# IMT 573: Problem Set 6 - Regression

*Naga Soundari Balamurugan*

*Due: Tuesday, November 13, 2018*

**Collaborators: Jayashree Raman**

**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. If you are using more than just a standard function that you found from another source, please credit the source in the comments. For example:
4. Collaboration on problem sets is acceptable, and even encouraged, but students must turn in an individual write-up in their own words and their 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` or `Knit Word`, rename the R Markdown file to `YourLastName_YourFirstName_ps6.Rmd`, knit a PDF or DOC and submit both the PDF/DOC and the Rmd 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
library(kableExtra)
library(ISLR)
library(leaps)
library(ggplot2)
```

**Housing Values in Suburbs of Boston**

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.**

```
#Read in Boston housing data
Boston_data <- MASS::Boston
```

```
str(Boston_data)
```

```
## '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 ...
```

```
no_of_rows <- nrow(Boston_data)
no_of_cols <- ncol(Boston_data)

#Rename the column headers for more readability
colnames(Boston_data) <- c("CrimeRate", "BigLots_Proportion", "Business_Proportion",
                           "CharlesRiver", "NO_Concentration", "Avg_Num_rooms", "Owner_Prop",
                           "Employ_Distance", "Highway_Access", "Taxrate", "Teacher_Ratio",
                           "Black_Proportion", "Lower_Status", "Median_Owner")
```

The Boston data frame has 506 rows and 14 columns. This data frame contains the following columns:

CrimeRate - per capita crime rate by town. BigLots_Proportion - proportion of residential land zoned for lots over 25,000 sq.ft. Business_Proportion - proportion of non-retail business acres per town. CharlesRiver - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). NO_Concentration - nitrogen oxides concentration (parts per 10 million). Avg_Num_rooms - average number of rooms per dwelling. Owner_Prop - proportion of owner-occupied units built prior to 1940. Employ_Distance - weighted mean of distances to five Boston employment centres. Highway_Access - index of accessibility to radial highways. Taxrate - full-value property-tax rate per $10,000. Teacher_Ratio - pupil-teacher ratio by town. Black_Proportion - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town. Lower_Status - lower status of the population (percent). Median_Owner - median value of owner-occupied homes in $1000s.

**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.**

```
#Find the correlation between each variables
corr_Matrix <- cor(Boston_data)

corr_Matrix
```

```
##                      CrimeRate BigLots_Proportion Business_Proportion
## CrimeRate            1.00000000        -0.20046922          0.40658341
## BigLots_Proportion  -0.20046922         1.00000000         -0.53382819
## Business_Proportion  0.40658341        -0.53382819          1.00000000
## CharlesRiver        -0.05589158        -0.04269672          0.06293803
## NO_Concentration     0.42097171        -0.51660371          0.76365145
## Avg_Num_rooms       -0.21924670         0.31199059         -0.39167585
```

```
## Owner_Prop           0.35273425        -0.56953734          0.64477851
## Employ_Distance     -0.37967009         0.66440822         -0.70802699
## Highway_Access       0.62550515        -0.31194783          0.59512927
## Taxrate              0.58276431        -0.31456332          0.72076018
## Teacher_Ratio        0.28994558        -0.39167855          0.38324756
## Black_Proportion    -0.38506394         0.17552032         -0.35697654
## Lower_Status         0.45562148        -0.41299457          0.60379972
## Median_Owner        -0.38830461         0.36044534         -0.48372516
##                     CharlesRiver NO_Concentration Avg_Num_rooms
## CrimeRate           -0.055891582       0.42097171   -0.21924670
## BigLots_Proportion  -0.042696719      -0.51660371    0.31199059
## Business_Proportion  0.062938027       0.76365145   -0.39167585
## CharlesRiver         1.000000000       0.09120281    0.09125123
## NO_Concentration     0.091202807       1.00000000   -0.30218819
## Avg_Num_rooms        0.091251225      -0.30218819    1.00000000
## Owner_Prop           0.086517774       0.73147010   -0.24026493
## Employ_Distance     -0.099175780      -0.76923011    0.20524621
## Highway_Access      -0.007368241       0.61144056   -0.20984667
## Taxrate             -0.035586518       0.66802320   -0.29204783
## Teacher_Ratio       -0.121515174       0.18893268   -0.35550149
## Black_Proportion     0.048788485      -0.38005064    0.12806864
## Lower_Status        -0.053929298       0.59087892   -0.61380827
## Median_Owner         0.175260177      -0.42732077    0.69535995
##                      Owner_Prop Employ_Distance Highway_Access     Taxrate
## CrimeRate            0.35273425     -0.37967009    0.625505145  0.58276431
## BigLots_Proportion  -0.56953734      0.66440822   -0.311947826 -0.31456332
## Business_Proportion  0.64477851     -0.70802699    0.595129275  0.72076018
## CharlesRiver         0.08651777     -0.09917578   -0.007368241 -0.03558652
## NO_Concentration     0.73147010     -0.76923011    0.611440563  0.66802320
## Avg_Num_rooms       -0.24026493      0.20524621   -0.209846668 -0.29204783
## Owner_Prop           1.00000000     -0.74788054    0.456022452  0.50645559
## Employ_Distance     -0.74788054      1.00000000   -0.494587930 -0.53443158
## Highway_Access       0.45602245     -0.49458793    1.000000000  0.91022819
## Taxrate              0.50645559     -0.53443158    0.910228189  1.00000000
## Teacher_Ratio        0.26151501     -0.23247054    0.464741179  0.46085304
## Black_Proportion    -0.27353398      0.29151167   -0.444412816 -0.44180801
## Lower_Status         0.60233853     -0.49699583    0.488676335  0.54399341
## Median_Owner        -0.37695457      0.24992873   -0.381626231 -0.46853593
##                     Teacher_Ratio Black_Proportion Lower_Status
## CrimeRate               0.2899456      -0.38506394    0.4556215
## BigLots_Proportion     -0.3916785       0.17552032   -0.4129946
## Business_Proportion     0.3832476      -0.35697654    0.6037997
## CharlesRiver           -0.1215152       0.04878848   -0.0539293
## NO_Concentration        0.1889327      -0.38005064    0.5908789
## Avg_Num_rooms          -0.3555015       0.12806864   -0.6138083
## Owner_Prop              0.2615150      -0.27353398    0.6023385
## Employ_Distance        -0.2324705       0.29151167   -0.4969958
## Highway_Access          0.4647412      -0.44441282    0.4886763
## Taxrate                 0.4608530      -0.44180801    0.5439934
## Teacher_Ratio           1.0000000      -0.17738330    0.3740443
## Black_Proportion       -0.1773833       1.00000000   -0.3660869
## Lower_Status            0.3740443      -0.36608690    1.0000000
## Median_Owner           -0.5077867       0.33346082   -0.7376627
##                     Median_Owner
```

```
## CrimeRate             -0.3883046
## BigLots_Proportion     0.3604453
## Business_Proportion   -0.4837252
## CharlesRiver           0.1752602
## NO_Concentration      -0.4273208
## Avg_Num_rooms          0.6953599
## Owner_Prop            -0.3769546
## Employ_Distance        0.2499287
## Highway_Access        -0.3816262
## Taxrate               -0.4685359
## Teacher_Ratio         -0.5077867
## Black_Proportion       0.3334608
## Lower_Status          -0.7376627
## Median_Owner           1.0000000
```

```
# #Displays the correlation matrix
# kable(corr_Matrix, "latex") %>% kable_styling(bootstrap_options =
#                                                 c("striped", "hover", "scale_down"))
```

Out of the above data, **CrimeRate is the response variable**. At a first glance, every variable except the Charles river dummy variable, looks like a predictor variable of the crime rate. After a deeper look, I expected the predictor variables to be Teacher_Ratio, BigLots_Proportion, Black_Proportion etc., The reason behind is a lower Teacher-student ratio might lead to lesser educational status and higher crime rate. Also plenty of bigger lots could to lesser security and high chance of crime. But after running the correlation test, the possible predictor variables(ones with comparatively high correlation rate) could be Business_Proportion, NO_Concentration, Owner_Prop, Employ_Distance, Highway_Access, Taxrate, Black_Proportion, Lower_Status, Median_Owner.

**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.**

```
#Linear regression model for Crime Rate vs BigLots_Proportion
lm_BigLots_Proportion <- lm(CrimeRate ~ BigLots_Proportion, data = Boston_data)
summary(lm_BigLots_Proportion)
```

```
##
## Call:
## lm(formula = CrimeRate ~ BigLots_Proportion, data = Boston_data)
##
## 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 ***
## BigLots_Proportion -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
```

```
#Linear regression model for Crime Rate vs Business_Proportion
lm_Business_Proportion <- lm(CrimeRate ~ Business_Proportion, data = Boston_data)
summary(lm_Business_Proportion)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Business_Proportion, data = Boston_data)
##
## 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 **
## Business_Proportion  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
```

```
#Linear regression model for Crime Rate vs CharlesRiver
lm_CharlesRiver <- lm(CrimeRate ~ CharlesRiver, data = Boston_data)
summary(lm_CharlesRiver)
```

```
##
## Call:
## lm(formula = CrimeRate ~ CharlesRiver, data = Boston_data)
##
## 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 ***
## CharlesRiver  -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
```

```
#Linear regression model for Crime Rate vs Nitrogen oxide concentration
lm_NOX_Con <- lm(CrimeRate ~ NO_Concentration, data = Boston_data)
summary(lm_NOX_Con)
```

```
##
## Call:
## lm(formula = CrimeRate ~ NO_Concentration, data = Boston_data)
##
```

```
## 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 ***
## NO_Concentration   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
```

```
#Linear regression model for Crime Rate vs Avg_Num_rooms
lm_Avg_Num_rooms <- lm(CrimeRate ~ Avg_Num_rooms, data = Boston_data)
summary(lm_Avg_Num_rooms)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Avg_Num_rooms, data = Boston_data)
##
## 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 ***
## Avg_Num_rooms   -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
```

```
#Linear regression model for Crime Rate vs Owner_Prop
lm_Owner_Prop <- lm(CrimeRate ~ Owner_Prop, data = Boston_data)
summary(lm_Owner_Prop)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Owner_Prop, data = Boston_data)
##
## 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 ***
## Owner_Prop   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
```

```
#Linear regression model for Crime Rate vs Employ_Distance
lm_Employ_Distance <- lm(CrimeRate ~ Employ_Distance, data = Boston_data)
summary(lm_Employ_Distance)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Employ_Distance, data = Boston_data)
##
## 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 ***
## Employ_Distance  -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
```

```
#Linear regression model for Crime Rate vs Highway_Access
lm_Highway_Access <- lm(CrimeRate ~ Highway_Access, data = Boston_data)
summary(lm_Highway_Access)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Highway_Access, data = Boston_data)
##
## 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 ***
## Highway_Access  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
```

```
#Linear regression model for Crime Rate vs Taxrate
lm_Taxrate <- lm(CrimeRate ~ Taxrate, data = Boston_data)
summary(lm_Taxrate)
```

```
##
```

```
## Call:
## lm(formula = CrimeRate ~ Taxrate, data = Boston_data)
##
## 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 ***
## Taxrate      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
```

```
#Linear regression model for Crime Rate vs Teacher_Ratio
lm_Teacher_Ratio <- lm(CrimeRate ~ Teacher_Ratio, data = Boston_data)
summary(lm_Teacher_Ratio)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Teacher_Ratio, data = Boston_data)
##
## 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 ***
## Teacher_Ratio   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
```

```
#Linear regression model for Crime Rate vs Black_Proportion
lm_Black_Proportion <- lm(CrimeRate ~ Black_Proportion, data = Boston_data)
summary(lm_Black_Proportion)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Black_Proportion, data = Boston_data)
##
## 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 ***
```

```
## Black_Proportion -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
```

```
#Linear regression model for Crime Rate vs Lower_Status
lm_Lower_Status <- lm(CrimeRate ~ Lower_Status, data = Boston_data)
summary(lm_Lower_Status)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Lower_Status, data = Boston_data)
##
## 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 ***
## Lower_Status  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
```

```
#Linear regression model for Crime Rate vs Median_Owner
lm_Median_Owner<- lm(CrimeRate ~ Median_Owner, data = Boston_data)
summary(lm_Median_Owner)
```

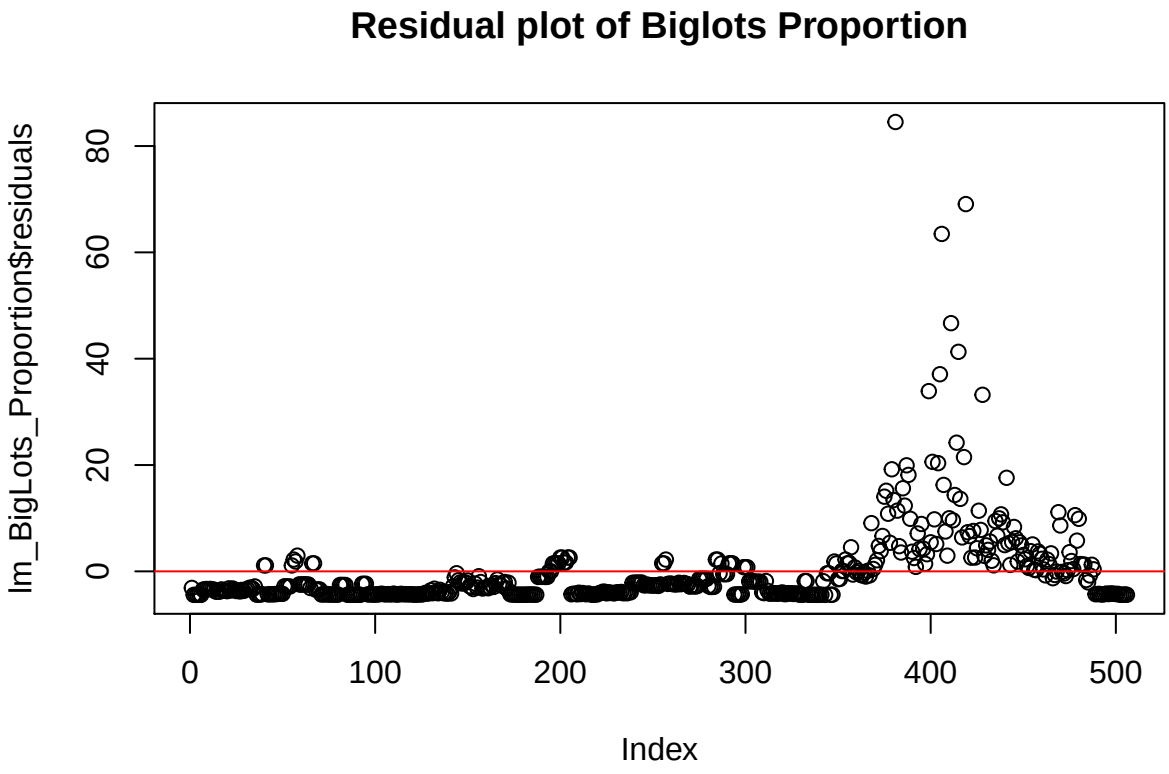
```
##
## Call:
## lm(formula = CrimeRate ~ Median_Owner, data = Boston_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.79654    0.93419   12.63   <2e-16 ***
## Median_Owner -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
```

Among the above linear models, all of them are statistically significant except the CharlesRiver variable. But the models with Owner_Prop, Taxrate, Black_Proportion and Median_Owner does

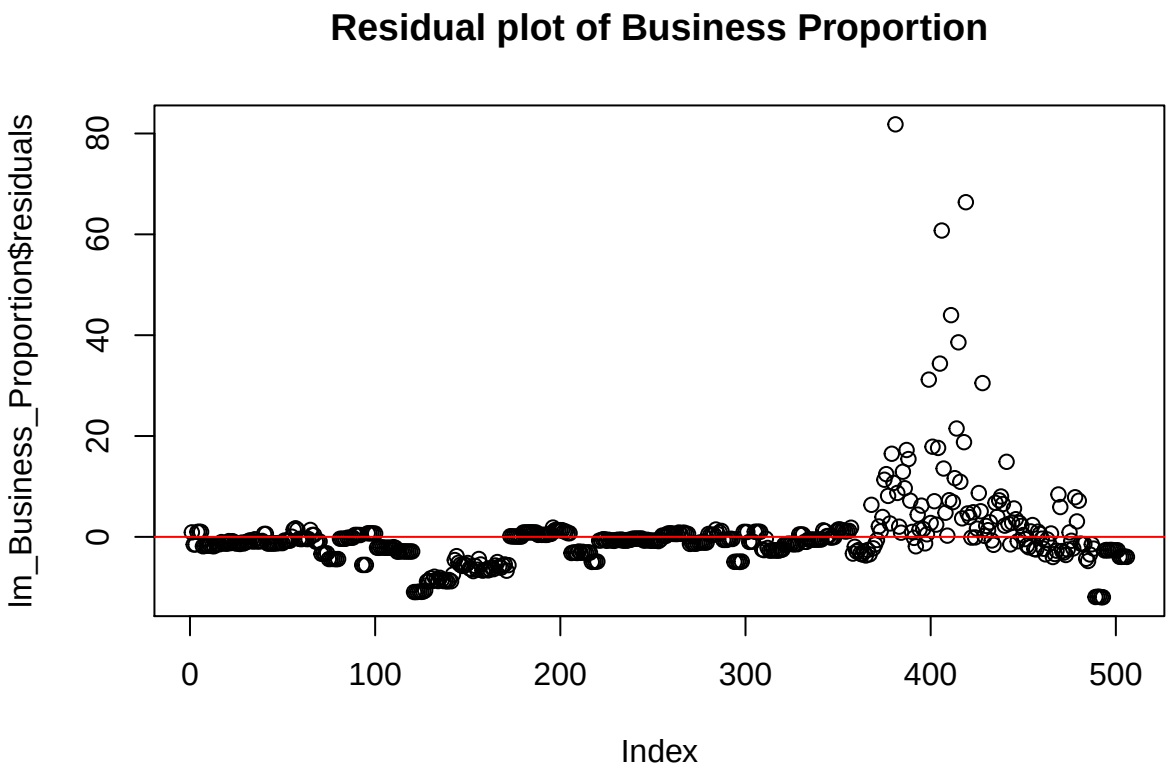
not show high association(Bad coefficient values) with the CrimeRate variable. This leaves us with the variables - Business_Proportion, NO_Concentration, Employ_Distance, Highway_Access and Lower_Status. Let us find the correlation between these variables.

Thus these variable could be fitted against the residuals to check the correctness of the model

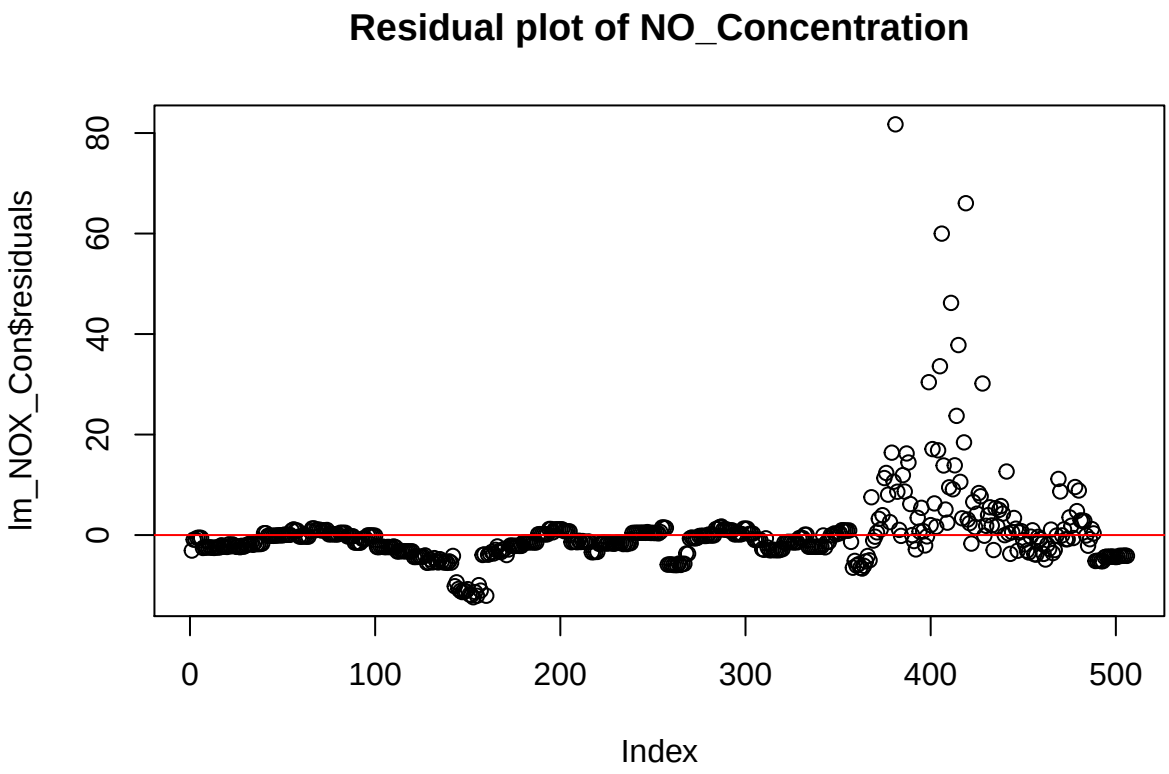
```
plot(lm_BigLots_Proportion$residuals, main = "Residual plot of Biglots Proportion")
abline(h = 0, col = "red")
```



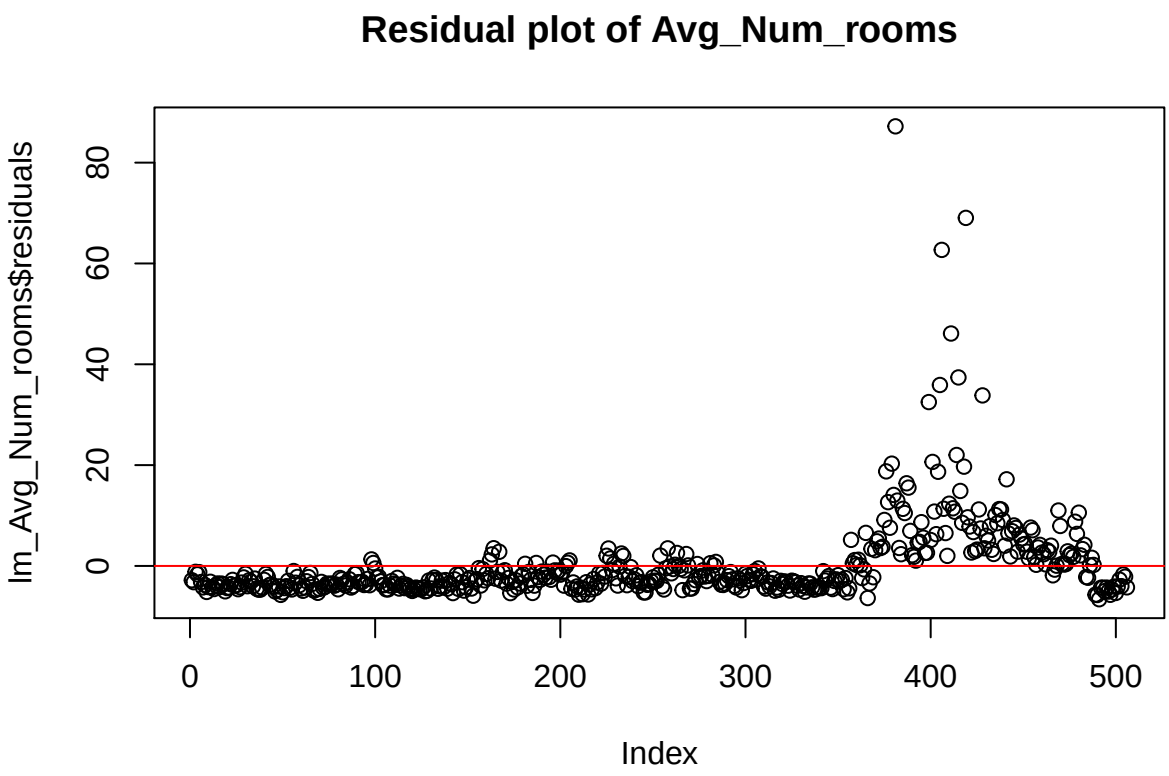

```
plot(lm_Business_Proportion$residuals, main = "Residual plot of Business Proportion")
abline(h = 0, col = "red")
```



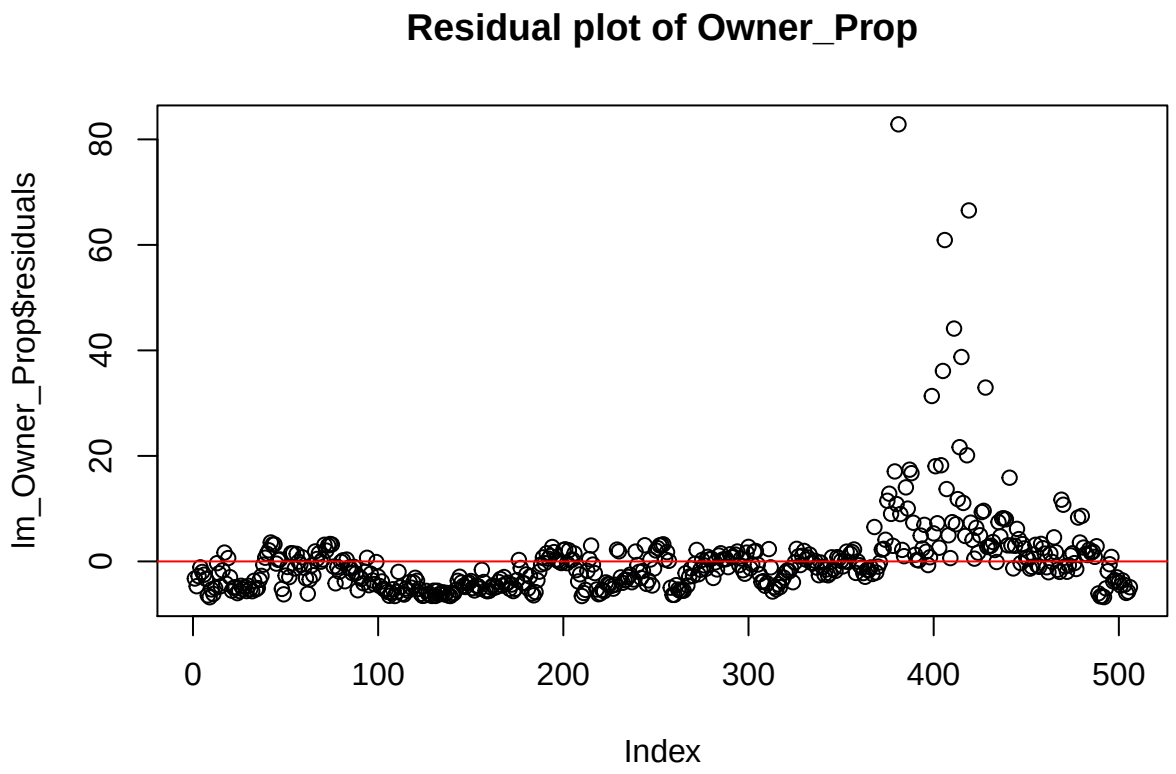

```
plot(lm_NOX_Con$residuals, main = "Residual plot of NO_Concentration")
abline(h = 0, col = "red")
```




```
plot(lm_Avg_Num_rooms$residuals, main = "Residual plot of Avg_Num_rooms")
abline(h = 0, col = "red")
```

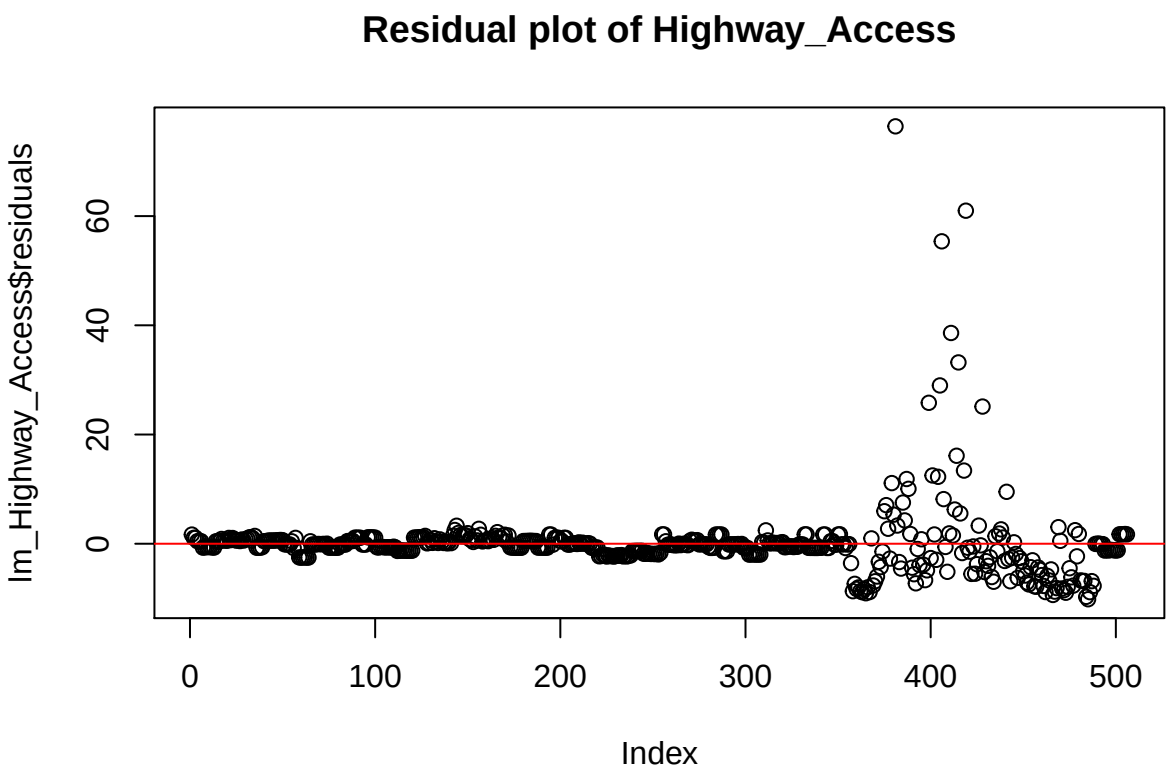



```r
plot(lm_Owner_Prop$residuals, main = "Residual plot of Owner_Prop")
abline(h = 0, col = "red")
```

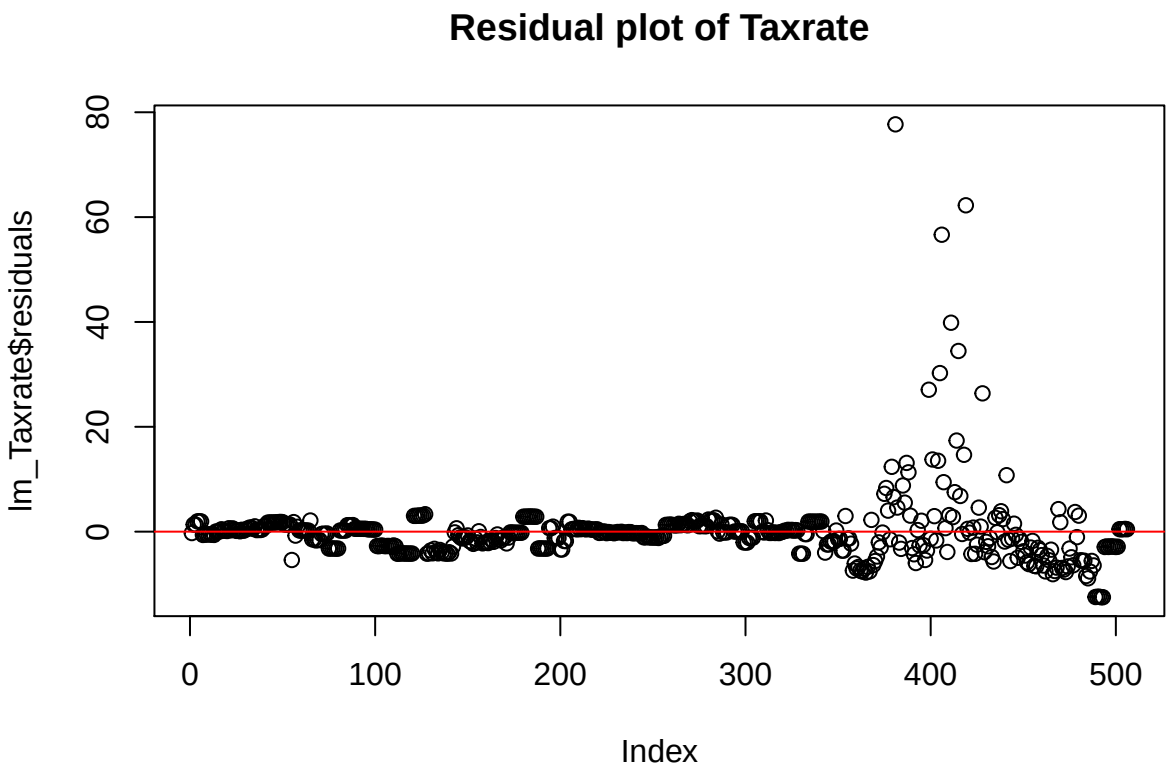



```r
plot(lm_Employ_Distance$residuals, main = "Residual plot of Employ_Distance")
abline(h = 0, col = "red")
```

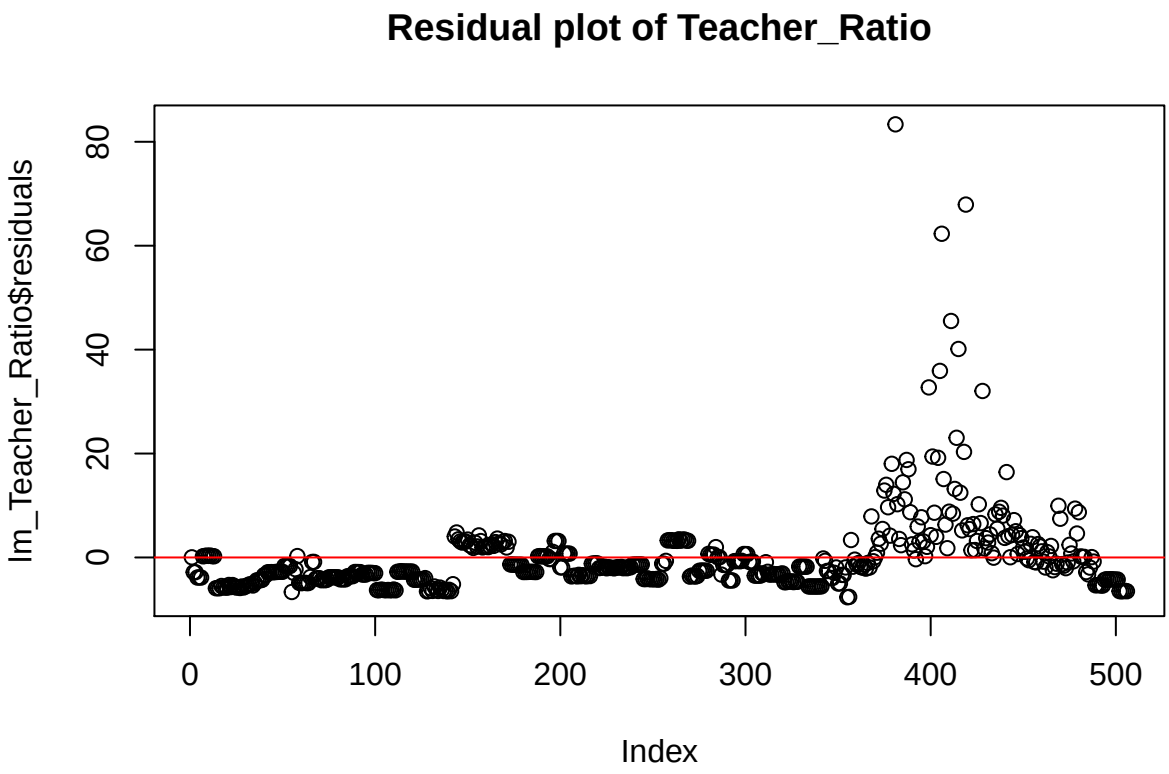

```r
plot(lm_Highway_Access$residuals, main = "Residual plot of Highway_Access")
abline(h = 0, col = "red")
```



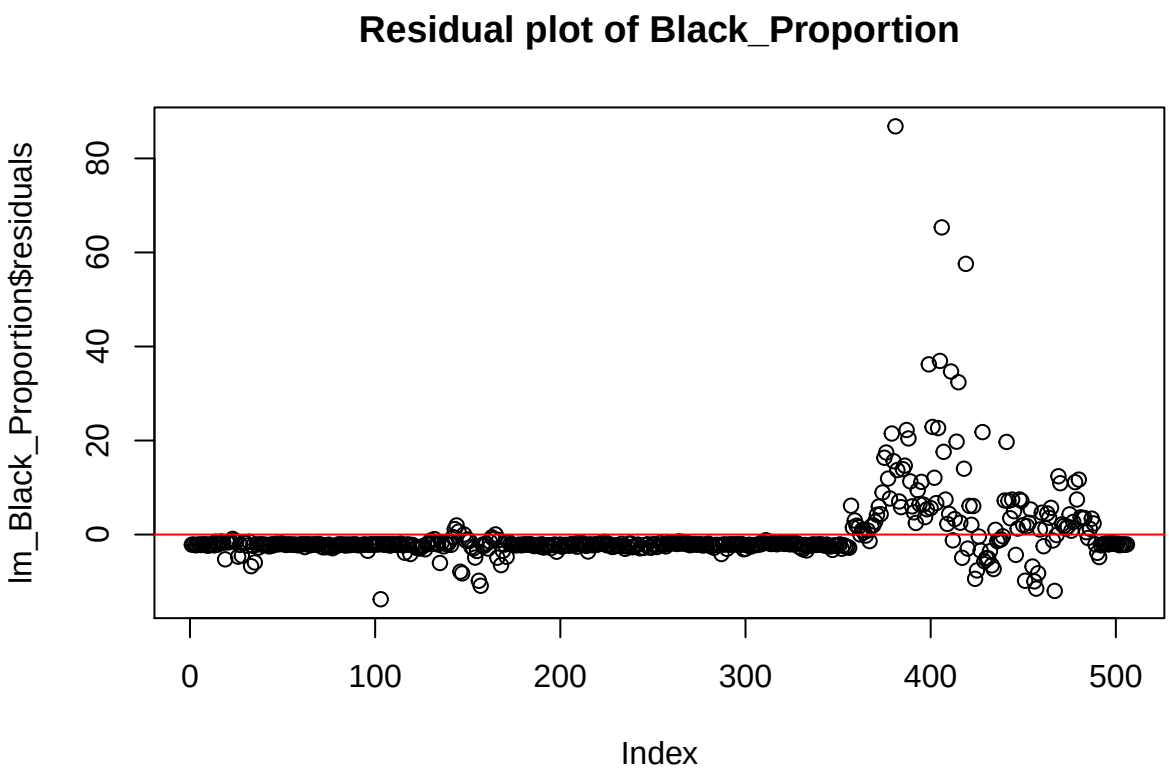

```r
plot(lm_Taxrate$residuals, main = "Residual plot of Taxrate")
abline(h = 0, col = "red")
```



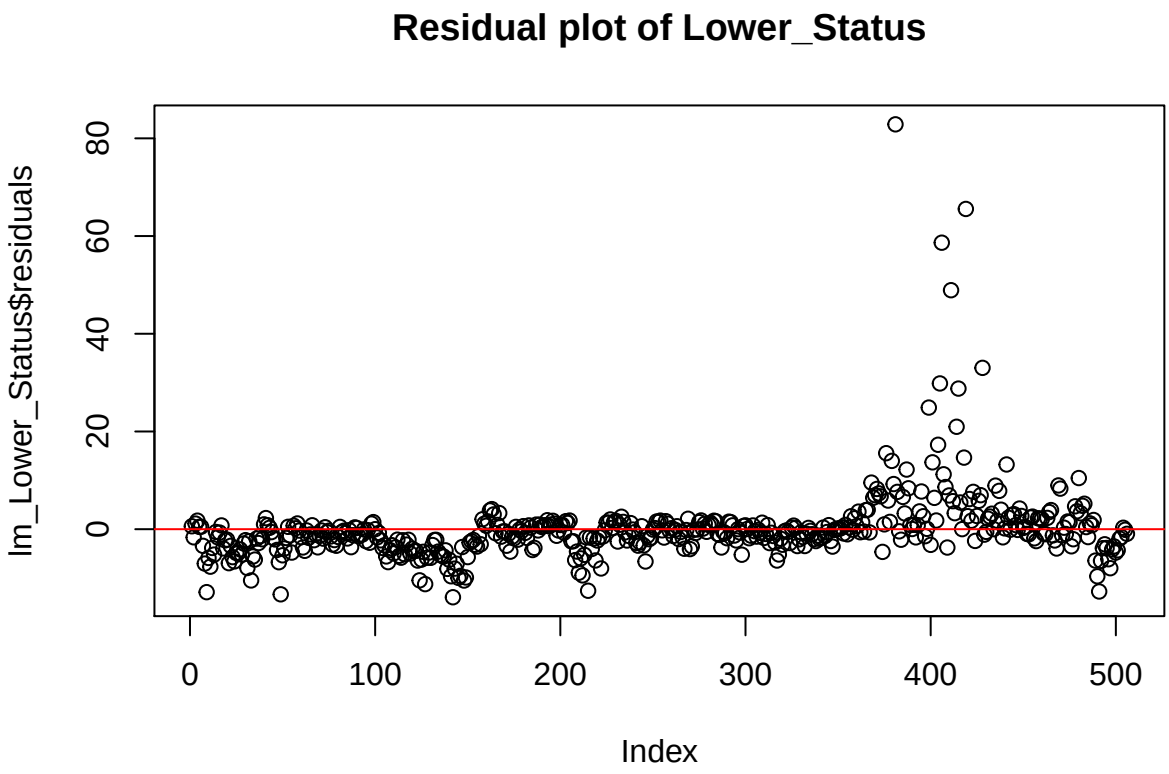

```
plot(lm_Teacher_Ratio$residuals, main = "Residual plot of Teacher_Ratio")
abline(h = 0, col = "red")
```



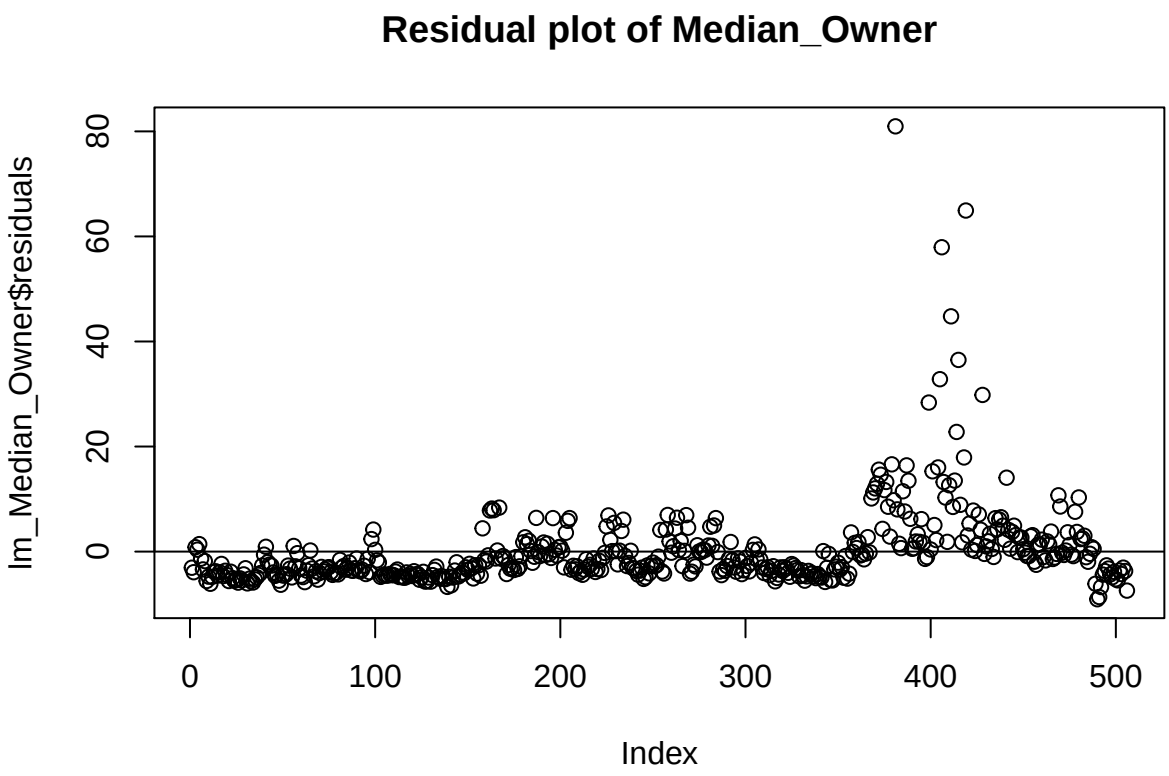

```r
plot(lm_Black_Proportion$residuals, main = "Residual plot of Black_Proportion")
abline(h = 0, col = "red")
```




```r
plot(lm_Lower_Status$residuals, main = "Residual plot of Lower_Status")
abline(h = 0, col = "red")
```




```r
plot(lm_Median_Owner$residuals, main = "Residual plot of Median_Owner")
abline(h = 0)
```




All the above models have a good residual plot with the data points consistently spread across 0. Though there are few datapoints at 400 that are far away from 0, it might not affect significantly as the proportion is less comparatively low.

**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 $H_0 : \beta_j = 0$?**

```
lm_multiple <- lm(CrimeRate ~ ., data = Boston_data)
summary(lm_multiple)
```

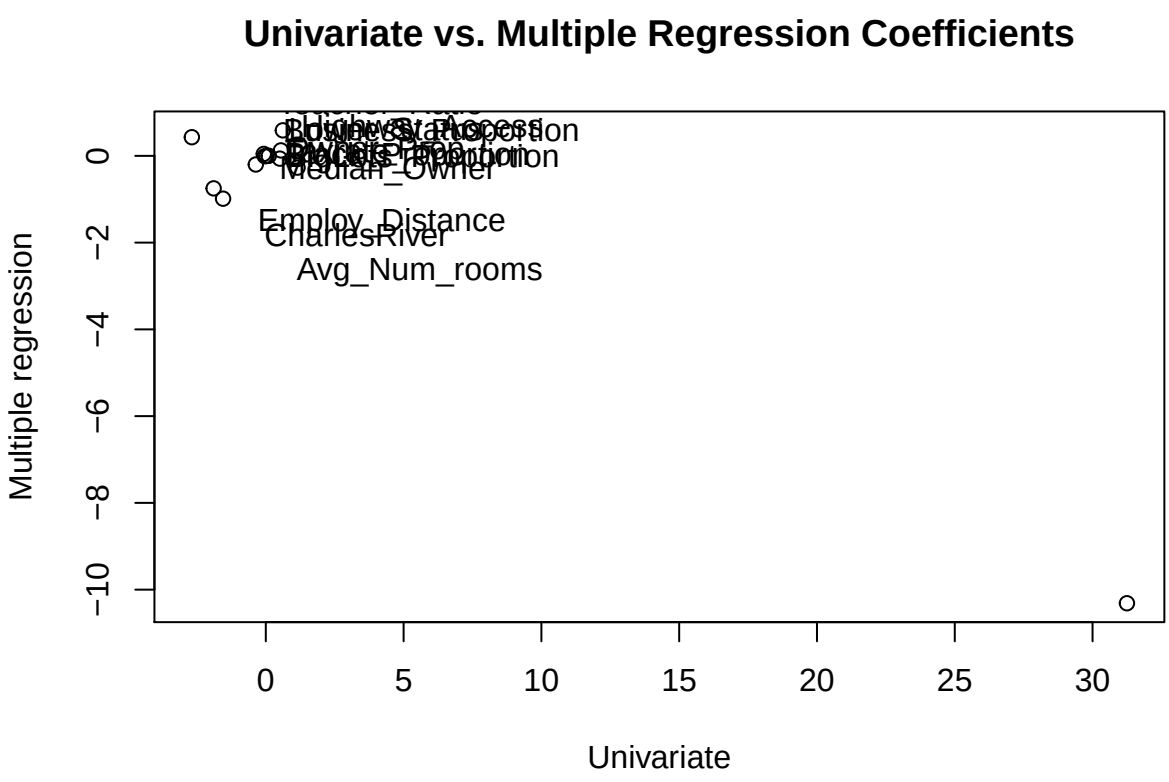
```
##
## Call:
## lm(formula = CrimeRate ~ ., data = Boston_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          17.033228   7.234903   2.354 0.018949 *
## BigLots_Proportion    0.044855   0.018734   2.394 0.017025 *
## Business_Proportion  -0.063855   0.083407  -0.766 0.444294
## CharlesRiver         -0.749134   1.180147  -0.635 0.525867
## NO_Concentration    -10.313535   5.275536  -1.955 0.051152 .
```

```
## Avg_Num_rooms          0.430131   0.612830   0.702 0.483089
## Owner_Prop             0.001452   0.017925   0.081 0.935488
## Employ_Distance       -0.987176   0.281817  -3.503 0.000502 ***
## Highway_Access         0.588209   0.088049   6.680 6.46e-11 ***
## Taxrate               -0.003780   0.005156  -0.733 0.463793
## Teacher_Ratio         -0.271081   0.186450  -1.454 0.146611
## Black_Proportion      -0.007538   0.003673  -2.052 0.040702 *
## Lower_Status           0.126211   0.075725   1.667 0.096208 .
## Median_Owner          -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
```

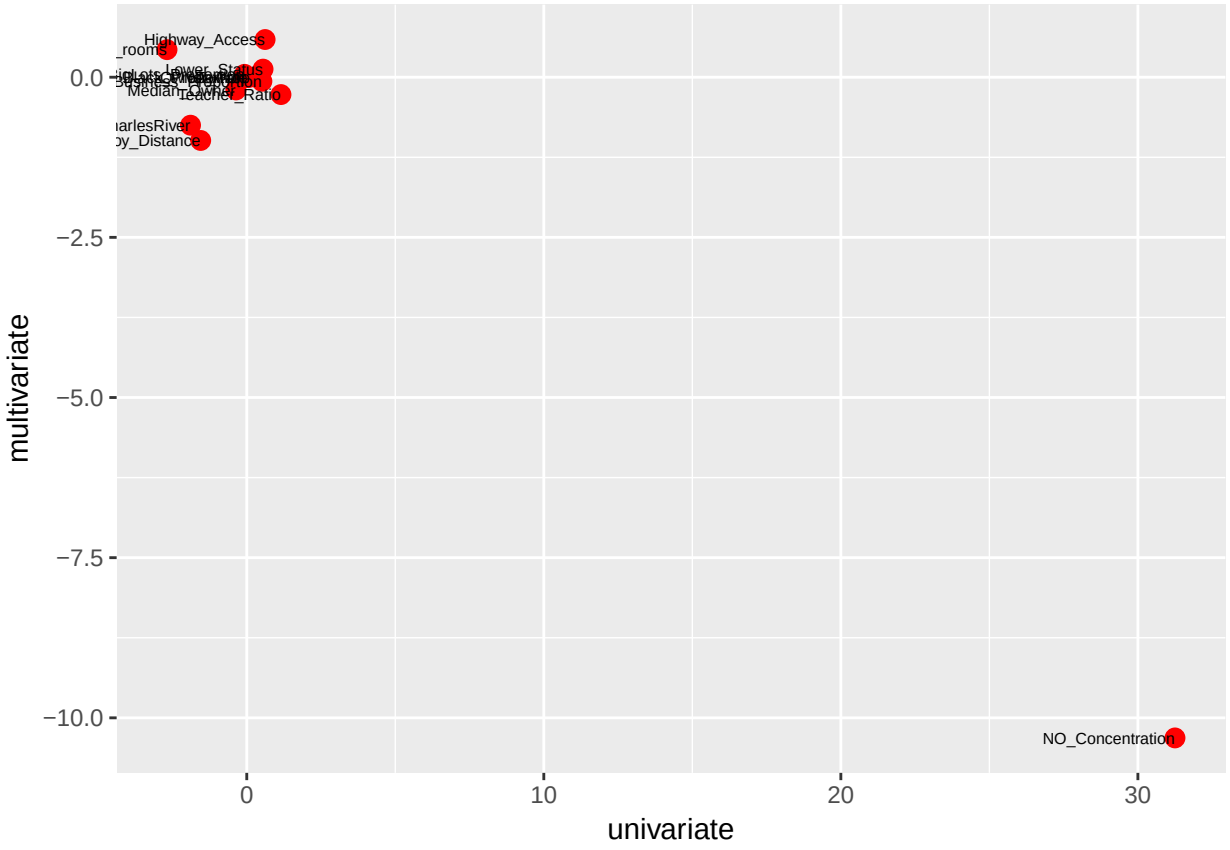
From the multiple regression model, We can reject the null hypothesis for the predictors whose p-values are significant. Those are (in order of high significance), Employ_Distance, Highway_Access, Median_Owner, Business_Proportion, BigLots_Proportion, Black_Proportion and NO_Concentration repectively. As the p-values for all the other coefficients are higher than 0.05, the null hypothesis cannot be rejected.

**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.**

```r
#Create a list consisting of all the univariates from the linear regression model
univariate <- lm_BigLots_Proportion$coefficients[2]
univariate <- append(univariate, lm_Business_Proportion$coefficients[2])
univariate <- append(univariate, lm_CharlesRiver$coefficients[2])
univariate <- append(univariate, lm_NOX_Con$coefficients[2])
univariate <- append(univariate, lm_Avg_Num_rooms$coefficients[2])
univariate <- append(univariate, lm_Owner_Prop$coefficients[2])
univariate <- append(univariate, lm_Employ_Distance$coefficients[2])
univariate <- append(univariate, lm_Highway_Access$coefficients[2])
univariate <- append(univariate, lm_Taxrate$coefficients[2])
univariate <- append(univariate, lm_Teacher_Ratio$coefficients[2])
univariate <- append(univariate, lm_Black_Proportion$coefficients[2])
univariate <- append(univariate, lm_Lower_Status$coefficients[2])
univariate <- append(univariate, lm_Median_Owner$coefficients[2])

#Create a list consisting of all the multivariates from the multiple regression model
multivariate <- lm_multiple$coefficients[2:14]

variatedf <-  as.data.frame(cbind(univariate, multivariate))

#Plot univariates against multivariates
plot(univariate, multivariate, main = "Univariate vs. Multiple Regression Coefficients",
    xlab = "Univariate", ylab = "Multiple regression")
with(variatedf, text(variatedf$univariate ~ variatedf$multivariate,
                     labels = row.names(variatedf), pos = 4))
```




```
ggplot(variatedf, aes(univariate, multivariate)) +
  geom_point(stat = "identity", colour = "red", size = 3) +
  geom_text(label = rownames(variatedf), size = 2, hjust = 1)
```




By comparing linear and multiple regression models, we see that most of the significant variables are similar except the Lower_Status variable. But after running the multiple regression model, we found more variables that are significant like Black_Proportion, Median_Owner and BigLots_Proportion. Also from the plot, we see that most of the datapoints are close to 0 except the NO_Concentration variable.

**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:**

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

```
#Non-linear regression model for Crime Rate vs BigLots_Proportion
nlm_BigLots_Proportion <- lm(CrimeRate ~ BigLots_Proportion + I(BigLots_Proportion^2) +
                             I(BigLots_Proportion^3), data = Boston_data)
summary(nlm_BigLots_Proportion)
```

```
##
## Call:
## lm(formula = CrimeRate ~ BigLots_Proportion + I(BigLots_Proportion^2) +
##     I(BigLots_Proportion^3), data = Boston_data)
##
## 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 ***
## BigLots_Proportion       -3.322e-01  1.098e-01  -3.025  0.00261 **
## I(BigLots_Proportion^2)   6.483e-03  3.861e-03   1.679  0.09375 .
## I(BigLots_Proportion^3) -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
```

```
#Non-linear regression model for Crime Rate vs Business_Proportion
nlm_Business_Proportion <- lm(CrimeRate ~ Business_Proportion + I(Business_Proportion^2) +
                              I(Business_Proportion^3), data = Boston_data)
summary(nlm_Business_Proportion)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Business_Proportion + I(Business_Proportion^2) +
##     I(Business_Proportion^3), data = Boston_data)
##
## 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 *
## Business_Proportion      -1.9652129  0.4819901  -4.077 5.30e-05 ***
## I(Business_Proportion^2)  0.2519373  0.0393221   6.407 3.42e-10 ***
## I(Business_Proportion^3) -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
```

```
#Non-linear regression model for Crime Rate vs CharlesRiver
nlm_CharlesRiver <- lm(CrimeRate ~ CharlesRiver + I(CharlesRiver^2) +
                       I(CharlesRiver^3), data = Boston_data)
summary(nlm_CharlesRiver)
```

```
##
## Call:
## lm(formula = CrimeRate ~ CharlesRiver + I(CharlesRiver^2) + I(CharlesRiver^3),
##     data = Boston_data)
##
## 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 ***
```

```
## CharlesRiver       -1.8928     1.5061  -1.257    0.209
## I(CharlesRiver^2)       NA         NA      NA       NA
## I(CharlesRiver^3)       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
```

```
#Non-Linear regression model for Crime Rate vs Nitrogen oxide concentration
nlm_NOX_Con <- lm(CrimeRate ~ NO_Concentration + I(NO_Concentration^2) +
                    I(NO_Concentration^3), data = Boston_data)
summary(nlm_NOX_Con)
```

```
##
## Call:
## lm(formula = CrimeRate ~ NO_Concentration + I(NO_Concentration^2) +
##     I(NO_Concentration^3), data = Boston_data)
##
## 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 ***
## NO_Concentration       -1279.37     170.40  -7.508 2.76e-13 ***
## I(NO_Concentration^2)   2248.54     279.90   8.033 6.81e-15 ***
## I(NO_Concentration^3) -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
```

```
#Non-Linear regression model for Crime Rate vs Avg_Num_rooms
nlm_Avg_Num_rooms <- lm(CrimeRate ~ Avg_Num_rooms + I(Avg_Num_rooms^2) +
                          I(Avg_Num_rooms^3), data = Boston_data)
summary(nlm_Avg_Num_rooms)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Avg_Num_rooms + I(Avg_Num_rooms^2) +
##     I(Avg_Num_rooms^3), data = Boston_data)
##
## 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 .
## Avg_Num_rooms       -39.1501    31.3115  -1.250   0.2118
## I(Avg_Num_rooms^2)    4.5509     5.0099   0.908   0.3641
```

```
## I(Avg_Num_rooms^3)  -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
```

```
#Non-Linear regression model for Crime Rate vs Owner_Prop
nlm_Owner_Prop <- lm(CrimeRate ~ Owner_Prop + I(Owner_Prop^2) +
                       I(Owner_Prop^3), data = Boston_data)
summary(nlm_Owner_Prop)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Owner_Prop + I(Owner_Prop^2) + I(Owner_Prop^3),
##     data = Boston_data)
##
## 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
## Owner_Prop       2.737e-01  1.864e-01   1.468  0.14266
## I(Owner_Prop^2) -7.230e-03  3.637e-03  -1.988  0.04738 *
## I(Owner_Prop^3)  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
```

```
#Non-Linear regression model for Crime Rate vs Employ_Distance
nlm_Employ_Distance <- lm(CrimeRate ~ Employ_Distance + I(Employ_Distance^2) +
                            I(Employ_Distance^3), data = Boston_data)
summary(nlm_Employ_Distance)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Employ_Distance + I(Employ_Distance^2) +
##     I(Employ_Distance^3), data = Boston_data)
##
## 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 ***
## Employ_Distance       -15.5543     1.7360  -8.960  < 2e-16 ***
## I(Employ_Distance^2)    2.4521     0.3464   7.078 4.94e-12 ***
## I(Employ_Distance^3)   -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
```

```
#Non-Linear regression model for Crime Rate vs Highway_Access
nlm_Highway_Access <- lm(CrimeRate ~ Highway_Access + I(Highway_Access^2) +
                           I(Highway_Access^3), data = Boston_data)
summary(nlm_Highway_Access)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Highway_Access + I(Highway_Access^2) +
##     I(Highway_Access^3), data = Boston_data)
##
## 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
## Highway_Access       0.512736   1.043597   0.491    0.623
## I(Highway_Access^2) -0.075177   0.148543  -0.506    0.613
## I(Highway_Access^3)  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
```

```
#Non-Linear regression model for Crime Rate vs Taxrate
nlm_Taxrate <- lm(CrimeRate ~ Taxrate + I(Taxrate^2) +
                    I(Taxrate^3), data = Boston_data)
summary(nlm_Taxrate)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Taxrate + I(Taxrate^2) + I(Taxrate^3),
##     data = Boston_data)
##
## 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
## Taxrate      -1.533e-01  9.568e-02  -1.602    0.110
## I(Taxrate^2)  3.608e-04  2.425e-04   1.488    0.137
## I(Taxrate^3) -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
```

```r
#Non-Linear regression model for Crime Rate vs Taxrate
nlm_Teacher_Ratio <- lm(CrimeRate ~ Teacher_Ratio + I(Teacher_Ratio^2) +
                          I(Teacher_Ratio^3), data = Boston_data)
summary(nlm_Teacher_Ratio)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Teacher_Ratio + I(Teacher_Ratio^2) +
##     I(Teacher_Ratio^3), data = Boston_data)
##
## 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 **
## Teacher_Ratio      -82.36054   27.64394  -2.979  0.00303 **
## I(Teacher_Ratio^2)   4.63535    1.60832   2.882  0.00412 **
## I(Teacher_Ratio^3)  -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
```

```r
#Non-Linear regression model for Crime Rate vs Black_Proportion
nlm_Black_Proportion <- lm(CrimeRate ~ Black_Proportion + I(Black_Proportion^2) +
                             I(Black_Proportion^3), data = Boston_data)
summary(nlm_Black_Proportion)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Black_Proportion + I(Black_Proportion^2) +
##     I(Black_Proportion^3), data = Boston_data)
##
## 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 ***
## Black_Proportion       -8.356e-02  5.633e-02  -1.483    0.139
## I(Black_Proportion^2)   2.137e-04  2.984e-04   0.716    0.474
## I(Black_Proportion^3) -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
```

```r
#Non-Linear regression model for Crime Rate vs Lower_Status
nlm_Lower_Status <- lm(CrimeRate ~ Lower_Status + I(Lower_Status^2) +
                         I(Lower_Status^3), data = Boston_data)
summary(nlm_Lower_Status)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Lower_Status + I(Lower_Status^2) + I(Lower_Status^3),
##     data = Boston_data)
##
## 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
## Lower_Status       -0.4490656  0.4648911  -0.966   0.3345
## I(Lower_Status^2)   0.0557794  0.0301156   1.852   0.0646 .
## I(Lower_Status^3) -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
```

```r
#Non-Linear regression model for Crime Rate vs Median_Owner
nlm_Median_Owner<- lm(CrimeRate ~ Median_Owner + I(Median_Owner^2) +
                        I(Median_Owner^3), data = Boston_data)
summary(nlm_Median_Owner)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Median_Owner + I(Median_Owner^2) + I(Median_Owner^3),
##     data = Boston_data)
##
## 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 ***
## Median_Owner      -5.0948305  0.4338321 -11.744  < 2e-16 ***
## I(Median_Owner^2)  0.1554965  0.0171904   9.046  < 2e-16 ***
## I(Median_Owner^3) -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
```

The first thing to note is that with the CharlesRiver variable, we get NA values for the squared

and cubed term. This makes sense as CharlesRiver is a dummy variable, composed of only 0s and 1s, and these values will not change if they are squared or cubed.

With the variables Business_Proportion, NO_Concentration, Employ_Distance, Teacher_Ratio and Median_Owner, there is evidence of a non-linear relationship, as each of these variables squared and cubed terms is found to be statistically signficant (we reject the null hypothesis that the coeffecients on these exponentiated variables are zero). Owner_Prop also appears to have a non-linear relationship, as once squared-age and cubed-age are brought into the model, linear age becomes statistically insignficant.

For every other variable, we do not find evidence of a non-linear relationship between the predictor and outcome variables.

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

```
fwd_stepwise_fit <- regsubsets(CrimeRate ~. , data = Boston_data, nvmax = 13, method = "forward")
summary(fwd_stepwise_fit)
```

```
## Subset selection object
## Call: regsubsets.formula(CrimeRate ~ ., data = Boston_data, nvmax = 13,
##     method = "forward")
## 13 Variables  (and intercept)
##                      Forced in Forced out
## BigLots_Proportion       FALSE      FALSE
## Business_Proportion      FALSE      FALSE
## CharlesRiver             FALSE      FALSE
## NO_Concentration         FALSE      FALSE
## Avg_Num_rooms            FALSE      FALSE
## Owner_Prop               FALSE      FALSE
## Employ_Distance          FALSE      FALSE
## Highway_Access           FALSE      FALSE
## Taxrate                  FALSE      FALSE
## Teacher_Ratio            FALSE      FALSE
## Black_Proportion         FALSE      FALSE
## Lower_Status             FALSE      FALSE
## Median_Owner             FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: forward
##           BigLots_Proportion Business_Proportion CharlesRiver
## 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 ) "*"                "*"                 "*"
##           NO_Concentration Avg_Num_rooms Owner_Prop Employ_Distance
```

```
## 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 ) "*"              "*"           "*"         "*"
##           Highway_Access Taxrate Teacher_Ratio Black_Proportion
## 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 ) "*"            "*"     "*"           "*"
##           Lower_Status Median_Owner
## 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 ) "*"          "*"
```

```
bkwd_stepwise_fit <- regsubsets(CrimeRate ~. , data = Boston_data, nvmax = 13, method = "backward")
summary(bkwd_stepwise_fit)
```

```
## Subset selection object
## Call: regsubsets.formula(CrimeRate ~ ., data = Boston_data, nvmax = 13,
##     method = "backward")
## 13 Variables  (and intercept)
##                      Forced in Forced out
## BigLots_Proportion       FALSE      FALSE
## Business_Proportion      FALSE      FALSE
## CharlesRiver             FALSE      FALSE
## NO_Concentration         FALSE      FALSE
## Avg_Num_rooms            FALSE      FALSE
```

```
## Owner_Prop              FALSE      FALSE
## Employ_Distance         FALSE      FALSE
## Highway_Access          FALSE      FALSE
## Taxrate                 FALSE      FALSE
## Teacher_Ratio           FALSE      FALSE
## Black_Proportion        FALSE      FALSE
## Lower_Status            FALSE      FALSE
## Median_Owner            FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: backward
##           BigLots_Proportion Business_Proportion CharlesRiver
## 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 ) "*"                "*"                 "*"
##           NO_Concentration Avg_Num_rooms Owner_Prop Employ_Distance
## 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 ) "*"              "*"           "*"        "*"
##           Highway_Access Taxrate Teacher_Ratio Black_Proportion
## 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 ) "*"            "*"     "*"           "*"
##           Lower_Status Median_Owner
## 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 ) "*"          "*"
```

For instance, we see that using forward stepwise selection, the best one-variable model contains only Highway_Access, and the best two-variable model additionally includes Lower_Status. In the backward selection process, the best-one variable model contains the same Highway_Access variable whereas, the best two-variable model includes the Median_Owner.

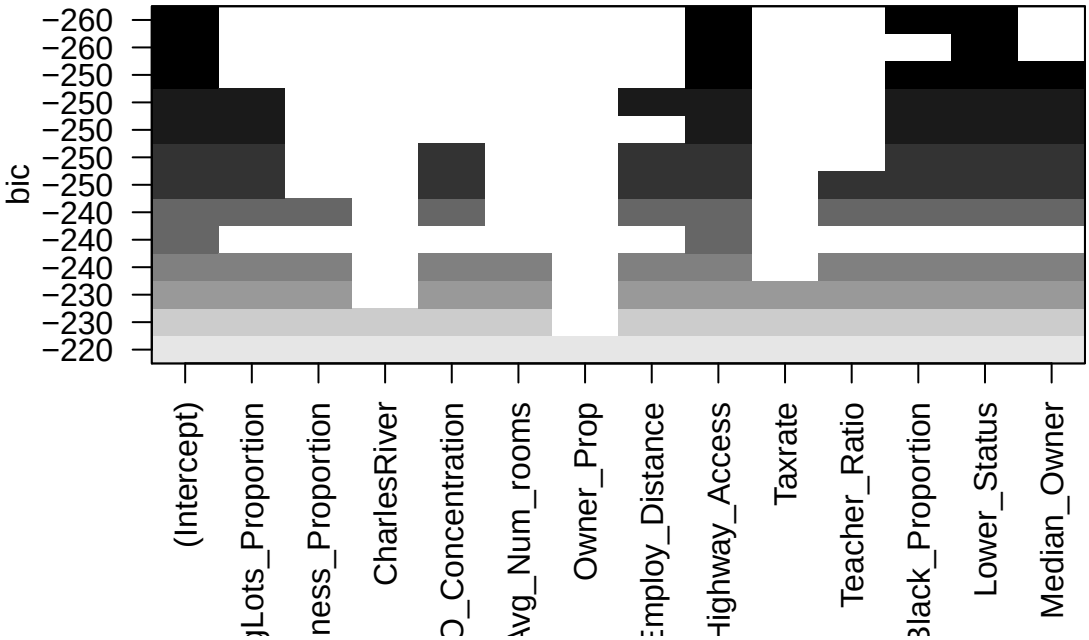
Rest of all the variables differ from each other in both the models. From the **forward stepwise selection** process, the best fit model could have the variables - **Highway_Access, Lower_Status, Black_Proportion, Median_Owner, BigLots_Proportion** respectively. Similarly, from the **backward stepwise selection** process, the best fit model could have the variables - **Highway_Access, Median_Owner, Employ_Distance, BigLots_Proportion, Black_Proportion** respectively. Thus only the Median_Owner and Employ_Distance differ in the top 5 best fit variables from both the approaches.

From the **multiple regression model** we fit in (4), the variables with high significance are - **Employ_Distance, Highway_Access, Median_Owner, Business_Proportion, BigLots_Proportion, Black_Proportion and NO_Concentration**. Almost all the variables that had high significance in multiple regression model is same as that of the stepwise selection model except the buiness_Proportion and NO_Concentration variable.

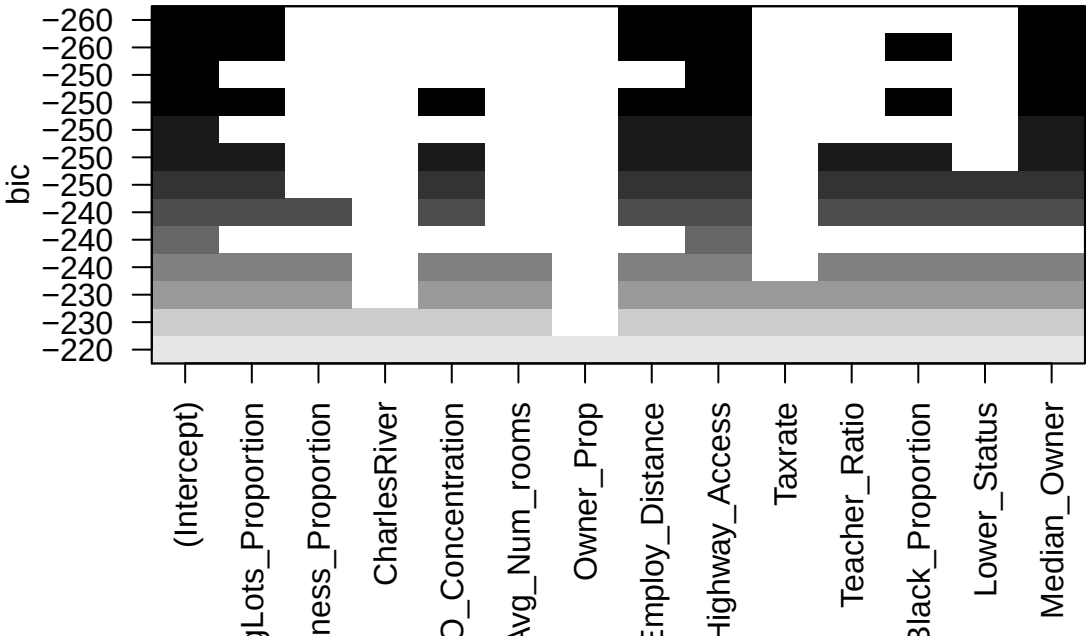
These are significant in the multiple regression model whereas its ranked with lower significance in the stepwise selection approach.

**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.**

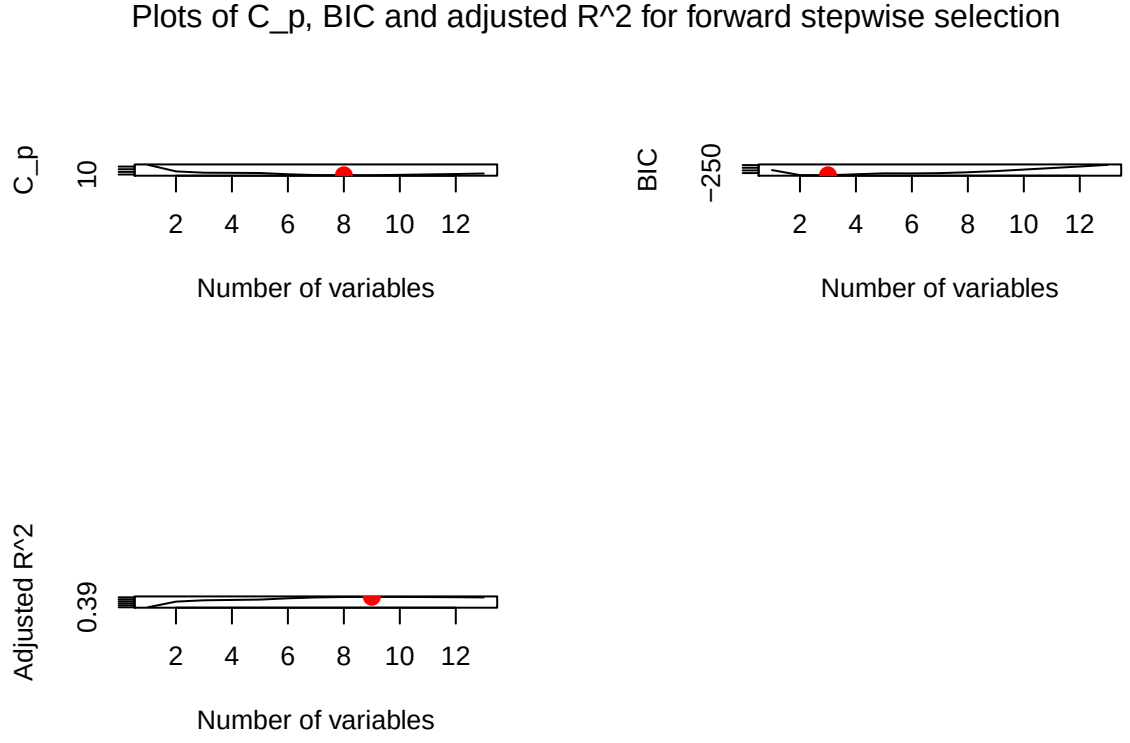
```
plot(fwd_stepwise_fit)
```



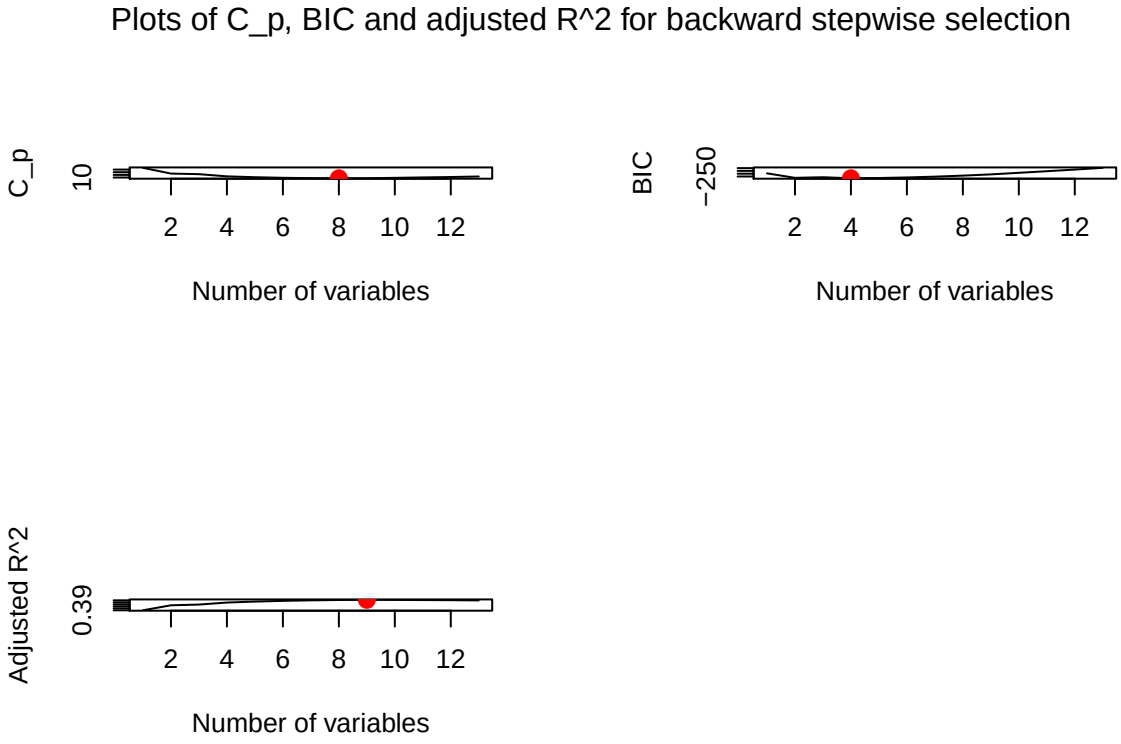

```
plot(bkwd_stepwise_fit)
```




```
#######Forward selection
summary_fwd_stepwise_fit <- summary(fwd_stepwise_fit)
par(mfrow = c(2, 2))

plot(summary_fwd_stepwise_fit$cp, xlab = "Number of variables", ylab = "C_p", type = "l")
points(which.min(summary_fwd_stepwise_fit$cp),
       summary_fwd_stepwise_fit$cp[which.min(summary_fwd_stepwise_fit$cp)],
       col = "red", cex = 2, pch = 20)

plot(summary_fwd_stepwise_fit$bic, xlab = "Number of variables", ylab = "BIC", type = "l")
points(which.min(summary_fwd_stepwise_fit$bic),
       summary_fwd_stepwise_fit$bic[which.min(summary_fwd_stepwise_fit$bic)],
       col = "red", cex = 2, pch = 20)

plot(summary_fwd_stepwise_fit$adjr2, xlab = "Number of variables", ylab = "Adjusted R^2", type = "l")
points(which.max(summary_fwd_stepwise_fit$adjr2),
       summary_fwd_stepwise_fit$adjr2[which.max(summary_fwd_stepwise_fit$adjr2)],
       col = "red", cex = 2, pch = 20)
mtext("Plots of C_p, BIC and adjusted R^2 for forward stepwise selection", side = 3,
      line = -2, outer = TRUE)


#######Backward selection
summary_bkwd_stepwise_fit <- summary(bkwd_stepwise_fit)
par(mfrow = c(2,2))
```




```r
plot(summary_bkwd_stepwise_fit$cp, xlab = "Number of variables", ylab = "C_p", type = "l")
points(which.min(summary_bkwd_stepwise_fit$cp),
       summary_bkwd_stepwise_fit$cp[which.min(summary_bkwd_stepwise_fit$cp)],
       col = "red", cex = 2, pch = 20)

plot(summary_bkwd_stepwise_fit$bic, xlab = "Number of variables", ylab = "BIC", type = "l")
points(which.min(summary_bkwd_stepwise_fit$bic),
       summary_bkwd_stepwise_fit$bic[which.min(summary_bkwd_stepwise_fit$bic)],
       col = "red", cex = 2, pch = 20)

plot(summary_bkwd_stepwise_fit$adjr2, xlab = "Number of variables", ylab = "Adjusted R^2",
     type = "l")
points(which.max(summary_bkwd_stepwise_fit$adjr2),
       summary_bkwd_stepwise_fit$adjr2[which.max(summary_bkwd_stepwise_fit$adjr2)],
       col = "red", cex = 2, pch = 20)
mtext("Plots of C_p, BIC and adjusted R^2 for backward stepwise selection", side = 3,
      line = -2, outer = TRUE)
```




From the forward selection process, based on CP value we pick a 8 variable model, on BIC we pick a 3 variable model and on Adjusted R^2, we pick a 9 variable model. It is similar for the backward selection process except we select a 4-model variable on BIC values. Almost all of the adjusted R squared values(12 variables) are around 0.4 for both the forward and backward selection process, whereas from the residual plot we could pick only 9 variables.