INFX 573: Problem Set 6 - Regression

Pierre Augustamar

Due: Tuesday, November 15, 2016

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset6.Rmd` file from Canvas. Open `problemset6.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset6.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps6.Rmd`, knit a PDF and submit the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(MASS) # Modern applied statistics functions
```

In this problem we will use the Boston dataset that is available in the `MASS` package. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. Load this data and use it to answer the following questions.

1. Describe the data and variables that are part of the `Boston` dataset. Tidy data as necessary.

```
dim(Boston) # get number of records and number of columns associated with this data set
```

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

```
str(Boston) # get detail info about the type of data and the column name
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```r
head(Boston) # get the top 5 data record
```

```
##      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
```

```r
summary(Boston) #produces result summary
```

```
##       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
```

Response - Describe data

crim - per capita crime rate by town. zn - proportion of residential land zoned for lots over 25,000 sq.ft. indus - proportion of non-retail business acres per town. chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). nox - nitrogen oxides concentration (parts per 10 million). rm - average number of rooms per dwelling. age - proportion of owner-occupied units built prior to 1940. dis - weighted mean of distances to five Boston employment centres. rad - index of accessibility to radial highways. tax - full-value property-tax rate per $10,000. ptratio - pupil-teacher ratio by town. black - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town. lstat - lower status of the population (percent). medv - median value of owner-occupied homes in $1000

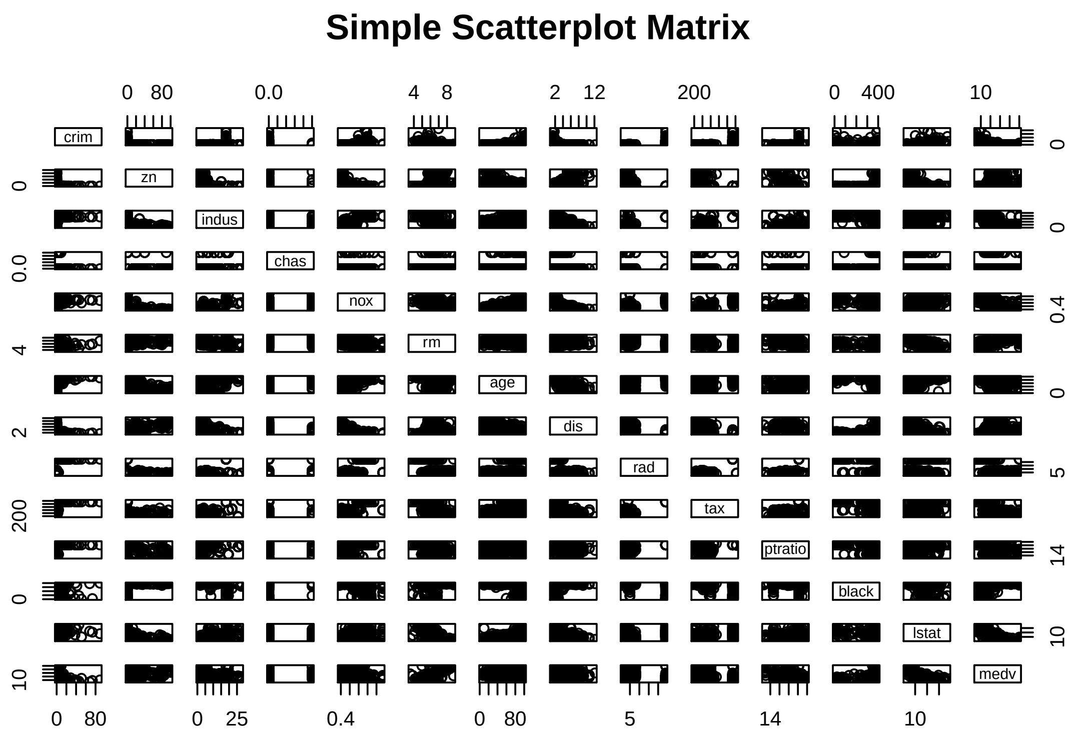
```
#check if any of the variables contain na values
apply(Boston, 2, function(x) any(is.na(x)))
```

```
##    crim      zn   indus    chas     nox      rm     age     dis     rad
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
##     tax ptratio   black   lstat    medv
##   FALSE   FALSE   FALSE   FALSE   FALSE
```

In order to tidy the data, we first check if there are any na value. Our investigation reveals that none of the variables appear to have any na values. Thus, at this time no cleanups will be needed. But we may need to investigate each reported data for each of the variables to check for any unsual values or data patterns.

2. Consider this data in context, what is the response variable of interest? Discuss how you think some of the possible predictor variables might be associated with this response.

```
#scatter plots matrix
pairs(Boston, main="Simple Scatterplot Matrix")
```



Before I answered the following questions, I generated a matrix that shows an overall representation of possible medv (median home values) and most of the other variables except for variable, Charles River, which does not seem to have any correlation with any of the other variables.

3. For each predictor, fit a simple linear regression model to predict the response. 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.

Response

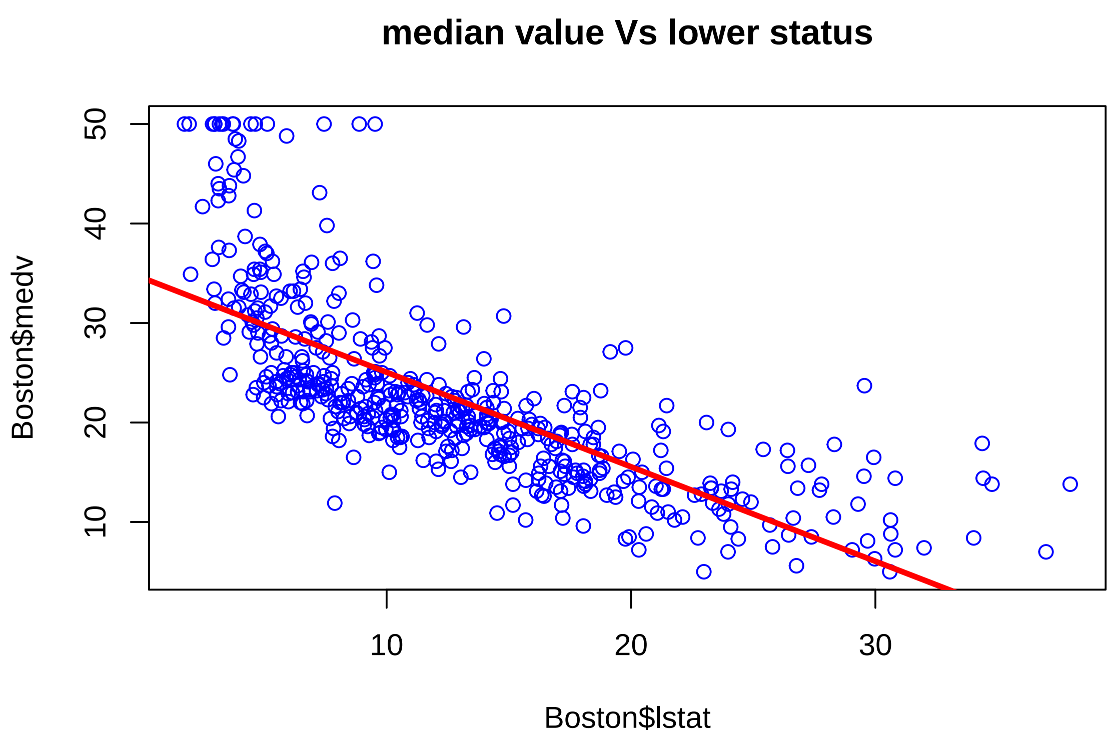
Using median house or medv as the response to generate a regression model

```
#Linear regression between median value and lower status of the population
lstat.lm = lm(Boston$medv ~ Boston$lstat, data=Boston)
summary(lstat.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.55384    0.56263   61.41   <2e-16 ***
## Boston$lstat -0.95005    0.03873  -24.53   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
# plot the response and the predictor
plot(Boston$lstat, Boston$medv,   col = "blue", main="median value Vs lower status")

# use abline() to display the least squares regression line
abline(lstat.lm, col = "red", lwd=3)
```



In this analysis, we are predicting the median home value at a unit of $1000 in function of lower status of the population. The data shows that lsat is statistically significant at 0.05 significance level. Also, The lsat coefficient variable indicates that for every one percent increase, the response variable medv decreases by 0.95.

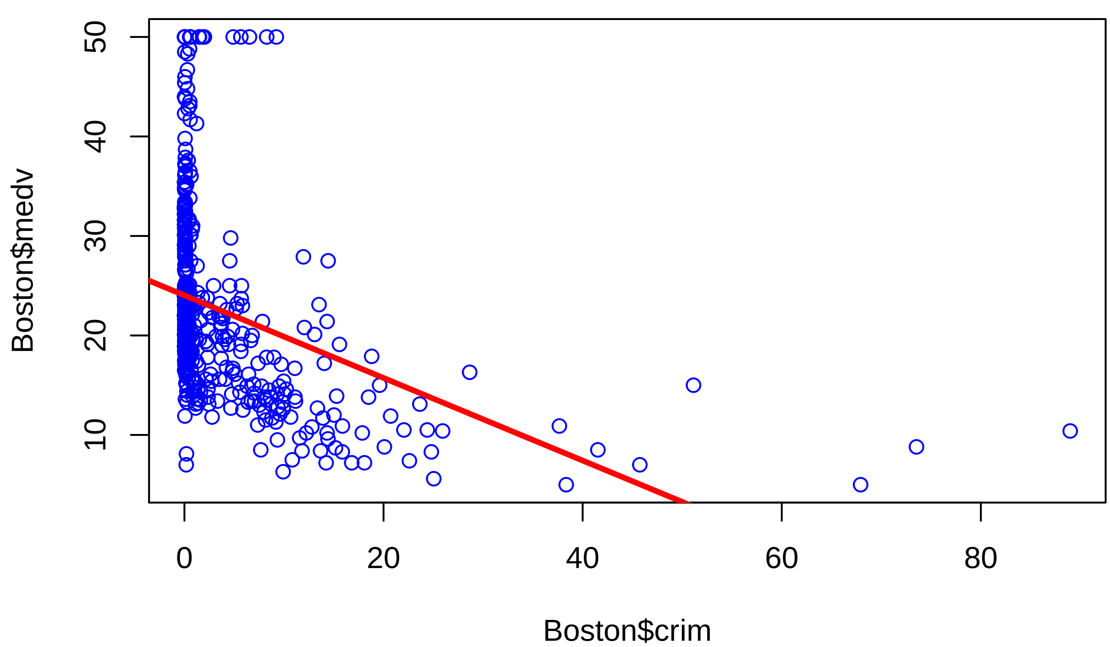
The graph shows that there are is a linear downhill relationship between median value home and lower status.

```
#Linear regression between median value and crime
crim.lm = lm(Boston$medv ~ Boston$crim, data=Boston)
summary(crim.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$crim, data = Boston)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.957  -5.449  -2.007   2.512  29.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.03311    0.40914   58.74   <2e-16 ***
## Boston$crim -0.41519    0.04389   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 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 the response and the predictor
plot(Boston$crim, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(crim.lm, col = "red", lwd=3)
```



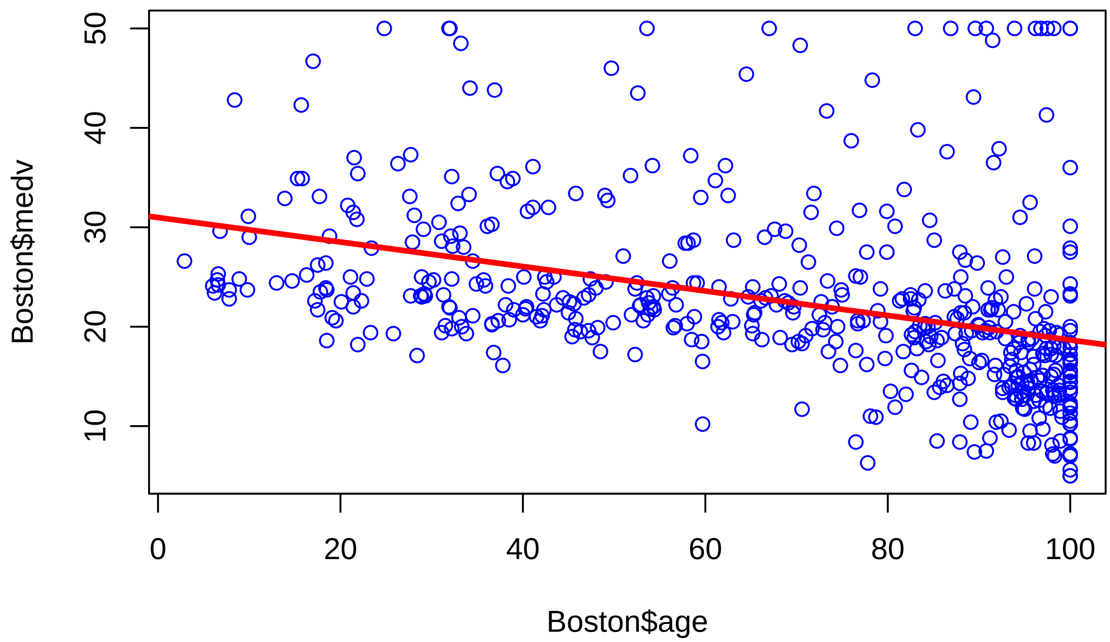
In this analysis, we are predicting the median home value at a unit of $1000 in function of crime. The data shows that crime is statistically significant at 0.05 significance level. Also, The crime coefficient variable indicates that for every single increase in crime, the response variable medv decreases by 0.41.

The graph shows that there is a linear downhill relationship betweebn median value home and crime. Also, it shows that there seems to be a higher level of crime in areas with lower median housing values.

```r
#Linear regression between median value and lower status of the population
age.lm = lm(Boston$medv ~ Boston$age, data=Boston)
summary(age.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  31.338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97868    0.99911  31.006   <2e-16 ***
## Boston$age  -0.12316    0.01348  -9.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
## Multiple R-squared:  0.1421, Adjusted R-squared:  0.1404
## F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
# plot the response and the predictor
plot(Boston$age, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(age.lm, col = "red", lwd=3)
```



In this analysis, we are predicting the median home value at a unit of $1000 in function of the proportion's age of the occupied owner. The data shows that age is statistically significant at 0.05 significance level. Also, The crime coefficient variable indicates that for every single increase in age, the response variable medv decreases by 0.12.

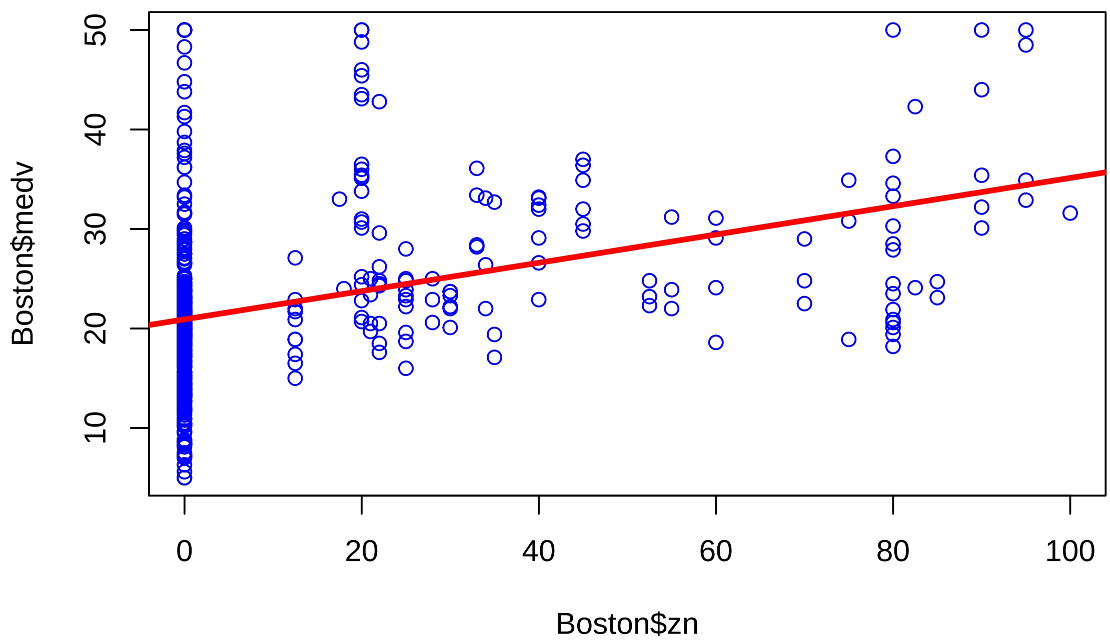
The graph shows that there is a linear downhill relationship between median value home and proportion's age of the occupied owner. Also, it shows a number of outliers.

```
#Linear regression between median value and lower status of the population
zn.lm = lm(Boston$medv ~ Boston$zn, data=Boston)
summary(zn.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$zn, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.918  -5.518  -1.006   2.757  29.082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.91758    0.42474  49.248   <2e-16 ***
## Boston$zn    0.14214    0.01638   8.675   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.587 on 504 degrees of freedom
```

```
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.1282
## F-statistic: 75.26 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
# plot the response and the predictor
plot(Boston$zn, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(zn.lm, col = "red", lwd=3)
```



In this analysis, we are predicting the median home value at a unit of $1000 in function of zn (the proportion of residential land zoned for lots over 25,000 sq.ft) The data shows that zn is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in zn, the response variable medv decreases by 0.14. Also, there are a large set of data that has a zn value = zero. Thus, there are a number of median value homed in $1000 that has no land zoned for lots over 25,000 sq.ft.

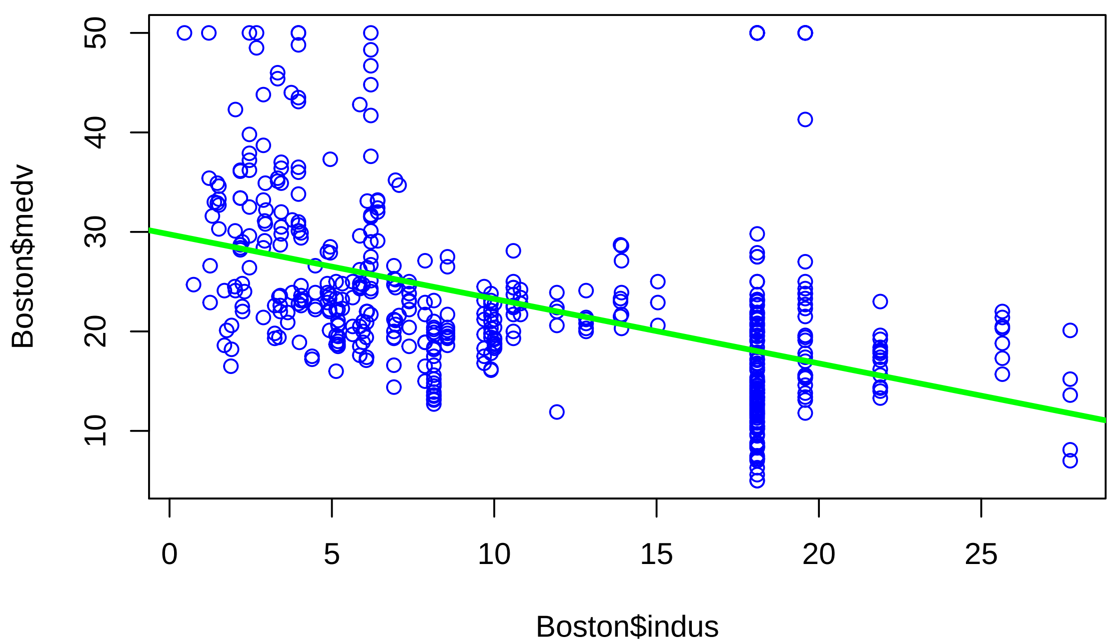
The graph shows that there is a linear uphill relationship between median value home and proportion's of residential land zoned for lots over 25,000 sq.ft.

```
#Linear regression between median value and lower status of the population
indus.lm = lm(Boston$medv ~ Boston$indus, data=Boston)
summary(indus.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$indus, data = Boston)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -13.017  -4.917  -1.457   3.180  32.943
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.75490    0.68345   43.54   <2e-16 ***
## Boston$indus -0.64849    0.05226  -12.41   <2e-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:  0.234,  Adjusted R-squared:  0.2325
## F-statistic:   154 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
# plot the response and the predictor
plot(Boston$indus, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(indus.lm, col = "green",lwd=3)
```



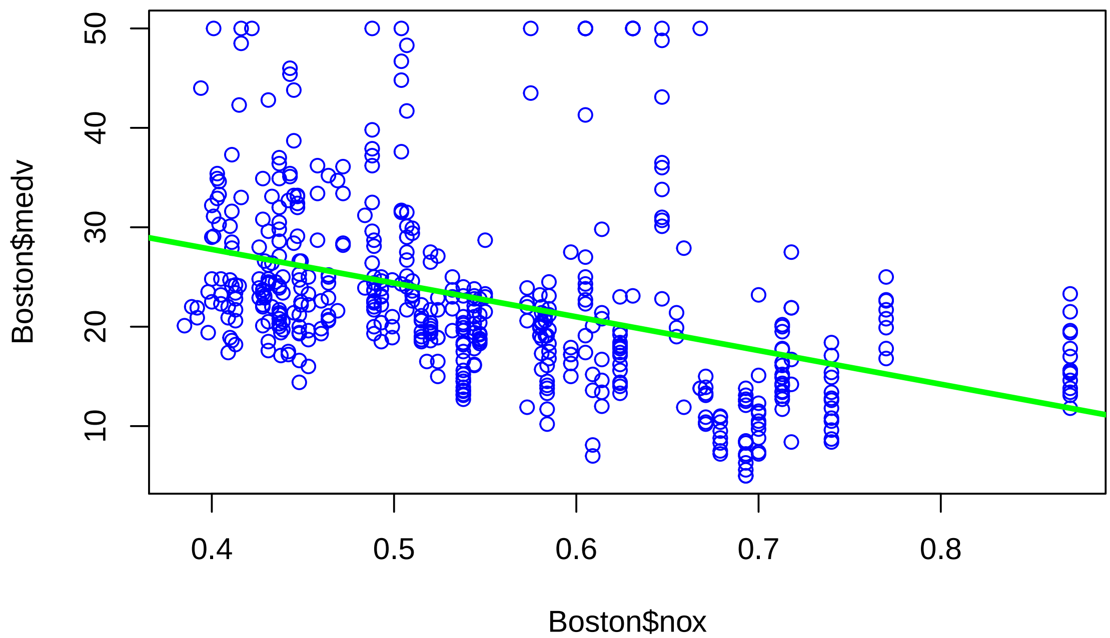
In this analysis, we are predicting the median home value at a unit of $1000 in function of indus (proportion of non-retail business acres per town) The data shows that indus is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in indus, the response variable medv decreases by 0.64.

The graph shows that there are is a linear downhill relationship between median value home and proportion's of non-retail business acres per town.

```r
#Linear regression between median value and lower status of the population
nox.lm = lm(Boston$medv ~ Boston$nox, data=Boston)
summary(nox.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$nox, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.691  -5.121  -2.161   2.959  31.310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.346      1.811   22.83   <2e-16 ***
## Boston$nox   -33.916      3.196  -10.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.323 on 504 degrees of freedom
## Multiple R-squared:  0.1826, Adjusted R-squared:  0.181
## F-statistic: 112.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
# plot the response and the predictor
plot(Boston$nox, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(nox.lm, col = "green",lwd=3)
```



In this analysis, we are predicting the median home value at a unit of $1000 in function of nox (nitrogen oxides concentration (parts per 10 million)) The data shows that nox is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in nox, the response variable medv decreases by 33.916.

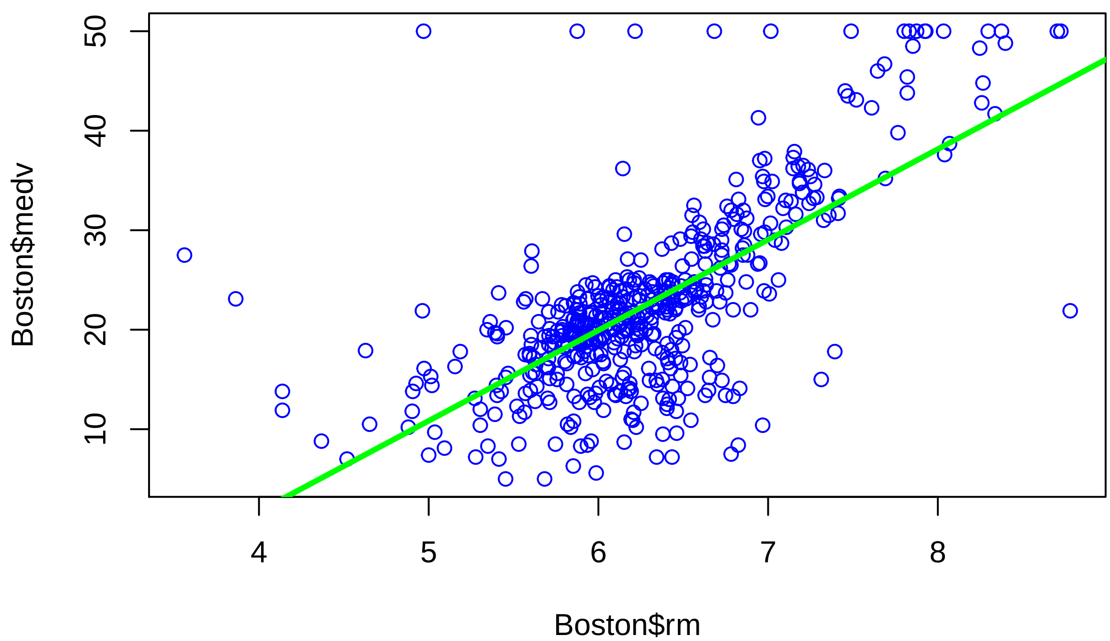
The graph shows that there is a linear downhill relationship between median value home and nitrogen oxides concentration (parts per 10 million).

```
#Linear regression between median value and lower status of the population
rm.lm = lm(Boston$medv ~ Boston$rm, data=Boston)
summary(rm.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$rm, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## Boston$rm      9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
```

```
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
# plot the response and the predictor
plot(Boston$rm, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(rm.lm, col = "green", lwd=3)
```



In this analysis, we are predicting the median home value at a unit of $1000 in function of rm (average number of rooms per dwelling) The data shows that rm is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in rm, the response variable medv increases by 9.102.

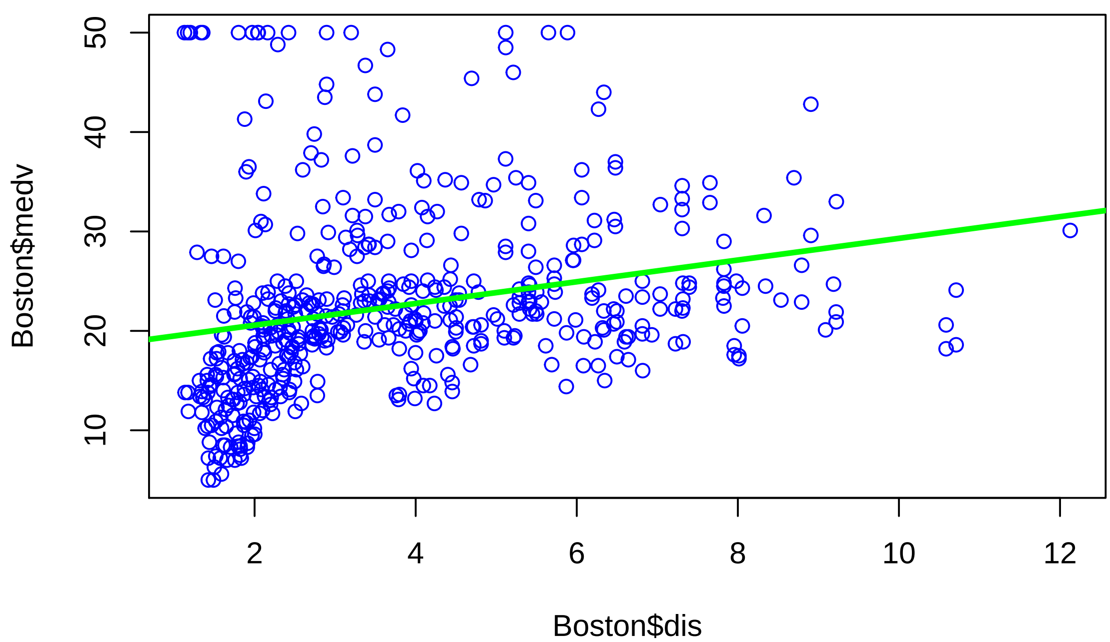
The graph shows that there is a linear uphill relationship between median value home and nitrogen oxides concentration (parts per 10 million.

```r
#Linear regression between median value and lower status of the population
dis.lm = lm(Boston$medv ~ Boston$dis, data=Boston)
summary(dis.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$dis, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.016  -5.556  -1.865   2.288  30.377
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.3901     0.8174  22.499  < 2e-16 ***
## Boston$dis    1.0916     0.1884   5.795 1.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 504 degrees of freedom
## Multiple R-squared:  0.06246,    Adjusted R-squared:  0.0606
## F-statistic: 33.58 on 1 and 504 DF,  p-value: 1.207e-08
```

```
# plot the response and the predictor
plot(Boston$dis, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(dis.lm, col = "green",lwd=3)
```



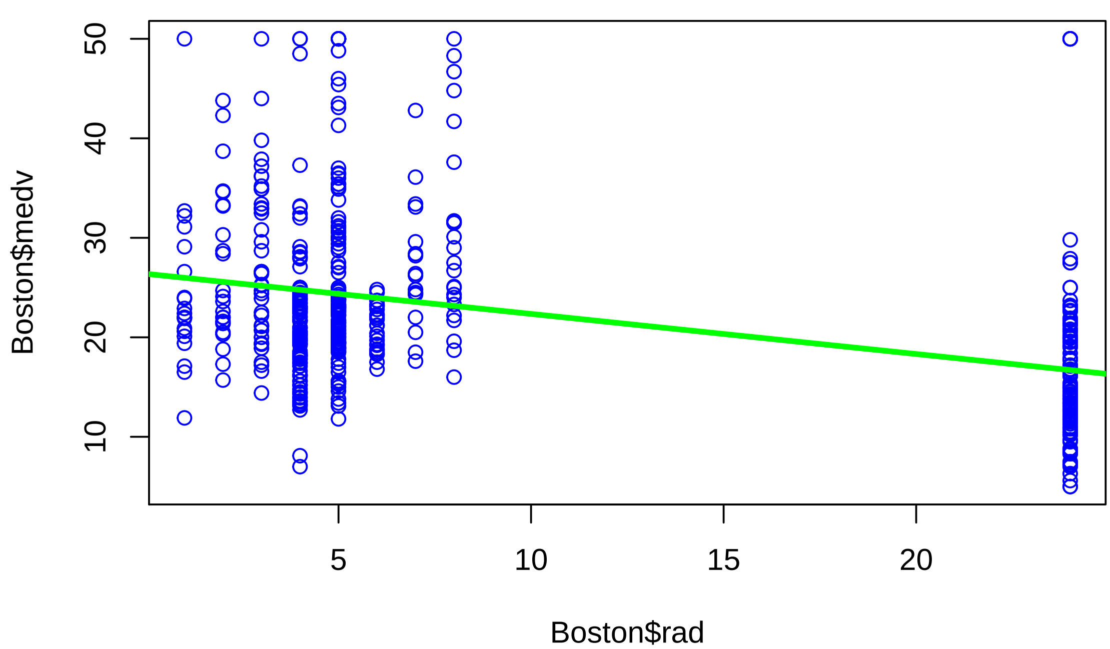
In this analysis, we are predicting the median home value at a unit of $1000 in function of dis (weighted mean of distances to five Boston employment centres) The data shows that dis is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in dis, the response variable medv increases by 1.09

The graph shows that there is a linear uphill relationship between median value home and weighted mean of distances to five Boston employment centres.

```
#Linear regression between median value and lower status of the population
rad.lm = lm(Boston$medv ~ Boston$rad, data=Boston)
summary(rad.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.770  -5.199  -1.967   3.321  33.292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.38213    0.56176  46.964   <2e-16 ***
## Boston$rad  -0.40310    0.04349  -9.269   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.509 on 504 degrees of freedom
## Multiple R-squared:  0.1456, Adjusted R-squared:  0.1439
## F-statistic: 85.91 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
# plot the response and the predictor
plot(Boston$rad, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(rad.lm, col = "green",lwd=3)
```



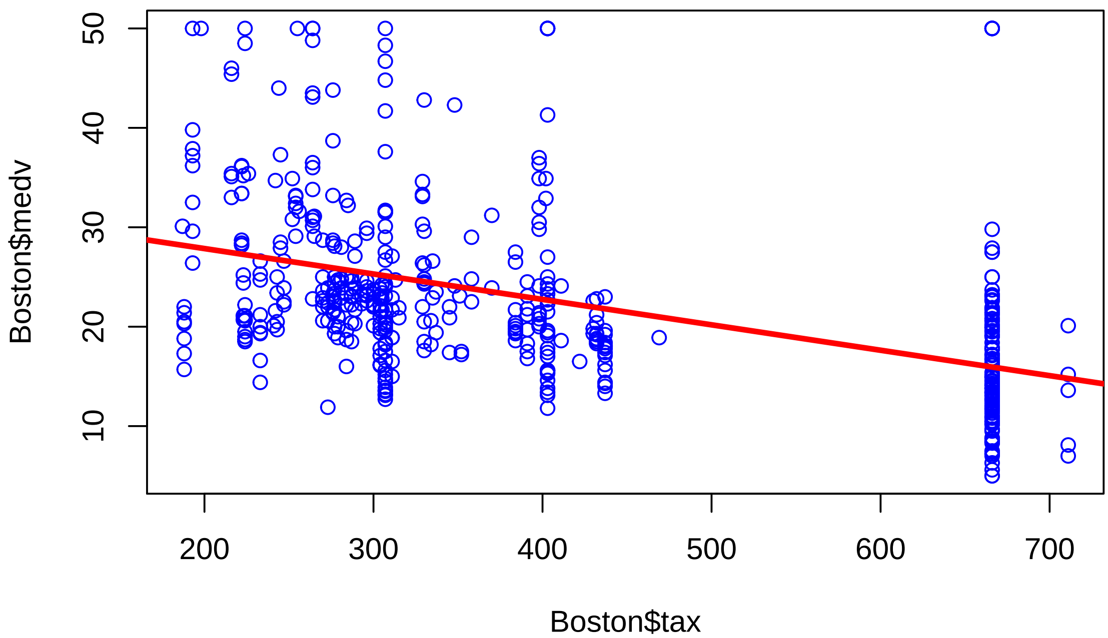
In this analysis, we are predicting the median home value at a unit of $1000 in function of rad (index of accessibility to radial highways) The data shows that rad is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in rad, the response variable medv decreases by 0.40.

The graph shows that there is a linear downhill relationship between median value home and index of accessibility to radial highways.

```
#Linear regression between median value and lower status of the population
tax.lm = lm(Boston$medv ~ Boston$tax, data=Boston)
summary(tax.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$tax, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.091  -5.173  -2.085   3.158  34.058
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.970654   0.948296   34.77   <2e-16 ***
## Boston$tax  -0.025568   0.002147  -11.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.133 on 504 degrees of freedom
## Multiple R-squared:  0.2195, Adjusted R-squared:  0.218
## F-statistic: 141.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
# plot the response and the predictor
plot(Boston$tax, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(tax.lm, col = "red",lwd=3)
```



In this analysis, we are predicting the median home value at a unit of $1000 in function of tax (full-value property-tax rate per $10,000.) The data shows that tax is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in tax, the response variable medv decreases by 0.025.

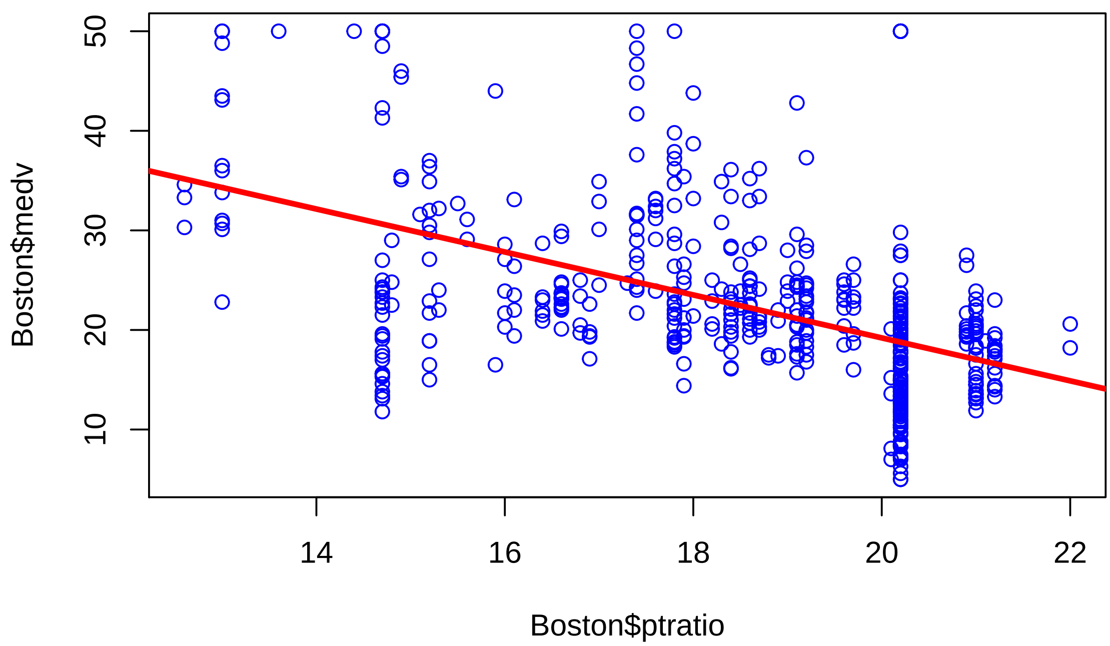
The graph shows that there is linear downhill relationship between median value home and full-value property-tax rate per $10,000.

```
#Linear regression between median value and lower status of the population
ptRatio.lm = lm(Boston$medv ~ Boston$ptratio, data=Boston)
summary(ptRatio.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$ptratio, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8342  -4.8262  -0.6426   3.1571  31.2303
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      62.345      3.029   20.58   <2e-16 ***
## Boston$ptratio   -2.157      0.163  -13.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.931 on 504 degrees of freedom
```

```
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.2564
## F-statistic: 175.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
# plot the response and the predictor
plot(Boston$ptratio, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(ptRatio.lm, col = "red",lwd=3)
```



In this analysis, we are predicting the median home value at a unit of $1000 in function of ptratio (pupil-teacher ratio by town.) The data shows that ptratio is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in ptratio, the response variable medv decreases by 2.157.

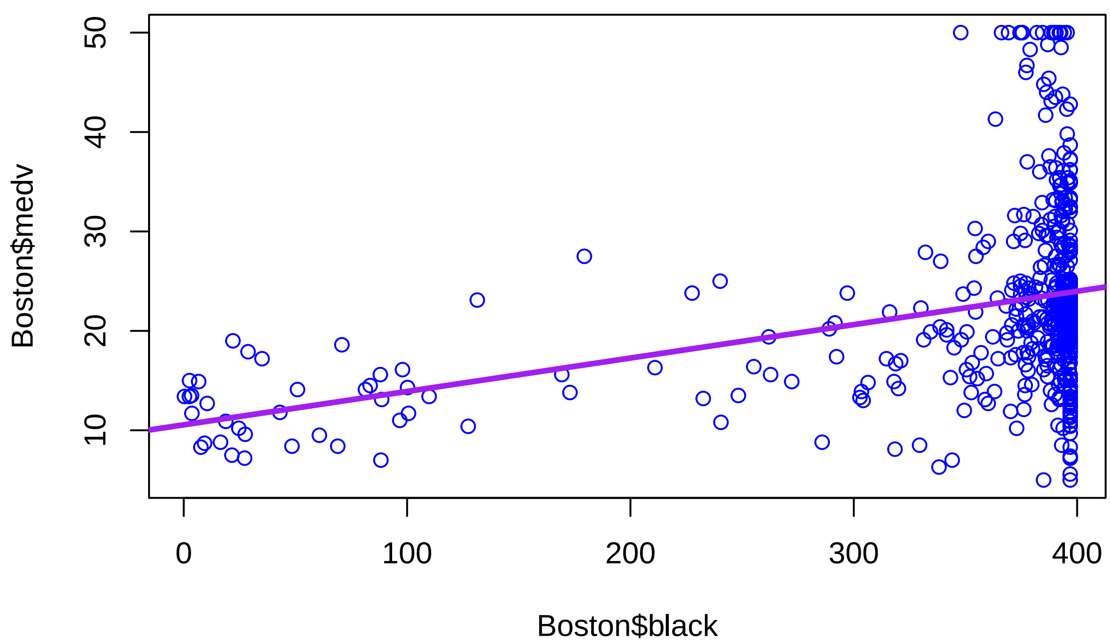
The graph shows that there is linear downhill relationship between median value home and pupil-teacher ratio by town.

```r
#Linear regression between median value and lower status of the population
black.lm = lm(Boston$medv ~ Boston$black, data=Boston)
summary(black.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$black, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.884  -4.862  -1.684   2.932  27.763
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.551034   1.557463   6.775 3.49e-11 ***
## Boston$black  0.033593   0.004231   7.941 1.32e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.679 on 504 degrees of freedom
## Multiple R-squared:  0.1112, Adjusted R-squared:  0.1094
## F-statistic: 63.05 on 1 and 504 DF,  p-value: 1.318e-14
```

```
# plot the response and the predictor
plot(Boston$black, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(black.lm, col = "purple",lwd=3)
```



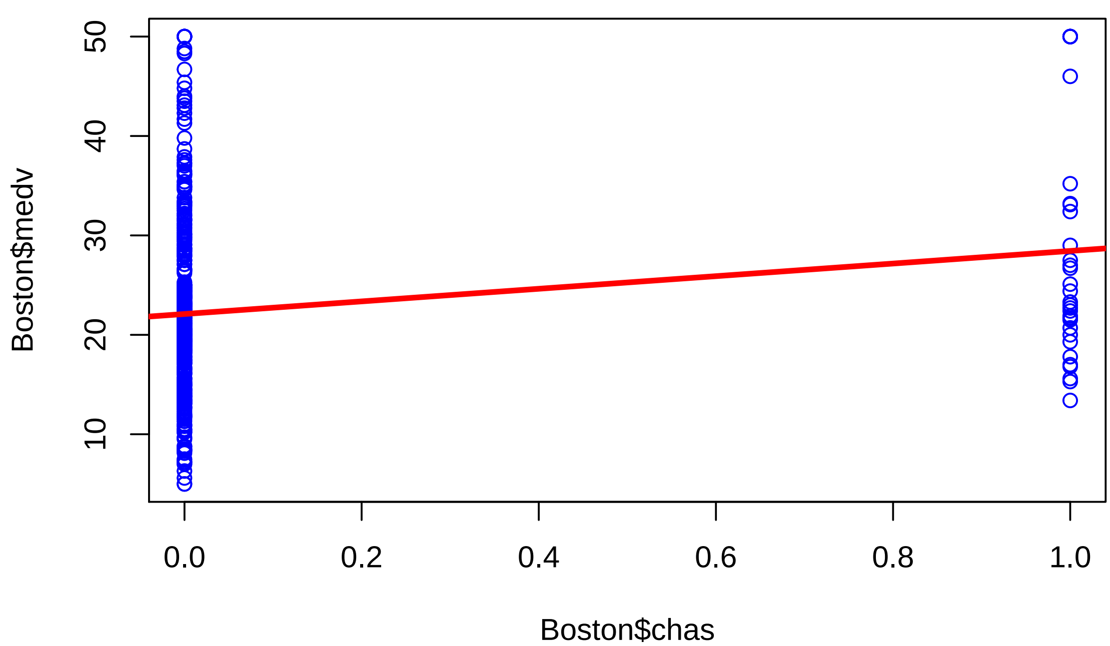
In this analysis, we are predicting the median home value at a unit of $1000 in function of black (1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.) The data shows that black is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in proportion of blacks by town, the response variable medv increases by 0.033.

The graph shows that there is a linear uphill relationship between median value home and 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

```
#Linear regression between median value and lower status of the population
chas.lm = lm(Boston$medv ~ Boston$chas, data=Boston)
summary(chas.lm)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$chas, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0938     0.4176  52.902  < 2e-16 ***
## Boston$chas   6.3462     1.5880   3.996 7.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.064 on 504 degrees of freedom
## Multiple R-squared:  0.03072,    Adjusted R-squared:  0.02879
## F-statistic: 15.97 on 1 and 504 DF,  p-value: 7.391e-05
```

```r
# plot the response and the predictor
plot(Boston$chas, Boston$medv,   col = "blue")

# use abline() to display the least squares regression line
abline(chas.lm, col = "red", lwd=3)
```



In this analysis, we are predicting the median home value at a unit of $1000 in function of chas (Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). The data shows that chas is statistically significant at 0.05 significance level. Also, The coefficient variable indicates that for every single increase in proportion of chas, the response variable medv decreases by 0.64.

The graph shows that there are is a linear downhill relationship between median value home and Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

4. 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 $H_0 : \beta_j = 0$?

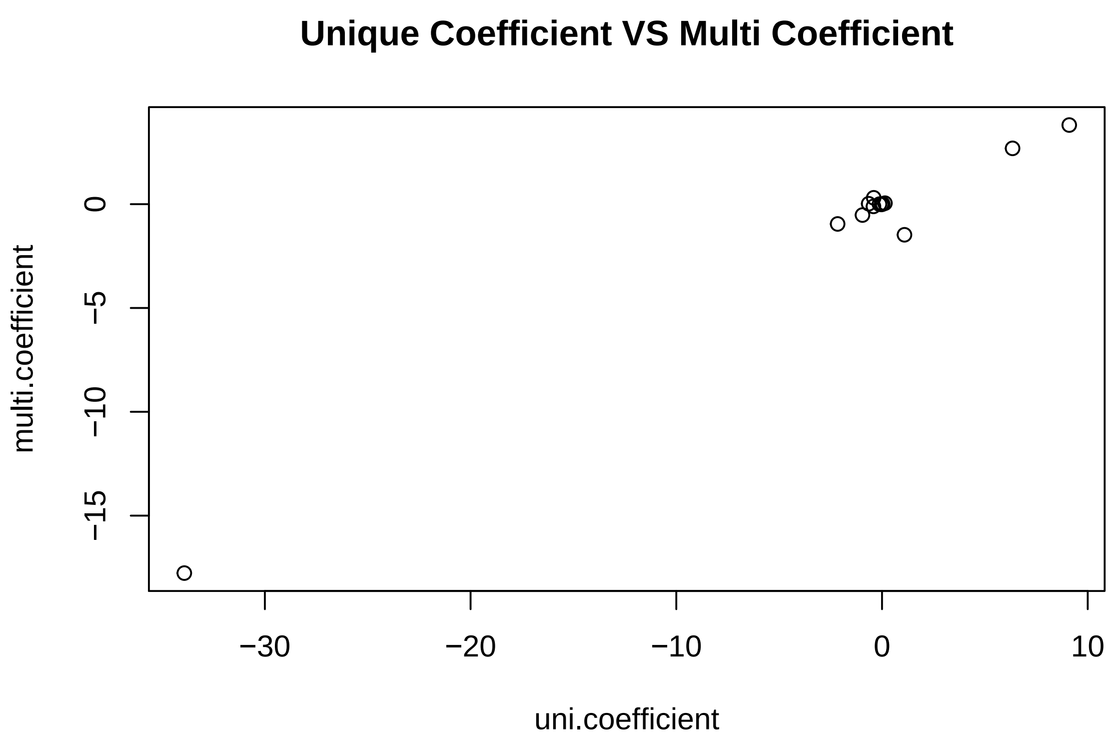
```r
#multiple regression model
multi.lm = lm(medv ~., data = Boston)
summary(multi.lm)
```

```
##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

A low p-value(<0.005) signifies that we can reject the null hypothesis. Thus, any changes in the predictor's value are related to changes in the response variable. Alternatively, large significant p-value signifies that changes in the predictor are not associated with changes in the response. Based on this, when looking at the result under the coefficient, using the median home value at a unit of $1000 as the intercecept, we can reject the null hypothesis for the following predictors: age, and indus.

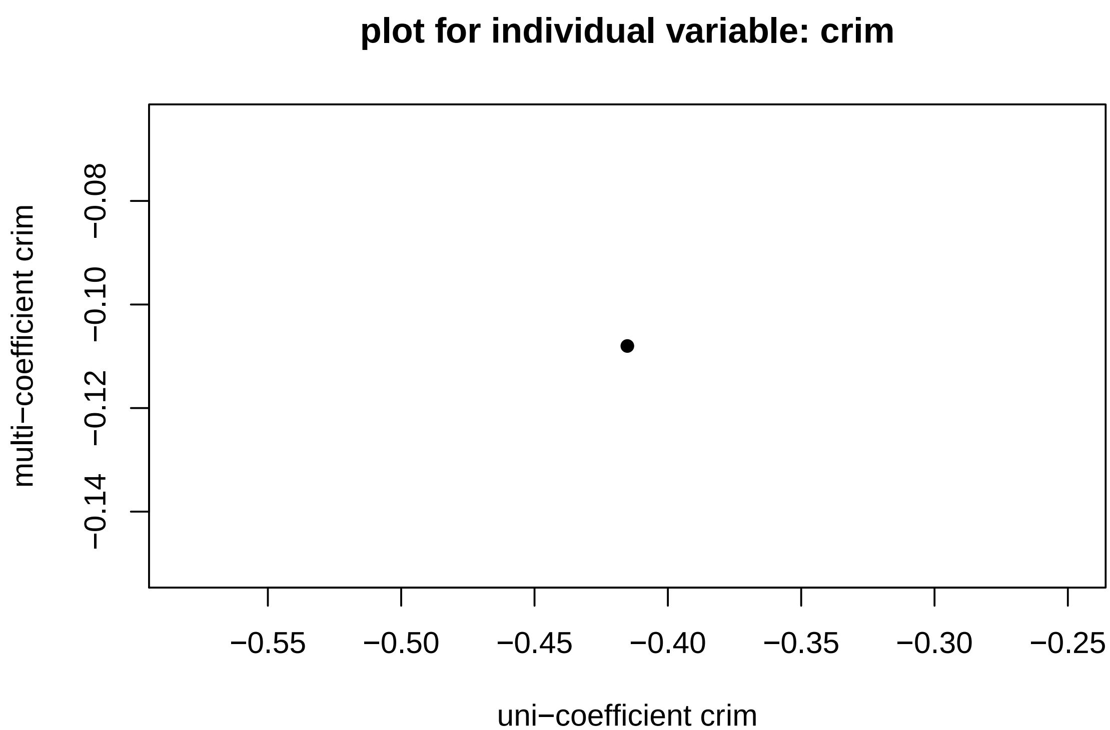
5. How do your results from (3) compare to your results from (4)? Create a plot displaying the univariate regression coefficients from (3) on the x-axis and the multiple regression coefficients from part (4) on the y-axis. Use this visualization to support your response.

```
# add the unique coefficient of each variables into a vector
uni.coefficient = c(coefficients(crim.lm)[2],
                    coefficients(zn.lm)[2],
                    coefficients(indus.lm)[2],
                    coefficients(chas.lm)[2],
                    coefficients(nox.lm)[2],
                    coefficients(rm.lm)[2],
                    coefficients(age.lm)[2],
                    coefficients(dis.lm)[2],
                    coefficients(rad.lm)[2],
                    coefficients(tax.lm)[2],
                    coefficients(ptRatio.lm)[2],
                    coefficients(black.lm)[2],
                    coefficients(lstat.lm)[2]
                    )
#add the coefficients into a vector
multi.coefficient = coefficients(multi.lm)[-1]

# graphs plotting unique vs multi coefficient
plot(uni.coefficient, multi.coefficient, main = "Unique Coefficient VS Multi Coefficient")
```

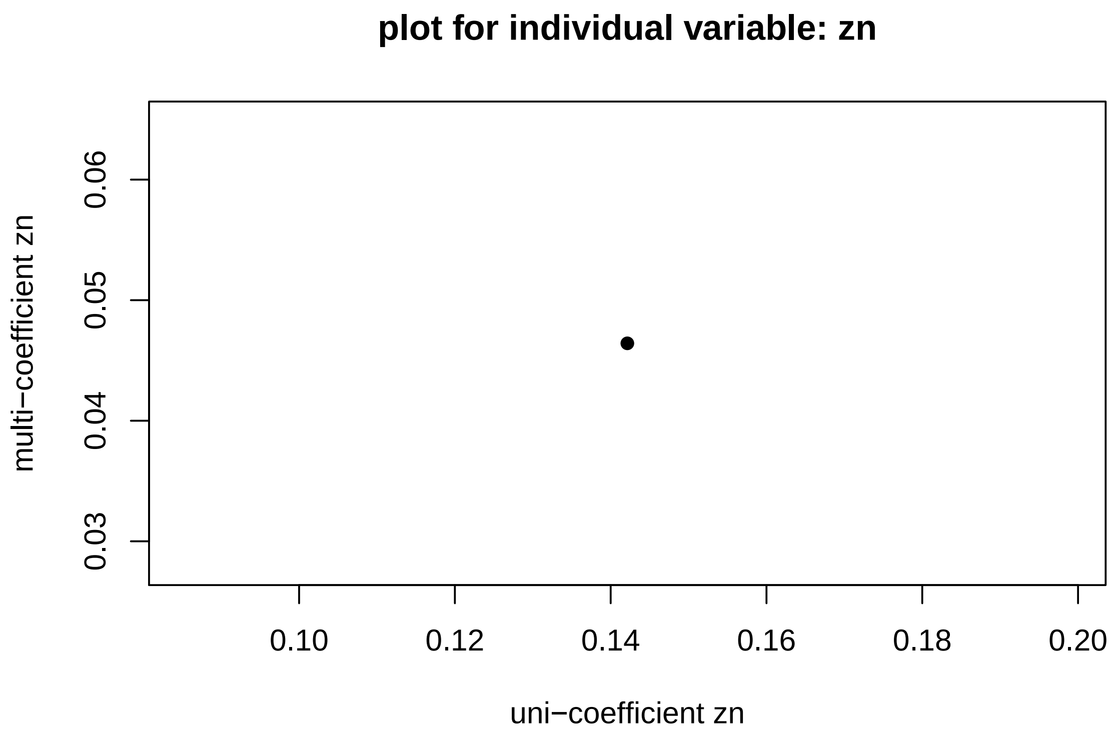



```
#function to calculate variations between unique regression and multi regression values
varationsCalculator = function(){
  graphValues = c('crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad', 'tax', 'ptratio'


    for(i in 1:length(uni.coefficient)){
```

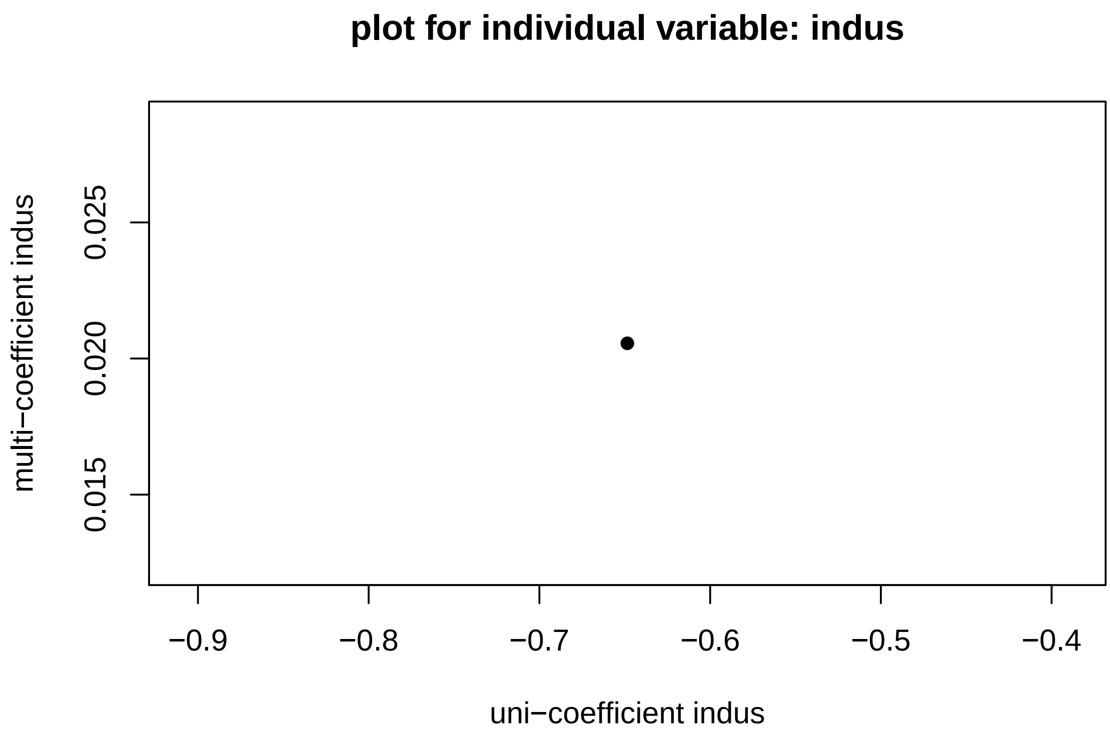
```r
        xCord = paste('uni-coefficient',graphValues[i])
        ycord = paste('multi-coefficient', graphValues[i])
        mainlabel = paste('plot for individual variable:', graphValues[i])

        # plot individual variables
        plot(uni.coefficient[i], multi.coefficient[i],pch=16, xlab = xCord,
              ylab = ycord, main = mainlabel)


        #print the variations from unique regression to multi regression
        print('varation')
        print(uni.coefficient[i] - multi.coefficient[i])
    }
}

varationsCalculator()
```



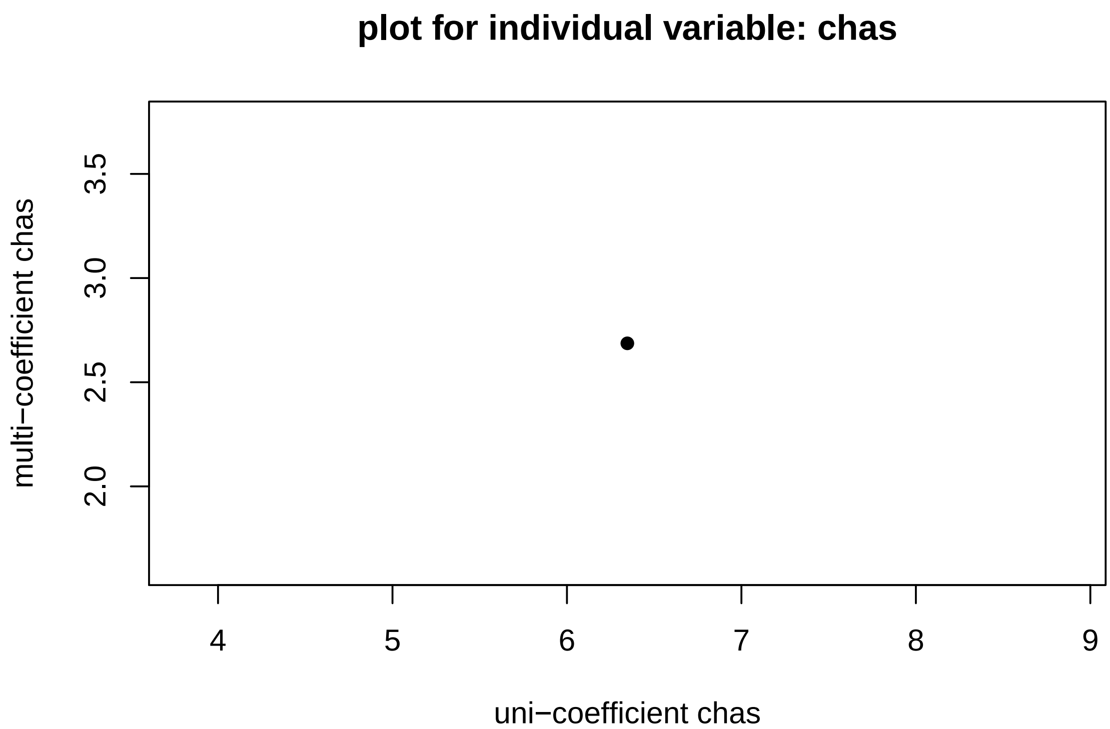

```
## [1] "varation"
## Boston$crim
##  -0.3071789
```



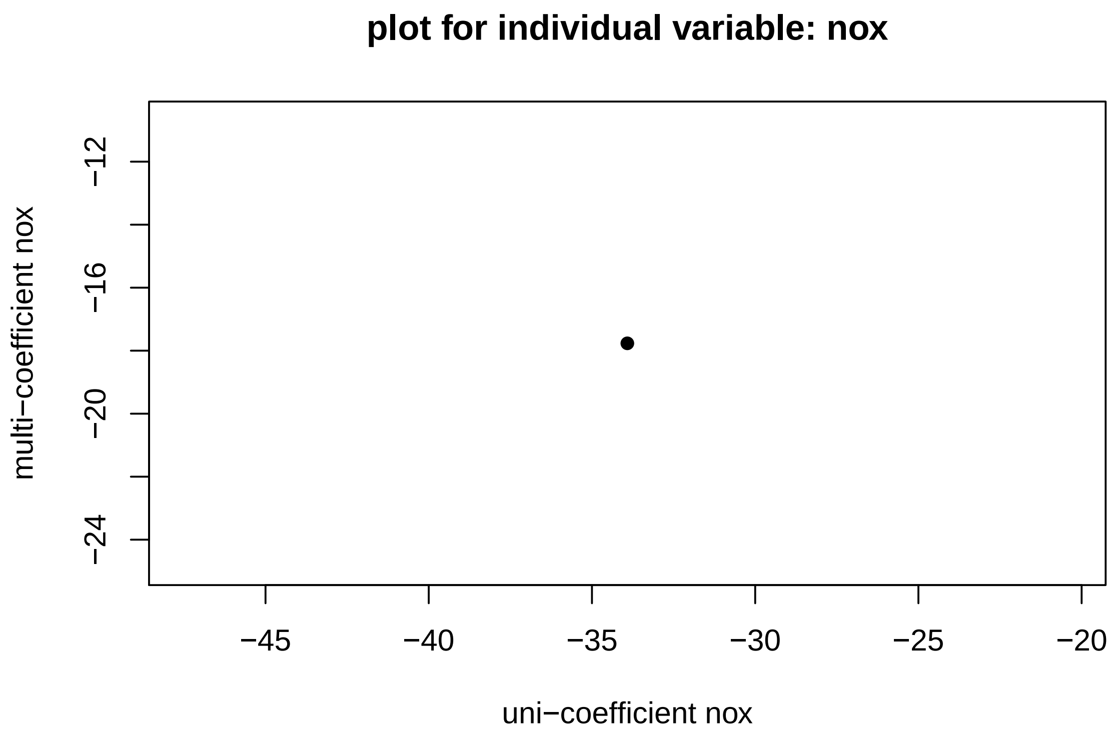

```
## [1] "varation"
##  Boston$zn
## 0.09571954
```



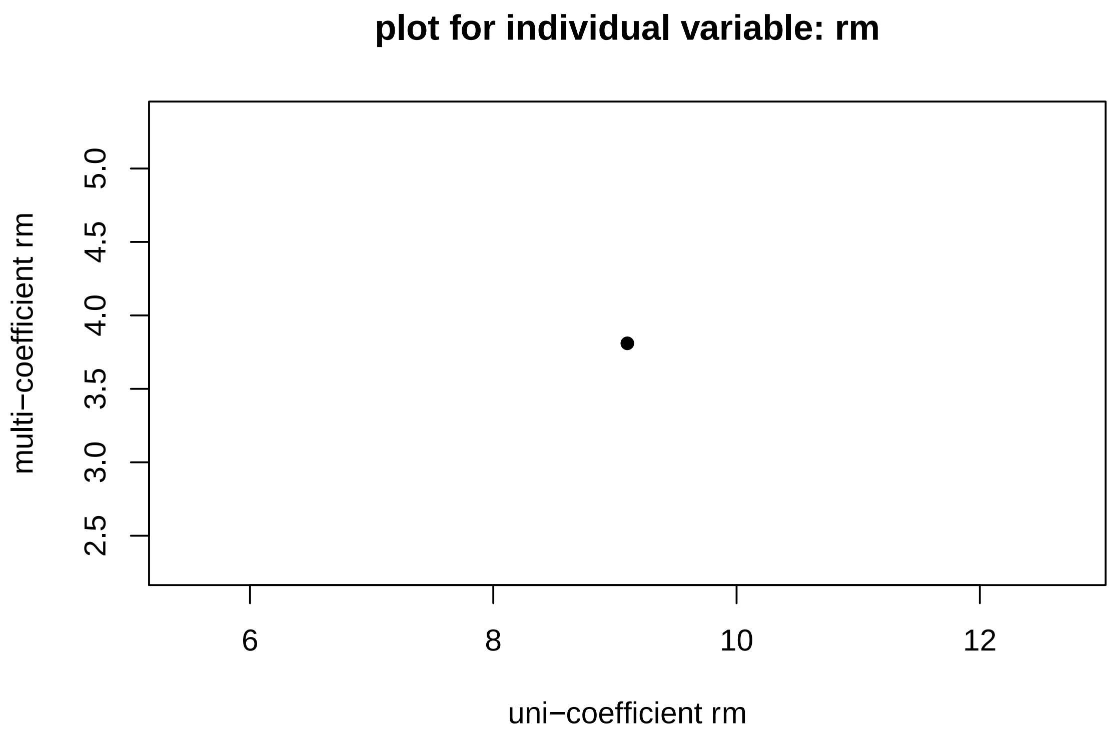

```
## [1] "varation"
## Boston$indus
##   -0.6690487
```



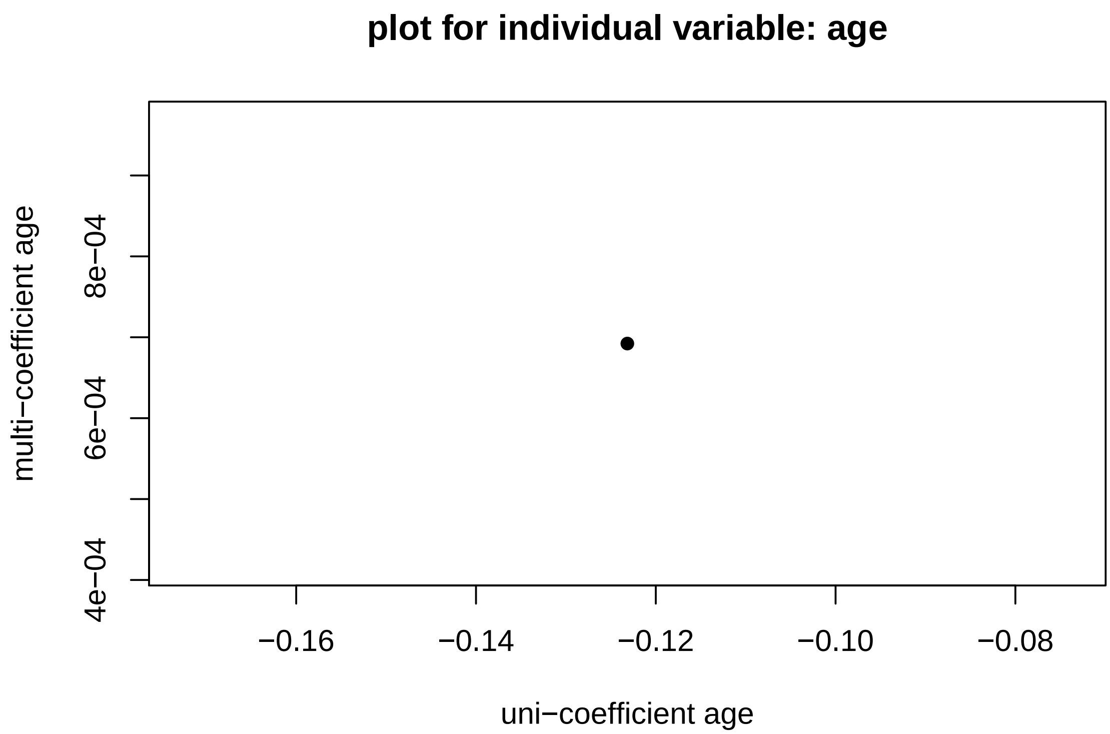

```
## [1] "varation"
## Boston$chas
##    3.659423
```



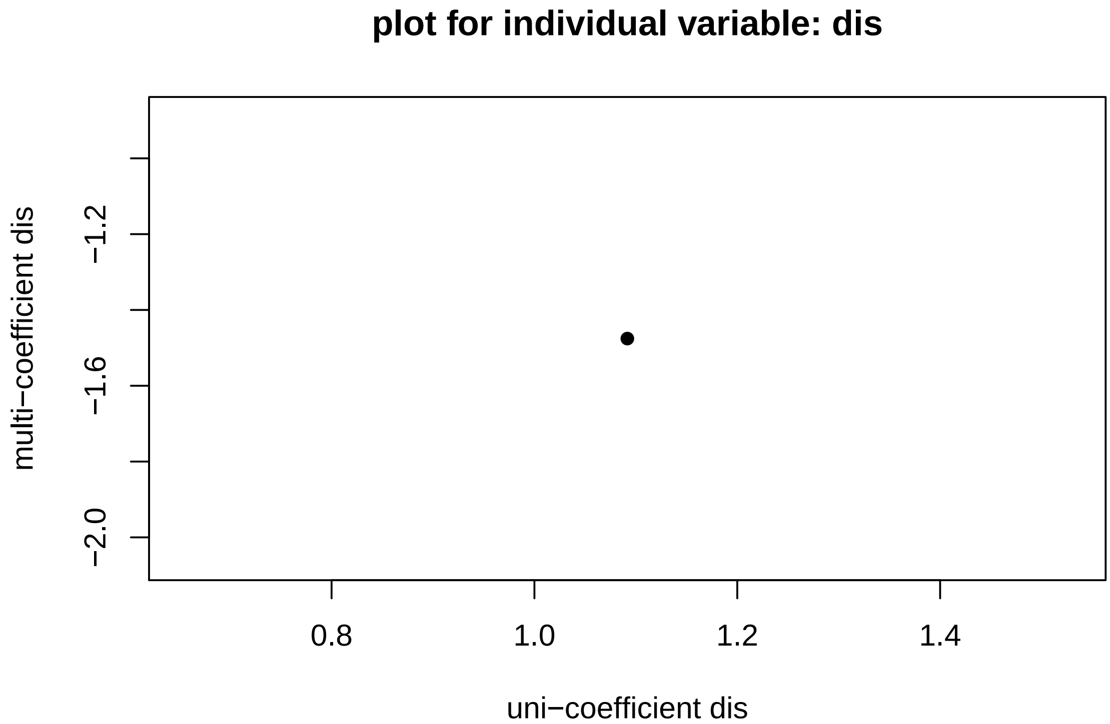

```
## [1] "varation"
## Boston$nox
##  -16.14944
```



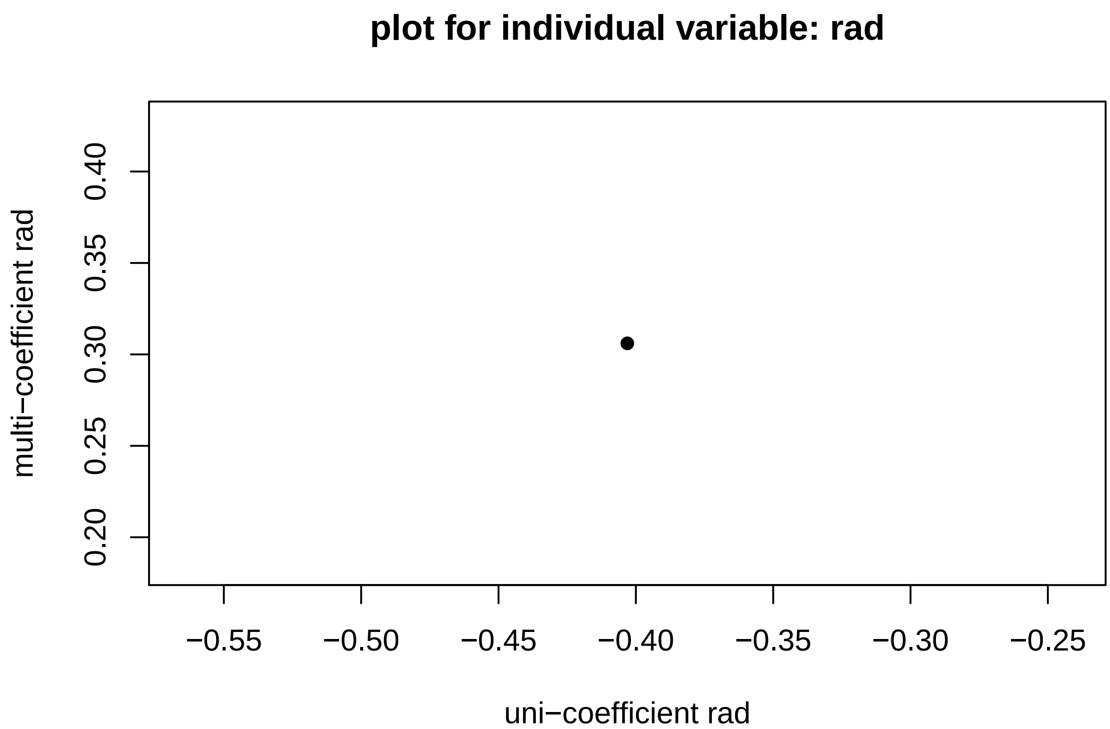

```
## [1] "varation"
## Boston$rm
##  5.292244
```



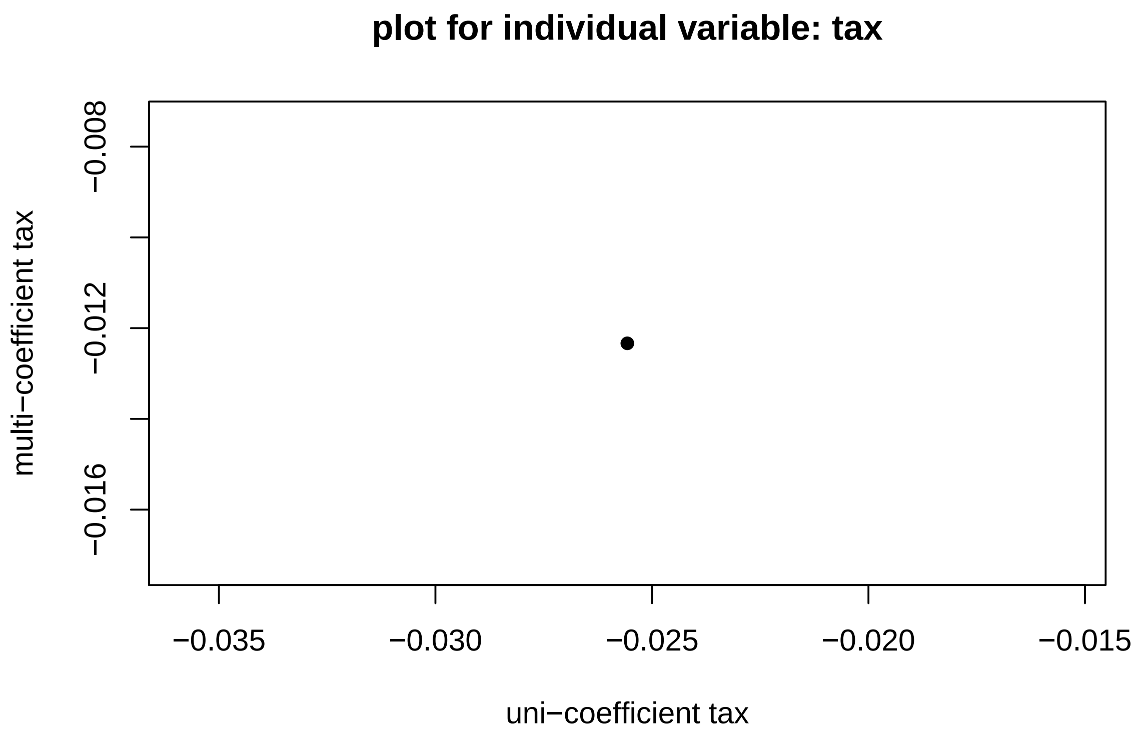

```
## [1] "varation"
## Boston$age
## -0.1238549
```



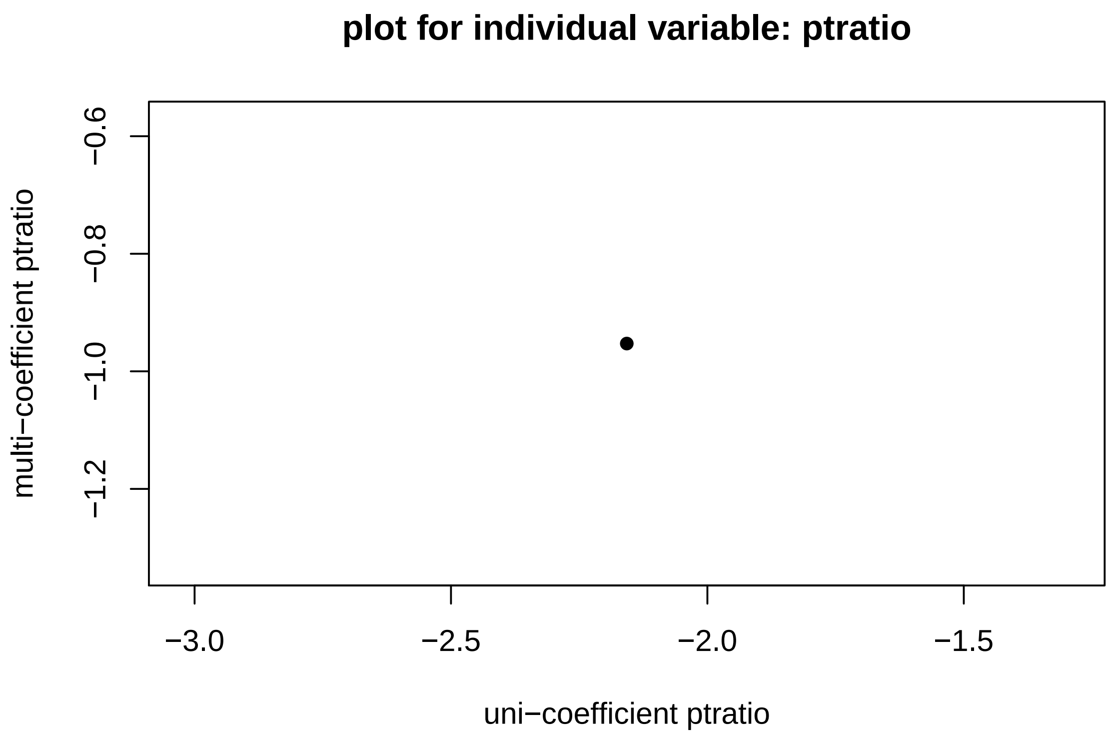

```
## [1] "varation"
## Boston$dis
##    2.56718
```



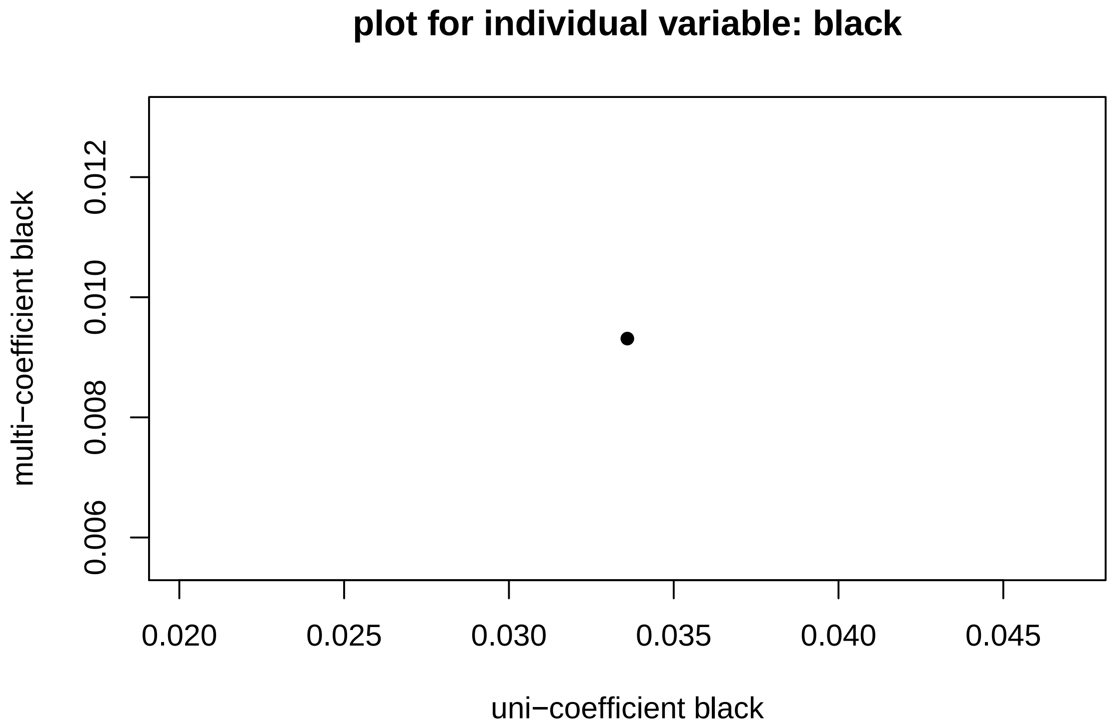

```
## [1] "varation"
## Boston$rad
## -0.7091449
```



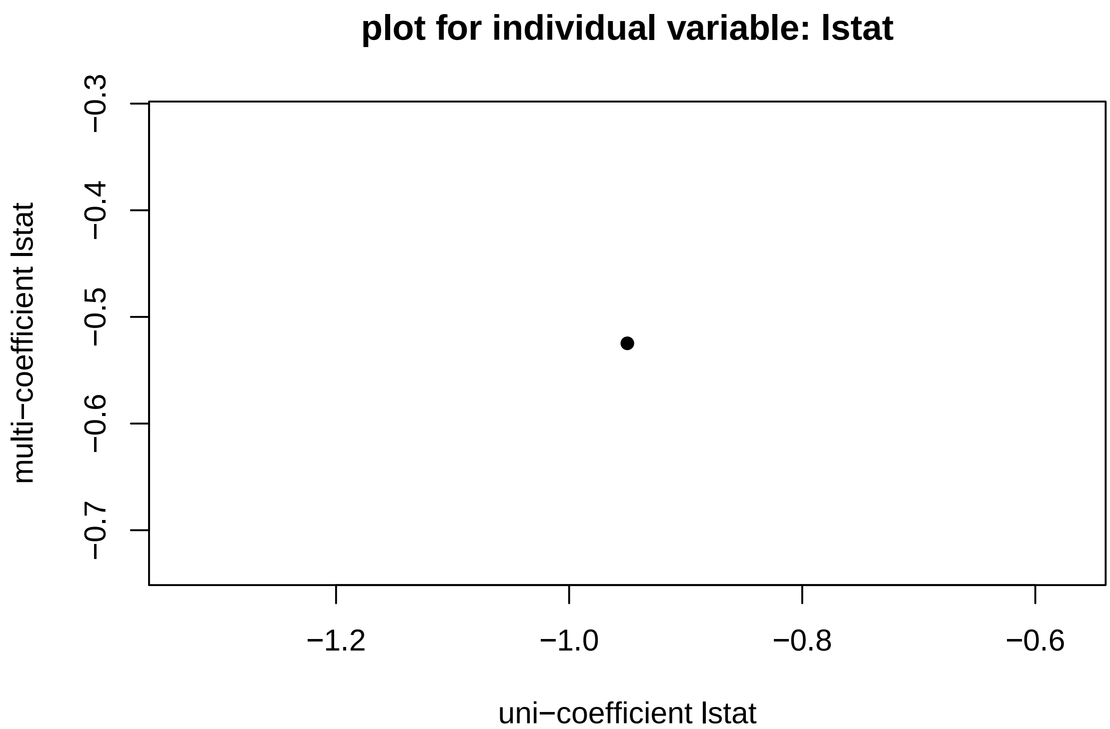

```
## [1] "varation"
##  Boston$tax
## -0.01323351
```




```
## [1] "varation"
## Boston$ptratio
##      -1.204428
```




```
## [1] "varation"
## Boston$black
##   0.02428138
```




```
## [1] "varation"
## Boston$lstat
##    -0.425291
```

The data show that a number of the predictors have greatly changed when calculating them as unique coefficients compare to when calculating them as part of the multi coefficient group.

- crim predictor changed from -0.4151903 as unique coefficient to -0.1080114 when being part of the multi coefficient group. It's a variation of -0.307
- zn predictor changed from 0.14214 as unique coefficient to 0.04642046 when being part of the multi coefficient group. It's a variation of +0.0957
- indus predictor changed from -0.6484901 as unique coefficient to 0.02055863 when being part of the multi coefficient group. It's a variation of -0.669
- chas predictor changed from +6.346157 as unique coefficient to 2.686734 when being part of the multi coefficient group. It's a variation of 3.659
- nox predictor changed from -33.91606 as unique coefficient to -17.76661 when being part of the multi coefficient group. It's a variation of -16.149
- rm predictor changed from +9.102109 as unique coefficient to 3.809865 when being part of the multi coefficient group. It's a variation of +5.29
- age predictor changed from -0.1231627 as unique coefficient to 0.0006922246 when being part of the multi coefficient group. It's a variation of -0.123
- dis predictor changed from +1.091613 as unique coefficient to -1.475567 when being part of the multi coefficient group. It's a variation of 2.567
- rad predictor changed from -0.4030954 as unique coefficient to 0.3060495 when being part of the multi coefficient group. It's a variation of -0.709
- tax predictor changed from -0.0255681 as unique coefficient to -0.01233459 when being part of the multi coefficient group. It's a variation of -0.0132

- ptratio predictor changed from -2.157175 as unique coefficient to -0.9527472 when being part of the multi coefficient group. It's a variation of -1.204
- black predictor changed from +0.03359306 as unique coefficient to 0.009311683 when being part of the multi coefficient group. It's a variation of 0.024
- lstat predictor changed from -0.9500494 as unique coefficient to -0.5247584 when being part of the multi coefficient group. It's a variation of 0.42

6. Is there evidence of a 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:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

```
#generating the polynomial result of this model at the 3rd degree
fit.lstat.3b <- lm(Boston$medv ~ poly(Boston$lstat, 3, raw=TRUE))
summary(fit.lstat.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$lstat, 3, raw = TRUE))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5441  -3.7122  -0.5145   2.4846  26.4153
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        48.6496253  1.4347240  33.909  < 2e-16
## poly(Boston$lstat, 3, raw = TRUE)1 -3.8655928  0.3287861 -11.757  < 2e-16
## poly(Boston$lstat, 3, raw = TRUE)2  0.1487385  0.0212987   6.983 9.18e-12
## poly(Boston$lstat, 3, raw = TRUE)3 -0.0020039  0.0003997  -5.013 7.43e-07
##
## (Intercept)                        ***
## poly(Boston$lstat, 3, raw = TRUE)1 ***
## poly(Boston$lstat, 3, raw = TRUE)2 ***
## poly(Boston$lstat, 3, raw = TRUE)3 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.396 on 502 degrees of freedom
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.6558
## F-statistic: 321.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 48.649625-3.8655928x + 0.1487385x^2 -0.0020039x^3

All the coefficients are significant, thus, the best model is the polynomial of 3rd degree.

```
#generating the polynomial result of this model at the 3rd degree
fit.crim.3b <- lm(Boston$medv ~ poly(Boston$crim, 3, raw=TRUE))
summary(fit.crim.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$crim, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.983  -4.975  -1.940   2.881  33.391
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        2.519e+01  4.355e-01  57.846  < 2e-16
## poly(Boston$crim, 3, raw = TRUE)1 -1.136e+00  1.444e-01  -7.868 2.24e-14
## poly(Boston$crim, 3, raw = TRUE)2  2.378e-02  6.808e-03   3.494 0.000518
## poly(Boston$crim, 3, raw = TRUE)3 -1.489e-04  6.641e-05  -2.242 0.025411
##
## (Intercept)                       ***
## poly(Boston$crim, 3, raw = TRUE)1 ***
## poly(Boston$crim, 3, raw = TRUE)2 ***
## poly(Boston$crim, 3, raw = TRUE)3 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.159 on 502 degrees of freedom
## Multiple R-squared:  0.2177, Adjusted R-squared:  0.213
## F-statistic: 46.57 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 2.519e+01 - 1.136e+00x + 2.378e-02x^2 - 1.489e-04x^3

The coefficient beta3 is not significant, thus, the best model is the polynomial of 2nd degree.

```r
#generating the polynomial result of this model at the 3rd degree
fit.age.3b <- lm(Boston$medv ~ poly(Boston$age, 3, raw=TRUE))
summary(fit.age.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$age, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.443  -4.909  -2.234   2.185  32.944
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.893e+01  2.992e+00   9.668   <2e-16
## poly(Boston$age, 3, raw = TRUE)1 -1.224e-01  2.014e-01  -0.608    0.544
## poly(Boston$age, 3, raw = TRUE)2  2.355e-03  3.930e-03   0.599    0.549
## poly(Boston$age, 3, raw = TRUE)3 -2.318e-05  2.279e-05  -1.017    0.310
##
## (Intercept)                      ***
## poly(Boston$age, 3, raw = TRUE)1
## poly(Boston$age, 3, raw = TRUE)2
```

```
## poly(Boston$age, 3, raw = TRUE)3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.472 on 502 degrees of freedom
## Multiple R-squared:  0.1566, Adjusted R-squared:  0.1515
## F-statistic: 31.06 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 2.893e+01 -1.224e-0x + 2.355e-03x^2 + 0.0001257x^3

All the coefficients are significant, thus, the best model is the polynomial of 3rd degree.

```r
#generating the polynomial result of this model at the 3rd degree
fit.zn.3b <- lm(Boston$medv ~ poly(Boston$zn, 3, raw=TRUE))
summary(fit.zn.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$zn, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.449  -5.549  -1.049   3.225  29.551
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     20.4485972  0.4359536  46.905  < 2e-16 ***
## poly(Boston$zn, 3, raw = TRUE)1  0.6433652  0.1105611   5.819 1.06e-08 ***
## poly(Boston$zn, 3, raw = TRUE)2 -0.0167646  0.0038872  -4.313 1.94e-05 ***
## poly(Boston$zn, 3, raw = TRUE)3  0.0001257  0.0000316   3.978 7.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.43 on 502 degrees of freedom
## Multiple R-squared:  0.1649, Adjusted R-squared:  0.1599
## F-statistic: 33.05 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 20.4485972 + 0.6433652x - 0.0167646x^2 + 0.0001257x^3

All the coefficients are significant, thus, the best model is the polynomial of 3rd degree.

```r
#generating the polynomial result of this model at the 3rd degree
fit.indus.3b <- lm(Boston$medv ~ poly(Boston$indus, 3, raw=TRUE))
summary(fit.indus.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$indus, 3, raw = TRUE))
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -15.760  -4.725  -1.009   2.932  32.038 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    37.080160   1.663326  22.293  < 2e-16
## poly(Boston$indus, 3, raw = TRUE)1 -2.806994   0.509349  -5.511 5.71e-08
## poly(Boston$indus, 3, raw = TRUE)2  0.140462   0.041554   3.380 0.000781
## poly(Boston$indus, 3, raw = TRUE)3 -0.002399   0.001011  -2.373 0.018026
##                                    
## (Intercept)                        ***
## poly(Boston$indus, 3, raw = TRUE)1 ***
## poly(Boston$indus, 3, raw = TRUE)2 ***
## poly(Boston$indus, 3, raw = TRUE)3 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.844 on 502 degrees of freedom
## Multiple R-squared:  0.2768, Adjusted R-squared:  0.2725 
## F-statistic: 64.06 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 37.080160 - 2.806994x + 0.140462x^2 - 0.002399x^3

The coefficient beta3 is not significant, thus, the best model is the polynomial of 2nd degree.

```
#generating the polynomial result of this model at the 3rd degree
fit.nox.3b <- lm(Boston$medv ~ poly(Boston$nox, 3, raw=TRUE))
summary(fit.nox.3b)
```

```
## 
## Call:
## lm(formula = Boston$medv ~ poly(Boston$nox, 3, raw = TRUE))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -13.104  -5.020  -2.144   2.747  32.416 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                     -22.49      38.52  -0.584   0.5596  
## poly(Boston$nox, 3, raw = TRUE)1   315.10     195.10   1.615   0.1069  
## poly(Boston$nox, 3, raw = TRUE)2  -615.83     320.48  -1.922   0.0552 .
## poly(Boston$nox, 3, raw = TRUE)3   350.19     170.92   2.049   0.0410 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.282 on 502 degrees of freedom
## Multiple R-squared:  0.1939, Adjusted R-squared:  0.189 
## F-statistic: 40.24 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = -22.49 + 315.10x - 615.83x^2 + 350.19x^3

The coefficients are not significant, thus, there is not an option to come up with the best model

```
#generating the polynomial result of this model at the 3rd degree
fit.rm.3b <- lm(Boston$medv ~ poly(Boston$rm, 3, raw=TRUE))
summary(fit.rm.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$rm, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.102  -2.674   0.569   3.011  35.911
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      241.3108    47.3275   5.099 4.85e-07 ***
## poly(Boston$rm, 3, raw = TRUE)1 -109.3906    22.9690  -4.763 2.51e-06 ***
## poly(Boston$rm, 3, raw = TRUE)2   16.4910     3.6750   4.487 8.95e-06 ***
## poly(Boston$rm, 3, raw = TRUE)3   -0.7404     0.1935  -3.827 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.11 on 502 degrees of freedom
## Multiple R-squared:  0.5612, Adjusted R-squared:  0.5586
## F-statistic:   214 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 241.3108 - 109.3906x + 16.4910x^2 - 0.7404x^3

All the coefficients are significant, thus, the best model is the polynomial of 3rd degree.

```
#generating the polynomial result of this model at the 3rd degree
fit.dis.3b <- lm(Boston$medv ~ poly(Boston$dis, 3, raw=TRUE))
summary(fit.dis.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$dis, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.571  -5.242  -2.037   2.397  34.769
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       7.03789    2.91134   2.417  0.01599 *
## poly(Boston$dis, 3, raw = TRUE)1  8.59284    2.06633   4.158 3.77e-05 ***
## poly(Boston$dis, 3, raw = TRUE)2 -1.24953    0.41235  -3.030  0.00257 **
## poly(Boston$dis, 3, raw = TRUE)3  0.05602    0.02428   2.307  0.02146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.727 on 502 degrees of freedom
## Multiple R-squared:  0.105,  Adjusted R-squared:  0.09968
## F-statistic: 19.64 on 3 and 502 DF,  p-value: 4.736e-12
```

Model

Y = 7.03789 + 8.59284x - 1.24953x^2 + 0.05602x^3

The coefficients intercept, and beta3 are not significant, thus, the best model is the polynomial of 2nd degree.

```
#generating the polynomial result of this model at the 3rd degree
fit.rad.3b <- lm(Boston$medv ~ poly(Boston$rad, 3, raw=TRUE))
summary(fit.rad.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$rad, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.630  -5.151  -2.017   3.169  33.594
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      30.251303   2.567860  11.781  < 2e-16 ***
## poly(Boston$rad, 3, raw = TRUE)1 -3.799454   1.307156  -2.907 0.003815 **
## poly(Boston$rad, 3, raw = TRUE)2  0.616347   0.186057   3.313 0.000991 ***
## poly(Boston$rad, 3, raw = TRUE)3 -0.020086   0.005717  -3.514 0.000482 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.37 on 502 degrees of freedom
## Multiple R-squared:  0.1767, Adjusted R-squared:  0.1718
## F-statistic: 35.91 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 30.251303 - 3.799454x + 0.616347x^2 - 0.020086x^3

All the coefficients are significant, thus, the best model is the polynomial of 3rd degree.

```
#generating the polynomial result of this model at the 3rd degree
fit.tax.3b <- lm(Boston$medv ~ poly(Boston$tax, 3, raw=TRUE))
summary(fit.tax.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$tax, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.109  -4.952  -1.878   2.957  33.694
##
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.222e+01  1.397e+01   3.739 0.000206
## poly(Boston$tax, 3, raw = TRUE)1 -1.635e-01  1.133e-01  -1.443 0.149646
## poly(Boston$tax, 3, raw = TRUE)2  3.029e-04  2.872e-04   1.055 0.292004
## poly(Boston$tax, 3, raw = TRUE)3 -2.079e-07  2.236e-07  -0.930 0.353061
##
## (Intercept)                     ***
## poly(Boston$tax, 3, raw = TRUE)1
## poly(Boston$tax, 3, raw = TRUE)2
## poly(Boston$tax, 3, raw = TRUE)3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.115 on 502 degrees of freedom
## Multiple R-squared:  0.2261, Adjusted R-squared:  0.2215
## F-statistic: 48.89 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 5.222e+01 - 1.635e-01x + 3.029e-04x^2 - 2.079e-07x^3

The coefficients beta1, beta2, and beta3 are not significant, thus, the best model is the polynomial of intercept.

```r
#generating the polynomial result of this model at the 3rd degree
fit.ptRatio.3b <- lm(Boston$medv ~ poly(Boston$ptratio, 3, raw=TRUE))
summary(fit.ptRatio.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$ptratio, 3, raw = TRUE))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7795  -5.0364  -0.9778   3.4766  31.1636
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         312.28642  152.48693   2.048   0.0411
## poly(Boston$ptratio, 3, raw = TRUE)1 -48.69114   26.88441  -1.811   0.0707
## poly(Boston$ptratio, 3, raw = TRUE)2   2.83995    1.56413   1.816   0.0700
## poly(Boston$ptratio, 3, raw = TRUE)3  -0.05686    0.03005  -1.892   0.0590
##
## (Intercept)                         *
## poly(Boston$ptratio, 3, raw = TRUE)1 .
## poly(Boston$ptratio, 3, raw = TRUE)2 .
## poly(Boston$ptratio, 3, raw = TRUE)3 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.898 on 502 degrees of freedom
## Multiple R-squared:  0.2669, Adjusted R-squared:  0.2625
## F-statistic: 60.91 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model

Y = 312.28642 - 48.69114x + 2.83995x^2 - 0.05686x^3

The coefficients are not significant,thus, there is not an option to come up with the best model.

```r
#generating the polynomial result of this model at the 3rd degree
fit.black.3b <- lm(Boston$medv ~ poly(Boston$black, 3, raw=TRUE))
summary(fit.black.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$black, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.005  -4.802  -1.613   2.852  28.051
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1.260e+01  2.517e+00   5.006  7.7e-07
## poly(Boston$black, 3, raw = TRUE)1 -1.703e-02  6.150e-02  -0.277    0.782
## poly(Boston$black, 3, raw = TRUE)2  2.036e-04  3.258e-04   0.625    0.532
## poly(Boston$black, 3, raw = TRUE)3 -2.224e-07  4.765e-07  -0.467    0.641
##
## (Intercept)                        ***
## poly(Boston$black, 3, raw = TRUE)1
## poly(Boston$black, 3, raw = TRUE)2
## poly(Boston$black, 3, raw = TRUE)3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.685 on 502 degrees of freedom
## Multiple R-squared:  0.1135, Adjusted R-squared:  0.1082
## F-statistic: 21.43 on 3 and 502 DF,  p-value: 4.463e-13
```

Model

Y = 1.260e+01 - 1.703e-02x + 2.036e-04x^2 - 2.224e-07x^3

The coefficients beta1, beta2, and beta3 are not significant,thus, the best model is the polynomial of intercept.

```r
fit.chas.3b <- lm(Boston$medv ~ poly(Boston$chas, 3, raw=TRUE))
summary(fit.chas.3b)
```

```
##
## Call:
## lm(formula = Boston$medv ~ poly(Boston$chas, 3, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
```

```
## Coefficients: (2 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     22.0938     0.4176  52.902  < 2e-16 ***
## poly(Boston$chas, 3, raw = TRUE)1   6.3462     1.5880   3.996 7.39e-05 ***
## poly(Boston$chas, 3, raw = TRUE)2       NA         NA      NA       NA
## poly(Boston$chas, 3, raw = TRUE)3       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.064 on 504 degrees of freedom
## Multiple R-squared:  0.03072,    Adjusted R-squared:  0.02879
## F-statistic: 15.97 on 1 and 504 DF,  p-value: 7.391e-05
```

Model

chas contains binary data (0 and 1), thus, it is not efficient to generate a polynomial.

7. Consider performing a stepwise model selection procedure to determine the bets fit model. Discuss your results. How is this model different from the model in (4)?

```r
#generating fit model using step function
step.fit = step(multi.lm, direction="both")
```

```
## Start:  AIC=1589.64
## medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + black + lstat
##
##           Df Sum of Sq   RSS    AIC
## - age      1      0.06 11079 1587.7
## - indus    1      2.52 11081 1587.8
## <none>                 11079 1589.6
## - chas     1    218.97 11298 1597.5
## - tax      1    242.26 11321 1598.6
## - crim     1    243.22 11322 1598.6
## - zn       1    257.49 11336 1599.3
## - black    1    270.63 11349 1599.8
## - rad      1    479.15 11558 1609.1
## - nox      1    487.16 11566 1609.4
## - ptratio  1   1194.23 12273 1639.4
## - dis      1   1232.41 12311 1641.0
## - rm       1   1871.32 12950 1666.6
## - lstat    1   2410.84 13490 1687.3
##
## Step:  AIC=1587.65
## medv ~ crim + zn + indus + chas + nox + rm + dis + rad + tax +
##     ptratio + black + lstat
##
##           Df Sum of Sq   RSS    AIC
## - indus    1      2.52 11081 1585.8
## <none>                 11079 1587.7
## + age      1      0.06 11079 1589.6
## - chas     1    219.91 11299 1595.6
## - tax      1    242.24 11321 1596.6
## - crim     1    243.20 11322 1596.6
```

```
## - zn       1    260.32 11339 1597.4
## - black    1    272.26 11351 1597.9
## - rad      1    481.09 11560 1607.2
## - nox      1    520.87 11600 1608.9
## - ptratio  1   1200.23 12279 1637.7
## - dis      1   1352.26 12431 1643.9
## - rm       1   1959.55 13038 1668.0
## - lstat    1   2718.88 13798 1696.7
##
## Step:  AIC=1585.76
## medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
##     black + lstat
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## + indus    1      2.52 11079 1587.7
## + age      1      0.06 11081 1587.8
## - chas     1    227.21 11309 1594.0
## - crim     1    245.37 11327 1594.8
## - zn       1    257.82 11339 1595.4
## - black    1    270.82 11352 1596.0
## - tax      1    273.62 11355 1596.1
## - rad      1    500.92 11582 1606.1
## - nox      1    541.91 11623 1607.9
## - ptratio  1   1206.45 12288 1636.0
## - dis      1   1448.94 12530 1645.9
## - rm       1   1963.66 13045 1666.3
## - lstat    1   2723.48 13805 1695.0
```

```
step.fit
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
##     tax + ptratio + black + lstat, data = Boston)
##
## Coefficients:
## (Intercept)         crim           zn         chas          nox
##   36.341145    -0.108413     0.045845     2.718716   -17.376023
##          rm          dis          rad          tax      ptratio
##    3.801579    -1.492711     0.299608    -0.011778    -0.946525
##       black        lstat
##    0.009291    -0.522553
```

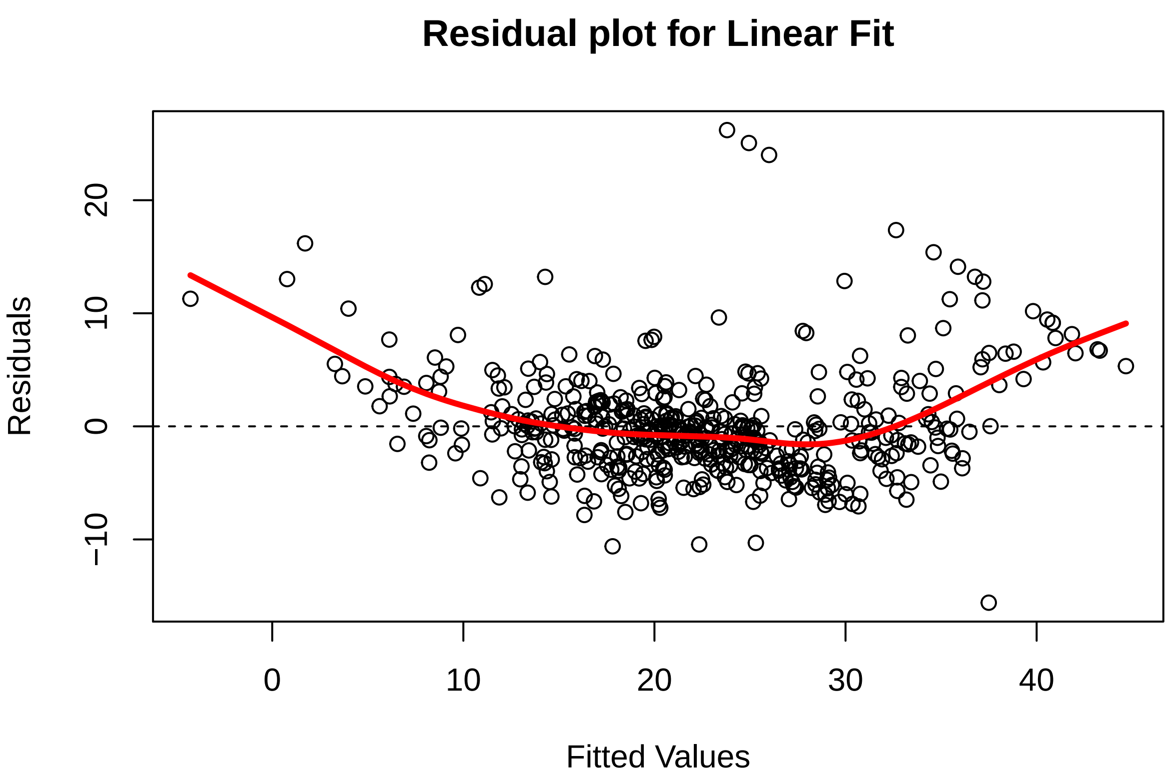
When looking at the results generated by the step function, we want to identify the section that has the lowest AIC. In this model, the step result with AIC = 1585.76 will be the result that we will analyze as this has the lowest AIC. This result shows that according to the step function, the best model is the one that includes the following variables: crim, zn, chas, nox, rm, dis, rad, tax, ptratio, black, and lstat. This result means that indus, and age variables should be excluded from generating the best model for this data set. If either indus and age were to be included in the model, AIC will be at 1587.7 and 1587.8 respectively. Thus, these AIC values will be above the overall AIC 1585.76.

The result generated from the step function matches the conclusion found in question 4 where predictors indus, and age were found to have a significant value greater than 0.005. Thus, they were both rejected.

In question 4 we use the p-value to identify the predictors that will fit our model. On the other hand, the step model uses the F-statistics to select predictors that would satisfy the best fit model. Both models identify the same variables to be removed (indus, age) so that we can have the best fit model.

8. Evaluate the statistical assumptions in your regression analysis from (7) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

```
#generating a residual vs fitted values graph which also contains zero horizontal line and smoothing li
plot(fitted(multi.lm), residuals(multi.lm), xlab = "Fitted Values", ylab = "Residuals",
     main="Residual plot for Linear Fit")
  abline(h=0, lty=2)
  lines(smooth.spline(fitted(multi.lm), residuals(multi.lm)), col="red", lwd=3)
```



Ideally, the smoothing line should be approximately straight and horizontal around zero. Basically it should overlay the horizontal zero line. The fact that we are seeing a non-linearity is an indication that there are some predictors that do not have a relationship with the response. However, the non-linearity is not big enough to invalidate the model.

As identified in the previous section, both the indus and the age predictors do not seem to satisfy the best fit model based on the p-value and the AIC results. We could assume that the removal of these two variables in the model would result to a residual that has a more linear smooth line that would overlay over the horizontal line.