1 Recitation Exercises

Chapter 2

Exercise-1

1. The sample size n is extremely large, and the number of predictors p is small?
   Answer:
   For large dataset, Performance of the model will be high. So the flexible model try to fit data and perform better while inflexible model leads overfitting of data in case of large data set.

2. The number of predictors p is extremely large, and the number of observations n is small?
   Answer:
   Here dataset is small so performance of the model will be low and overfit data. So inflexible model would perform compared to flexible model.
3. The relationship between the predictors and response is highly non-linear?
   Answer:
   Flexible models are good where relations between the predictors and response is non linear.so the flexible model perform better than inflexible model which may result in underfitting value due to non-linear relationship.
4. The variance of the error terms, i.e., σ2 = Var(ϵ), is extremely high?
   Answer:
   For flexible method, high value of variance will result in overfitting of data due to present of noise. So inflexible model performs better than flexible.

Exercise-2

Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

1. We are interested in understanding factors affecting CEO's Salary
   Answer:
   n=500 and p=3
   This is regression Problem as salary is continuous variable.
   It is inference as we are interested in understanding how salary affected by other independent variables.
2. it will be a success or a failure.
   Answer:
   n=20 and p=13
   It is classification and prediction Problem as we are interested in knowing whether it will be success or failure.
3. The % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets
   Answer:
   n=52 and p=3
   It is Regression Problem as % of change is continuous value
   It is prediction problem as well because we want to know % change in the USD/Euro.

Exercise-4

You will now think of some real-life applications for statistical learning.

a.

1. Classification model will be useful to classify whether student eligible to get admission or not based on parameters like previous course work, grade, language score, financial situation
   Response: Eligible/Not Eligible
   Predictors: previous course work, grade, language score, financial situation
   Prediction as we are interested in getting admission or not.

2. Classification model will be useful to classify whether a person should buy car or not based on parameters like age, requirement, salary, price, maintance,insurance
   Response: Yes/No
   Predictors: age, requirement, salary, price, quality maintenance, insurance
   Prediction as we are interested in should buy a car or not.
3. Classification model will be useful to classify whether technical event will be Useful or not.
   Response: Useful/Not
   Predictors: contents quality, topics to be cover, Place, speaker profile
   Prediction as we are interested in event will be useful or not.

b.

1. Result of any sports game

   Response: Team will win with what score
   Predictors: Player's profile, weather, practice, previous score records
   Inference

2. Weather prediction based on certain parameters

   Predictors: Temperature, Humidity, Pressure, Moisture

   Response: Percentage of Rainfall.

   prediction

4. Percentage of increment in salary
   Predictors: performance, task completed, achievements, behaviour, participation
   Response: percentage of hike
   Inference

c.

1. Clustering analysis used to identify viewers interest for any show based on similar behaviour. Based amount of time spend per days, total viewing episodes per week, unique show viewed per month.

2. Clustering analysis used to identify group of customers who use their health insurance in specific ways based on parameters like number of hospital visit, family size, average age of family members and can set premium according.

3. Identify peoples with same community based on their language, food, dressing style, tone of speech

Exercise-6

differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

Answer:

Parametric model - depends on statistical distribution of data and used fixed number of parameters to build the model.so model used to fit data known in advance.

Non-Parametric model- not depend on distribution of the data use flexible number of parameters to build the model. This model required large dataset to estimate function f.

Advantages: As due to different parameters in parametric model this model doesn't require a larger dataset like nonparametric model.

Disadvantage-if more flexible model caused inaccurate estimation of f. In general, More Flexible model cause wrong estimation of f or overfit the observation.

Exercise-7

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

a. Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0

Answer:

Index	x1	x2	x3	Y	Distance
1	0	3	0	red	3
2	2	0	0	red	2
3	0	1	3	red	3.16
4	0	1	2	green	2.23
5	-1	0	1	green	1.41
6	1	1	1	red	1.73

b. What is our prediction with K = 1? Why?

Answer:

For k=1 in above table we can see 5^{th} observation is near, so class of this observation and our prediction is Green.

c. What is our prediction with K = 3? Why?

Answer:

For k=3 in above table we can see 3rd observation is near, So class of this observation and our prediction is red.

d. If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?
For Higher values of k, Bayes decision boundary will almost linear. Here Bayes decision boundary in this problem is highly nonlinear which denote value of k to be small.

Chapter-3

Exercise-1

Answer:

The null hypotheses for TV shows that, in presence of radio ads and newspaper ads, TV ads have no effect on sales. In same way null hypotheses for radio states that, in presence of TV and newspaper ads, radio ads have no effect on sales. Similarly, the null hypothesis for newspaper shows that, in presence of TV and radio ads, newspaper ads have no effect on sales. Still because of the small p values of TV and radio, null hypotheses are rejected. While high p value of newspaper states that null hypotheses for newspaper holds true.

Exercise-3

a. Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

Answer:

Y=B0+B1X1+B2X2+B3X3+B4X4+B5X5

- ⇨ 50 + 20 * GPA + 0.07 * IQ + 35 * (Gender) + 0.01 *(GPA * IQ) - 10 * (GPA * Gender).
- ⇨ Gender=Male=0 and Female=1

Salary of men =Y= 50 + 20*(GPA) + 0.07*(IQ) + 0.01*(GPA*IQ)

Salary of women =Y= 85 + 10*(GPA) + 0.07*(IQ) + 35 + 0.01*(GPA*IQ)

From both equation we get GPA=3.5

So male earning more than Female if GPA is more than 3.5
Statement(iii) is correct.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

Answer:

As explained in the previous answer, we can not say that females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

Answer:

TRUE

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

Answer:

if the GPA is high, it asserts that men earn more than women. So (iv) is **FALSE**

b. Predict the salary of a female with IQ of 110 and a GPA of 4.0.

Answer:

Salary => 50 + 20GPA + 0.07IQ + 35 + 0.01(GPA * IQ) - 10GPA.
=> 50 + 20 * 4 + 0.07 * 110 + 35 + 0.01 * 4 * 110 - 10 * 4
=> 137.1

Unit is 1000’s dollar. Therefore, Salary is anticipated as 137100.

c. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Answer:

It is possible to have a plentiful of evidence for a small effect. Also, small coefficient does not imply that interaction effect is small. Therefor the above sentence is false.

Exercise-4

Collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression,

i.e. Y = β[0] +β[1]X +β[2]X^2 +β[3]X^3 + e.

a. Suppose that the true relationship between X and Y is linear, i.e. Y = β[0] + β[1]X + e. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Answer:
The relationship between x and y is linear, it can be assumed that least square line to be near to the linear regression. consequently, RSS for linear may be lower than cubic. if we use cubic regression then noise will be added. Which implies that RSS for cubic regression will be lower than for linear regression.

b. Answer (a) using a test rather than training RSS.

Answer:
It is assumed that the polynomial regression will be having a high-test RSS, because the Linear regression would have less error than the overfit from training. So, to provide any conclusion enough information not available.

c. Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Answer:
Polynomial regression has lower train RSS compared to linear fit because of great flexibility so more flexible model will more rapidly follow point and reduce tarin RSS.

d. Answer (c) using a test rather than training RSS.

Answer:
The information given is not enough to answer which test RSS would be lower.

2.1 Problem 1

Load the iris sample dataset into R using a dataframe (it is a built-in dataset). Create a boxplot of each of the 4 features and highlight the feature with the largest empirical IQR. Calculate the parametric standard deviation for each feature - do your results agree with the empirical values? Use the ggplot2 library from CRAN to create a coloured boxplot for each feature, with a box-whisker per flower species. Which flower type exhibits a significantly different Petal Length/Width once it is separated from the other classes?

Load Dataset into R

Library(datasets)

Iris=data.frame(iris)

Create boxplot of the features

Boxplot (iris ,

main =”Boxplot of features “ ,

xlab= “Attributes/Features” ,

ylab= “value”,

col=c(”Green”,”blue”),

border=” Orange”)

Calculate empirical interquartile Range (IQR)

IQR(iris$Sepal.Length)

IQR(iris$Sepal.Width)

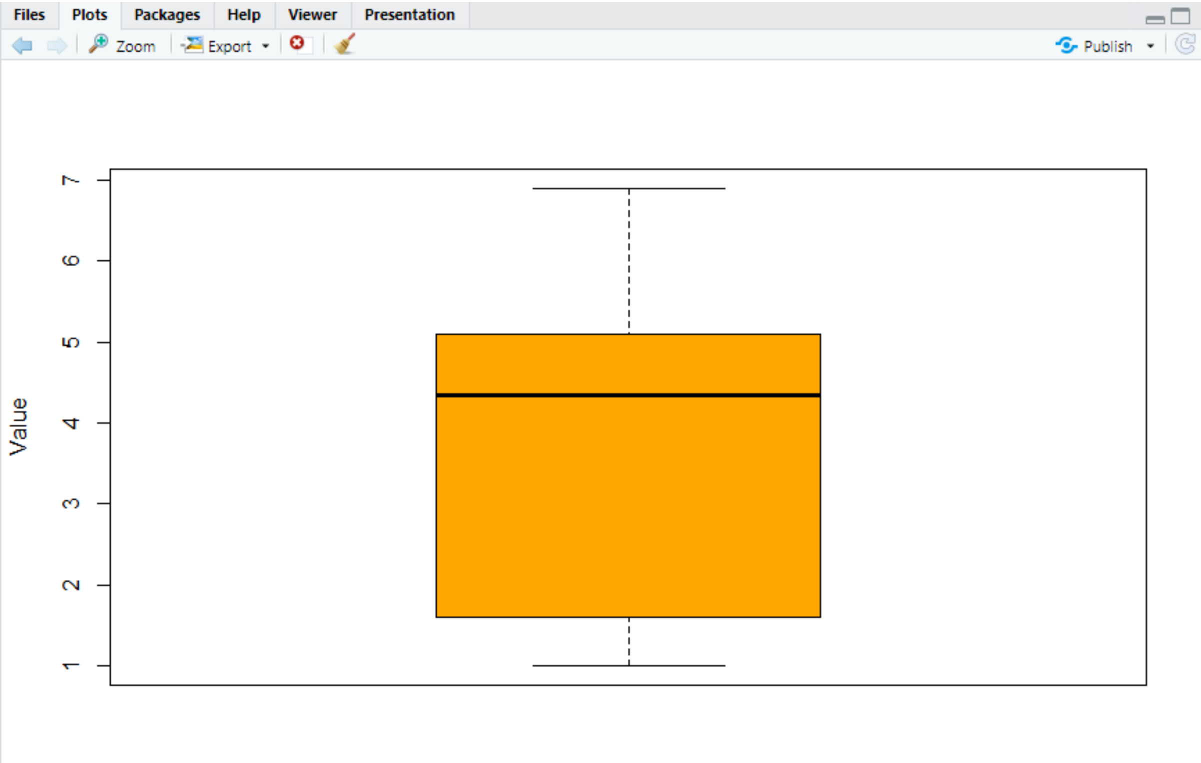
IQR(iris$Petal.Length)

IQR(iris$Petal.Width)

Petal Length has the highest IQR Value

Highlight Petal Length

boxplot(iris$Petal.Length, main="Maximum IQR 3.5", xlab="Petal Length",ylab ="Value", col="Orange")



Calculate the parametric standard deviation

SD(iris$Sepal.Length)

SD(iris$Sepal.Width)

SD(iris$Petal.Length)

SD(iris$Petal.Width)

Yes .Petal Length has the Maximum Interquartile range and Standard Deviation Value.

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function   Addins
Console  Terminal  Background Jobs
R 4.2.1 · ~/
> library(datasets)
> iris=data.frame(iris)
> boxplot(iris,main="Boxplot of features",xlab="Attributes/Features",ylab="values",col=c("Green","blue"),border="Orange")
> IQR(iris$Sepal.Length)
[1] 1.3
> IQR(iris$Sepal.Width)
[1] 0.5
> IQR(iris$Petal.Length)
[1] 3.5
> IQR(iris$Sepal.Width)
[1] 0.5
> boxplot(iris$Petal.Length,Main="Maximum IQR 3.5",xlab="Petal Length",ylab="Value",col="Orange")
> sd(iris$Sepal.Length)
[1] 0.8280661
> sd(iris$Sepal.Width)
[1] 0.4358663
> sd(iris$Petal.Length)
[1] 1.765298
> sd(iris$Petal.Width)
[1] 0.7622377
```

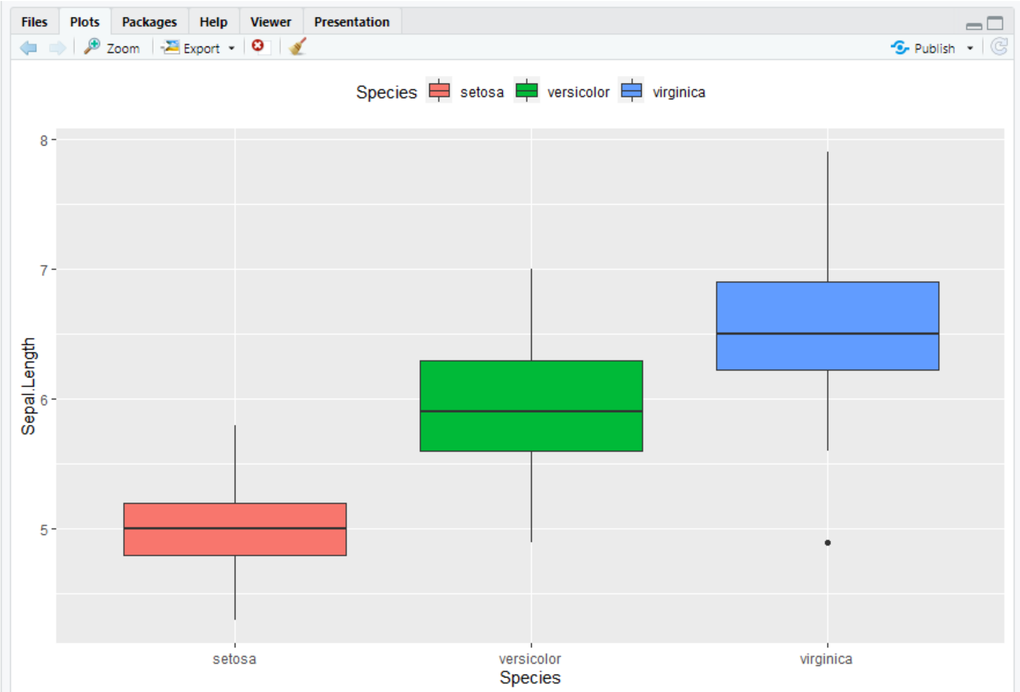
Use ggplot2 Library

Install. Packages('ggplot2')

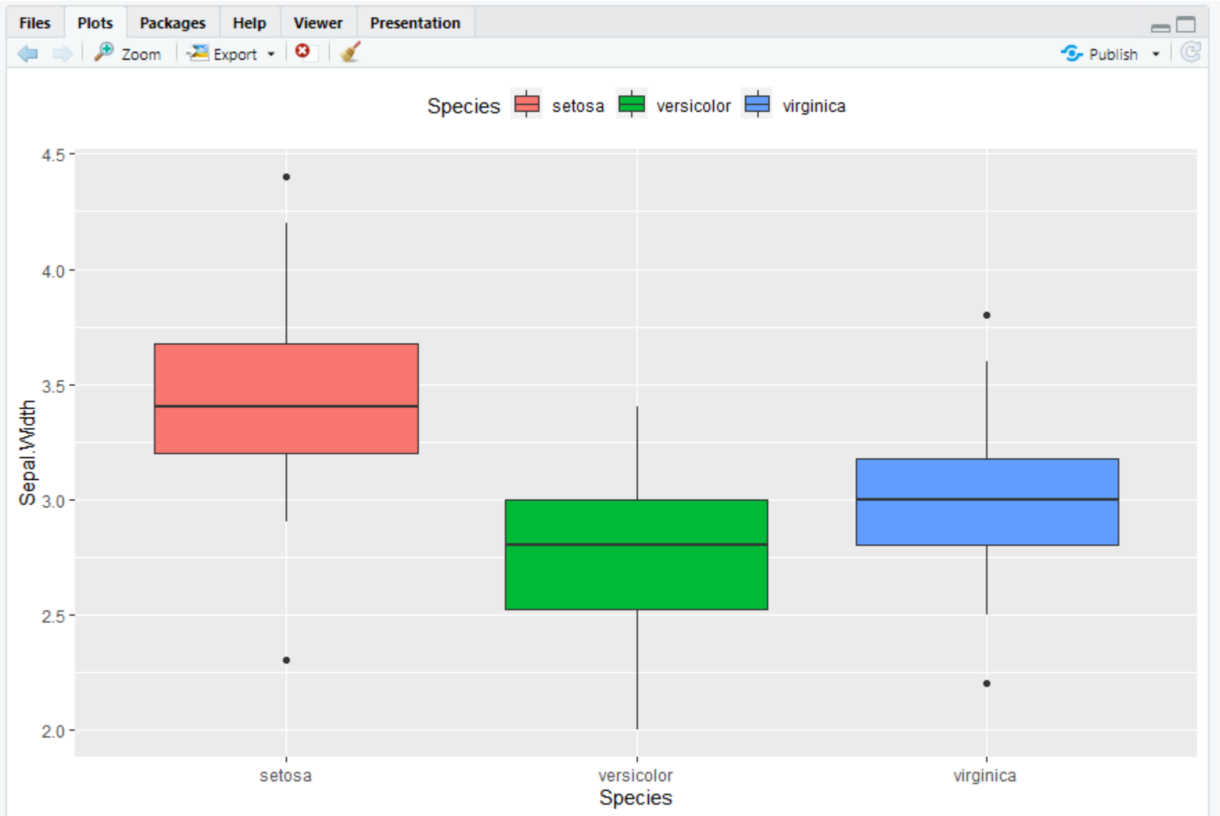
Library('ggplot2)

Create a coloured boxplot for each feature, with a box-whisker per flower species.

1) **Sepal Length**
   ggplot(data=iris, mapping=aes(x=Species,y=Sepal.Length,fill=Species))+geom_boxplot()+ theme(legend. Position = "top")

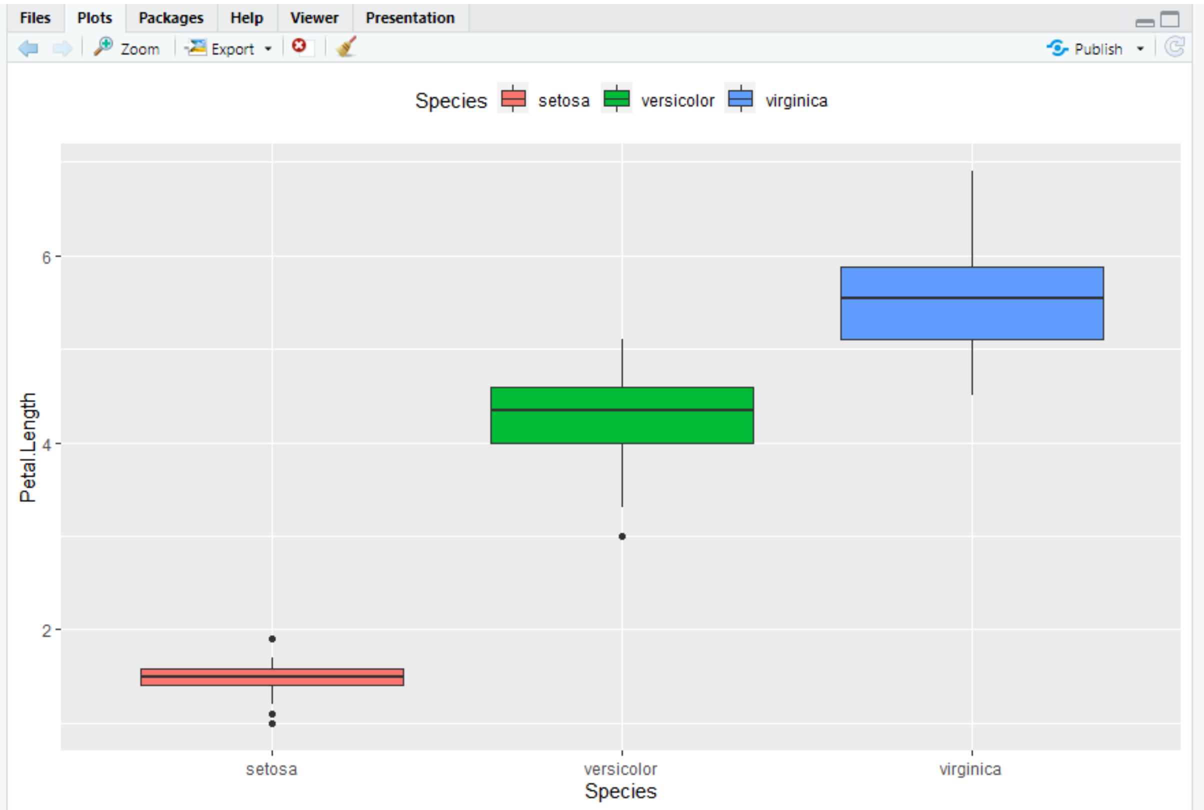


2) **Sepal Width**
   ggplot(data=iris, mapping=aes(x=Species,y=Sepal.Width,fill=Species))+geom_boxplot()+ theme(legend. Position = "top")



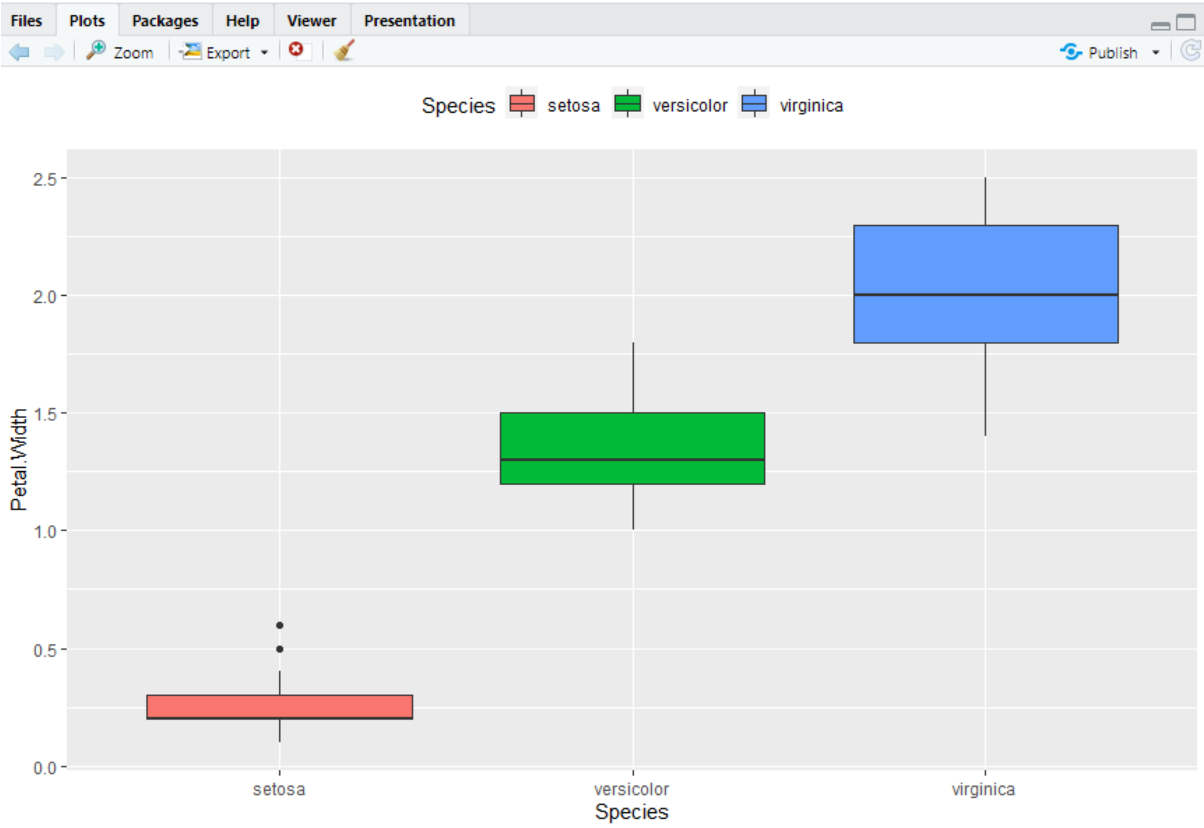
3) Petal Length

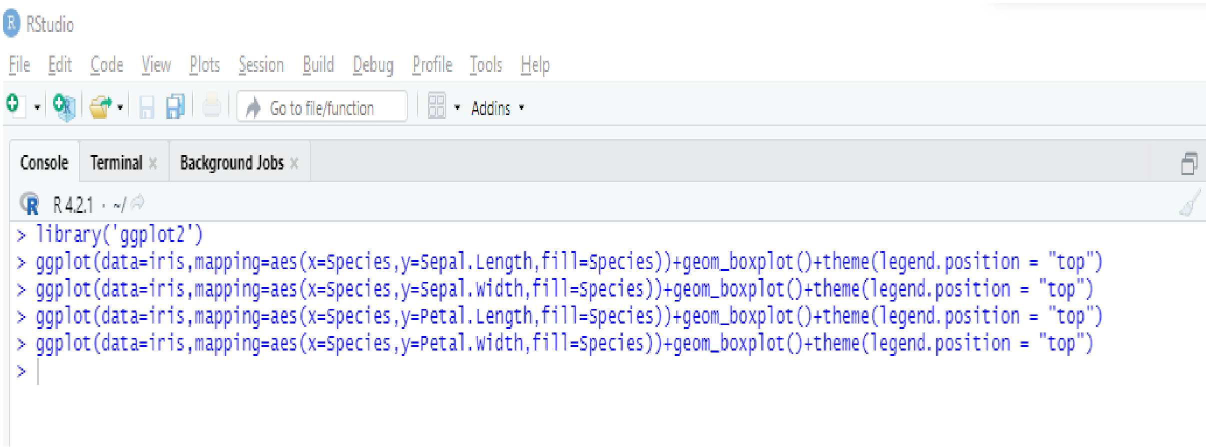
ggplot(data=iris,mapping=aes(x=Species,y=Petal.Length,fill=Species))+geom_boxplot()+theme(legend. Position = "top")



4) Petal Width

ggplot(data=iris,mapping=aes(x=Species,y=Petal.Width,fill=Species))+geom_boxplot()+theme(legend. Position = "top")





Setosa Species has the different values for Petal Length and Petal Width

2.2 Problem 2

Load the trees sample dataset into R using a dataframe (it is a built-in dataset) and produce a 5-number summary of each feature. Create a histogram of each variable - which variables appear to be normally distributed based on visual inspection? Do any variables exhibit positive or negative skewness? Install the moments library from CRAN use the skewness function to calculate the skewness of each variable. Do the values agree with the visual inspection?

Load the trees sample dataset into R

tree=data. frame(trees)

Summary of Each feature

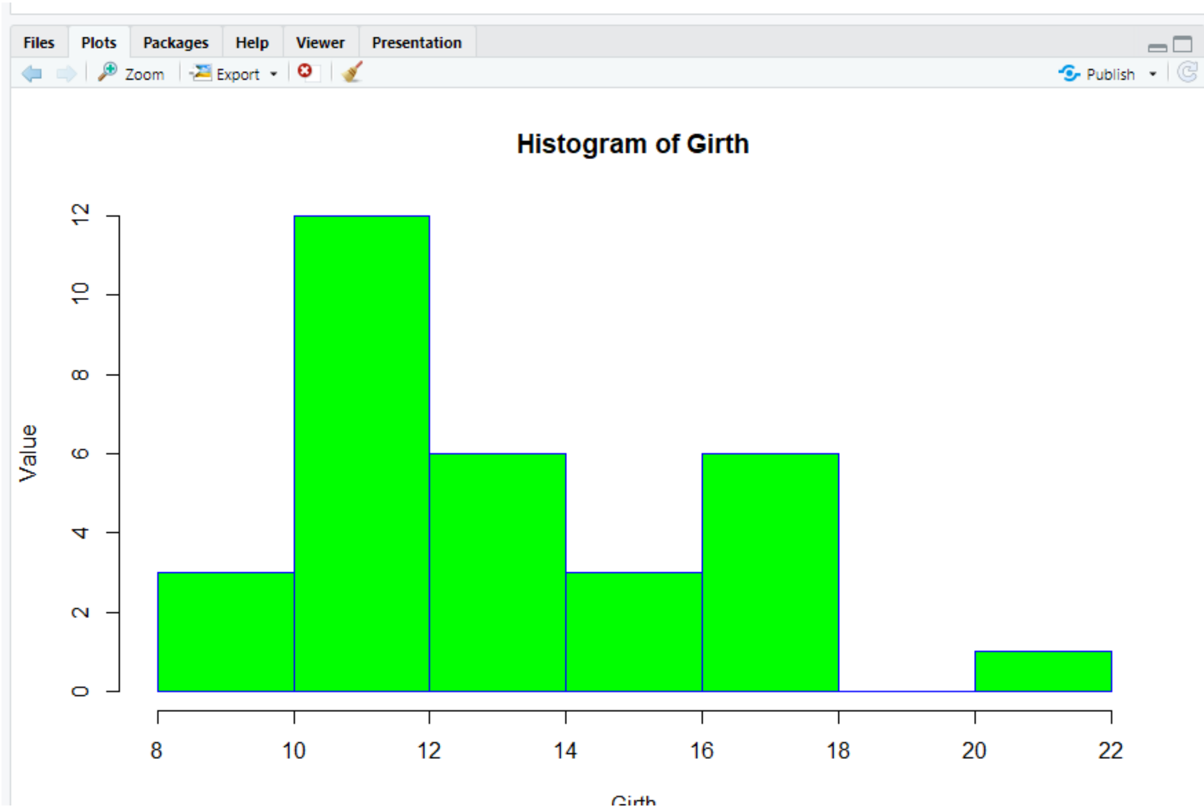
summary(trees)

```
> tree=data.frame(trees)
> summary(trees)
     Girth           Height       Volume
 Min.   : 8.30   Min.   :63   Min.   :10.20
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
 Median :12.90   Median :76   Median :24.20
 Mean   :13.25   Mean   :76   Mean   :30.17
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
 Max.   :20.60   Max.   :87   Max.   :77.00
>
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function   Addins
Console  Terminal  Background Jobs
R 4.2.1 · ~/
> tree=data.frame(trees)
> summary(trees)
     Girth           Height       Volume
 Min.   : 8.30   Min.   :63   Min.   :10.20
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
 Median :12.90   Median :76   Median :24.20
 Mean   :13.25   Mean   :76   Mean   :30.17
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
 Max.   :20.60   Max.   :87   Max.   :77.00
> #Alternative
> summary(trees$Girth)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.30   11.05   12.90   13.25   15.25   20.60
> summary(trees$Height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     63      72      76      76      80      87
> summary(trees$Volume)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.20   19.40   24.20   30.17   37.30   77.00
>
```

Create a histogram of each variable

1) .

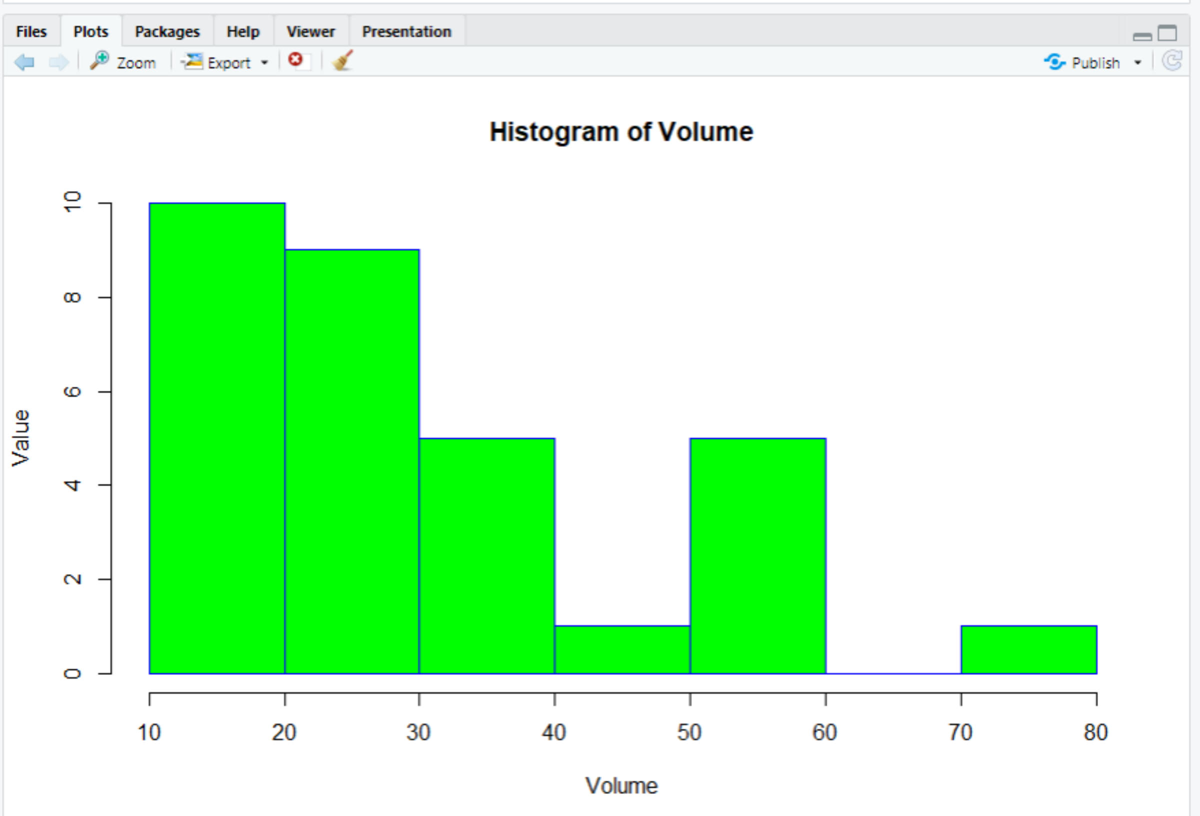


2) Girth

3) Height

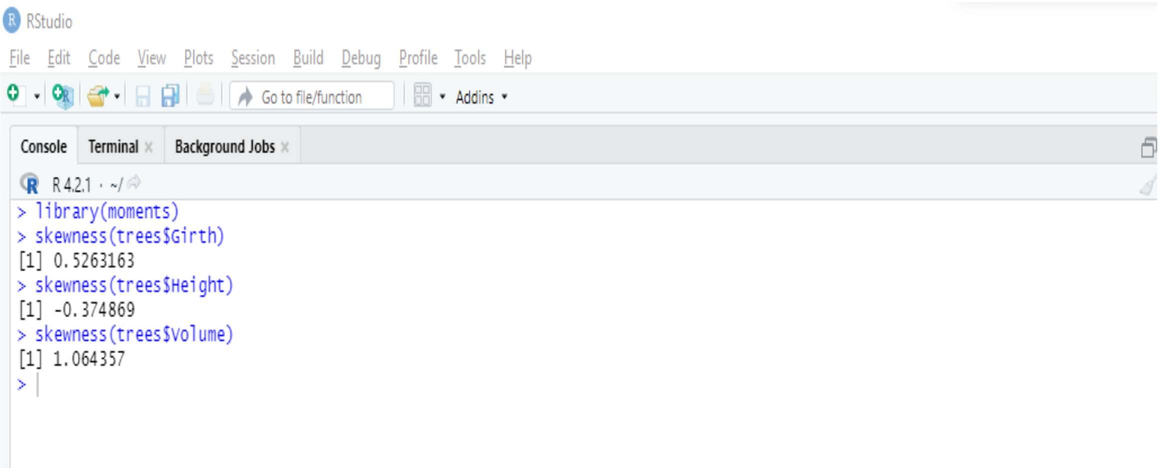

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Histogram of Height
Value
0
2
4
6
8
10
60
65
70
75
80
85
90
Height

4) Volume


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Histogram of Volume
Value
0
2
4
6
8
10
10
20
30
40
50
60
70
80
Volume

From Above Histogram we can Variable Height appears to be normally distributed.

Install Moments Library and Calculate Skewness for each Variable



As per the result displayed Girth and Volume has positive Skewness.

So from visual inspection we can conclude that Height has normal distribution and negative skewness.

2.3 Problem 3

Load the auto-mpg sample dataset from the UCI Machine Learning Repository (auto-mpg.data) into R using a dataframe (Hint: You will need to use read.csv with url, and set the appropriate values for header,as.is, and sep). The horsepower feature has a few missing values with a ? - and will be treated as a string. Use the as.numeric casting function to obtain the column as a numeric vector, and replace all NA values with the median. How does this affect the value obtained for the mean vs the original mean when the records were ignored?

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function   Addins
Console  Terminal  Background Jobs
R 4.2.1 · ~/
> Mpg_data=read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",sep="",header = F,string
sAsFactors = F)
> head(read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",sep="",header = F,stringsAsF
actors =F ))
  V1 V2  V3    V4   V5   V6 V7 V8                        V9
1 18  8 307 130.0 3504 12.0 70  1 chevrolet chevelle malibu
2 15  8 350 165.0 3693 11.5 70  1         buick skylark 320
3 18  8 318 150.0 3436 11.0 70  1        plymouth satellite
4 16  8 304 150.0 3433 12.0 70  1             amc rebel sst
5 17  8 302 140.0 3449 10.5 70  1               ford torino
6 15  8 429 198.0 4341 10.0 70  1          ford galaxie 500
> Mpg_Header=c("mpg","cylinder","displacement","horsepower","weight","acceleration","model_year","origin","car_name")
> colnames(Mpg_data)=Mpg_Header
> Mpg_data$horsepower=as.numeric(Mpg_data$horsepower)
Warning message:
NAs introduced by coercion
> head(Mpg_data)
  mpg cylinder displacement horsepower weight acceleration model_year origin                  car_name
1  18        8          307        130   3504         12.0         70      1 chevrolet chevelle malibu
2  15        8          350        165   3693         11.5         70      1         buick skylark 320
3  18        8          318        150   3436         11.0         70      1        plymouth satellite
4  16        8          304        150   3433         12.0         70      1             amc rebel sst
5  17        8          302        140   3449         10.5         70      1               ford torino
6  15        8          429        198   4341         10.0         70      1          ford galaxie 500
> Mpg_Mean_Before=mean(Mpg_data$horsepower,na.rm = T)
> print(Mpg_Mean_Before)
[1] 104.4694
> cat("\n blank values in horsepower field",sum(is.na(Mpg_data$horsepower)))

 blank values in horsepower field 6
> Mpg_median=median(Mpg_data$horsepower,na.rm = T)
> Mpg_data$horsepower[is.na(Mpg_data$horsepower)]=Mpg_median
> Mpg_Mean_After=mean(Mpg_data$horsepower,na.rm = T)
> print(Mpg_Mean_After)
[1] 104.304
> cat("Difference Before and After Replacement",Mpg_Mean_Before-Mpg_Mean_After)
Difference Before and After Replacement 0.1653677
> 
```

By replacing Na values by median mean value change from 104.4694 to 104.304

2.4 Problem 4

Load the Boston sample dataset into R using a dataframe (it is part of the MASS package). Use lm to fit a regression between medv and lstat - plot the resulting fit and show a plot of fitted values vs. residuals. Is there a possible non-linear relationship between the predictor and response? Use the predict function to calculate values response values for lstat of 5, 10, and 15 - obtain confidence intervals as well as prediction intervals for the results - are they the same? Why or why not? Modify the regression to include lstat2 (as well lstat itself) and compare the R2 between the linear and non-linear fit - use ggplot2 and stat smooth to plot the relationship.

Load the Boston sample dataset

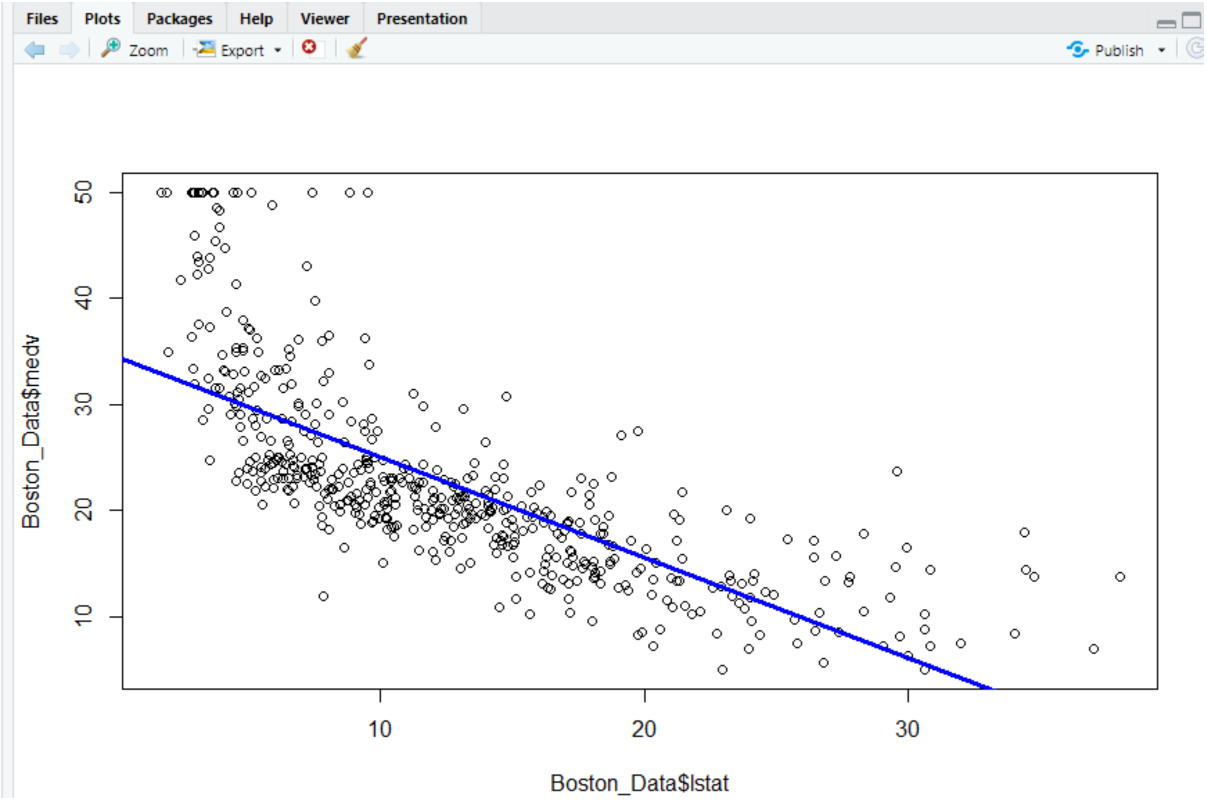
```
> library(MASS)
> Boston_Data=data.frame(Boston)
> head(Boston_Data)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
> summary(Boston_Data)
      crim                zn             indus            chas              nox               rm             age        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
      dis              rad              tax           ptratio          black            lstat            medv      
 Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 2.100   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 : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.: 5.188   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.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
> 
```

fit a regression Model

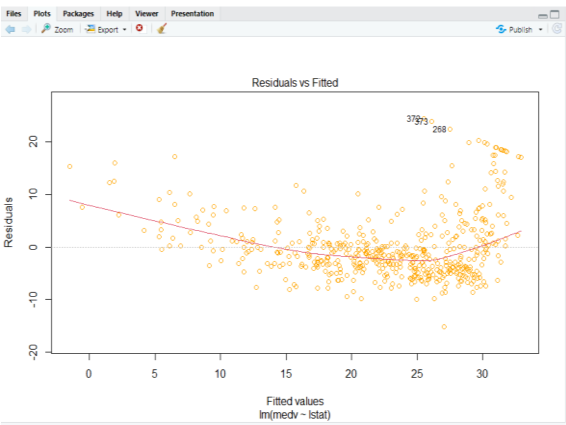
```
> Linear_Model=lm(medv~lstat ,data=Boston_Data)
> summary(Linear_Model)

Call:
lm(formula = medv ~ lstat, data = Boston_Data)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.168  -3.990  -1.318   2.034  24.500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,	Adjusted R-squared:  0.5432 
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16

> coef(Linear_Model)
(Intercept)       lstat 
 34.5538409  -0.9500494 
> confint(Linear_Model)
                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
> cat("r-Squared for Linear Model",summary(Linear_Model)$r.sq)
r-Squared for Linear Model 0.5441463
```

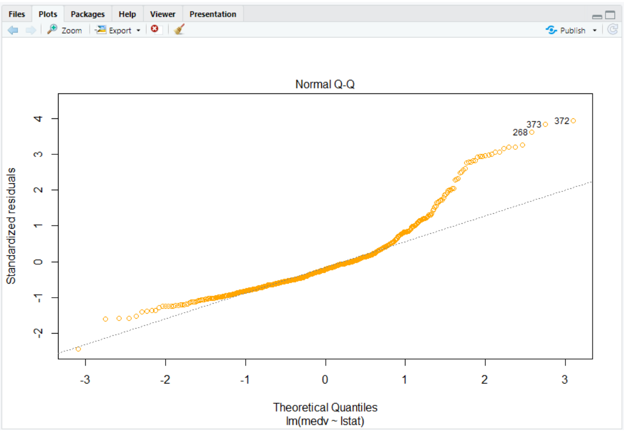
For linear Model R-square value R2 is 0.5441

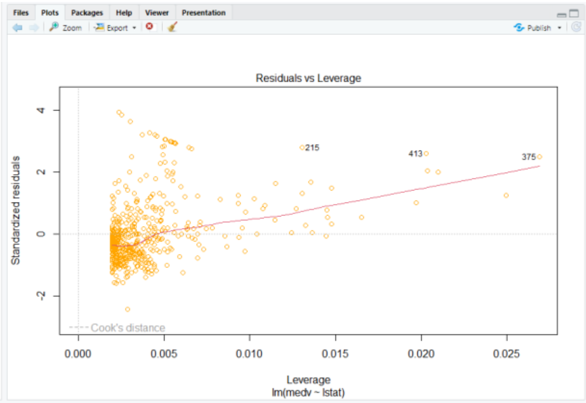
Visualization of Result

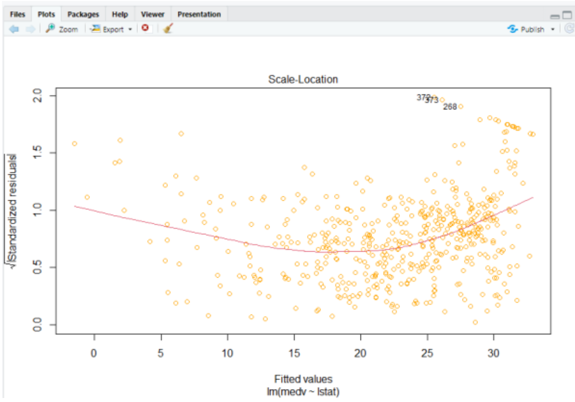



```
> plot(Boston_Data$lstat,Boston_Data$medv)
> abline(Linear_Model,lwd=3,col="blue")
> plot(Linear_Model,col="orange")
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
```









Plot for Non-Linear Fit:

RStudio

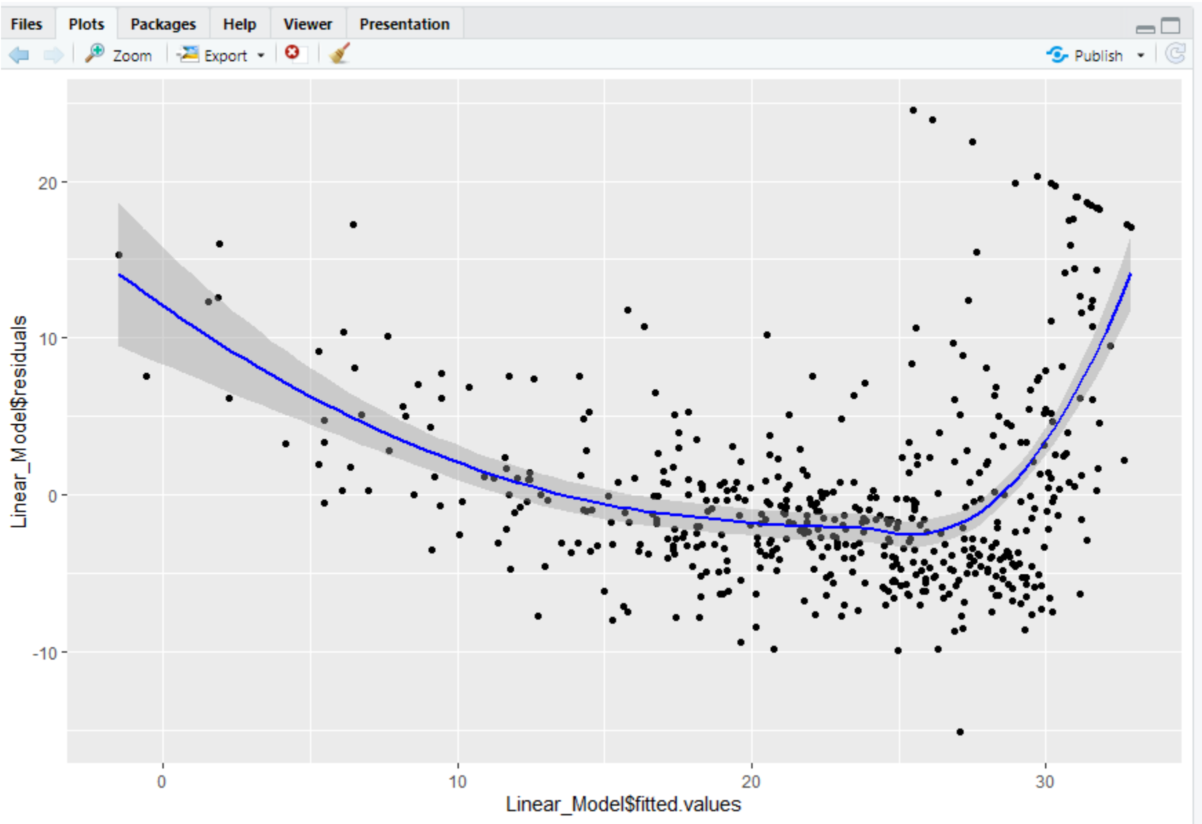
File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Console Terminal Background Jobs

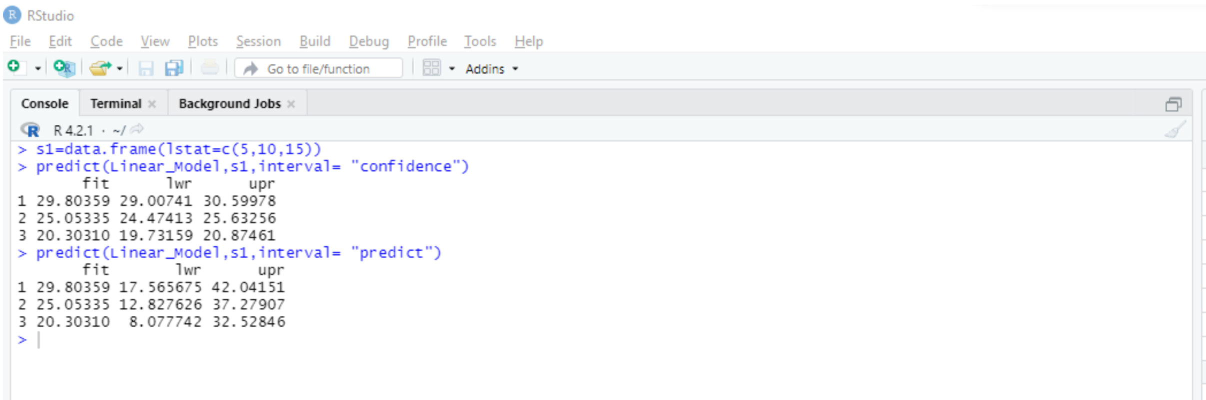
R 4.2.1 · ~/

```
> ggplot(data=Boston_Data,aes(x=Linear_Model$fitted.values,y=Linear_Model$residuals))+geom_point()+stat_smooth(col="blue")
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
> coef(Linear_Model)
(Intercept)       lstat 
 34.5538409  -0.9500494 
> |
```



From above graph we can say that predictors and response variables associated with nonlinear relationship.

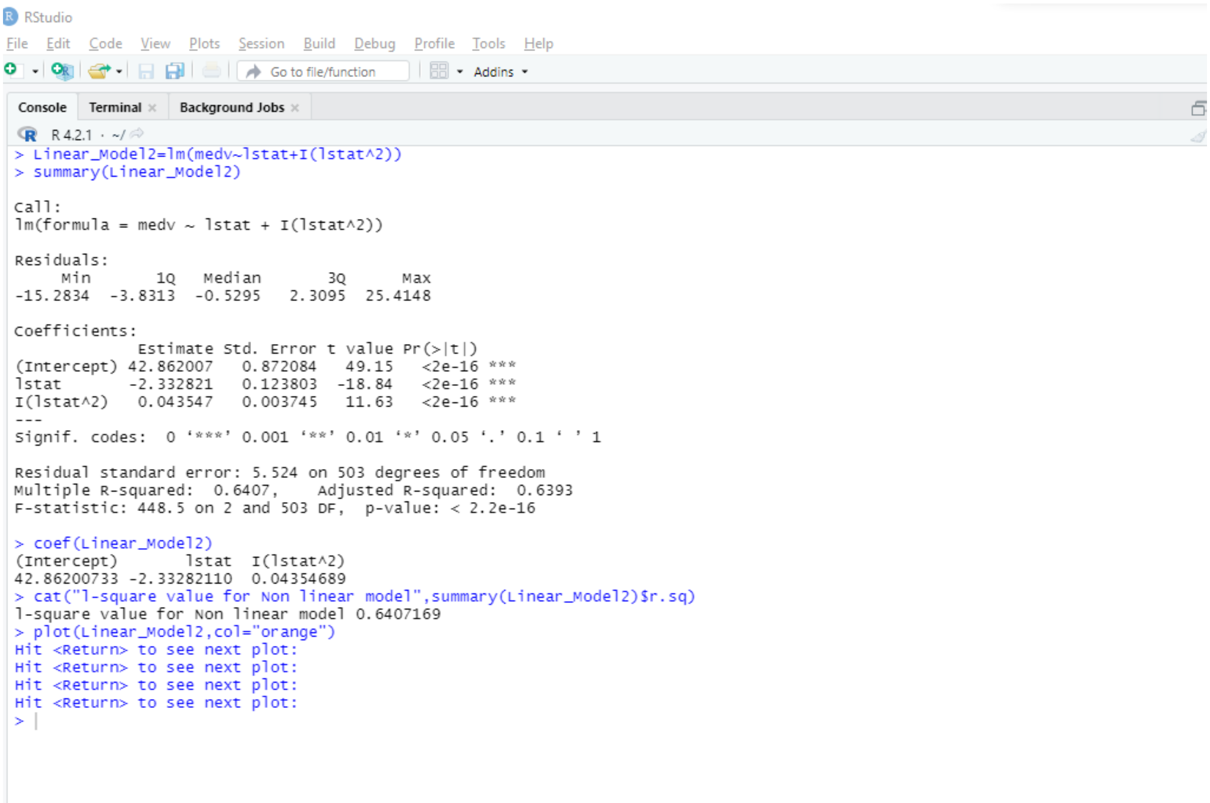
Use predict function to calculate values response values for lstat of 5, 10, and 15



```
> s1=data.frame(lstat=c(5,10,15))
> predict(Linear_Model,s1,interval= "confidence")
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
> predict(Linear_Model,s1,interval= "predict")
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
> 
```

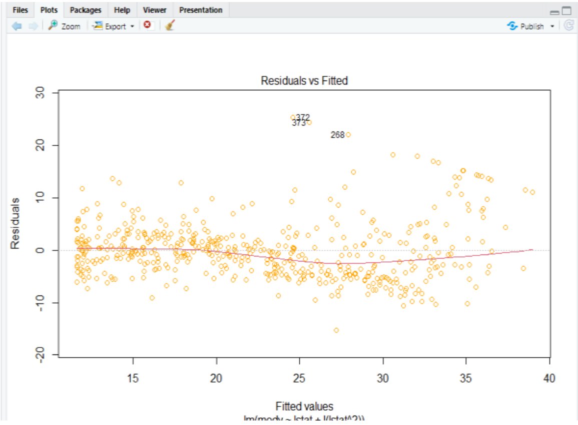
Form above we can conclude ,response values the interval confidence and predict are not same.for both interval we get same fitted values except range which is higher in prediction interval due to error.For prediction interval has uncertainty around single value whereas for confidence its around mean prediction.

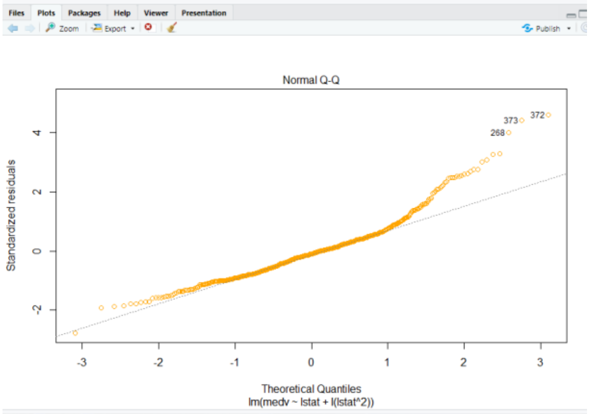
Modify the regression to include lstat2

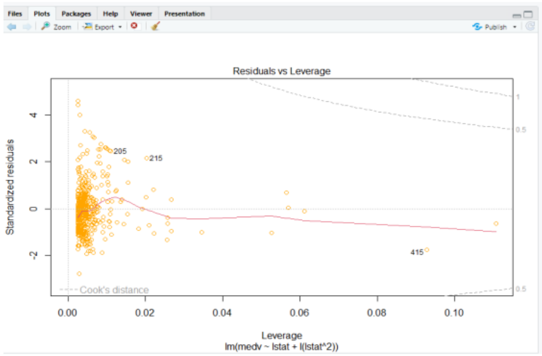


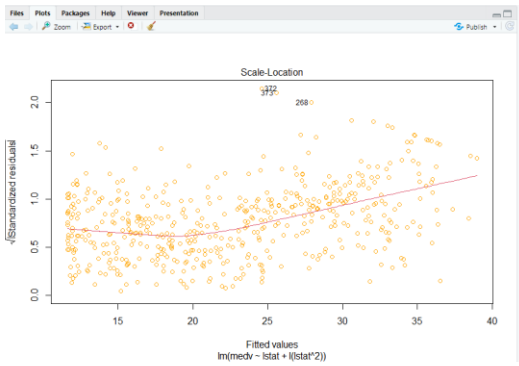
```
> Linear_Model2=lm(medv~lstat+I(lstat^2))
> summary(Linear_Model2)

Call:
lm(formula = medv ~ lstat + I(lstat^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2834  -3.8313  -0.5295   2.3095  25.4148 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,	Adjusted R-squared:  0.6393 
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16

> coef(Linear_Model2)
(Intercept)       lstat  I(lstat^2) 
42.86200733 -2.33282110  0.04354689 
> cat("l-square value for Non linear model",summary(Linear_Model2)$r.sq)
l-square value for Non linear model 0.6407169
> plot(Linear_Model2,col="orange")
Hit <Return> to see next plot: 
Hit <Return> to see next plot: 
Hit <Return> to see next plot: 
Hit <Return> to see next plot: 
> 
```

R-square value for linear model2 has increased from 54% to 64% means 10% more variance.so Performance of the model increased with high degree of polynomial.

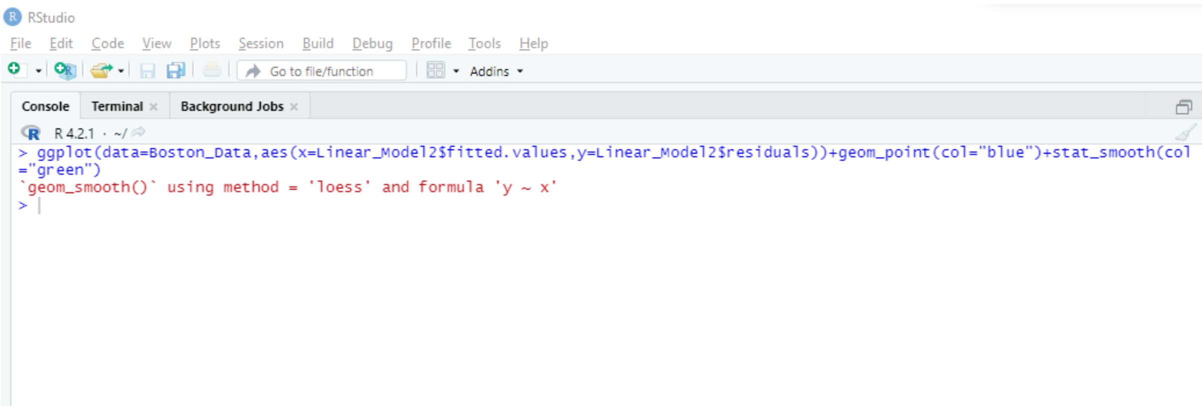


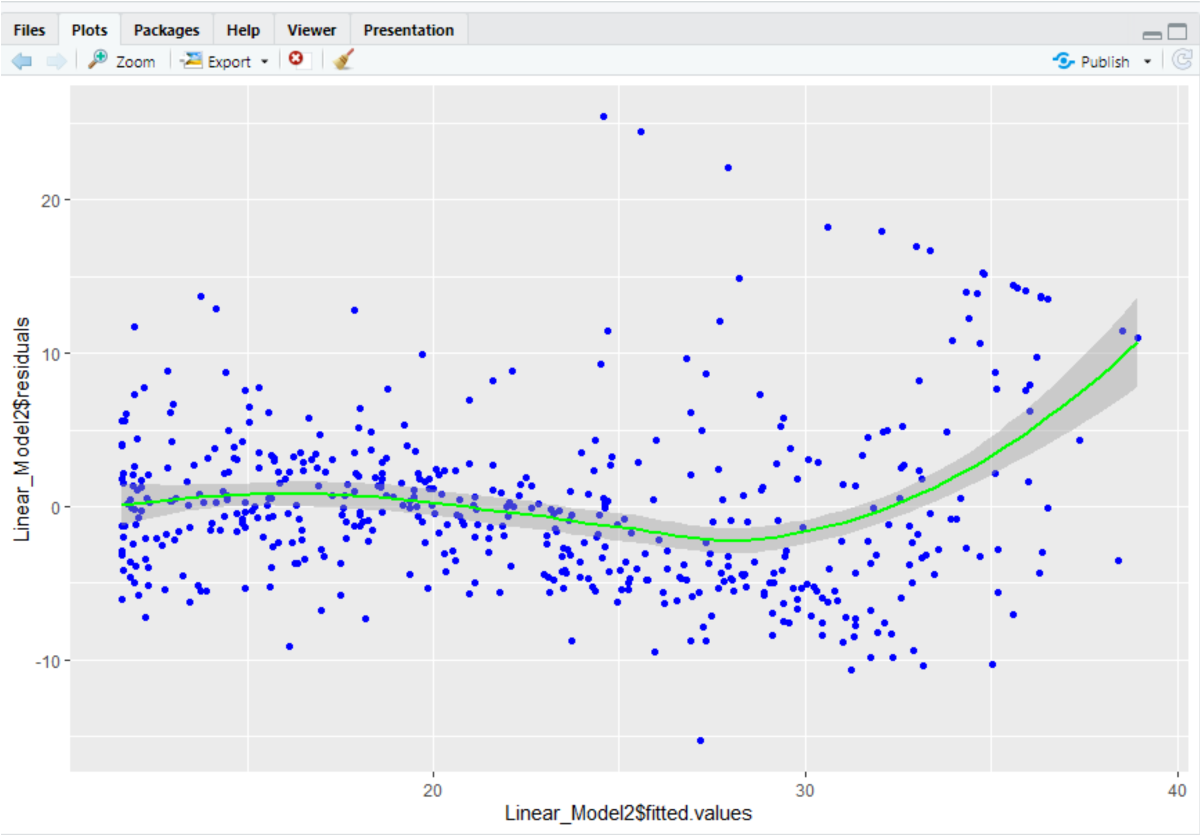




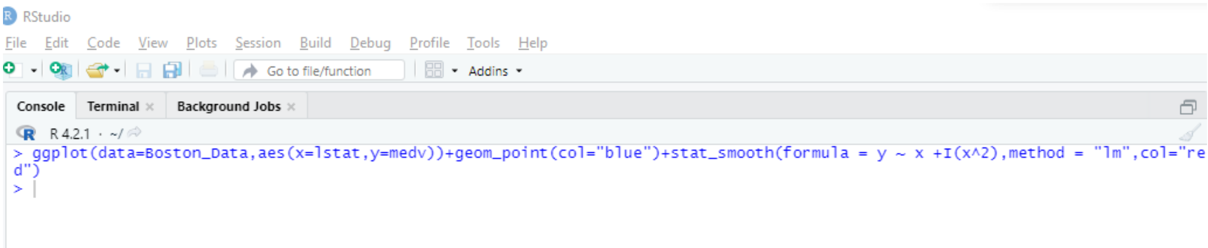


Plot for fitted value vs Residual values:

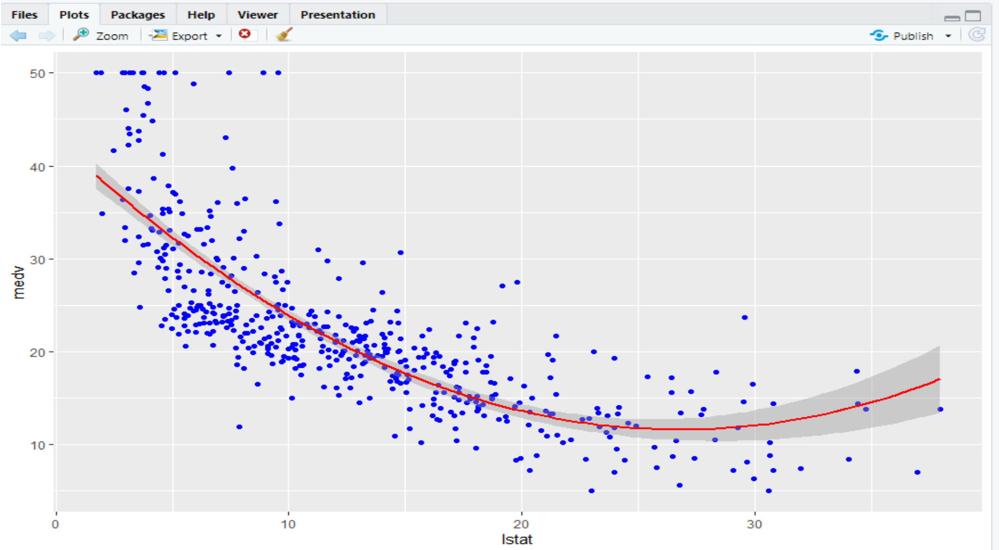




Plot for Non-linear fit:



SHRADDHA A20499171



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Console Terminal Background Jobs

R 4.2.1 · ~/

```
> anova(Linear_Model,Linear_Model2)
Analysis of Variance Table

Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    504 19472
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```