Regression Modelling in R – Boston Housing Dataset

Table of Contents

FITTING A SIMPLE LINEAR REGRESSION MODEL TO PREDICT THE RESPONSE VARIABLE

The Boston Dataset contains the following 14 variables:

```
Library(MASS)
Data(Boston)
```

```
names(Boston)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
"dis"
 [9] "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

We will fit a linear regression model using lm.fit command between the Per Capita Crime Rate ‘crim’ and each of the other 12 variables separately as follows in R:

Per Capita Crime Rate (crim) and proportion of residential land zoned for lots over 25,000 sq.ft. (zn).

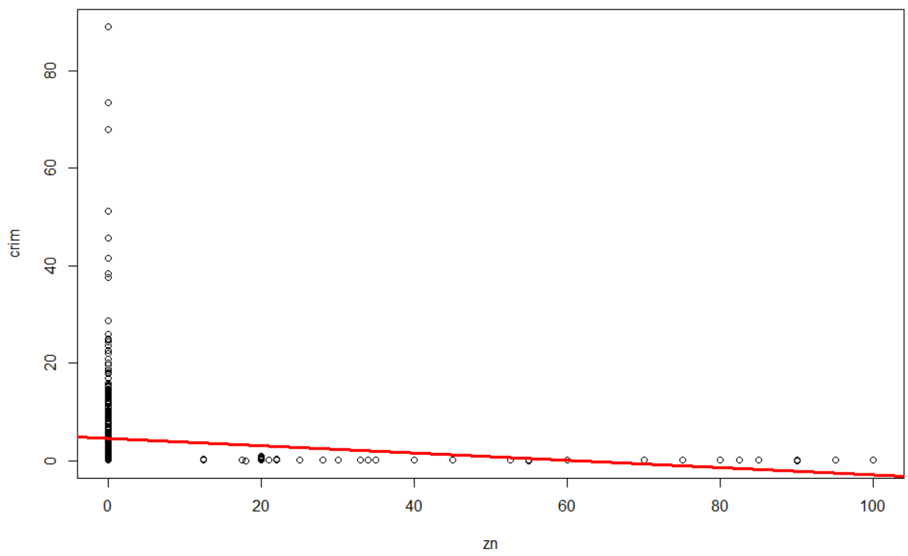
```
lm.fit.zn=lm(crim~zn, data=Boston)
```

```
summary(lm.fit.zn)
```

```
Call:
lm(formula = crim ~ zn)

Residuals:
   Min     1Q Median     3Q    Max 
-4.429 -4.222 -2.620  1.250 84.523 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
zn          -0.07393    0.01609  -4.594 5.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared:  0.04019, Adjusted R-squared:  0.03828 
F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

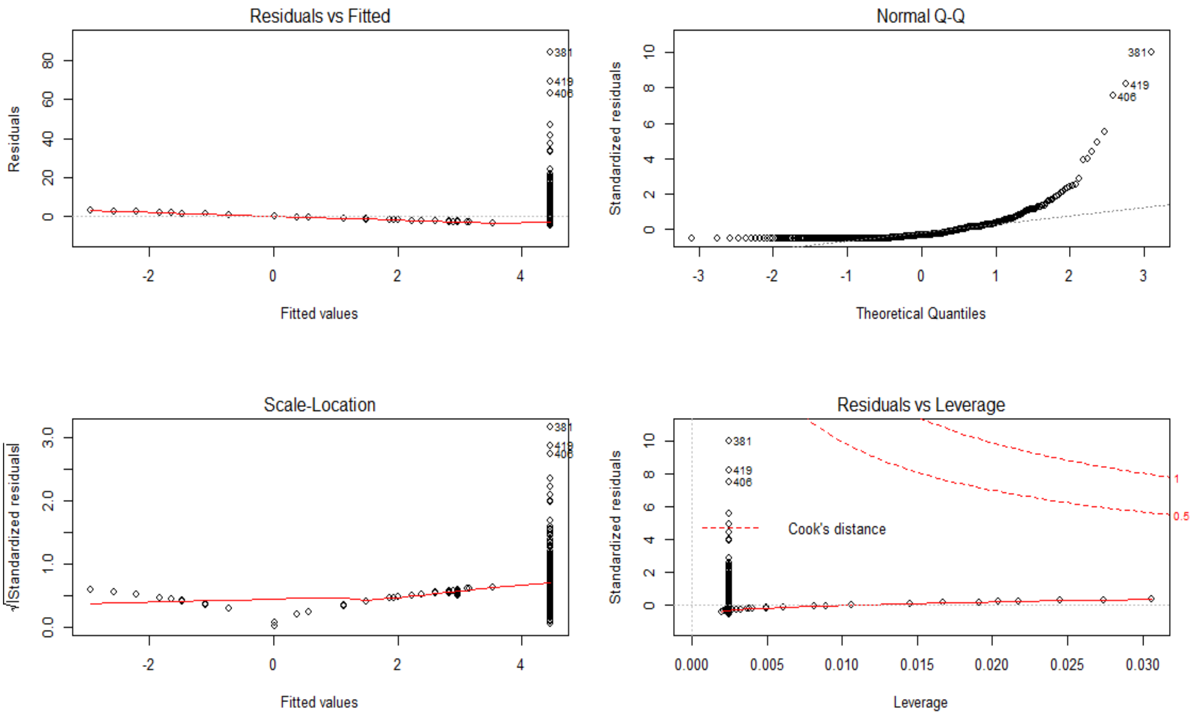
Plot of the linear regression model for crim and zn

```
plot(zn ,crim)
abline(lm.fit.zn,lwd=3,col="red")
```



Now I will examine some diagnostic plots using par (panels)

```
par(mfrow=c(2,2))
plot(lm.fit.zn)
```

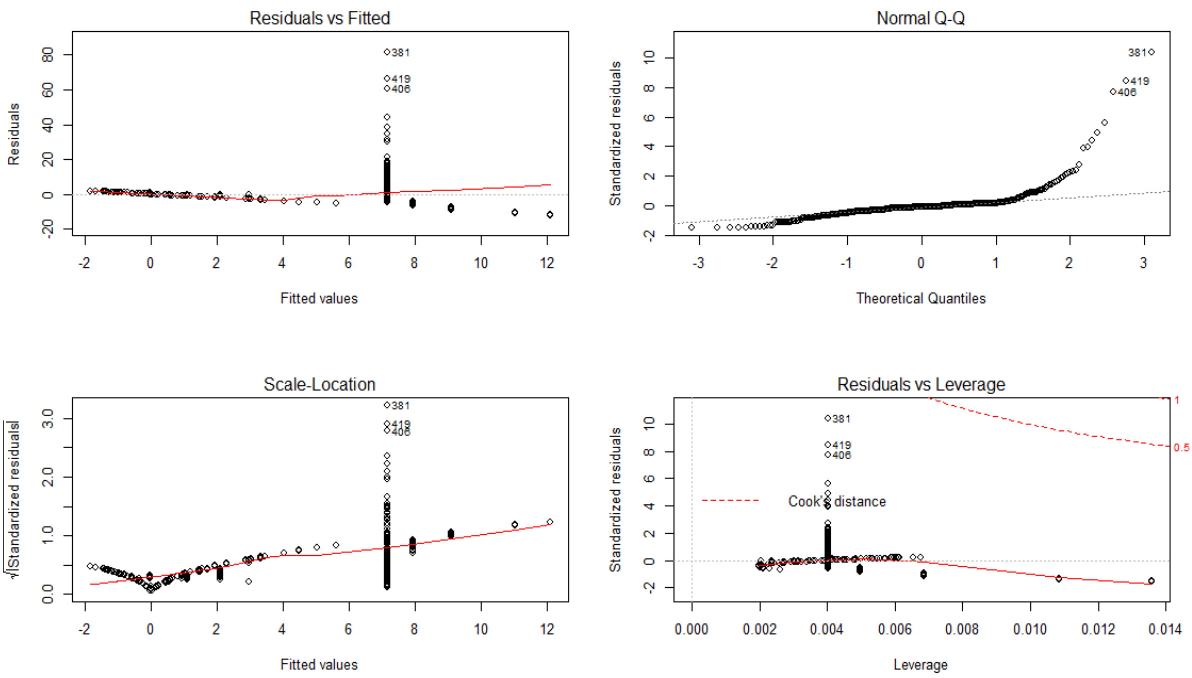


Per Capita Crime Rate (crim) and proportion of non-retail business acres per town (indus).

```
lm.fit.indus=lm(crim~indus, data=Boston)

summary(lm.fit.indus)

Call:
lm(formula = crim ~ indus)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.972  -2.698  -0.736   0.712  81.813 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.06374    0.66723  -3.093  0.00209 ** 
indus        0.50978    0.05102   9.991  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.866 on 504 degrees of freedom
Multiple R-squared:  0.1653,	Adjusted R-squared:  0.1637 
F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diagnostic plots using par (panels):

```
par(mfrow=c(2,2))
plot(lm.fit.indus)
```



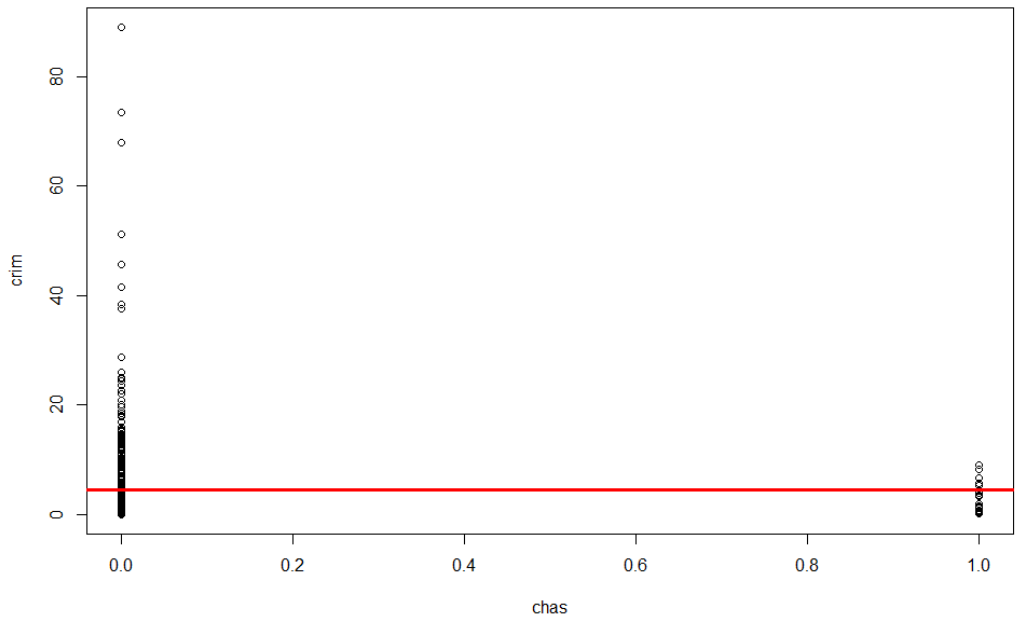
Per Capita Crime Rate (crim) and Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) (chas).

```
lm.fit.chas=lm(crim~chas, data=Boston)

summary(lm.fit.chas)
```

```
Call:
lm(formula = crim ~ chas)

Residuals:
   Min     1Q Median     3Q    Max 
-3.738 -3.661 -3.435  0.018 85.232 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.7444     0.3961   9.453   <2e-16 ***
chas         -1.8928     1.5061  -1.257    0.209    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared:  0.003124, Adjusted R-squared:  0.001146 
F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

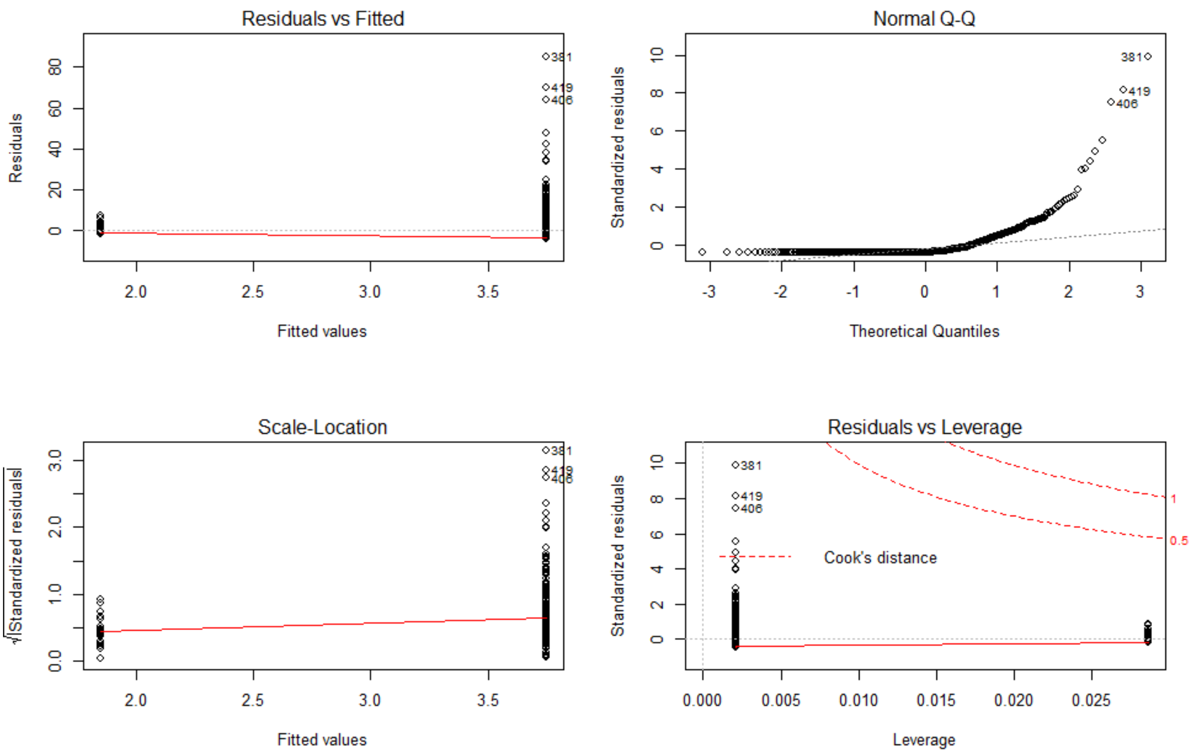
Plot of the linear regression model for crim and chas

```
plot(chas ,crim)
abline(lm.fit.chas,lwd=3,col="red")
```



Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.chas)
```



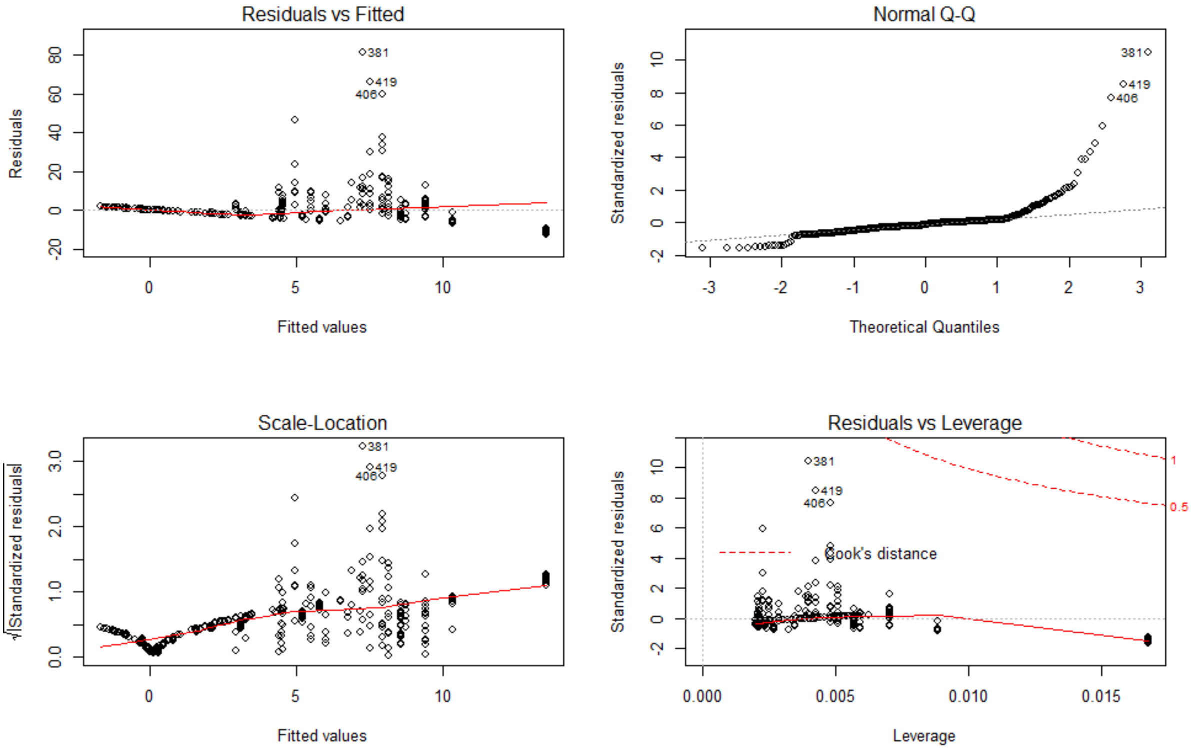
Per Capita Crime Rate (crim) and nitrogen oxides concentration (parts per 10 million) (nox).

```
lm.fit.nox=lm(crim~nox, data=Boston)

summary(lm.fit.nox)
```

```
Call:
lm(formula = crim ~ nox)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.371  -2.738  -0.974   0.559  81.728 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
nox           31.249      2.999  10.419  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.81 on 504 degrees of freedom
Multiple R-squared:  0.1772,	Adjusted R-squared:  0.1756 
F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.nox)
```



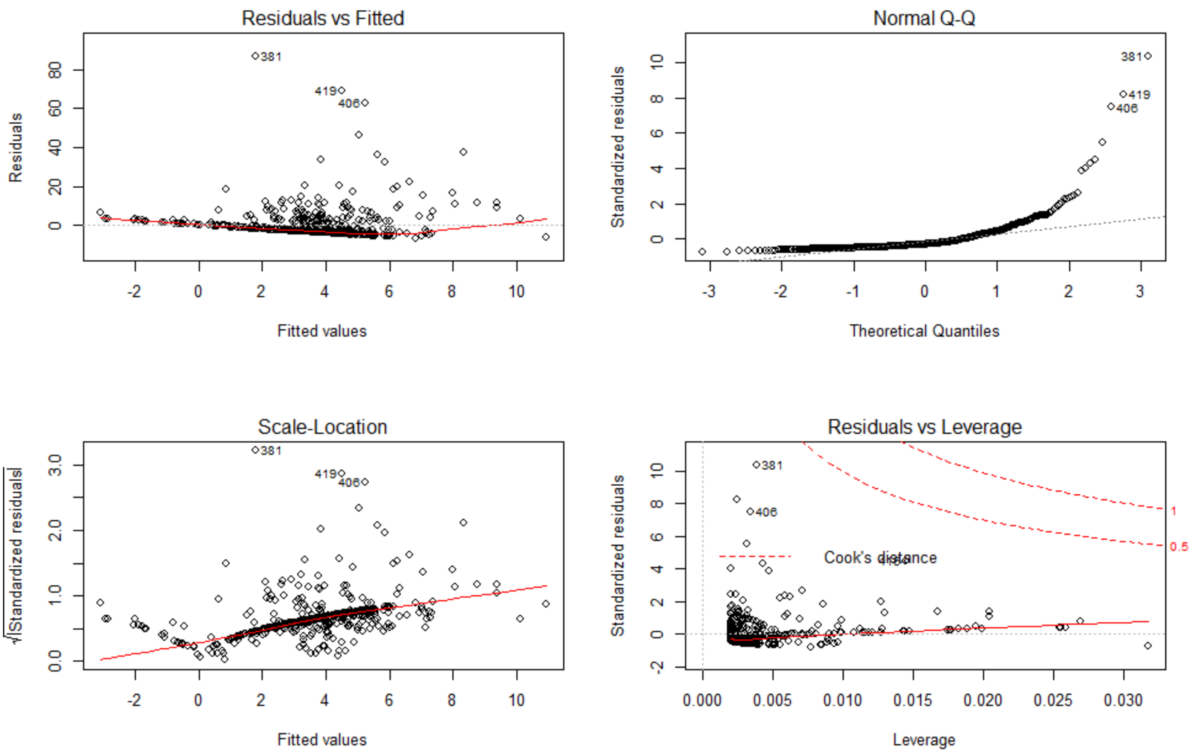
Per Capita Crime Rate (crim) and average number of rooms per dwelling (rm).

```
lm.fit.rm=lm(crim~rm, data=Boston)

summary(lm.fit.rm)
```

```
Call:
lm(formula = crim ~ rm)

Residuals:
   Min     1Q Median     3Q    Max 
-6.604 -3.952 -2.654  0.989 87.197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   20.482      3.365   6.088 2.27e-09 ***
rm            -2.684      0.532  -5.045 6.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.401 on 504 degrees of freedom
Multiple R-squared:  0.04807,	Adjusted R-squared:  0.04618 
F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.rm)
```



Per Capita Crime Rate (crim) and proportion of owner-occupied units built prior to 1940 (age).

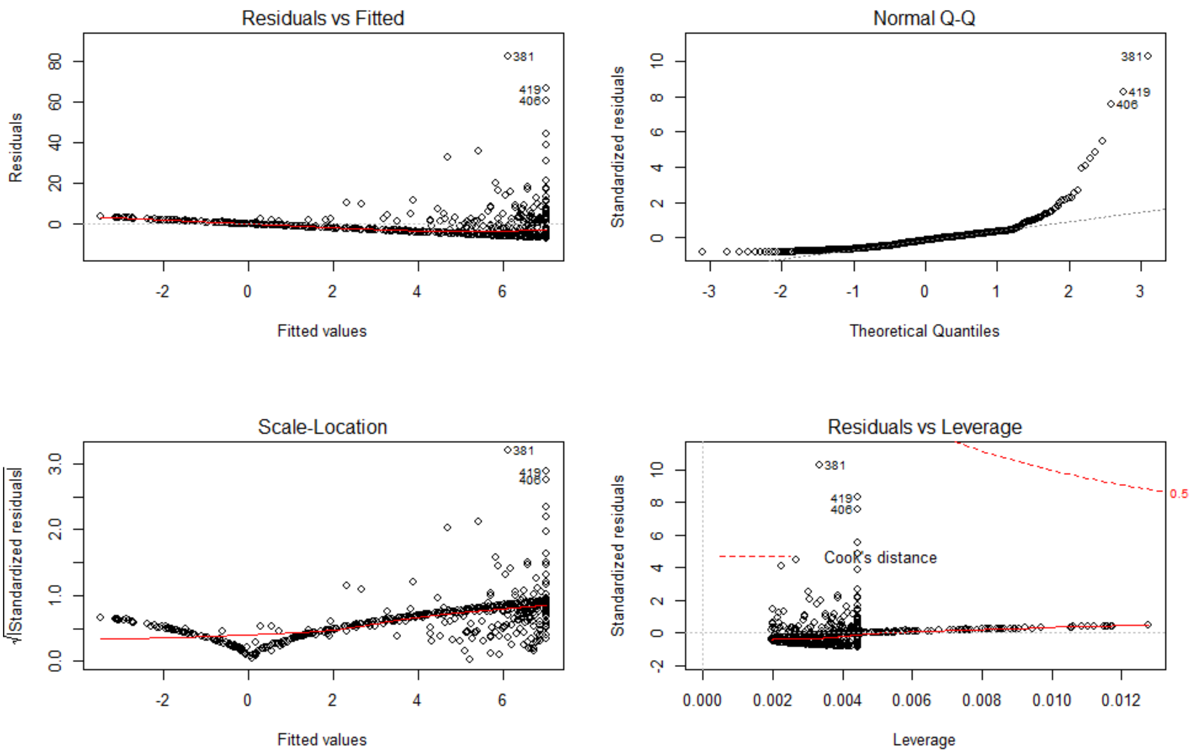
```
lm.fit.age=lm(crim~age, data=Boston)
```

```
summary(lm.fit.age)
```

```
Call:
lm(formula = crim ~ age)

Residuals:
   Min     1Q Median     3Q    Max 
-6.789 -4.257 -1.230  1.527 82.849 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
age          0.10779    0.01274   8.463 2.85e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.age)
```

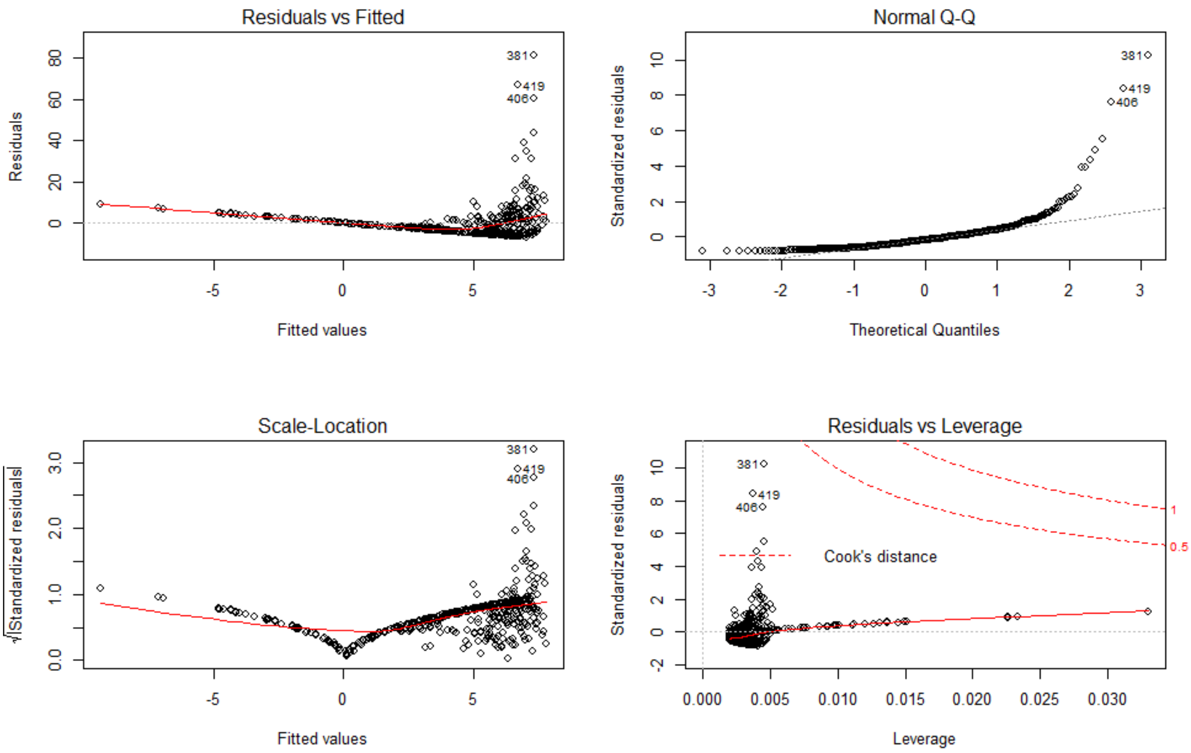


Per Capita Crime Rate (crim) and weighted mean of distances to five Boston employment centres (dis).

```
lm.fit.dis=lm(crim~dis, data=Boston)

summary(lm.fit.dis)

Call:
lm(formula = crim ~ dis)

Residuals:
   Min      1Q Median      3Q     Max 
-6.708 -4.134 -1.527  1.516 81.674 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.4993     0.7304  13.006   <2e-16 ***
dis          -1.5509     0.1683  -9.213   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.965 on 504 degrees of freedom
Multiple R-squared:  0.1441,	Adjusted R-squared:  0.1425 
F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.dis)
```



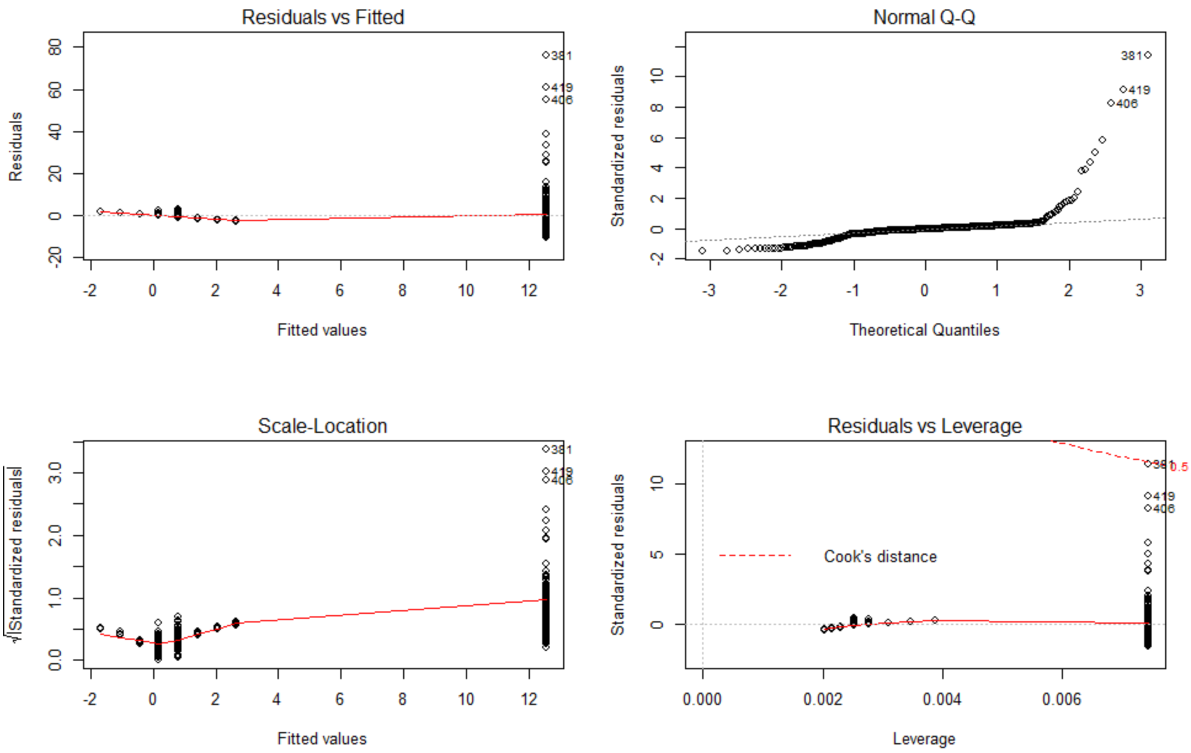
Per Capita Crime Rate (crim) and index of accessibility to radial highways (rad).

```
lm.fit.rad=lm(crim~rad, data=Boston)

summary(lm.fit.rad)
```

```
Call:
lm(formula = crim ~ rad)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.164  -1.381  -0.141   0.660  76.433 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
rad          0.61791    0.03433  17.998  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.718 on 504 degrees of freedom
Multiple R-squared:  0.3913,	Adjusted R-squared:   0.39 
F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.rad)
```



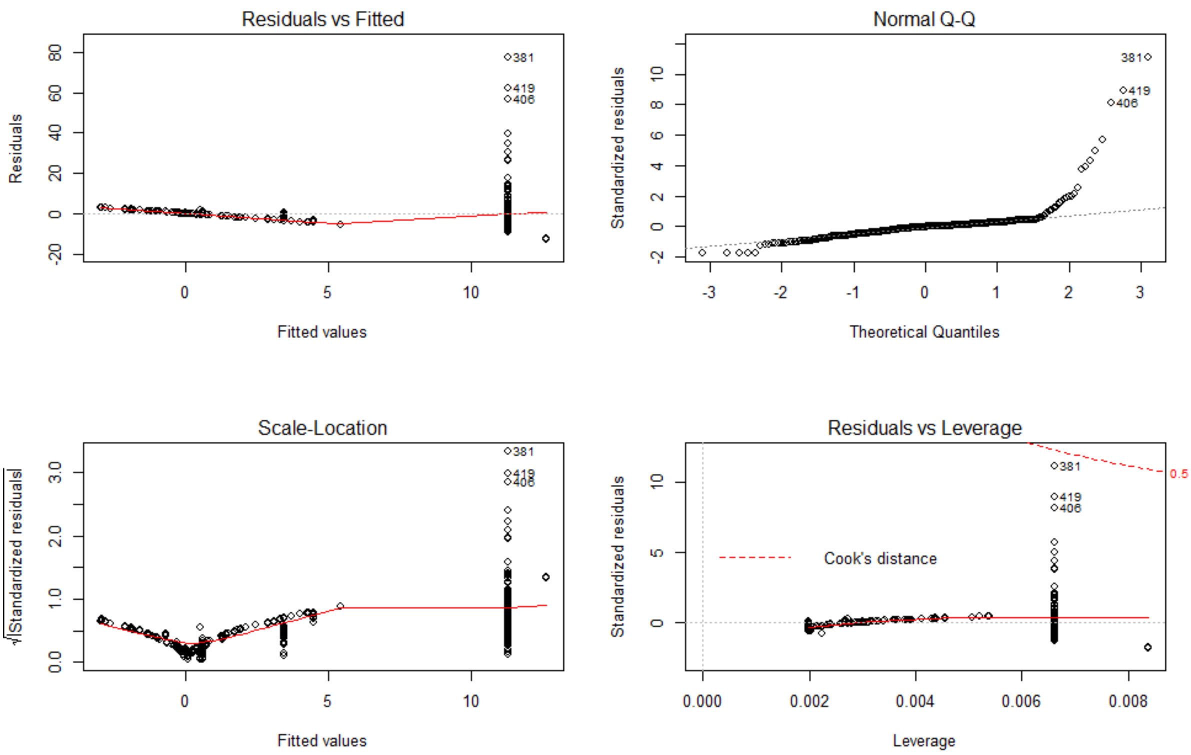
Per Capita Crime Rate (crim) and full-value property-tax rate per $10,000 (tax.

```
lm.fit.tax=lm(crim~tax, data=Boston)

summary(lm.fit.tax)
```

```
Call:
lm(formula = crim ~ tax)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.513  -2.738  -0.194   1.065  77.696 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
tax          0.029742   0.001847   16.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.997 on 504 degrees of freedom
Multiple R-squared:  0.3396,	Adjusted R-squared:  0.3383 
F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.tax)
```

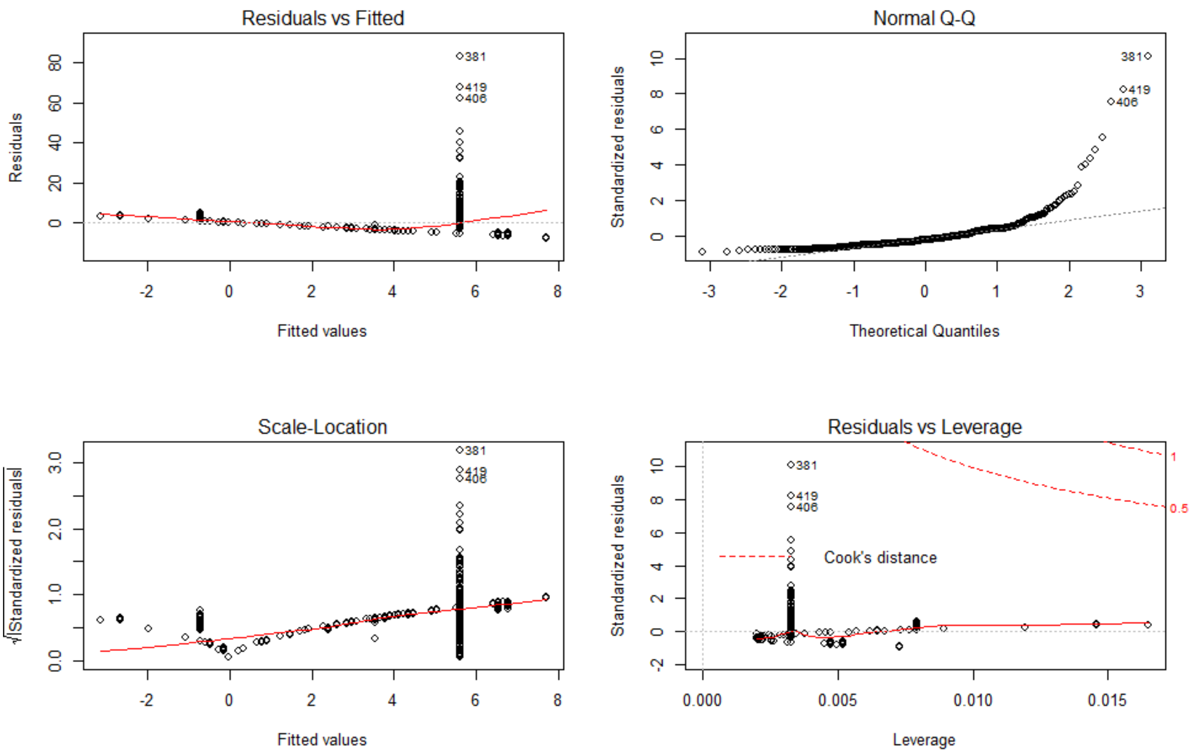


Per Capita Crime Rate (crim) and pupil-teacher ratio by town (ptratio).

```
lm.fit.ptratio=lm(crim~ptratio, data=Boston)

summary(lm.fit.ptratio)

Call:
lm(formula = crim ~ ptratio)

Residuals:
   Min     1Q Median     3Q    Max
-7.654 -3.985 -1.912  1.825 83.353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
ptratio       1.1520     0.1694   6.801 2.94e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.24 on 504 degrees of freedom
Multiple R-squared:  0.08407,  Adjusted R-squared:  0.08225
F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.ptratio)
```

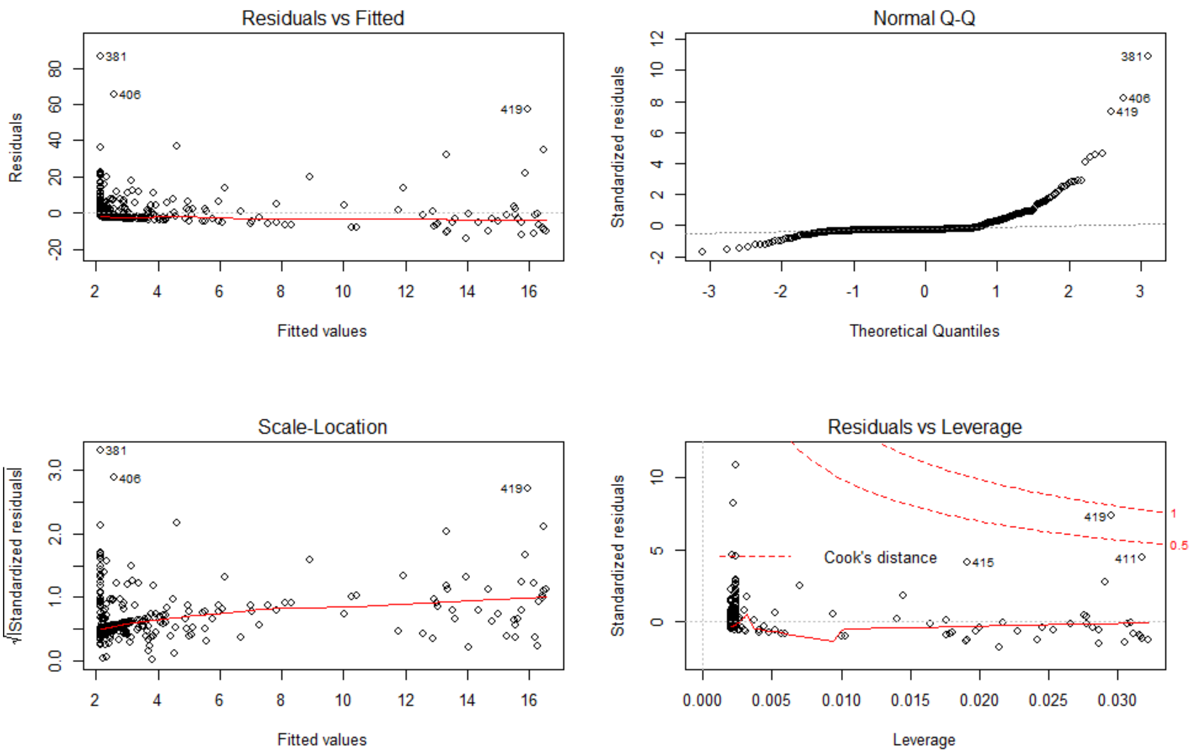


Per Capita Crime Rate (crim) and where Bk is the proportion of blacks by town (black).

```
lm.fit.black=lm(crim~black, data=Boston)

summary(lm.fit.black)

Call:
lm(formula = crim ~ black)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.756  -2.299  -2.095  -1.296  86.822 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.553529   1.425903  11.609   <2e-16 ***
black       -0.036280   0.003873  -9.367   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.946 on 504 degrees of freedom
Multiple R-squared:  0.1483,	Adjusted R-squared:  0.1466 
F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.black)
```



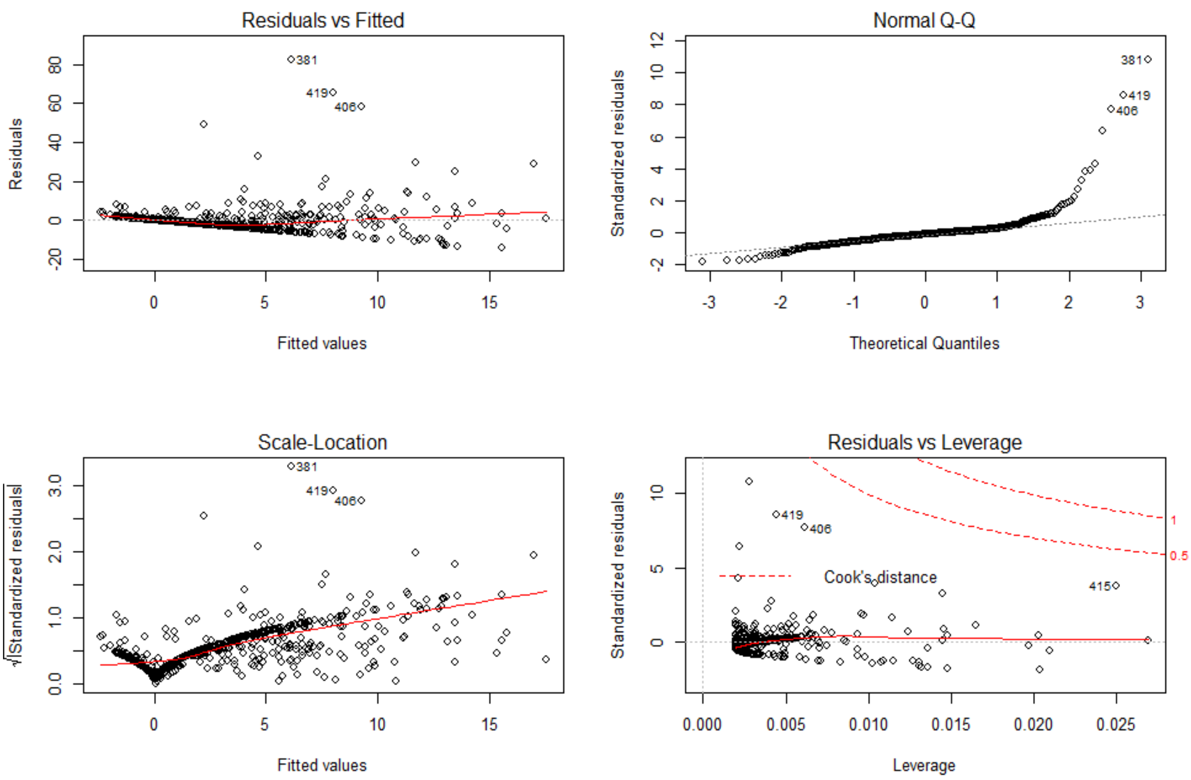
Per Capita Crime Rate (crim) and lower status of the population in percent (lstat).

```
lm.fit.lstat=lm(crim~lstat, data=Boston)

summary(lm.fit.lstat)
```

```
Call:
lm(formula = crim ~ lstat)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.925  -2.822  -0.664   1.079  82.862 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
lstat        0.54880    0.04776  11.491  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.664 on 504 degrees of freedom
Multiple R-squared:  0.2076,	Adjusted R-squared:  0.206 
F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.lstat)
```

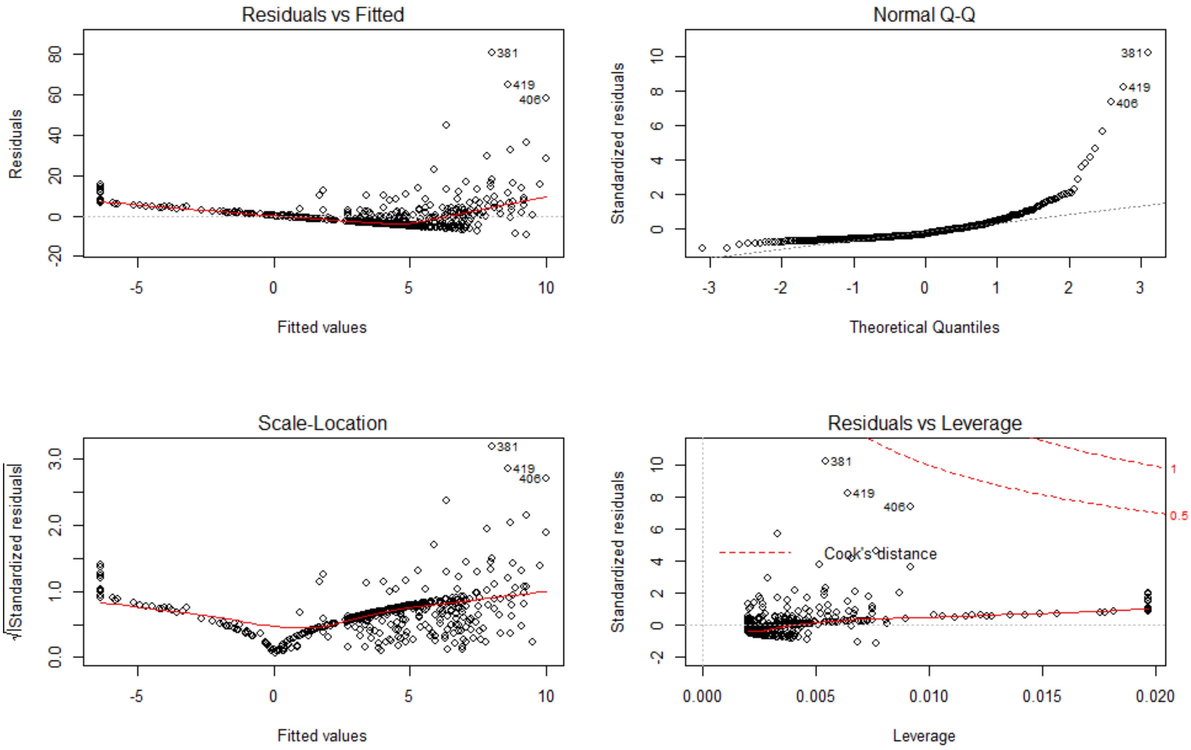


Per Capita Crime Rate (crim) and median value of owner-occupied homes in $1000s (medv).

```
lm.fit.medv=lm(crim~medv, data=Boston)

summary(lm.fit.medv)

Call:
lm(formula = crim ~ medv)

Residuals:
   Min     1Q Median     3Q    Max 
-9.071 -4.022 -2.343  1.298 80.957 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.79654    0.93419   12.63   <2e-16 ***
medv        -0.36316    0.03839   -9.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.934 on 504 degrees of freedom
Multiple R-squared:  0.1508,	Adjusted R-squared:  0.1491 
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(lm.fit.medv)
```



Results summarized as follows:

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Res std. error	Adj R-squared	F-stat	p-value
zn	-0.07393	0.01609	-4.594	5.51e-06 ***	8.435	0.03828	21.1	0.000005506
indus	0.50978	0.05102	9.991	< 2e-16 ***	7.866	0.1637	99.82	< 2.2e-16
chas	-1.8928	1.5061	-1.257	0.209	8.597	0.001146	1.579	0.2094
nox	31.249	2.999	10.419	< 2e-16 ***	7.81	0.1756	108.6	< 2.2e-16
rm	-2.684	0.532	-5.045	6.35e-07 ***	8.401	0.04618	25.45	6.347E-07
age	0.10779	0.01274	8.463	2.85e-16 ***	8.057	0.1227	71.62	2.855E-16
dis	-1.5509	0.1683	-9.213	<2e-16 ***	7.965	0.1425	84.89	< 2.2e-16
rad	0.61791	0.03433	17.998	< 2e-16 ***	6.718	0.39	323.9	< 2.2e-16
tax	0.029742	0.001847	16.1	<2e-16 ***	6.997	0.3383	259.2	< 2.2e-16
ptratio	1.152	0.1694	6.801	2.94e-11 ***	8.24	0.08225	46.26	2.943E-11
black	-0.03628	0.003873	-9.367	<2e-16 ***	7.946	0.1466	87.74	< 2.2e-16
lstat	0.5488	0.04776	11.491	< 2e-16 ***	7.664	0.206	132	< 2.2e-16
medv	-0.36316	0.03839	-9.46	<2e-16 ***	7.934	0.1491	89.49	< 2.2e-16

Comments:

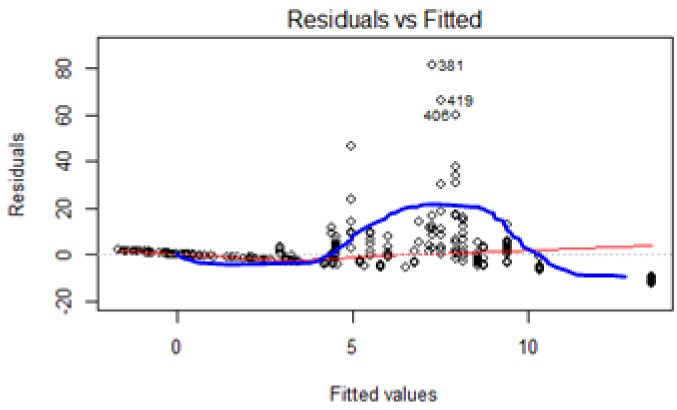
At a quick glance of the p-values we can see that all of the predictors except for 'chas' have a statistically significant association between them and the response 'crim' (at least in isolation). Chas's p-value of 0.209 falls into the acceptance region and therefore we would accept the null hypothesis for this model – it is statistically insignificant.

Furthermore, crim has the smallest R-squared value and F-Statistic of all the variables fitted against crim. Chas has an R-squared value of 0.001146 which is very close to zero meaning this linear model is not a good fit for crim against chas at all.

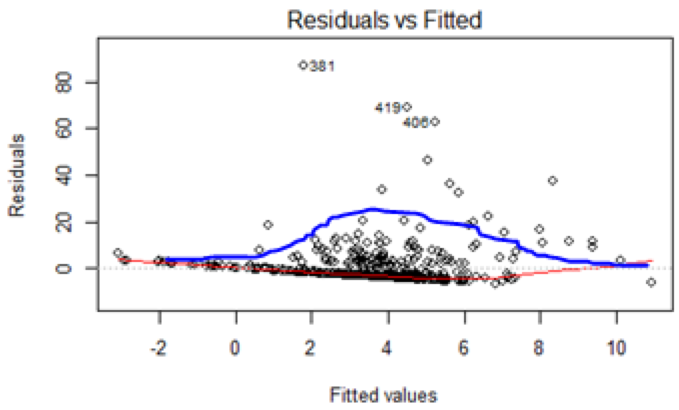
However, the R-squared values for all the other variables are also quite low and therefore this would indicate that these variables may also only describe a small % of variation in the response variable crim. We must be careful then with our initial assumption that all the other variables bar chas have a linear relationship, further investigating will be required.

Looking at the residual plots for each variable against crim some show collinearity but there are some that do not and show a large number of outliers and have some shape. Some of these variables are as follows (the blue line denoting the shape of the data):

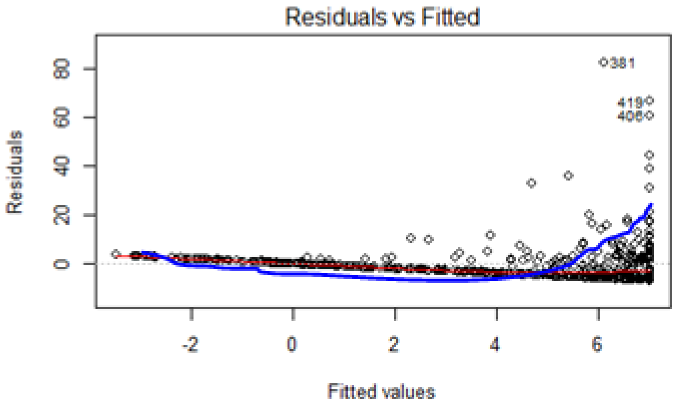
Nox Residual plot:



Rm Residual plot:



Age Residual plot:



FITTING A MULTIPLE REGRESSION MODEL TO PREDICT THE RESPONSE VARIABLE – USING ALL THE PREDICTORS.

```
lm.fit.mul=lm(crim~.,data=Boston)

summary(lm.fit.mul)

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
```

Comments:

Examining the p-values for each variable we can see that the null hypothesis can be rejected for only a few of the variables now: zn, dis, rad, black, and medv as these p-values are less than 0.05 (95% confidence level, typical CI used).

The R-squared value of 0.4396 for this multiple model is generally much higher than the R-squared results for each of the simple linear models we ran before meaning we can now explain a higher % of variance in the response crim using this multiple regression model.

As it seems only a small subset of predictors are helping to explain the response let's remove the insignificant variables and run the multiple model again against zn, dis, rad, black and medv only.

```
lm.fit.multi2=lm(crim~zn+dis+rad+black+medv,data=Boston)
summary(lm.fit.multi2)

Call:
lm(formula = crim ~ zn + dis + rad + black + medv, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.553  -1.869  -0.358   0.839  75.744 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.919933   1.778986   4.452 1.05e-05 ***
zn           0.051799   0.017329   2.989 0.002935 ** 
dis         -0.672189   0.202939  -3.312 0.000992 ***
rad          0.472306   0.042102  11.218  < 2e-16 ***
black       -0.008211   0.003615  -2.271 0.023562 *  
medv        -0.174219   0.036295  -4.800 2.10e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.473 on 500 degrees of freedom
Multiple R-squared:  0.4393,   Adjusted R-squared:  0.4337
F-statistic: 78.34 on 5 and 500 DF,  p-value: < 2.2e-16
```

All 5 variables are still showing signs of significance (very low p-values) and we have also improved the F-statistic on the first multiple model; from 31.47 to 78.34. We have also produced a very similar Adjusted R-squared value by removing the insignificant variables thus showing that all the insignificant variables were in fact that.

COMPARING LINEAR REGRESSION AND MULTIPLE REGRESSION RESULTS.

First, we output the coefficients for the variables under the simple liner model and then all the coefficients for the multiple model as follows in R:

```
coef(lm.fit.age)
coef(lm.fit.black)
coef(lm.fit.chas)
coef(lm.fit.dis)
coef(lm.fit.indus)
coef(lm.fit.lstat)
coef(lm.fit.medv)
coef(lm.fit.nox)
coef(lm.fit.ptratio)
coef(lm.fit.rad)
coef(lm.fit.rm)
coef(lm.fit.tax)
coef(lm.fit.zn)

coef(lm.fit.mul)
```

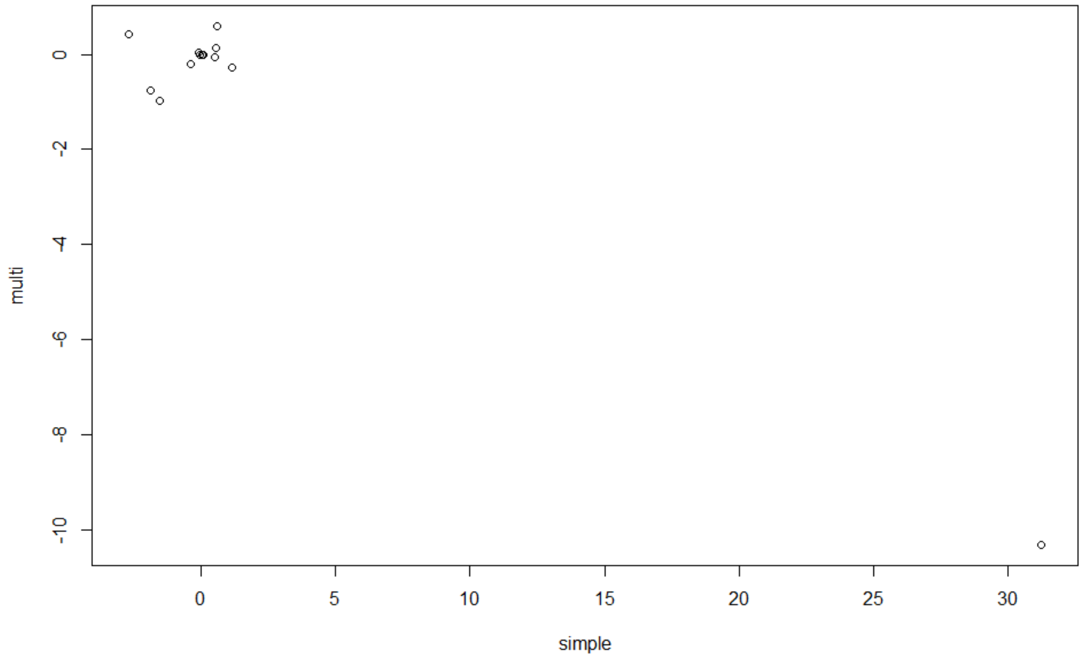
Coefficient results summarized as follows:

Variables	Simple Coefficients	Multiple Coefficients
Age	0.1077862	0.001451643
Black	-0.03627964	-0.007537505
Chas	-1.892777	-0.749133611
Dis	-1.550902	-0.987175726
Indus	0.5097763	-0.063854824
Lstat	0.5488048	0.126211376
Medv	-0.3631599	-0.198886821
Nox	31.24853	-10.313534912
Ptratio	1.151983	-0.271080558
Rad	0.6179109	0.588208591
Rm	-2.684051	0.430130506
Tax	0.02974225	-0.003780016
Zn	-0.07393498	0.044855215

Next we will plot both sets of coefficients plotting the simple linear coefficients on the X-axis and the multiple results on the Y-axis as follows in R:

```
simple = c(coef(lm.fit.zn)[2],
        coef(lm.fit.indus)[2],
        coef(lm.fit.chas)[2],
        coef(lm.fit.nox)[2],
        coef(lm.fit.rm)[2],
        coef(lm.fit.age)[2],
        coef(lm.fit.dis)[2],
        coef(lm.fit.rad)[2],
        coef(lm.fit.tax)[2],
        coef(lm.fit.ptratio)[2],
        coef(lm.fit.black)[2],
        coef(lm.fit.lstat)[2],
```

```
        coef(lm.fit.medv)[2])
multi = coef(lm.fit.mul)[2:14]
plot(simple, multi)
```



Comments:

There are some changes regarding the coefficients when comparing the single linear model to the multiple model as follows:

Some coefficients change from having a positive effect to having a negative one and vice versa:

Variables	Simple Coefficients	Multiple Coefficients
Indus	0.5097763	-0.063854824
Ptratio	1.151983	-0.271080558
Rm	-2.684051	0.430130506
Zn	-0.07393498	0.044855215

In the main though as we can see from the plot above there are only slight changes in the coefficient results between the models.

However, nox is the standout coefficient in each model:

Nox	31.24853	-10.313534912

So, this suggests that for a one unit increase in nox we should get a 31 unit increase in crim using the single regression model. However, using the multiple regression model, we expect a 10 unit decrease in crim; conflicting results when both models are compared.

CHECK FOR NON-LINEAR ASSOCIATIONS BETWEEN PREDICTORS AND THE RESPONSE VARIABLE.

To help we will fit a model in the form below for each predictor X:

$$Y=\beta 0+\beta 1\ X+\ \beta 2X^2+\beta 3X^3+\ \varepsilon$$

We will use the poly function to fit this model form for each variable except for the variable chas. Chas's outputs are zero or one so a polynomial model would not fit this variable against crim.

```
lm.fit.zn2=lm(crim~poly(zn,3))
summary(lm.fit.zn2)

Call:
lm(formula = crim ~ poly(zn, 3))

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)    3.6135     0.3722   9.709  < 2e-16 ***
poly(zn, 3)1 -38.7498     8.3722  -4.628  4.7e-06 ***
poly(zn, 3)2  23.9398     8.3722   2.859  0.00442 ** 
poly(zn, 3)3 -10.0719     8.3722  -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

lm.fit.indus2=lm(crim~poly(indus,3))
summary(lm.fit.indus2)

Call:
lm(formula = crim ~ poly(indus, 3))

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)        3.614      0.330  10.950  < 2e-16 ***
poly(indus, 3)1   78.591      7.423  10.587  < 2e-16 ***
poly(indus, 3)2  -24.395      7.423  -3.286  0.00109 ** 
poly(indus, 3)3  -54.130      7.423  -7.292  1.2e-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
```

```
lm.fit.nox2=lm(crim~poly(nox,3))
summary(lm.fit.nox2)

Call:
lm(formula = crim ~ poly(nox, 3))

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)     3.6135     0.3216  11.237  < 2e-16 ***
poly(nox, 3)1  81.3720     7.2336  11.249  < 2e-16 ***
poly(nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
poly(nox, 3)3 -60.3619     7.2336  -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

lm.fit.rm2=lm(crim~poly(rm,3))
summary(lm.fit.rm2)

Call:
lm(formula = crim ~ poly(rm, 3))

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)    3.6135     0.3703   9.758  < 2e-16 ***
poly(rm, 3)1 -42.3794     8.3297  -5.088 5.13e-07 ***
poly(rm, 3)2  26.5768     8.3297   3.191  0.00151 ** 
poly(rm, 3)3  -5.5103     8.3297  -0.662  0.50858    
---
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

lm.fit.age2=lm(crim~poly(age,3))
summary(lm.fit.age2)

Call:
lm(formula = crim ~ poly(age, 3))

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)     3.6135     0.3485  10.368  < 2e-16 ***
poly(age, 3)1  68.1820     7.8397   8.697  < 2e-16 ***
poly(age, 3)2  37.4845     7.8397   4.781 2.29e-06 ***
```

```
poly(age, 3)3  21.3532     7.8397   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

lm.fit.dis2=lm(crim~poly(dis,3))
summary(lm.fit.dis2)

Call:
lm(formula = crim ~ poly(dis, 3))

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)     3.6135     0.3259  11.087  < 2e-16 ***
poly(dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
poly(dis, 3)3 -42.6219     7.3315  -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

lm.fit.rad2=lm(crim~poly(rad,3))
summary(lm.fit.rad2)

Call:
lm(formula = crim ~ poly(rad, 3))

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)     3.6135     0.2971  12.164  < 2e-16 ***
poly(rad, 3)1 120.9074     6.6824  18.093  < 2e-16 ***
poly(rad, 3)2  17.4923     6.6824   2.618  0.00912 **
poly(rad, 3)3   4.6985     6.6824   0.703  0.48231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

lm.fit.tax2=lm(crim~poly(tax,3))
summary(lm.fit.tax2)

Call:
lm(formula = crim ~ poly(tax, 3))

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)     3.6135     0.3047  11.860  < 2e-16 ***
poly(tax, 3)1 112.6458     6.8537  16.436  < 2e-16 ***
poly(tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
poly(tax, 3)3  -7.9968     6.8537  -1.167    0.244    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

```
lm.fit.ptratio2=lm(crim~poly(ptratio,3))
summary(lm.fit.ptratio2)
```

```
Call:
lm(formula = crim ~ poly(ptratio, 3))

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)          3.614      0.361  10.008  < 2e-16 ***
poly(ptratio, 3)1   56.045      8.122   6.901 1.57e-11 ***
poly(ptratio, 3)2   24.775      8.122   3.050  0.00241 ** 
poly(ptratio, 3)3  -22.280      8.122  -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
```

```
lm.fit.black2=lm(crim~poly(black,3))
summary(lm.fit.black2)
```

```
Call:
lm(formula = crim ~ poly(black, 3))

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)       3.6135     0.3536  10.218   <2e-16 ***
poly(black, 3)1 -74.4312     7.9546  -9.357   <2e-16 ***
poly(black, 3)2   5.9264     7.9546   0.745    0.457    
poly(black, 3)3  -4.8346     7.9546  -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
```

```
lm.fit.lstat2=lm(crim~poly(lstat,3))
```

```
summary(lm.fit.lstat2)

Call:
lm(formula = crim ~ poly(lstat, 3))

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)       3.6135     0.3392  10.654   <2e-16 ***
poly(lstat, 3)1  88.0697     7.6294  11.543   <2e-16 ***
poly(lstat, 3)2  15.8882     7.6294   2.082   0.0378 *  
poly(lstat, 3)3 -11.5740     7.6294  -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

lm.fit.medv2=lm(crim~poly(medv,3))
summary(lm.fit.medv2)

Call:
lm(formula = crim ~ poly(medv, 3))

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)       3.614      0.292  12.374  < 2e-16 ***
poly(medv, 3)1  -75.058      6.569 -11.426  < 2e-16 ***
poly(medv, 3)2   88.086      6.569  13.409  < 2e-16 ***
poly(medv, 3)3  -48.033      6.569  -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
```

Observations summarized as follows:

zn	Up to the second order leads to improvement
indus	Up to the third order leads to improvement
nox	Up to the third order leads to improvement
Rm	Up to the second order leads to improvement
age	Up to the third order leads to improvement
dis	Up to the third order leads to improvement
rad	Up to the second order leads to improvement
tax	Up to the second order leads to improvement
ptratio	Up to the third order leads to improvement
Black	For x only

Lstat	Up to the second order leads to improvement
Medv	Up to the third order leads to improvement

For indus, nox, age, dis, ptratio, and medv there is significant evidence of non-linearity as the p-values for each term, squared and cubed are statistically significant.

We can go one step further and compare the simple linear models we had earlier on with these new quadratic models using the anova function – analysis of variance. This will help quantify whether the quadratic models provide a better fit than the simple linear models. In the ANOVA f-test the null hypothesis is that there is no significant difference between the two models tested and the alternative being there is a difference.

Summarized ANOVA results as follows:

ANOVA test models	p-values
anova(lm.fit.zn, lm.fit.zn2)	0.008512 **
anova(lm.fit.indus, lm.fit.indus2)	8.409e-14 ***
anova(lm.fit.nox, lm.fit.nox2)	< 2.2e-16 ***
anova(lm.fit.rm, lm.fit.rm2)	0.005229 **
anova(lm.fit.age, lm.fit.age2)	4.125e-07 ***
anova(lm.fit.dis, lm.fit.dis2)	< 2.2e-16 ***
anova(lm.fit.rad, lm.fit.rad2)	0.02608 *
anova(lm.fit.tax, lm.fit.tax2)	1.144e-05 ***
anova(lm.fit.ptratio, lm.fit.ptratio2)	0.0002542 ***
anova(lm.fit.black, lm.fit.black2)	0.6302
anova(lm.fit.lstat, lm.fit.lstat2)	0.03698 *
anova(lm.fit.medv, lm.fit.medv2)	< 2.2e-16 ***

Looking at the p-value results here indus, nox, age, dis, ptratio and medv have extremely low values and thus are better suited to the quadratic model. We have further quantified that there is evidence of non-linearity between these variables and the response crim.

However looking at the results of some of the other ANOVA tests we can see that to a slightly lesser extent variables such as zn, rm, rad, lstat, and tax are also showing signs that the quadratic model is a slightly better fit than the linear model. Again as stated earlier we must keep in mind that the residuals of some of the other variables still show some shape and outliers are clearly evident. At this point we are unsure how heavily or not these outliers are affecting the regression models we have applied here. Further questions arise then such as should we remove some of these outliers, should we investigate other methods of regression or are we certain that all the data offered here has been accurately captured? I guess this is the life of an analyst.