Homework 2 - STAT602

Snigdha Peddi

Question 3.7.5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y_i} = \; x_i\hat{\beta}$$

where

$$\hat{\beta} = (\sum_{i=1}^{n}(x_iy_i))/(\sum_{i'=1}^{n}(x_{i'}^2))$$

Show that we can write

$$\hat{y_i} = \sum_{i'=1}^{n} a_{i'}y_{i'}$$

What is

$$a_{i'}?$$

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.**

Answer 3.7.5.

$$\hat{y_i} = \; x_i\hat{\beta}$$

$$= \; x_i\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i'=1}^{n}x_{i'}^2}$$

$$= \frac{\sum_{i=1}^{n}x_i^2y_i}{\sum_{i'=1}^{n}x_{i'}^2}$$

Multiplying and dividing by sum of y_i' values from i'=1 to n

$$= \frac{\sum_{i=1}^{n}x_i^2y_i}{\sum_{i'=1}^{n}x_{i'}^2}\frac{\sum_{i'=1}^{n}y_{i'}}{\sum_{i'=1}^{n}y_{i'}}$$

Summation over i and i' are independent with respect to each other

$$= \sum_{i'=1}^{n}\frac{\sum_{i=1}^{n}x_i^2y_i}{\sum_{i'=1}^{n}x_{i'}^2\sum_{i'=1}^{n}y_{i'}}y_{i'}$$

$$\hat{y_i} = \sum_{i'=1}^{n} a_{i'}y_{i'}$$

where,

$$a_{i'} = \frac{\sum_{i=1}^{n} x_i^2 y_i}{\sum_{i'=1}^{n} x_{i'}^2 \sum_{i'=1}^{n} y_{i'}}$$

References

- R Markdown for Scientist blog, Chapter 11 Math.(https://rmd4sci.njtierney.com/math)

Question 3.7.10. This question should be answered using the Carseats data set.

```
##
##  Dimensions of dataset: 400 11
```

```
##
##  col names: Sales CompPrice Income Advertising Population Price ShelveLoc Age Education Urban US
```

```
##
##  Number of missing values in dataset: 0
```

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
## [1] "Summary of Lineal Model1:"
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

From the the coefficients obtained above we can interpret that, - with a one unit increase in Price the sales decrease by 0.05 units on average given other variables are constant. - if the store is present in an urban area then the sales decrease by 0.02 units on average given other variables are constant. However, as the p value is higher it indicates that there is not enough evidence to say that presence of store in an urban area would effect the sales (no eveidenc e that a relationship exists between Sales and Urban) - if the store is present in the US then the sales increase by 1.20 units on average given other variables are constant.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

$$Sales = 13.043469 - 0.054459.Price - 0.021916.UrbanYes + 1.200573.USYes \tag{1}$$

(d) For which of the predictors can you reject the null hypothesis H0 : j = 0?

From the p values of model summary we can reject the null hypothesis for variables Price and USYes(lower p values). However, there is no evidence of relationship between UrbanYes and Sales (higher p value) to reject Null hypothesis.

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
## [1] "Summary of Lineal Model2:"
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

(f) How well do the models in (a) and (e) fit the data?

The Adjusted R2 value of Model 1 and Model 2 are pretty close, 0.2335 and 0.2354 respectively with a similar R-squared value (though the model 1 has one variable that has no significance). However, the F-statistic of Model 2 (62.43) is higher than Model 1(41.52) which indicates that the significance of overall model is well explained by Model 2.

(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

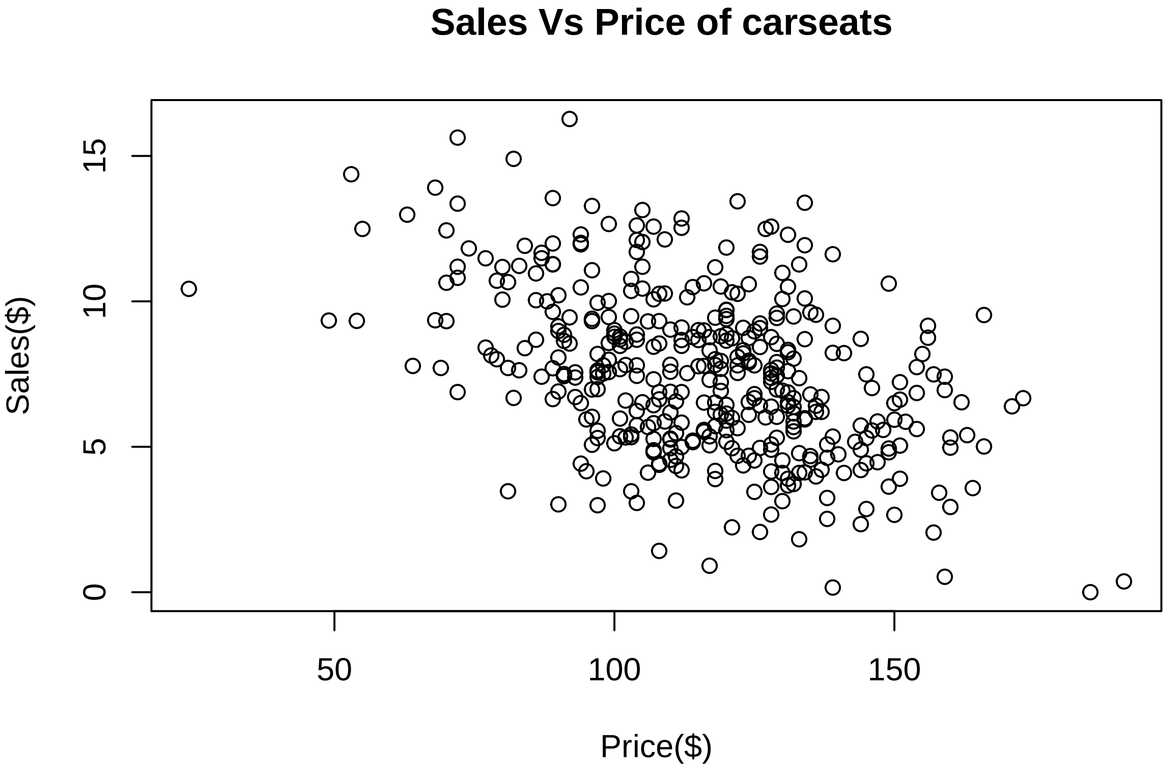
```
## 95% confidence intervals for the coefficients of Model 2:
```

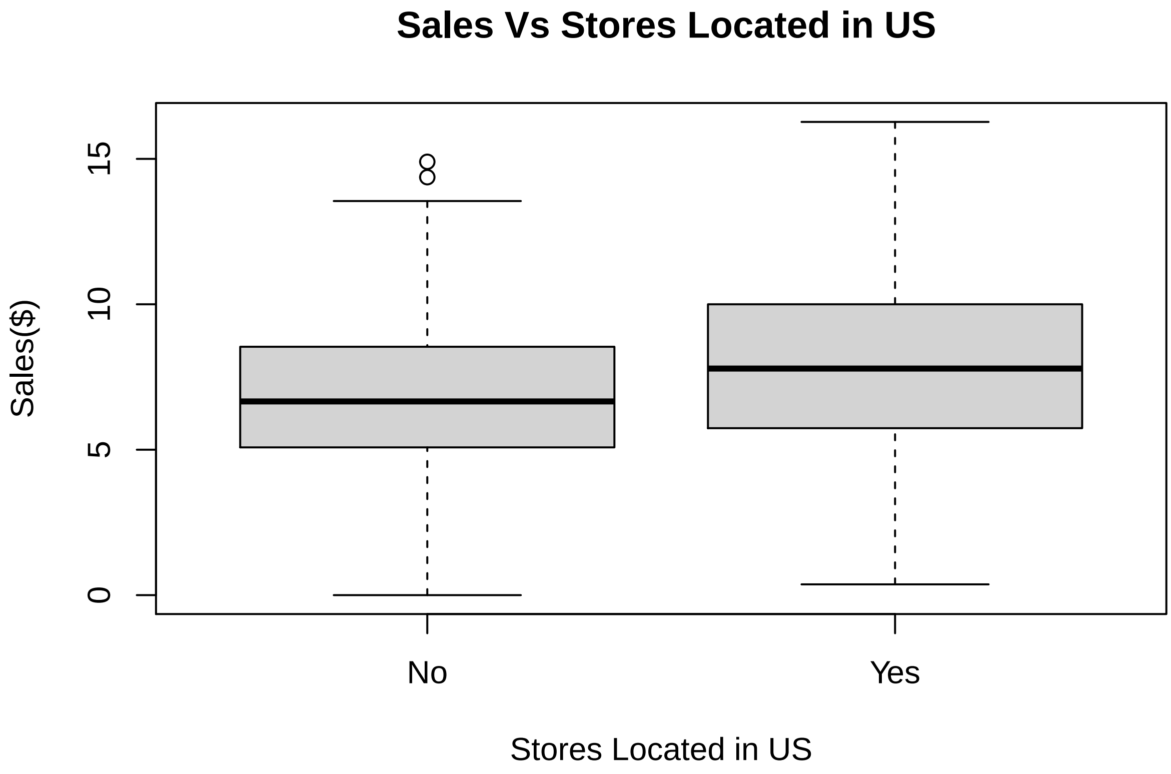
```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

(h) Is there evidence of outliers or high leverage observations in the model from (e)?

- Outliers are the points that fall far from the other data points. These are the points that are extreme in someway.
- If the parameter estimates change a great deal when a point is removed from the calculations, the point is said to be influential.
- Points with extreme values of X are said to have high leverage. High Leverage points have a greater ability to move the line.If these points fall outside the overall pattern, they can be influential.

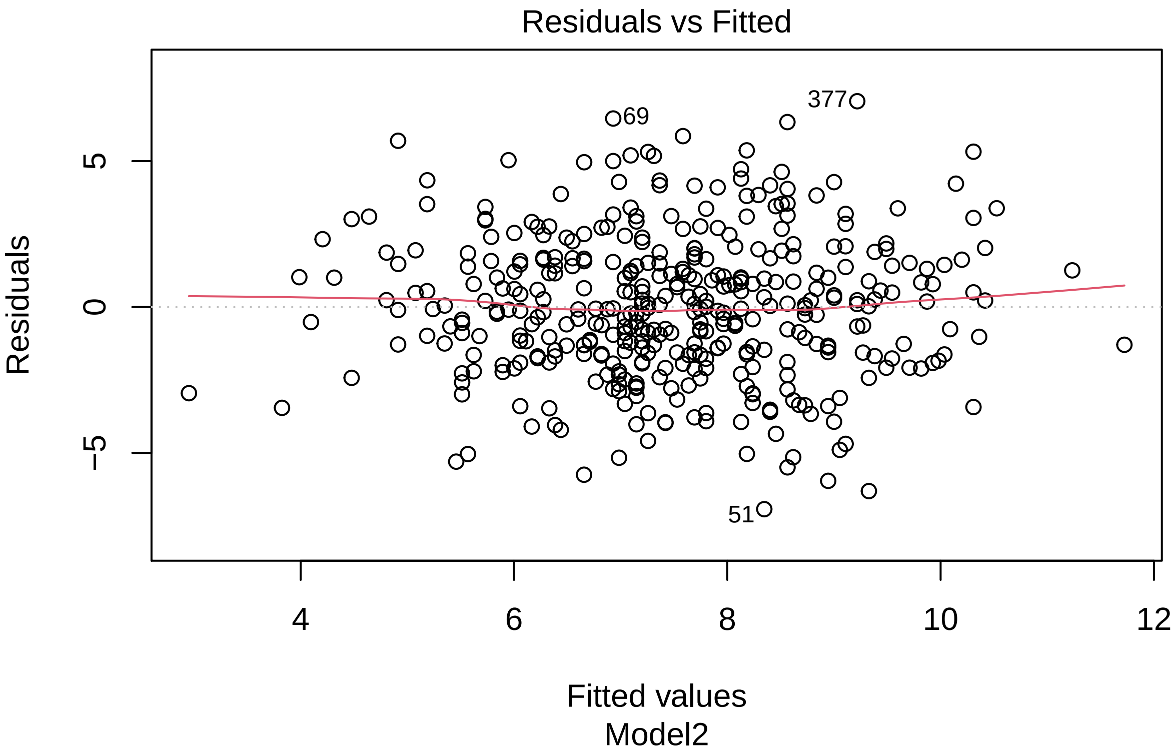
  Considering the Linear model fit before with variables Price and stores located in US the following interpretations are made.

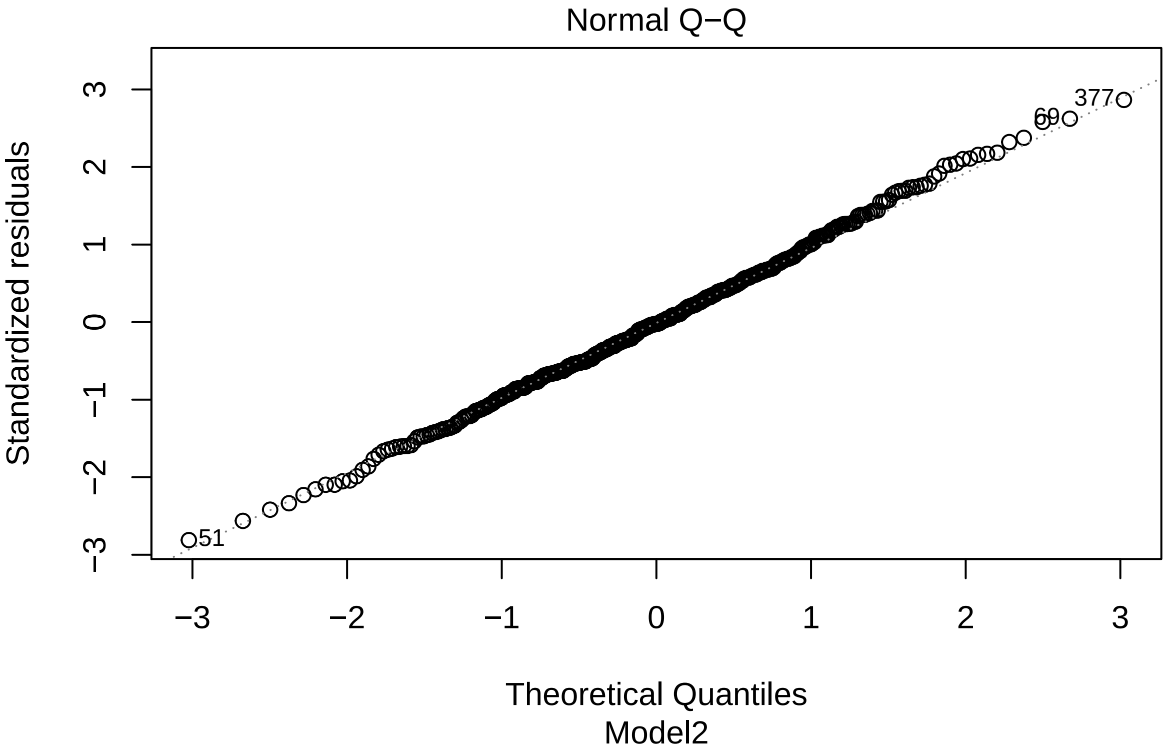


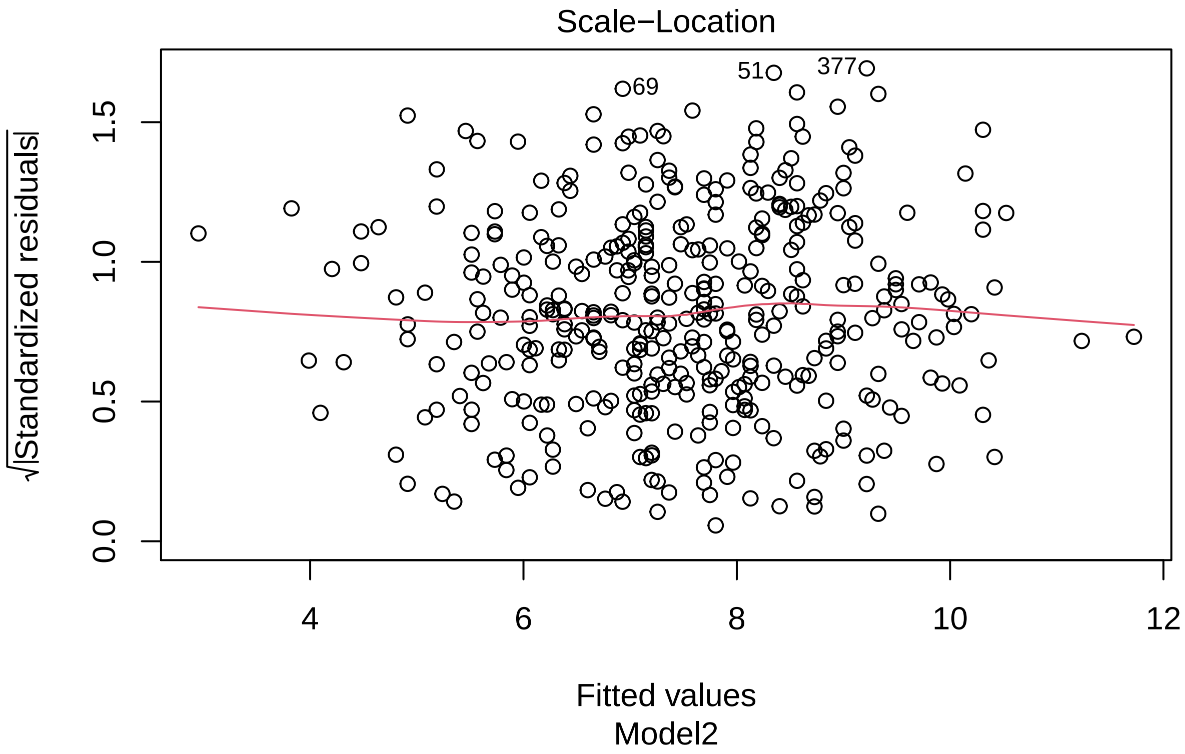


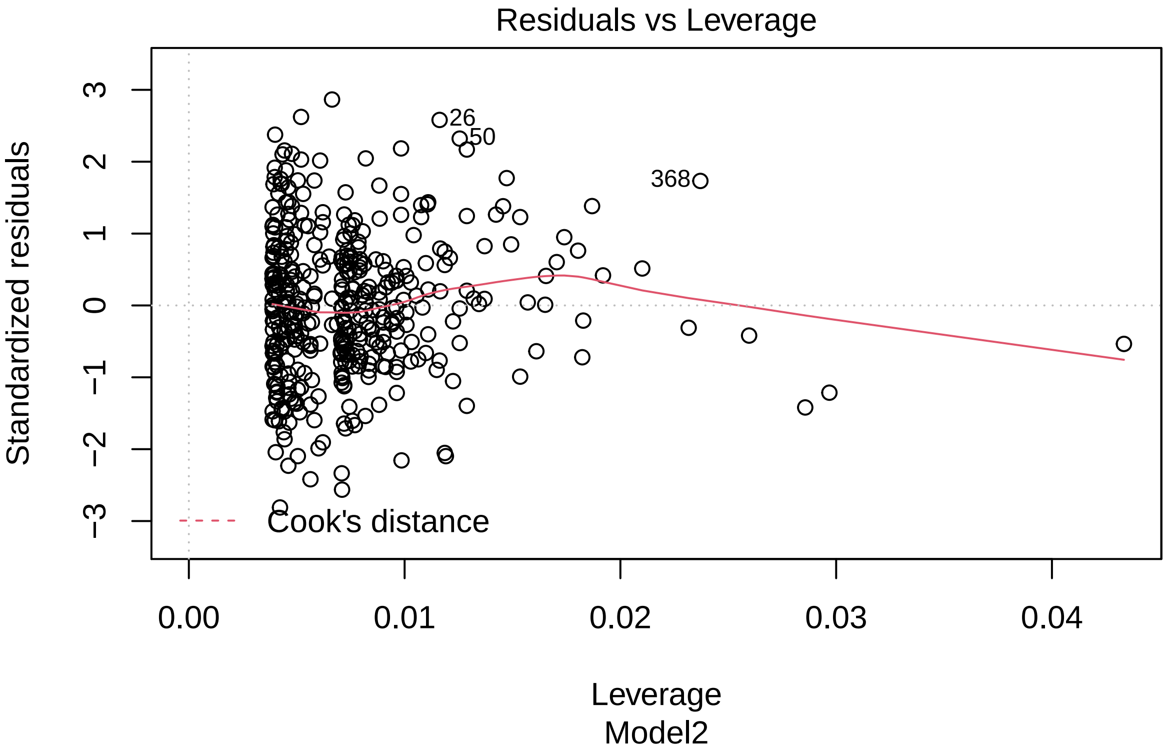
```
## Observed outliers of Price variable:

##    Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 43 10.43        77     69           0         25    24    Medium  50        18
##    Urban US
## 43   Yes No

## Observed outliers of Price variable:

##     Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 175  0.00       139     24           0        358   185    Medium  79        15
## 166  0.37       147     58           7        100   191       Bad  27        15
##     Urban  US
## 175    No  No
## 166   Yes Yes

## Observed outliers of US variable:

##     Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 26  14.90       139     32           0        176    82      Good  54        11
## 368 14.37        95    106           0        256    53      Good  52        17
##     Urban US
## 26     No No
## 368   Yes No
```

Visual observation of the plots and the extracted data shows that the rows 43,175,166,26,368 are outliers. However,the Q-Q plot of Model2 shows that all the data points follow the trend and there are no outliers.




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
377
69
51
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Model2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
4
6
8
10
12
Fitted values
Model2
69
51
377

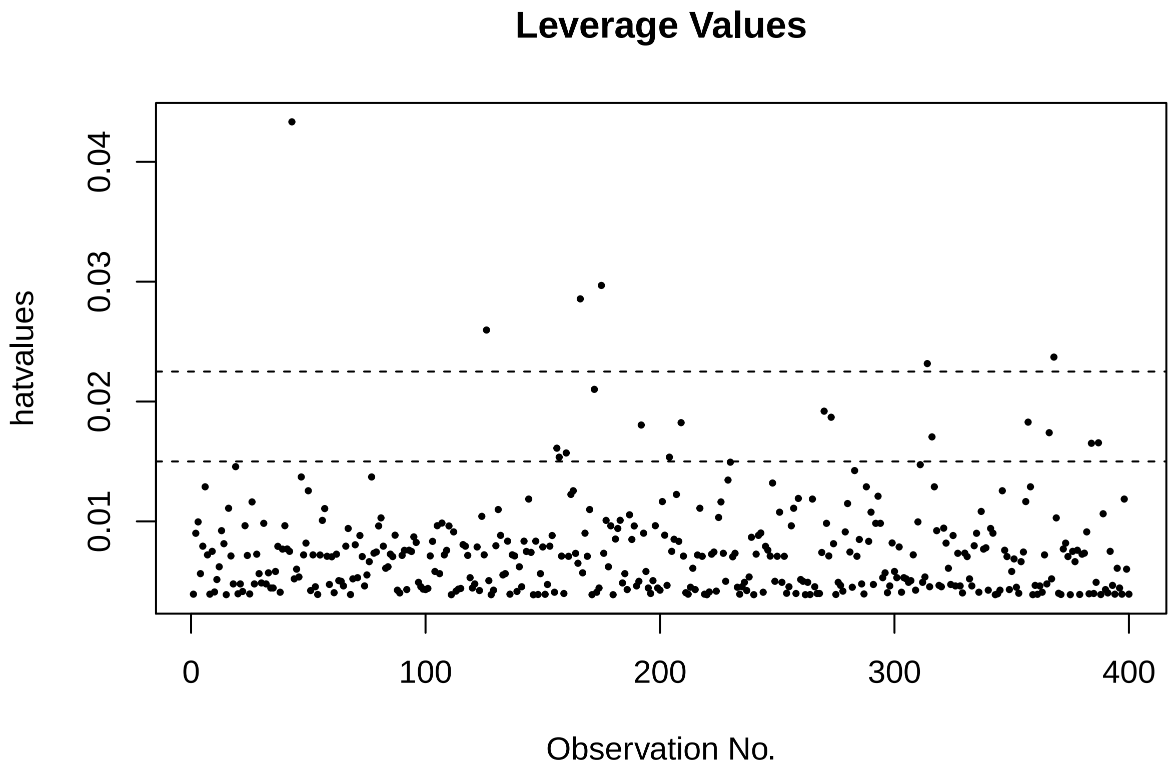


Leverage points can be calculated from hat values.Average value of leverage points for the observations is given by (p+1)/n. Where, p is the number of predictors in the model and n is total number of observations. For our model we have an average leverage value of 0.0075. 3 times the average leverage value gives the high leverage points. The Residual vs Leverage plot show that most of the standardized residuals fall between -2 and 2 indicating that there are not many high leverage observations.

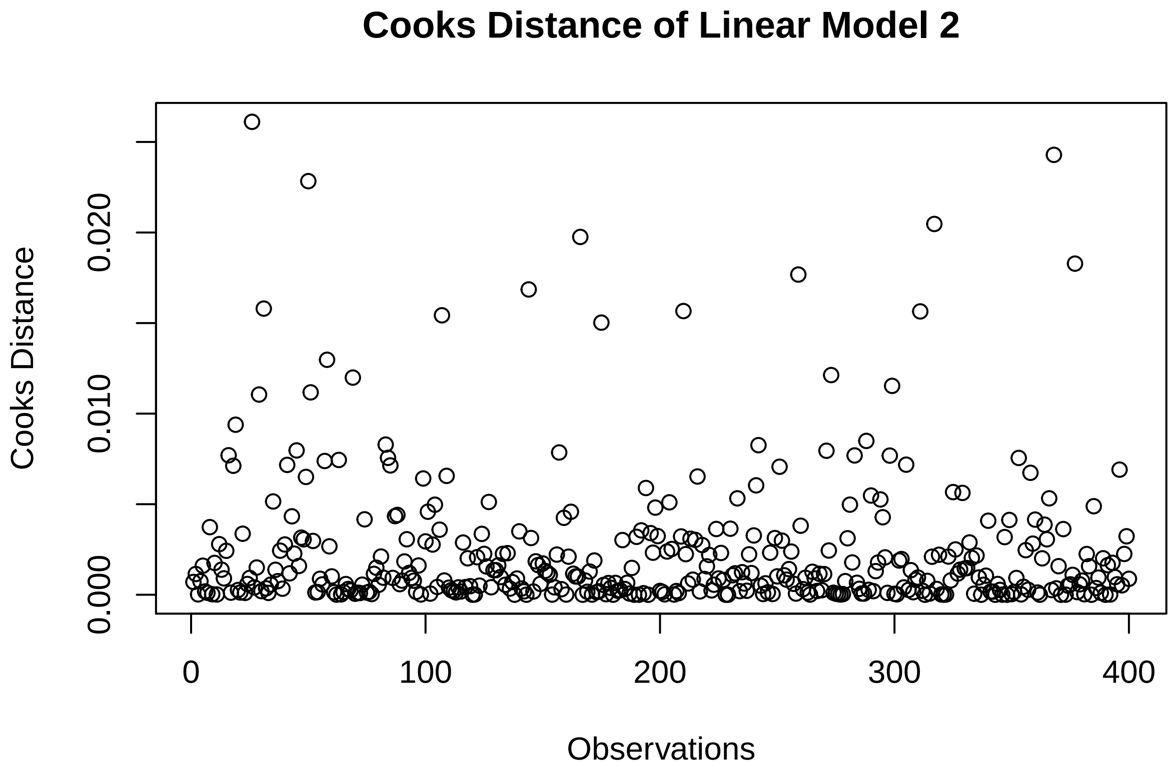
```
##  Average Leverage value: 0.0075
```

```
##     High_Leverage_values
## 43            0.04333766
## 126           0.02596614
## 166           0.02856661
## 175           0.02968672
## 314           0.02316470
## 368           0.02370705
```

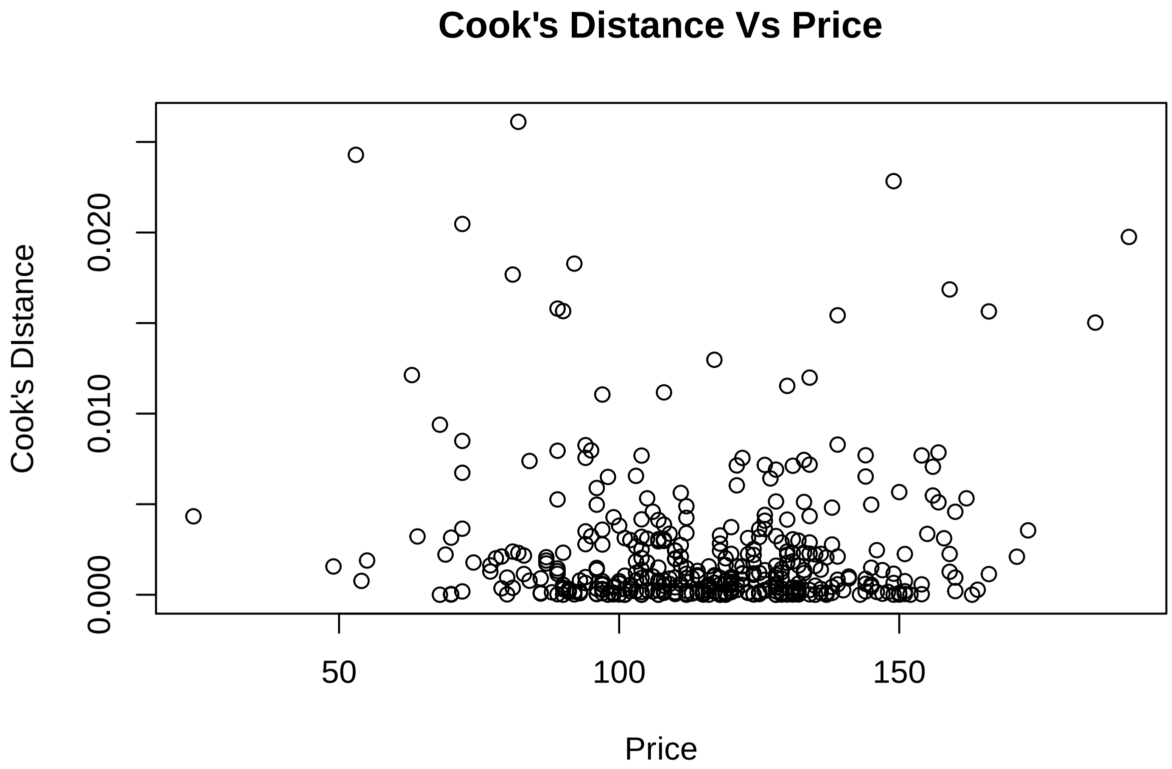
The plot againest the total obseravtions and hat values from the model confirms the above high leverage points.



To verify if these high leverage points are influential we will calculate Cooks distance. If the cooks distance is greater than 1 then the points influence the parameter estimates. Cooks distance give information on how far x-values are from the mean of the x's and how far is y from the regression line. From the plots below it is clear that no observation has a cook's value greater than 1 and hence there are no influential points.

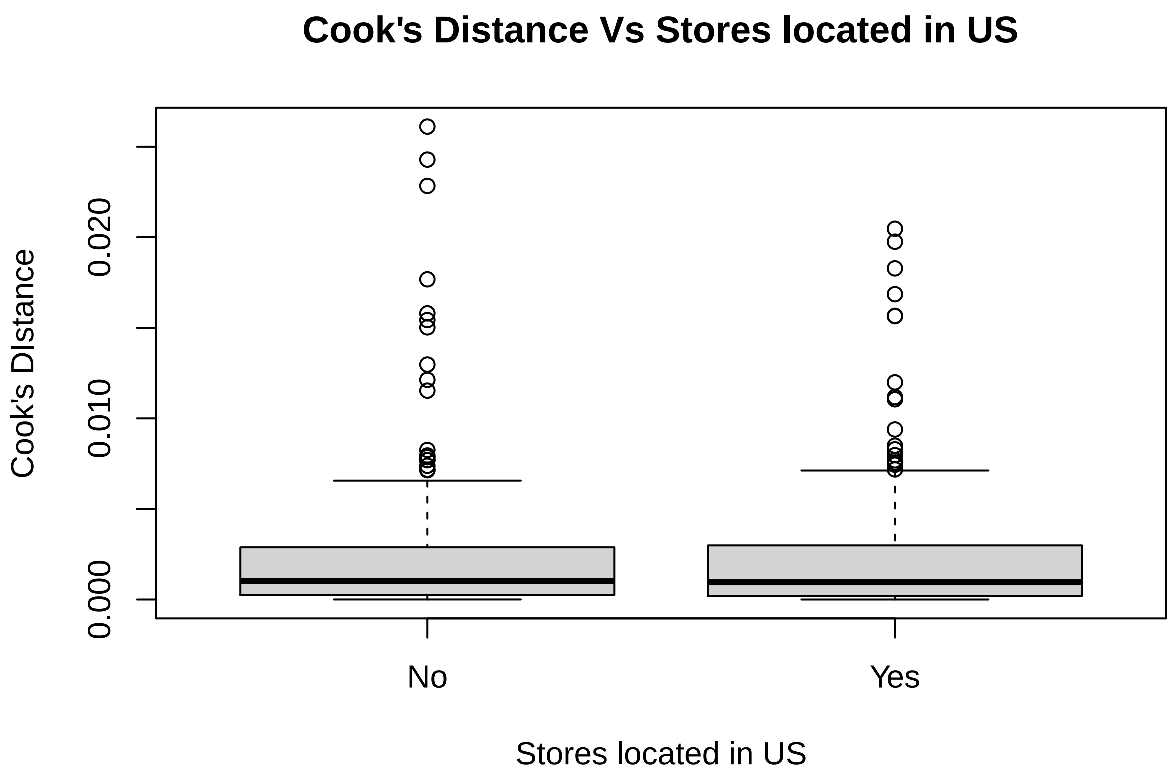


```
## Cooks distance for Price Variable:
```



```
## NULL
```

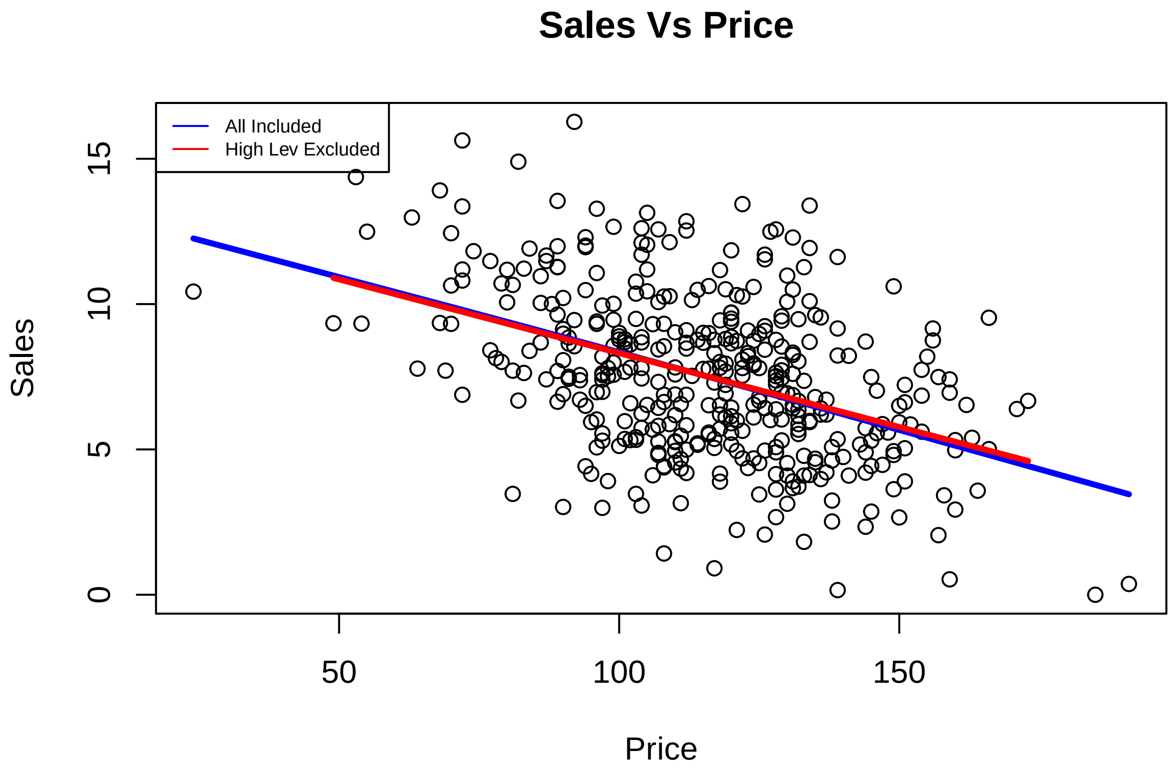
```
## Cooks distance for US Variable:
```



```
## $stats
##              [,1]         [,2]
## [1,] 5.236050e-07 1.343255e-08
## [2,] 2.502538e-04 1.978290e-04
## [3,] 1.011029e-03 9.513108e-04
## [4,] 2.880999e-03 2.987165e-03
## [5,] 6.562834e-03 7.121553e-03
##
## $n
## [1] 142 258
##
## $conf
##              [,1]         [,2]
## [1,] 0.0006622173 0.0006769336
## [2,] 0.0013598417 0.0012256880
##
## $out
##          16          26          31          41          50          57
## 0.007700068 0.026109457 0.015798963 0.007167711 0.022835461 0.007384844
##          58          85         107         157         175         242
## 0.012972301 0.007138939 0.015428866 0.007855441 0.015024314 0.008257311
##         259         271         273         283         299         368
## 0.017677163 0.007950211 0.012128209 0.007690521 0.011531871 0.024287363
##          19          29          45          51          63          69
## 0.009387831 0.011054314 0.007967239 0.011173381 0.007441574 0.011988429
##          83          84         144         166         210         288
## 0.008289860 0.007558845 0.016856674 0.019755042 0.015662819 0.008493346
##         298         305         311         317         353         377
## 0.007682583 0.007181584 0.015643789 0.020470465 0.007552057 0.018282191
##
## $group
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##
## $names
## [1] "No"  "Yes"
```

```
## [1] "Summary of Lineal Model3:"
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = car)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9165 -1.6079 -0.0761  1.5598  7.1066
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.697875   0.672035  18.895  < 2e-16 ***
## Price       -0.051682   0.005553  -9.307  < 2e-16 ***
## USYes        1.220297   0.260747   4.680 3.96e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.463 on 391 degrees of freedom
## Multiple R-squared:  0.2128, Adjusted R-squared:  0.2087
## F-statistic: 52.84 on 2 and 391 DF,  p-value: < 2.2e-16
```



To confirm that the high leverage points are not influential a new linear model is fit after removing the high leverage points. Both the Red (Excluding high leverage points) and Blue(all observations included)regression lines are very close and none of the points influenced the regression lines.Adjusted R Square of Model 3 (Excluding high leverage points is 0.21(Model3) where as adjusted R sqaure of Model 2 with all observations included is 0.24.

In conclusion, there are no outliers that are extreme to the regression line(Q-Q plot). There are few High leverage points that have low influence on the Regression line.

REFERENCES

- Lecture by Jonathan Brown,*Interpreting R Output For Simple Linear Regression Part 1(EPSY 5262)*, Jan 31,2016.
- Lecture by Jonathan Brown,*Interpreting R Output For Simple Linear Regression Part 2(EPSY 5262)*, Feb 8,2016.
- Lecture from jbstatistics,*Leverage and influential points in simple linear regression*,Dec 23,2012.
- Lecture by Phil Chan,*Statistics with R: Example of outlier and leverage analysis part 1 of 3*,Nov 7,2012.
- Lecture by Phil Chan,*Statistics with R: Example of outlier and leverage analysis part 2 of 3*,Nov 7,2012.
- Lecture by Phil Chan,*Statistics with R: Example of outlier and leverage analysis part 3 of 3*,Nov 7,2012.
- PDF document, 22s:152 Applied Linear Regression, University of Iowa, *Chapter 11 Diagnostics: Unusual and Influential Data:Outliers,Leverage, and Influence.*

Question 3.7.15. This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

```
##
## Dimensions of dataset: 506 14
```

```
##
## column names of dataset: crim zn indus chas nox rm age dis rad tax ptratio black lstat medv
```

```
##
## Number of missing values in dataset: 0
```

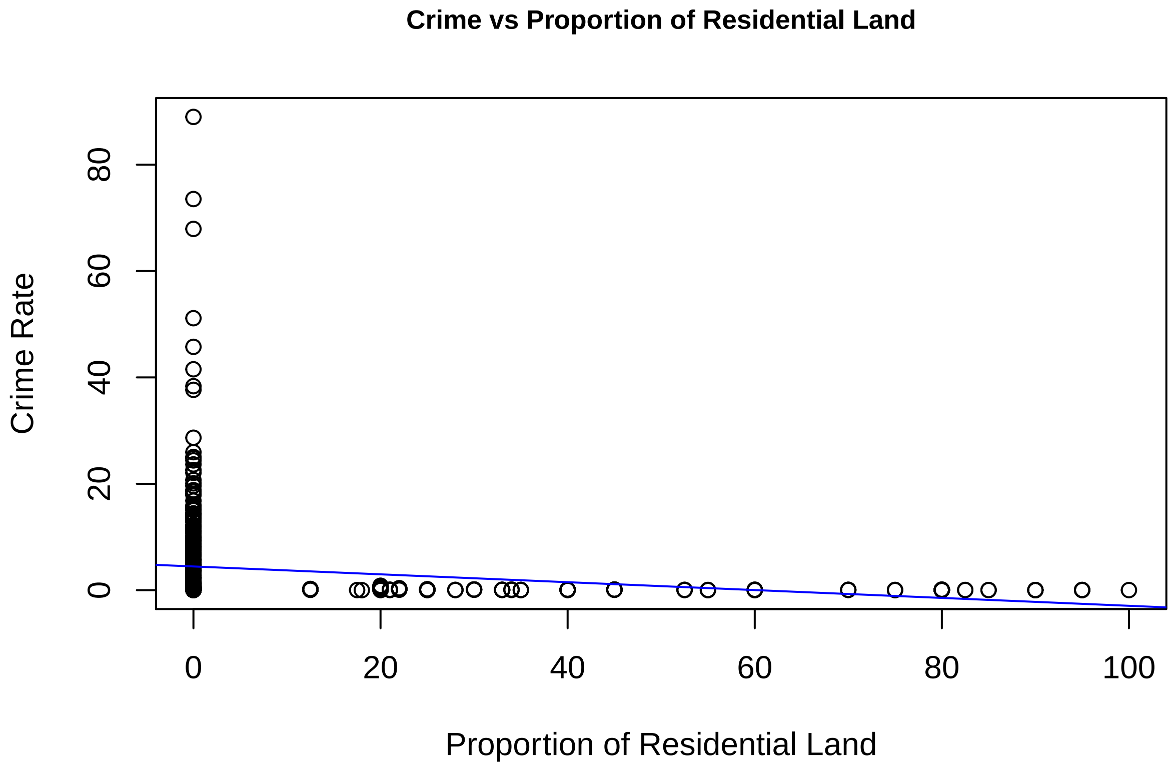
(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

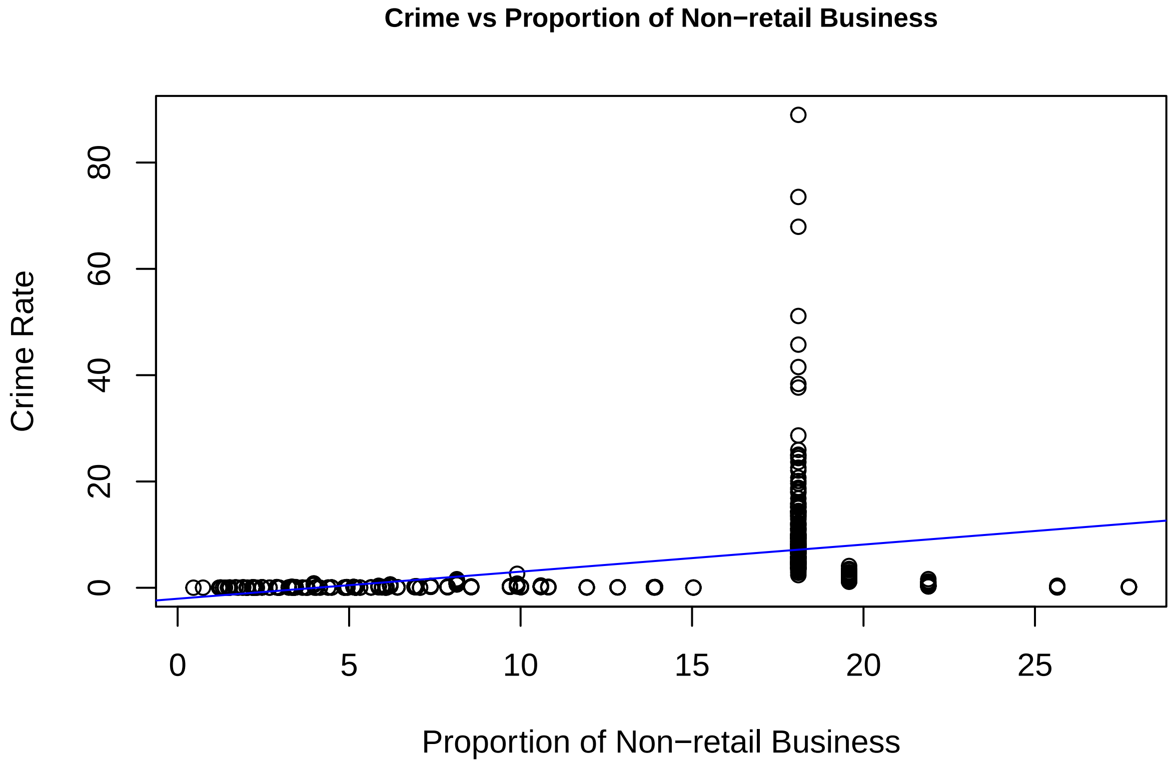
A linear model of form

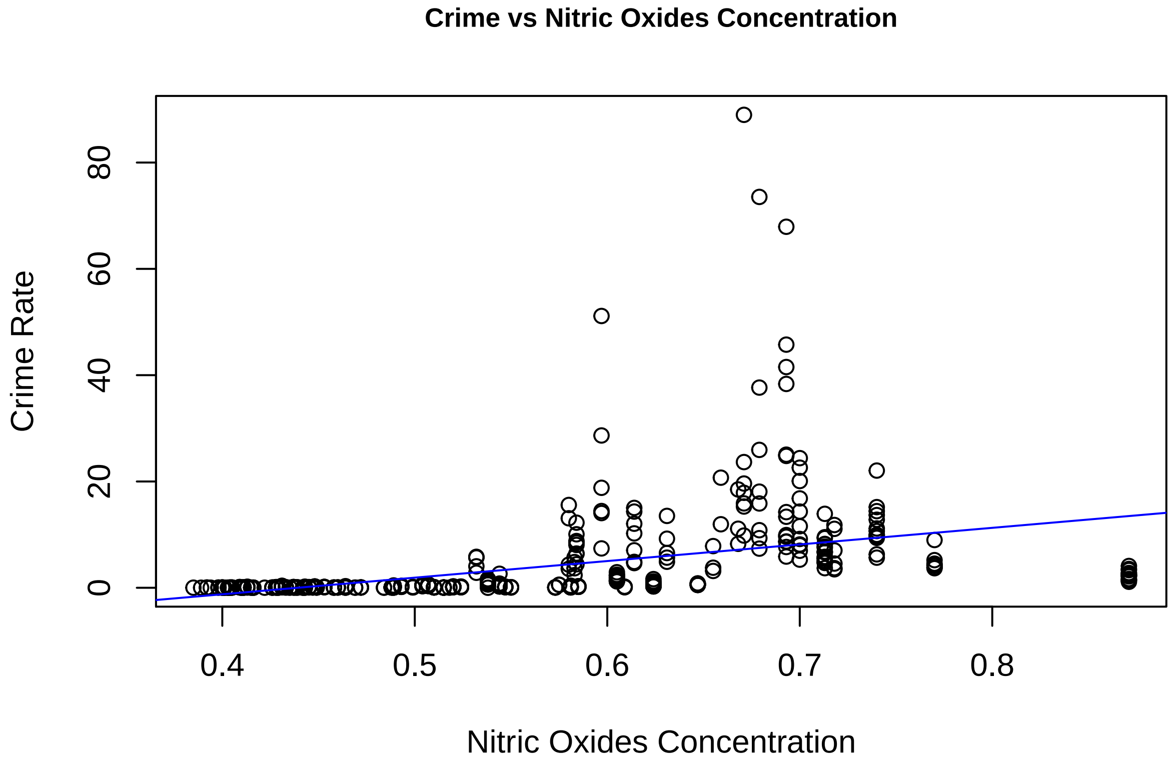
$$lm(y\ x, data = Boston) \tag{2}$$

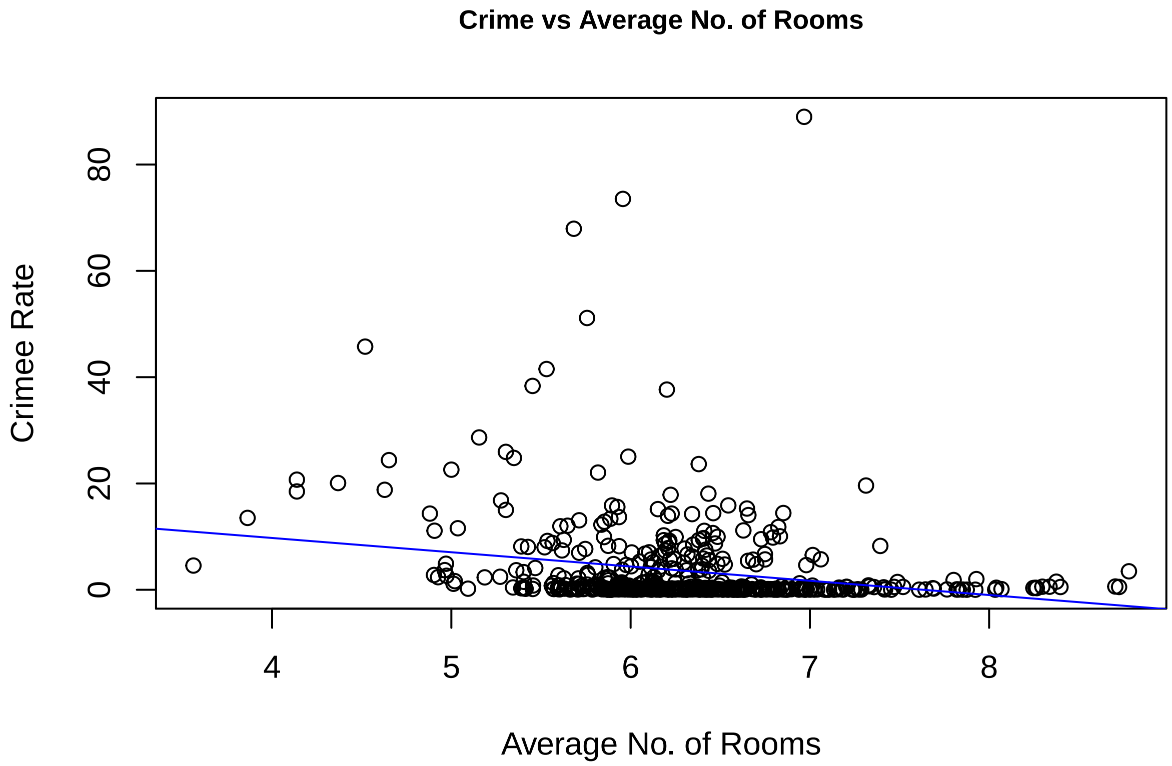
is fit for individual predictors against the Per capita crime rate by town(crim) and the p value of each linear model were extracted and presented in the table. From the p values it is clear that except for the "chas" variable all other predictors have a very low p value at 95% confidence interval and are significant indicating a statistically significant association.

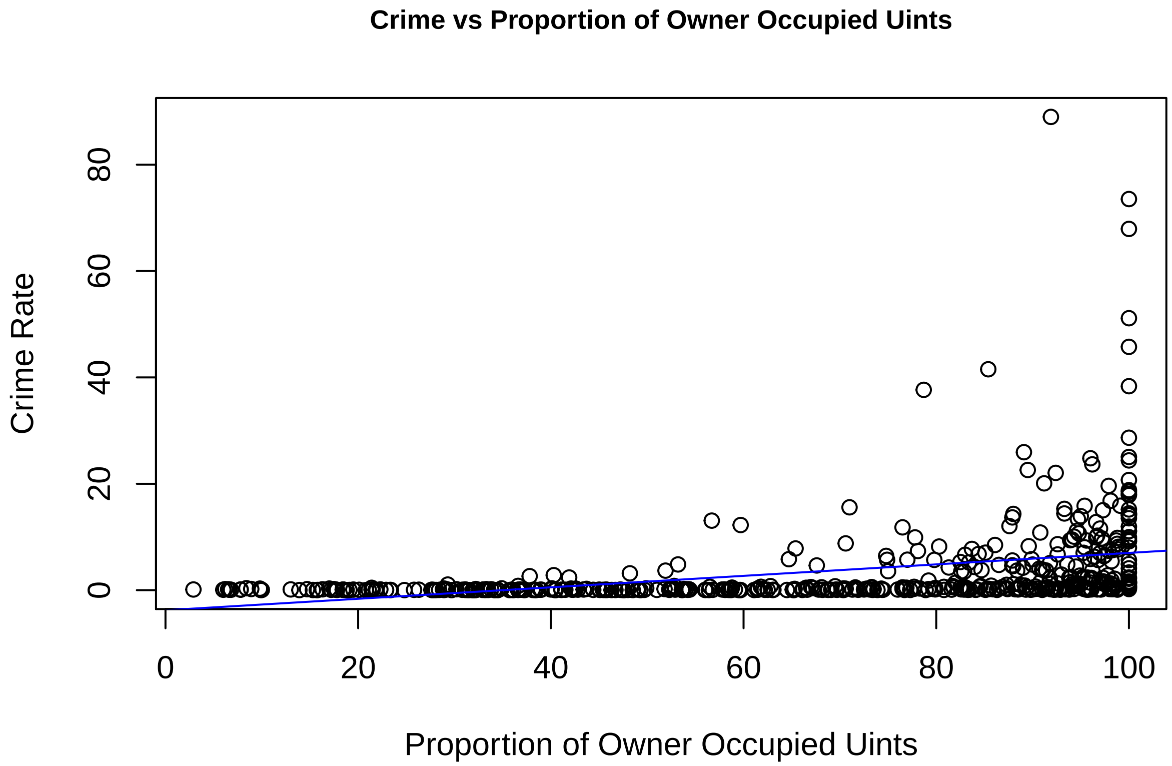
```
##    Predictor     Pvalue    rsq
## 1       chas 0.20943450 0.0031
## 2         zn 0.00000551 0.0402
## 3         rm 0.00000063 0.0481
## 4      indus 0.00000000 0.1653
## 5        nox 0.00000000 0.1772
## 6        age 0.00000000 0.1244
## 7        dis 0.00000000 0.1441
## 8        rad 0.00000000 0.3913
## 9        tax 0.00000000 0.3396
## 10   ptratio 0.00000000 0.0841
## 11     black 0.00000000 0.1483
## 12     lstat 0.00000000 0.2076
## 13      medv 0.00000000 0.1508
## 14      crim         NA     NA
```

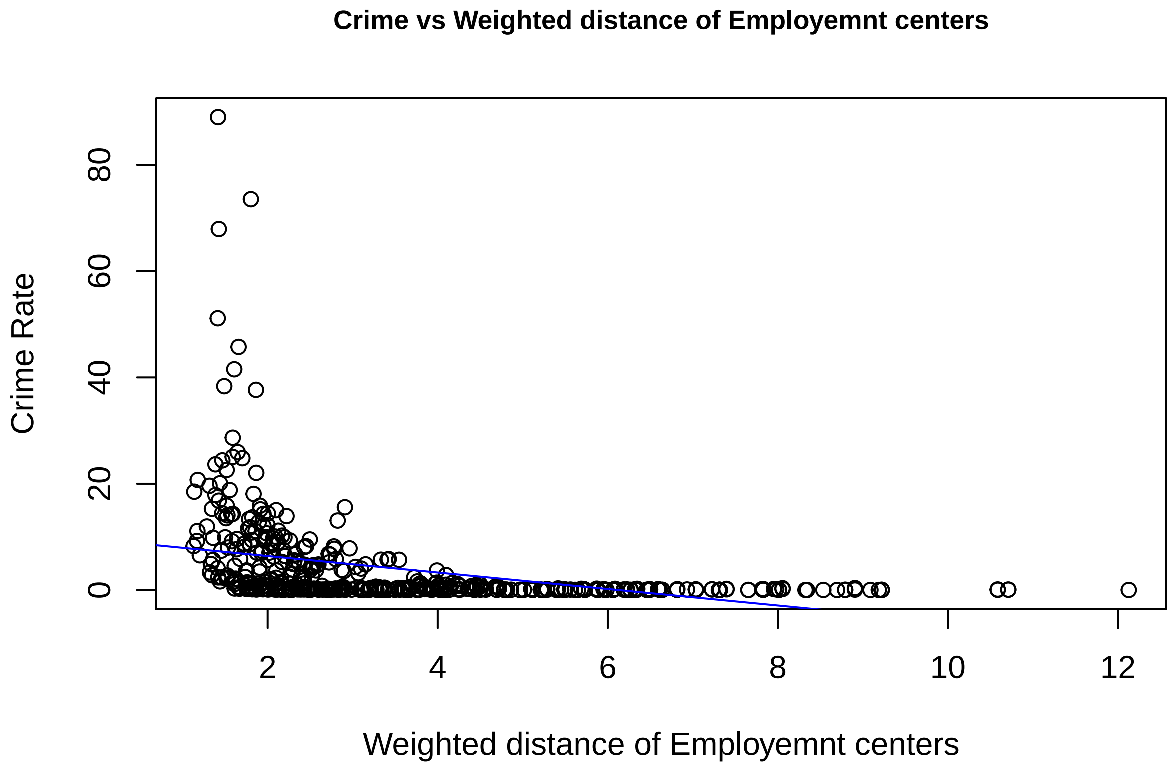

Crime vs Proportion of Residential Land
Crime Rate
0
20
40
60
80
Proportion of Residential Land
0
20
40
60
80
100

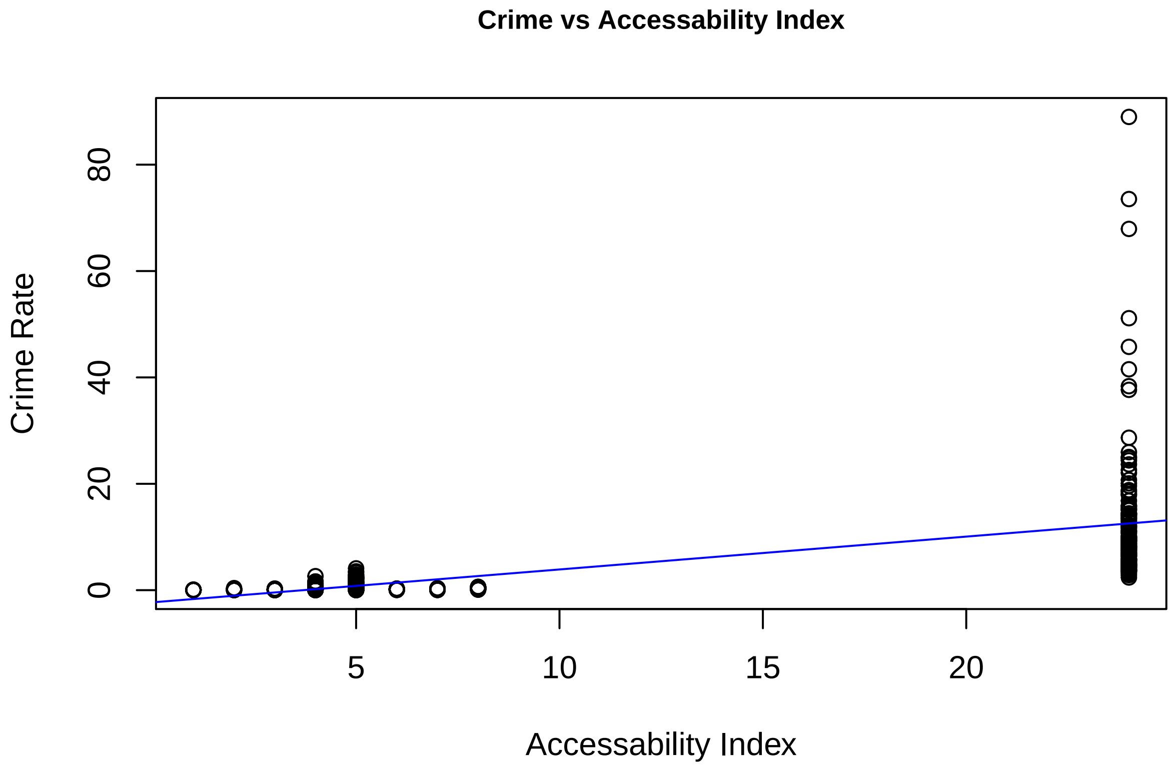

Crime vs Proportion of Non−retail Business
Crime Rate
0
20
40
60
80
0
5
10
15
20
25
Proportion of Non−retail Business

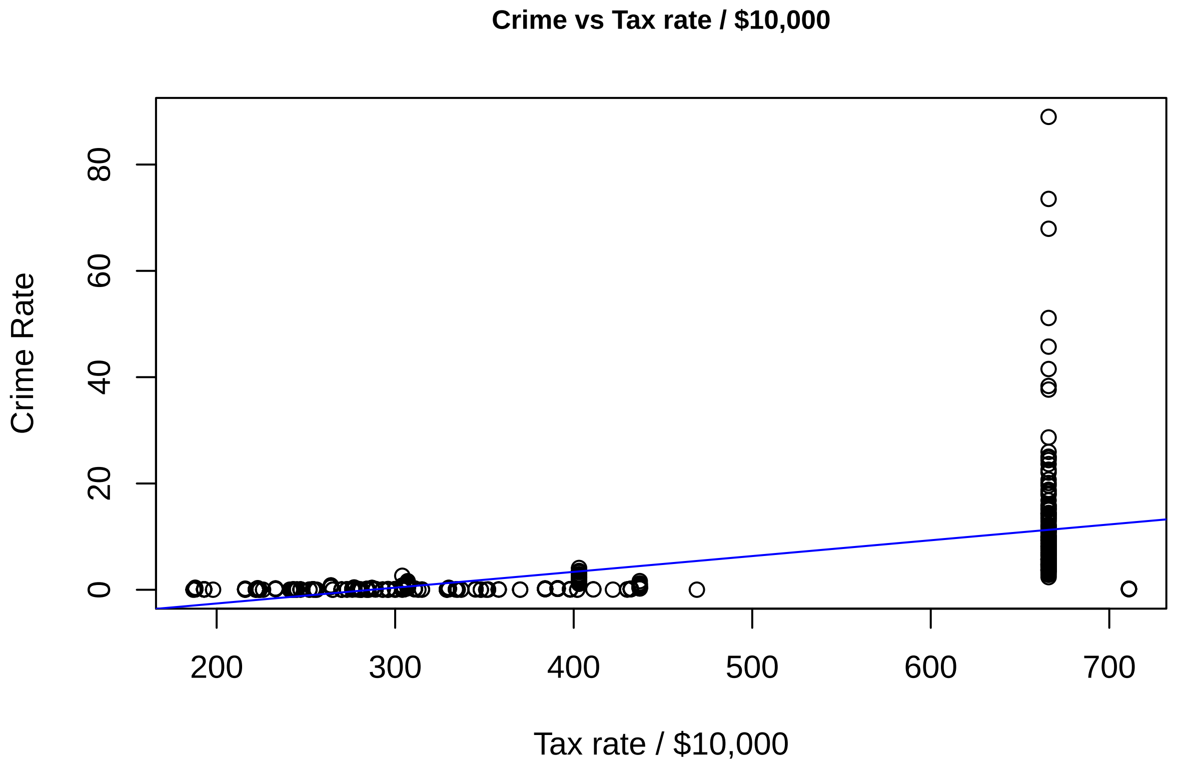

Crime vs Nitric Oxides Concentration
Crime Rate
0
20
40
60
80
0.4
0.5
0.6
0.7
0.8
Nitric Oxides Concentration

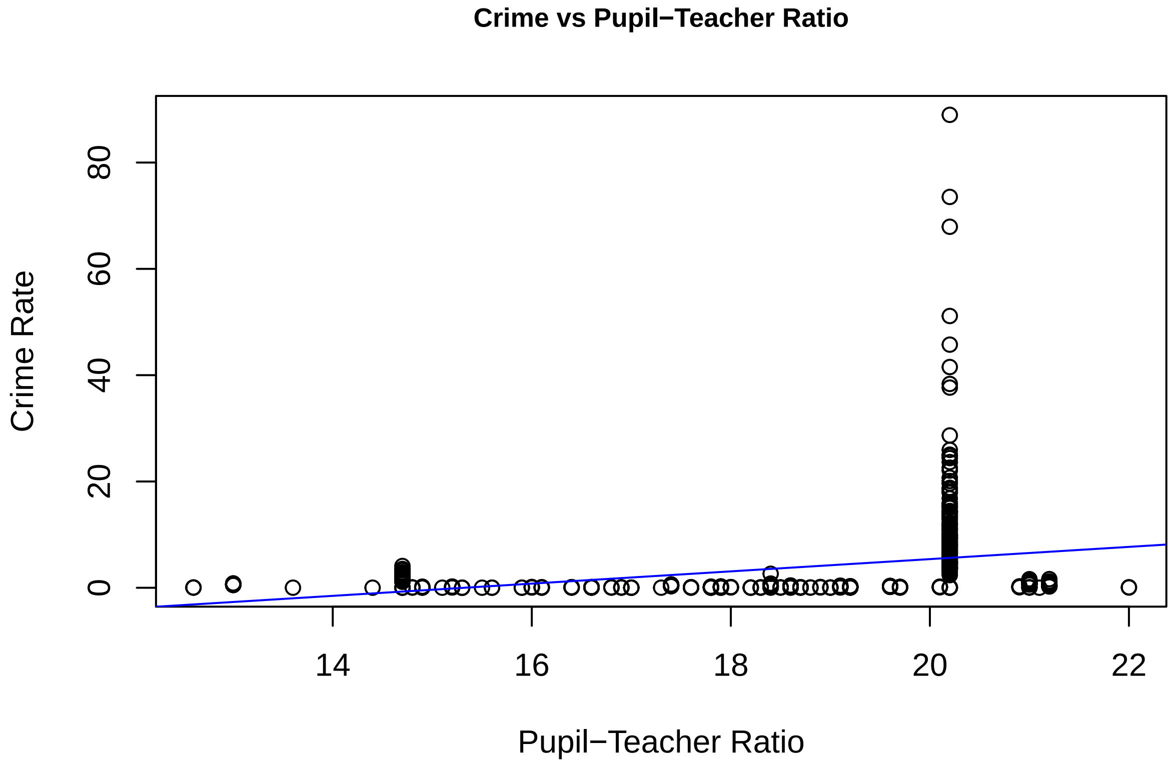

Crime vs Average No. of Rooms
Crimee Rate
0
20
40
60
80
4
5
6
7
8
Average No. of Rooms

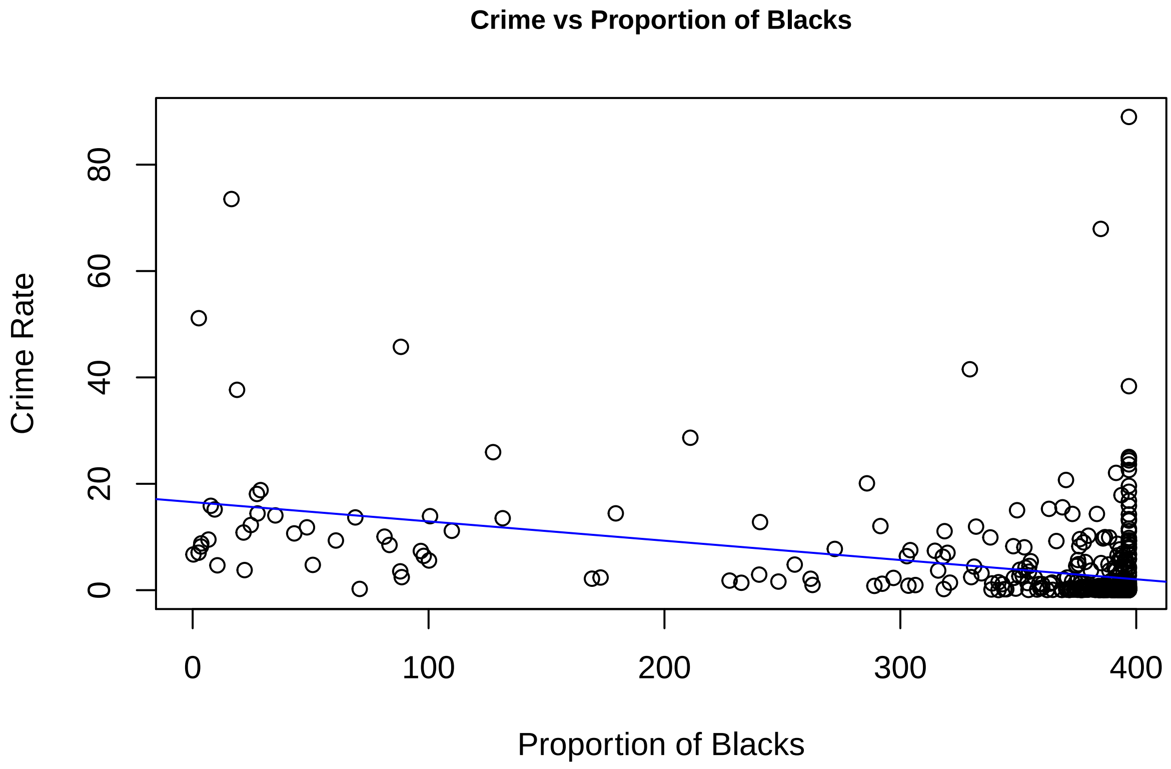

Crime vs Proportion of Owner Occupied Uints
Crime Rate
0
20
40
60
80
0
20
40
60
80
100
Proportion of Owner Occupied Uints

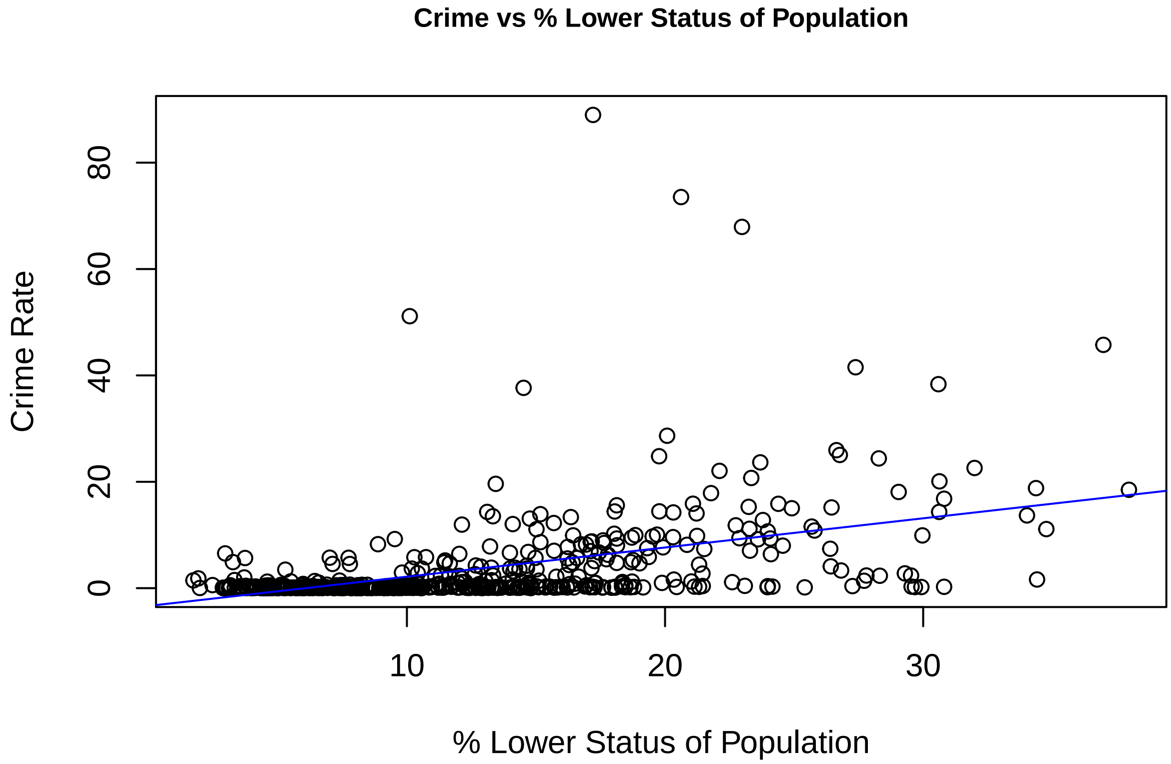

Crime vs Weighted distance of Employemnt centers
Crime Rate
0
20
40
60
80
2
4
6
8
10
12
Weighted distance of Employemnt centers

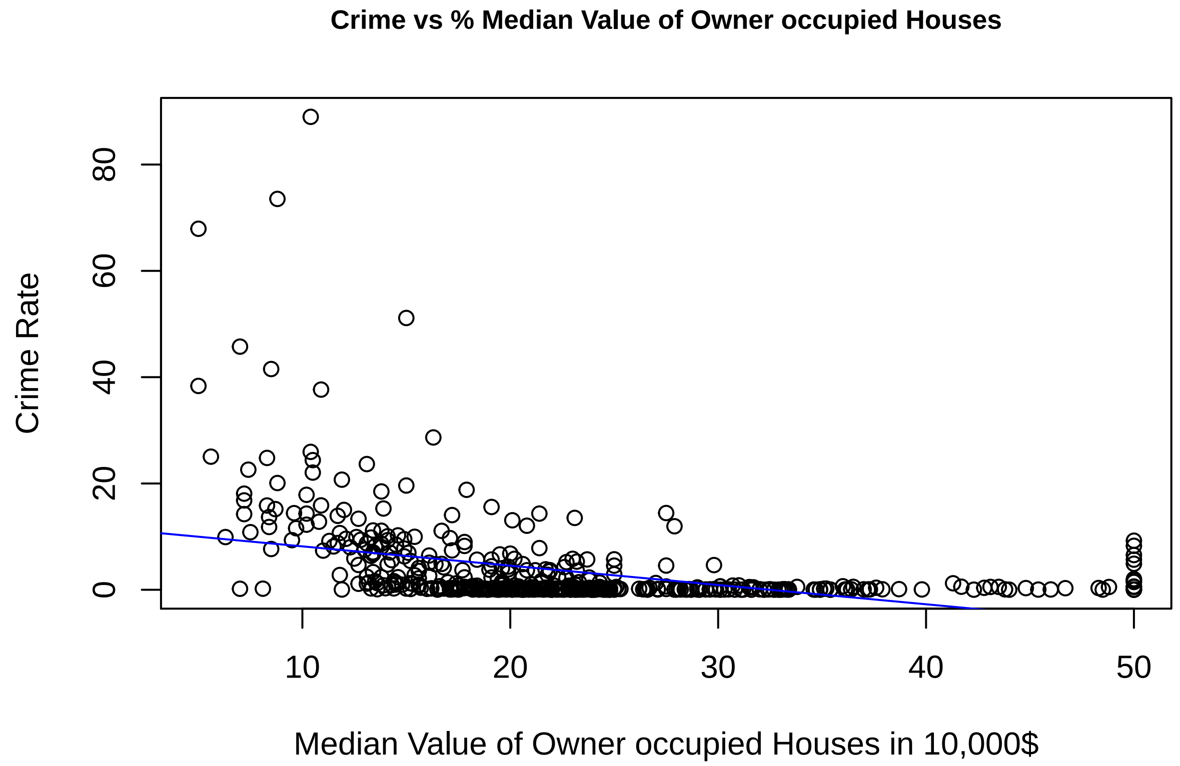

Crime vs Accessability Index
Crime Rate
0
20
40
60
80
5
10
15
20
Accessability Index


Crime vs Tax rate / $10,000
Crime Rate
0
20
40
60
80
200
300
400
500
600
700
Tax rate / $10,000


Crime vs Pupil−Teacher Ratio
Crime Rate
0
20
40
60
80
14
16
18
20
22
Pupil−Teacher Ratio


Crime vs Proportion of Blacks
Crime Rate
0
20
40
60
80
0
100
200
300
400
Proportion of Blacks


Crime vs % Lower Status of Population
Crime Rate
0
20
40
60
80
10
20
30
% Lower Status of Population



The scatter plots of individual predictor and the crime rate show the linear relationship. They have positive and negative slopes.Their relationship is explained by the linear models statistically. The Adjust R square values of these models are very low which are correlating to the regression lines fit over the scatter plots.

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

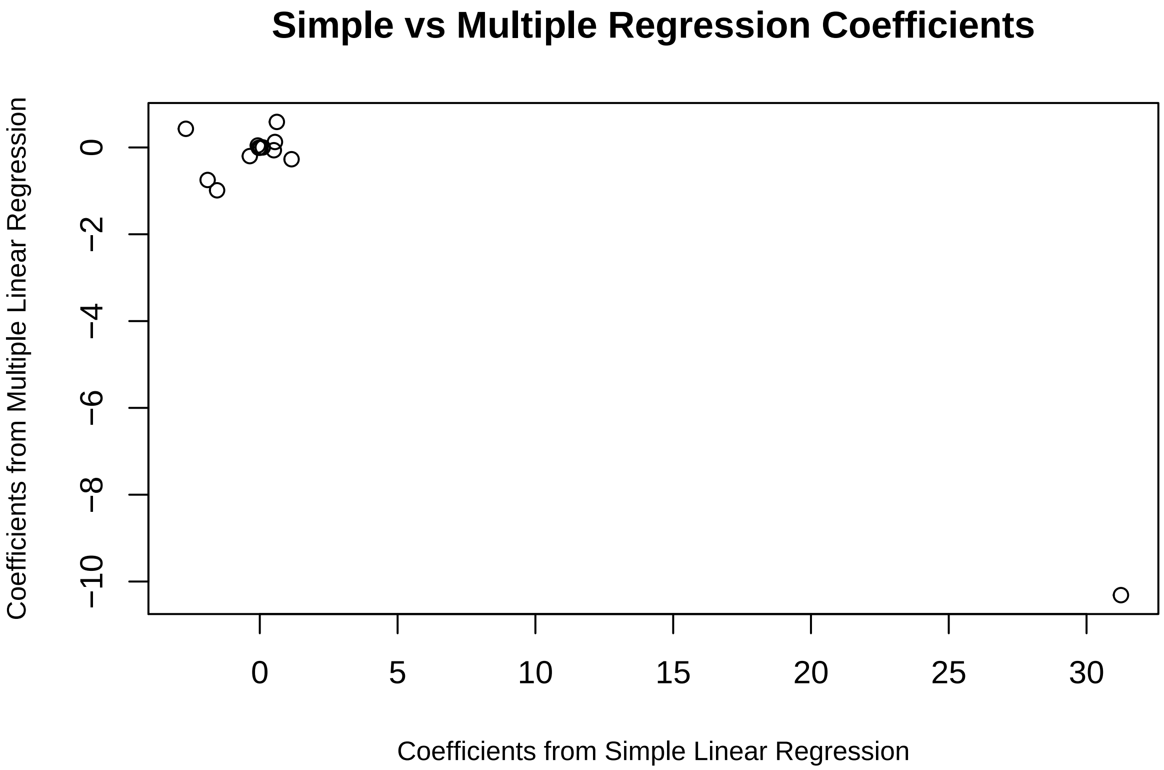
The null hypothesis can be rejected for the predictors *zn*- proportion of residential land zoned for lots over 25,000 sq.ft,*dis*- weighted distances of 5 Boston employment centers,*rad*-index of accessibility to radial highways,*black*-Proportion of Blacks by town and *medv*-median value of owner occupied homes. They have a low p value at 95% confidence level. There is no sufficient evidence to rejct the null hypothesis for remaining predictors.

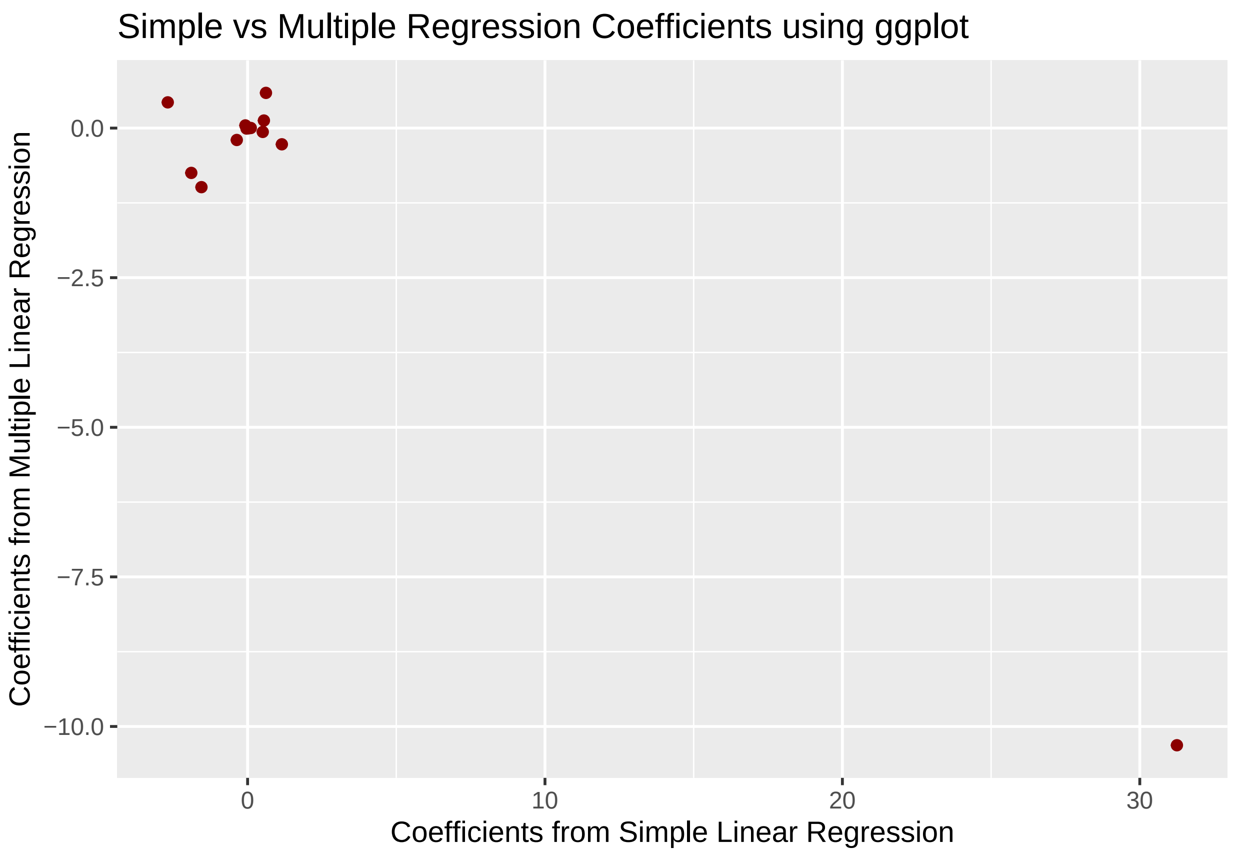
```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

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

The coefficients of the univariate models of all predictors and the coefficients of all predictor from multivariate models are pooled together and presented in the below table.These estimates were plotted by taking Univariate coefficients on X axis and multivariate coefficients on Y axis.Both base R plot and ggplot were fit.

```
##    Predictor uni_coeff Multi_coeff
## 1        nox   31.2485    -10.3135
## 2    ptratio    1.1520     -0.2711
## 3        rad    0.6179      0.5882
## 4      lstat    0.5488      0.1262
## 5      indus    0.5098     -0.0639
## 6        age    0.1078      0.0015
## 7        tax    0.0297     -0.0038
## 8      black   -0.0363     -0.0075
## 9         zn   -0.0739      0.0449
## 10      medv   -0.3632     -0.1989
## 11       dis   -1.5509     -0.9872
## 12      chas   -1.8928     -0.7491
## 13        rm   -2.6841      0.4301
```


Simple vs Multiple Regression Coefficients
Coefficients from Multiple Linear Regression
0
−2
−4
−6
−8
−10
0
5
10
15
20
25
30
Coefficients from Simple Linear Regression



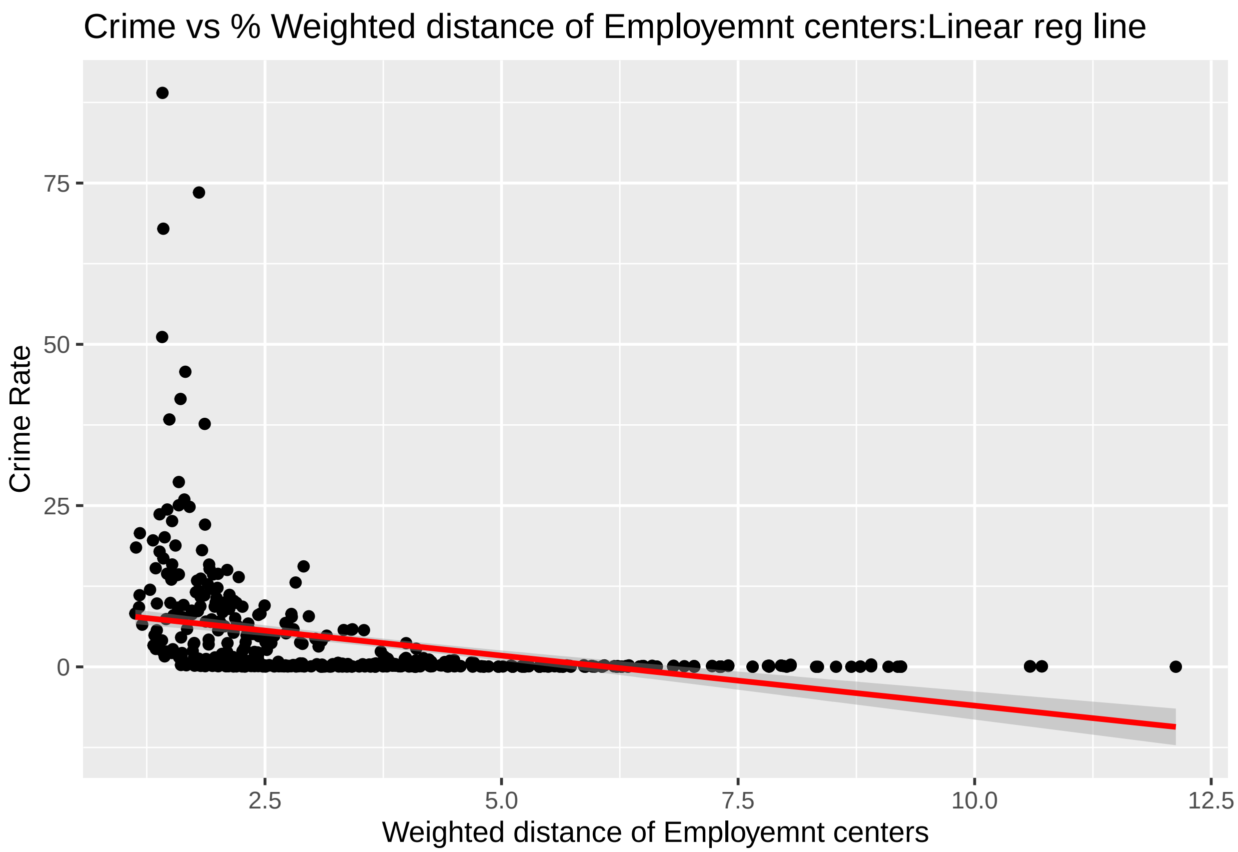
(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form Y = 0 + 1X + 2X2 + 3X3 + .

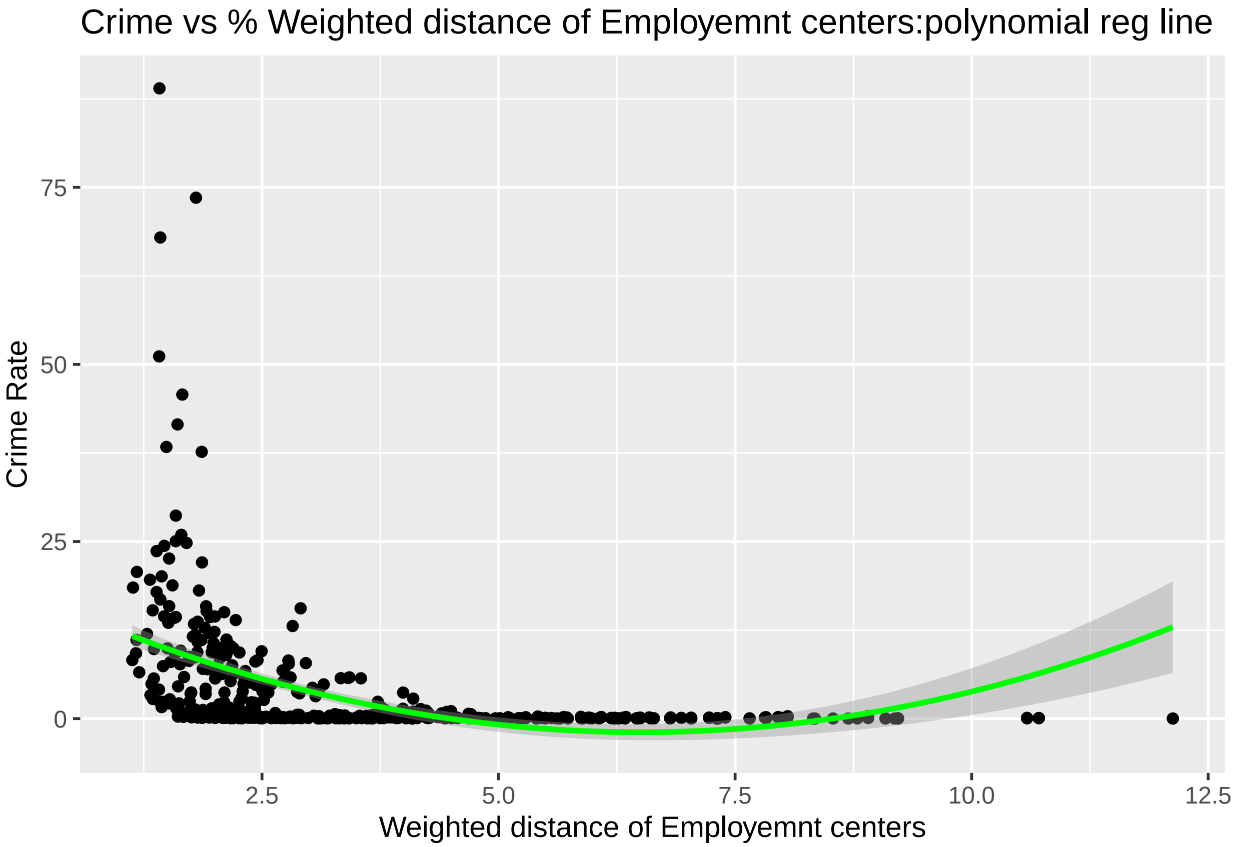
A polynomial model of form

$$lm(y\ x + I(x^2) + I(x^3), data = Boston) \tag{3}$$

is fit individually for all the predictors except chas variable. And the pvalues of all orders of polynomials were pooled together for all the predictors and presented in the below table. Significantly lower pvalues for 3rd order polynomials of indus,nox,medv,dis,ptratio,age variables indicate that there is a non-linear relationship between these variables and crime rate at 95% confidence interval.zn variable has a low p value for the first order parameter inidicating a linear relationship.

```
##    Predictor    P1value    P2value    P3value
## 1      indus 0.00005297 0.00000000 0.00000000
## 2        nox 0.00000000 0.00000000 0.00000000
## 3       medv 0.00000000 0.00000000 0.00000000
## 4        dis 0.00000000 0.00000000 0.00000001
## 5    ptratio 0.00302866 0.00411955 0.00630051
## 6        age 0.14266083 0.04737733 0.00667992
## 7      lstat 0.33452999 0.06458736 0.12989059
## 8         zn 0.00261230 0.09375050 0.22953862
## 9        tax 0.10970752 0.13746816 0.24385068
## 10       rad 0.62341752 0.61300988 0.48231377
## 11        rm 0.21175641 0.36410939 0.50857511
## 12     black 0.13858713 0.47417508 0.54361718
## 13      crim         NA         NA         NA
```


Crime vs % Weighted distance of Employemnt centers:Linear reg line
Crime Rate
75
50
25
0
2.5
5.0
7.5
10.0
12.5
Weighted distance of Employemnt centers



I have compared the linear regression plot and the Polynomial regression plot for dis variable to confirm that the cubic order fit better explains this variable.

References

- Stackoverflow blog,*pull out p-values and r-squared from a linear regression*.
- STHDA Articles-Regression Analysis by kassambara,*Nonlinear Regression Essentials in R: Polynimial and spline Regression Models*, Nov 3,2018.