Wooten_Reece_Exam

Reece Wooten

8/4/2017

Question 10 From Chapter 2

(a)

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

```
set.seed(1)

library(MASS)
x=Boston
?Boston
cat("The Number of Rows",nrow(x))
```

```
## The Number of Rows 506
```

```
cat("The Number of Columns",ncol(x))
```

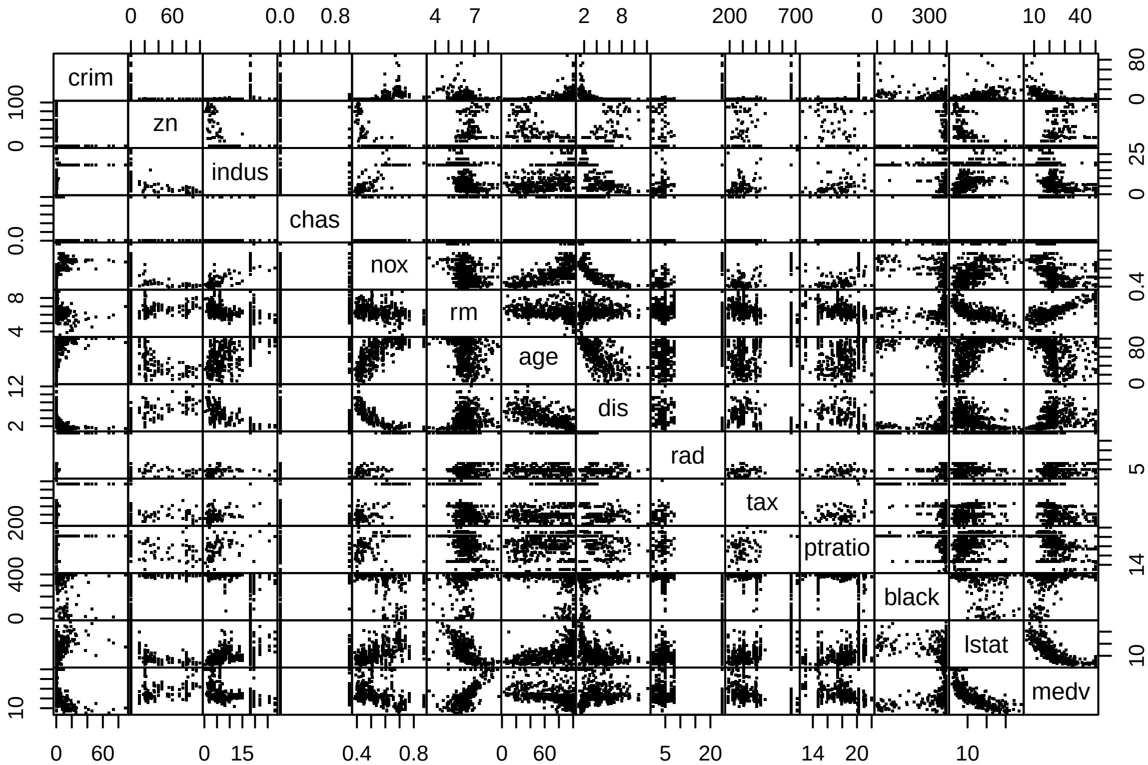
```
## The Number of Columns 14
```

- The Rows of the Boston data set are neighborhoods/suburbs in the Boston area.
- The Columns of the data set are the various attributes of those suburbs, crime rate, demographic factors, etc…

(b)

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

```
set.seed(1)

pairs(x[, ], gap = 0, pch = ".")
```



- The pair wise plots above show a positive relationship between crime rate nitrogen oxide concentration, age of the suburbs, and percent of lower status people in the population. There also seems to be a negative relationship between crime rate and the weighted mean of distances to five Boston employment centers and the median value of owner occupied homes.
- nitrogen oxide content appears to have a positive relationship with the age of the neighborhood and a negative relationship between the weighted mean of distances to five Boston employment centers.
- rm appears to have a positive relationship with median value of owner occupied homes and a negative relationship with the percent of lower status people in the population.
- age appears to have a negative relationship with the mean distance to five Boston employment centers, and a positive relationship with the percent of lower status people in the population.
- The percent of lower status people in the population appears to have a negative relationship with the median value of owner occupied homes.

(c)

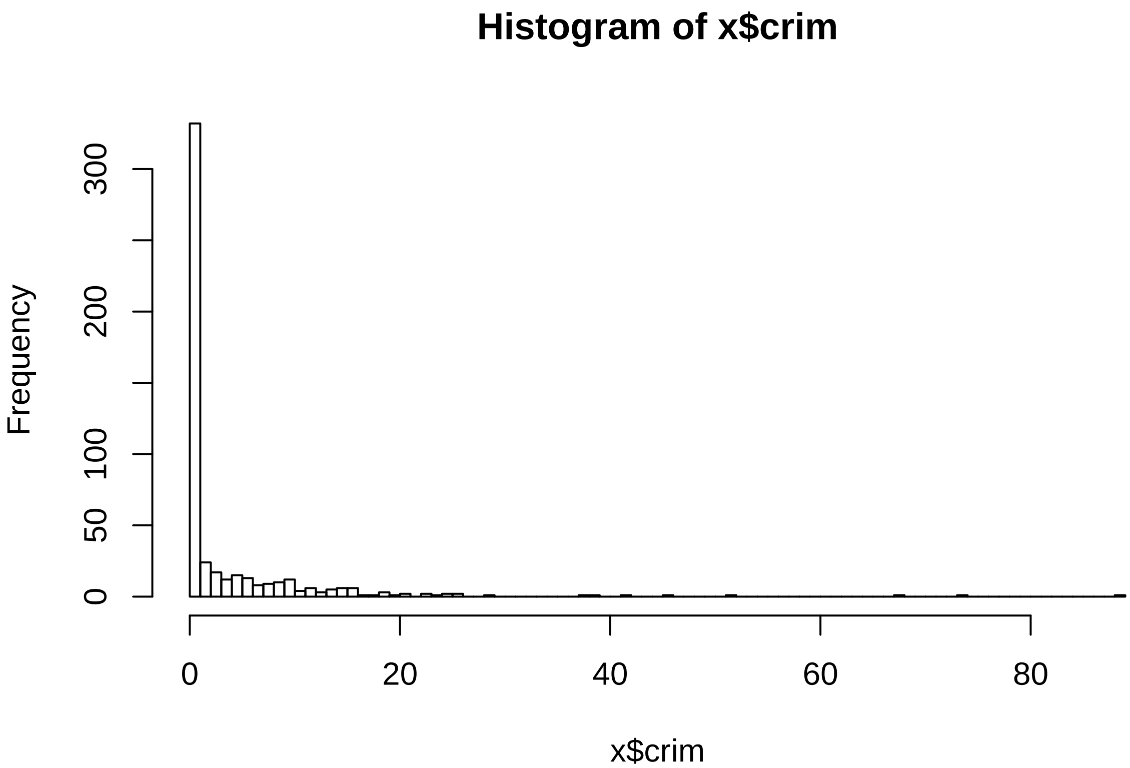
Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

- The pair wise plots above show a positive relationship between crime rate and nitrogen oxide concentration, age of the suburbs, and percent of lower status people in the population. There also seems to be a negative relationship between crime rate and the weighted mean of distances to five Boston employment centers and the median value of owner occupied homes. There could also be a relationship between crime rate and rooms, but it is not clear from the pairwise graph if its positive or negative.

(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.

```r
set.seed(1)

hist(x$crim,breaks=100)
```




```r
cat('Range of Crime Rate:',range(x$crim))
```

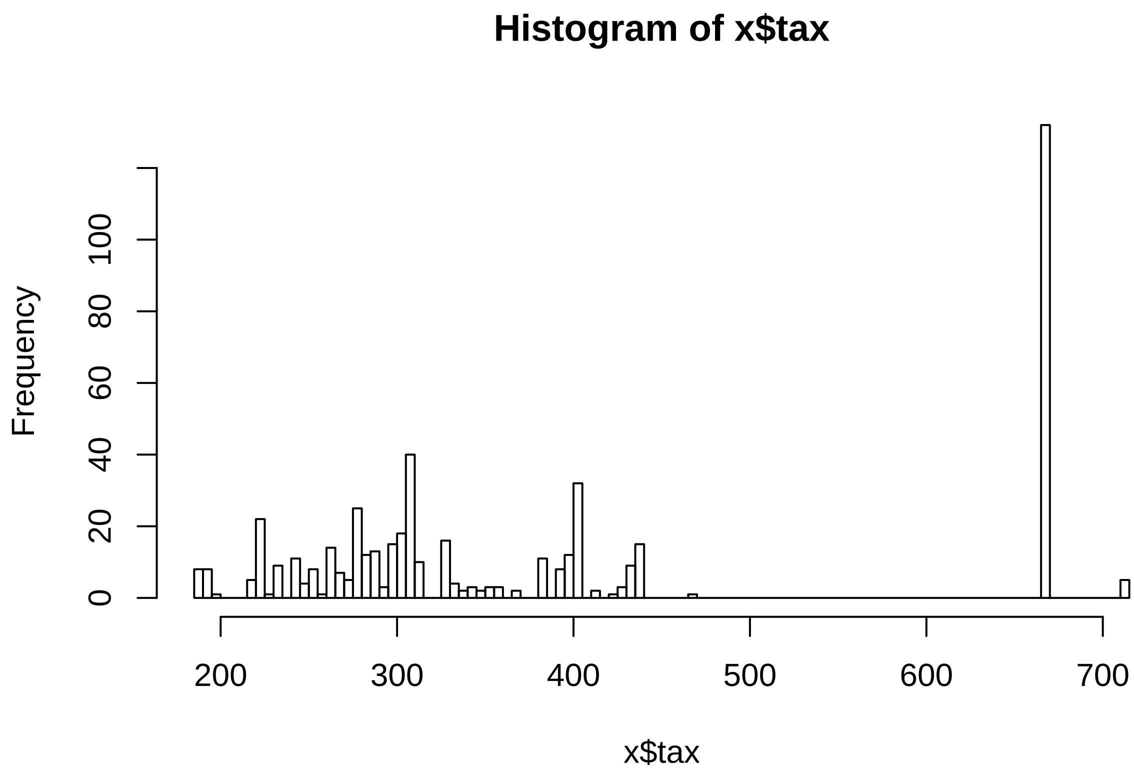
```
## Range of Crime Rate: 0.00632 88.9762
```

```r
summary(x$crim)
```

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

- There does seem to be a number of suburbs that have a per capita crime rate above 20%, up to the 3rd quartile being below 3.67, but a max of 88.98.

```r
hist(x$tax,breaks=100)
```




```r
cat('Range of Tax Breaks:',range(x$tax))
```

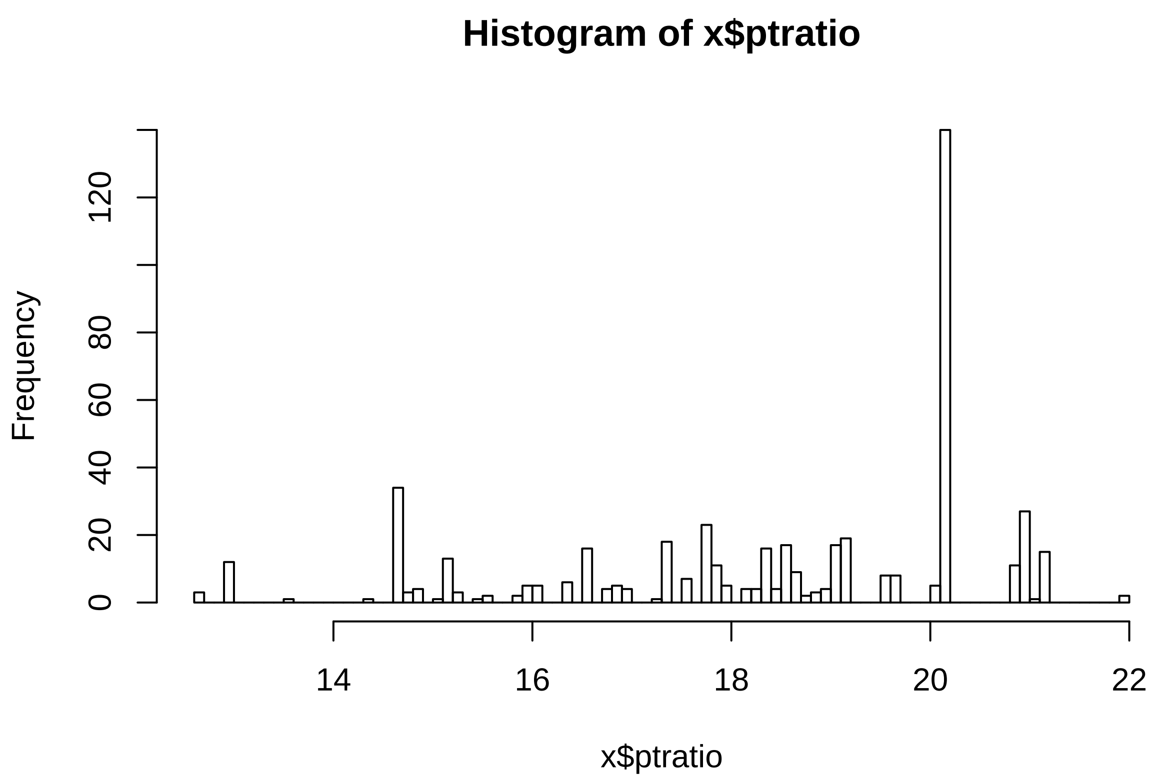
```
## Range of Tax Breaks: 187 711
```

```r
summary(x$tax)
```

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

- Half of the full value property tax rates per $10,000 fall of the suburbs fall below 330. while some reach a max of 711 which is significantly higher than the median.

```r
hist(x$ptratio,breaks=100)
```




```r
cat('Range of Pupil-Teacher Ratio:',range(x$ptratio))
```

```
## Range of Pupil-Teacher Ratio: 12.6 22
```

```r
summary(x$ptratio)
```

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

- The median pupil to teacher ratio in the Boston suburbs is 19.05, with a max of 22, and a min of 12.60. This variable does not have as extreme of a range as the other two variables mentioned.

(e)

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

```r
set.seed(1)

cat("Suburbs That Bound the Charles River Table:",table(x$chas))
```

```
## Suburbs That Bound the Charles River Table: 471 35
```

- The number of suburbs that bound the Charles River is 35

(f)

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

```r
set.seed(1)

summary(x$ptratio)
```

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

- The Median Pupil to Teacher Ratio among the towns is 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.

```r
set.seed(1)

which(x$medv==5)
```

```
## [1] 399 406
```

```r
which(x$medv<5)
```

```
## integer(0)
```

- There are two suburbs that share the lowest median value of owner-occupied homes which are suburbs 399 and 406.

```r
set.seed(1)

print('Suburb 309:')
```

```
## [1] "Suburb 309:"
```

```r
x[399,]
```

```
##        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    20.2 396.9
##     lstat medv
## 399 30.59    5
```

```r
print('Suburb 406:')
```

```
## [1] "Suburb 406:"
```

```r
x[406,]
```

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

- The output above shows the values of the other predictors for suburbs 399 and 406.

```r
set.seed(1)

cat('crim Range',range(x$crim))
```

```
## crim Range 0.00632 88.9762
```

- Both Suburbs have relatively high crime rates compared to the range, especially suburb 406. This is likely due to the neighborhood being extremely impoverished.

```
set.seed(1)

cat('zn Range:',range(x$zn))
```

```
## zn Range: 0 100
```

- Both Suburbs have the minimum amount of residential land zoned lots over 25,000 sq.fit. This is likely related to the median value of owner occupied homes being low. large residential land zoned lots are more likely in richer neighborhoods.

```
set.seed(1)

cat('indus Range:',range(x$indus))
```

```
## indus Range: 0.46 27.74
```

- Both Suburbs have the same amount of non-retail business acres per town, from the range it seems they are in the middle of the range. It seems as though having low median value of owner occupied homes does not depend on how much non-retail business acres are allotted in the town.

```
set.seed(1)

cat('chas Range:',range(x$chas))
```

```
## chas Range: 0 1
```

- Both suburbs aren't bounding the Charles River.

```
set.seed(1)

cat('nox range:',range(x$nox))
```

```
## nox range: 0.385 0.871
```

- The nitrogen oxide concentration for both suburbs are the same and is slightly higher than the minimum.

```
set.seed(1)

cat('rm Range:',range(x$rm))
```

```
## rm Range: 3.561 8.78
```

- Both suburbs have the same average number of rooms per dwelling, and is in the middle of the range.

```
set.seed(1)

cat('age Range:',range(x$age))
```

```
## age Range: 2.9 100
```

- The age of the suburbs dwellings are the same and all the dwellings are built prior to 1940. This indicated the neighborhoods are aging and have not been developed in a while, potentially discouraging business owners to build businesses in the area.

```
set.seed(1)

cat('dis Range:',range(x$dis))
```

```
## dis Range: 1.1296 12.1265
```

- Both suburbs have the same weighted mean of distances to five Boston employment centers and they are both close to the minimum of the range. This indicates an attempt to put employment centers close to the poor neighborhoods in the Boston area to find people jobs.

```
set.seed(1)

cat('rad Range:',range(x$rad))
```

```
## rad Range: 1 24
```

- Both suburbs have the same amount of accessibility to radial highways which is at the max of the range.

```
set.seed(1)

cat('tax Range:',range(x$tax))
```

```
## tax Range: 187 711
```

- Both suburbs have the same property tax rate, which is relatively high in the range.

```
set.seed(1)

cat('ptratio Range:',range(x$ptratio))
```

```
## ptratio Range: 12.6 22
```

- Both suburbs have the same pupil teacher ratio and is relatively high in the range, a high pupil-teacher ratio indicated poor management/funding in the school districts which typically perpetuates poor neighborhoods.

```
set.seed(1)

cat('black Range:',range(x$black))
```

```
## black Range: 0.32 396.9
```

- Suburb 399 has the max black proportion, and suburb 406 also has a relatively high black proportion compared to the range. This could indicate that black citizens are being displaced, or disadvantaged in the Boston surrounding area.

```
set.seed(1)

cat('lstat Range:',range(x$lstat))
```

```
## lstat Range: 1.73 37.97
```

- Both Suburbs have relatively high lower status population compared to the range. This is in line with the median value of owner occupied homes being low also.

(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.

```
set.seed(1)

table(x$rm>7)
```

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

```r
table(x$rm>8)
```

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

```r
rm(list=ls())
```

- The number of suburbs that average more than 7 rooms per dwelling is 64, and the number of suburbs that average more than 8 rooms per dwelling is 13.
- All the suburbs with an average of 8 or more rooms have crime rates below 1, The majority of the suburbs have a median value of owner occupied homes above 40, and the majority of the suburbs have a lower status percent of their population below 5. These attributes seem to indicate that these suburbs are of the more affluent in the Boston area.

Question 15 From Chapter 3

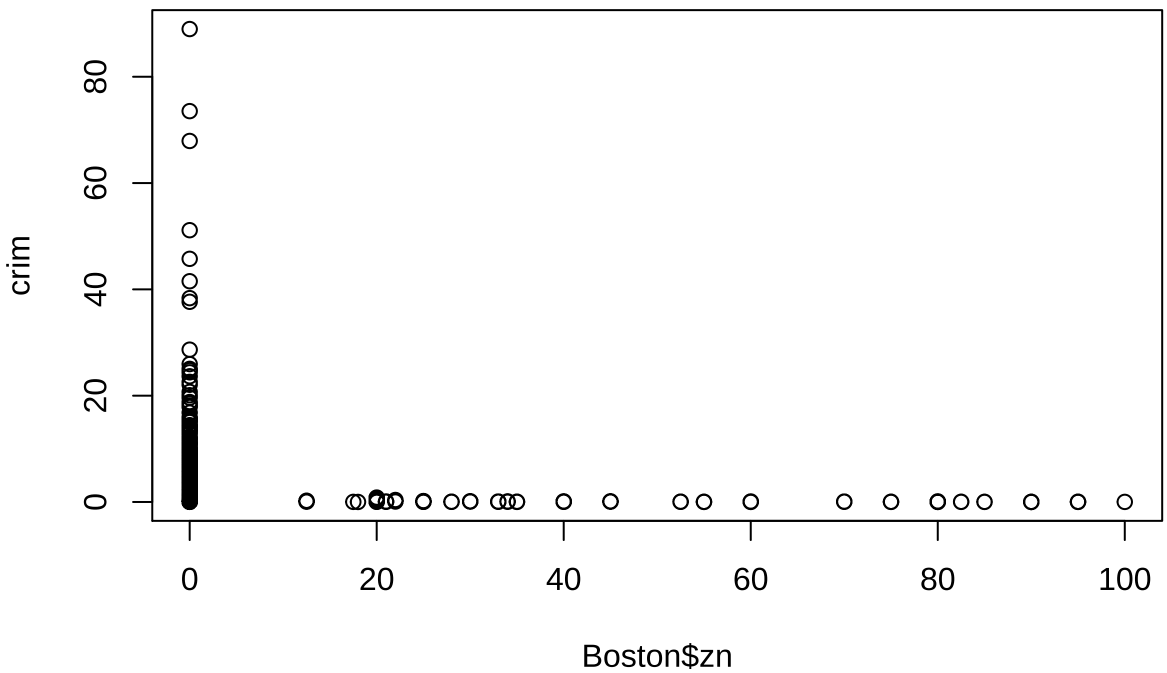
(a)

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.

```r
set.seed(1)

library("MASS")
attach(Boston)
Boston=Boston
fit1=lm(Boston$crim~Boston$zn)
summary(fit1)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$zn)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.429 -4.222 -2.620  1.250 84.523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
## Boston$zn   -0.07393    0.01609  -4.594 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
## F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

```r
plot(Boston$zn,crim)
```



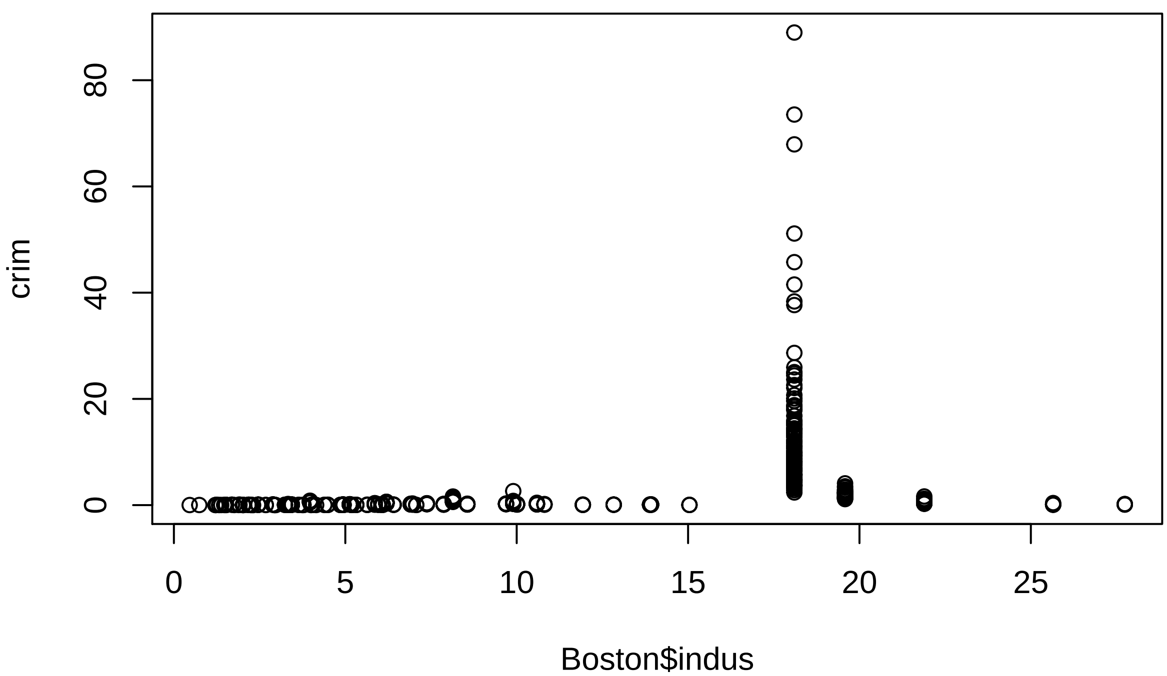
- A one percentage point increase in the residential land zoned for lots over 25,000 ft will decrease crime by .074 percentage points per capita on average. This variable is statistically significant at the 5% level.

```r
set.seed(1)

Boston$zn2=zn^2
Boston$zn3=zn^3
Boston$indus2=indus^2
Boston$indus3=indus^3
Boston$chas2=chas^2
Boston$chas3=chas^3
Boston$nox2=nox^2
Boston$nox3=nox^3
Boston$rm2=rm^2
Boston$rm3=rm^3
Boston$age2=age^2
Boston$age3=age^3
Boston$dis2=dis^2
Boston$dis3=dis^3
Boston$rad2=rad^2
Boston$rad3=rad^3
Boston$tax2=tax^2
Boston$tax3=tax^3
Boston$ptratio2=ptratio^2
Boston$ptratio3=ptratio^3
Boston$black2=black^2
Boston$black3=black^3
Boston$lstat2=lstat^2
Boston$lstat3=lstat^3
Boston$medv2=medv^2
Boston$medv3=medv^3
```

```r
set.seed(1)

fit2=lm(crim~indus)
summary(fit2)
```

```
##
## Call:
## lm(formula = crim ~ indus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.972  -2.698  -0.736   0.712  81.813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.06374    0.66723  -3.093  0.00209 **
## indus        0.50978    0.05102   9.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.866 on 504 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637
## F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
plot(Boston$indus,crim)
```

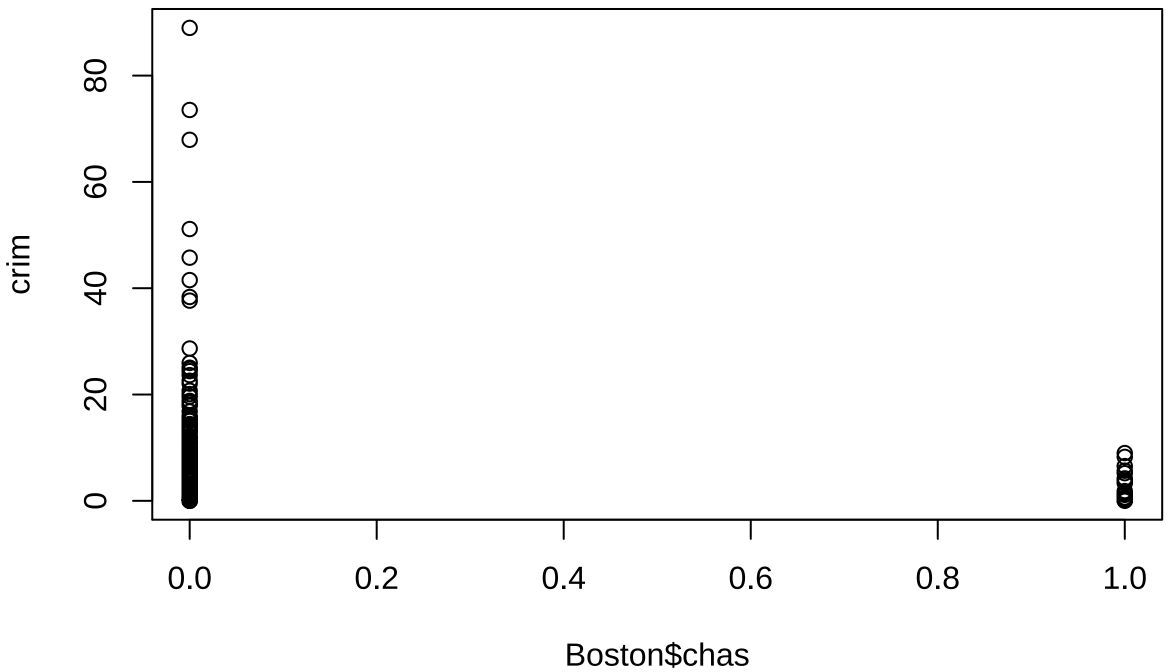


- A one percentage point increase in non-retail business acres per town will decrease crime by 2.06 percentage points on average. This variable is statistically significant at the 5% level.

```r
set.seed(1)

fit3=lm(crim~Boston$chas)
summary(fit3)
```

```
##
## Call:
## lm(formula = crim ~ Boston$chas)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7444     0.3961   9.453   <2e-16 ***
## Boston$chas  -1.8928     1.5061  -1.257    0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.597 on 504 degrees of freedom
## Multiple R-squared:  0.003124,   Adjusted R-squared:  0.001146
## F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

```
plot(Boston$chas,crim)
```



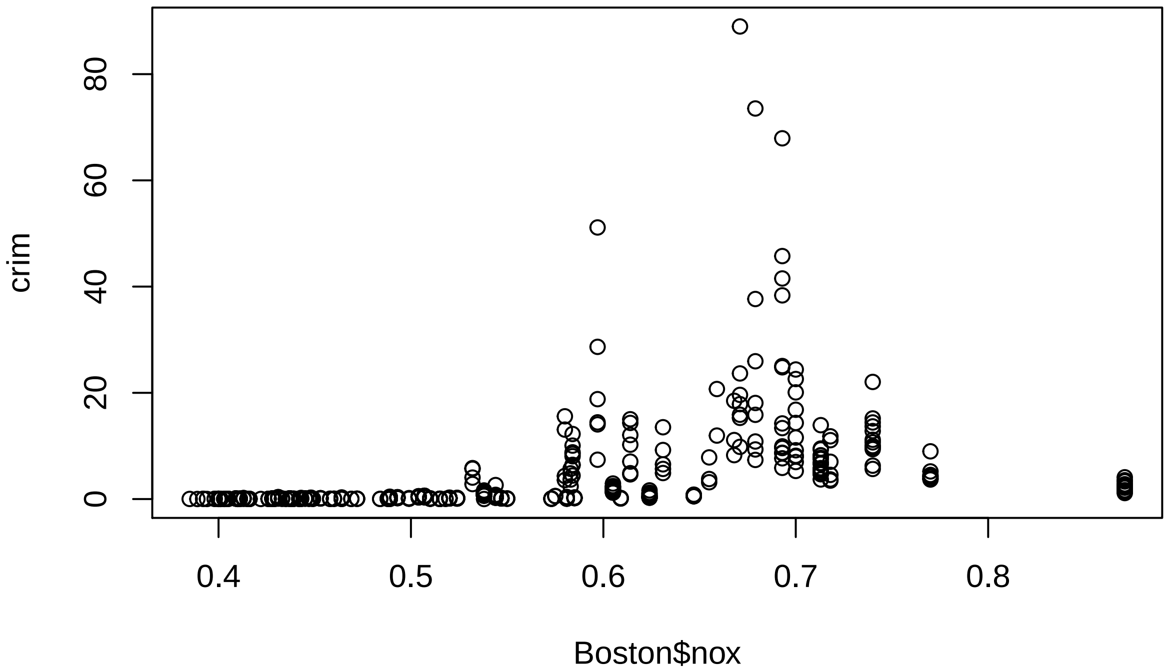
- Suburbs that bound the Charles River will have a 1.89 percentage point less crime rate than the ones that don't on average. This variable is not statistically significant at the 5% level.

```
set.seed(1)

fit4=lm(crim~Boston$nox)
summary(fit4)
```

```
##
## Call:
## lm(formula = crim ~ Boston$nox)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -12.371  -2.738  -0.974   0.559  81.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
## Boston$nox    31.249      2.999  10.419  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
plot(Boston$nox,crim)
```



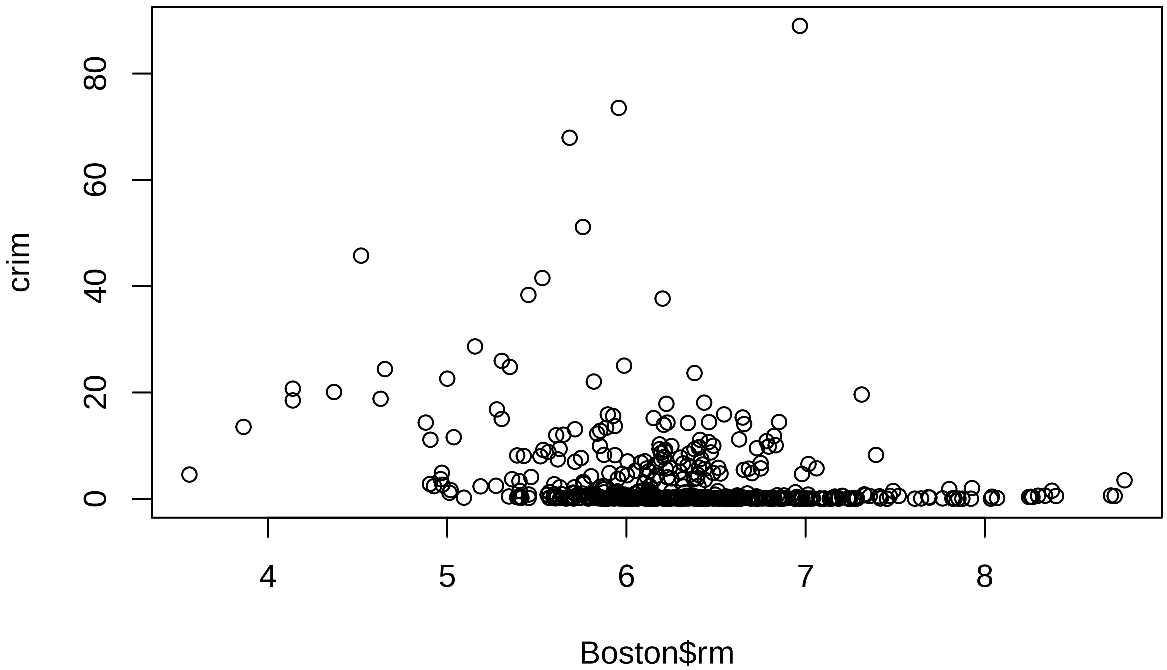
- A one unit increase in the nitrogen oxide concentration in a suburb will increase the crime rate by 31.249 on average. This variable is statistically significant at the 5% level.

```
set.seed(1)

fit5=lm(crim~Boston$rm)
summary(fit5)
```

```
##
## Call:
## lm(formula = crim ~ Boston$rm)
##
## Residuals:
##     Min      1Q Median     3Q    Max
## -6.604 -3.952 -2.654  0.989 87.197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   20.482      3.365   6.088 2.27e-09 ***
## Boston$rm     -2.684      0.532  -5.045 6.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.401 on 504 degrees of freedom
## Multiple R-squared:  0.04807,    Adjusted R-squared:  0.04618
## F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

```r
plot(Boston$rm,crim)
```



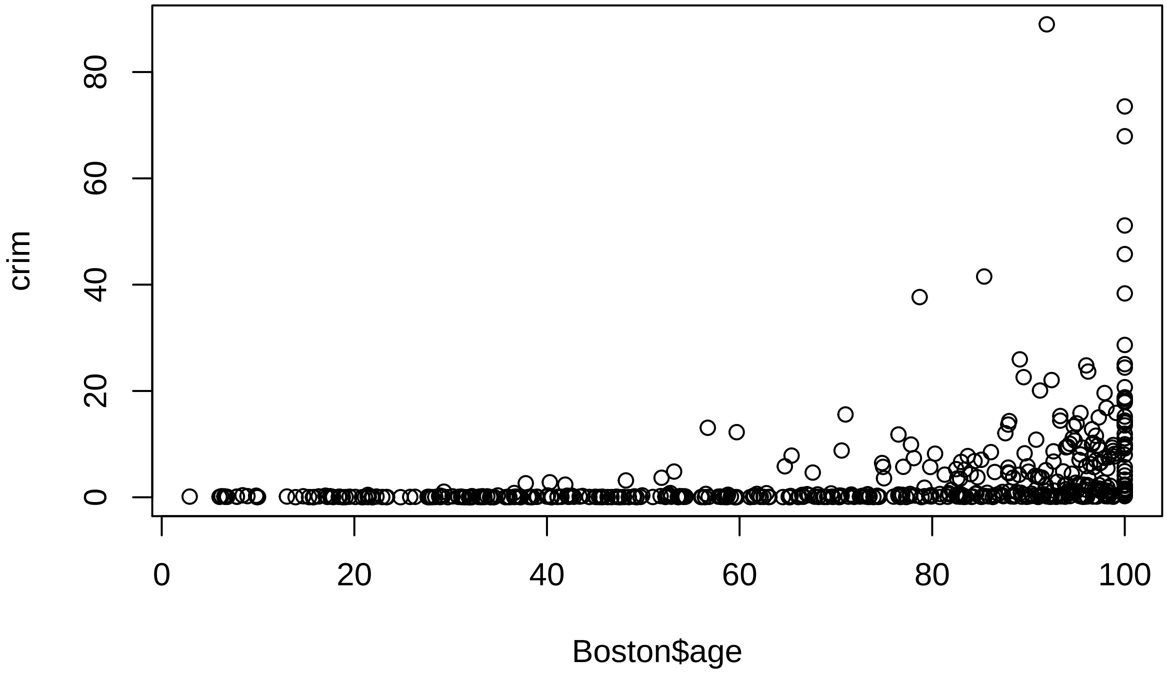
- A one room increase in the average number of rooms per dwelling will decrease the crime rate by 2.684% on average. This variable is statistically significant at the 5% level.

```r
set.seed(1)

fit6=lm(crim~Boston$age)
summary(fit6)
```

```
##
## Call:
## lm(formula = crim ~ Boston$age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.789 -4.257 -1.230  1.527 82.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
## Boston$age   0.10779    0.01274   8.463 2.85e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
## F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

```r
plot(Boston$age,crim)
```

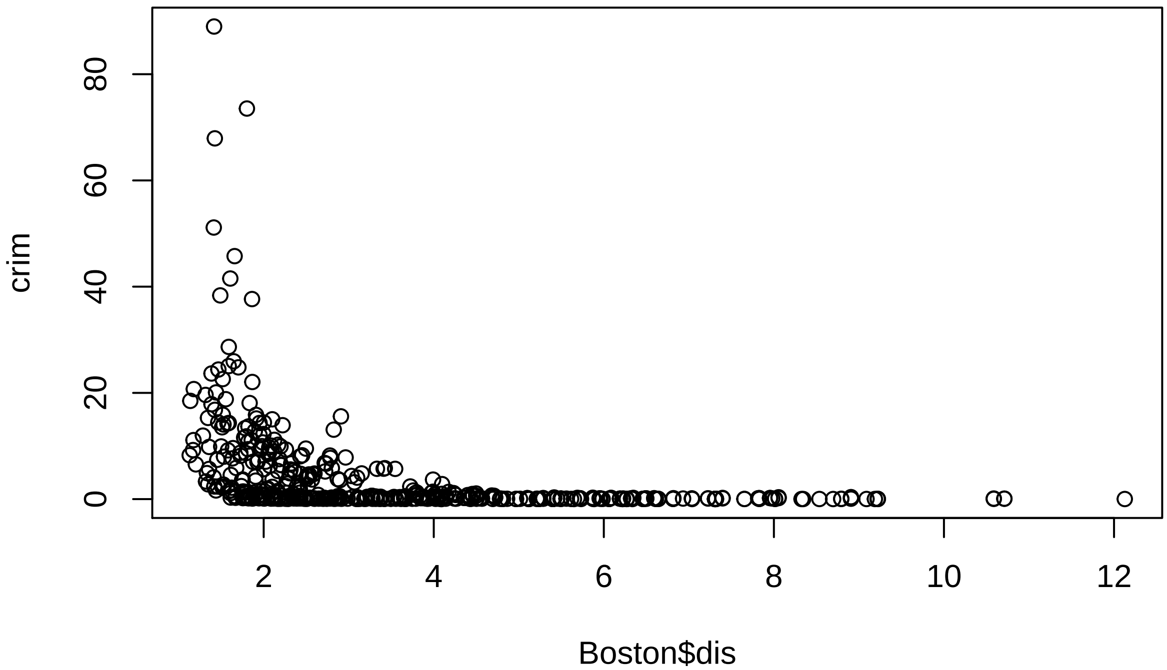


- A one percentage point increase in the proportion of owner occupied units built prior to 1940 will increase the crime rate by .1078 percentage points on average. This variable is statistically significant at the 5% level

```r
set.seed(1)

fit7=lm(crim~Boston$dis)
summary(fit7)
```

```
##
## Call:
## lm(formula = crim ~ Boston$dis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.708 -4.134 -1.527  1.516 81.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4993     0.7304  13.006   <2e-16 ***
## Boston$dis   -1.5509     0.1683  -9.213   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.965 on 504 degrees of freedom
## Multiple R-squared:  0.1441, Adjusted R-squared:  0.1425
## F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
plot(Boston$dis,crim)
```

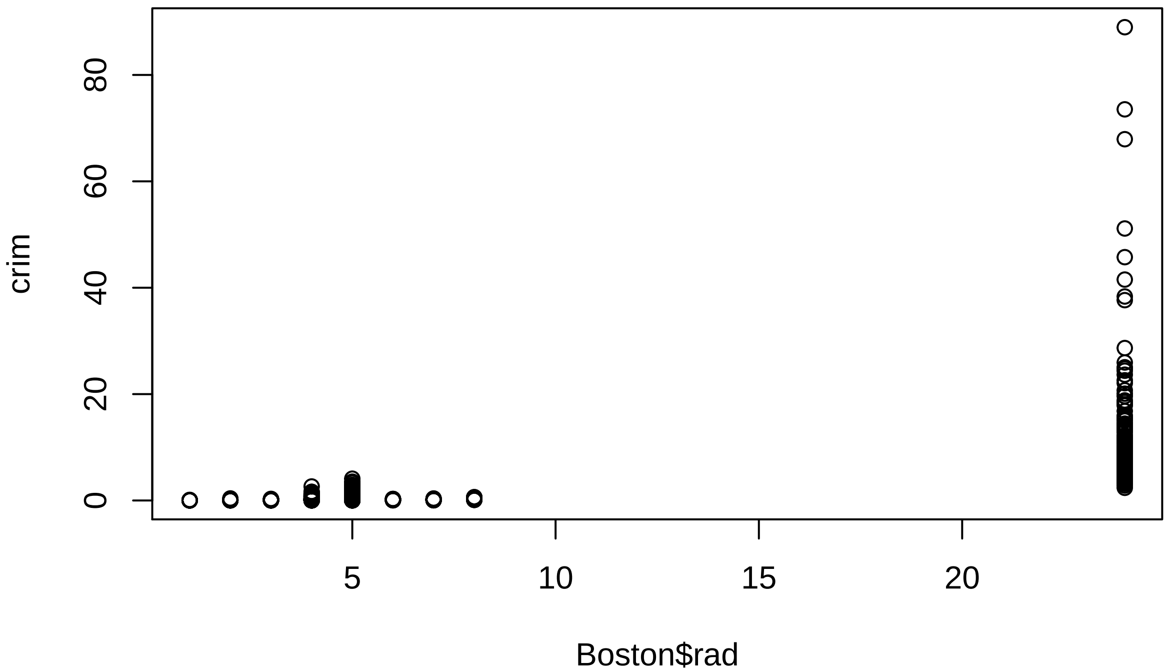


- A one unit increase in the weighted mean of distances to five Boston employment centers will decrease the crime rate by 1.55% on average. This variable is statistically significant at the 5% level.

```r
set.seed(1)

fit8=lm(crim~Boston$rad)
summary(fit8)
```

```
##
## Call:
## lm(formula = crim ~ Boston$rad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.164  -1.381  -0.141   0.660  76.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
## Boston$rad   0.61791    0.03433  17.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.718 on 504 degrees of freedom
## Multiple R-squared:  0.3913, Adjusted R-squared:   0.39
## F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
plot(Boston$rad,crim)
```

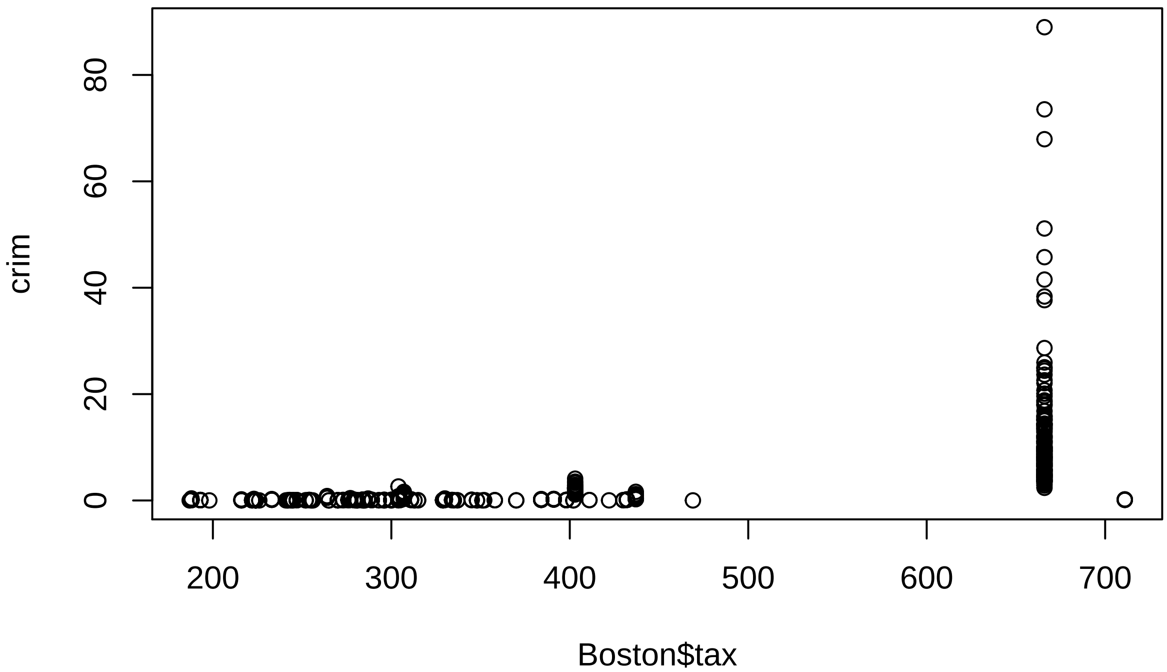


- A one unit increase in the index of accessibility to radial highways will increase the crime rate by .618% on average. This variable is statistically significant at the 5% level.

```r
set.seed(1)

fit9=lm(crim~Boston$tax)
summary(fit9)
```

```
##
## Call:
## lm(formula = crim ~ Boston$tax)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.513  -2.738  -0.194   1.065  77.696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
## Boston$tax   0.029742   0.001847   16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.997 on 504 degrees of freedom
## Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
## F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
plot(Boston$tax,crim)
```

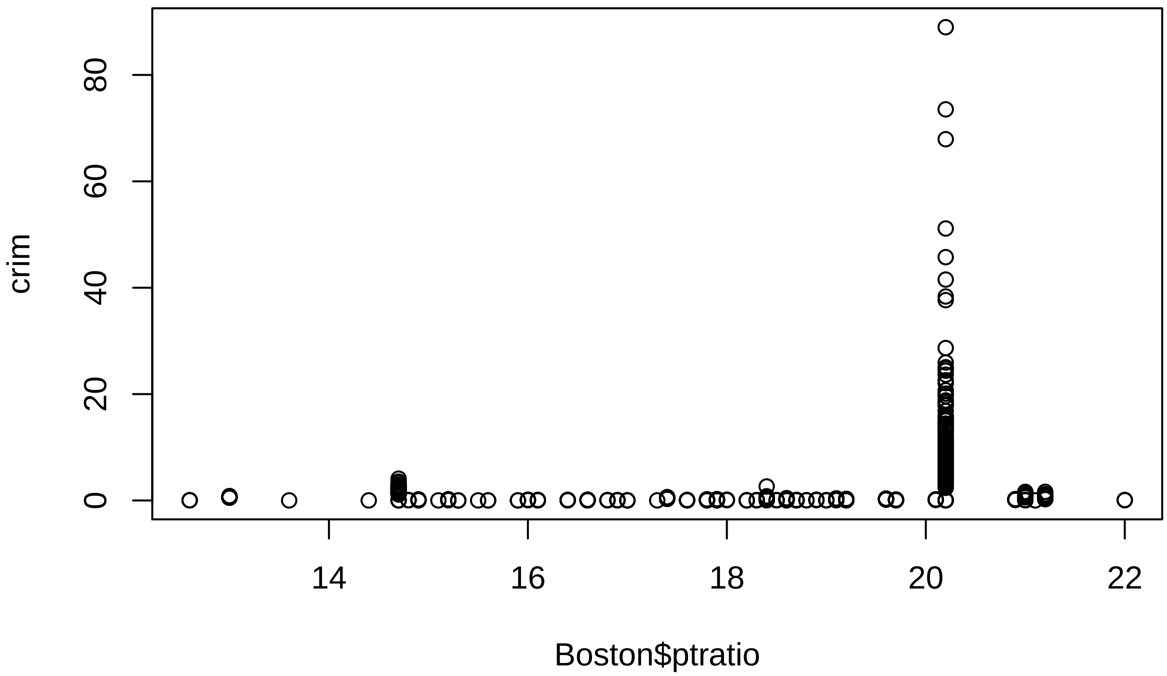


- A one percentage point increase in the full-value property-tax rate will increase crime by .0297 percentage points on average. This value is statistically significant at the 5% level.

```r
set.seed(1)

fit10=lm(crim~Boston$ptratio)
summary(fit10)
```

```
##
## Call:
## lm(formula = crim ~ Boston$ptratio)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.654 -3.985 -1.912  1.825 83.353
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -17.6469     3.1473  -5.607 3.40e-08 ***
## Boston$ptratio   1.1520     0.1694   6.801 2.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.24 on 504 degrees of freedom
## Multiple R-squared:  0.08407,    Adjusted R-squared:  0.08225
## F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

```r
plot(Boston$ptratio,crim)
```

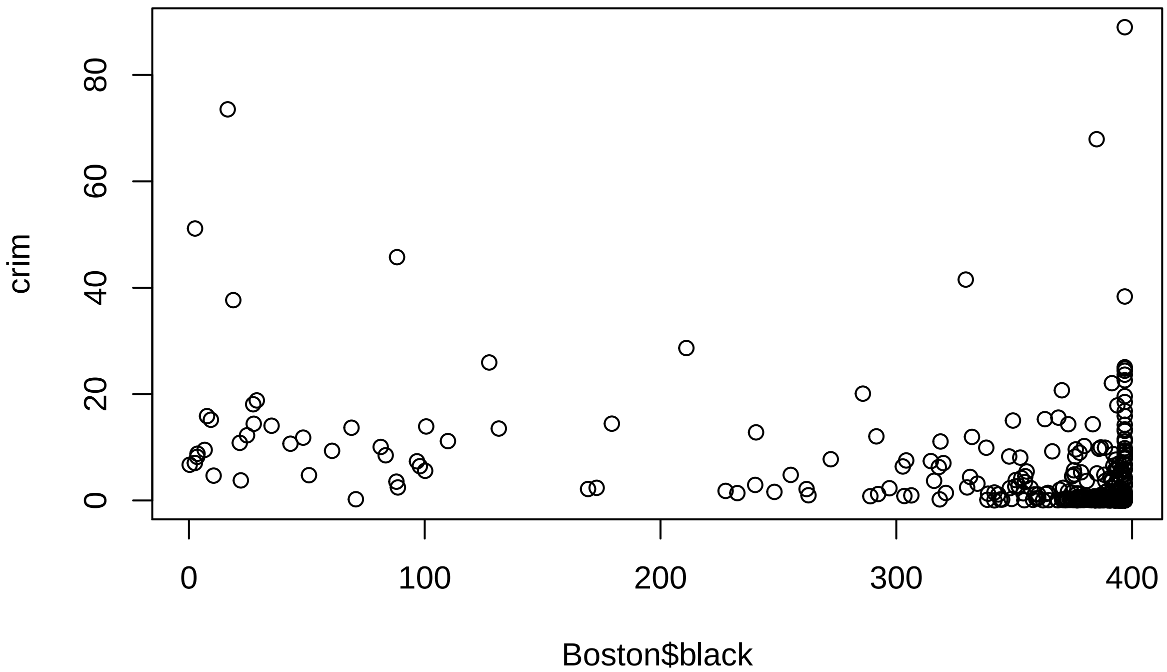


- A one percentage point increase in the pupil-teacher ratio will increase the crime rate by 1.15 percentage points on average. This variable is statistically significant at the 5% level.

```
set.seed(1)

fit11=lm(crim~Boston$black)
summary(fit11)
```

```
##
## Call:
## lm(formula = crim ~ Boston$black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.756  -2.299  -2.095  -1.296  86.822
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.553529   1.425903  11.609   <2e-16 ***
## Boston$black -0.036280   0.003873  -9.367   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.946 on 504 degrees of freedom
## Multiple R-squared:  0.1483, Adjusted R-squared:  0.1466
## F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
plot(Boston$black,crim)
```

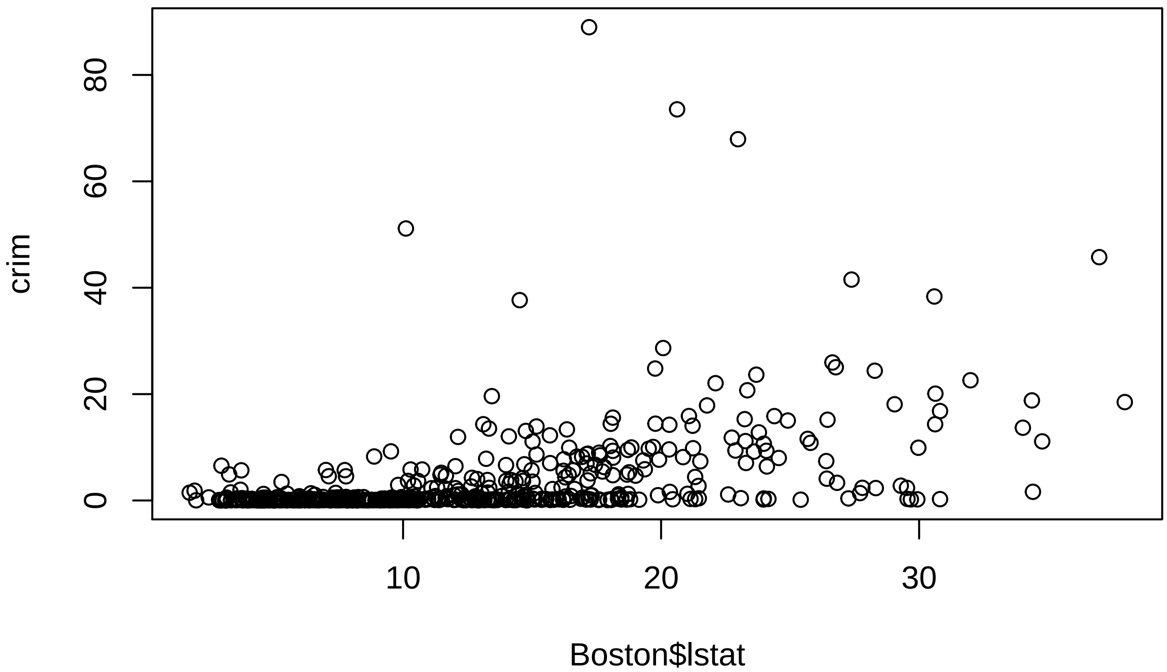


- A one percentage point increase in the black proportion in the suburbs will decrease the crime rate by .0362 percentage points on average. This variable is statistically significant at the 5% level

```
set.seed(1)

fit12=lm(crim~Boston$lstat)
summary(fit12)
```

```
##
## Call:
## lm(formula = crim ~ Boston$lstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.33054    0.69376  -4.801 2.09e-06 ***
## Boston$lstat  0.54880    0.04776  11.491  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
## F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
plot(Boston$lstat,crim)
```

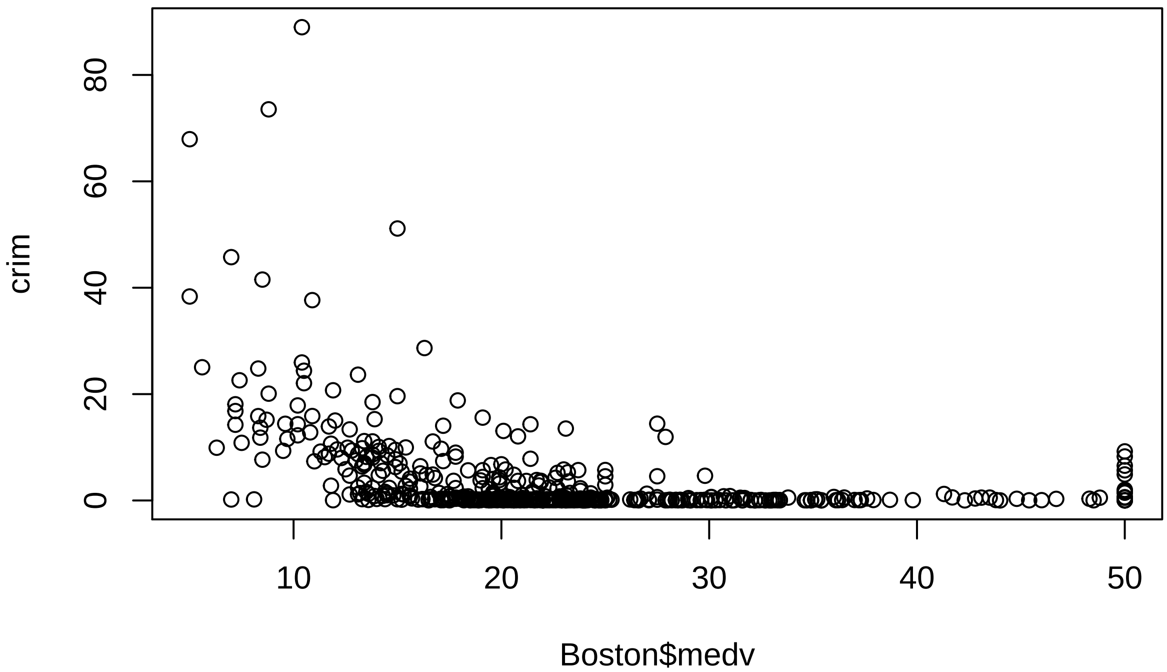


- A one percentage point increase in the percent of lower status population will increase crime by .548 percentage points on average. This variable is statistically significant at the 5% level.

```r
set.seed(1)

fit13=lm(crim~Boston$medv)
summary(fit13)
```

```
##
## Call:
## lm(formula = crim ~ Boston$medv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## Boston$medv -0.36316    0.03839   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
plot(Boston$medv,crim)
```



- A 1000 dollar increase in the median value of owner occupied homes will decrease the crime rate by .363% on average. This variable is statistically significant at the 5% level.

(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?

```
set.seed(1)

fit14=lm(crim~Boston$zn+Boston$indus+Boston$chas+Boston$nox+Boston$rm+Boston$age+Boston$dis+Boston$rad+I
summary(fit14)
```

```
##
## Call:
## lm(formula = crim ~ Boston$zn + Boston$indus + Boston$chas +
##     Boston$nox + Boston$rm + Boston$age + Boston$dis + Boston$rad +
##     Boston$tax + Boston$ptratio + Boston$black + Boston$lstat +
##     Boston$medv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     17.033228   7.234903   2.354 0.018949 *
## Boston$zn        0.044855   0.018734   2.394 0.017025 *
## Boston$indus    -0.063855   0.083407  -0.766 0.444294
## Boston$chas     -0.749134   1.180147  -0.635 0.525867
## Boston$nox     -10.313535   5.275536  -1.955 0.051152 .
## Boston$rm        0.430131   0.612830   0.702 0.483089
```

```
## Boston$age        0.001452   0.017925   0.081 0.935488
## Boston$dis       -0.987176   0.281817  -3.503 0.000502 ***
## Boston$rad        0.588209   0.088049   6.680 6.46e-11 ***
## Boston$tax       -0.003780   0.005156  -0.733 0.463793
## Boston$ptratio  -0.271081   0.186450  -1.454 0.146611
## Boston$black    -0.007538   0.003673  -2.052 0.040702 *
## Boston$lstat     0.126211   0.075725   1.667 0.096208 .
## Boston$medv     -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

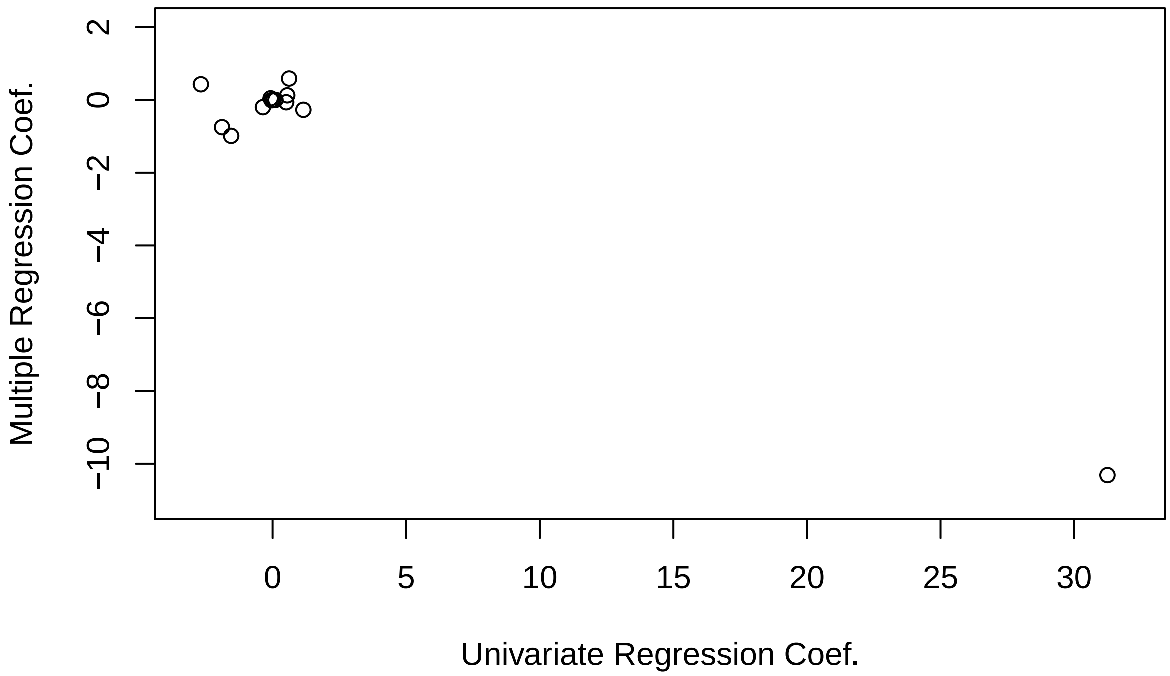
- zn, dis, rad, black, and medv, are all statistically significant at the 5% level, therefore they reject the null hypothesis that the true values of those coefficients are 0. The interpretation for the significant variables are:
- A one percentage point increase in the proportion of residential land zoned for lots over 25,000 sq.ft increase the crime rate by .045 percentage points on average and all else equal.
- A one unit increase in the weighted mean of distances to five Boston employment centers decrease the crime rate by .987% on average and all else equal
- A one unit increase in the index of accessibility to radial highways increase the crime rate by .588%
- A one percentage point increase in the proportion of the black population decreases crime by .007 percentage points
- A one unit increase in the median value of owner occupied homes will decrease crime by 19.8%

(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 regres- sion model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

```r
set.seed(1)

par(mfrow=c(1,1))
plot(1, type="n", xlab="Univariate Regression Coef.", ylab="Multiple Regression Coef.", xlim=c(-3, 32),
points(summary(fit1)$coef[2],summary(fit14)$coef[2])
points(summary(fit2)$coef[2],summary(fit14)$coef[3])
points(summary(fit3)$coef[2],summary(fit14)$coef[4])
points(summary(fit4)$coef[2],summary(fit14)$coef[5])
points(summary(fit5)$coef[2],summary(fit14)$coef[6])
points(summary(fit6)$coef[2],summary(fit14)$coef[7])
points(summary(fit7)$coef[2],summary(fit14)$coef[8])
points(summary(fit8)$coef[2],summary(fit14)$coef[9])
points(summary(fit9)$coef[2],summary(fit14)$coef[10])
points(summary(fit10)$coef[2],summary(fit14)$coef[11])
points(summary(fit11)$coef[2],summary(fit14)$coef[12])
```

```r
points(summary(fit12)$coef[2],summary(fit14)$coef[13])
points(summary(fit13)$coef[2],summary(fit14)$coef[14])
```



- Most of the coef. stayed relatively the same between the multivariate model and the univariate models, however most of the coef. in the multivariate model are not statistically significant from 0 and reside in the coef. range of -2,2 while it appears the coef. range in the univariate models were between -3,3. There is one outlier where the variable nox had a very large univariate coef., but in the multivariate model, was decreased 3 fold. The multivariate model should be used over the univariate models because it allows for controls in the crime rate.

(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.

```r
set.seed(1)

fitzn=lm(Boston$crim~Boston$zn+(Boston$zn2)+(Boston$zn3))
summary(fitzn)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$zn + (Boston$zn2) + (Boston$zn3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.821 -4.614 -1.294  0.473 84.130
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.846e+00  4.330e-01  11.192  < 2e-16 ***
## Boston$zn   -3.322e-01  1.098e-01  -3.025  0.00261 **
```

```
## Boston$zn2   6.483e-03  3.861e-03   1.679  0.09375 .
## Boston$zn3  -3.776e-05  3.139e-05  -1.203  0.22954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.372 on 502 degrees of freedom
## Multiple R-squared:  0.05824,	Adjusted R-squared:  0.05261
## F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06
```

- There is no evidence of a non-linear factor at the 5% significance level for the variable zn.

```r
set.seed(1)

fitindus=lm(Boston$crim~Boston$indus+(Boston$indus2)+(Boston$indus3))
summary(fitindus)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$indus + (Boston$indus2) + (Boston$indus3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.278 -2.514  0.054  0.764 79.713
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.6625683  1.5739833   2.327   0.0204 *
## Boston$indus  -1.9652129  0.4819901  -4.077 5.30e-05 ***
## Boston$indus2  0.2519373  0.0393221   6.407 3.42e-10 ***
## Boston$indus3 -0.0069760  0.0009567  -7.292 1.20e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.423 on 502 degrees of freedom
## Multiple R-squared:  0.2597,	Adjusted R-squared:  0.2552
## F-statistic: 58.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

- There is evidence of a non-linear factor at the 5% significance level at both the squared and cubed transformations.

```r
set.seed(1)

fitchas=lm(Boston$crim~Boston$chas+(Boston$chas2)+(Boston$chas3))
summary(fitchas)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$chas + (Boston$chas2) + (Boston$chas3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.7444     0.3961   9.453   <2e-16 ***
```

```
## Boston$chas   -1.8928     1.5061  -1.257    0.209
## Boston$chas2       NA         NA      NA       NA
## Boston$chas3       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.597 on 504 degrees of freedom
## Multiple R-squared:  0.003124,   Adjusted R-squared:  0.001146
## F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

- Dummy variables do not have non-linear factors alone.

```r
set.seed(1)

fitnox=lm(Boston$crim~Boston$nox+(Boston$nox2)+(Boston$nox3))
summary(fitnox)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$nox + (Boston$nox2) + (Boston$nox3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.110 -2.068 -0.255  0.739 78.302
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   233.09      33.64   6.928 1.31e-11 ***
## Boston$nox  -1279.37     170.40  -7.508 2.76e-13 ***
## Boston$nox2  2248.54     279.90   8.033 6.81e-15 ***
## Boston$nox3 -1245.70     149.28  -8.345 6.96e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.234 on 502 degrees of freedom
## Multiple R-squared:  0.297,  Adjusted R-squared:  0.2928
## F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

- There is evidence of a non-linear factor at the 5% significance level at both the squared and cubed transformations

```r
set.seed(1)

fitrm=lm(Boston$crim~Boston$rm+(Boston$rm2)+(Boston$rm3))
summary(fitrm)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$rm + (Boston$rm2) + (Boston$rm3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.485  -3.468  -2.221  -0.015  87.219
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 112.6246    64.5172   1.746   0.0815 .
## Boston$rm   -39.1501    31.3115  -1.250   0.2118
## Boston$rm2    4.5509     5.0099   0.908   0.3641
## Boston$rm3   -0.1745     0.2637  -0.662   0.5086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.33 on 502 degrees of freedom
## Multiple R-squared:  0.06779,    Adjusted R-squared:  0.06222
## F-statistic: 12.17 on 3 and 502 DF,  p-value: 1.067e-07
```

- There is no evidence of a non-linear factor at the 5% significance level.

```r
set.seed(1)

fitage=lm(Boston$crim~Boston$age+(Boston$age2)+(Boston$age3))
summary(fitage)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$age + (Boston$age2) + (Boston$age3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.762 -2.673 -0.516  0.019 82.842
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.549e+00  2.769e+00  -0.920  0.35780
## Boston$age   2.737e-01  1.864e-01   1.468  0.14266
## Boston$age2 -7.230e-03  3.637e-03  -1.988  0.04738 *
## Boston$age3  5.745e-05  2.109e-05   2.724  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.84 on 502 degrees of freedom
## Multiple R-squared:  0.1742, Adjusted R-squared:  0.1693
## F-statistic: 35.31 on 3 and 502 DF,  p-value: < 2.2e-16
```

- There is evidence of a non-linear factor at the 5% significance level at both the squared and cubed transformations, but interestingly the non-transformed variable lost its statistical significance.

```r
set.seed(1)

fitdis=lm(Boston$crim~Boston$dis+(Boston$dis2)+(Boston$dis3))
summary(fitdis)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$dis + (Boston$dis2) + (Boston$dis3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.757  -2.588   0.031   1.267  76.378
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.0476     2.4459  12.285  < 2e-16 ***
## Boston$dis  -15.5543     1.7360  -8.960  < 2e-16 ***
## Boston$dis2   2.4521     0.3464   7.078 4.94e-12 ***
## Boston$dis3  -0.1186     0.0204  -5.814 1.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.331 on 502 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2735
## F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16
```

- There is evidence of a non-linear factor at the 5% level at both the squared and cubed transformations.

```r
set.seed(1)

fitrad=lm(Boston$crim~Boston$rad+(Boston$rad2)+(Boston$rad3))
summary(fitrad)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$rad + (Boston$rad2) + (Boston$rad3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.381  -0.412  -0.269   0.179  76.217
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.605545   2.050108  -0.295    0.768
## Boston$rad   0.512736   1.043597   0.491    0.623
## Boston$rad2 -0.075177   0.148543  -0.506    0.613
## Boston$rad3  0.003209   0.004564   0.703    0.482
##
## Residual standard error: 6.682 on 502 degrees of freedom
## Multiple R-squared:    0.4,  Adjusted R-squared:  0.3965
## F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

- There is no evidence of a non-linear factor at the 5% level, nor is the original variable statistically significant at the 5% level.

```r
set.seed(1)

fittax=lm(Boston$crim~Boston$tax+(Boston$tax2)+(Boston$tax3))
summary(fittax)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$tax + (Boston$tax2) + (Boston$tax3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.273  -1.389   0.046   0.536  76.950
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.918e+01  1.180e+01   1.626    0.105
## Boston$tax  -1.533e-01  9.568e-02  -1.602    0.110
## Boston$tax2  3.608e-04  2.425e-04   1.488    0.137
## Boston$tax3 -2.204e-07  1.889e-07  -1.167    0.244
##
## Residual standard error: 6.854 on 502 degrees of freedom
## Multiple R-squared:  0.3689, Adjusted R-squared:  0.3651
## F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

- There is no evidence of a non-linear factor at the 5% level, nor is the original variable statistically significant at the 5% level.

```r
set.seed(1)

fitpt=lm(Boston$crim~Boston$ptratio+(Boston$ptratio2)+(Boston$ptratio3))
summary(fitpt)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$ptratio + (Boston$ptratio2) +
##     (Boston$ptratio3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.833 -4.146 -1.655  1.408 82.697
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     477.18405  156.79498   3.043  0.00246 **
## Boston$ptratio  -82.36054   27.64394  -2.979  0.00303 **
## Boston$ptratio2   4.63535    1.60832   2.882  0.00412 **
## Boston$ptratio3  -0.08476    0.03090  -2.743  0.00630 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.122 on 502 degrees of freedom
## Multiple R-squared:  0.1138, Adjusted R-squared:  0.1085
## F-statistic: 21.48 on 3 and 502 DF,  p-value: 4.171e-13
```

- There is evidence of a non-linear factor at the 5% level for both the squared and cube terms.

```r
set.seed(1)

fitblack=lm(Boston$crim~Boston$black+(Boston$black2)+(Boston$black3))
summary(fitblack)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$black + (Boston$black2) + (Boston$black3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.096  -2.343  -2.128  -1.439  86.790
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    1.826e+01  2.305e+00   7.924  1.5e-14 ***
## Boston$black  -8.356e-02  5.633e-02  -1.483    0.139
## Boston$black2  2.137e-04  2.984e-04   0.716    0.474
## Boston$black3 -2.652e-07  4.364e-07  -0.608    0.544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.955 on 502 degrees of freedom
## Multiple R-squared:  0.1498, Adjusted R-squared:  0.1448
## F-statistic: 29.49 on 3 and 502 DF,  p-value: < 2.2e-16
```

- There is no evidence that there is a non-linear factor at the 5% level, nor is the original variable statistically significant.

```r
set.seed(1)

fitls=lm(Boston$crim~Boston$lstat+(Boston$lstat2)+(Boston$lstat3))
summary(fitls)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$lstat + (Boston$lstat2) + (Boston$lstat3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.234  -2.151  -0.486   0.066  83.353
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.2009656  2.0286452   0.592   0.5541
## Boston$lstat  -0.4490656  0.4648911  -0.966   0.3345
## Boston$lstat2  0.0557794  0.0301156   1.852   0.0646 .
## Boston$lstat3 -0.0008574  0.0005652  -1.517   0.1299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.629 on 502 degrees of freedom
## Multiple R-squared:  0.2179, Adjusted R-squared:  0.2133
## F-statistic: 46.63 on 3 and 502 DF,  p-value: < 2.2e-16
```

- There is no evidence of a non-linear factor at the 5% level for both the squared and cubed terms, nor is the original variable statistically significant.

```r
set.seed(1)

fitme=lm(Boston$crim~Boston$medv+(Boston$medv2)+(Boston$medv3))
summary(fitme)
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$medv + (Boston$medv2) + (Boston$medv3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.427  -1.976  -0.437   0.439  73.655
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.1655381  3.3563105  15.840  < 2e-16 ***
## Boston$medv  -5.0948305  0.4338321 -11.744  < 2e-16 ***
## Boston$medv2  0.1554965  0.0171904   9.046  < 2e-16 ***
## Boston$medv3 -0.0014901  0.0002038  -7.312 1.05e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.569 on 502 degrees of freedom
## Multiple R-squared:  0.4202, Adjusted R-squared:  0.4167
## F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
remove(Boston)
rm(list=ls())
```

- There is evidence of a non-linear factor for both the cubed and squared terms at the 5% significance level.

Question 9 From Chapter 6

```r
set.seed(1)

library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.3.2
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-10
```

```r
library(ISLR)
attach(College)
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 3.3.2
```

```
##
## Attaching package: 'Metrics'
```

```
## The following object is masked from 'package:glmnet':
##
##     auc
```

```r
college=College
```

(a)

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

```r
set.seed(1)

index     <- 1:nrow(college)
testindex <- sample(index, trunc(length(index)/2))
```

```
testset   <- college[testindex,]
trainset  <- college[-testindex,]
```

(b)

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

```
set.seed(1)

fit1=lm(Apps~.,data=trainset)
pred_fit1=predict(fit1,testset)
mean((pred_fit1-testset$Apps)^2)
```

```
## [1] 1354497
```

```
sqrt(mean((pred_fit1-testset$Apps)^2))
```

```
## [1] 1163.829
```

- The test MSE for the model is 1354497 and the test RMSE is 1163.829

(c)

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

```
set.seed(1)

library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.3.2
```

```
x=model.matrix(Apps~.,college)[,-1]
y=college$Apps
train=sample(1:nrow(x), nrow(x)/2)
test=(-train)
y.test=y[test]
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x[train,],y[train],alpha=0,lambda=grid,thresh=1e-12)
cv.ridge.mod=cv.glmnet(x[train,],y[train],alpha=0,lambda=grid,thresh=1e-12)
ridge.pred=predict(ridge.mod ,s=4, newx=x[test,])
mean((ridge.pred-y.test)^2)
```

```
## [1] 1103208
```

```
sqrt(mean((ridge.pred-y.test)^2))
```

```
## [1] 1050.337
```

```
bestlam=cv.ridge.mod$lambda.min
bestlam
```

```
## [1] 0.01321941
```

- The test error MSE was 1103208 and a test RMSE of 1050.337, and the best lamba was .0132

(d)

Fit a lasso model on the training set, with lamdba chosen by cross- validation. Report the test error obtained, along with the num-ber of non-zero coefficient estimates.

```
set.seed(1)


lasso.mod=glmnet(x[train,],y[train],alpha=1,lambda=grid,thresh=1e-12)
cv.lasso.mod=cv.glmnet(x[train,],y[train],alpha=0,lambda=grid,thresh=1e-12)

lasso.pred=predict(lasso.mod ,s=4, newx=x[test,])
mean((lasso.pred-y.test)^2)
```

```
## [1] 1090101
```

```
sqrt(mean((lasso.pred-y.test)^2))
```

```
## [1] 1044.079
```

```
bestlam=cv.lasso.mod$lambda.min
bestlam
```

```
## [1] 0.01
```

```
lasso.coef=predict(lasso.mod,type="coefficients",s=bestlam)
lasso.coef
```

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                        1
## (Intercept)   77.95965673
## PrivateYes  -757.12240527
## Accept         1.67976412
## Enroll        -0.62381617
## Top10perc     67.44884886
## Top25perc    -22.36946351
## F.Undergrad   -0.06121978
## P.Undergrad    0.04741724
## Outstate      -0.09225906
## Room.Board     0.24510403
## Books          0.09083099
## Personal       0.05883457
## PhD           -8.88831388
## Terminal      -1.72024559
## S.F.Ratio     -5.74868051
## perc.alumni   -1.46717637
## Expend         0.03486892
## Grad.Rate      7.57372582
```

- The test error MSE was 1,09010 and a test RMSE of 1044.079 and the best lamba was .01, the number of non-zero coefficients estimates was 17 not including the intercept.

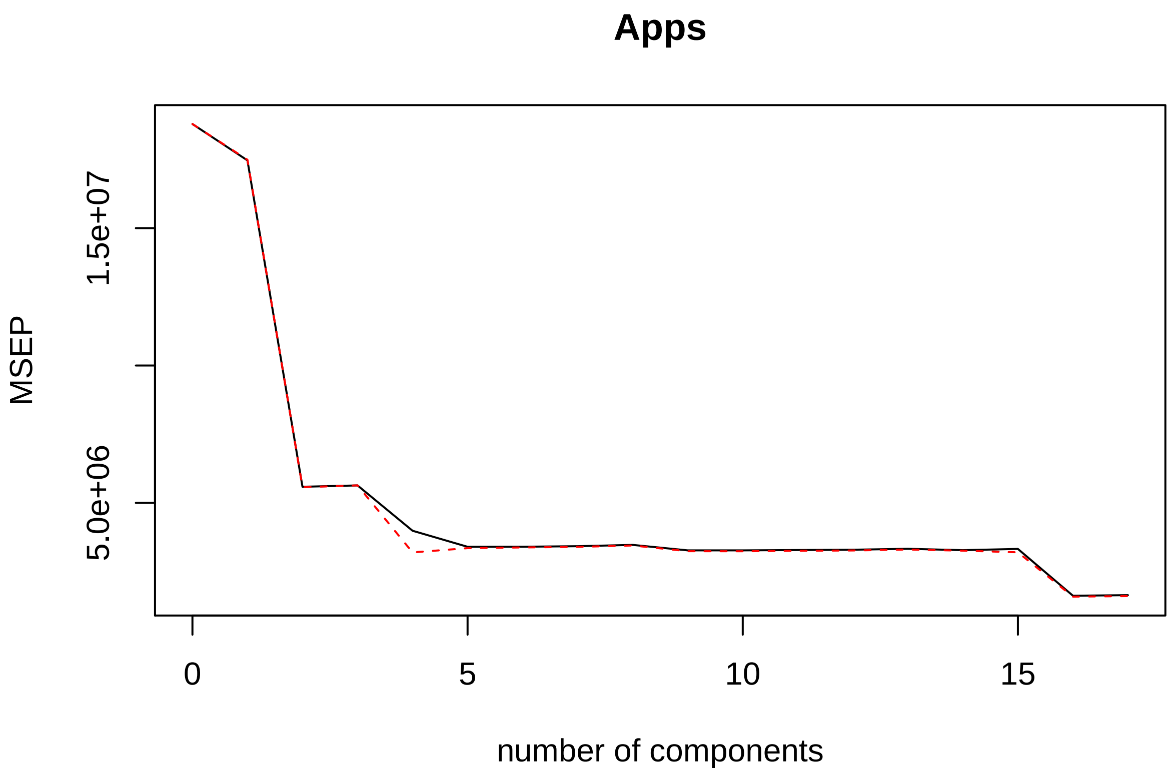
(e)

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

```r
set.seed(1)

library("pls")
```

```
## Warning: package 'pls' was built under R version 3.3.2
##
## Attaching package: 'pls'
## The following object is masked from 'package:stats':
##
##     loadings
```

```r
pcr.fit=pcr(Apps~., data=college, subset=train, scale=TRUE, validation ="CV")


validationplot(pcr.fit,val.type="MSEP")
```




```r
summary(pcr.fit)
```

```
## Data:    X dimension: 388 17
##  Y dimension: 388 1
## Fit method: svdpc
## 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            4335     4179     2364     2374     1996     1844     1845
## adjCV         4335     4182     2360     2374     1788     1831     1838
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
```

```
## CV         1850      1863      1809       1809       1812       1815       1825
## adjCV      1844      1857      1801       1800       1804       1808       1817
##        14 comps  15 comps  16 comps  17 comps
## CV          1810      1823      1273      1281
## adjCV       1806      1789      1260      1268
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X      31.216    57.68    64.73    70.55    76.33    81.30    85.01
## Apps    6.976    71.47    71.58    83.32    83.44    83.45    83.46
##       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X       88.40    91.16     93.36     95.38     96.94     97.96     98.76
## Apps    83.47    84.53     84.86     84.98     84.98     84.99     85.24
##       15 comps  16 comps  17 comps
## X        99.40     99.87    100.00
## Apps     90.87     93.93     93.97
```

```
pcr.pred=predict(pcr.fit,x[test,],ncomp=17)
mean((pcr.pred-y.test)^2)
```

```
## [1] 1108531
```

```
sqrt(mean((pcr.pred-y.test)^2))
```

```
## [1] 1052.868
```

- The Test error was 1,108,531 and a test RMSE of 1052.868, and the value of M that minimized the cross validation error was using all of the variables M=17

(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.

```
set.seed(1)


pls.fit=plsr(Apps~.,data=college,subset=train,scale=TRUE,validation="CV")
summary(pls.fit)
```

```
## Data:    X dimension: 388 17
##  Y dimension: 388 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            4335     2176     1893     1725     1613     1406     1312
## adjCV         4335     2171     1884     1715     1578     1375     1295
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1297     1285     1280      1278      1279      1282      1281
## adjCV     1281     1271     1267      1265      1266      1269      1268
##        14 comps  15 comps  16 comps  17 comps
## CV         1281      1281      1281      1281
```

```
## adjCV      1267      1267      1268      1268
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X       26.91    43.08    63.26    65.16    68.50    73.75    76.10
## Apps    76.64    83.93    87.14    91.90    93.49    93.85    93.91
##       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X       79.03    81.76     85.41     89.03     91.38     93.31     95.43
## Apps    93.94    93.96     93.96     93.96     93.97     93.97     93.97
##       15 comps  16 comps  17 comps
## X        97.41     98.78    100.00
## Apps     93.97     93.97     93.97
```

```r
pls.pred=predict(pls.fit,x[test,],ncomp=16)
mean((pls.pred-y.test)^2)
```

```
## [1] 1108502
```

```r
sqrt(mean((pls.pred-y.test)^2))
```

```
## [1] 1052.854
```

- The test error was 1,108,502 and a test RMSE of 1052.854, and the value of M that minimized the cross validation error was the full model M=17

(g)

Comment on the results obtained. How accurately can we pre- dict the number of college applications received? Is there much difference among the test errors resulting from these five ap- proaches?

```r
set.seed(1)

sqrt(mean((pred_fit1-testset$Apps)^2))
```

```
## [1] 1163.829
```

```r
sqrt(mean((ridge.pred-y.test)^2))
```

```
## [1] 1050.337
```

```r
sqrt(mean((lasso.pred-y.test)^2))
```

```
## [1] 1044.079
```

```r
sqrt(mean((pcr.pred-y.test)^2))
```

```
## [1] 1052.868
```

```r
sqrt(mean((pls.pred-y.test)^2))
```

```
## [1] 1052.854
```

```r
summary(Apps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      81     776    1558    3002    3624   48090
```

```
rm(list=ls())
```

- All the models predict with similar amounts of accuracy according to their respective test RMSE's. The only model that deviates from the other models is the linear regression using all of the predictors. The models do not do very well compared to the summary statistics above. The only universities that would be able to use this model would possible be the ones in fourth quartile.

```
set.seed(1)

library(MASS)
attach(Boston)
```

```
## The following objects are masked from Boston (pos = 11):
##
##     age, black, chas, crim, dis, indus, lstat, medv, nox, ptratio,
##     rad, rm, tax, zn
```

```
library(glmnet)
Boston=Boston
```

Question 11 From Chapter 6

(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.

```
set.seed(1)
index     <- 1:nrow(Boston)
testindex <- sample(index, trunc(length(index)/2))
testset   <- Boston[testindex,]
trainset  <- Boston[-testindex,]
```

```
set.seed(1)
library(leaps)
regfit.full=regsubsets(crim~.,Boston,nvmax = 13)
summary(regfit.full)
```

```
## Subset selection object
## Call: regsubsets.formula(crim ~ ., Boston, nvmax = 13)
## 13 Variables  (and intercept)
##         Forced in Forced out
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas        FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## black       FALSE      FALSE
## lstat       FALSE      FALSE
```

```
## medv         FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: exhaustive
##           zn  indus chas nox rm  age dis rad tax ptratio black lstat medv
## 1  ( 1 )  " " " "   " "  " " " " " " " " "*" " " " "     " "   " "   " " 
## 2  ( 1 )  " " " "   " "  " " " " " " " " "*" " " " "     " "   "*"   " " 
## 3  ( 1 )  " " " "   " "  " " " " " " " " "*" " " " "     "*"   "*"   " " 
## 4  ( 1 )  "*" " "   " "  " " " " " " "*" "*" " " " "     " "   " "   "*" 
## 5  ( 1 )  "*" " "   " "  " " " " " " "*" "*" " " " "     "*"   " "   "*" 
## 6  ( 1 )  "*" " "   " "  "*" " " " " "*" "*" " " " "     "*"   " "   "*" 
## 7  ( 1 )  "*" " "   " "  "*" " " " " "*" "*" " " "*"     "*"   " "   "*" 
## 8  ( 1 )  "*" " "   " "  "*" " " " " "*" "*" " " "*"     "*"   "*"   "*" 
## 9  ( 1 )  "*" "*"   " "  "*" " " " " "*" "*" " " "*"     "*"   "*"   "*" 
## 10  ( 1 ) "*" "*"   " "  "*" "*" " " "*" "*" " " "*"     "*"   "*"   "*" 
## 11  ( 1 ) "*" "*"   " "  "*" "*" " " "*" "*" "*" "*"     "*"   "*"   "*" 
## 12  ( 1 ) "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"   "*" 
## 13  ( 1 ) "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"   "*"
```

```
reg.summary=summary(regfit.full)
reg.summary$rsq
```

```
##  [1] 0.3912567 0.4207965 0.4286123 0.4334892 0.4392738 0.4440173 0.4476594
##  [8] 0.4504606 0.4524408 0.4530572 0.4535605 0.4540031 0.4540104
```

```
par(mfrow=c(2,2))
plot(reg.summary$rss ,xlab="Number of Variables ",ylab="RSS",
type="l")
plot(reg.summary$adjr2 ,xlab="Number of Variables ",
ylab="Adjusted RSq",type="l")
which.max(reg.summary$adjr2)
```

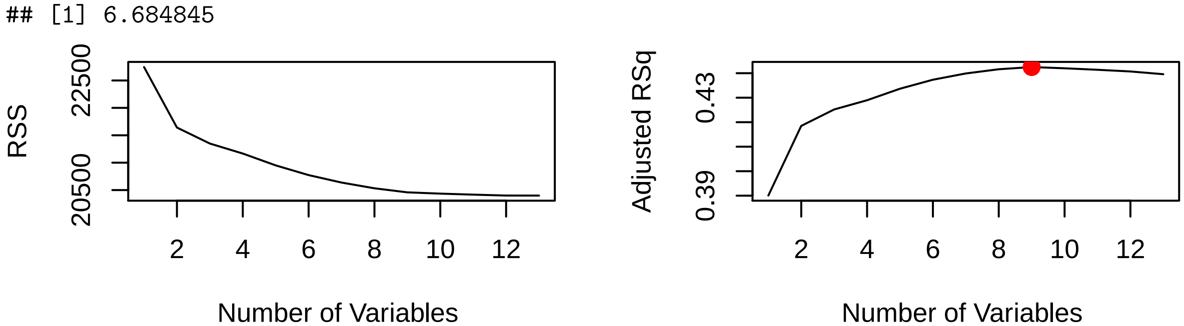
```
## [1] 9
```

```
points(9,reg.summary$adjr2[9], col="red",cex=2,pch=20)

fit1=lm(crim~zn+indus+nox+age+dis+rad+ptratio+black+lstat+medv,data=trainset)

pred_fit1=predict(fit1,testset)
mean((pred_fit1-testset$crim)^2)
```

```
## [1] 44.68716
```

```
sqrt(mean((pred_fit1-testset$crim)^2))
```

```
## [1] 6.684845
```



- Using the best subset selection and determining that nine variables minimized the RSS. The test RMSE of those nine variables in a linear regression was 6.684.

```r
set.seed(1)
library(leaps)
x=model.matrix(crim~.,Boston)[,-1]
y=Boston$crim
train=sample(1:nrow(x), nrow(x)/2)
test=(-train)
y.test=y[test]
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x[train,],y[train],alpha=0,lambda=grid,thresh=1e-12)
cv.ridge.mod=cv.glmnet(x[train,],y[train],alpha=0,lambda=grid,thresh=1e-12)
ridge.pred=predict(ridge.mod ,s=4, newx=x[test,])
mean((ridge.pred-y.test)^2)
```

```
## [1] 39.71717
```

```r
sqrt(mean((ridge.pred-y.test)^2))
```

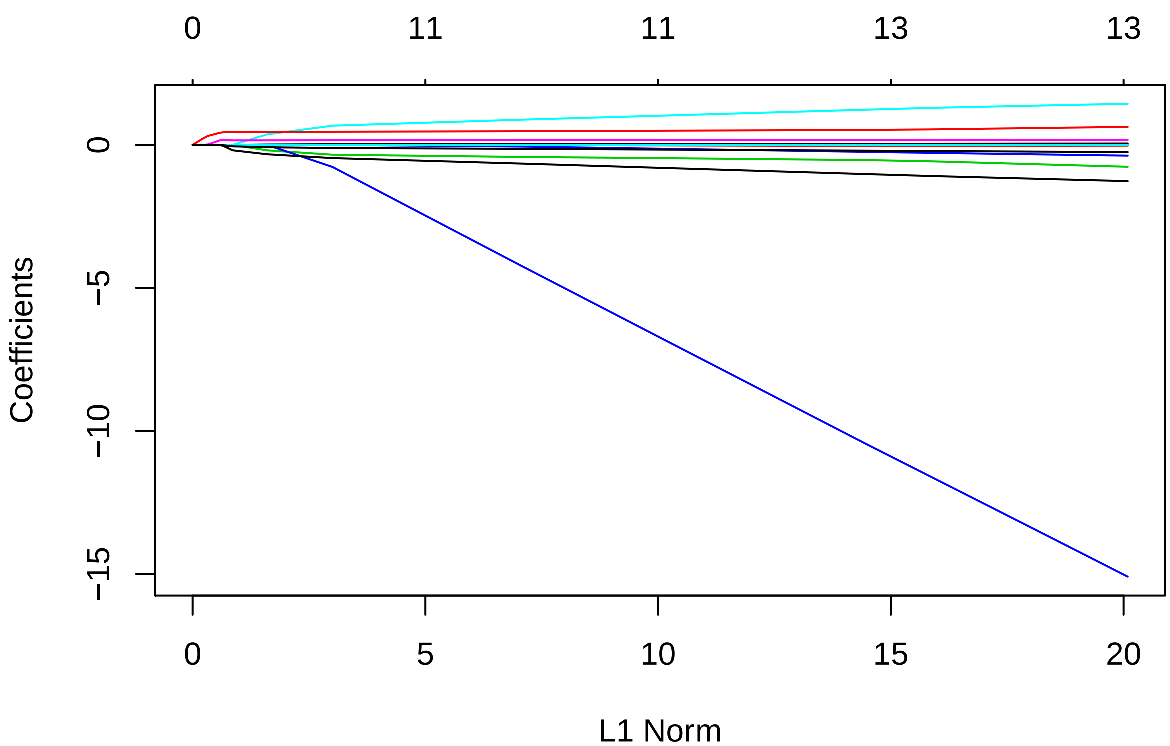
```
## [1] 6.302156
```

```r
bestlam=cv.ridge.mod$lambda.min
bestlam
```

```
## [1] 0.4977024
```

- After trying a ridge regression with a best lambda of .497. The Test RMSE is 6.3 which is significantly better than using the best subset selection/linear regression.

```r
set.seed(1)
lasso.mod=glmnet(x[train,],y[train],alpha=1,lambda=grid)
plot(lasso.mod)
```




```r
cv.lasso.mod=cv.glmnet(x[train,],y[train],alpha=1,lambda=grid)
```

```r
bestlam=cv.lasso.mod$lambda.min
bestlam
```

```
## [1] 0.09326033
```

```r
lasso.pred=predict(lasso.mod,s=bestlam,newx=x[test,])

lasso.coef=predict(lasso.mod,type="coefficients",s=bestlam)[1:14,]
lasso.coef
```

```
##  (Intercept)           zn        indus         chas          nox
##  6.426849545  0.036274165 -0.030379671 -0.504093148 -8.954417394
##           rm          age          dis          rad          tax
##  1.145825464  0.000000000 -0.932738275  0.511842261  0.000000000
##      ptratio        black        lstat         medv
## -0.199416479 -0.002437602  0.174973017 -0.186376367
```

```r
lasso_lm_fit=lm(crim~zn+indus+chas+nox+rm+dis+rad+ptratio+black+lstat+medv ,data=trainset)

lass_lm_pred=predict(lasso_lm_fit,data=testset)

sqrt(mean((lass_lm_pred-y.test)^2))
```

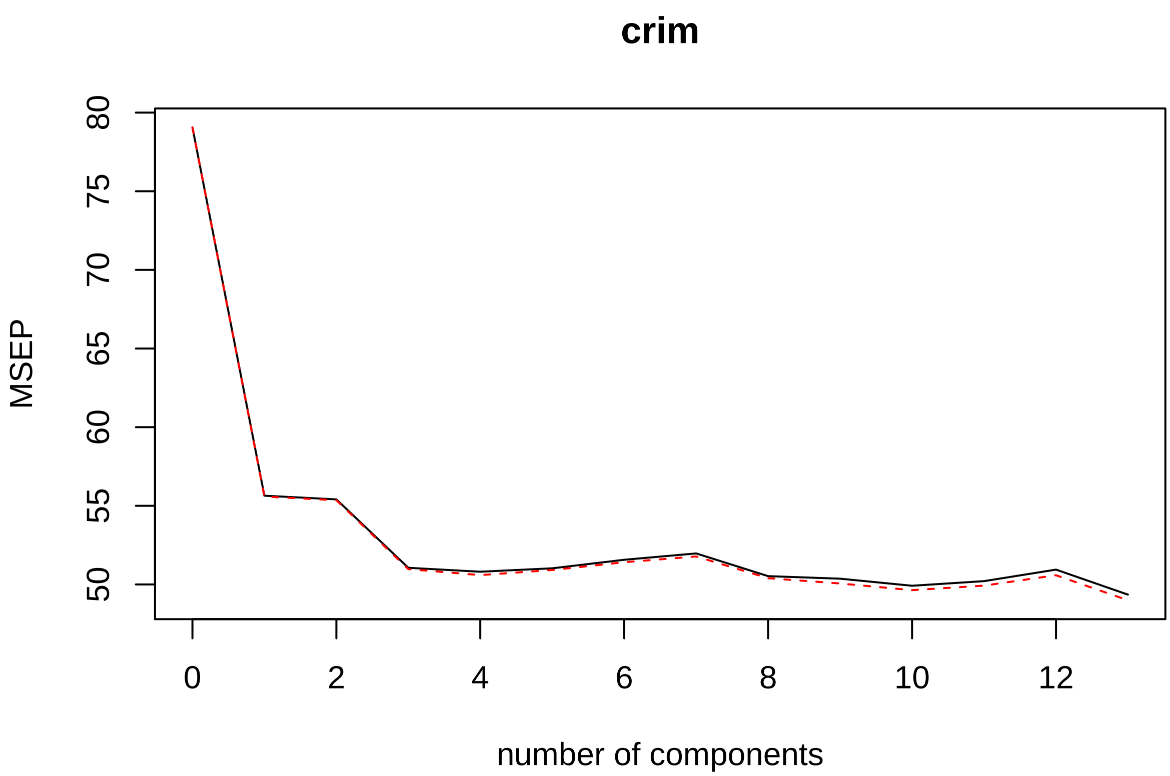
```
## [1] 6.043508
```

```r
sqrt(mean((lasso.pred-y.test)^2))
```

```
## [1] 6.19128
```

- After trying a lasso regression with a best lambda of .093 , and seeing the coef. age and tax going to 0, the test RMSE of a linear model, using the lasso as parameter selection, was 6.04, and using the predict function within the lasso model gave a test RMSE of 6.19 which is fairly similar, but the linear model should be used. This model is better than the ridge regression.

```r
set.seed(1)
library("pls")
pcr.fit=pcr(crim~., data=Boston, subset=train, scale=TRUE, validation ="CV")


validationplot(pcr.fit,val.type="MSEP")
```




```r
summary(pcr.fit)
```

```
## Data:    X dimension: 253 13
##  Y dimension: 253 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           8.892    7.459    7.444    7.146    7.128    7.143    7.181
## adjCV        8.892    7.456    7.440    7.140    7.113    7.136    7.170
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       7.209    7.108    7.097     7.065     7.086     7.137     7.025
## adjCV    7.196    7.099    7.075     7.045     7.066     7.112     7.000
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X       49.04    60.72    69.75    76.49    83.02    88.40    91.73
## crim    30.39    30.93    36.63    37.31    37.35    37.98    38.85
##       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X       93.77    95.73     97.36     98.62     99.57    100.00
## crim    39.94    41.89     42.73     42.73     43.55     45.48
```

```r
pcr.pred=predict(pcr.fit,x[test,],ncomp=12)
mean((pcr.pred-y.test)^2)
```

```
## [1] 39.40753
```

```r
sqrt(mean((pcr.pred-y.test)^2))
```

```
## [1] 6.277542
```

```r
rm(list=ls())
detach(Boston)
```

- Using PCR, and using 12 Principle components resulted in a test RMSE of 6.27, which is on par with the ridge model but greater than the lasso parameter selection model.

(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.

- I propose to use the Lasso model as parameter selection and then a linear model for predicting crime rates in Boston suburbs. The Lasso provide the smallest test RMSE, therefore the model should have the strongest predictive power.

(c)

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

- The chosen model uses all the variables in the model expect for age and tax. The Lasso regression shrinks the predictor space by a value lambda which is chosen through cross validation, but only two of the predictors reach 0 so all predictors except for age and tax were used in linear model.

```r
library(tree)
library(ISLR)
attach(Carseats)
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
Carseats=Carseats
set.seed(1)
```

Question 8 From Chapter 8

(a)

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

```r
set.seed(1)
train = sample(1:nrow(Carseats), nrow(Carseats)/2)
test=Carseats[-train ,"Sales"]
```

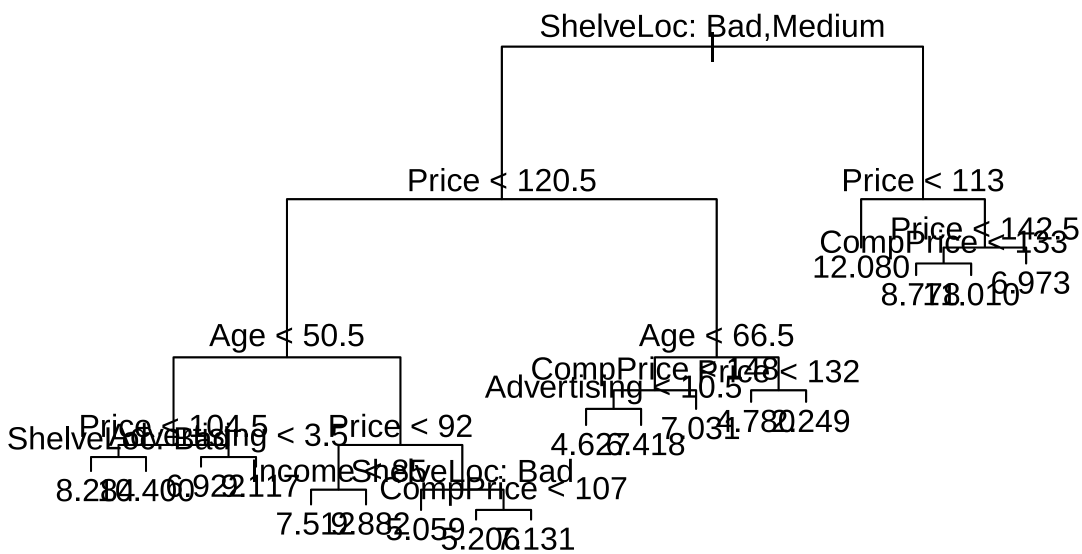
(b)

Fit a regression tree to the training set. Plot the tree, and inter- pret the results. What test error rate do you obtain?

```
set.seed(1)
tree.Carseats=tree(Sales~.,Carseats ,subset=train)
summary(tree.Carseats)
```

```
##
## Regression tree:
## tree(formula = Sales ~ ., data = Carseats, subset = train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "Advertising" "Income"
## [6] "CompPrice"
## Number of terminal nodes:  18
## Residual mean deviance:  2.36 = 429.5 / 182
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.2570 -1.0360  0.1024  0.0000  0.9301  3.9130
```

```
plot(tree.Carseats)
text(tree.Carseats ,pretty=0)
```




```
yhat=predict(tree.Carseats ,newdata=Carseats[-train ,])
mean((yhat-test)^2)
```

```
## [1] 4.148897
```

```
sqrt(mean((yhat-test)^2))
```

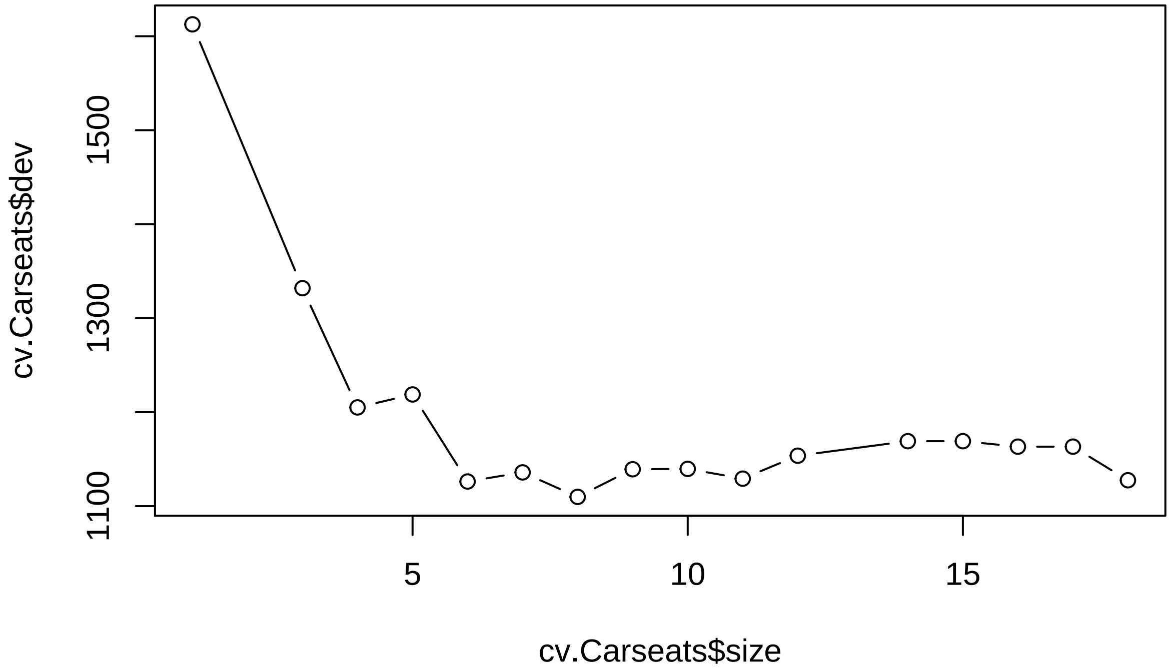
```
## [1] 2.036884
```

- The mean squared error test rate is 4.148, and the RMSE is 2.036, which is fairly decent. The variable with the most information for sales is shelvloc: bad/medium which is shown from the plot of the tree.

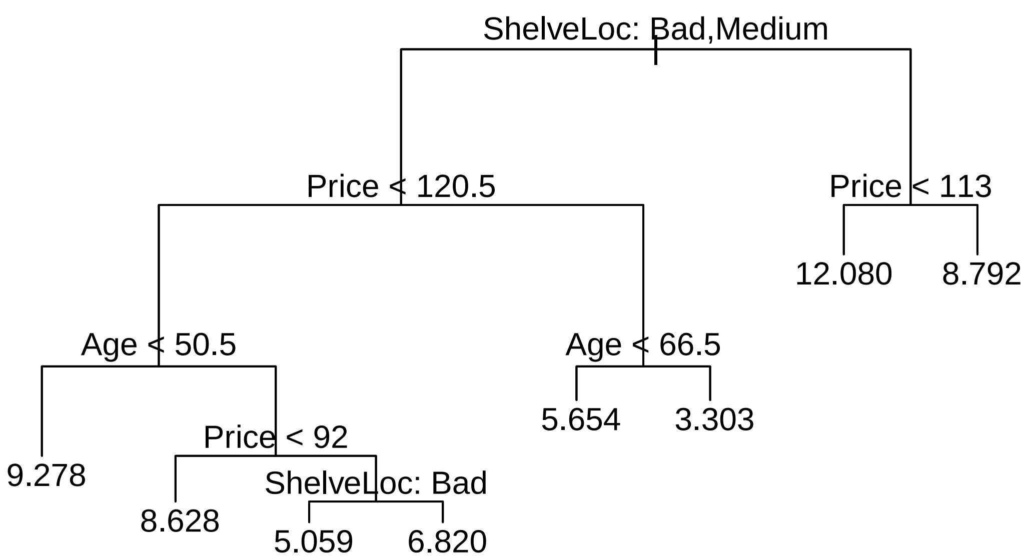
(c)

Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test error rate?

```r
set.seed(1)

cv.Carseats=cv.tree(tree.Carseats,FUN=prune.tree )
plot(cv.Carseats$size ,cv.Carseats$dev ,type='b')
```




```r
prune.Carseats=prune.tree(tree.Carseats ,best=8)
plot(prune.Carseats)
text(prune.Carseats ,pretty=0)
```




```r
yhat=predict(prune.Carseats ,newdata=Carseats[-train ,])
mean((yhat-test)^2)
```

```
## [1] 5.09085
```

```r
sqrt(mean((yhat-test)^2))
```

```
## [1] 2.256291
```

- The test RMSE is 2.256, which actually increases marginally, but the tree is much more interpret-able with only 8 terminal nodes.

(d)

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

```r
set.seed(1)

bag.car=randomForest(Sales~.,data=Carseats,subset=train,mtry=10,importance =TRUE)
bag.car
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = Carseats, mtry = 10,      importance = TRUE, subset = trai
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 2.825834
##                     % Var explained: 62.98
```

```r
yhat=predict(bag.car ,newdata=Carseats[-train ,])
mean((yhat-test)^2)
```

```
## [1] 2.554292
```

```r
sqrt(mean((yhat-test)^2))
```

```
## [1] 1.598215
```

```r
importance(bag.car)
```

```
##               %IncMSE IncNodePurity
## CompPrice   14.032030    129.568747
## Income       5.523038     75.448682
## Advertising 13.571285    131.246840
## Population   1.968853     63.042648
## Price       56.863812    504.158108
## ShelveLoc   44.720455    323.055042
## Age         22.225468    194.915976
## Education    4.823966     40.810991
## Urban       -1.902185      8.746566
## US           6.632887     14.599565
```

- The test RMSE is 1.59 which is substantially better than the previous tree models. The top 3 most important variables are shelveloc, price, and age.

(e)

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

```
set.seed(1)

bag.car=randomForest(Sales~.,data=Carseats,subset=train,mtry=3,importance =TRUE)
bag.car
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = Carseats, mtry = 3,      importance = TRUE, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 3.26604
##                     % Var explained: 57.21
```

```
yhat=predict(bag.car ,newdata=Carseats[-train ,])
mean((yhat-test)^2)
```

```
## [1] 3.30763
```

```
sqrt(mean((yhat-test)^2))
```

```
## [1] 1.818689
```

```
bag.car=randomForest(Sales~.,data=Carseats,subset=train,mtry=5,importance =TRUE)
bag.car
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = Carseats, mtry = 5,      importance = TRUE, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##           Mean of squared residuals: 2.940785
##                     % Var explained: 61.47
```

```
yhat=predict(bag.car ,newdata=Carseats[-train ,])
mean((yhat-test)^2)
```

```
## [1] 2.814854
```

```
sqrt(mean((yhat-test)^2))
```

```
## [1] 1.677753
```

```
bag.car=randomForest(Sales~.,data=Carseats,subset=train,mtry=7,importance =TRUE)
bag.car
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = Carseats, mtry = 7,      importance = TRUE, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 7
```

```
##
##           Mean of squared residuals: 2.881145
##                     % Var explained: 62.25
```

```
yhat=predict(bag.car ,newdata=Carseats[-train ,])
mean((yhat-test)^2)
```

```
## [1] 2.676364
```

```
sqrt(mean((yhat-test)^2))
```

```
## [1] 1.63596
```

```
bag.car=randomForest(Sales~.,data=Carseats,subset=train,mtry=9,importance =TRUE)
bag.car
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = Carseats, mtry = 9,      importance = TRUE, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 9
##
##           Mean of squared residuals: 2.894683
##                     % Var explained: 62.07
```

```
yhat=predict(bag.car ,newdata=Carseats[-train ,])
mean((yhat-test)^2)
```

```
## [1] 2.592225
```

```
sqrt(mean((yhat-test)^2))
```

```
## [1] 1.610039
```

```
importance(bag.car)
```

```
##                %IncMSE IncNodePurity
## CompPrice   13.8613559    130.399757
## Income       5.0714754     78.546932
## Advertising 15.8484911    129.797525
## Population  -0.5814597     62.100670
## Price       52.8140381    506.176451
## ShelveLoc   44.6158853    312.876187
## Age         22.9423123    193.642513
## Education    3.6120385     42.302945
## Urban       -2.5805163      8.201527
## US           6.4404356     14.613714
```

```
detach(Carseats)
rm(list=ls())
```

- As the number of variables in the random forest model increases(mtry) at each split, the lower the test RMSE gets, which indicates that this data needs more complex trees for predictive power. The Random forest is marginally better than the bagging model in the previous question which used all 10 variables. The variables which were most important were shelveloc, price, and age which were the same variables in the previous bagging problem.

```r
library(ISLR)
Caravan=Caravan
Caravan$Purchase <- ifelse(Caravan$Purchase == "Yes", 1, 0)
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 3.3.2
## Loading required package: survival
## Warning: package 'survival' was built under R version 3.3.2
## Loading required package: lattice
## Loading required package: splines
## Loading required package: parallel
## Loaded gbm 2.1.3
```

```r
attach(Caravan)
set.seed(1)
```

Question 11 From Chapter 8

(a)

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

```r
set.seed(1)

train = sample(1:nrow(Caravan), 1000)
train.set=Caravan[train,]
test=Caravan[-train, ]
```

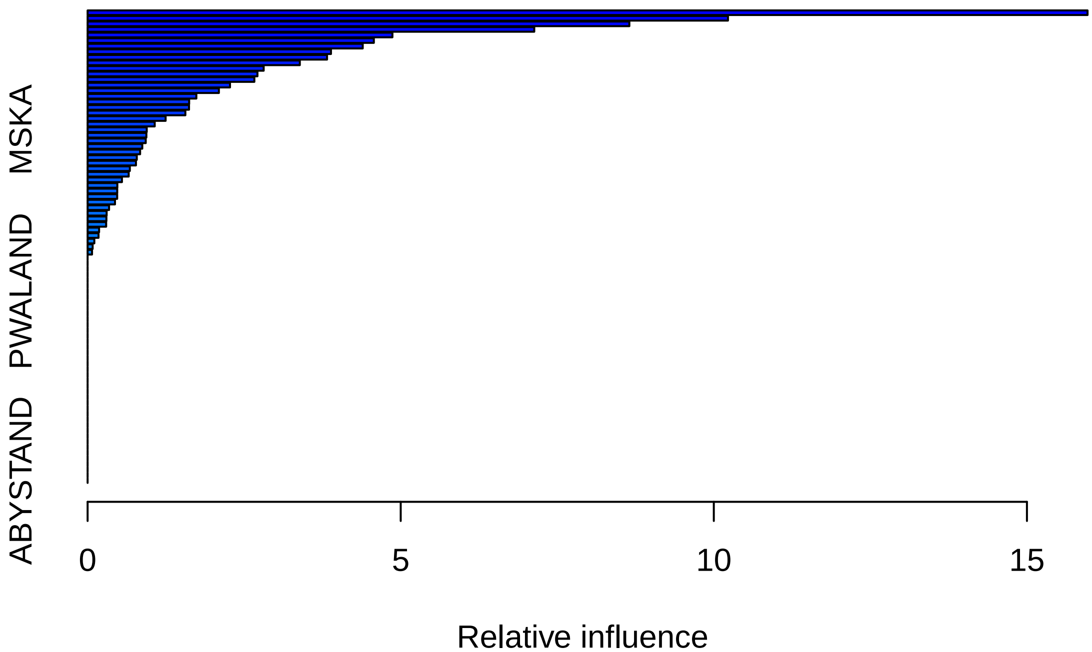
(b)

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?

```r
set.seed(1)

boost.caravan=gbm(Purchase~.,n.trees=1000,distribution="gaussian",data=train.set,shrinkage = .01)
```

```
## Warning in gbm.fit(x, y, offset = offset, distribution = distribution, w =
## w, : variable 50: PVRAAUT has no variation.

## Warning in gbm.fit(x, y, offset = offset, distribution = distribution, w =
## w, : variable 60: PZEILPL has no variation.

## Warning in gbm.fit(x, y, offset = offset, distribution = distribution, w =
## w, : variable 71: AVRAAUT has no variation.

## Warning in gbm.fit(x, y, offset = offset, distribution = distribution, w =
## w, : variable 81: AZEILPL has no variation.
```

```r
summary(boost.caravan)
```




```
##                 var     rel.inf
## APERSAUT APERSAUT 15.96810420
## MOSTYPE   MOSTYPE 10.22612374
## PPERSAUT PPERSAUT  8.65124494
## MBERMIDD MBERMIDD  7.13258302
## MINKGEM   MINKGEM  4.86733635
## MRELGE     MRELGE  4.57093132
## PWAPART   PWAPART  4.39364327
## MINK7512 MINK7512  3.88622571
## PBRAND     PBRAND  3.82357715
## MOPLHOOG MOPLHOOG  3.38870380
## MFGEKIND MFGEKIND  2.81196175
## MGODOV     MGODOV  2.71239150
## MINK3045 MINK3045  2.66280111
## MOPLMIDD MOPLMIDD  2.27415799
## MOSHOOFD MOSHOOFD  2.09631331
## MBERARBO MBERARBO  1.73682059
## MSKC         MSKC  1.62149415
## MBERZELF MBERZELF  1.61942846
## ALEVEN     ALEVEN  1.56097805
## MBERARBG MBERARBG  1.24607984
## MINK4575 MINK4575  1.07198447
## MSKA         MSKA  0.94435636
## MGODRK     MGODRK  0.93854816
## MSKB1       MSKB1  0.92831203
## MGODGE     MGODGE  0.87279432
## MOPLLAAG MOPLLAAG  0.83847068
## MGEMLEEF MGEMLEEF  0.78587272
## MAUT1       MAUT1  0.77198256
## MAUT0       MAUT0  0.67443431
## MFALLEEN MFALLEEN  0.65610416
```

```
## MHKOOP      MHKOOP  0.54845026
## MZFONDS    MZFONDS  0.47499474
## MKOOPKLA MKOOPKLA  0.47313781
## MGODPR      MGODPR  0.47228477
## MINKM30    MINKM30  0.43724266
## PLEVEN      PLEVEN  0.34280871
## MFWEKIND MFWEKIND  0.30400581
## MSKB2        MSKB2  0.30024351
## MZPART      MZPART  0.29646397
## MHHUUR      MHHUUR  0.18349512
## MRELOV      MRELOV  0.17466922
## MBERHOOG MBERHOOG  0.10724638
## MGEMOMV    MGEMOMV  0.08034933
## MINK123M MINK123M  0.07084767
## MAANTHUI MAANTHUI  0.00000000
## MRELSA      MRELSA  0.00000000
## MBERBOER MBERBOER  0.00000000
## MSKD          MSKD  0.00000000
## MAUT2        MAUT2  0.00000000
## PWABEDR    PWABEDR  0.00000000
## PWALAND    PWALAND  0.00000000
## PBESAUT    PBESAUT  0.00000000
## PMOTSCO    PMOTSCO  0.00000000
## PVRAAUT    PVRAAUT  0.00000000
## PAANHANG PAANHANG  0.00000000
## PTRACTOR PTRACTOR  0.00000000
## PWERKT      PWERKT  0.00000000
## PBROM        PBROM  0.00000000
## PPERSONG PPERSONG  0.00000000
## PGEZONG    PGEZONG  0.00000000
## PWAOREG    PWAOREG  0.00000000
## PZEILPL    PZEILPL  0.00000000
## PPLEZIER PPLEZIER  0.00000000
## PFIETS      PFIETS  0.00000000
## PINBOED    PINBOED  0.00000000
## PBYSTAND PBYSTAND  0.00000000
## AWAPART    AWAPART  0.00000000
## AWABEDR    AWABEDR  0.00000000
## AWALAND    AWALAND  0.00000000
## ABESAUT    ABESAUT  0.00000000
## AMOTSCO    AMOTSCO  0.00000000
## AVRAAUT    AVRAAUT  0.00000000
## AAANHANG AAANHANG  0.00000000
## ATRACTOR ATRACTOR  0.00000000
## AWERKT      AWERKT  0.00000000
## ABROM        ABROM  0.00000000
## APERSONG APERSONG  0.00000000
## AGEZONG    AGEZONG  0.00000000
## AWAOREG    AWAOREG  0.00000000
## ABRAND      ABRAND  0.00000000
## AZEILPL    AZEILPL  0.00000000
## APLEZIER APLEZIER  0.00000000
## AFIETS      AFIETS  0.00000000
## AINBOED    AINBOED  0.00000000
```

```
## ABYSTAND ABYSTAND  0.00000000
```

- The top 3 most important variables are MOSTYPE,APERSAUT, and PPERSAUT

(c)

Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated prob-ability of purchase is greater than 20 %. Form a confusion ma-trix. 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?

```r
set.seed(1)

yhat.boost=predict(boost.caravan,newdata=test, n.trees=1000)


glm.pred=rep("No",4822)
glm.pred[yhat.boost >.2]="Yes"
```

```r
glm_table=table(test$Purchase,glm.pred)
glm_table
```

```
##    glm.pred
##       No  Yes
##   0 4393  134
##   1  269   26
```

```r
a=glm_table[4]/(glm_table[4]+glm_table[3])
a
```

```
## [1] 0.1625
```

- a is the percent of people who are predicted to make a purchase using a boosting model who do in fact make one.

```r
set.seed(1)

library(class)
train.Direction =Purchase[train]
knn.pred=knn(train.set,test,train.Direction ,k=3)

knn_table=table(test$Purchase,knn.pred)
knn_table
```

```
##    knn.pred
##        0    1
##   0 4499   28
##   1  291    4
```

```r
b=knn_table[4]/(knn_table[4]+knn_table[3])
b
```

```
## [1] 0.125
```

```r
detach(Caravan)
rm(list=ls())
```

- b is the percent of people who are predicted to make a purchase using a knn model who do in fact make one.
- Boosting consistently makes better predictions than the KNN model.

```
data=read.csv("BeautyData.csv")
attach(data)
set.seed(1)
```

Problem 1 From the Exam

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.

```
set.seed(1)

lm.fit3=lm(CourseEvals~.,data=data)
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = CourseEvals ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31385 -0.30202  0.01011  0.29815  1.04929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.06542    0.05145  79.020  < 2e-16 ***
## BeautyScore  0.30415    0.02543  11.959  < 2e-16 ***
## female      -0.33199    0.04075  -8.146 3.62e-15 ***
## lower       -0.34255    0.04282  -7.999 1.04e-14 ***
## nonenglish  -0.25808    0.08478  -3.044  0.00247 **
## tenuretrack -0.09945    0.04888  -2.035  0.04245 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4273 on 457 degrees of freedom
## Multiple R-squared:  0.3471, Adjusted R-squared:  0.3399
## F-statistic: 48.58 on 5 and 457 DF,  p-value: < 2.2e-16
```

```
detach(data)
rm(list=ls())
```

- The model above includes course evals being the response variable, beauty score, a dummy variable for gender, a dummy variable for course level, a dummy variable for english speaking, and a dummy variable for tenured tracked professors as predictors. A one unit increase in the beauty score of a professor on average and all else equal is associated with a .304 unit increase in course evaluations. This variable is statistically significant at the 5% level. This finding indicates that more attractive professors receive better course evaluations. There are a host of other predictors that would make this model more robust in its controls. To name a few: raters college, professors course college, ethnicity of rater and professor, etc… These other determinants would provide more controls for the model which would isolate the causal relationship between beauty score and course evaluations.

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?

- In reference to Dr. Hamermesh's paper, he is referring to disentangling the effect that the effect of being attractive has a larger effect on male professors, than female professors. The reason why Dr. Hamermesh says its probably impossible to isolate if this result is due to productivity or discrimination is because there are too many endogenous factors within the study he ran to know for sure if the effect is discrimination or productivity. There could be key unobserved variables within the determinate that determine course evaluations that aren't observed in this study. There would be a randomized experiment in which the same course was taught to the same students at the same time of day, and the two professors had the same beauty level, and the only difference between the two experiments were that one professor was male and one was female. This is a near impossible experiment, which could be reduced to an instrumental variable if there was one. For Dr. Hamermesh's paper, and his model, to isolate productivity or discrimination was statistically impossible.

```r
data=read.csv("MidCity.csv")
set.seed(1)
```

Problem 2 From the Exam

1.

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

```r
data=read.csv("MidCity.csv")
data$Nbhd=factor(data$Nbhd)
fit=lm(Price~.,data=data)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27897.8  -6074.8    -48.7   5551.8  27536.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2037.726   8911.501   0.229 0.819524
## Home          -11.456     25.387  -0.451 0.652616
## Nbhd2       -1729.613   2433.756  -0.711 0.478675
## Nbhd3       20534.706   3176.051   6.465 2.33e-09 ***
## Offers      -8350.128   1103.693  -7.566 8.96e-12 ***
## SqFt           53.634      5.926   9.051 3.30e-15 ***
## BrickYes    17313.540   1988.548   8.707 2.12e-14 ***
## Bedrooms     4136.461   1621.775   2.551 0.012023 *
```

```
## Bathrooms    7975.157   2133.831   3.737 0.000287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10050 on 119 degrees of freedom
## Multiple R-squared:  0.8688, Adjusted R-squared:   0.86
## F-statistic: 98.54 on 8 and 119 DF,  p-value: < 2.2e-16
```

```r
set.seed(1)
```

- Holding all else equal, the variable for Brick houses is statistically significant and has a positive value. Which indicates that on average and all else equal, brick houses increase price by $17,313.540 compared to non brick houses.

2.

Is there a premium for houses in neighborhood 3?

- After converting the neighborhood variable into factor levels, on average and all else equal, neighborhood 3 is $20,534 more expensive than neighborhood 1.

3.

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

```r
set.seed(1)

fit=lm(Price~Home+Nbhd+Offers+SqFt+Brick+Bedrooms+Bathrooms+(Nbhd*Brick),data=data)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ Home + Nbhd + Offers + SqFt + Brick + Bedrooms +
##     Bathrooms + (Nbhd * Brick), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27843.9  -5544.3   -526.9   4167.3  28237.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3593.645   8860.065   0.406  0.68578
## Home             -12.410     24.975  -0.497  0.62020
## Nbhd2          -1527.046   2721.268  -0.561  0.57577
## Nbhd3          16807.264   3466.191   4.849 3.86e-06 ***
## Offers         -8470.621   1086.489  -7.796 2.91e-12 ***
## SqFt              54.427      5.866   9.278 1.10e-15 ***
## BrickYes       12033.113   4097.033   2.937  0.00399 **
## Bedrooms        4660.752   1608.651   2.897  0.00449 **
## Bathrooms       6554.909   2176.681   3.011  0.00319 **
## Nbhd2:BrickYes  2781.540   5090.237   0.546  0.58580
## Nbhd3:BrickYes 12019.217   5360.949   2.242  0.02685 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9879 on 117 degrees of freedom
## Multiple R-squared:  0.8755, Adjusted R-squared:  0.8648
## F-statistic: 82.24 on 10 and 117 DF,  p-value: < 2.2e-16
```

```r
rm(list=ls())
```

- The interaction between neighborhood 3 and the brick variable allows the model to check for the premium in brick houses in neighborhood 3 compared to neighborhood 1. The variable is statistically significant at the 5% level and on average holding all else equal a brick house in neighborhood 3 will cost 12,019 more than a brick house in neighborhood 1.

4.

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

- From the previous questions output, we see that neighborhood two does not have a 5% statistically significant effect on the price of a house compared to neighborhood one. Because of this result, we could combine neighborhood one and two since we can assume the effect of neighborhood two on price is no different than neighborhood one.

Problem 3 From the Exam

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)

- Most likely when cities have more crime they will increase the amount of cops accordingly, so there is already a positive relationship between crime and cops that is naturally observed in cities. So there needs to be a way to isolate the effects of an increase in cops on crime where the increase in cops isn't related to an increase in crime.

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.

- The researchers needed to find data where a lot of police were added to a city that weren't related to that cities crime levels.The way they were able to do this was in DC the monitored the terrorism alert system, when the terror alert level goes to orange DC positions more cops in the city. This increase in cops is a perfect way to measure the effect of an increase in cops on crime, when the increase in cops is unrelated to the cities crime level.
- From Table 2, they found that when DC is on high alert and more cops are positioned in the city, on average and when holding midday ridership constant there is a 6.04 decrease in the number of daily crimes.

3.

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

- They needed to check rider levels to see if on high alert days less people were just in DC which might have decreased crime, which would have over emphasized the effect of an increase cops on crime levels. These variables are statistically significant at the 5% level.

4.

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

- The model that is being estimated is the effect of more cops due to a high alert on different police districts within DC. On average and all else equal, when DC is under high alert, police district one has a 2.62 more decrease in daily crime than the base line police district. Also one average and all else equal when DC is under high alert, other police districts has a .571 more decrease in daily crime than the baseline district.

Problem 4 From the Exam

1.

Describe your contribution to the "Pricing Cars" group project (1 page max)

- My contribution to the Pricing Cars group project consisted of four parts: meetings, R assignments, write up, and presentation. Group 8 met numerous times throughout the preparation process, I scheduled the first meeting where we went over the game plan for approaching the project, we laid out the timeline of the project and broke it down into steps which had deadlines. The second meeting consisted of assigning models from the textbook to run on our data. I was assigned the support vector machine, and brought the results to the third meeting where we all compared our results. Throughout the entire project, I tried numerous models past the one I was assigned, and also presented the results of those models to compare to my teammates models. Once we decided our model, which was random forest, Gaby had already ran the code in which she sent it to everyone to check, and run it for ourselves to check for any errors. Once we had our final model we began to do the write up. I was assigned the initial data analysis section where I discussed the small amount of data munging we did, and which variables appeared to provide the most information to car price just from a visual analysis. Once I finished my part in the write up, I also helped write the section that talked about the boosting models we ran, and the limitation section of the paper. Once our write up was completed we started working on the presentation, we each presented on the section we wrote about in the write up. I made numerous plots for the initial data analysis showing why we dropped some variables and our reasoning behind putting importance on certain variables. Since I was the first one to speak I also introduced the problem and introduced the team in the presentation. The weekend before the presentation the group met again to practice the presentation and make the final draft of the write up and presentation.