Linear Regression in R

Linear Regression is a very simple approach for supervised learning. It is a useful tool for predicting a quantitative response and most widely statistical learning method. linear regression is important because many fancy statistical learning approaches can be seen as generalizations or extensions of linear regression.

In this blog I will work on Boston dataset which is a default dataset in R package 'MASS'. Dataset includes 13 explanatory variables and a response variable and contains 506 rows and 14 columns.

Here are a few important questions that we might seek to address:

1. Is there a relationship between explanatory variables and a response variable?

Our first goal should be to determine whether the data provide evidence of an association between explanatory variable and response variable. If the evidence is week then one might argue that this variable is statistically insignificant.

1. How strong is the relationship between explanatory variables and the response variable?

Assuming that there is a relationship between explanatory variables and a response variable. We would like to show the strength of this relationship. In other words, given a certain values of explanatory variables, can we predict the response (medv) with a high level of accuracy.

1. Which explanatory variable contribute to Response variable (medv)?

Do all the explanatory variables contributes to the response variable (medv), or do just one or two of the media contribute? To answer this question, we must find a way to separate out the individual effects of each explanatory variable.

1. How accurately can we estimate the effect of each explanatory variable on Response variable?

For every explanatory variable, what is the individual effect of these variable on response variable and how accurately can we predict the effect of increase?

1. How accurately can we predict future Response (medv)?

For any given values of predictors, we can predict future values for Response variable (medv).

1. Is the relationship linear?

If there is approximately a straight-line relationship between explanatory variable and a response variable, then linear regression is an appropriate tool. If not, then it may still be possible to transform the predictors or the response so that linear regression can be used.

**Simple linear regression in R**

#install desired packages

install.packages("MASS")

install.packages("ISLR")

#call the desired packages

library(MASS)

library(ISLR)

#Reading Boston Dataset in R

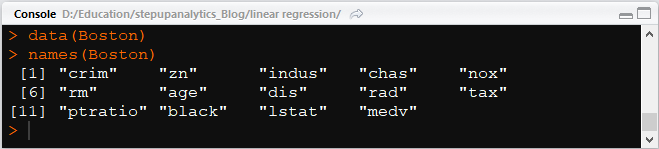
data(Boston)

#Dimension of Dataset Boston

dim(Boston)

#names of the variables in the dataset Boston

names(Boston)



# To learn more about the definition of each variable, type the below code in your R console,

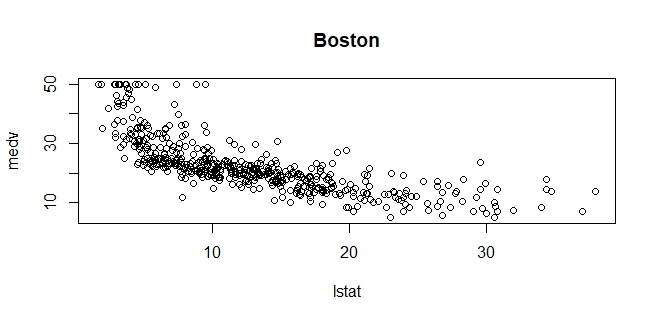
?Boston or help(Boston)

#to view the dataset Boston

View(Boston)

#Before starting with the analysis, it is often useful to understand the data by visualizing it. For this dataset, we can use a scatter plot to visualize the data, since it has only two properties to plot (medv and lstat). Many other problems that we encounter in real life are multi-dimensional and can’t be plotted on a 2-d plot.

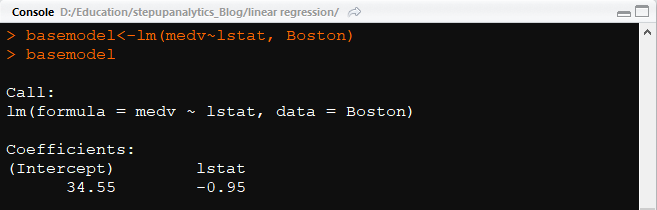
plot(medv~lstat, main = "Boston", data = Boston)



#response~predictor (response is modeled as predictor)

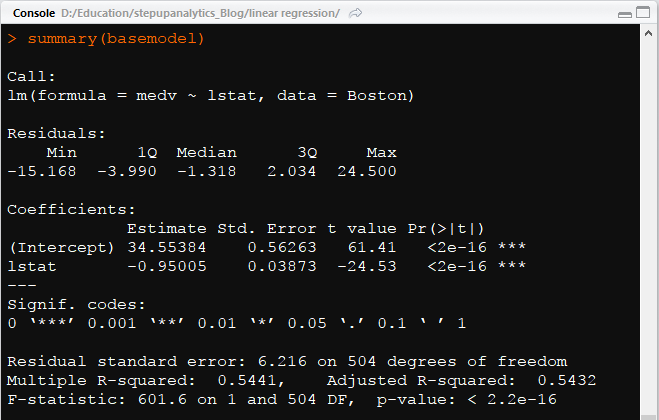
Basemodel <- lm(medv~lstat, Boston)

Basemodel



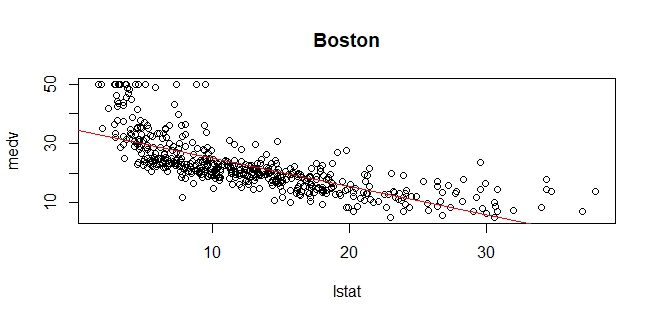
summary(basemodel)

#summary function will let us know the descriptive statistic like minimum, quartiles, median value etc.



#add a line to the fit

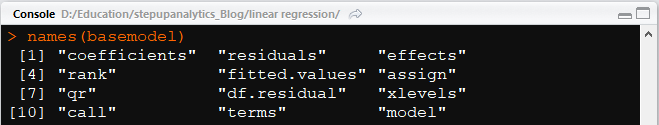
abline(basemodel,col="red")



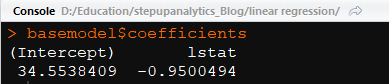
#see the components of fit

#access any one of these like "basemodel$coefficients" etc.

names(basemodel)



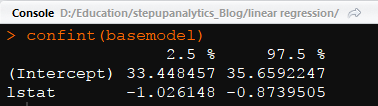
basemodel$coefficients



The intercept value is the value at which the straight line cuts the y-axis.

#95% confidence interval

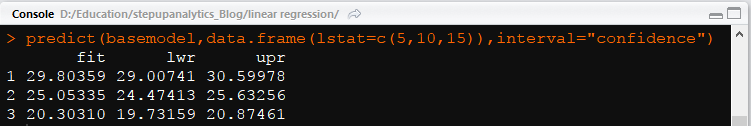
confint(basemodel)



#predict medv (response) for these 3 values of lstat (predictor).

#also show confidence intervals

predict(basemodel,data.frame(lstat=c(5,10,15)),interval="confidence")



From the above, we can see that for each prediction there are corresponding lower and upper values or interval.

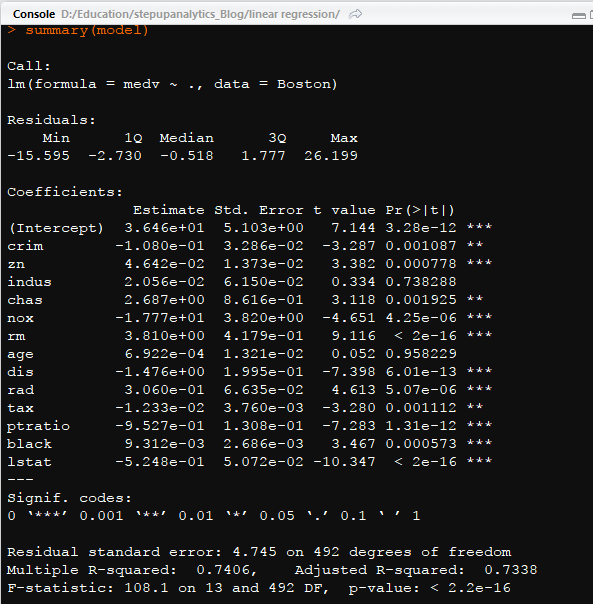
**Multiple linear Regression in R**

# fitting the regression model

model <- lm(medv~.,data = Boston)

# summary of the fitted model

summary(model)

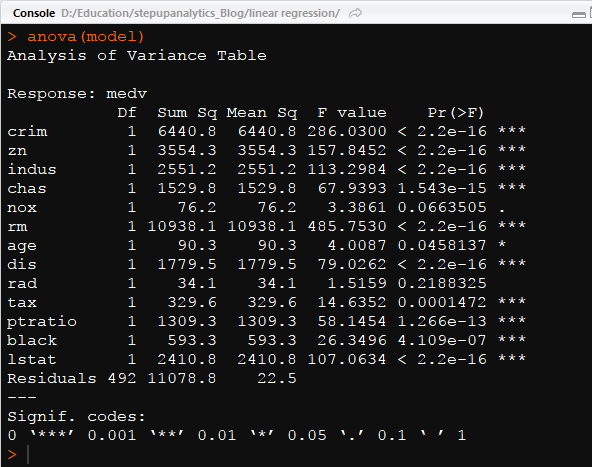


Interpretation:

Clearly we can see that indus and age are statistically insignificant. Since the values are greater then 0.05. Removing these variables would cause no change in the fitted model.

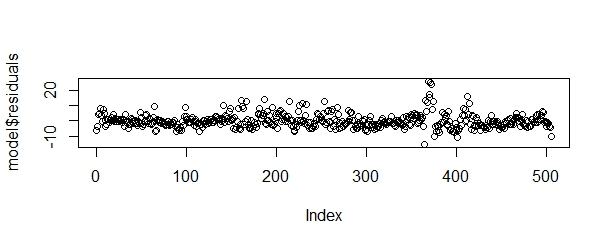
# anova of the fitted model

anova(model)



# plot of the fitted model

plot(model$residuals)

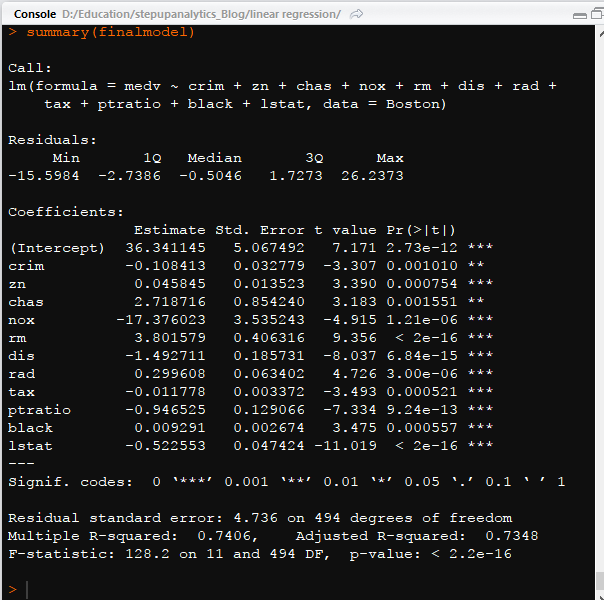


#fitting the regression model with only significant variables

finalmodel <- update(model,~.-age-indus, data = Boston)

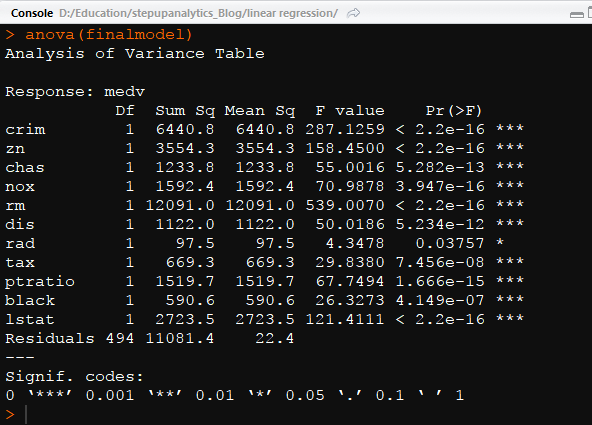
# summary of fitted model

summary(finalmodel)



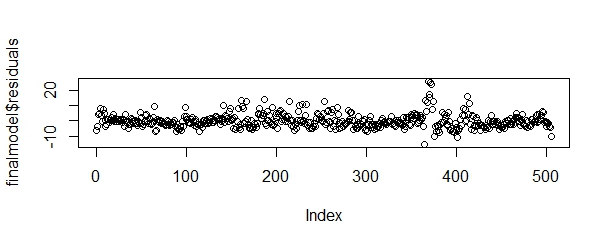
# anova of fitted model

anova(finalmodel)



# plotting of the finalmodel

plot(finalmodel$residuals)



Skill-Up Scale-Up !!!