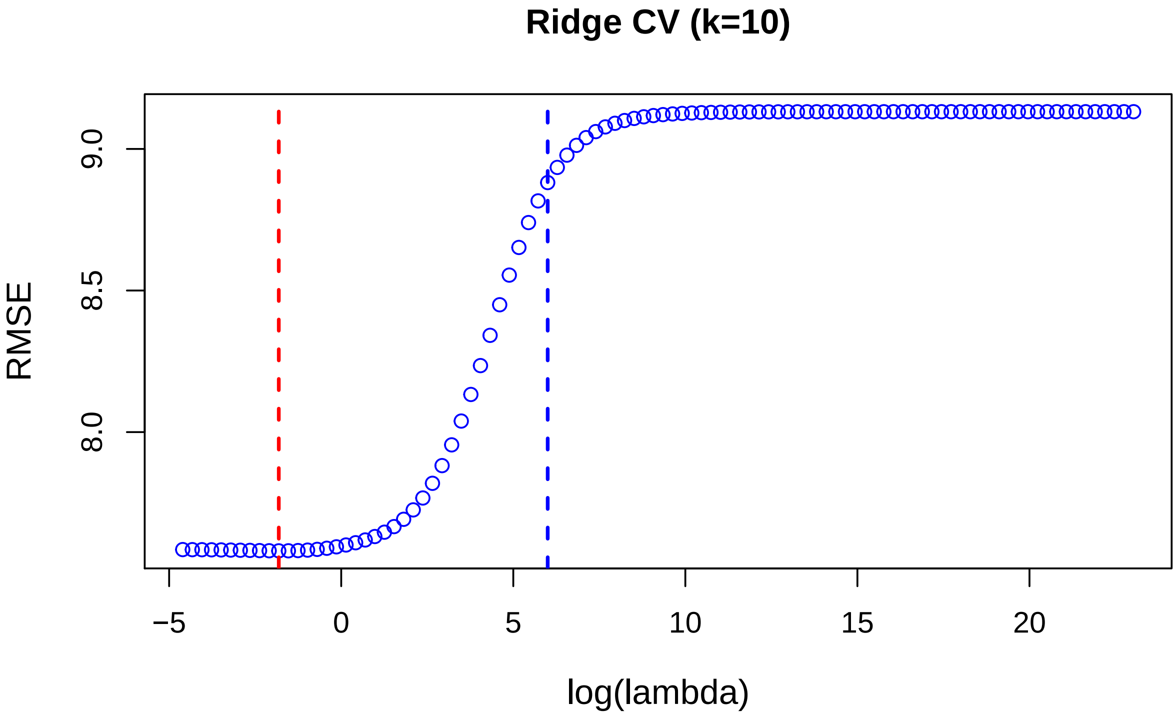
```
## [1] 6.45343
```

Ridge and Lasso

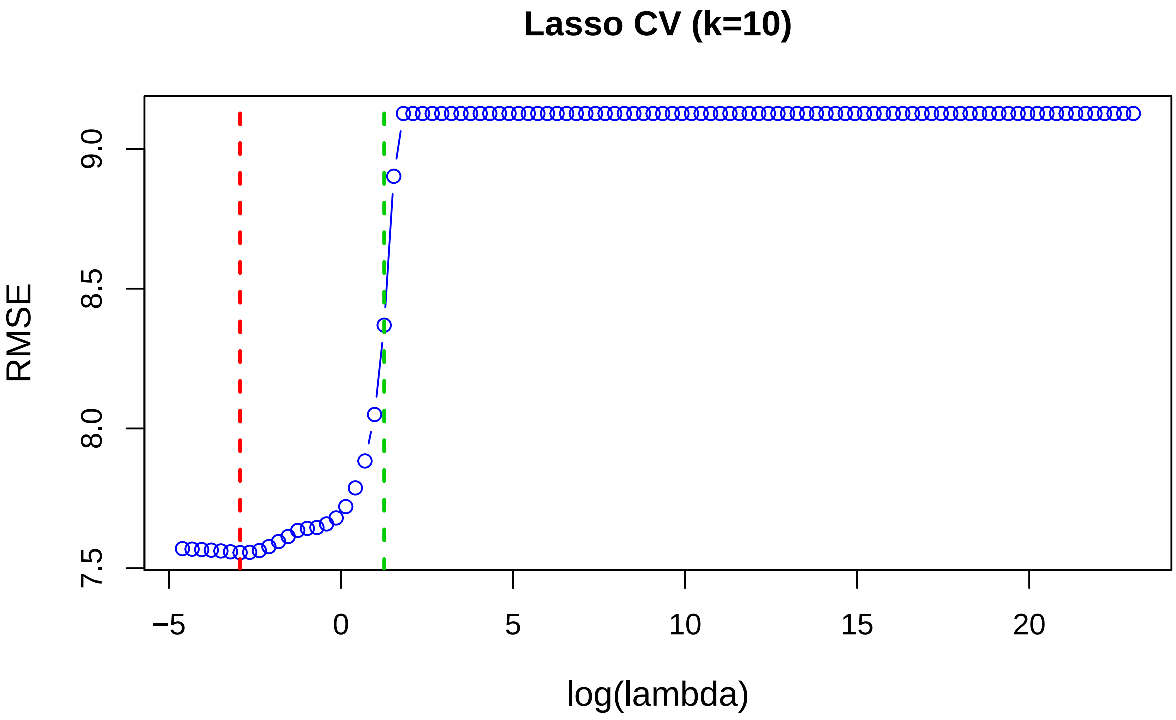
For Ridge and Lasso fit of the model, I have split the data in Test and train. Then, I fit the Ridge regression on train using cv.glmnet function from (glmnet) library. A set of lambdas and the corresponding errors can be obtained from this. Then, I selected lambda for which the error is the minimum. I also checked lambda.1se but the error was less for Lambda.min.

Although, for Lasso regression, I chose lambda.1se(highest lambda whose error is within 1 standard error of the minimum error). I selected lambda.1se because with only 1 variable, I am getting an error of 6.60.





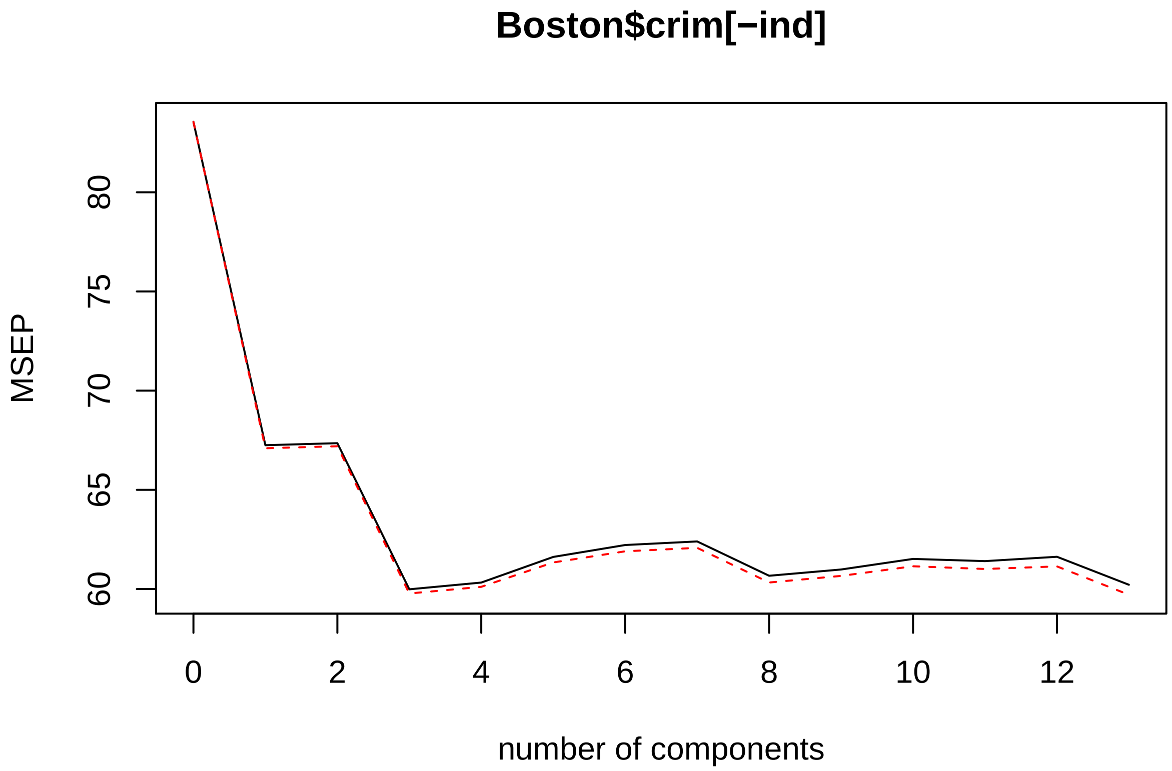
```
## Test Error from Ridge:  5.020304
```



```
## Test Error from Lasso(minimum error lambda):  4.963033

## Test Error from Lasso(one SE lambda):  6.603986

## Number of non zero coefficients(1se lambda):  1

## Number of non zero coefficients(min lambda):  12
```

PCR

I fit the model using PCR and checked the summary to identify the principal component for lowest error. Using that Principal component, I made predictions on test data and calculated corresponding RMSE. For PCR, Principal components upto 3 are able to reduce error to 5.19.



```
## Data:    X dimension: 306 13
##  Y dimension: 306 1
## Fit method: svdpc
## Number of components considered: 13
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           9.141    8.201    8.207    7.745    7.767    7.850    7.888
## adjCV        9.141    8.191    8.197    7.731    7.753    7.832    7.868
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       7.899    7.789    7.809     7.843     7.836     7.850     7.760
## adjCV    7.879    7.767    7.789     7.820     7.811     7.819     7.727
##
## TRAINING: % variance explained
##                    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X                    47.38    60.73    70.25    76.87    82.74    87.59
## Boston$crim[-ind]    22.44    22.48    31.50    31.69    31.69    31.79
##                    7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
## X                     90.9    93.40    95.36     97.13     98.49     99.49
## Boston$crim[-ind]     31.9    33.84    33.94     34.02     34.33     36.10
##                    13 comps
## X                    100.00
## Boston$crim[-ind]     38.14

## Test RMSE from PCR =  5.198544
## M selected by cross validation: 3
```

PLS

I fit the model using PLS and checked the summary to identify the principal component for lowest error. Using that Principal component, I made predictions on test data and calculated corresponding RMSE. For PLS, Principal components upto 8 are able to reduce error to 5.13. Note that in case of PLS, less principal components are required since the PCs capture the dependence of response variables on independent variables while computing principal components. thus, the variance in Repsonse variable is captured in less number of principal components.

```
## Data:    X dimension: 306 13
##  Y dimension: 306 1
## Fit method: kernelpls
## Number of components considered: 13
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           9.141    7.927    7.622    7.659    7.631    7.568    7.568
## adjCV        9.141    7.922    7.613    7.634    7.611    7.548    7.546
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       7.559    7.548    7.551     7.552     7.553     7.553     7.553
## adjCV    7.538    7.528    7.531     7.532     7.533     7.533     7.533
##
## TRAINING: % variance explained
##                    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X                    46.74    56.67    60.14    70.89    76.00    80.05
## Boston$crim[-ind]    26.27    33.74    36.48    37.01    37.64    37.98
##                    7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
## X                    83.92    86.16    88.96     91.94     96.46     98.18
## Boston$crim[-ind]    38.06    38.13    38.14     38.14     38.14     38.14
##                    13 comps
## X                    100.00
## Boston$crim[-ind]     38.14

## Test Error from PLS=  5.130156

## M selected by cross validation: 8
```

(b) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross validation, or some other reasonable alternative, as opposed to using training error.

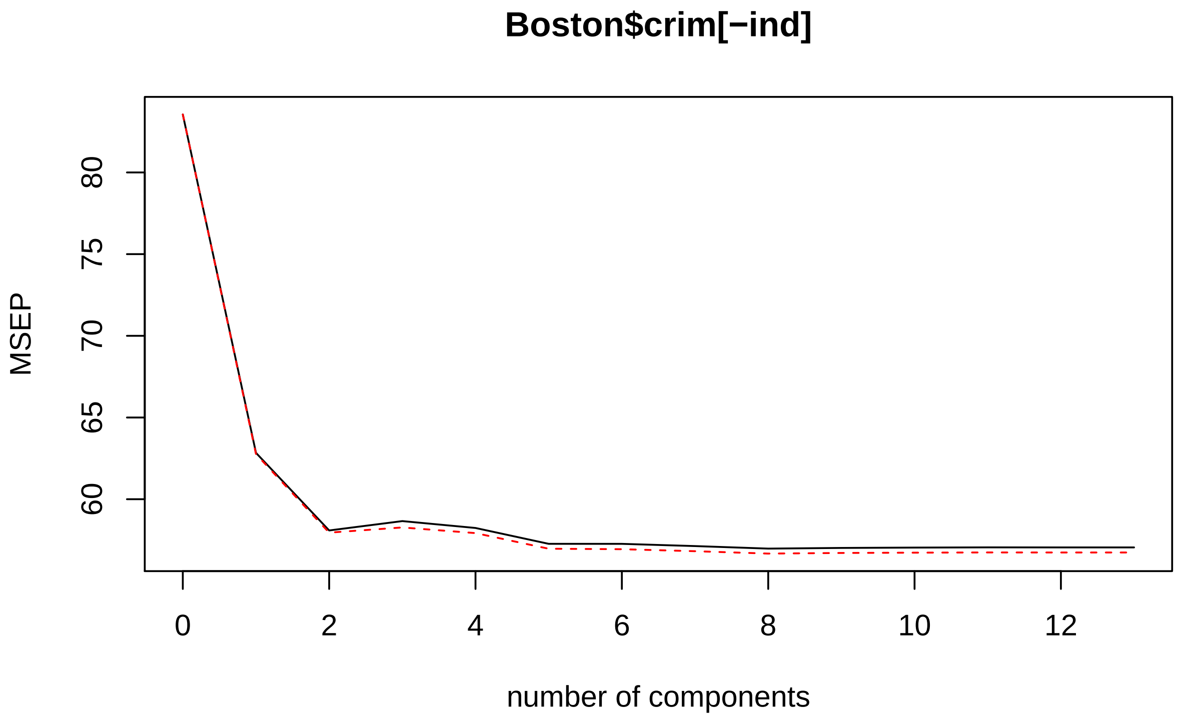
Various models and their test RMSEs from cross validation are as below:

- Subset Selection : 6.45
- Ridge:5.02
- Lasso: 6.60
- PCR: 5.19
- PLS: 5.13

For this problem, I would suggest to go for PLS model which uses only 8 components with a very low error. Although the lowest error is being provided by Ridge compression, it includes all the variables of the model. PLS has reduced features while capturing most of the features of the model.

(c) Does your chosen model involve all of the features in the data set? Why or why not?

No, PLS includes the first 8 Principal components. If we look at the plot below, it can be seen that for 8 components, minimum error is being achieved.



Book Questions: Chapter 8

Question 8

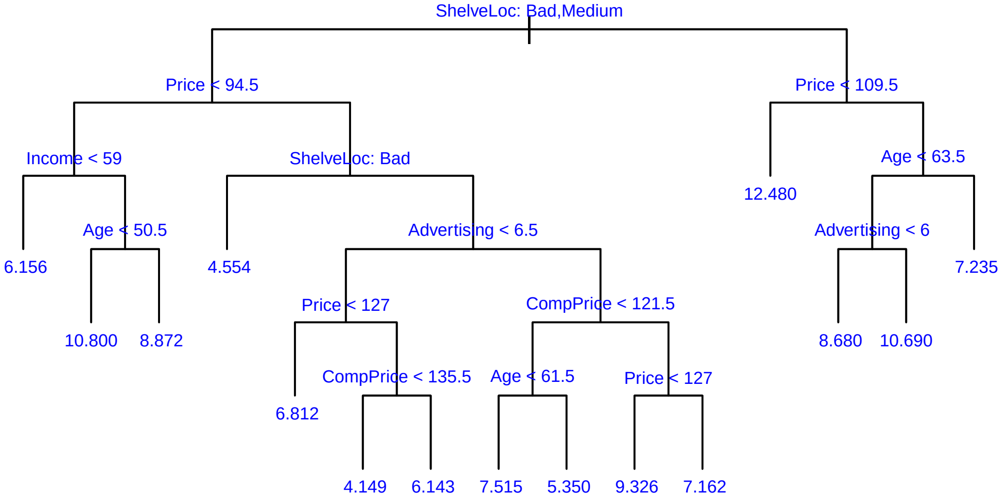
In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

(a) Split the data set into a training set and a test set.

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

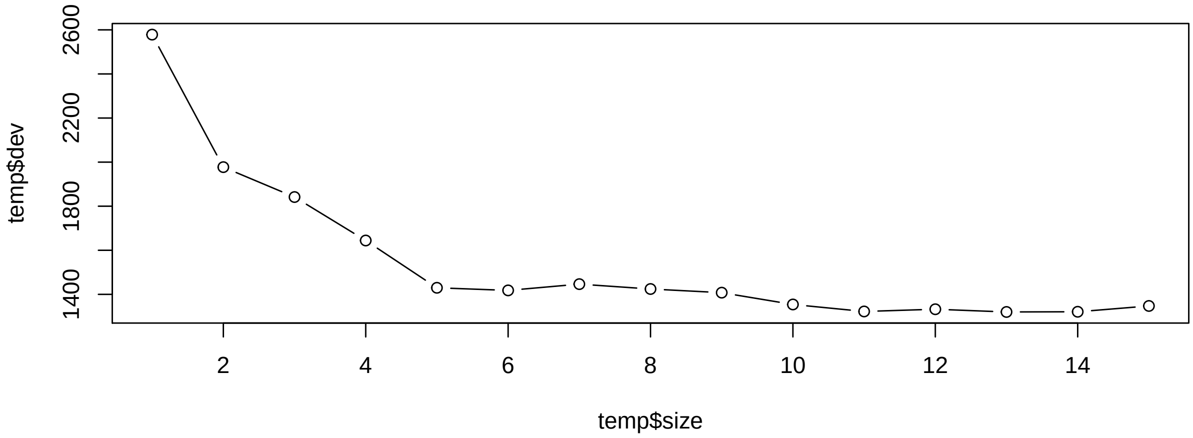
> Looking at the frame from summary of the tree, it can be derived that Price and Shelving location are the two variables which explain a lot of deviance in different nodes. Thus, they can be considered as the most important variables of the tree. The Test MSE obtained from this tree is 5.6 which is 2.36 RMSE.This means that overall Sales will vary by ~ +/- 5 units from the prediction.Pruning this tree reduces the error to +/-4 units.

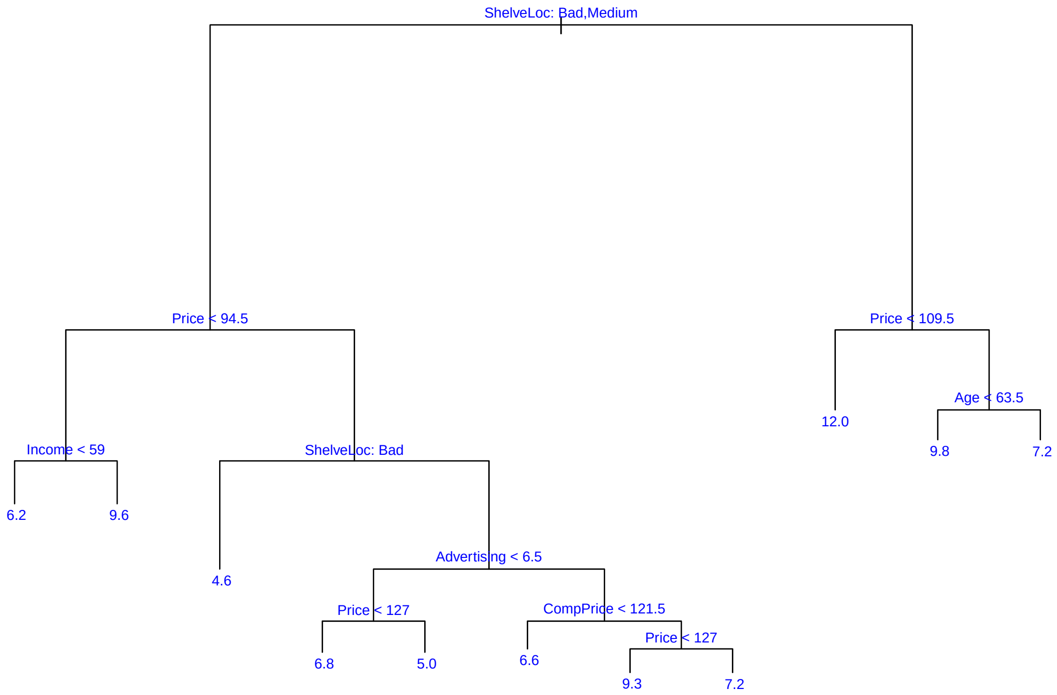


```
## Test MSE:  5.61023

## Test MSE After Pruning:  4.836681
```

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

After using cross validation, the best size came out to be 11. Yes, pruning the tree helps to improve the error in this case although the drcrease in test MSE is very less (5.38 from 5.6). The new MSE after Pruning is 5.38.





```
## Test Error from simple tree with pruning and cross validation :  5.380093
```

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance () function to determine which variables are most important.

Bagging technique of building trees produces an MSE of 3.51 with Shelving location and Price as the most important variables.

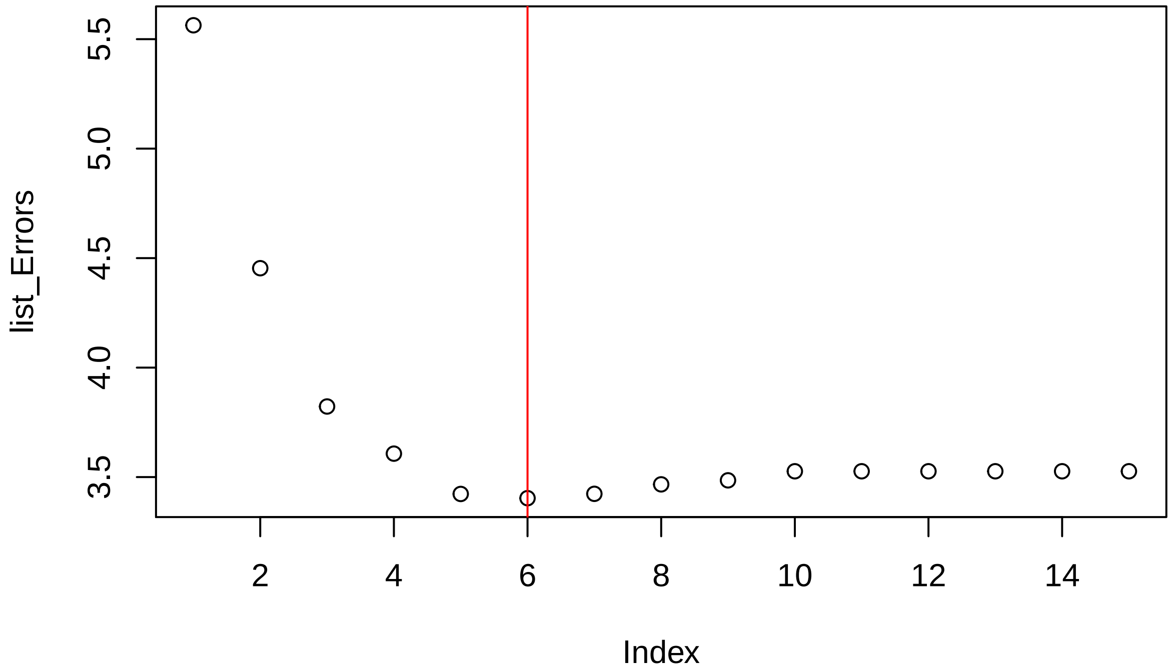
```
##             IncNodePurity
## CompPrice       49.720570
## Income          19.697448
## Advertising     52.986537
## Population       6.713099
## Price          509.394967
## ShelveLoc      770.568633
## Age             72.445509
## Education        1.374885
## Urban            0.000000
## US               3.209238
```

```
## Test MSE From Bagging:  3.51935
```

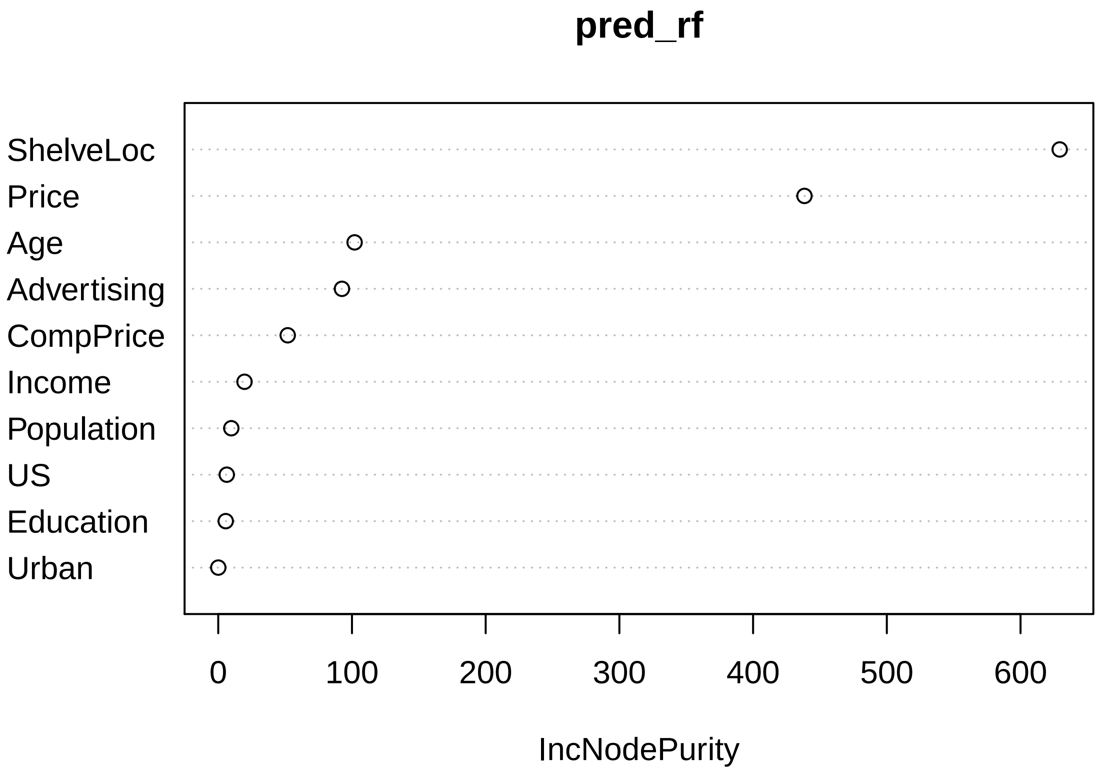
(e) Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

Using Random Forests to build tree produces a test error of 3.4, which is the least among the other methods tried before. Also, for this estimate I took the value for mtry = 6 (derived from the Error vs Mtry graph below), ntree=600 and nodesize=50.The most important variable comes out to be Shelving location, followed by price.

I also tried various values of m to capture its effect on the Test MSE obtained.



```
##              IncNodePurity
## CompPrice     51.96270845
## Income        19.63730272
## Advertising   92.50621186
## Population     9.75699773
## Price        438.40536787
## ShelveLoc    629.29339062
## Age          101.97336125
## Education      5.59234220
## Urban          0.04197248
## US             6.37130083
```



```
## Test MSE From Bagging:  3.414118
```

Question 11

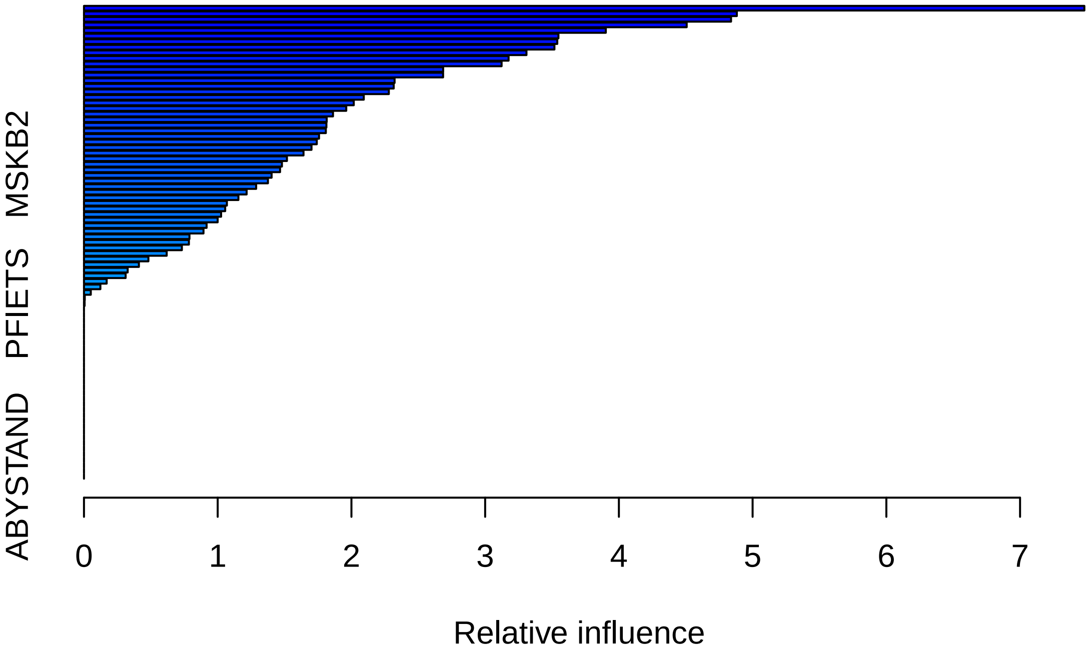
This question uses the Caravan data set.

(a) Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.

(b) Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important?

> With a bernoulli distribution(since its a categorical response variable, and interaction depth of 4, 'PPERSAUT' and 'MOPLHOOG' appear to be the most important variables for this model followed by 'MGODGE' and 'MKOOPKLA'.

```
## Top 6 predictive variables are:
```



```
##                var  rel.inf
## PPERSAUT PPERSAUT 7.480819
## MOPLHOOG MOPLHOOG 4.882054
## MGODGE     MGODGE 4.838870
## MKOOPKLA MKOOPKLA 4.507280
## MOSTYPE   MOSTYPE 3.902338
## MGODPR     MGODPR 3.547892
```

(c) Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20 %. Form a confusion matrix. What fraction of the people predicted to make a purchase do in fact make one? How does this compare with the results obtained from applying KNN or logistic regression to this data set?

> From Gradient Boosting, 29% of the people who have been predicted to make a purchase do actually make one.
>
> For Logistic regression, around 14% people who were predicted to make a purchase actually made one.
>
> And lastly, for KNN, 17% people who were predicted to make a purchase do actually make one.
>
> The Boosting method classifies the categorical response variable the best amongst all in for this problem.

```
## 
## yhat_new    0    1
##        0 4509  279
##        1   24   10
```

```
## 
## yhat_logit_test2    0    1
##                0 4183  231
##                1  350   58
```

```
## 
## knn_pred    0    1
##        0 4528  288
##        1    5    1
```