Analytics for Business Intelligence

Project 1

Name: Anusha Savaram
Email: as344@scarletmail.rutgers.edu
Net ID: as344
RU ID: 205003732

Life expectancy

1. What is the best model to predict life expectancy with the data provided in the following two links?

- https://www.worldometers.info/demographics/life-expectancy/#countries-ranked-by-life-expectancy (Links to an external site.)
- https://www.worldometers.info/world-population/population-by-country/

Life Expectancy

1. Import data set 'Life_Expectancy_per_Country 'uploaded in the midterm section of modules in the canvas. Alternatively, we can use the links provided in the project 1 section of Assignment in canvas.

   Open 'https://www.worldometers.info/demographics/life-expectancy/' link and download the webpage using save as (webpage, html only).

   Read the saved file on R using:

   url1<- "/Users/anusha._s/Desktop/Rutgers/Sem 1/ABI/Project 1/Life Expectancy by Country and in the World (2021) - Worldometer.html"

   From XML package we use readHTMLTable to read the table.
   LifeExpec<- readHTMLTable(url1, which = 2, stringsAsFactors = FALSE,colClasses =c("numeric","character",rep("numeric",5)))

   Ensure that tables are imported into R
   View(LifeExpec)

   We can see that Life Expectancy table is imported into R.

   Similarly, repeat the steps with Population dataset as well.

url2<- "/Users/anusha._s/Desktop/Rutgers/Sem 1/ABI/Project 1/Population by Country (2021) - Worldometer.html"

Population<-readHTMLTable(url2, which = 1, stringsAsFactors = FALSE, coClasses= c("numeric","character", rep("numeric",10)))

View(Population)

Notes:
- Which is selecting the table from the webpage
- stringsAsFactors allows us to customize the datatypes.
- Colclasses is used to customize the datatype.

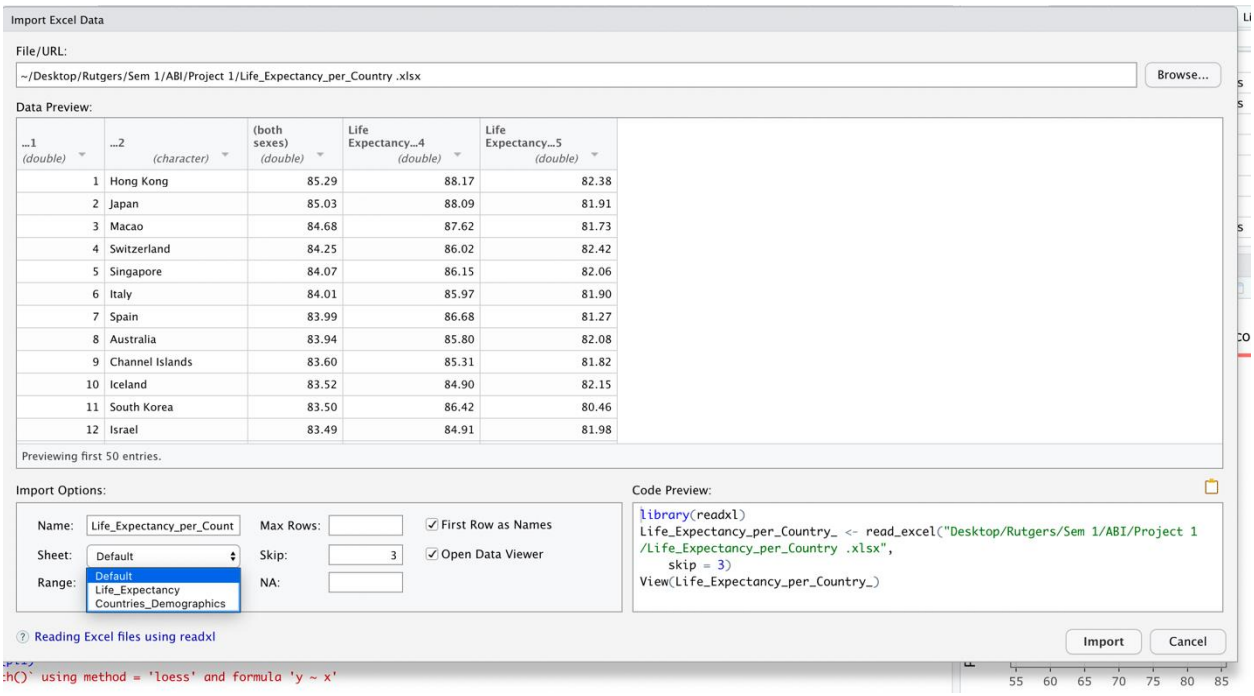
⇨ To find the best model to predict Life Expectancy of the Population. We should merge the Population and Life Expectancy table, study and analyze it.

⇨ We prepare the tables to merge. Autocorrelation helps us understand the influence of one attribute on another. This understanding enables us to build regression models.

⇨ To find autocorrelation, we convert all the columns with numbers from character datatype to numerical data type.

For Population Table:
I import the Population dataset after making the following changes. We have skip 3 rows as it has null values does not have any significant data . Range= Countries_Demographics

Population<-Life_Expectancy_per_Country_



Verify the datatypes using
str(Population)
str(LifeExpec)

Prepare to merge

In population table, I changed country (or dependency) to country. Country name in Life Expectancy table and Population table is same. We merger tables by country.

/*Column name is changed to country from Country(or dependency) in the population table*/
colnames(Population)[1]<-"Country"

Before merging delete # columns for both the tables by
LifeExpec<-LifeExpec[,-1]
Population<-Population[,-1] //for deleting rows
Population<-Population[-1,] //for deleting columns

⇨ We now merge Population and Life Expectancy table

/*Merging Life Expectancy and Population Table*/
Population_Story<-merge(LifeExpec,Population,by="Country")

A new merged dataset names Population_Story is created.
View(Population_Story)

Remove Null values from the dataset.

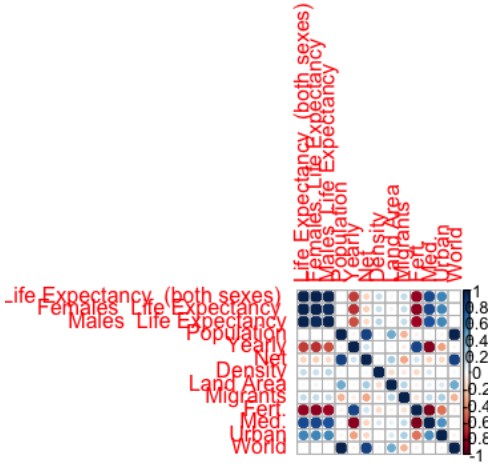
```
numberOfNA <- length(which(is.na(Population_Story)==T))
if(numberOfNA>0) {
  Population_Story <- Population_Story[complete.cases(Population_Story),]
}
```

⇨ **Prepare to analyze**

```
/*Preparing Dataset for analysis*/
train<-Population_Story[1:97,]
test<-Population_Story[98:195,]
```

We now plot the correlation of Population_Story to understand the influence of one attribute over other.

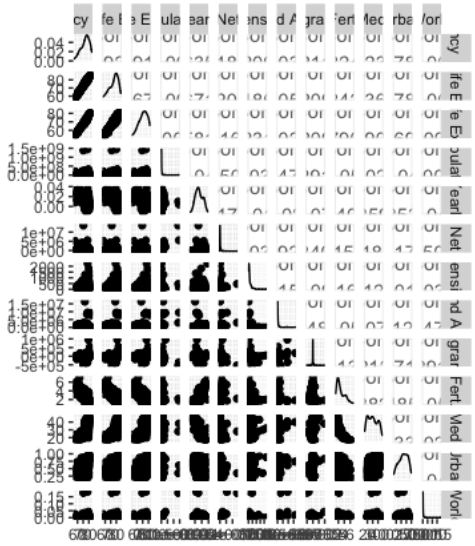
```
/*Correlation of Population Story
corrplot(cor(Population_Story[,-1])).     //We remove country name and plot the correlation as finding correlation between countries make no sense.
```



This plot shows that Female Life Expectancy and Male Life Expectancy have highest correlation with Life Expectancy (both sexes). So I built a model using these attributes.

Correlation Plot 2

```
library("GGally")
ggpairs(Population_Story[,-1])
```



This correlation plot shows a better detailing on attribute relation with one another.

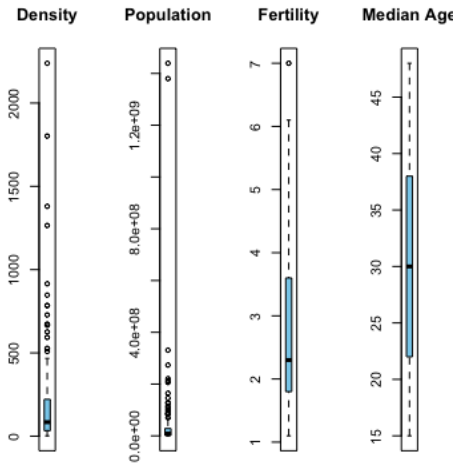
```
summary(Population_Story)
```

```
> summary(Population_Story)
   Country          Life Expectancy  (both sexes)  Females  Life Expectancy  Males  Life Expectancy   Population
 Length:195         Min.   :54.36                 Min.   :56.58            Min.   :52.16          Min.   :9.793e+04
 Class :character   1st Qu.:68.24                 1st Qu.:70.47            1st Qu.:66.22          1st Qu.:1.794e+06
 Mode  :character   Median :74.65                 Median :77.89            Median :71.75          Median :8.737e+06
                    Mean   :73.45                 Mean   :75.86            Mean   :71.06          Mean   :3.972e+07
                    3rd Qu.:78.99                 3rd Qu.:81.34            3rd Qu.:76.36          3rd Qu.:2.841e+07
                    Max.   :85.03                 Max.   :88.09            Max.   :82.42          Max.   :1.439e+09
     Yearly              Net              Density          Land Area           Migrants           Fert.            Med.
 Min.   :-0.01350   Min.   : -383840   Min.   :   2.0   Min.   :     180   Min.   :-532687   Min.   :1.100   Min.   :15.00
 1st Qu.: 0.00435   1st Qu.:    4354   1st Qu.:  33.0   1st Qu.:   24945   1st Qu.: -10024   1st Qu.:1.800   1st Qu.:22.00
 Median : 0.01090   Median :   62206   Median :  84.0   Median :  112760   Median :  -1000   Median :2.300   Median :30.00
 Mean   : 0.01233   Mean   :  416964   Mean   : 182.1   Mean   :  660265   Mean   :   3322   Mean   :2.731   Mean   :30.26
 3rd Qu.: 0.02040   3rd Qu.:  362964   3rd Qu.: 221.0   3rd Qu.:  504845   3rd Qu.:   8282   3rd Qu.:3.600   3rd Qu.:38.00
 Max.   : 0.03840   Max.   :13586631   Max.   :2239.0   Max.   :16376870   Max.   : 954806   Max.   :7.000   Max.   :48.00
     Urban            World
 Min.   :0.1300   Min.   :0.000000
 1st Qu.:0.4300   1st Qu.:0.000200
 Median :0.6000   Median :0.001100
 Mean   :0.5968   Mean   :0.005095
 3rd Qu.:0.7850   3rd Qu.:0.003650
 Max.   :1.0000   Max.   :0.184700
```

Box Plots to understand the outliers present in the attributed that majorly influenceLife Expectancy.

```
par(mfrow = c(1, 4))
boxplot(Population_Story$Density, main='Density',col='Sky Blue')
boxplot(Population_Story$Population, main='Population',col='Sky Blue')
boxplot(Population_Story$Fert., main='Fertility',col='Sky Blue')
```

boxplot(Population_Story$Med., main='Median Age',col='Sky Blue')



Population and Population Density has large outliers. We design the model keeping this in mind.

Change the column names as follows

View(train)
colnames(train)[2]<-"Life_Expectancy_Both_Sexes"
colnames(test)[2]<-"Life_Expectancy_Both_Sexes"
colnames(train)[3]<-"Female_Life_Expectancy"
colnames(test)[3]<-"Female_Life_Expectancy"
colnames(train)[4]<-"Male_Life_Expectancy"
colnames(test)[4]<-"Male_Life_Expectancy"

Iteration 1

Population_Multiple<-
lm(Life_Expectancy_Both_Sexes~Female_Life_Expectancy+Male_Life_Expectancy, data = train)

summary(Population_Multiple)

```
> summary(Population_Multiple)

Call:
lm(formula = Life_Expectancy_Both_Sexes ~ Female_Life_Expectancy +
    Male_Life_Expectancy, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30729 -0.03393  0.00229  0.03432  0.17393

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.164315   0.076297  -2.154   0.0338 *
Female_Life_Expectancy  0.512911   0.004048 126.705   <2e-16 ***
Male_Life_Expectancy    0.488570   0.004207 116.138   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07444 on 94 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 4.763e+05 on 2 and 94 DF,  p-value: < 2.2e-16
```

F stat value is 4.76e+05 and Multiple R-Squared = Adjusted R-Squared = 0.99999. Which is ideal. Further studying the data, it looks like there is multicollinearity between the attributes used to build this model.

Hence the above prediction is not true.

So we go for iteration 2

Iteration 2

Model:

Population_Multiple2<-lm(Life_Expectancy_Both_Sexes~Population-Migrants+Fert.+Density,data=train)

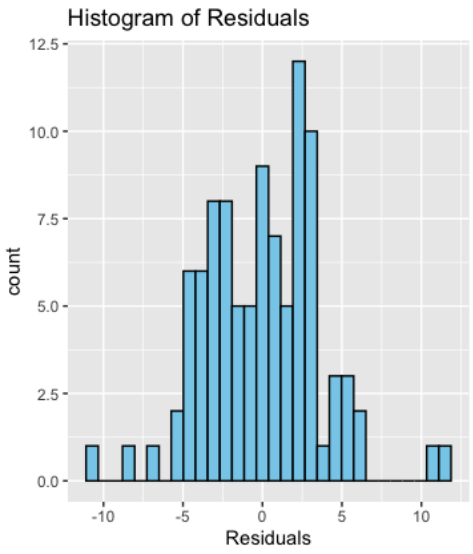
summary(Population_Multiple2)

```
> summary(Population_Multiple2)

Call:
lm(formula = Life_Expectancy_Both_Sexes ~ Population - Migrants +
    Fert. + Density, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9546  -2.7271  -0.0169   2.4845  11.2430

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.757e+01  9.831e-01  89.083   <2e-16 ***
Population  -2.246e-09  1.883e-09  -1.193    0.236
Fert.       -5.206e+00  3.115e-01 -16.716   <2e-16 ***
Density      7.783e-04  1.313e-03   0.593    0.555
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.721 on 93 degrees of freedom
Multiple R-squared:  0.7561,    Adjusted R-squared:  0.7482
F-statistic: 96.08 on 3 and 93 DF,  p-value: < 2.2e-16
```

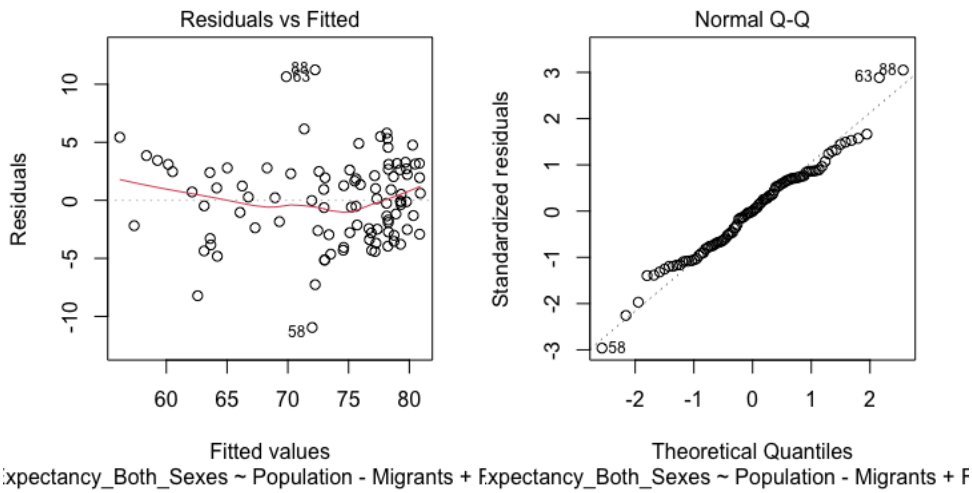
F- Statastic is at 96.08 which significantly high. Multiple R-squared= 0.756 and Adjusted R-squared= 0.7482. These squared values can be improved further. So, we go forward for next iteration.

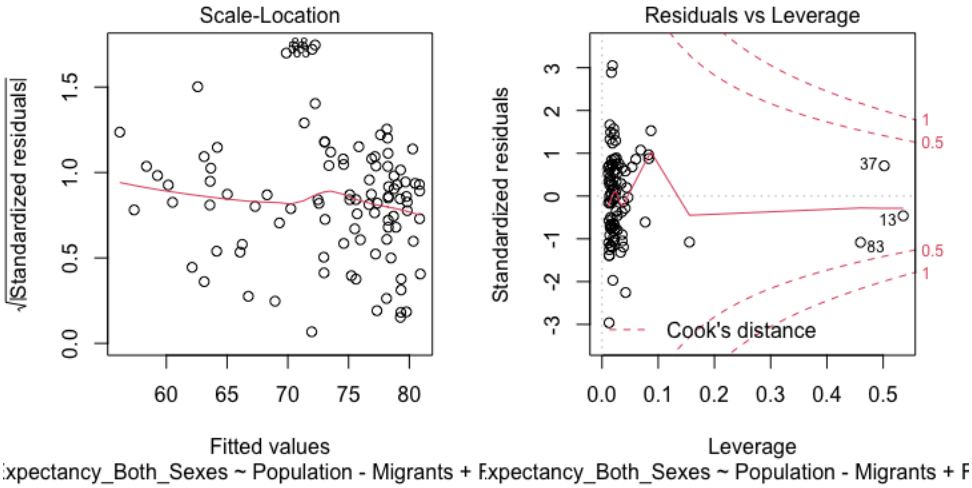
Studying Residuals.

Population2_residuals <- data.frame('Residuals' = Population_Multiple2$residuals)
ggplot(Population2_residuals, aes(x=Residuals)) + geom_histogram(color='black', fill='skyblue') + ggtitle('Histogram of Residuals')

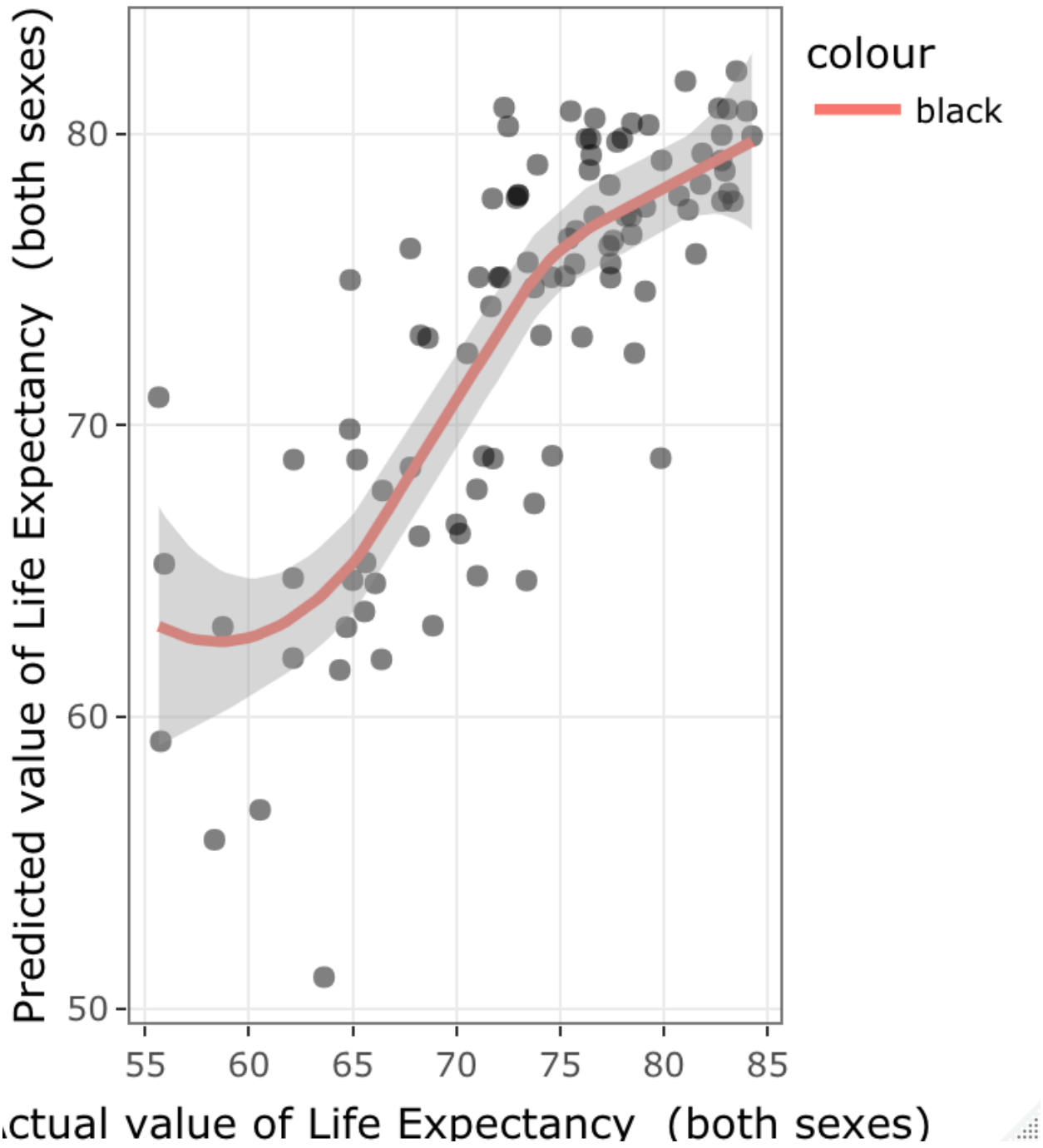


This histogram does not even slightly match the bell curve. This means we can design a better model than Iteration 2





Residual plots show that data points are all over the place and aligning with regression line. There is a need to build better data model. Hence we go for Iteration 3



Data points are farther away from the regression line implying data model needs to be improved. There is a need for another iteration.

Finding error values

```
> accuracy(test$Life_Expectancy_Both_Sexes, test$pred_pop2)
                  ME     RMSE     MAE        MPE     MAPE
Test set -0.07502393 4.521385 3.58418 -0.2836412 5.019649
>
```

Iteration 3:

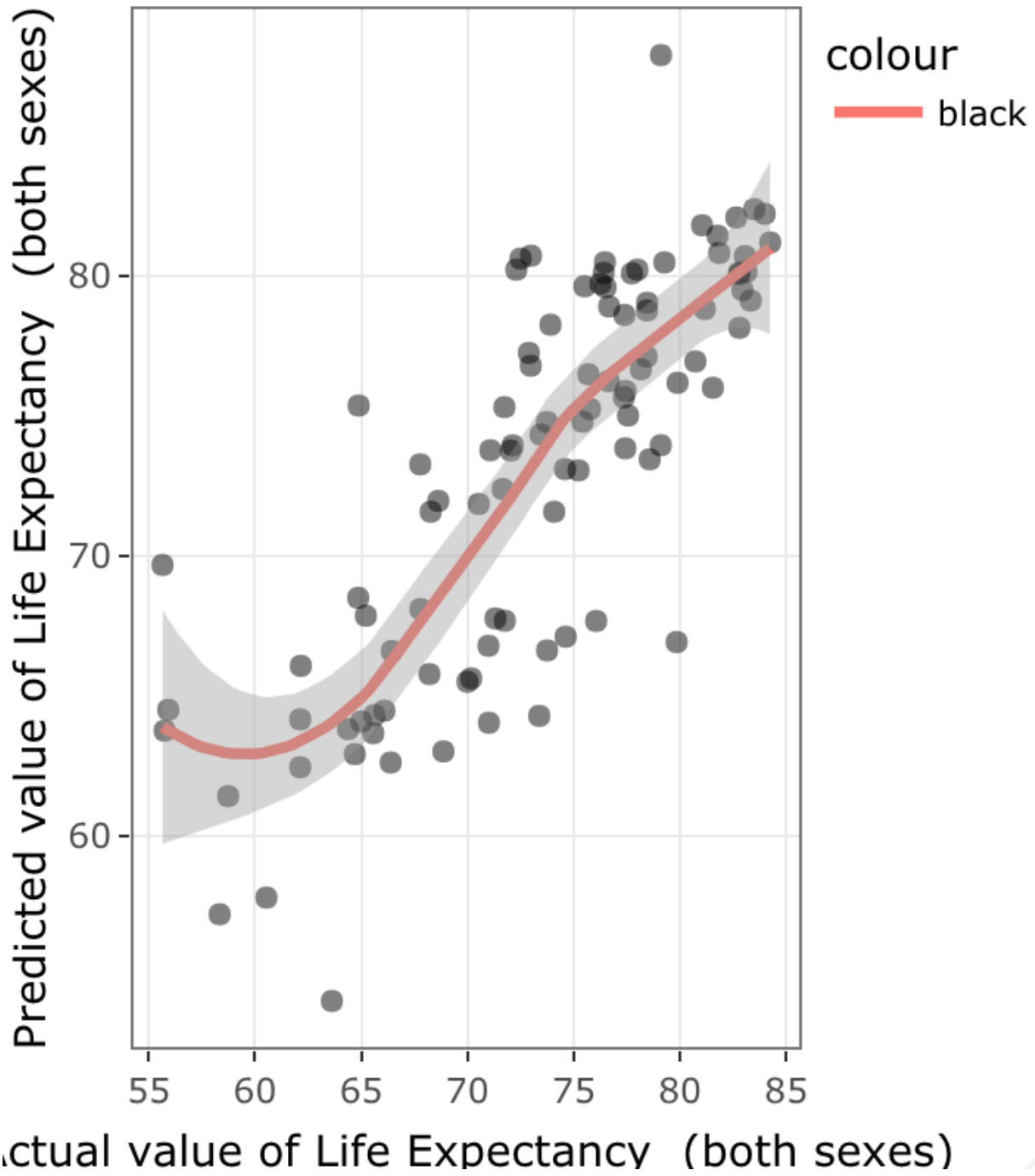
Population_Multiple3<-
lm(Life_Expectancy_Both_Sexes~Population+Migrants+Fert.+Med.+Population*Fert.,data=train)

summary(Population_Multiple3)

```
> summary(Population_Multiple3)

Call:
lm(formula = Life_Expectancy_Both_Sexes ~ Population + Migrants +
    Fert. + Med. + Population * Fert., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0363  -2.4553   0.4396   1.9954  11.3732

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.599e+01  4.282e+00  17.747  < 2e-16 ***
Population       -9.438e-09  1.129e-08  -0.836  0.40551
Migrants          9.743e-06  4.890e-06   1.992  0.04919 *
Fert.            -3.745e+00  6.490e-01  -5.771 9.86e-08 ***
Med.              2.401e-01  8.845e-02   2.715  0.00787 **
Population:Fert.  5.587e-09  5.733e-09   0.974  0.33229
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.711 on 95 degrees of freedom
Multiple R-squared:  0.7697,    Adjusted R-squared:  0.7575
F-statistic: 63.48 on 5 and 95 DF,  p-value: < 2.2e-16

>
```

Multiple R-squared = 0.7697 and Adjusted R Squared = 0.7575. p is also significantly low lesser than 0.05 showing auto correlation. We observe significant drop in F-statastic value from 96.08 in iteration 2 to 63.48 in iteration 3 implying ieration 3 has got a better model over iteration 2.

```
test$pred_pop3=predict(Population_Multiple3,newdata = test)

library(plotly)
pl1 <-test %>%
  ggplot(aes(Life_Expectancy_Both_Sexes,pred_pop3)) +
  geom_point(alpha=0.5) +
  stat_smooth(aes(colour='black')) +
  xlab('Actual value of Life Expectancy  (both sexes)') +
  ylab('Predicted value of Life Expectancy  (both sexes)') +
  theme_bw()
ggplotly(pl1)
```



The regression line is raising up to be a straight line and the data points got closer in comparison to the data points in the iteration 2. Regression line rose up to be nearly straight line and has better linearity over regression line in iteration 2 reiterating the fact that Iteration 3 has a better data model over iteration 2.

Understanding Errors

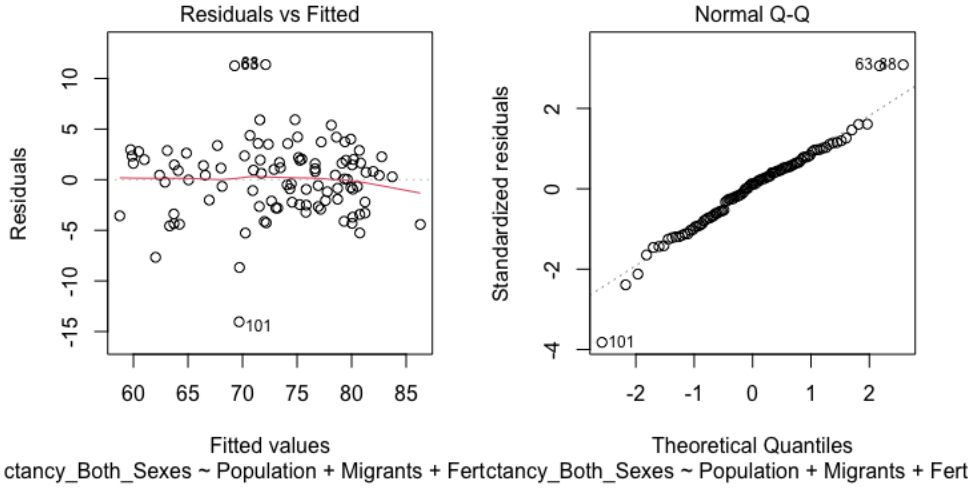
accuracy(test$Life_Expectancy_Both_Sexes, test$pred_pop3)

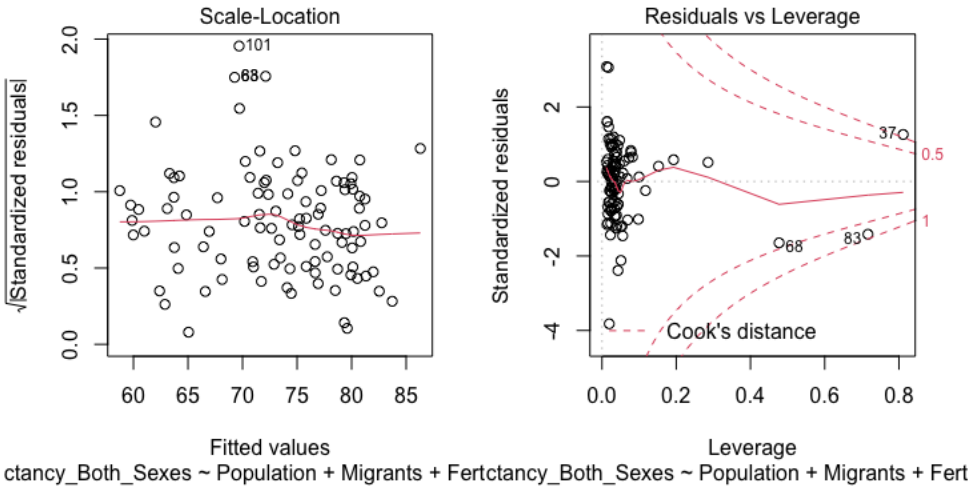
```
> accuracy(test$Life_Expectancy_Both_Sexes, test$pred_pop3)
                    ME      RMSE       MAE        MPE      MAPE
Test set -0.2480304 4.475722 3.475372 -0.527476 4.872586
>
```

Errors have dropped in iteration 3 in comparison to iteration 2.

We will understand the residual plots.

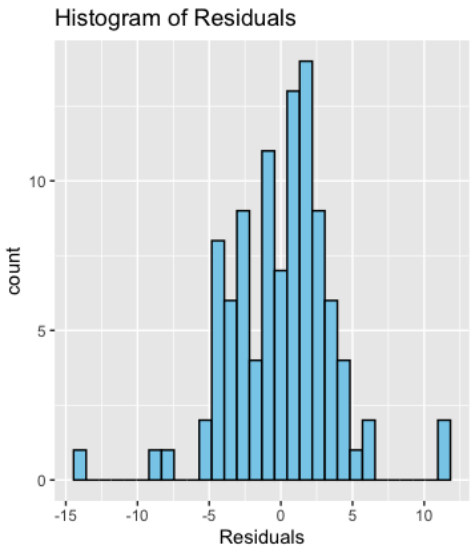
Residuals study





We observe that data points are in a way aligning to the regression line in the residual plot a little better than Iteration 2.

```
Population3_residuals <- data.frame('Residuals' = Population_Multiple3$residuals)
ggplot(Population3_residuals, aes(x=Residuals)) + geom_histogram(color='black',
fill='skyblue') + ggtitle('Histogram of Residuals')
```



Though the graph is slightly skewed towards right, it is closer towards normal distribution which means Iteration 3 data model is a perfect fit.

Conclusion: Iteration 3 is a better regression model compared to iteration 2. This could be so because of the introduction of Median age and emphasis on population and its fertility into data modeling of iteration 3. The higher the fertility rate implies higher overall life expectancy.

2. Which other variables could you include in the model to improve predictions of life expectancy?

Life Expectancy also depends on if the population is near an industrial area. The more the proximity of a factory, the higher the fatalty rate which in turn implies lower life expectancy. Radial distance from nearest industry is the required variable.

Access to quality health centers like affordable Multi Specialty Hospitals also plays major role in determining the life expectancy numbers. The greater the number of good hospitals at affordable rates in a country the higher would be the life expectancy. The number of affordable hospitals in a country is the required variable.

An account of tobacco smokers or drug influenced population strongly impacts life expectancy numbers. The higher the narcotic influence the lower the Life Expectancy. Number of narcotics influenced population country wise needs to be given.

Literacy rate is another variable which is determines life expectancy. The more literate the person is the better chances of him being aware of the best practices to stay away from disease causing microorganism. Hence higher literacy rate implies Literacy rate.

Median Value of Own houses in Boston area

- What is the best model to predict the median value of the houses in the Boston area?

To design the best model to predict the median value of the houses in the Boston area, we first need to prepare the data set. We do that by splitting the dataset into testing and training data.

########Preparing Boston Dataset######

```
train<-Boston[1:253,]
test<-Boston[254:506,]

install.packages("corrplot")
library(corrplot)
str(Boston)
```

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

Chas has a type factor. In order to plot correlation graph all the attributes should be numeric. We convert char to numeric

Summary(Boston)

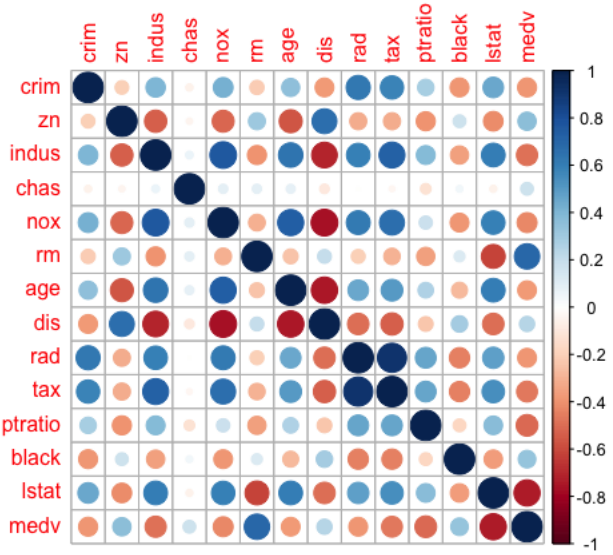
```
> summary(Boston)
      crim                zn              indus            chas              nox              rm            age              dis        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :1.000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:1.000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :1.000   Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :1.069   Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:1.000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :2.000   Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
      rad              tax             ptratio          black            lstat            medv      
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

Boston$chas<-as.numeric(Boston$chas)

We plot correlation plot to understand correlation between medv and other variables. Variabkes with better correltion with medv are chosen to perform regression.

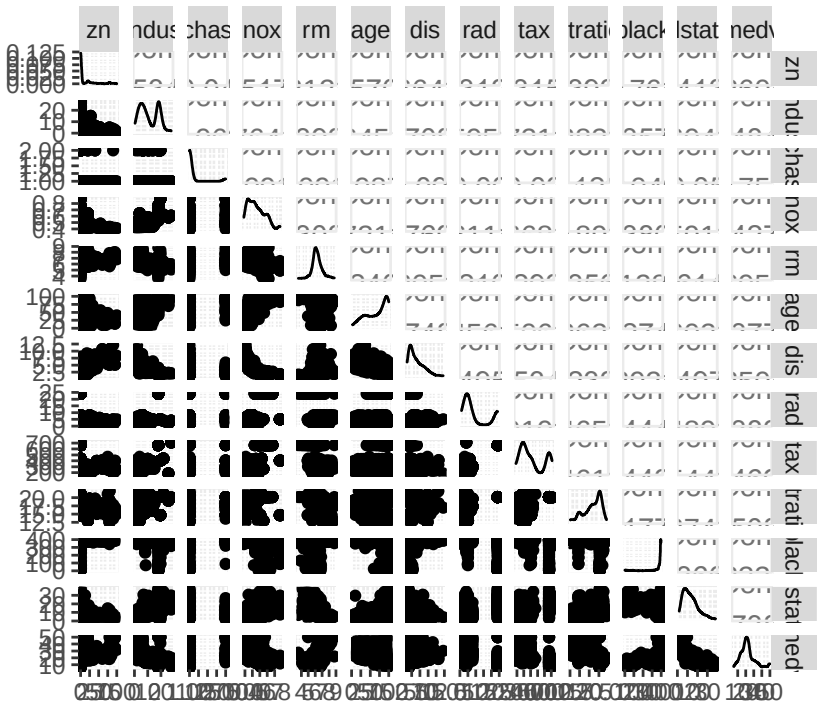
Correlation Plot

corrplot(cor(Boston))



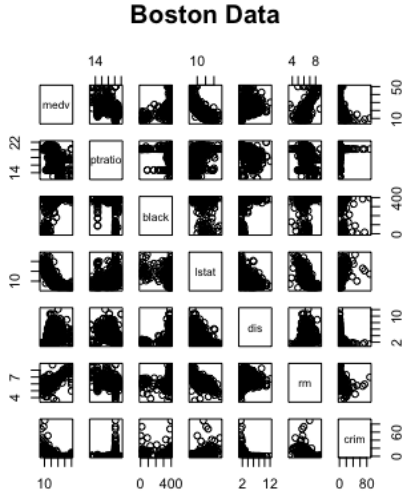
Correlation Plot 2

ggpairs(Boston[,-1])



Correlation Plot 3:

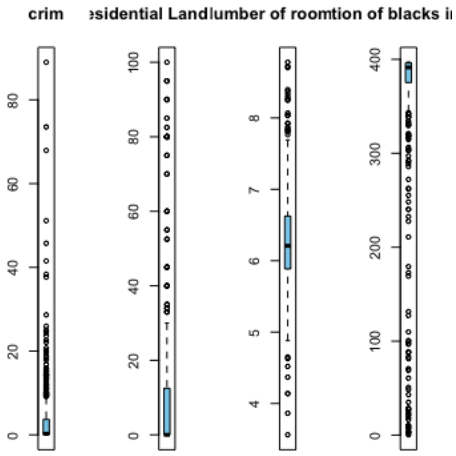
We draw multiple correlation plots for better understanding of relationship between attributes and aesthetics



This is a pairs plot depicting correlation between one attribute to every other attribute. Rm and lstat have better correlation with medv whereas crim black donot have great correlation with medv. We build our model accordingly by removing insignificant variables.

We do box plotting to better understand the outliers in different attributes. So that it will be easier for us to model data by removing insignificant variables.

```
/*Box Plot*/
par(mfrow = c(1, 4))
boxplot(Boston$crim, main='crim',col='Sky Blue')
boxplot(Boston$zn, main='Residential Land zones',col='Sky Blue')
boxplot(Boston$rm, main='Average Number of rooms per dwelling',col='Sky Blue')
boxplot(Boston$black, main='Proportion of blacks in town Age',col='Sky Blue')
```



We can clearly say crim, zn, rm and black have lot of outliers and must be avoided by modelling data.

Iteration 1

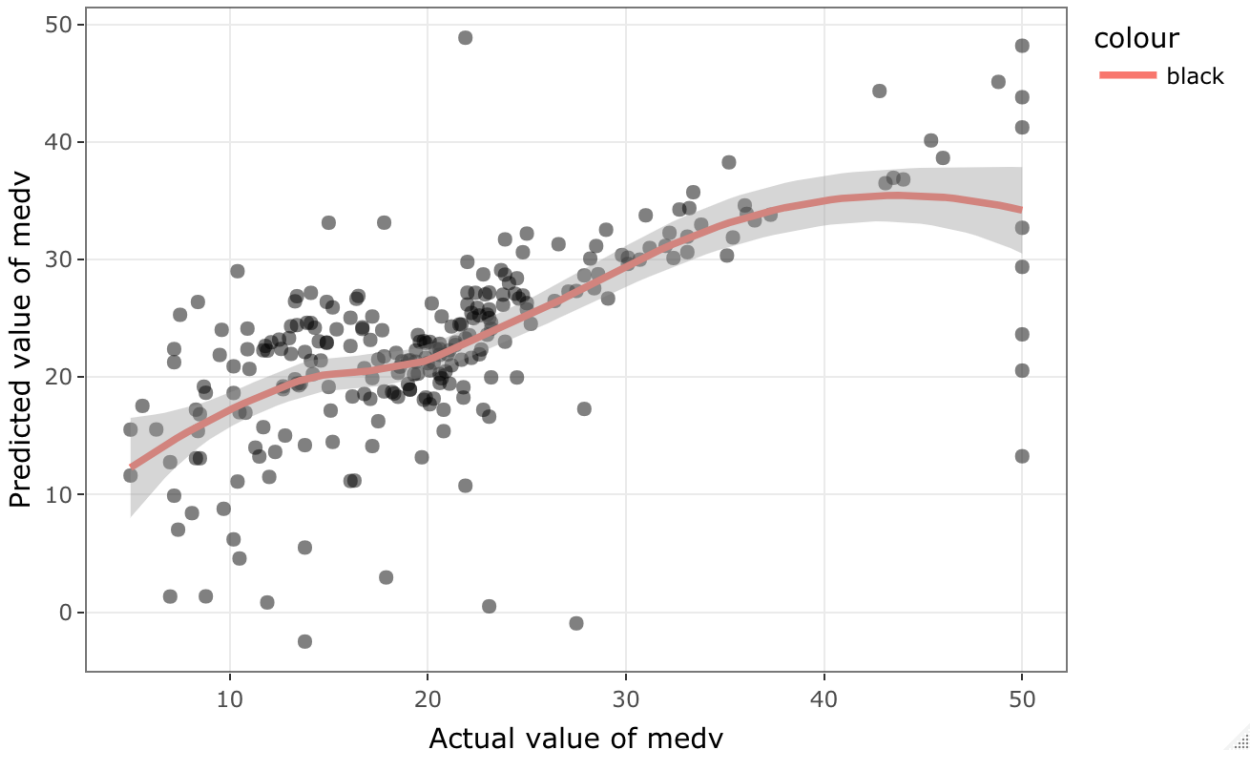
Let us use Multiple Regression to model this data.
Rm and lstat appear to have better correlation with medv. Hence we perform multiple regression on them.

```
/*Multiple Regression*/
Boston_Multiple<-lm(medv~lstat+rm,data=train)
summary(Boston_Multiple)
```

```
lm(formula = medv ~ lstat + rm, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4354 -2.5501 -0.3141  1.9053 13.2313

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -33.04114    3.78530  -8.729 3.73e-16 ***
lstat        -0.22727    0.05432  -4.184 3.97e-05 ***
rm            9.46575    0.52333  18.087  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 250 degrees of freedom
Multiple R-squared:  0.785,     Adjusted R-squared:  0.7832
F-statistic: 456.3 on 2 and 250 DF,  p-value: < 2.2e-16
```

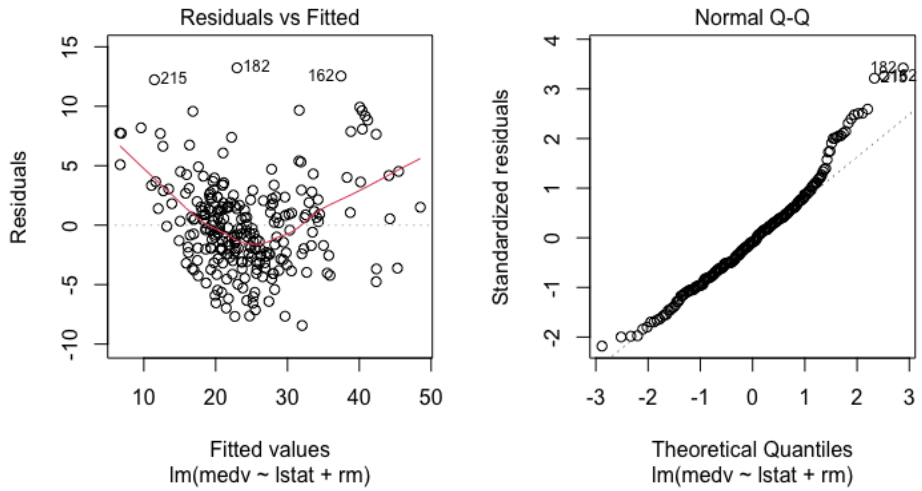
Maximum and Minimum values are wide apart marking high range which says Multiple regression between medv and lstat, rm is not a good model for Boston Data set. R-Squared value should be closer to 1 which is not the case here. F-static value is higher (ideally F-static should be as lower as possible).

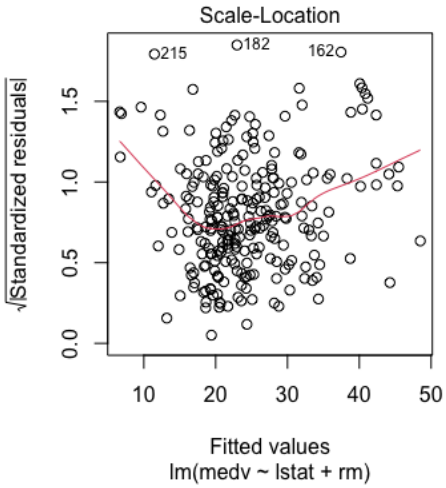
```
test$pred_1=predict(Boston_Multiple,newdata = test)

library(plotly)
pl1 <-test %>%
  ggplot(aes(medv,pred_1)) +
  geom_point(alpha=0.5) +
  stat_smooth(aes(colour='black')) +
  xlab('Actual value of medv') +
  ylab('Predicted value of medv') +
  theme_bw()
ggplotly(pl1)
```

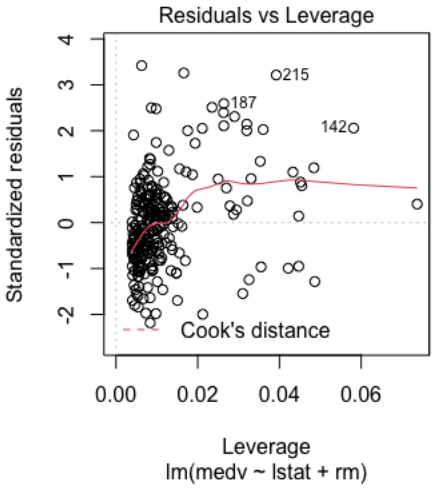


The data points are more scattered over the regression line and hence iteration 1 is not a great model to predict medv values in Boston data set

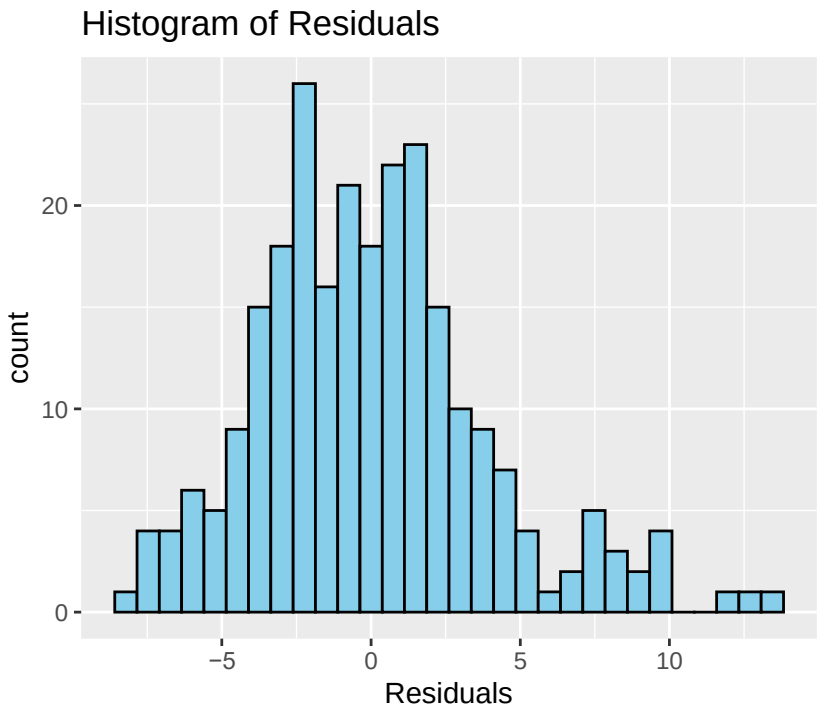
Understanding Residuals








```
Boston1_residual <- data.frame('Residuals' = Boston_Multiple$residuals)
ggplot(Boston1_residual, aes(x=Residuals)) + geom_histogram(color='black', fill='skyblue') +
ggtitle('Histogram of Residuals')
```



Ideally the residual histogram should be a bell curve. The above residual histogram does not even closely approximate to a bell curve. This data model needs improvement.

Understanding Errors

accuracy(test$medv,test$pred_1)

```
> View(test)
> accuracy(test$medv,test$pred_1)
                ME     RMSE      MAE       MPE     MAPE
Test set 2.078329 7.896188 5.526807 -5.618953 67.67076
>
```

Iteration 2

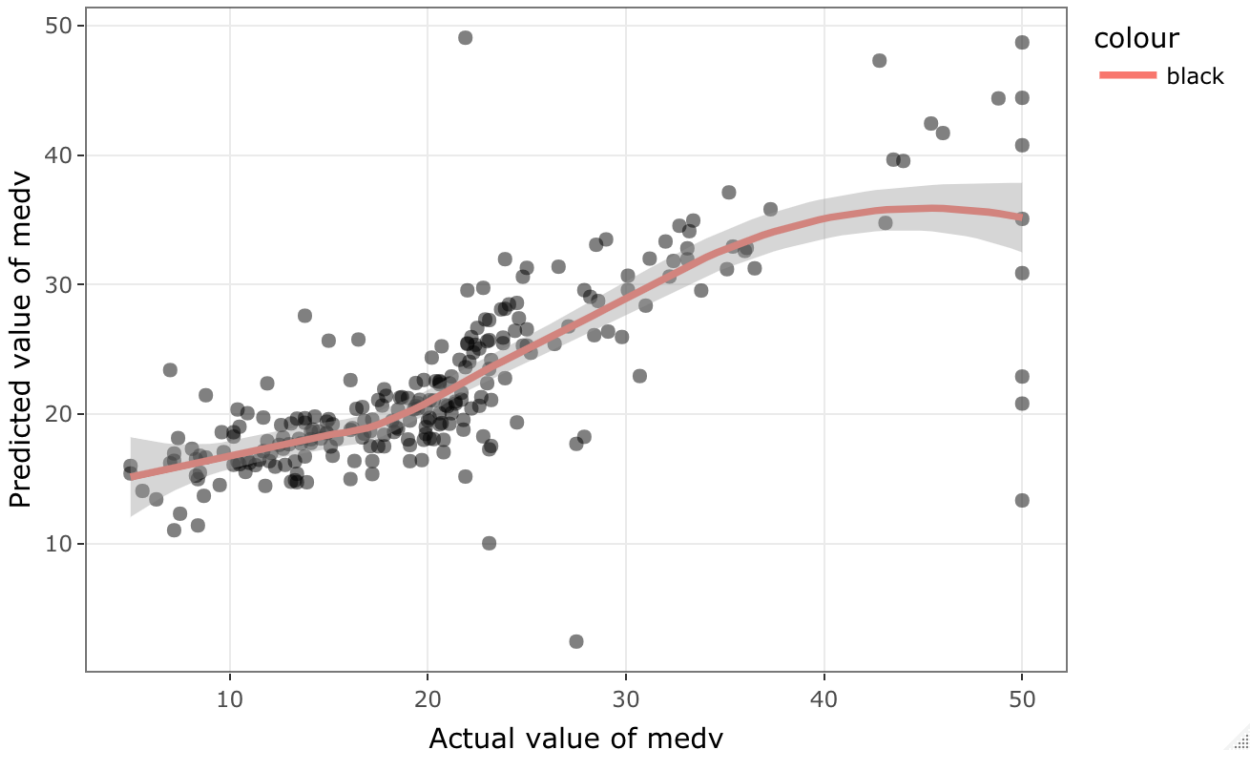
Boston_Multiple2 <- lm(medv~rm*lstat,data=train)
summary(Boston_Multiple2)

```
> summary(Boston_Multiple2)

Call:
lm(formula = medv ~ rm * lstat, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8014 -2.0150 -0.2286  1.5608 13.7390

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.80702    3.40339  -14.34   <2e-16 ***
rm           12.39044    0.50192   24.69   <2e-16 ***
lstat         3.08879    0.30038   10.28   <2e-16 ***
rm:lstat     -0.58642    0.05254  -11.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.176 on 249 degrees of freedom
Multiple R-squared:  0.8567,    Adjusted R-squared:  0.8549
F-statistic: 496.1 on 3 and 249 DF,  p-value: < 2.2e-16
```

R-Squared value is 0.8549 for Iteration 2 which is better than iteration 1 which is at 0.7832. Let us try to improve the prediction method and get a better R Squared value.

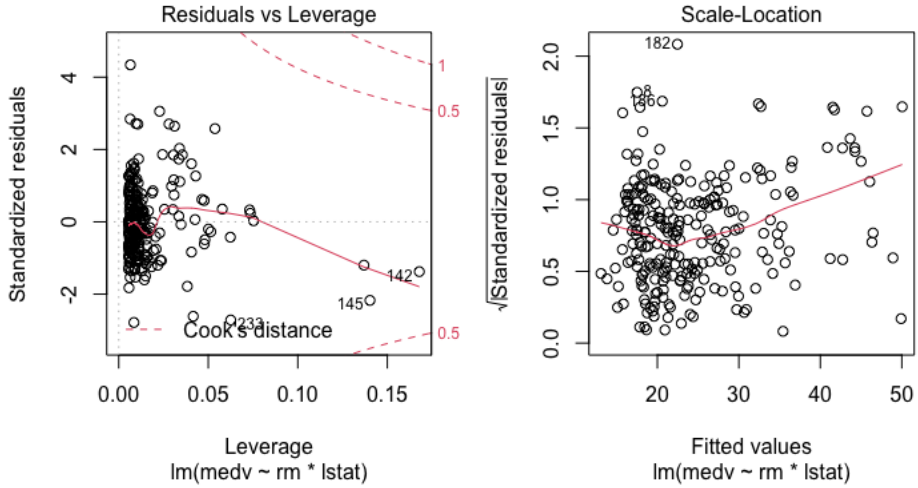
Plotting Regression Line

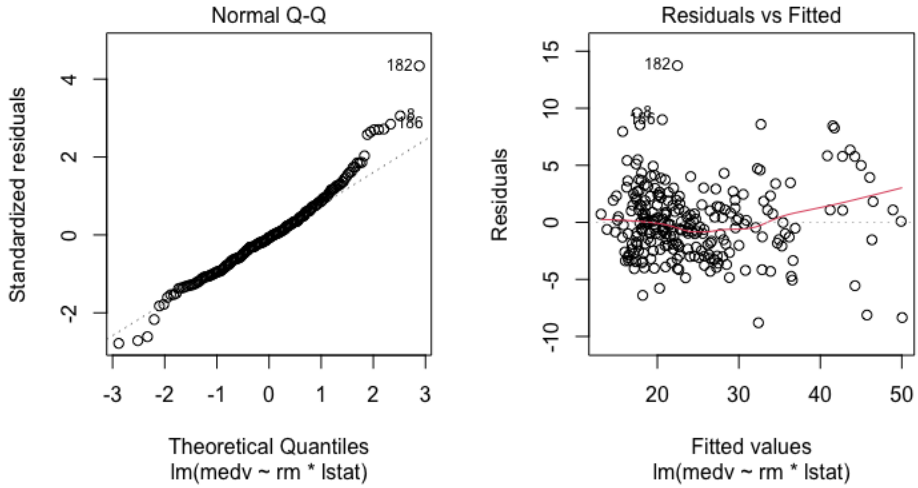
```
test$pred_2=predict(Boston_Multiple2,newdata = test)

library(plotly)
pl1 <-test %>%
  ggplot(aes(medv,pred_2)) +
  geom_point(alpha=0.5) +
  stat_smooth(aes(colour='black')) +
  xlab('Actual value of medv') +
  ylab('Predicted value of medv') +
  theme_bw()
ggplotly(pl1)
```

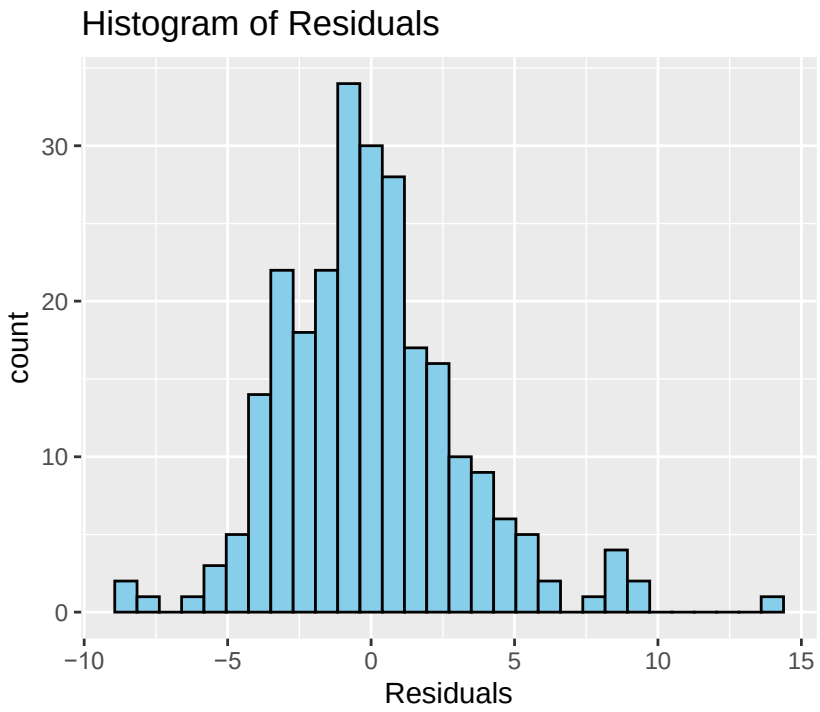


Iteration 2 has data points closer to the regression line. We observe that the regression line is a bit straighter than the iteration 1. We can conclude that Iteration 2 is a better model to predict medv values over iteration 1

Understanding Residuals


Residuals vs Leverage
Standardized residuals
Leverage
lm(medv ~ rm * lstat)
Cook's distance
1
0.5
0.5
142
145
233
Scale-Location
√|Standardized residuals|
Fitted values
lm(medv ~ rm * lstat)
182
8
186


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(medv ~ rm * lstat)
182
8
186
Residuals vs Fitted
Residuals
Fitted values
lm(medv ~ rm * lstat)
182
8
186



The above residual histogram is in a way inclined to be a bell curve. We can build a better model to make Residual histogram even closer to a bell curve. Hence we build iteration 3

Understanding errors

```
> accuracy(test$medv,test$pred_2)
                ME      RMSE      MAE      MPE      MAPE
Test set 1.565439 6.456826 4.38126 4.53777 25.25834
```

Iteration 3

Boston_Multiple3<-lm(medv~.-indus-age-zn+rm*lstat-black+lstat*rad, data=train)
summary(Boston_Multiple3)
plot(Boston_Multiple3)

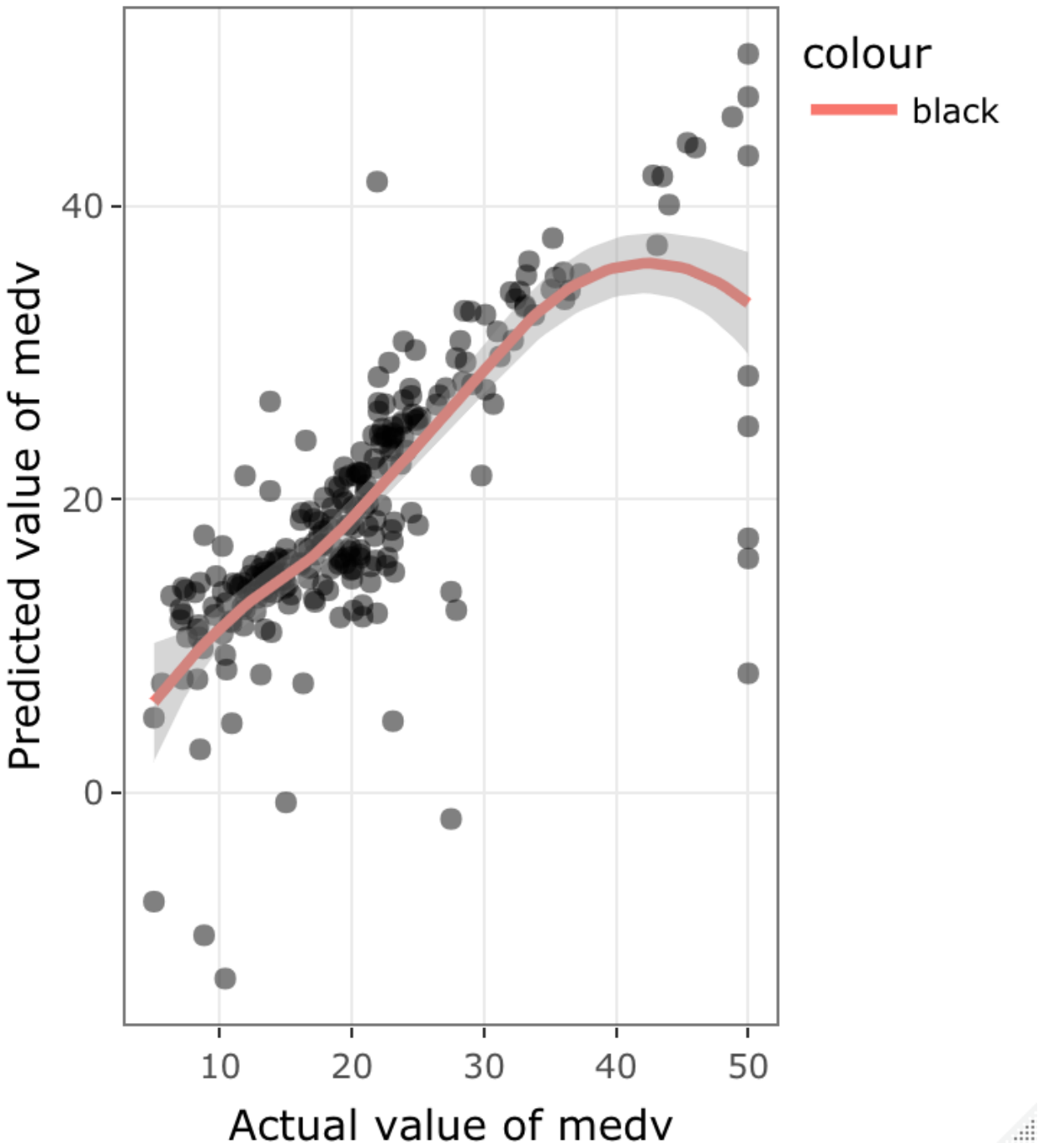
```
> summary(Boston_Multiple3)

Call:
lm(formula = medv ~ . - indus - age - zn + rm * lstat - black +
    lstat * rad, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5421 -1.6301 -0.1025  1.4826 10.7948

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.867516   5.004968  -4.169 4.26e-05 ***
crim         -0.366899   0.433419  -0.847  0.39810
chas1         0.762312   0.707837   1.077  0.28258
nox         -11.022928   3.640224  -3.028  0.00273 **
rm           11.498394   0.485221  23.697  < 2e-16 ***
dis          -0.730025   0.144496  -5.052 8.63e-07 ***
rad          -0.045591   0.222143  -0.205  0.83756
tax          -0.010481   0.003179  -3.297  0.00112 **
ptratio      -0.586712   0.102014  -5.751 2.67e-08 ***
lstat         2.863489   0.293980   9.740  < 2e-16 ***
rm:lstat     -0.549757   0.050537 -10.878  < 2e-16 ***
rad:lstat     0.013758   0.022527   0.611  0.54197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.83 on 241 degrees of freedom
Multiple R-squared:  0.8898,    Adjusted R-squared:  0.8848
F-statistic:   177 on 11 and 241 DF,  p-value: < 2.2e-16
```

R-Squared value for Iteration 3 is 0.8848 which is better than iteration 2 which is at 0.8549. We can say regression model used in Iteration 3 is better than 2. We always have scope of improvement and pull R Squared value closer to 1. We will stop at Iteration 3.

Let us plot a regression plot for iteration 3.

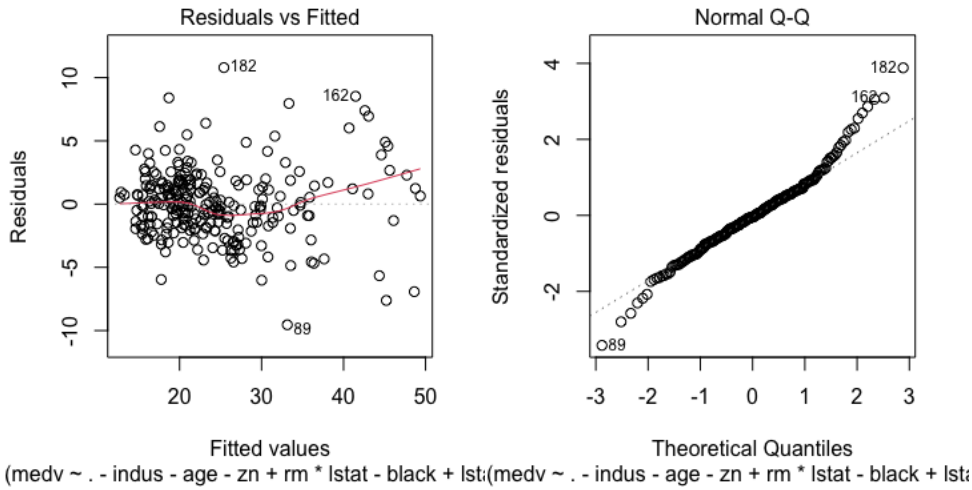
```
test$pred_3=predict(Boston_Multiple3,newdata = test)

library(plotly)
pl1 <-test %>%
 ggplot(aes(medv,pred_3)) +
 geom_point(alpha=0.5) +
 stat_smooth(aes(colour='black')) +
 xlab('Actual value of medv') +
 ylab('Predicted value of medv') +
 theme_bw()
```

ggplotly(pl1)



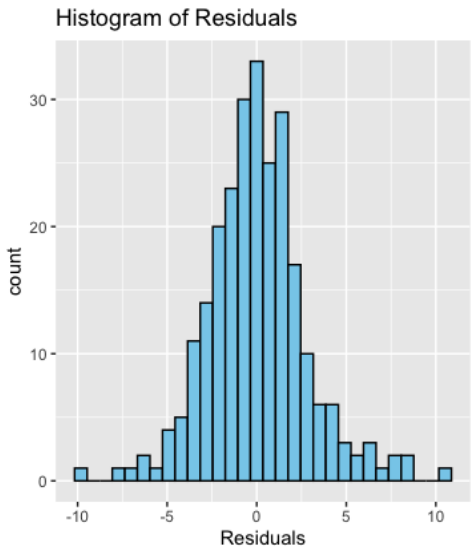
The data points are coinciding with the regression line implying iteration 3 is a better model to predict medv values. We can also observe that the regression line is straighter when compared with iteration 1 and 2. We can generate better medv prediction models by analysing the attributes

Studying Residuals



Data points in the residual charts of iteration 3 are closer to the regression line. Implying iteration 3 is a better model over iteration 1 and 2. Iteration 3 can better predict the values of medv.

Boston3_residuals <- data.frame('Residuals' = Boston_Multiple3$residuals)
ggplot(Boston3_residuals, aes(x=Residuals)) + geom_histogram(color='black', fill='skyblue') + ggtitle('Histogram of Residuals')



The above residuals for histogram is nearly a bell curve. Which implies Iteration 3 is a good model for predicting med v values of Boston.

Understanding Errors

```
> accuracy(test$medv, test$pred_3)
                ME      RMSE       MAE       MPE      MAPE
Test set -1.105095 6.689022 3.763305 9.124159 40.21799
>
```

Conclusion: Iteration 3 is a better model to predict medv values in Boston dataset.