Summative CA02 -1 (.30)

Applied Statistics - Modelling

Submitted by

Shekhar Ramesh Vasudev

10383982

Contents

Introduction:

The Objective of this study is to apply regression modelling on the Boston Housing data set to predict the per capita crime rate using other variables in the data set.

The per capital crime rate is the response (y) and all the other variables are predictors (x)

The Boston data set contains Housing values in suburbs of Boston, with 506 rows and 14 columns. The Table below explains the meaning of each of the elements in the columns.

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

All the predictors will be referred to by their abbreviations.

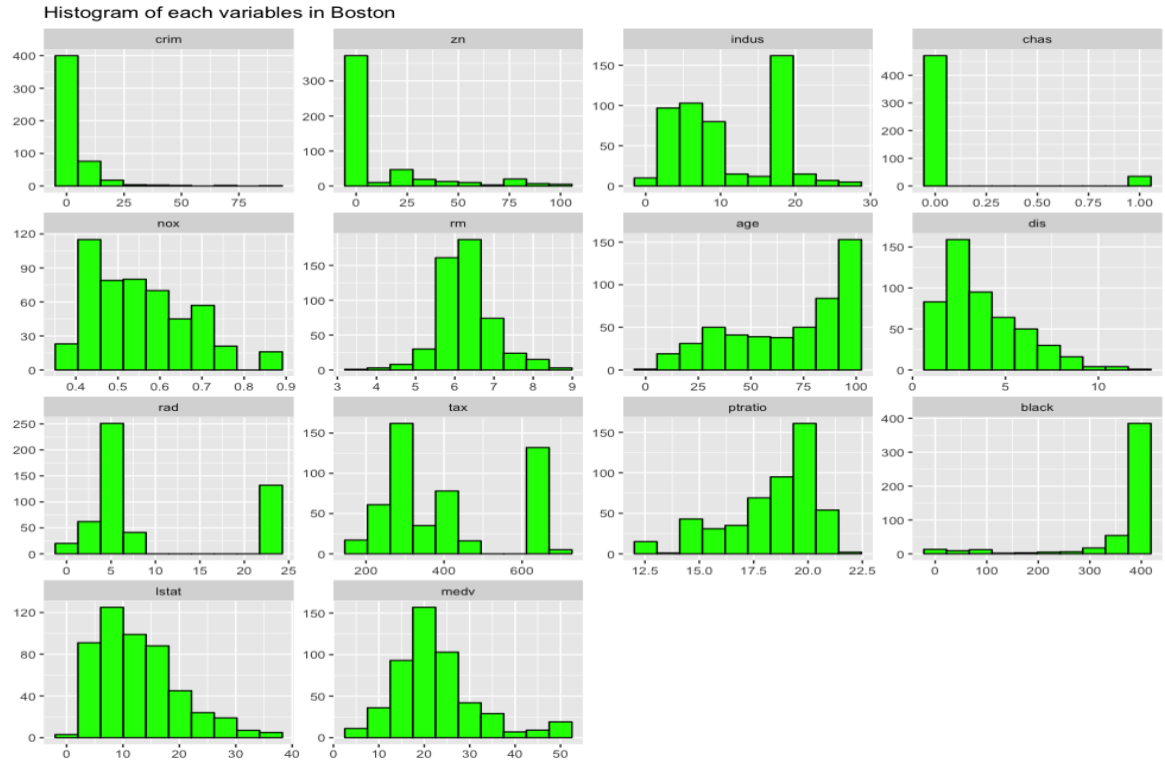
a. Descriptive Statistics

Using the summary function of the Boston data set we can get a descriptive summary (Minimum, Maximum, Median, 1st Quartile and the 3rd Quartile) of all 14 variables in the data set.

```
> summary(Boston)
      crim                zn              indus            chas              nox              rm        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561  
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780  
      age              dis              rad              tax            ptratio          black       
 Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
 1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
 Median : 77.50   Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
 Mean   : 68.57   Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
 3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
 Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
     lstat            medv      
 Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 6.95   1st Qu.:17.02  
 Median :11.36   Median :21.20  
 Mean   :12.65   Mean   :22.53  
 3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :37.97   Max.   :50.00  
```

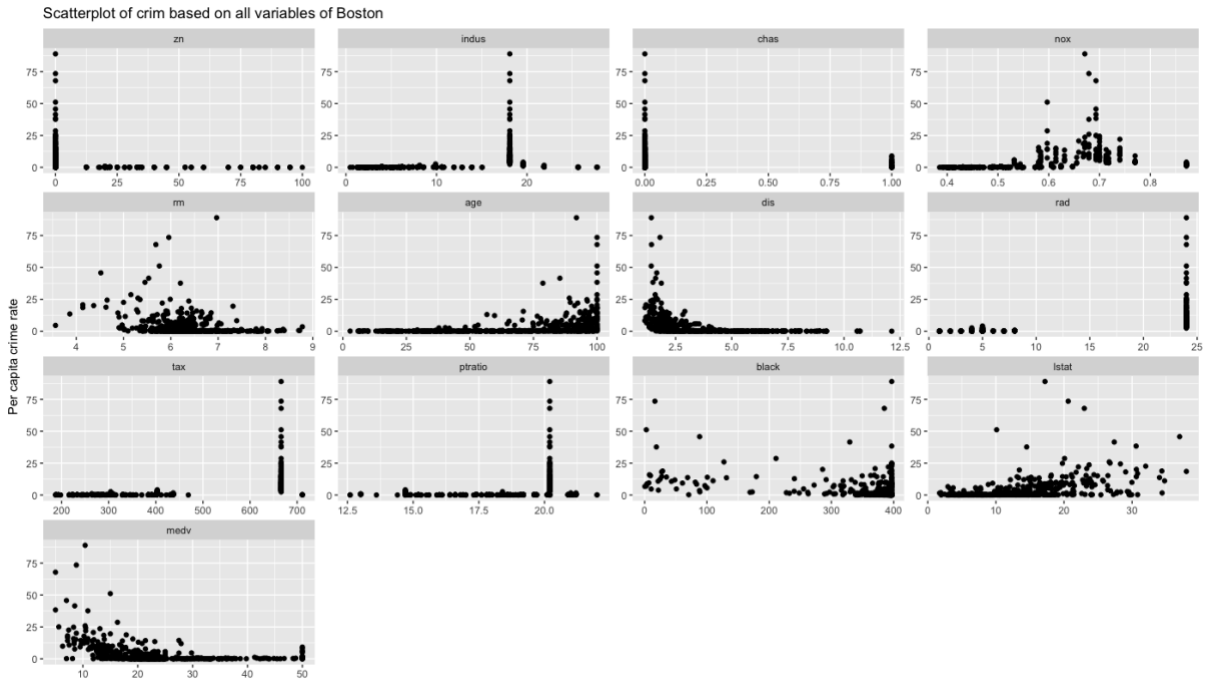
The Per capita crime rate (crim: response variable) has a minimum value of 0.00632 and a maximum of 88.97620. Among the 14 variables, chas is the only categorical variable.

Histogram of all the variables of the Boston data set



From the ggplot histograms, it's evident that none of the variables have a normal distribution and each of the variable has its own distinctive distribution. Chas being a categorical variable has 2 extreme values.

Scatterplot of crim based on all the variables of Boston Housing data set.



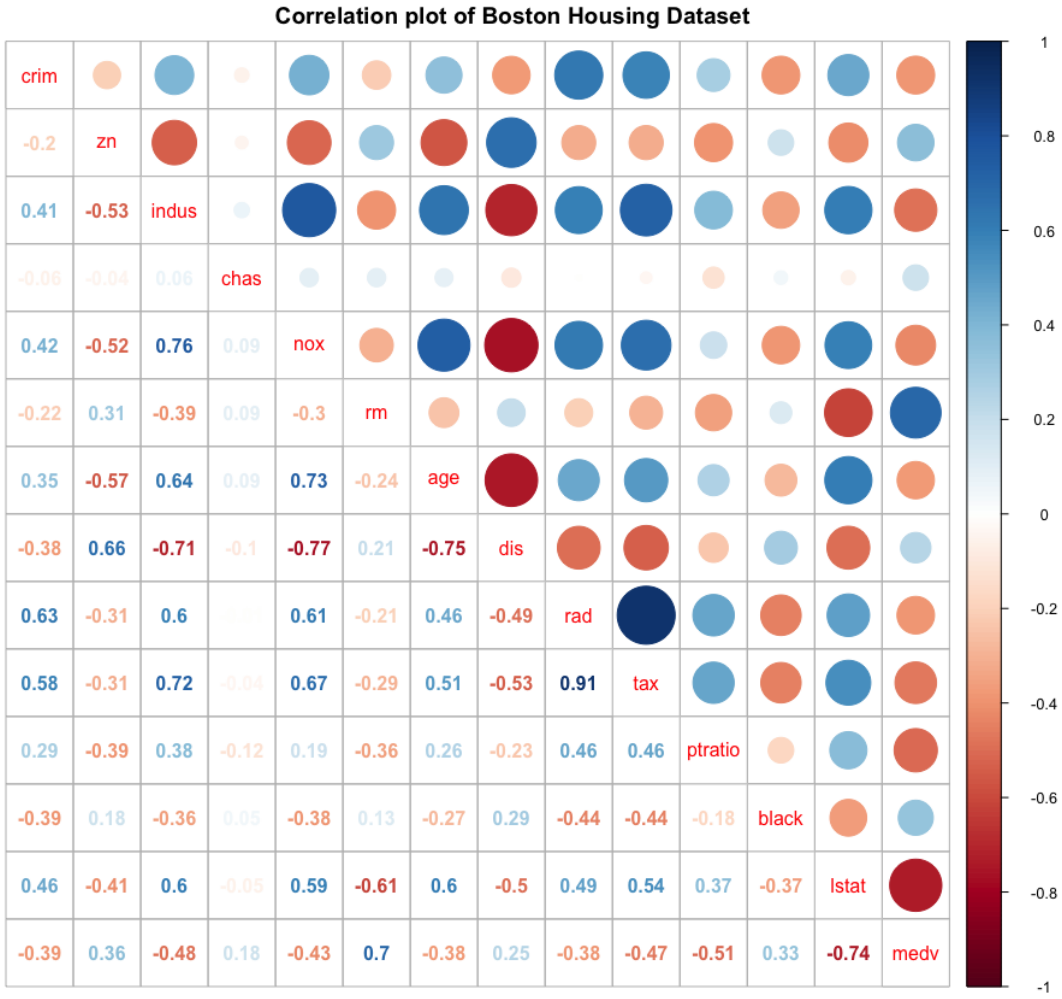
Using ggplot we can create a scatterplot of each predictor with the response variable (Y:crim) Per capita crime rate it is evident that none of the scatter plot shows a linear relationship between the predictors and the response variable.

Calculate the Pearson coefficient of correlation of the predictor (y: crim) with each response (x)

cor (y,x)	r value	Relation
Crim, zn	- 0.20046922	Weak negative
Crim, indus	0.40658341	weak positive
Crim, chas	- 0.05589158	Very weak negative / no relation
Crim, nox	0.42097171	Weak positive
Crim, rm	- 0.21924670	Weak negative
Crim, age	0.35273425	Weak positive
Crim, dis	- 0.37967009	Weak negative
Crim, rad	0.62550515	Strong positive
Crim, tax	0.58276431	Strong positive
Crim, ptratio	0.28994558	Weak positive
Crim, black	- 0.38506394	Weak negative
Crim, lstat	0.45562148	Moderately strong
Crim, medv	- 0.38830461	Weak negative

Rad returns a r value of **.62550515** and **tax** has a r value of **.58276431** which states that rad and tax have a strong positive correlation with crim when compared to the other predictors in the Boston Data set.

Visualising the correlation coefficient's using corrplot function in R.



From the correlation plot it is evident that the strongest correlation is between **rad** and c**rim** followed by **tax, lstat and nox** when compared to the other predictors.

b. Simple Linear Regression

Using R, we can fit a simple linear regression using **crim** as the response (y) variable and each of the other elements of the Boston as the predictor (x) to find on which of these models is there a statistically significant association between the predictor and the response.

Response (y): crim
Predictor (x): zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv

Using the model

$Y = \beta_0 + \beta_1 X + \varepsilon$

1. Fitting a simple linear regression using crim as the response and zn as the predictor.

This simple linear regression model will predict the per capita crime rate based on the proportion of residential land zoned for lots over 25000 sq.ft.

Response (y): crim
Predictor (x): zn

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-4.429	-4.222	-2.620	1.250	84.523

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	4.45369	0.41722	10.675	< 2e-16	***
zn (β_{1x1})	-0.07393	0.01609	-4.594	5.51e-06	***

Residual standard error: 8.435 on 504 degrees of freedom

Multiple R-squared: 0.04019 | **Adjusted R-squared:** 0.03828

F Statistics: 21.1 on 1 and 504 DF | **P value:** 5.506e-06

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = 4.4539 – 0.07393X + 8.435**

From the R output, it is shown that the Residuals vary from -4.429 to 84.523 with a median of -2.620.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **4.45369** which means that the per capital crime will be at **4.45369** when the zn (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.41722** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **10.675** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **-0.07393** which states that the Per capita crime rate will decrease by **– 0.07393** for every one unit increase in zn. The Std. Error is the variability

of the slope estimate value i.e. it can vary $\pm$**0.01609** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **- 4.594** which returns a p-value close to **0**.

When we get a negative slope, we know that the response variable and the predictor have a negative relation.

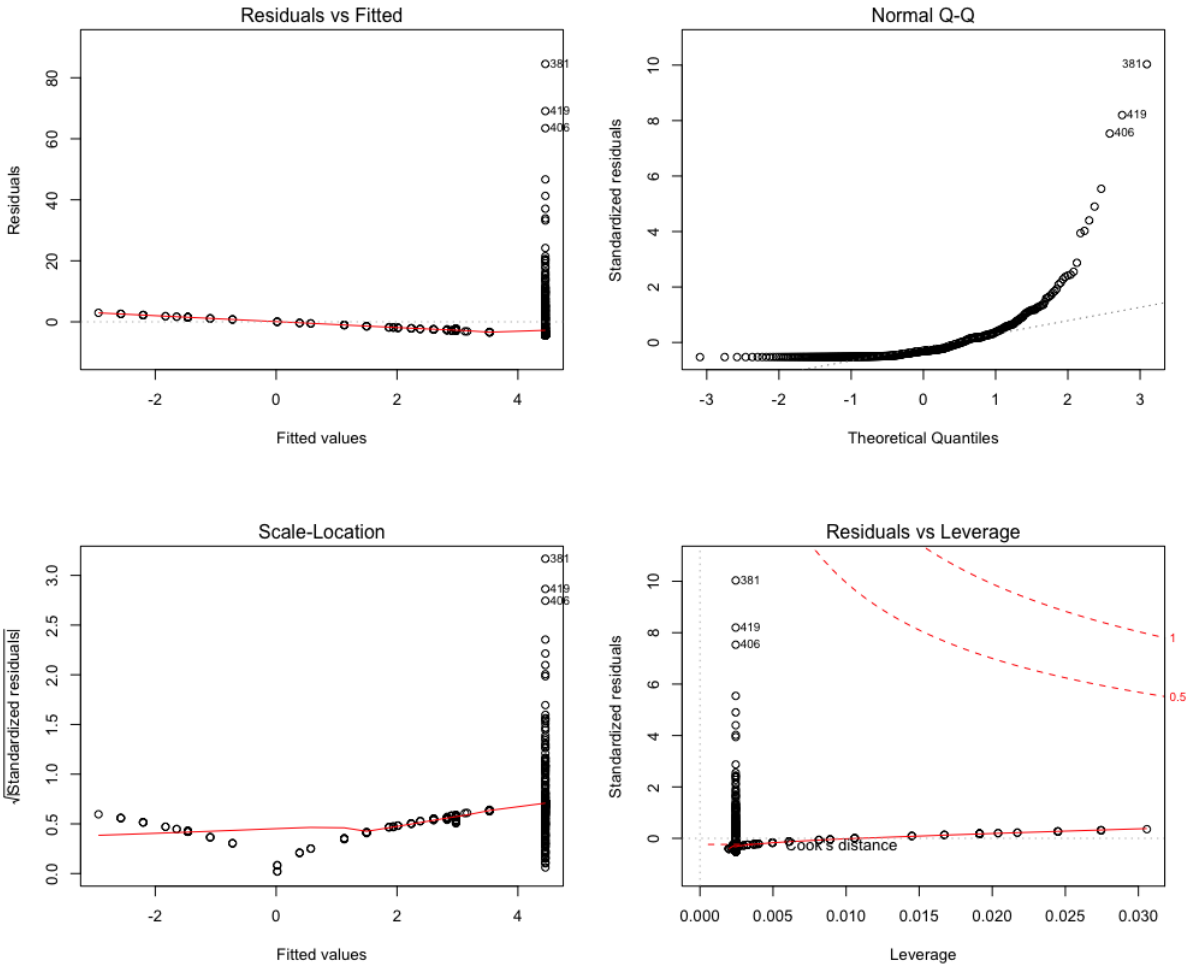
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **zn** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm$**8.435**. It shows how well the regression line fits the data.

The Multiple R-squared of **0.04019** states that only **4.019%** of the variation in Per capita crime rate is explained by the variation in zn.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model (3.828%).

The F-statistics returns a value of 21.1 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point are left out on the right side of the plot.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on zn.

2. <u>Fitting a simple linear regression using crim as the response and indus as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): indus (proportion of non-retail business acres per town.)

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-11.972	-2.698	-0.736	0.712	81.813

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	-2.06374	0.66723	-3.093	0.00209	**
indus (β_{1x1})	0.50978	0.05102	9.991	< 2e-16	***

Residual standard error: 7.866 on 504 degrees of freedom

Multiple R-squared: 0.1653 **Adjusted R-squared:** 0.1637

F Statistics: 99.82 on 1 and 504 DF **P value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = -2.06374 + 0.50978X + 7.866**

From the R output, it is shown that the Residuals vary from -11.972 to 81.813 with a median of -0.736.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **-2.06374** which means that the per capital crime will be at **-2.06374** when the **indus** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.66723** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-3.093** which returns a p-value of **0.00209**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **.50978** which states that the Per capita crime rate will increase by **.50978** for every one unit increase in **indus**. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.05102** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **9.991** which returns a p-value close to **0**.

When we get a positive slope, we know that the response variable and the predictor have a positive relation.

Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor

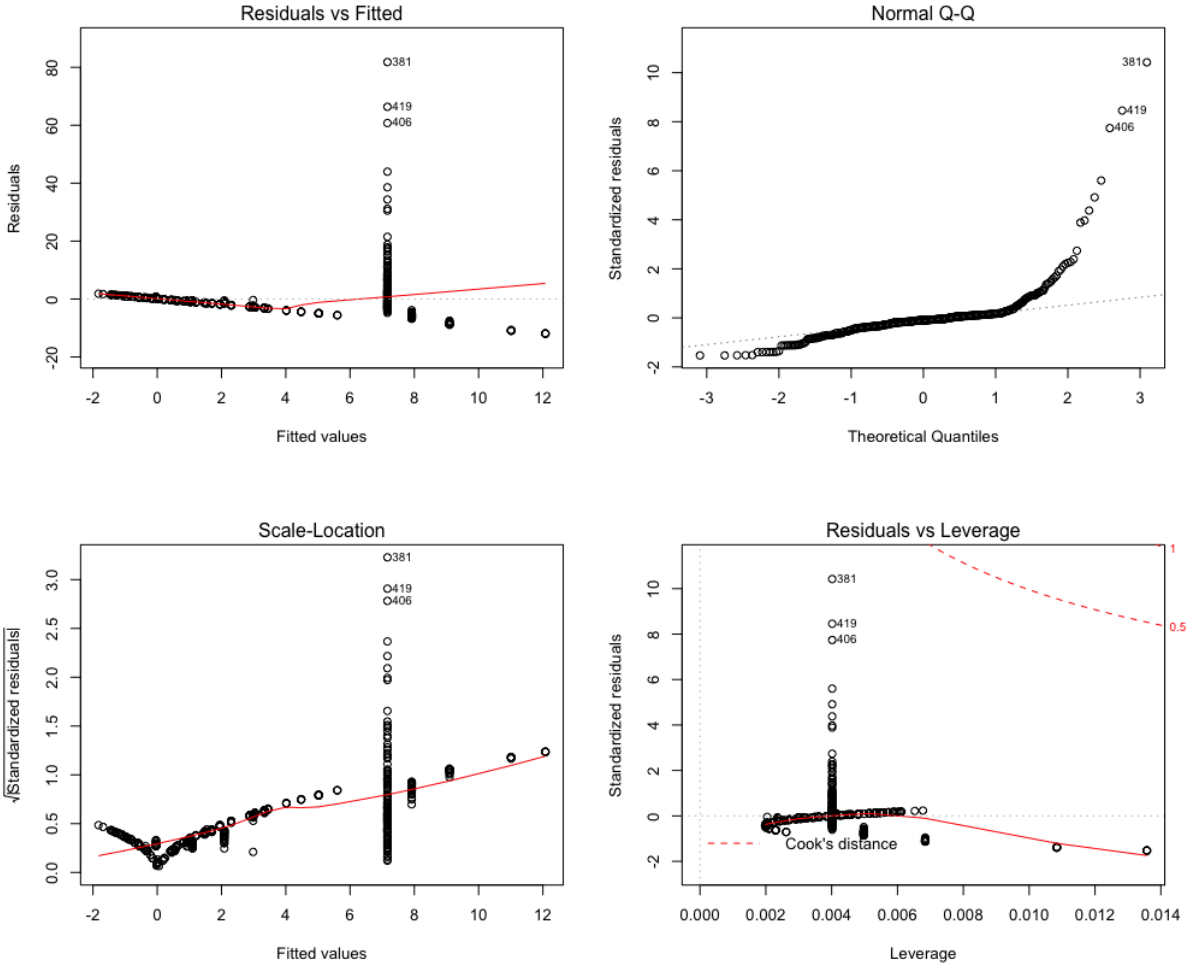
indus is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm \mathbf{7.866}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.1653** states that only **16.53%** of the variation in Per capita crime rate is explained by the variation in indus.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model. (16.37%)

The F-statistics returns a value of 99.82 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point are left out on the right side of the plot.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on indus.

3. <u>Fitting a simple linear regression using crim as the response and chas as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): chas

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-3.738	-3.661	- 3.435	0.018	85.232

Coefficients:					
	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	3.7444	0.3961	9.453	<2e-16	***
chas (β_{1x1})	-1.8928	1.5061	-1.257	0.209	

Residual standard error: 8.597 on 504 degrees of freedom						
Multiple R-squared: 0.003124			**Adjusted R-squared:** 0.001146			
F Statistics: 1.579 on 1 and 504 DF				**P value:** 0.2094		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = 3.7444 – 1.8928X + 8.597**

From the R output, it is shown that the Residuals vary from **-3.661** to **85.232** with a median of - **3.435**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **3.7444** which means that the per capital crime will be at **3.7444** when the chas (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.3961** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **9.453** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **-1.8928** which states that the Per capita crime rate will decrease by **-1.8928** for every one unit increase in chas. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**1.5061** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-1.257** which returns a p-value close of **0.209**.

When we get a negative slope, we know that the response variable and the predictor have a negative relation.

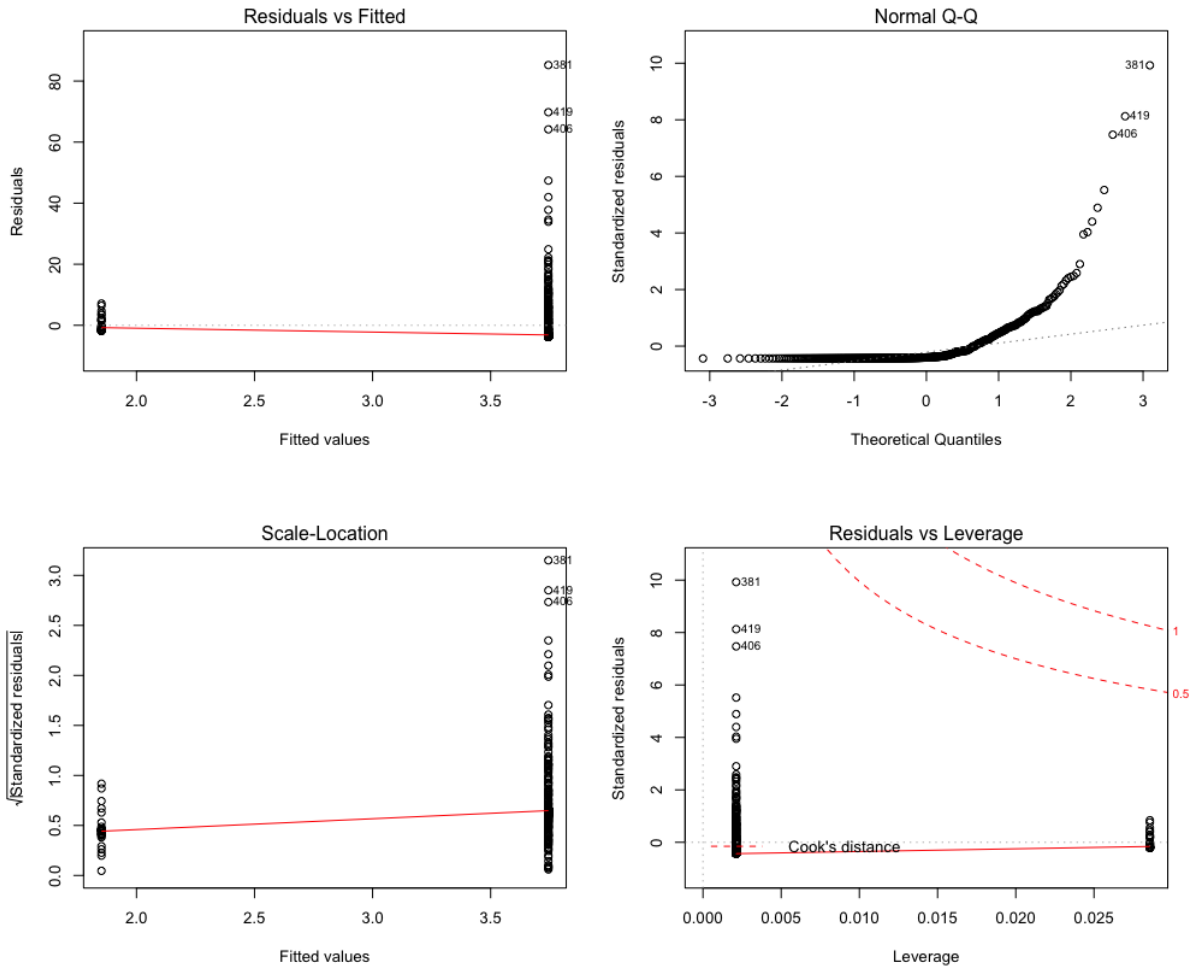
Using these p-values we can accept the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **chas** is not statistically significant (there is no relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{8.597}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.003124**states that only **0.3124%** of the variation in Per capita crime rate is explained by the variation in chas.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model (0.1146%).

Since there is no relationship the Response variable and predictor the F-statistics returns a value of 1.579 on 1 and 504 degrees of freedom and a p-value close of 0.2094, at a .05 level of significance we can state that the overall model is not significant.



- From the plots, it is evident that residuals and the fitted points fall on away from the regression line.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points isn't constant.

From our findings, the F-statistics states that the overall model isn't significant we can conclude that it can't be used for predicting the Per capita crime rate based on indus.

4. <u>Fitting a simple linear regression using crim as the response and nox as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): nox

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-12.371	-2.738	-0.974	0.559	81.728

Coefficients:					
	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	-13.720	1.699	-8.073	5.08e-15	***
nox (β_{1x1})	31.249	2.999	10.419	< 2e-16	***

Residual standard error: 7.81 on 504 degrees of freedom						
Multiple R-squared: 0.1772			**Adjusted R-squared:** 0.1756			
F Statistics: 108.6 on 1 and 504 DF				**P value:** < 2.2e-16		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = -13.720 + 31.249 X + 7.81**

From the R output, it is shown that the Residuals vary from **-12.371** to **81.728** with a median of **-0.974**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **-13.720** which means that the per capital crime will be at **-13.720** when the **nox** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**1.699** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-8.073** which returns a p-value of **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **31.249** which states that the Per capita crime rate will increase by **31.249** for every one unit increase in **nox**. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**1.699** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-8.073** which returns a p-value close to **0**.

When we get a positive slope, we know that the response variable and the predictor have a positive relation.

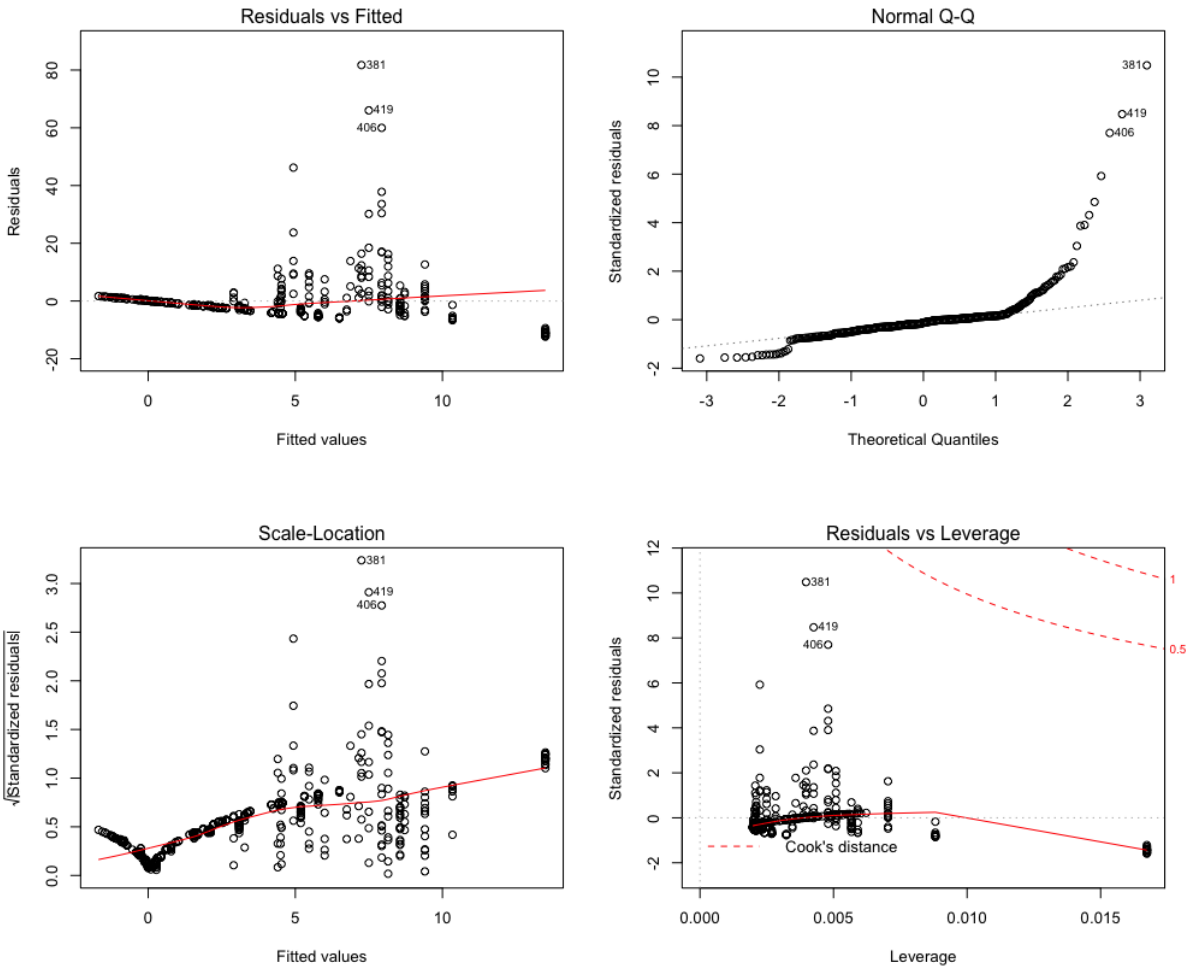
Using these p-values we can reject the null hypothesis $\boldsymbol{H_0}$: $\boldsymbol{\beta_j = 0}$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **nox** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{7.81}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.1772** states that only **17.72%** of the variation in **Per capita crime rate** is explained by the variation in **nox**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model. (17.56%)

The F-statistics returns a value of **108.6** on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point are left out of the regression line.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on nox.

5. <u>Fitting a simple linear regression using crim as the response and rm as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): rm

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-6.604	-3.952	-2.654	0.989	87.197

Coefficients:					
	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	20.482	3.365	6.088	2.27e-09	***
rm (β_{1x1})	-2.684	0.532	-5.045	6.35e-07	***

Residual standard error: 8.401 on 504 degrees of freedom						
Multiple R-squared: 0.04807			**Adjusted R-squared:** 0.04618			
F Statistics: 25.45 on 1 and 504 DF				**P value:** 6.347e-07		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = 20.482 – 2.684 X + 8.401**

From the R output, it is shown that the Residuals vary from **-6.604** to **87.197** with a median of **-2.654**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **20.482** which means that the per capital crime will be at **20.482** when the **rm** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**3.365** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **6.088** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **-2.684** which states that the Per capita crime rate will decrease by **– 2.684** for every one unit increase in rm. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.532** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-5.045** which returns a p-value close to **0**.

When we get a negative slope, we know that the response variable and the predictor have a negative relation.

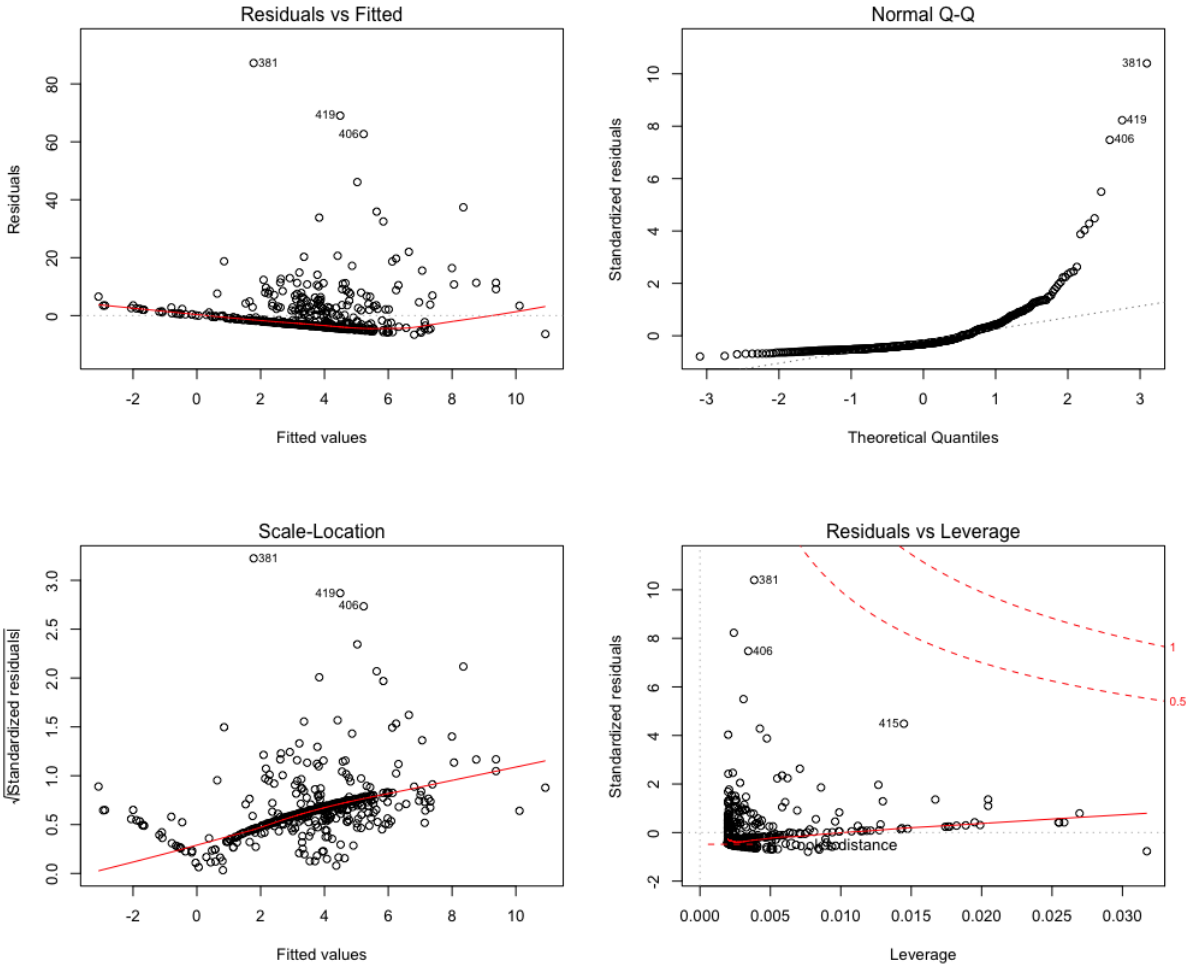
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **rm** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{8.401}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.04807** states that only **4. 807%** of the variation in Per capita crime rate is explained by the variation in **rm**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model (4.618%).

The F-statistics returns a value of 25.45 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point are left out of the regression line.

- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.

- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on rm.

6. <u>Fitting a simple linear regression using crim as the response and age as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): age

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-6.789	-4.257	-1.230	1.527	82.849

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	-3.77791	0.94398	-4.002	7.22e-05	***
age (β_{1x1})	0.10779	0.01274	8.463	2.85e-16	***

Residual standard error: 8.057 on 504 degrees of freedom

Multiple R-squared: 0.1244 **Adjusted R-squared:** 0.1227

F Statistics: 71.62 on 1 and 504 DF **P value:** 2.855e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = -3.77791 + 0.10779 X + 8.057**

From the R output, it is shown that the Residuals vary from **-6.789** to **82.849** with a median of **-1.230** .

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **-3.77791** which means that the per capital crime will be at **-3.77791** when the **age** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.94398** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-4.002** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **0.10779** which states that the Per capita crime rate will increase by **.50978** for every one unit increase in **age**. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.01274** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **8.463** which returns a p-value close to **0**.

When we get a positive slope, we know that the response variable and the predictor have a positive relation.

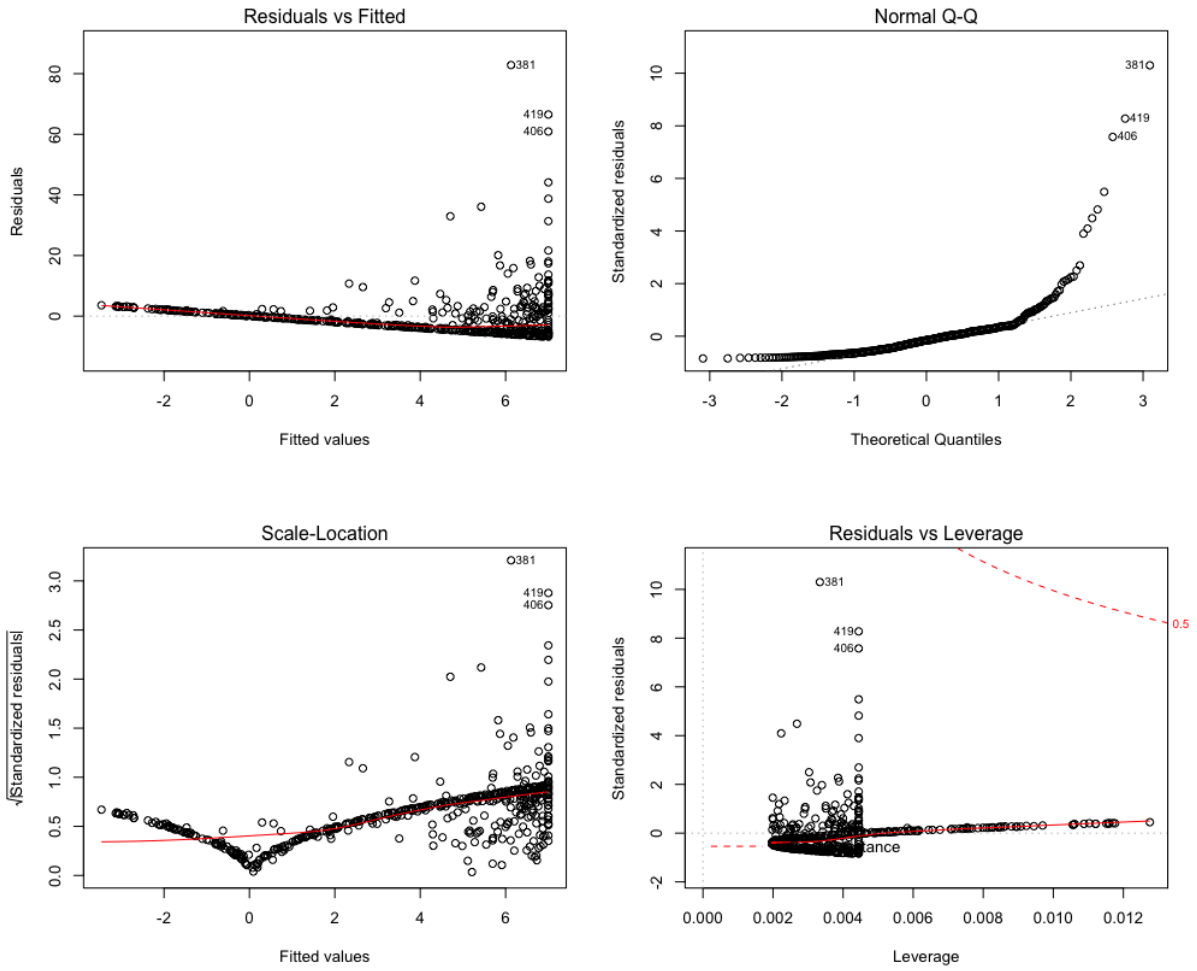
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **age** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{8.057}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.1244**states that only **12.44%** of the variation in Per capita crime rate is explained by the variation in **age**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model. (12.27%)

The F-statistics returns a value of 71.62 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the right side of the plot.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on age.

7. Fitting a simple linear regression using crim as the response and dis as the predictor.

Response (y): crim (per capita crime rate by town.)
Predictor (x): dis

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-6.708	-4.134	-1.527	1.516	81.674

Coefficients:					
	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	9.4993	0.7304	13.006	<2e-16	***
dis (β_{1x1})	-1.5509	0.1683	-9.213	<2e-16	***

Residual standard error: 7.965 on 504 degrees of freedom						
Multiple R-squared: 0.1441			**Adjusted R-squared:** 0.1425			
F Statistics: 84.89 on 1 and 504 DF				**P value:** < 2.2e-16		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = 9.4993 – 1.5509 X + 7.965**

From the R output, it is shown that the Residuals vary from **-6.708** to **81.674** with a median of **-1.527**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **9.4993** which means that the per capital crime will be at **9.4993** when the **dis** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.7304** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **13.006** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **-1.5509** which states that the Per capita crime rate will decrease by **-1.5509** for every one unit increase in dis. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.1683** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-9.213** which returns a p-value close to **0**.

When we get a negative slope, we know that the response variable and the predictor have a negative relation.

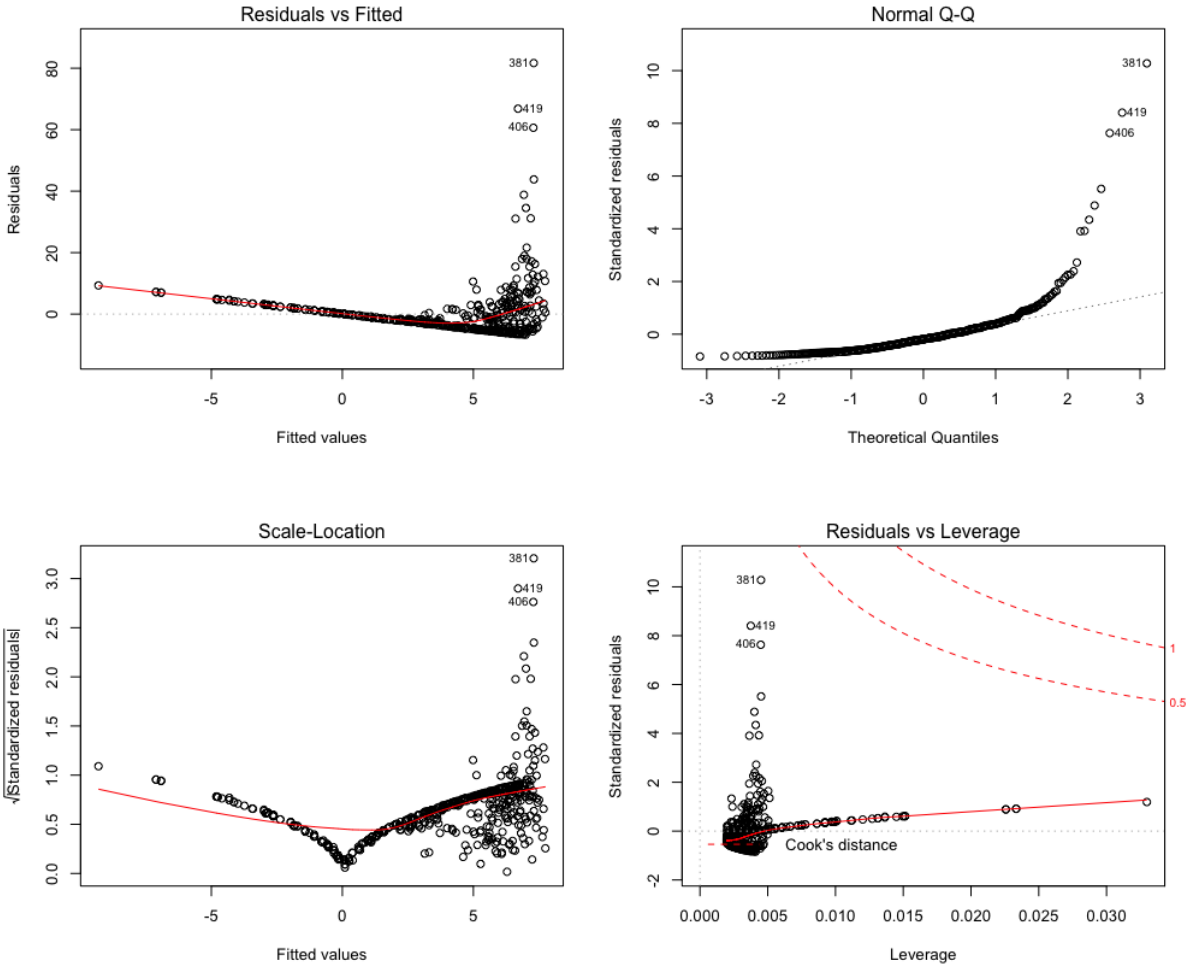
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **dis** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{7.965}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.1441**states that only 1**4. 41%** of the variation in Per capita crime rate is explained by the variation in **dis**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model (14.25%).

The F-statistics returns a value of 84.89 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the right side of the plot.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on dis.

8. <u>Fitting a simple linear regression using crim as the response and rad as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): rad

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-10.164	-1.381	-0.141	0.660	76.433

Coefficients:					
	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	-2.28716	0.44348	-5.157	3.61e-07	***
rad (β_{1x1})	0.61791	0.03433	17.998	< 2e-16	***

Residual standard error: 6.718 on 504 degrees of freedom						
Multiple R-squared: 0.3913			**Adjusted R-squared:** 0.39			
F Statistics: 323.9 on 1 and 504 DF				**P value:** < 2.2e-16		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = -2.28716 + 0.61791 X + 6.718**

From the R output, it is shown that the Residuals vary from **-10.164** to **76.433** with a median of **-0.141** .

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **-2.28716** which means that the per capital crime will be at **-2.28716** when the **rad** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.44348** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-5.157** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **0.61791** which states that the Per capita crime rate will increase by **0.61791** for every one unit increase in **rad**. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.03433** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **17.998** which returns a p-value close to **0**.

When we get a positive slope, we know that the response variable and the predictor have a positive relation.

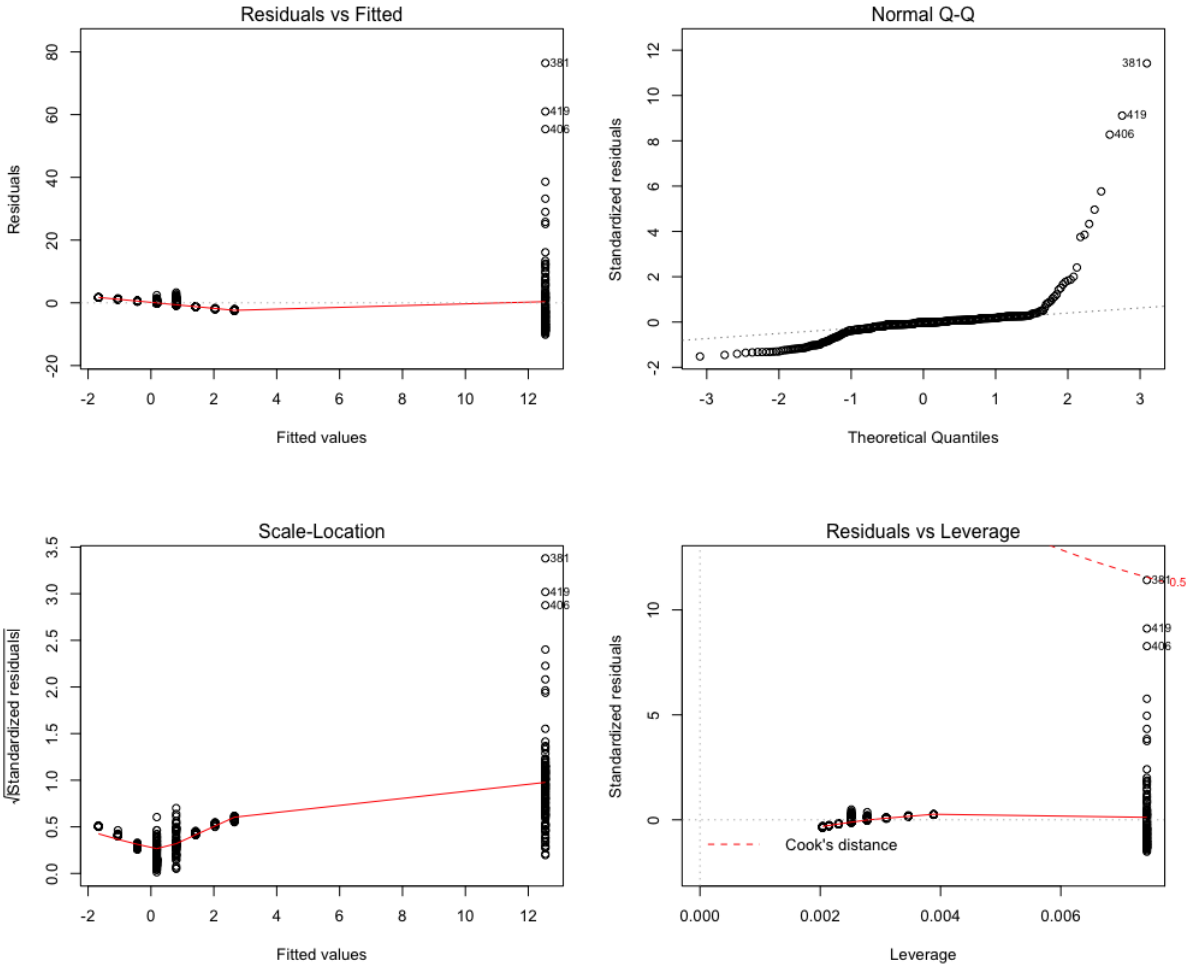
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **rad** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{6.718}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.3913** states that only **39.13%** of the variation in Per capita crime rate is explained by the variation in **rad**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model. (39%)

The F-statistics returns a value of 323.9 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the right side of the plot.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on rad.

9. <u>Fitting a simple linear regression using crim as the response and tax as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): tax

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-12.513	-2.738	-0.194	1.065	77.696

Coefficients:					
	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	-8.528369	0.815809	-10.45	<2e-16	***
tax (β_{1x1})	0.029742	0.001847	16.10	<2e-16	***

Residual standard error: 6.997 on 504 degrees of freedom						
Multiple R-squared: 0.3396			**Adjusted R-squared:** 0.3383			
F Statistics: 259.2 on 1 and 504 DF				**P value:** < 2.2e-16		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = -8.528369 + 0.029742 X + 6.997**

From the R output, it is shown that the Residuals vary from **-12.513** to **77.696** with a median of **-0.194**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **-8.528369** which means that the per capital crime will be at **-8.528369** when the **tax** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.815809** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-10.45** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **0.029742** which states that the Per capita crime rate will increase by **0.029742** for every one unit increase in **tax**. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.001847** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **16.10** which returns a p-value close to **0**.

When we get a positive slope, we know that the response variable and the predictor have a positive relation.

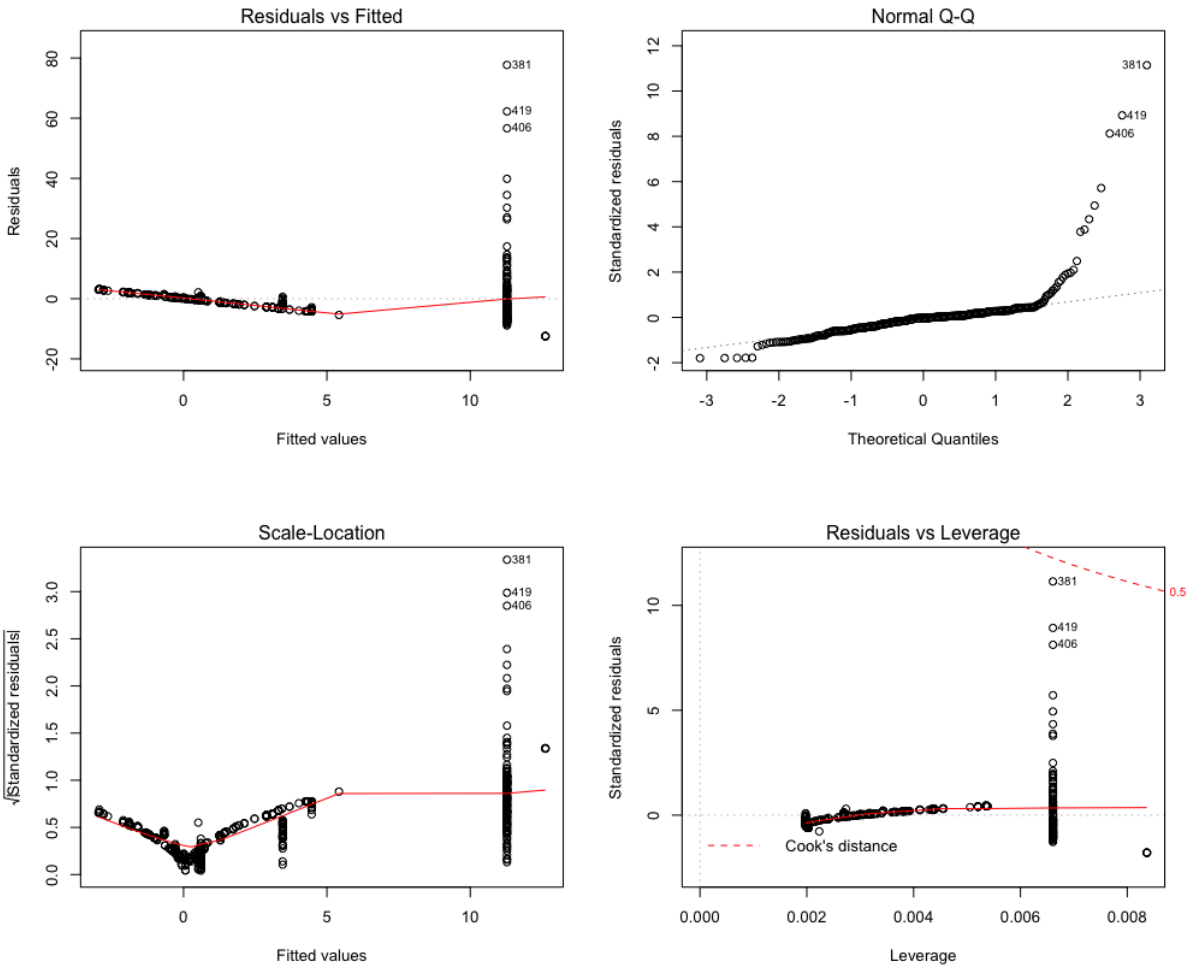
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **tax** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{6.997}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.3396** states that only **33.96%** of the variation in Per capita crime rate is explained by the variation in **tax**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model. (33.83%)

The F-statistics returns a value of 259.2 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the right side of the plot.

- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.

- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on tax.

10. <u>Fitting a simple linear regression using crim as the response and ptratio as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): ptratio

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-7.654	-3.985	-1.912	1.825	83.353

Coefficients:					
	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	-17.6469	3.1473	-5.607	3.40e-08	***
ptratio(β_{1x1})	1.1520	0.1694	6.801	2.94e-11	***

Residual standard error: 8.24 on 504 degrees of freedom						
Multiple R-squared: 0.08407			**Adjusted R-squared:** 0.08225			
F Statistics: 46.26 on 1 and 504 DF				**P value:** 2.943e-11		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = -17.6469 + 1.1520 X + 8.24**

From the R output, it is shown that the Residuals vary from **-7.654** to 83.353 with a median of **-1.912**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **-17.6469** which means that the per capital crime will be at **-17.6469** when the **ptratio** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**3.1473** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-5.607** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **1.1520** which states that the Per capita crime rate will increase by **1.1520** for every one unit increase in **ptratio**. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.1694** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **6.801** which returns a p-value close to **0**.

When we get a positive slope, we know that the response variable and the predictor have a positive relation.

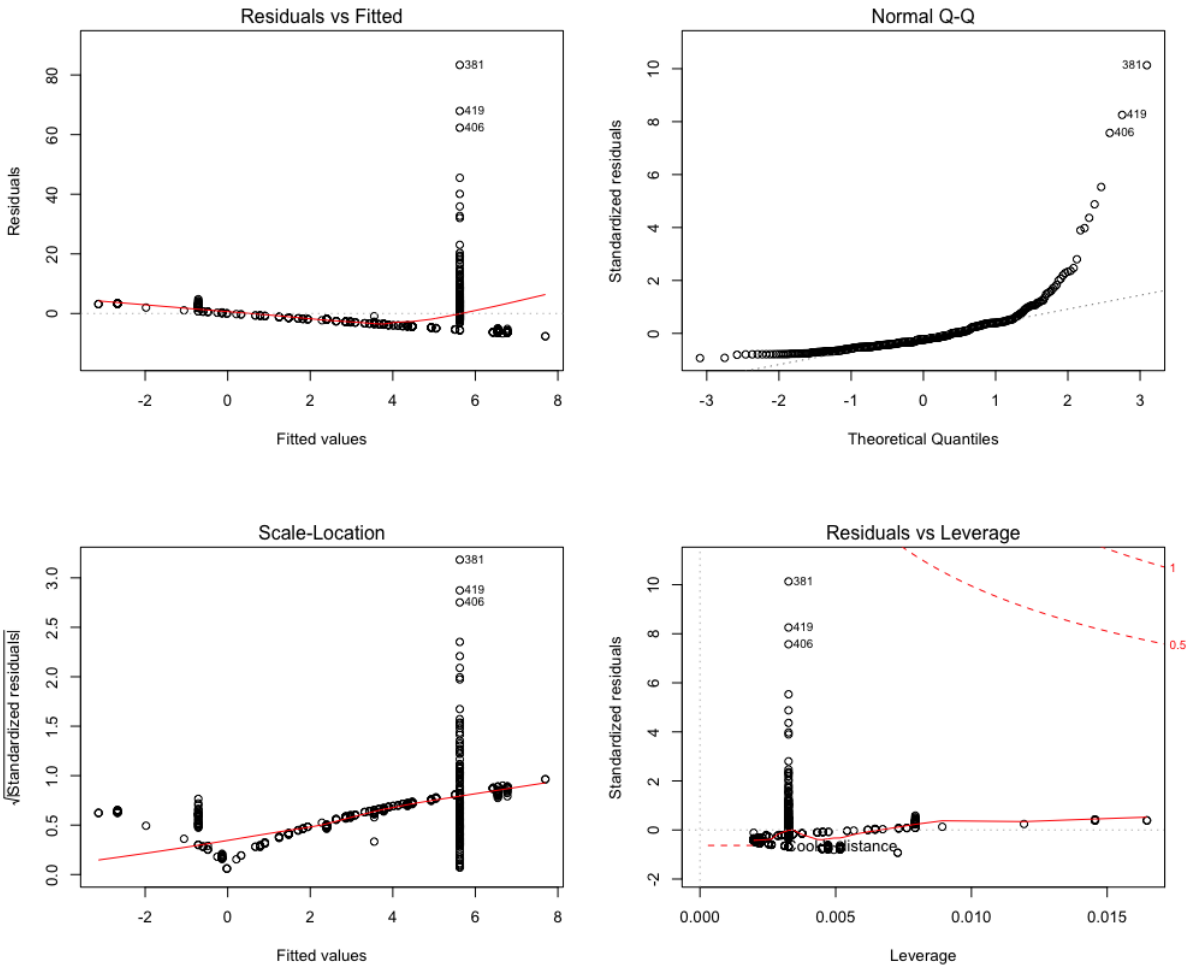
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **ptratio** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{8.24}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.08407** states that only **8.407%** of the variation in Per capita crime rate is explained by the variation in **ptratio**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model. (8.225%)

The F-statistics returns a value of 46.26 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the right side of the plot.

- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.

- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on ptratio.

11. <u>Fitting a simple linear regression using crim as the response and black as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): black

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-13.756	-2.299	-2.095	-1.296	86.822

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	16.553529	1.425903	11.609	<2e-16	***
black (β_{1x1})	-0.036280	0.003873	-9.367	<2e-16	***

Residual standard error: 7.946 on 504 degrees of freedom

Multiple R-squared: 0.1483 **Adjusted R-squared:** 0.1466

F Statistics: 87.74 on 1 and 504 DF **P value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = 16.553529 – 0.036280 X + 7.946**

From the R output, it is shown that the Residuals vary from **-13.756** to **86.822** with a median of **-2.095**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **16.553529** which means that the per capital crime will be at **16.553529** when the **black** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**1.425903** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **11.609** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **-0.036280** which states that the Per capita crime rate will decrease by **-0.036280** for every one unit increase in black. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.003873** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-9.367** which returns a p-value close to **0**.

When we get a negative slope, we know that the response variable and the predictor have a negative relation.

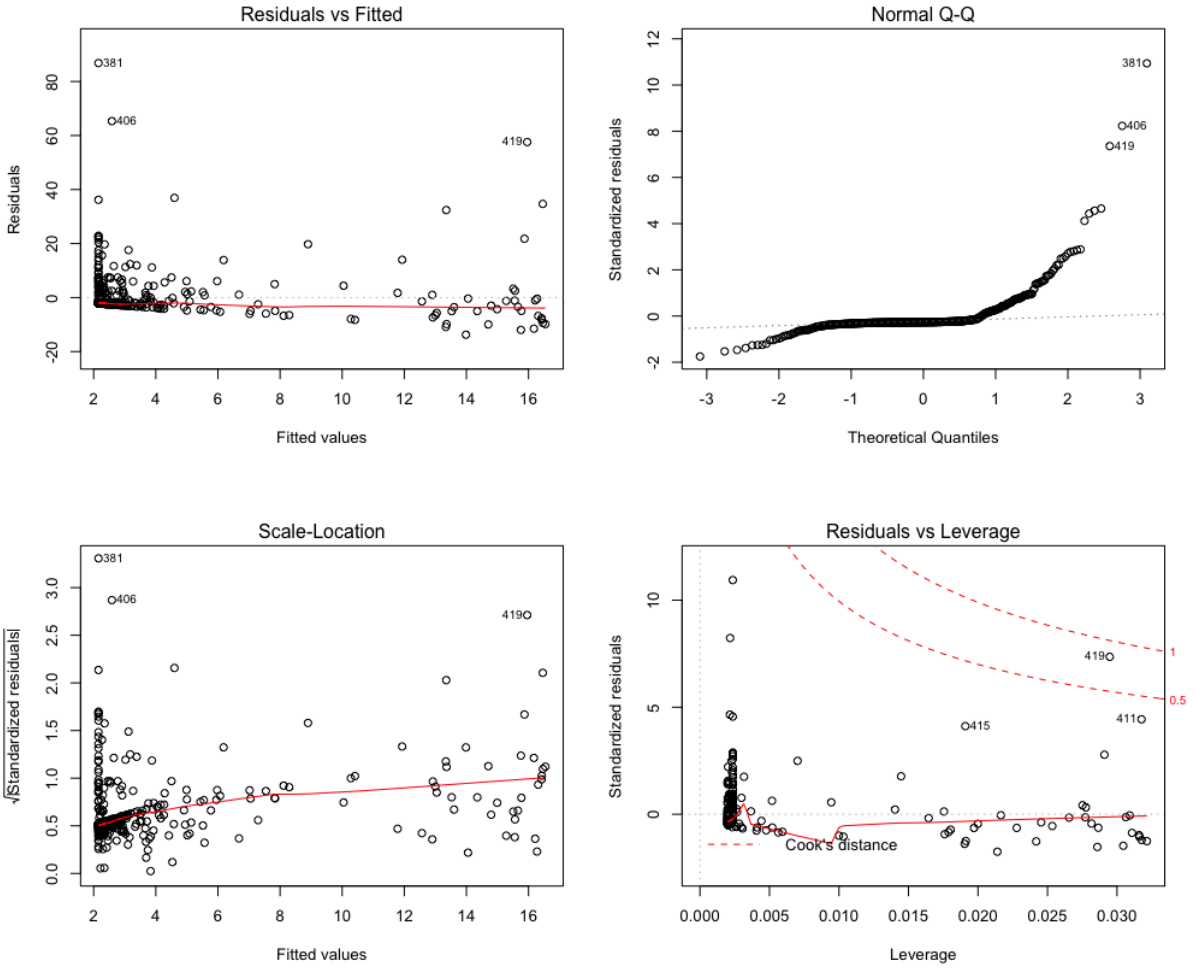
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **black** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{7.946}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.1483** states that only **14.83%** of the variation in Per capita crime rate is explained by the variation in **black**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model (14.66%).

The F-statistics returns a value of 87.74 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the left side of the plot.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points isn't constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on black.

12. <u>Fitting a simple linear regression using crim as the response and lstat as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): lstat

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-13.925	-2.822	-0.664	1.079	82.862

Coefficients:					
	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	-3.33054	0.69376	-4.801	2.09e-06	***
lstat (β_{1x1})	0.54880	0.04776	11.491	< 2e-16	***

Residual standard error: 7.664 on 504 degrees of freedom						
Multiple R-squared: 0.2076			**Adjusted R-squared:** 0.206			
F Statistics: 132 on 1 and 504 DF				**P value**: < 2.2e-16		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = -3.33054 + 0.54880 X + 7.664**

From the R output, it is shown that the Residuals vary from **-13.925** to **82.862** with a median of **-0.664**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **-3.33054** which means that the per capital crime will be at **-3.33054** when the **lstat** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.69376** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-4.801** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **0.54880** which states that the Per capita crime rate will increase by **0.54880** for every one unit increase in **lstat**. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.04776** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **11.491** which returns a p-value close to **0**.

When we get a positive slope, we know that the response variable and the predictor have a positive relation.

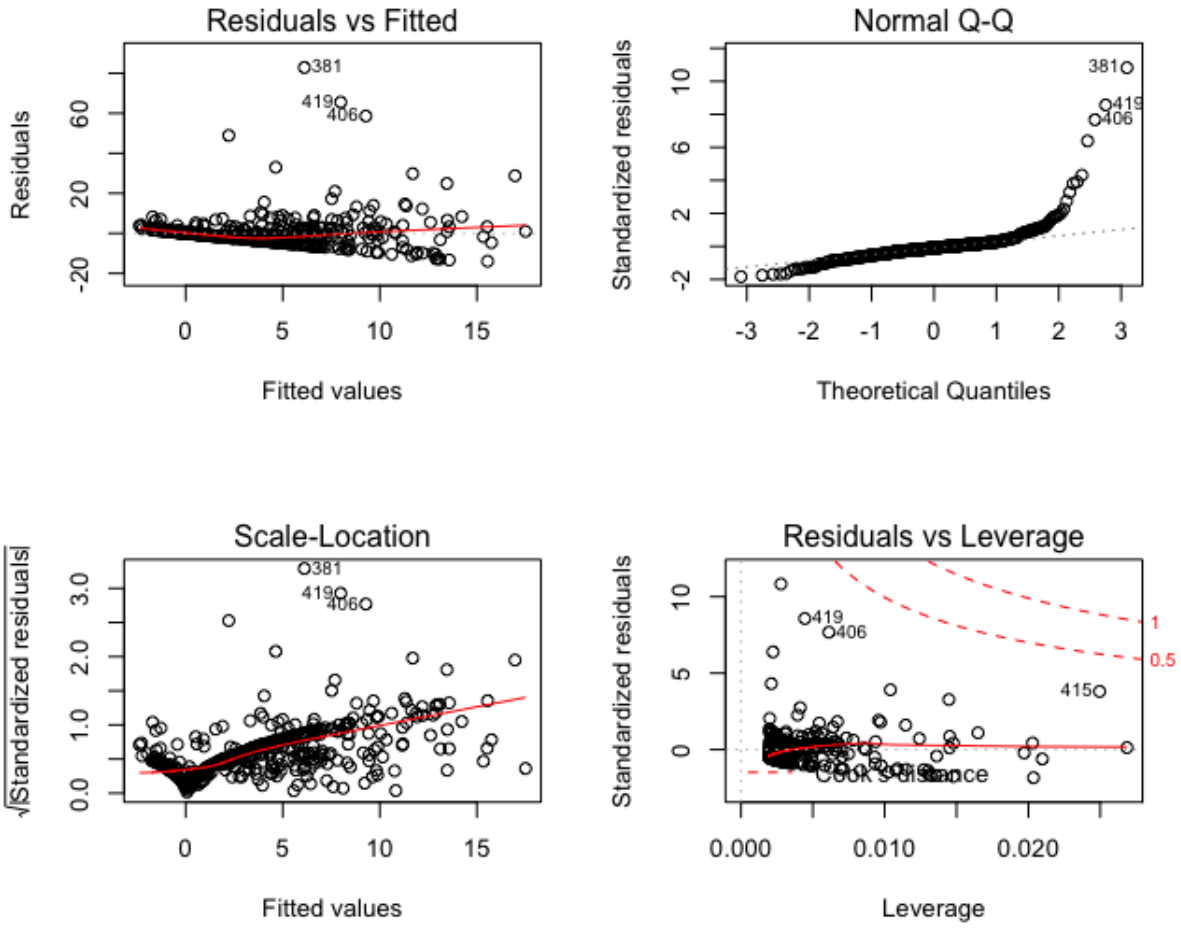
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **lstat** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{7.664}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.2076** states that only **20.76%** of the variation in Per capita crime rate is explained by the variation in **lstat**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model. (20.6%)

The F-statistics returns a value of 132 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the right side of the plot.
- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.
- The variance among the points is non-constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on lstat.

13. <u>Fitting a simple linear regression using crim as the response and medv as the predictor.</u>

Response (y): crim (per capita crime rate by town.)
Predictor (x): medv

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-9.071	-4.022	-2.343	1.298	80.957

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	**Signif. Code**
Intercept(β_0)	11.79654	0.93419	12.63	<2e-16	***
medv (β_{1x1})	-0.36316	0.03839	-9.46	<2e-16	***

Residual standard error: 7.934 on 504 degrees of freedom

Multiple R-squared: 0.1508 **Adjusted R-squared:** 0.1491

F Statistics: 89.49 on 1 and 504 DF **P value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation **Y = 11.79654 – 0.36316 X + 7.934**

From the R output, it is shown that the Residuals vary from **-9.071** to **80.957** with a median of **-2.343**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **11.79654** which means that the per capital crime will be at **11.79654** when the **medv** (predictor) is held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**0.93419** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **12.63** which returns a p-value close to **0**.

The $\boldsymbol{\beta_{1X1}}$ **(The slope)** has an estimate of **-0.36316** which states that the Per capita crime rate will decrease by **-0.36316** for every one unit increase in medv. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.03839** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **-9.46** which returns a p-value close to **0**.

When we get a negative slope, we know that the response variable and the predictor have a negative relation.

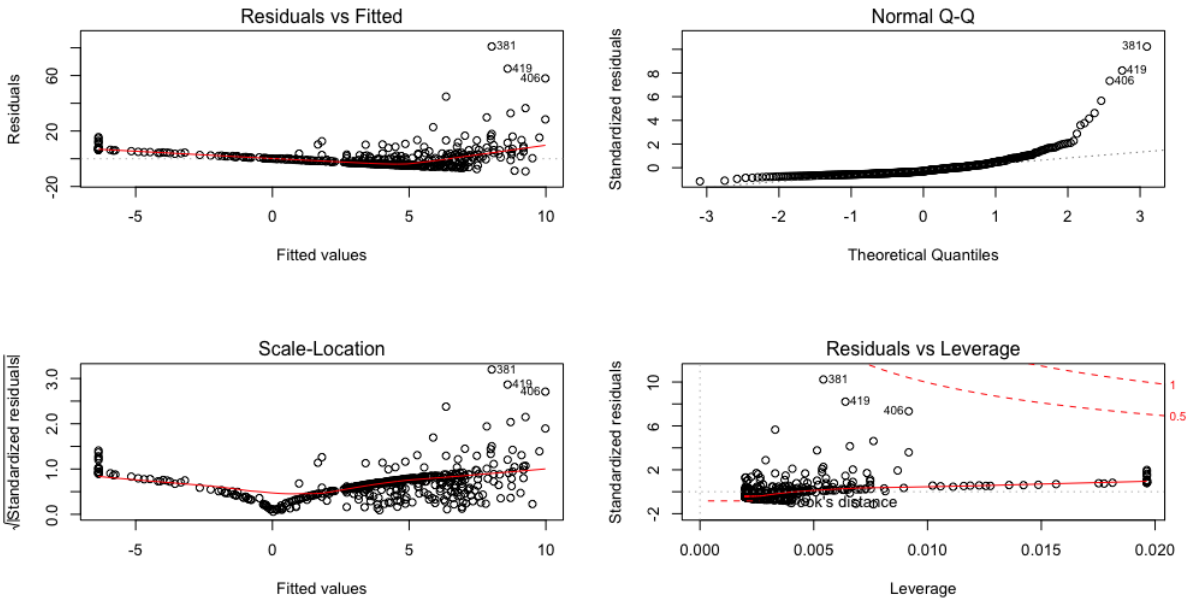
Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a .05 level of significance, which means the predictor **medv** is statistically significant (there is some relation between the response variable and the predictor).

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm\mathbf{7.936}$. It shows how well the regression line fits the data.

The Multiple R-squared of **0.1508** states that only **15.08%** of the variation in Per capita crime rate is explained by the variation in **medv**.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model (14.91%).

The F-statistics returns a value of 89.49 on 1 and 504 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the right side of the plot.

- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.

- The variance among the points is non-constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on medv.

From the Simple linear regression model, it is evident that none of the predictors can fit the regression line for prediction. All the simple linear models can explain only a part of the variation in Per capita crime rate. Which states that simple linear regression isn't a good model for prediction.

c. Multiple Regression Model

Using R, we can fit a Multiple linear regression using **crim** as the response (y) variable and all the other elements of the Boston as the predictors (x) to find on which of these models is there a statistically significant association between the predictors and the response.

Using the model

$$Y = \beta_0 + \beta_{1X_1} + \beta_{2X_2} + \beta_{3X_3} + \cdots + \beta_{nX_n} + \varepsilon$$

Response (y): crim (per capita crime rate by town.)
Predictors (x): zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv

For this model, we have 13 predictors.

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-9.924	-2.120	-0.353	1.019	75.051

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	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	**

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

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

For this model, we can fit the equation:

$$\mathbf{Y = 17.033228 + 0.044855\ X_1 - 0.063855\ X_2 - 0.749134\ X_3 - 10.313535\ X_4 + 0.430131\ X_5 + 0.001452\ X_6 - 0.987176\ X_7 + 0.588209\ X_8 - 0.003780\ X_9 - 0.271081\ X_{10} - 0.007538\ X_{11} + 0.126211\ X_{12} - 0.198887\ X_{13} + 6.439}$$

From the R output, it is shown that the Residuals vary from **-9.071** to **75.051** with a median of **-0.353**.

Under the coefficients, the **Intercept** ($\boldsymbol{\beta_0}$) returns a value **17.033228** which means that the per capital crime will be at **17.033228** when the all the predictors (zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv) are held constant at 0. The Std. Error is the variability of the Intercept estimate value i.e. it can vary $\pm$**7.234903** the value. The t-value is calculated by dividing the Estimate by the Std. Error which is **2.354** which returns a p-value close to **0.01**.

Using these p-values of the predictors we can reject the null hypothesis $H_0: \beta_j = 0$ (which states that the intercept value is equal to zero) for the predictors **rad, dis, medv, black, zn** at a confidence level of 95% (.05 level of significance), we can conclude that the 5 predictors statistically significant (there is some relation between the response variable and the predictors).

Using these p-values of the predictors we can accept the null hypothesis $H_0: \beta_j = 0$ (which states that the intercept value is equal to zero) For the predictors **indus, chas, nox, rm, age, tax, ptratio, lstat** we can conclude that these 8 predictors aren't as statistically significant as the other predictors.

Looking at the confidence intervals (95%) on the intercept and the coefficients for all the variables

	2.5%	97.5%
Intercept($\boldsymbol{\beta_0}$)	2.818109179	31.2483458660
zn	0.008046562	0.0816638671
indus	-0.227733150	0.1000235023
chas	-3.067882868	1.5696156471
nox	-20.678894713	0.0518248891
rm	-0.773956866	1.6342178774
age	-0.033767600	0.0366708869
dis	-1.540889544	-0.4334619069
rad	0.415209611	0.7612075719
tax	-0.013909700	0.0063496670
ptratio	-0.637417996	0.0952568794
black	-0.014754837	-0.0003201725
lstat	-0.022572584	0.2749953365
medv	-0.317788478	-0.0799851646

The corresponding confidence interval provides the uncertainty in the estimate. That means in repeated random sampling, the computed confidence interval straddles the true but unknown coefficient 95% of the time.

From the t-test results it is evident that the confidence intervals of predictors (**dis, rad, medv, black, zn**) doesn't contain zero. Whereas the confidence intervals of the other predictors (**indus, chas, nox, rm, age, tax, ptratio, lstat**) straddles zero.

The predictor **rad** has an ($\boldsymbol{\beta_{1X1}}$) estimate of **0.588209** which states that the Per capita crime rate will increase by **0.588209** for every one unit increase in **rad**. The Std. Error is the variability of the slope estimate value i.e. it can vary $\pm$**0.088049** the value.

The RSE(ε) shows the standard deviation of the residuals i.e. the residuals can vary from the actual value by $\pm \mathbf{6.439}$. It shows how well the regression line fits the data.

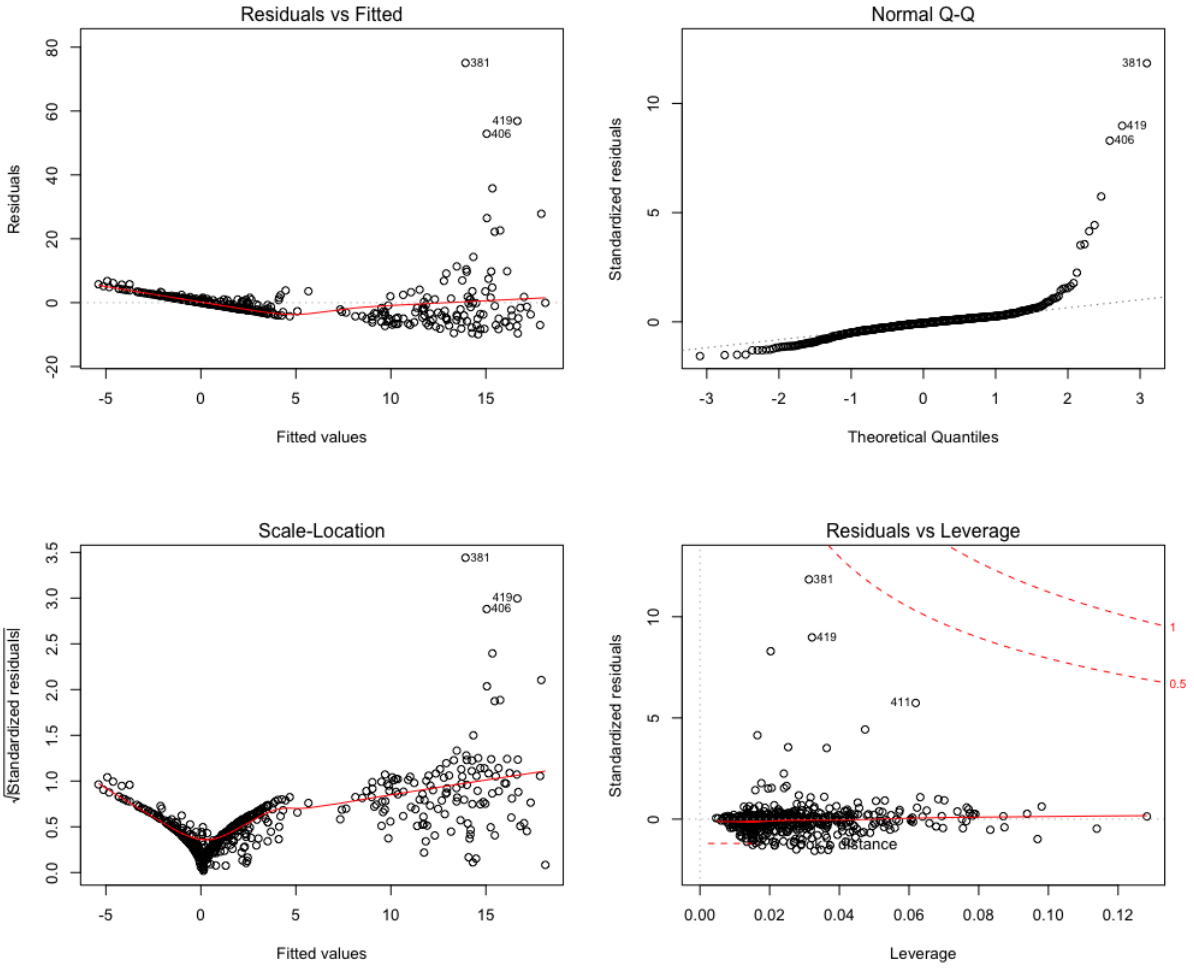
The Multiple R-squared of **0.454** states that 45.4**%** of the variation in Per capita crime rate is explained by the variation in the predictors **(zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv)**.

When comparing to all the simple linear regression models it is evident that the multiple regression model has a better Multiple R-squared, it is because we are taking into count all the 13 predictors (R-squared can be increased by adding more variables to the model).

Just adding more variables to explain a given data set but not to improve the explanatory nature of the model is known as overfitting. To address the possibility of overfitting the data, the adjusted R-squared accounts for the number of parameters included in the regression model.

The adjusted R-square is the same R-squared value adjusted for the number of variables in a model (43.96%).

The F-statistics returns a value 31.47 on 13 and 492 degrees of freedom and a p-value close to 0, at a .05 level of significance we can state that the overall model is significant.



- From the plots, it is evident that the regression line can fit only a few points and majority of the point fall on the right side of the plot.

- The QQ plot shows that the residuals don't have a normal distribution hence the mean will not be equal to 0.

- The variance among the points is non-constant.

From our findings, we can conclude that even though the F-statistics states that this is a good model, it can't be used for predicting the Per capita crime rate based on the predictors (zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv).

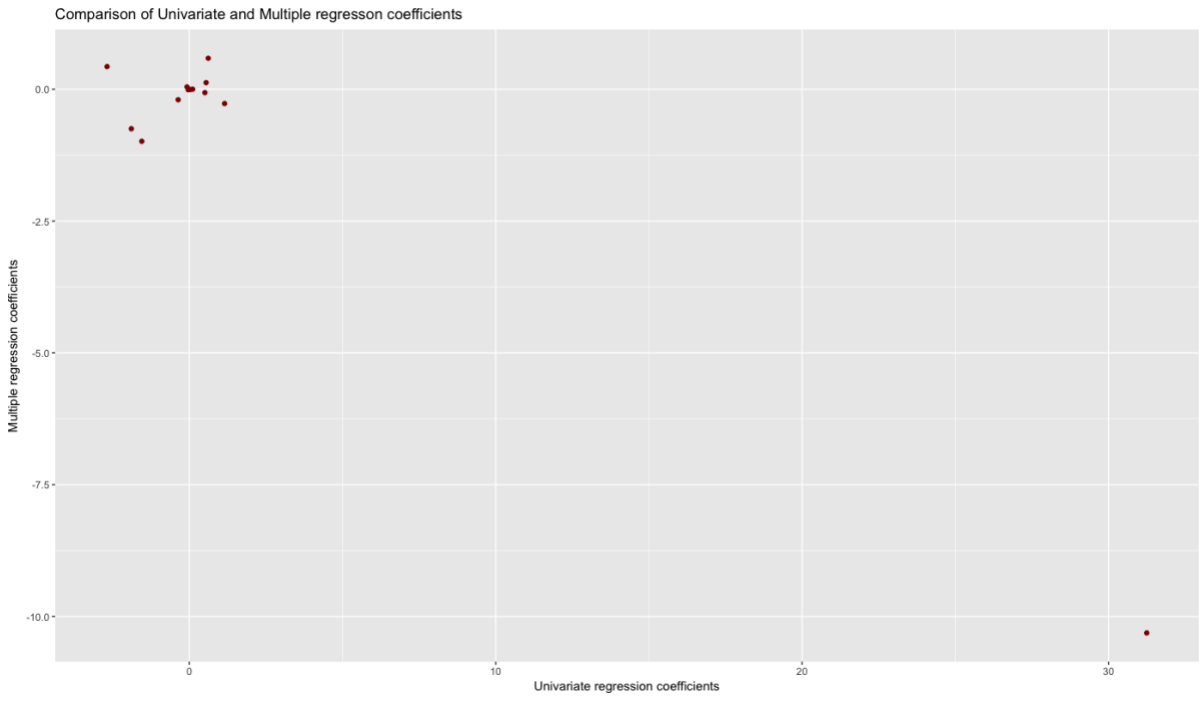
From the Multiple regression model, it is evident that using all the variables as predictors the regression line can only explain 43.96% of the variation in Per capita crime rate. Which states that multiple regression isn't a good model for prediction.

d. Comparison of Univariate and Multiple regression coefficients

Comparing the coefficients of the predictors in the simple regression model and the multiple regression model.

Coefficients of Simple linear regression		Coefficients of Multiple regression	
zn	-0.07393498	**zn**	0.044855215
indus	0.50977633	**indus**	-0.063854824
chas	-1.89277655	**chas**	-0.749133611
nox	31.24853120	**nox**	-10.313534912
rm	-2.68405122	**rm**	0.430130506
age	0.10778623	**age**	0.001451643
dis	-1.55090168	**dis**	-0.987175726
rad	0.61791093	**rad**	0.588208591
tax	0.02974225	**tax**	-0.003780016
ptratio	1.15198279	**ptratio**	-0.271080558
black	-0.03627964	**black**	-0.007537505
lstat	0.54880478	**lstat**	0.126211376
medv	-0.36315992	**medv**	-0.198886821

Using ggplot, Plotting the coefficients of the predictors in the simple linear regression on the x-axis and the coefficients of the predictors in the multiple regression on the y-axis, it will return a single point on the plot for each predictor.



From the plot, it is evident that majority of the predictors (**zn, indus, chas, rm, age, dis, rad, tax, ptratio, black, lstat, medv**) fall closer to zero forming a cluster except for **nox** which falls far away from the remaining predictors. That means for one unit change in **nox, crim** would increase by 31 units in the simple linear model and decrease by -10 units in the Multiple regression model which states the opposite. Other predictors have a better range.

e. Non-Linear Transformations of the Predictors

Using the model, we fit a cubic fit for each predictor

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

This model shows if there is any evidence of non- linear association between the predictors and the response.

H_0: There is no evidence of non linear relationship between the predictor and response
H_a: There is an evidence of non linear relationship between the predictor and response

1. Nonlinear transformation of the predictor zn

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-4.821	-4.614	-1.294	0.473	84.130

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	4.846e+00	4.330e-01	11.192	< 2e-16	***
zn	3.322e-01	1.098e-01	-3.025	0.00261	**
zn^2	6.483e-03	3.861e-03	1.679	0.09375	.
zn^3	-3.776e-05	3.139e-05	-1.203	0.22954	

Residual standard error: 8.372 on 502 degrees of freedom

Multiple R-squared: 0.05824 **Adjusted R-squared:** 0.05261

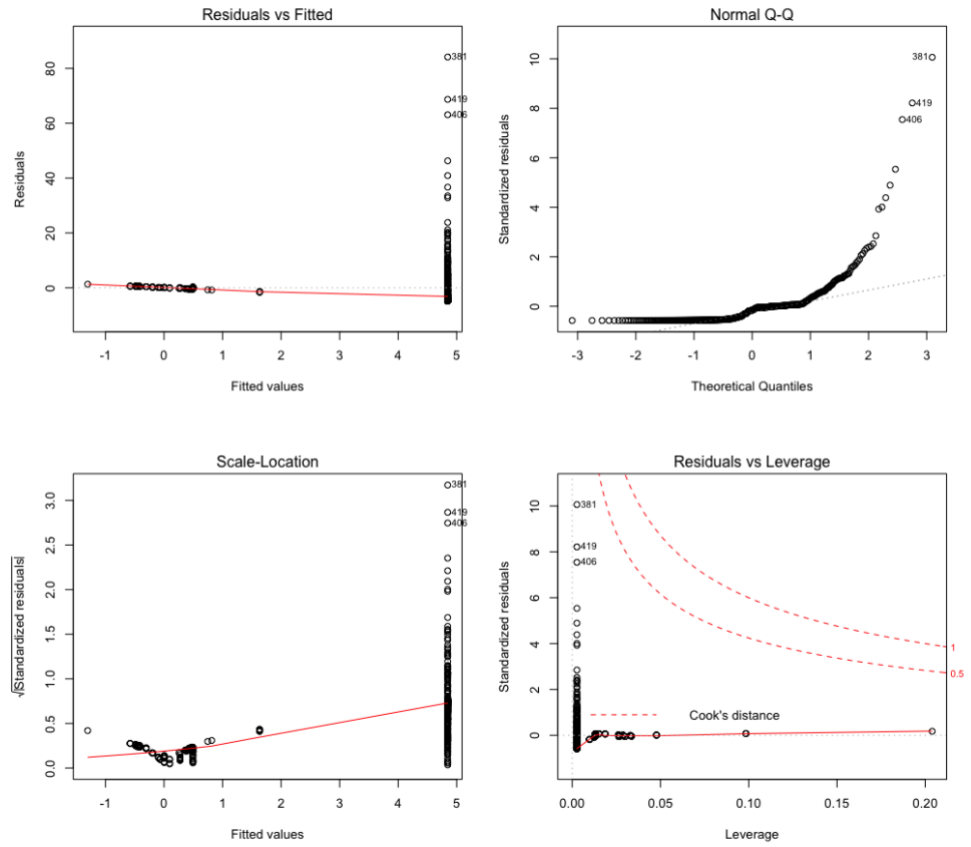
F-statistic: 10.35 on 3 and 502 DF **p-value:** 1.281e-06

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor is squared and cubed their significance dropped from the original state. Zn has a p-value close to 0, zn^2 has a p-value of .09375 and zn^3 has a p-value of .22954 at a confidence level of 95% we fail to reject the H_0, hence there is some evidence of nonlinear relationship between the predictor and the response.

The model only explains 5.824% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 10.32 on 3 and 502 degrees of freedom which states that the model is significant because one of coefficient is non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor zn.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	35862					
502	35187 2	2	674.56	4.8118	0.008512	**
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 4.8118 and p-value 0.008512 we reject H_0,hence we can say that the cubic model fits the data better for the predictor zn.

2. Nonlinear transformation of the predictor indus

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-8.278	-2.514	0.054	0.764	79.713

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	3.6625683	1.5739833	2.327	0.0204	*
indus	-1.9652129	0.4819901	-4.077	5.30e-05	***
$indus^2$	0.2519373	0.0393221	6.407	3.42e-10	***
$indus^3$	-0.0069760	0.0009567	-7.292	1.20e-12	***

Residual standard error: 7.423 on 502 degrees of freedom

Multiple R-squared: 0.2597 **Adjusted R-squared:** 0.2552

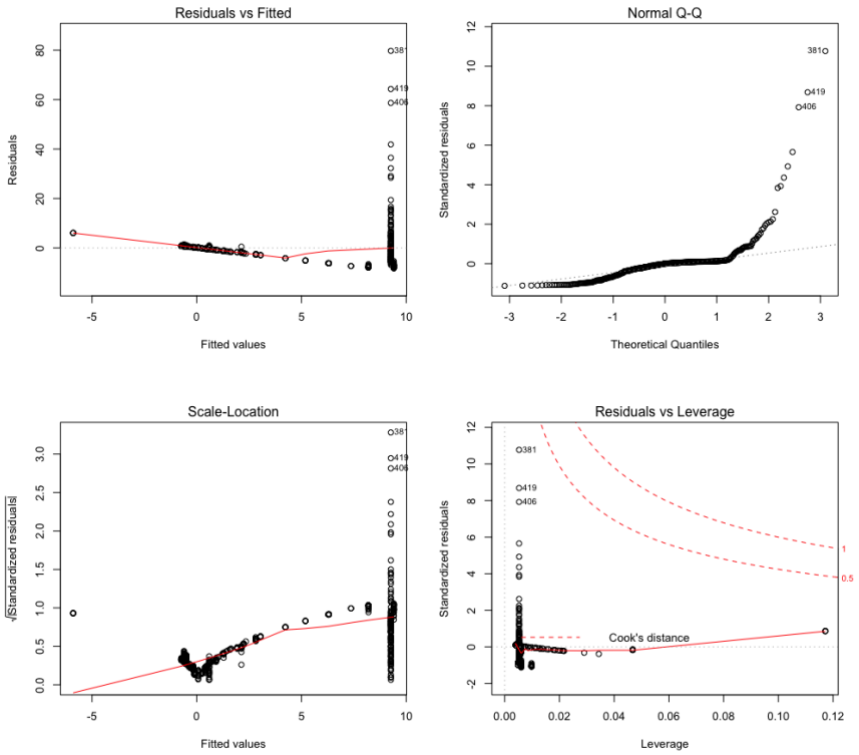
F-statistic: 58.69 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **indus** is squared and cubed they appeared to be statistically significant at a confidence level of 95% we reject the H_0, hence there is some evidence of nonlinear relationship between the predictor and the response.

The model only explains 25.97% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 58.69 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor indus.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	31187					
502	27662	2	3525.1	31.987	8.409e-14	***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 31.987 and p-value close 0 we reject H_0,hence we can say that the cubic model fits the data better for the predictor indus.

3. Nonlinear transformation of the predictor chas

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-3.738	-3.661	-3.435	0.018	85.232

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	3.7444	0.3961	9.453	<2e-16	***
chas	-1.8928	1.5061	-1.257	0.209	
$chas^2$	NA	NA	NA	NA	
$chas^3$	NA	NA	NA	NA	

Residual standard error: 8.597 on 504 degrees of freedom

Multiple R-squared: 0.003124 **Adjusted R-squared:** 0.001146

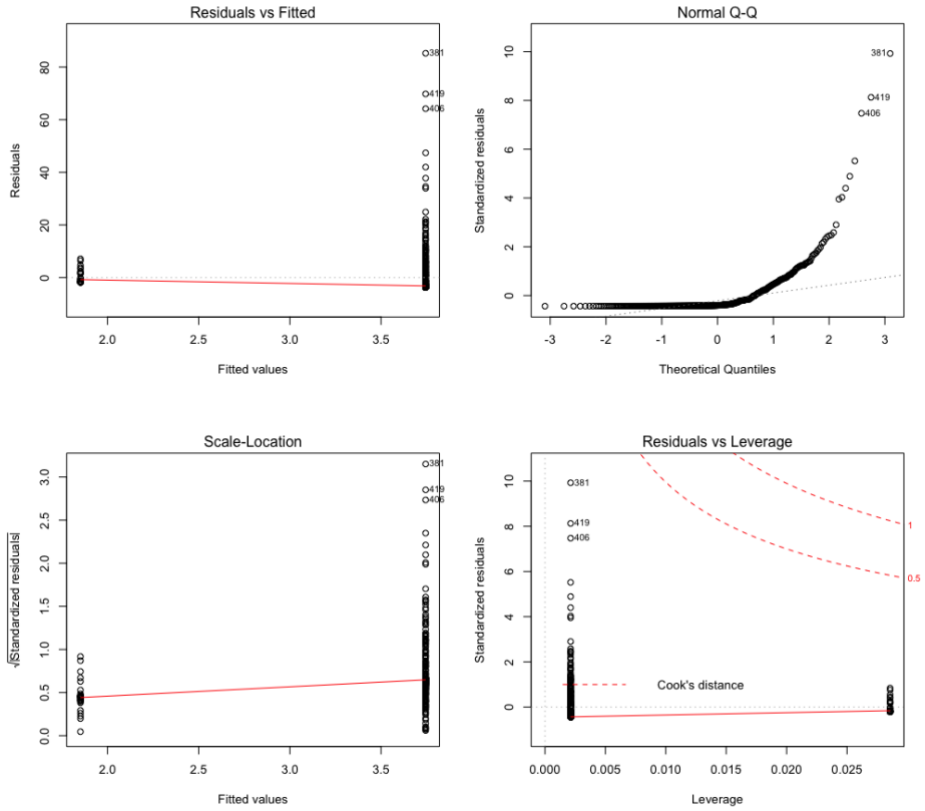
F-statistic: 1.579 on 1 and 504 DF **p-value:** 0.2094

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **chas** is squared and cubed their significance dropped from the original state. chas has a p-value .209, $chas^2$ and $chas^3$ returns NA value because chas is a dummy variable, at a confidence level of 95% we fail to reject the H_0, hence there is some evidence of nonlinear relationship between the predictor and the response.

The model only explains 0.3124% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 1.579 on 3 and 502 degrees of freedom with a p-value of .2094 which states that the model is not significant because none of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor ichas.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	37247					
504	37247	0	0			
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

Since chas is a dummy variable is it evident that the linear model is a better fit.

4. Nonlinear transformation of the predictor nox

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-9.110	-2.068	-0.255	0.739	78.302

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	233.09	33.64	6.928	1.31e-11	***
nox	-1279.37	170.40	-7.508	2.76e-13	***
nox^2	2248.54	279.90	8.033	6.81e-15	***
nox^3	-1245.70	149.28	-8.345	6.96e-16	***

Residual standard error: 7.234 on 502 degrees of freedom

Multiple R-squared: 0.297 **Adjusted R-squared:** 0.2928

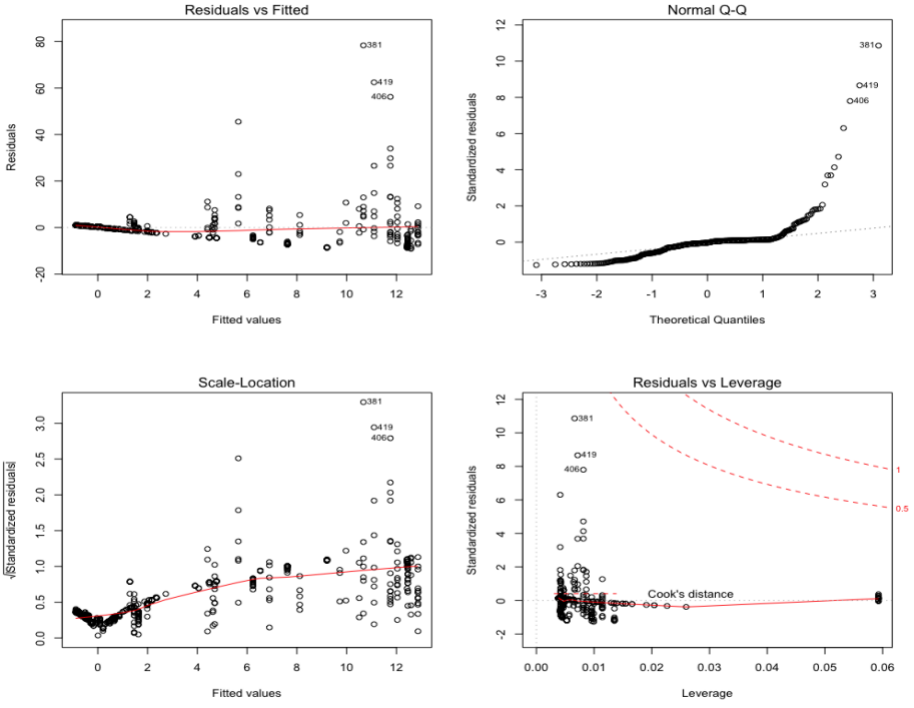
F-statistic: 70.69 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **nox** is squared and cubed they appeared to be statistically significant at a confidence level of 95% we reject the H_0, hence there is some evidence of nonlinear relationship between the predictor and the response.

The model only explains 29.28% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 70.69 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor nox.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	30742					
502	26267	2	4474.6	42.758	< 2.2e-16	***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 42.758 and p-value close 0 we reject H_0,hence we can say that the cubic model fits the data better for the predictor nox.

5. Nonlinear transformation of the predictor rm

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-18.485	-3.468	-2.221	-0.015	87.219

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	112.6246	64.5172	1.746	0.0815	.
rm	-39.1501	31.3115	-1.250	0.2118	
rm^2	4.5509	5.0099	0.908	0.3641	
rm^3	-0.1745	0.2637	-0.662	0.5086	

Residual standard error: 8.33 on 502 degrees of freedom

Multiple R-squared: 0.06779 **Adjusted R-squared:** 0.06222

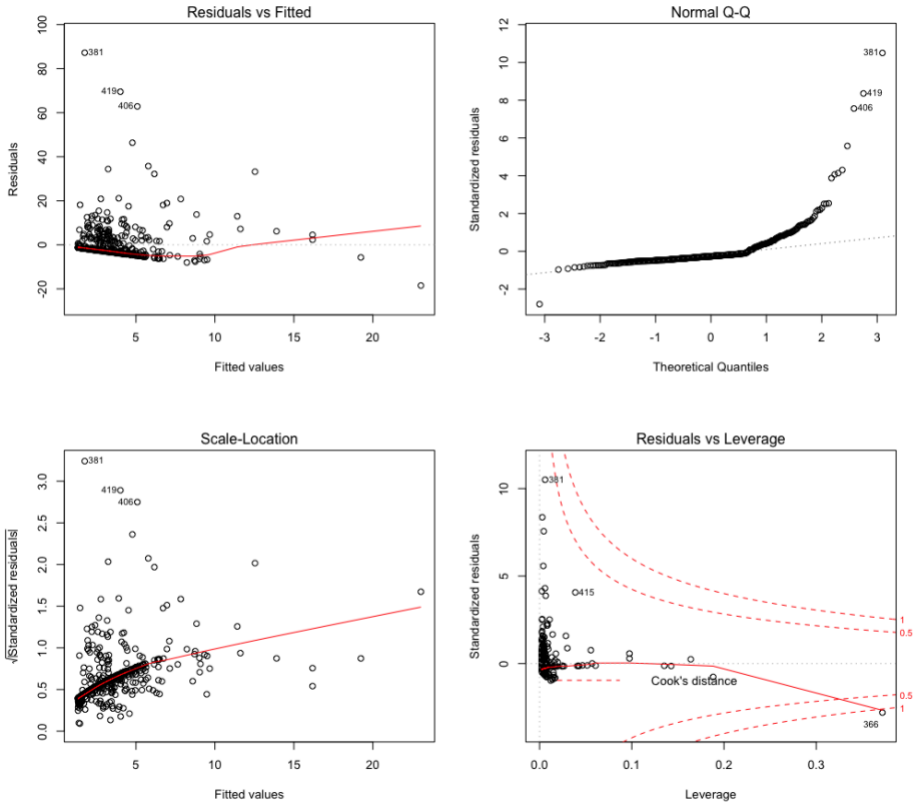
F-statistic: 12.17 on 3 and 502 DF **p-value:** 1.067e-07

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor is squared and cubed their significance dropped from the original state. rm has a p-value of 0.2118, rm^2has a p-value of .3641 and rm^3has a p-value of .5086 at a confidence level of 95% we fail to reject the H_0, hence there is no evidence of nonlinear relationship between the predictor and the response.

The model only explains 6.779% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 12.17 on 3 and 502 degrees of freedom which states that the model is significant because at least one of coefficient is non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor rm.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	35567					
502	34831	2	736.69	5.3088	0.005229	**
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 5.3088 and p-value 0.005229 we reject H_0 at a confidence level of 95%,hence we can say that the cubic model fits the data better for the predictor rm.

6. Nonlinear transformation of the predictor age

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-9.762	-2.673	-0.516	0.019	82.842

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	-2.549e+00	2.769e+00	-0.920	0.35780	
age	2.737e-01	1.864e-01	1.468	0.14266	
age^2	-7.230e-03	3.637e-03	-1.988	0.04738	*
age^3	5.745e-05	2.109e-05	2.724	0.00668	**

Residual standard error: 7.84 on 502 degrees of freedom

Multiple R-squared: 0.1742 **Adjusted R-squared:** 0.1693

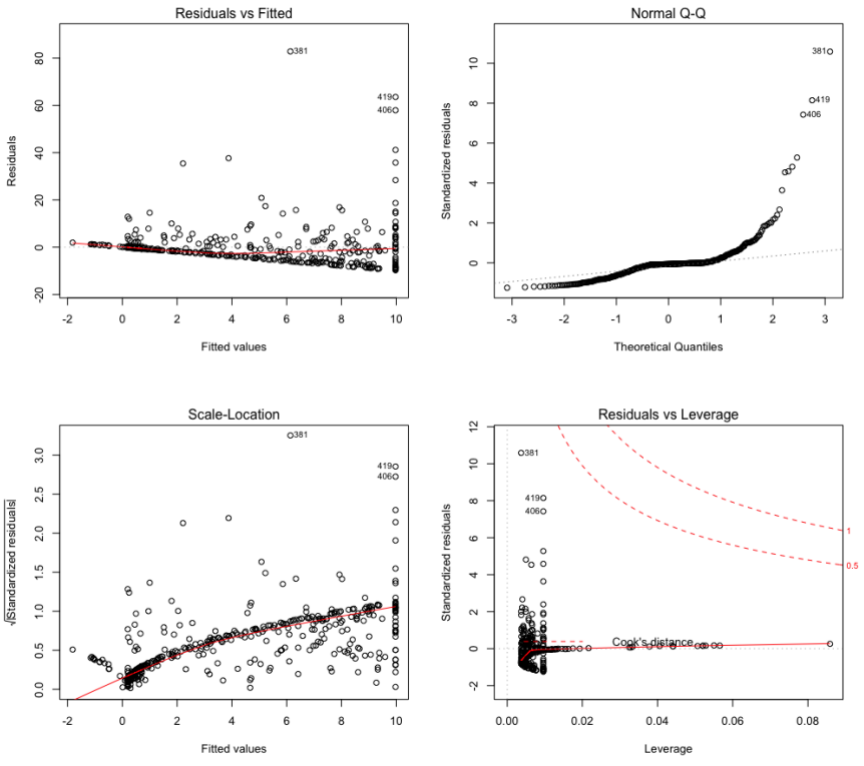
F-statistic: 35.31 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **age** is squared and cubed they appeared to be statistically significant at a confidence level of 95% we reject the H_0, hence there is some evidence of nonlinear relationship between the predictor and the response.

The model only explains 16.93% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 35.31 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor age.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	32714					
502	30853	2	1861	15.14	4.125e-07	***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 5.3088 and p-value close to 0, we reject H_0 confidence level of 95%, hence we can say that the cubic model fits the data better for the predictor age.

7. Nonlinear transformation of the predictor dis

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-10.757	-2.588	0.031	1.267	76.378

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	30.0476	2.4459	12.285	< 2e-16	***
dis	-15.5543	1.7360	-8.960	< 2e-16	***
dis^2	2.4521	0.3464	7.078	4.94e-12	***
dis^3	-0.1186	0.0204	-5.814	1.09e-08	***

Residual standard error: 7.331 on 502 degrees of freedom

Multiple R-squared: 0.2778 **Adjusted R-squared:** 0.2735

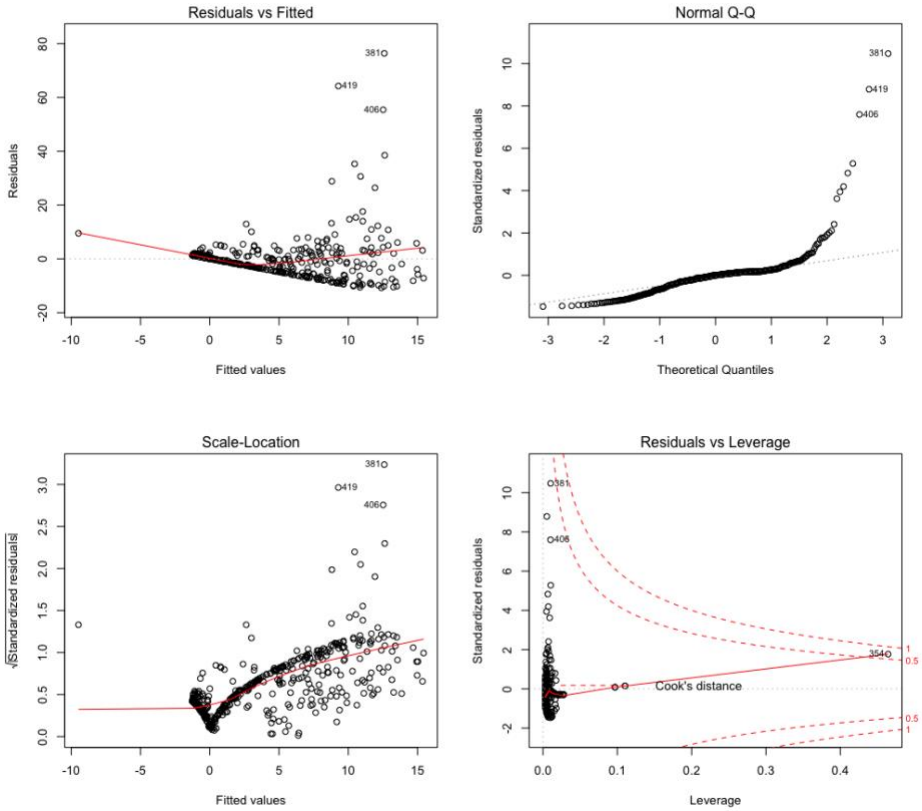
F-statistic: 64.37 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **dis** is squared and cubed they appeared to be statistically significant at a confidence level of 95% we reject the H_0, hence there is some evidence of nonlinear relationship between the predictor and the response.

The model only explains 27.35% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 64.37 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor dis.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	31977					
502	26983	2	4994.5	46.46	< 2.2e-16	***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 46.46 and p-value close to 0, we reject H_0 confidence level of 95%, hence we can say that the cubic model fits the data better for the predictor dis.

8. Nonlinear transformation of the predictor rad

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-10.381	-0.412	-0.269	0.179	76.217

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	-0.605545	2.050108	-0.295	0.768	
rad	0.512736	1.043597	0.491	0.623	
rad^2	-0.075177	0.148543	-0.506	0.613	
rad^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

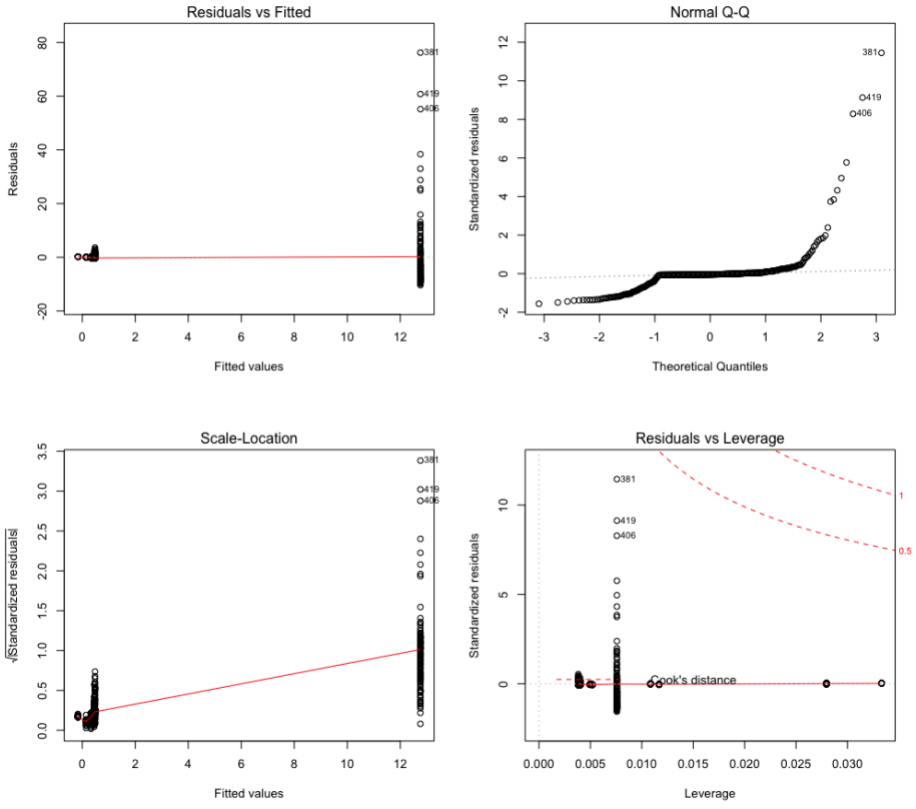
F-statistic: 111.6 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **rad** is squared and cubed they appeared to be statistically non-significant at a confidence level of 95% we fail to reject the H_0, hence there is no evidence of nonlinear relationship between the predictor and the response.

The model only explains 39.65% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 111.6 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor rad.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	22745					
502	22417	2	328.06	3.6733	0.02608	*
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 3.6733 and p-value close to 0, we reject H_0 confidence level of 95%, hence we can say that the cubic model fits the data better for the predictor rad.

9. Nonlinear transformation of the predictor tax

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-13.273	-1.389	0.046	0.536	76.950

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	1.918e+01	1.180e+01	1.626	0.105	
tax	-1.533e-01	9.568e-02	-1.602	0.110	
tax^2	3.608e-04	2.425e-04	1.488	0.137	
tax^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

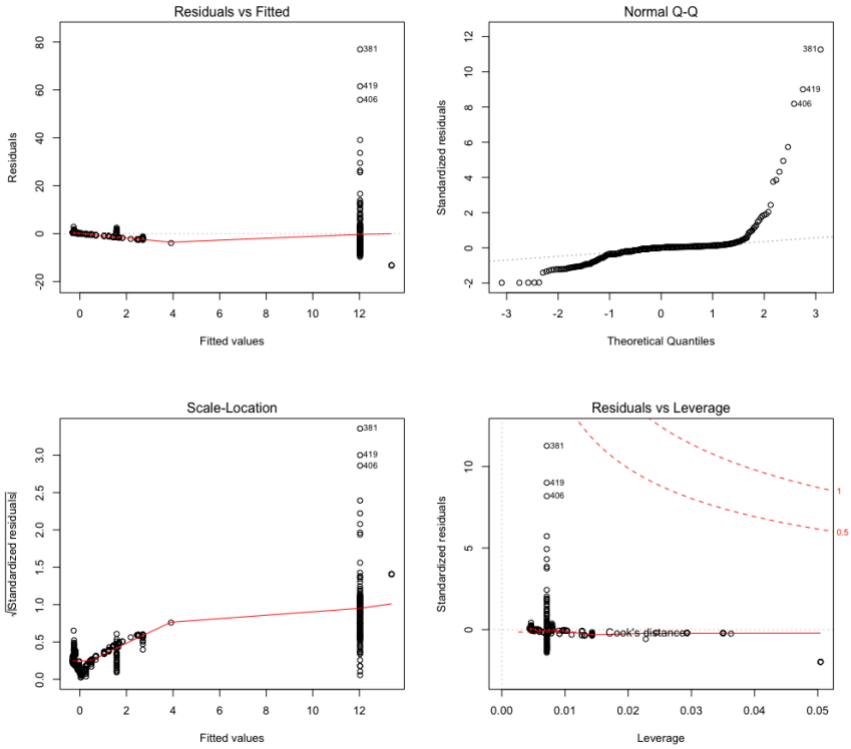
F-statistic: 97.8 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **tax** is squared and cubed they appeared to be statistically non-significant at a confidence level of 95% we fail to reject the H_0, hence there is no evidence of nonlinear relationship between the predictor and the response.

The model only explains 36.51% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 97.8 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor tax.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	24674					
502	23581	2	1093.5	11.64	1.144e-05	***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 11.64 and p-value close to 0, we reject H_0 confidence level of 95%, hence we can say that the cubic model fits the data better for the predictor tax.

10. Nonlinear transformation of the predictor ptratio

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-6.833	-4.146	-1.655	1.408	82.697

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	477.18405	156.79498	3.043	0.00246	**
ptratio	-82.36054	27.64394	-2.979	0.00303	**
ptratio^2	4.63535	1.60832	2.882	0.00412	**
ptratio^3	-0.08476	0.03090	-2.743	0.00630	**

Residual standard error: 8.122 on 502 degrees of freedom

Multiple R-squared: 0.1138 **Adjusted R-squared:** 0.1085

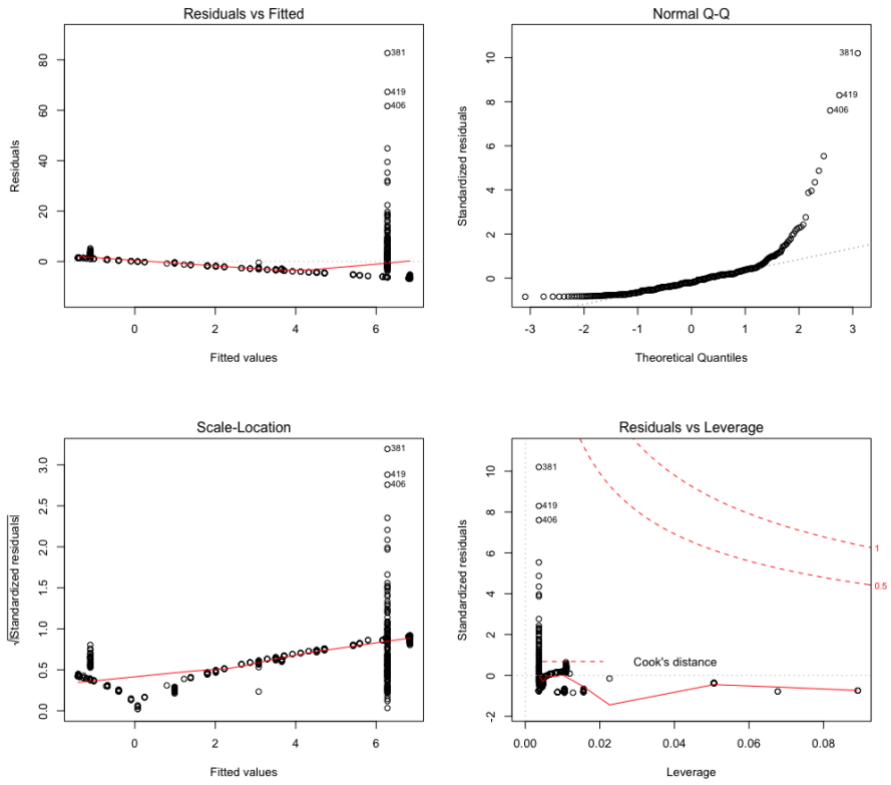
F-statistic: 21.48 on 3 and 502 DF **p-value:** 4.171e-13

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **ptratio** is squared and cubed they appeared to be statistically significant at a confidence level of 95% we reject the H_0, hence there is some evidence of nonlinear relationship between the predictor and the response.

The model only explains 10.85% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 21.48 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor ptratio.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	34222					
502	33112	2	1110.2	8.4155	0.0002542	***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 8.4155 and p-value close to 0, we reject H_0 confidence level of 95%, hence we can say that the cubic model fits the data better for the predictor ptratio.

11. Nonlinear transformation of the predictor black

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-13.096	-2.343	-2.128	-1.439	86.790

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	1.826e+01	2.305e+00	7.924	1.5e-14	***
black	-8.356e-02	5.633e-02	-1.483	0.139	
black^2	2.137e-04	2.984e-04	0.716	0.474	
black^3	-2.652e-07	4.364e-07	-0.608	0.544	

Residual standard error: 7.955 on 502 degrees of freedom

Multiple R-squared: 0.1498 **Adjusted R-squared:** 0.1448

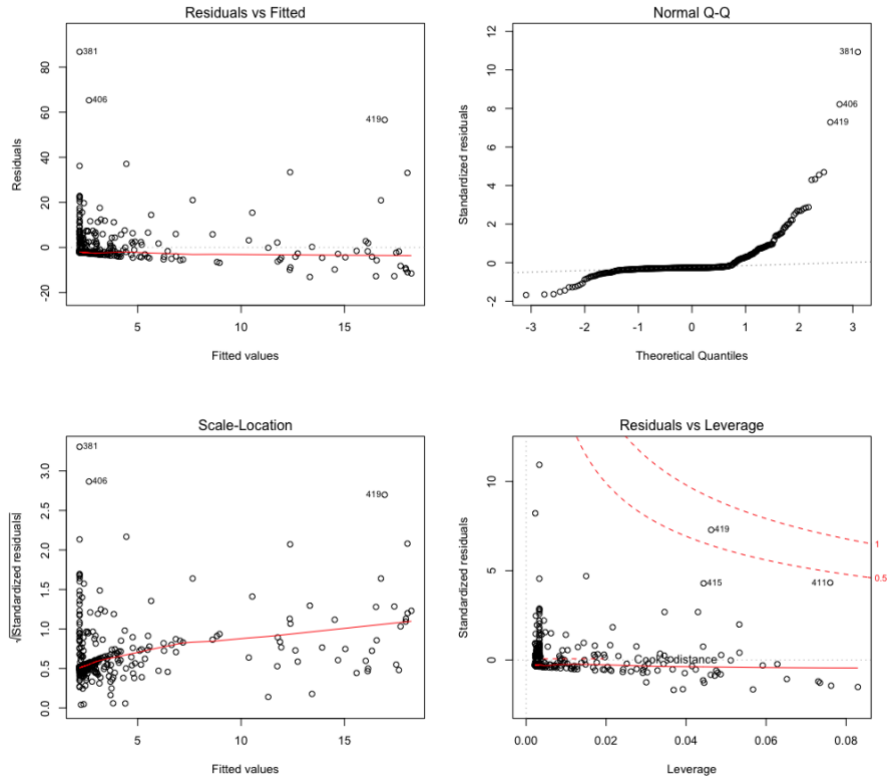
F-statistic: 29.49 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **black** is squared and cubed they appeared to be statistically non-significant at a confidence level of 95% we fail to reject the H_0, hence there is no evidence of nonlinear relationship between the predictor and the response.

The model only explains 14.48% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 29.49 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor black.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	31823					
502	31765	2	58.495	0.4622	0.6302	
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of .4622 and p-value of 0.6302, we fail to reject H_0 confidence level of 95%, hence we can say that both the models fit the data well for the predictor black.

12. Nonlinear transformation of the predictor lstat

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-15.234	-2.151	-0.486	0.066	83.353

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	1.2009656	2.0286452	0.592	0.5541	
lstat	-0.4490656	0.4648911	-0.966	0.3345	
lstat^2	0.0557794	0.0301156	1.852	0.0646	.
lstat^3	-0.0008574	0.0005652	-1.517	0.1299	

Residual standard error: 7.629 on 502 degrees of freedom

Multiple R-squared: 0.2179 **Adjusted R-squared:** 0.2133

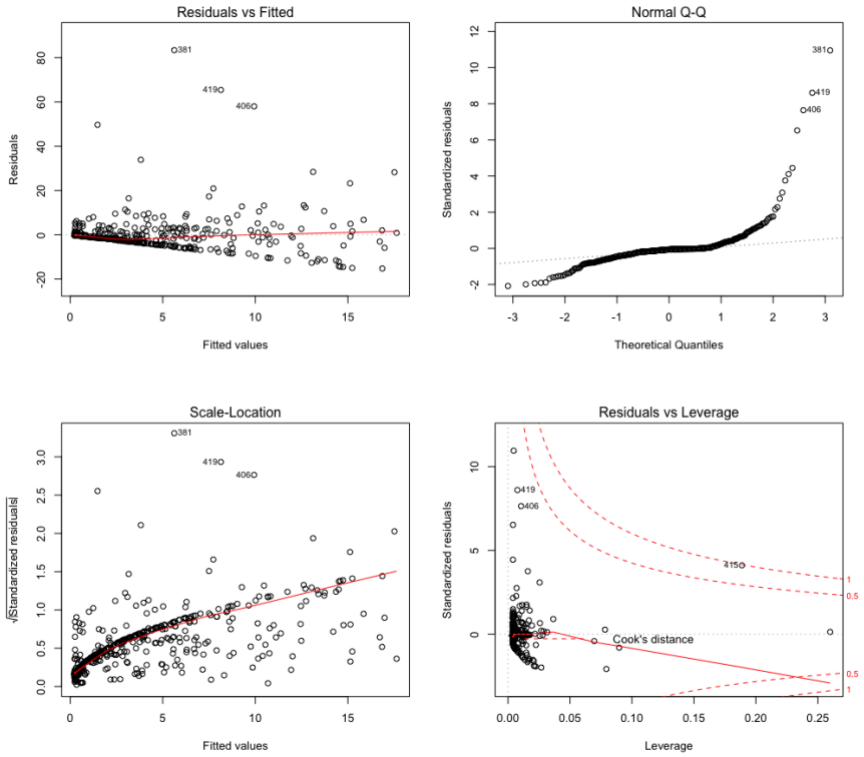
F-statistic: 46.63 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **lstat** is squared and cubed they appeared to be statistically non-significant at a confidence level of 95% we fail to reject the H_0, hence there is no evidence of nonlinear relationship between the predictor and the response.

The model only explains 21.33% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 44.63 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor lstat.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	29607					
502	29221	2	386.39	3.319	0.03698	*
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 3.319 and p-value of 0.03698, we reject H_0 confidence level of 95%, hence we can say that the cubic model fits the data better for the predictor tax.

13. Nonlinear transformation of the predictor medv

Residuals:

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-24.427	-1.976	-0.437	0.439	73.655

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif. Code**
Intercept(β_0)	53.1655381	3.3563105	15.840	< 2e-16	***
medv	-5.0948305	0.4338321	-11.744	< 2e-16	***
$medv^2$	0.1554965	0.0171904	9.046	< 2e-16	***
$medv^3$	-0.0014901	0.0002038	-7.312	1.05e-12	***

Residual standard error: 6.569 on 502 degrees of freedom

Multiple R-squared: 0.4202 **Adjusted R-squared:** 0.4167

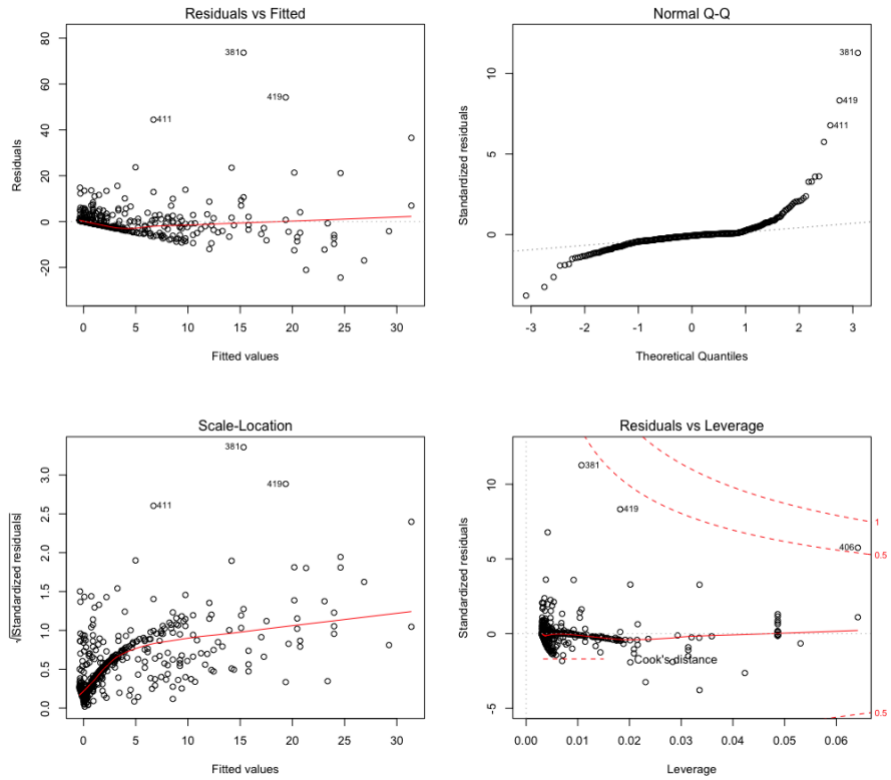
F-statistic: 121.3 on 3 and 502 DF **p-value:** < 2.2e-16

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

In this model, it is evident that when the predictor **medv** is squared and cubed they appeared to be statistically significant at a confidence level of 95% we reject the H_0, hence there is some evidence of nonlinear relationship between the predictor and the response.

The model only explains 41.67% of the variation in per capita crime rate, that means majority of the variation isn't explained by the model.

F-statistics returns a value of 121.3 on 3 and 502 degrees of freedom with a p-value close to 0 which states that the model is significant because all of coefficients are non-zero.



- The above plot shows that the cubic fit doesn't explain much of the variation in the per capita crime rate (y).
- The QQ plot has many points off the line which could be a suggestion of outliers.
- Both the Scale- location and Residuals vs Leverage plots show that the points are scattered away from the center which means that some points have excessive leverage.

We can fit an **anova** function to fit compare the simple linear model and the larger cubic model of the predictor medv.

H_0: Both the models fit the data equal well
H_a: The cubic model fits the data better than the simple linear model

Anova test

Res.DF	RSS	Df	Sum of Sq	F-value	P-value	Signif. Code
504	31730					
502	21663	2	10066	116.63	< 2.2e-16	***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

With a F-value of 8.4155 and p-value close to 0, we reject H_0 confidence level of 95%, hence we can say that the cubic model fits the data better for the predictor medv.

The predictors (indus,nox,age,dis,ptratio,medv) shows an evidence of non-linear relationship because when either of those are squared or cubed they appear statistically significant.

f) Logistic Regression & LDA

(i)

Logistic regression is used to predict the likelihood of **Per capita crime rate** based on a subset of predictors (**zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv**) using the model:

$$Y = \beta_0 + \beta_{1X_1} + \beta_{2X_2} + \beta_{3X_3} + \cdots + \beta_{nX_n} + \varepsilon$$

Per capita crime will be categorical variable, where anything **above the median** will be **1 (High per capita crime rate)** and **below the median** will be **0** (**Low per capita crime rate)**

Accessing the summary of the new data frame N.Boston where crim.cv is a categorical variable.

```
> summary(N.Boston)
       zn             indus            chas              nox              rm             age              dis        
 Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
 Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
 Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
 3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
 Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
      rad              tax            ptratio          black            lstat            medv          crim.cv     
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00   Min.   :0.0  
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02   1st Qu.:0.0  
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20   Median :0.5  
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53   Mean   :0.5  
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00   3rd Qu.:1.0  
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00   Max.   :1.0  
```

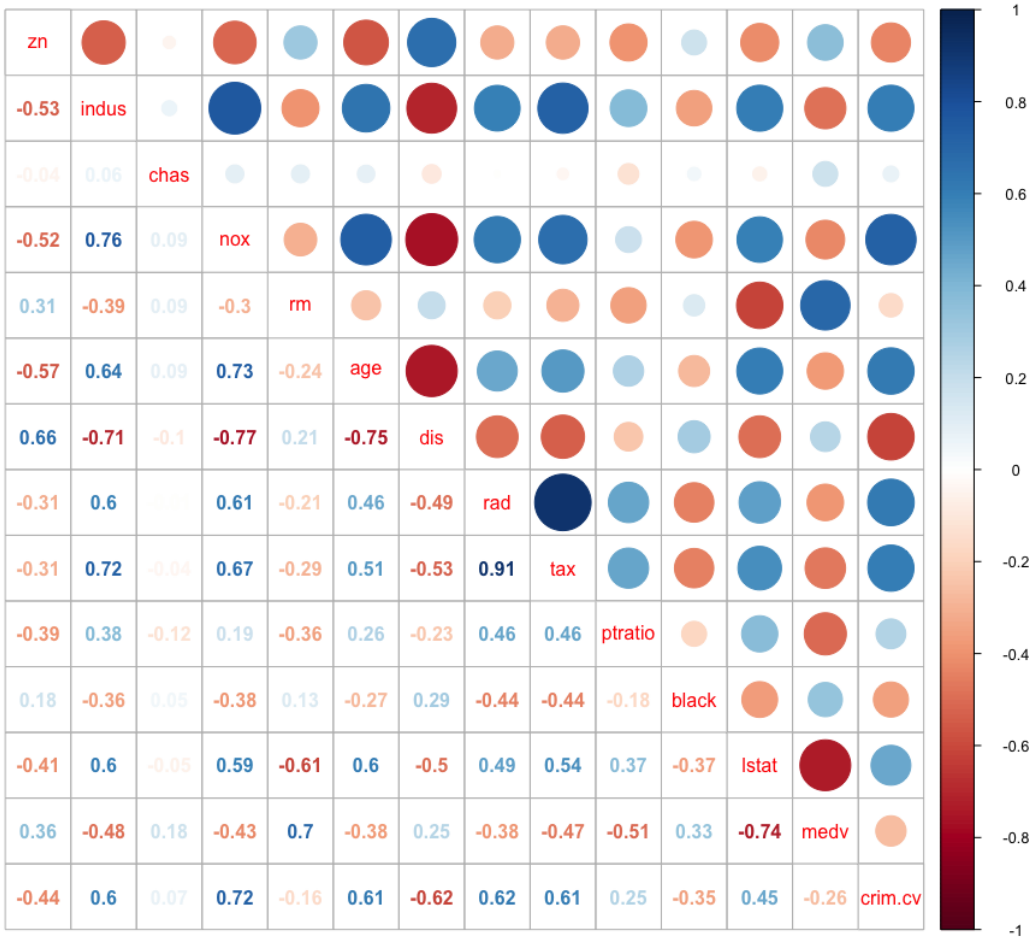
crime.cv returns a minimum of 0 and a maximum of 1 with a median of 0.5

Calculate the Pearson coefficient of correlation of the predictor (y: crim.cv) with each response (x)

cor (y,x)	r value	Relation
Crim.cv, zn	-0.43615103	Weak negative
Crim.cv, indus	0.60326017	Strong positive
Crim.cv, chas	0.07009677	Very weak positive / no relation
Crim.cv, nox	0.72323480	Strong positive
Crim.cv, rm	-0.15637178	Weak negative / no relation
Crim.cv, age	0.61393992	Strong positive
Crim.cv, dis	-0.61634164	Weak negative
Crim.cv, rad	0.61978625	Strong positive
Crim.cv, tax	0.60874128	Strong positive
Crim.cv, ptratio	0.25356836	Weak positive
Crim.cv, black	-0.35121093	Weak negative
Crim.cv, lstat	0.45326273	Moderately strong
Crim.cv, medv	-0.26301673	Weak negative

The predictors (indus,nox,age,rad,tax,lstat) have a strong positive relation with crim.cv when compare to the other predictors.

Visualising the correlation coefficient's using corrplot function in R.



From the correlation plot it is evident that the strongest correlation is between **nox** and c**rim.cv** followed by **rad, tax, rm, indus and lstat**.

Using R, we can fit a logistic regression model using **crim.cv** as the response (y) variable and each of the other elements of the Boston as the predictor (x) to find what is the likelihood of the response(y) based on the predictors.

Response (y): crim.cv
Predictor (x): zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv

$$\boldsymbol{Y = \beta_0 + \beta_{1X_1} + \beta_{2X_2} + \beta_{3X_3} + \cdots + \beta_{nX_n} + \varepsilon}$$

Deviance Residuals

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	Maximum
-2.3946	-0.1585	-0.0004	0.0023	3.4239

Coefficients:

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**Signif. Code**
(Intercept)	-34.103704	6.530014	-5.223	1.76e-07	***
zn	-0.079918	0.033731	-2.369	0.01782	*
indus	-0.059389	0.043722	-1.358	0.17436	
chas	0.785327	0.728930	1.077	0.28132	
nox	48.523782	7.396497	6.560	5.37e-11	***
rm	-0.425596	0.701104	-0.607	0.54383	
age	0.022172	0.012221	1.814	0.06963	.
dis	0.691400	0.218308	3.167	0.00154	**
rad	0.656465	0.152452	4.306	1.66e-05	***
tax	-0.006412	0.002689	-2.385	0.01709	*
ptratio	0.368716	0.122136	3.019	0.00254	**
black	-0.013524	0.006536	-2.069	0.03853	*
lstat	0.043862	0.048981	0.895	0.37052	
medv	0.167130	0.066940	2.497	0.01254	*
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	701.46 on 505 degrees of freedom
Residual deviance:	211.93 on 492 degrees of freedom
AIC: 239.93	

Using these p-values we can reject the null hypothesis H_0: $\beta_j = 0$ (which states that the intercept value is equal to zero) at a 95% confidence level (.05 level of significance) for the predictors (**zn, nox, dis, rad, tax, ptratio, black, medv**), which means that these predictors are statistically significant (there is some relation between the response variable and the predictors).

Using the frequentist approach (Confidence Intervals, Hypothesis Testing) we can select the significant predictors.

Considering the confidence intervals (95%) of the predictors we can assess their significance.

	2.5%	97.5%
Intercept($\boldsymbol{\beta_0}$)	-47.480389822	-21.699753794
zn	-0.152359922	-0.020567540
indus	-0.149113408	0.024168460
chas	-0.646429219	2.233443233
nox	34.967619055	64.088411260
rm	-1.811639107	0.950196261
age	-0.001231256	0.046865843
dis	0.280762523	1.140619391
rad	0.376833861	0.975898274
tax	-0.012038221	-0.001324887
ptratio	0.136910471	0.618725856
black	-0.029151201	-0.002990159
lstat	-0.053062947	0.139446105
medv	0.040925281	0.304379859

The corresponding confidence interval provides the uncertainty in the estimate. That means in repeated random sampling, the computed confidence interval straddles the true but unknown coefficient 95% of the time.

From the z-test results it is evident that the confidence intervals of predictors (**zn, nox, dis, rad, tax, ptratio, black, medv**) doesn't contain zero. Whereas the confidence intervals of the other predictors (**indus, chas, rm, age, lstat**) straddles zero.

Since the p-values of the coefficients of the predictors (**indus, chas, rm, age, lstat**) are quite large, these variables are dropped from the model.

Deviance Residuals

Minimum	**1st Quartile**	**Median**	**3rd Quartile**	Maximum
-2.4400	-0.1918	- 0.0008	0.0025	3.1885

Coefficients:					
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**Signif. Code**
(Intercept)	-28.347709	5.569863	-5.089	3.59e-07	***
zn	-0.074499	0.029975	-2.485	0.01294	*
nox	44.180443	6.289746	7.024	2.15e-12	***
dis	0.489849	0.194930	2.513	0.01197	*
rad	0.692116	0.137842	5.021	5.14e-07	***
tax	-0.007448	0.002428	-3.067	0.00216	**
ptratio	0.272145	0.107311	2.536	0.01121	*
black	-0.013484	0.006331	-2.130	0.03317	*
medv	0.087913	0.030787	2.856	0.00430	**
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '
(Dispersion parameter for binomial family taken to be 1)					
Null deviance:	701.46 on 505 degrees of freedom				
Residual deviance:	221.78 on 497 degrees of freedom				
AIC:	239.78				

The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

- For every one unit change in zn, the log odds of High Per capita crime rate decrease by -0.074499.
- For every one unit change in nox, the log odds of High Per capita crime rate increase by 44.180443.
- For every one unit change in dis, the log odds of High Per capita crime rate increase by 0.489849.
- For every one unit change in rad, the log odds of High Per capita crime rate increase by 0.692116.
- For every one unit change in tax, the log odds of High Per capita crime rate decrease by -0.007448.
- For every one unit change in ptratio, the log odds of High Per capita crime rate increase by 0.272145.
- For every one unit change in black, the log odds of High Per capita crime rate decrease by -0.013484.
- For every one unit change in medv, the log odds of High Per capita crime rate increase by 0.087913.

In logistic regression, **deviance** is defined to be -2*logL, where L is the maximized value of the likelihood function that was used to obtain the parameter estimates.

The **null deviance** is the value where the likelihood function is based only on the intercept term (y=β_0).
The **residual deviance** is the value where the likelihood function is based on the specified logistic model.

The pseudo-R^2 is a measure of how well the fitted model explains the data as compared to the default model of no predictor variables and only an intercept term. A pseudo-R^2 value near to 1 indicates a good fit over the simple null model.

pseudo-$R^2 = 1 - \frac{residual\ dev.}{null\ dev.} = \frac{null\ dev - res.dev.}{null\ dev.}$

$= \frac{701.46 - 221.78}{701.46} = 0.683830867$

Which means that the models a good fit over the simple null model.

Using the confusion matrix, we can compute the error rate.

Confusion Matrix

	Crim.cv	
Glm.pred	**0**	**1**
0	232	25
1	21	228

The diagonal elements of the confusion matrix indicate correct predictions while the off diagonals represent incorrect predictions. The model correctly predicted 232 areas in the

Boston will have Per capita crime rate below the median **(Low Per capita crime rate)** and 228 areas will have a Per capita crime rate above the median **(High Per capita crime rate)** out of the 506 observations in the Boston data set.

$$\frac{232 + 228}{506} = .90909$$

The logistic regression correctly predicted the Per capita crime rate using the predictors (zn, nox, dis, rad, tax, ptratio, black, medv) 90.90% of the time.

Using this value, we can calculate the sampling error of the model

1 - 0.90909 = 0.09091

The logistic model has a sampling error of 9.091%, which suggests that the model is accurate at prediction.

(ii) LDA

Using R, we can fit a Linear discriminant analysis (LDA) using **crim.cv** as the response (y) variable and a subset of other elements of the Boston data set as the predictor (x) to use means and the variances of each class to create a linear boundary between them.

Response (y): crim.cv
Predictor (x): zn, nox, dis, rad, tax, ptratio, black, medv

$$\boldsymbol{Y = \beta_0 + \beta_{1X_1} + \beta_{2X_2} + \beta_{3X_3} + \cdots + \beta_{nX_n} + \varepsilon}$$

Per capita crime will be categorical variable, where anything **above the median** will be **High per capita crime rate** and **below the median** will be **Low per capita crime rate**

Prior probabilities of groups:

0	**1**
0.5	0.5

Group means:

	zn	**nox**	**dis**	**rad**	**tax**	**ptratio**	**black**	**medv**
0	21.525692	0.4709711	5.091596	4.158103	305.7431	17.90711	388.7061	24.94941
1	1.201581	0.6384190	2.498489	14.940711	510.7312	19.00395	324.6420	20.11621

Coefficients of linear discriminants:

Variables	**LD1**
zn	-0.006320287
nox	9.320301191
dis	-0.043164803
rad	0.069028308
tax	-0.000546202
ptratio	0.056626313
black	-0.001073684
medv	0.030819347

The prior probabilities of the groups are the ones that exist within the sample Boston dataset i.e. 50% of the data corresponds to **Lower per capita crime rate** evaluated as **0** and 50% of the other data corresponds to **High per capita crime rate** evaluated as **1**.

The **group means** are the average of each predictor within each class (0,1). These values suggest that:

- The variable **zn** will have a higher influence on **Lower per capita crime rate (**21.525692**)** than on **High per capita crime rate (**1.201581**).**
- The variable **nox** will have a slightly higher influence on **High per capita crime rate (**0.6384190**)** than on **Lower per capita crime rate (**0.4709711**).**
- The variable **dis** will have a higher influence on **Lower per capita crime rate (**5.091596**)** than on **High per capita crime rate** (2.498489).
- The variable **rad** will have a higher influence on **High per capita crime rate (**14.940711**)** than on **Lower per capita crime rate** (4.158103).
- The variable **tax** will have a higher influence on **High per capita crime rate (**510.7312**)** than on **Lower per capita crime rate** (305.7431).
- The variable **ptratio** will have a slightly higher influence on **High per capita crime rate (**19.00395**)** than on **Lower per capita crime rate** (17.90711).
- The variable **black** will have a slightly higher influence on **Lower per capita crime rate (**388.7061**)** than on **High per capita crime rate (**324.6420**).**
- The variable **medv** will have a slightly higher influence on **Lower per capita crime rate (**24.94941**)** than on **High per capita crime rate (**20.11621**).**

The **Coefficients of linear discriminants** means that the boundary between the two different classes (0,1) will be specified by the formula:

Y = -0.006320287 * zn + 9.320301191 * nox -0.043164803 * dis + 0.069028308 * rad - 0.000546202 * tax + 0.056626313* ptratio -0.001073684 * black + 0.030819347 * medv

Using the confusion matrix, we can compute the error rate.

Confusion Matrix

	Crim.cv	
Classes	**0**	**1**
0	247	59
1	6	194

The diagonal elements of the confusion matrix indicate correct predictions while the off diagonals represent incorrect predictions. The model correctly predicted 247 areas in the Boston will have Per capita crime rate below the median **(Low Per capita crime rate)** and 194 areas will have a Per capita crime rate above the median **(High Per capita crime rate)** out of the 506 observations in the Boston data set.

$$\frac{247 + 194}{506} = .8715415$$

The LDA model correctly predicted the Per capita crime rate using the predictors (zn, nox, dis, rad, tax, ptratio, black, medv) 87.15415% of the time.

Using this value, we can calculate the sampling error of the model

1 - 0.8715415 = 0.1284585

The LDA model has a sampling error of 12.845%, which suggests that the model is accurate at prediction.

The LDA model returns the posterior probabilities, the no. of areas with a High per capita crime rate (above the median) as 306 and the no. of areas with Low crime rate as 200 out of the total 506 areas.

The **Logistic model** calculates the likelihood based on the frequentist approach using Confidence intervals and Hypothesis tests to find the significant predictors. Using that information, we can drop the non-significant predictors to build a good model. The **LDA** uses the Bayesian approach to statistics (using prior probability to determine the posterior probabilities) to create a linear boundary between the classes.

References

1. ‘Exploring bivariate numerical data’ (Accessed on: 2018-04-04): https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data

2. ‘Advanced Analytical Theory and Methods: Regression’ Data Science and Big Data Analytics – Discovering, Analysing, Visualizing and Presenting Data, EMC2

3. ‘Linear Regression’ pg. 59, Introduction to Statistical Learning with R, Stanford

4. ‘Classification’ pg. 127, Introduction to Statistical Learning with R, Stanford