**Assignment #2: Linear Regression**

**Submit through link: eCampus -> Assignments->Assignment 2 Submission**

**Deadline: September 28 (Friday) @11:59pm**

**# YOUR NAME: ....**

**#**

**# RULES:**

**#**

**# 1) You can discuss the homework with each other in general terms,**

**# but you must write your own solutions and not copy from anyone.**

**#**

**# 2) Edit your answers into this file following each problem. Write the answer Problem 1(b)**

**# right after Problem 1 part (b).**

**#**

**# 3) For problems 1&2 you do not need to use R.**

**#**

**# 4) Post your solutions to e-campus**

**#**

**# The filename should have this format: LastName-FirstName-hw02.doc**

**# An example would be: Eksin-Ceyhun-hw02.doc**

**Problem 1 (1pt)**

Assume that the mean height for women at a large university (to be viewed as a population) is 65 inches with a standard deviation of 3 inches.

1. If the women are placed randomly into classes of 36 each, what will be the standard deviation of the class means for height (i.e. the standard error of the mean)?

Answer= 0.5

1. Using your answer to part (a), what is the z score for a class whose average height is 64.2 inches? What is the two-tailed p-value for this class?

Answer Z= -1.6, p value= 0.1096

1. If you were testing the null hypothesis (i.e. μ=65), would you reject H0 for this class at the 0.05 level for a two-tailed test?

Answer - No, since z = -1.6 lies inside the 0.05 level (-1.96 to 1.96)

1. Repeat part (b) for a class whose mean height is 67.4 inches. Would you reject the null hypothesis for this class with a two-tailed test at the 0.05 level? At the 0.01 level?

Answer -Reject.

We will reject the Null hypothesis μ=65 for both these levels max (-2.58, 2.58) will be the confidence interval for z value. And z value is 4.8 which is outside the interval.

**Problem 2 (1pt)**

This exercise relies on the following data set: 1, 3, 6, 0, 1, 1, 2, 1, 4.

1. Perform a t test in order to decide whether you can reject the null hypothesis of μ=2.5 at the 0.05 level (two-tailed) for these data.

Answer - No. We can’t reject

since 2.5 lies in 95% confidence interval (0.65, 2.57)around the sample mean.

1. Redo your t test in part (a) for a null hypothesis of μ=6.0

Answer -Yes, We will reject it.

Since μ=6.0 lies out of the interval

1. Compute the 95% confidence interval (CI) for the population mean form which these data were drawn. Explain how this CI could be used to draw conclusions about the null hypothesis in parts (a) and (b).

Answer - (0.65, 2.57) since μ=2.5 is inside the interval so we cant reject it and μ=6.0

lies outside it so we can reject it.

**Problem 3 (2pt)**

Use the Auto data set to answer the following questions:

(a) Perform a simple linear regression with mpg as the response and horsepower as the predictor. Comment on the output. For example

i. Is there a relationship between the predictor and the response?

Answer - Yes, since the value of F statistic is quite large and P value is small we can say there is a relationship between predictor and response.

ii. How strong is the relationship between the predictor and the response?

Answer - The coefficient of horsepower in the estimator function is -0.15 so unit increase in horsepower decreases mpg by 0.15 and since the P value is quite low. The relationship is significant.

iii. Is the relationship between the predictor and the response positive or negative?

Answer -The relationship is negative. It can be seen by the -ve sign of the coefficient.

iv. How to interpret the estimate of the slope?

Answer -The coefficient of horsepower in the estimator function is -0.15 so unit increase in horsepower decreases mpg by 0.15.

v. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

Answer - confidence interval:

fit lwr upr

24.47 23.97 24.96

Prediction Interval:

fit lwr upr

24.47 14.81 34.12

(b) Plot the response and the predictor. Display the least squares regression line in the plot.

Answer - lm.fit = lm(mpg~horsepower, data= Auto)

plot(horsepower,mpg)

abline(lm.fit)

1. Produce the diagnostic plots of the least squares regression fit. Comment on each plot.

Answer -par(mfrow=c(2,2))

plot(lm.fit)

1. **Residual vs fitted** - it shows that there is some non-linearity between the mpg and horsepower relation
2. **Normal Q**-Q plot- It shows that the residuals are normally distributed which is good.
3. **Scale Location Plot**- It is also showing a hint of non-linearity in the fit.
4. **Residuals Vs leverage**- This graph tells us about any influential outlier if it is beyond cook’s distance. It can negatively influence our regression analysis.
5. Try a few different transformations of the predictor, such as , and repeat (a)-(c). Comment on your findings.

Answer - The relationship with (horsepower)^2 is quite significant and it is also reducing the non-linearity which is evident in the diagnostic plots. So it is a better fit that simple linear regression

**Problem 4 (2pt)**

Use the Auto data set to answer the following questions:

(a) Produce a scatterplot matrix which includes all of the variables in the data set. Which predictors appear to have an association with the response?

Answer- pairs(Auto), yes the cylinders, displacement, acceleration, horsepower and weight seem to have a association with the mpg.

(b) Compute the matrix of correlations between the variables (using the function cor()). You will need to exclude the name variable, which is qualitative.

Answer- Data<- Auto

Data$name<-NULL

Cor(Data)

(c) Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Comment on the output. For example,

lm.fit=lm(mpg~.-name, data=Auto)

summary(lm.fit)

1. Is there a relationship between the predictors and the response?

Answer- Yes there is a significant relationship between them as the F statistic is very large and P value is very low.

1. Which predictors have a statistically significant relationship to the response?

Answer- Displacement, weight ,year and origin

1. What does the coefficient for the year variable suggest?

Every year car mpg increased by 0.75

1. Produce diagnostic plots of the linear regression fit. Comment on each plot.

**1. Residual vs fitted** - it shows that there is some non-linearity between the response and predictors regressed relationship.

**2. Normal Q**-Q plot- It shows that the residuals are normally distributed which is good.

**3. Scale Location Plot**- It is also showing a hint of non-linearity in the fit.

**4. Residuals Vs leverage**- This graph tells us about any influential outlier if it is beyond cook’s distance. It can negatively influence our regression analysis. Point 14 has high leverage but still under cooks line so not much significant.

(e) Is there serious collinearity problem in the model? Which predictors are collinear?

Answer- using vif(lm.fit) from car library.

Yes a few vif values are high enough to say that there is a collinearity problem in the model.

(f) Fit linear regression models with interactions. Are any interactions statistically significant?

Answer- lm.fit1= lm(mpg~cylinders\*displacement+ displacement\*weight+cylinders\*weight, data= Auto)

summary(lm.fit1)

only displacement\*weight interaction has a low p value which is significant.

**Problem 5 (2pt)**

Use the Carseats data set to answer the following questions:

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

Answer- attach(Carseats)

lm.fit = lm(Sales~Price+Urban+US)

summary(lm.fit)

(b) Provide an interpretation of each coefficient in the model (note: some of the variables are qualitative).

Answer- **price** is negatively impacting sales as P value is low and slope is -ve.

**urbanYes** is not significant due to low p value which means that location of store in urban area or not doesn’t impact sales significantly.

**USYes** yes the location of store in USA will cause an extra 1200 unit sale

(c) Write out the model in equation form.

Sales = 13.04 + -0.05 Price + -0.02 UrbanYes + 1.20 USYes

**The sales value is given in 1000s**.

(d) For which of the predictors can you reject the null hypothesis ?

Answre- Price and USYes null hypothesis can be rejected as the p value is low and t statistic is large.

(e) On the basis of your answer to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the response.

Answer- lm.fit1 = lm(Sales~Price+US, data=Carseats)

summary(lm.fit1)

(f) How well do the models in (a) and (e) fit the data?

Answer- Model (e) fits better as R square value is higher and RSE value is lower

(g) Is there evidence of outliers or high leverage observations in the model from (e)?

Answer - plot(predict(lm.fit1), rstudent(lm.fit1))

all points lie within -3 to 3 interval so there are no outliers which can significantly alter our regression analysis.

**Problem 6 (10pt)**

It is 1970 and you are hired by a real estate company in Boston as a data consultant. The company would like to make price estimates based on different preferences that a customer can have, so that they can assess if a customer’s budget is realistic or not. Your task is to build a model, and write an executive summary that outlines important features that affect the price of a house.

Attach “Boston” data included in the MASS library. The data contains information from 504 geographic areas. There are 14 attributes in each area of the dataset. You may find the description of each attribute using help(Boston) code in R.

Your model should have the log transformation median house value (log(medv)) as the output variable. When you are searching for a meaningful regression equation, consider the following attributes:

1. Structural properties of a house (age, number of rooms, lot size allowed by zoning laws in the area)
2. Accessibility (distance to major employment centers and closeness to highways)
3. Neighborhood (crime rate, education quality, whether it is by the Charles river or not)

You may include other terms in your model but you should discuss a model that contains above terms as needed. When you discuss the effect of an attribute on price, you need to describe the meaning of the attribute and the ranges of values associated with the attribute. For example, if there is an effect of crime rate on house prices, it would be necessary to report not only the coefficient associated with the age but what is the worst crime rate and best crime rate across different areas. Decide on one model and use it to write an executive summary.

**Executive Summary**

Summarize your findings to the president of the real estate company at a level that a non-technical person (president or a realtor) can understand. The executive summary is **at most one pages**, **single spaced and in 12-point type.** Do not include any graphs or statistical concepts. Use the following page to write your executive summary.

In the executive summary,

1. Discuss important attributes that significantly affect the value of a house.
2. Introduce a common baseline scenario such as relatively young houses near decent schools with great accessibility, and discuss its estimated price and price range.
3. Describe the attributes that yield highest house values (houses that are in top 5% in value)
4. Discuss certain bargains, such as if you sacrifice a certain neighborhood attribute you can afford to live at a bigger house for a budget that is at the mean of house values.
5. Use rounded numbers that are memorable.

**Technical Summary**

In the technical summary you speak to your peers. You show them that you performed a reasonable analysis and that you interpreted the results competently. This should not be a step by step report of what you did in R, but a summary of the most important steps in logical, not chronological, order. Even in a technical summary it is not of interest to hear, for example, how you used R; it is simply assumed that you know how to execute with available software.

The technical narrative should explain what values were used in the executive summary and how they were rounded. In addition, it should explain what contributions to the model were neglected because their effect on house value is too small.

The technical summary should, among other things, explain the final fitted model, term by term and estimate by estimate. It should mention model diagnostics that were performed and their outcomes, possibly accompanied by plots. Report data points (geographical areas) that you may have removed, if any, and why you did so.

Technical summary should at most be **five pages including any figures/tables you choose to include.** It should follow your executive summary page.

**To:** President X

**From:** Data Consultant Y

**Subject:** Analysis of Boston Suburban Housing Values, Executive Summary’

Housing Prices in suburbs of Boston are influenced by structural properties of the house, accessibility to good education, distance from the major employment centres of the city, environmental conditions and many other factors which are further explained in this summary.

The analysis of this data shows that the no. of rooms and percentage of people in lower economic status greatly affects the median value of houses. An extra room in a house will increase the price by an average of 24% .

Top 5% high priced houses have a minimum of 5 rooms and goes upto 8. The percentage of lower status people are lesser in the areas where these high priced houses are located. unit increase in these people’s population causes the house prices in that town to go down depending upon the number of rooms in that particular house.

The Low priced houses have fewer than rooms as it is evident that no of rooms are directly positively related to the value of house. The population of lower socio-ecomonic status of people are quite high( more than 25%) in the areas of these low priced houses.

One of the most evident factor determining the price of the property is its nearness to Charles river. The houses near Charles river are priced higher than their counterparts which have the same amenities and other parameters except the vicinity to Charles river.

Another important feature is the quality education. The common relation of quality education is with the people to teachers ratio in a town and the towns having lesser no of teachers per student has lower property rates although not by a very significant margin. Accessibility to highways are also increasing the house prices.

What’s most surprising is that the percentage of old houses in a town is not really significantly affecting the house prices. Another not so significant factor for determining the price is distance the distance from the five employment centres of Boston. Which intuitively should affect the price but does not so according to this data. Amount of zoned land, non-retail business acres, nitrogen oxide concentration are not having much effect on the prices.

The crime rate is also affecting the prices negatively. High prices houses locality have considerably less crime rate than the low priced houses localities. The localities with lowest prices have crime rates near 80 per capita which is apparently very high.

For families having very high budget houses in localities having low crime rates Less nox levels, good educational facilities and near to Charles river can be suggested.

People having low budgets can be offered houses little far from radial highways and away from Charles river can be suggested. Sacrificing some criteria one can get a house in low prices.

**TECHNICAL SUMMARY**

The Boston data set contains information collected by US Census service concerning housing in the Boston. The data set captured some variables against each house such as crime rate per capita in the town, nitric oxide levels, No of rooms, Age of houses, accessibility to highways and employment centres, median value of houses. The data set is small with only 506 number of samples.

The task at hand is to predict the Median values of houses in Boston using the following attributes

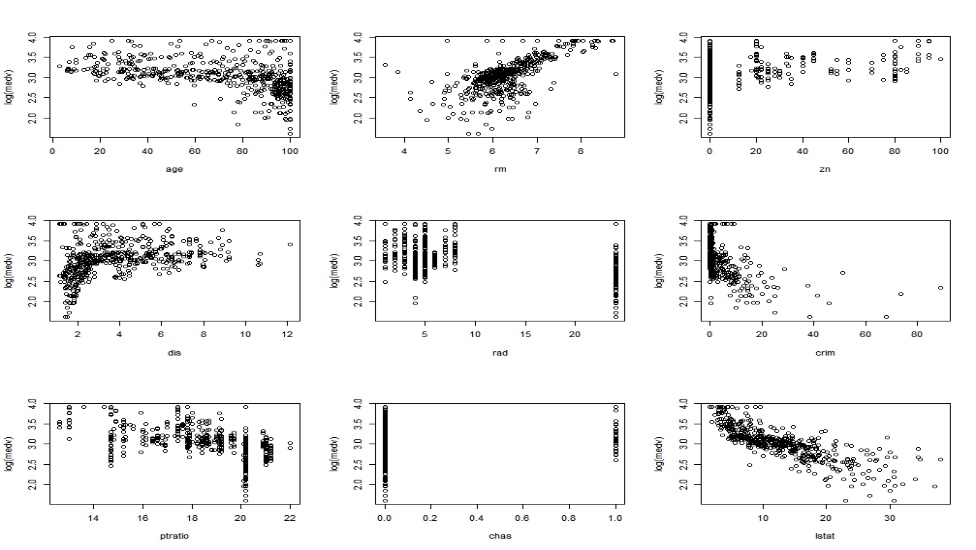
1. Structural properties of a house such as age, number of rooms, lot size allowed by zoning laws in the area
2. Measure of Accessibility such as distance to major employment centers and closeness to highways)
3. Neighborhood attributes such as crime rate, education quality, whether it is by the Charles river or not)

In the data set we have been provided the recorded variables which can be further categorized into attributes on which we have to base our model. Below are the list of variables which have been recorded in this dataset.

* *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* - nitric 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 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* - percentage lower status of the population
* *MEDV* - Median value of owner-occupied homes in $1000's.

**Exploratory data Visualization**

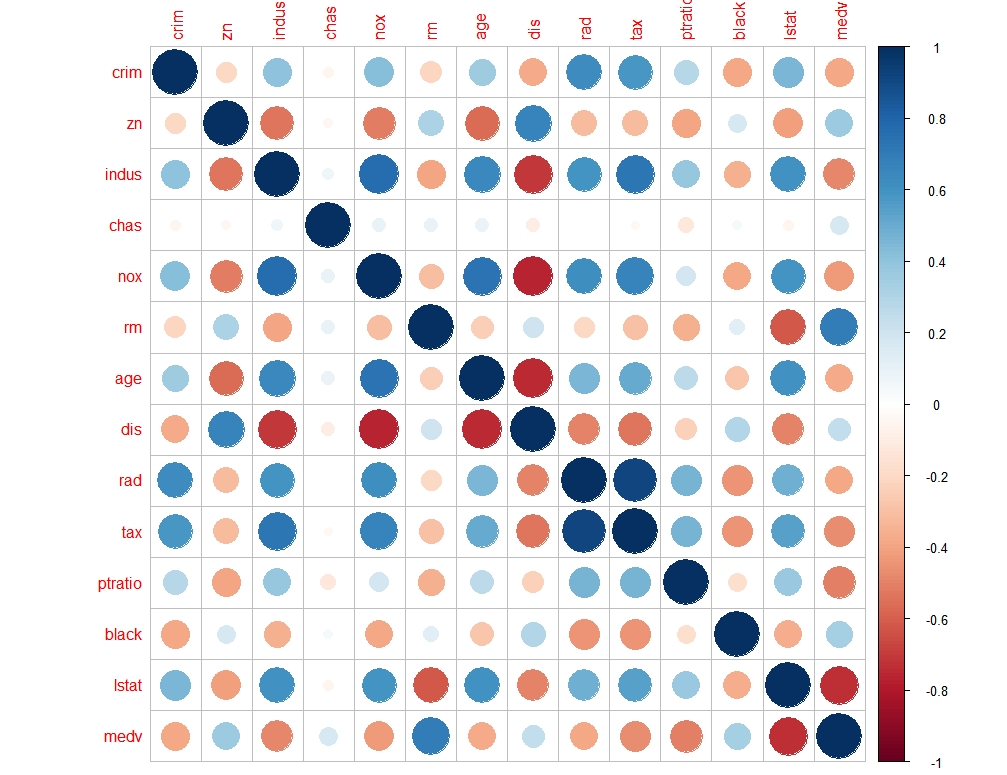
After plotting each of the predictor on X axis with log of median pricing on Y axis. We got the following plots for visualization from which we can get an idea of relationship between the price and the other variables.



*Scatter plot between the log(medv) vs different predictors*

Following inference can be made from the data By looking at the scatter plots:

* Median pricing seems to be heavily influenced by No of rooms(rm) , percentage of lower status of population(lstat) and crime rate (crim).
* Pricing is also showing slight relationships with some other predictors such as age(age), people to teachers ratio(ptratio), distance from employment centres and radial highway(dis and rad)

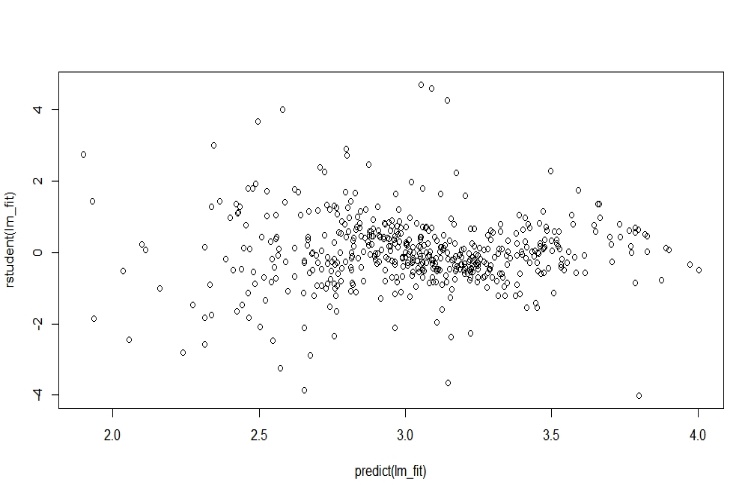
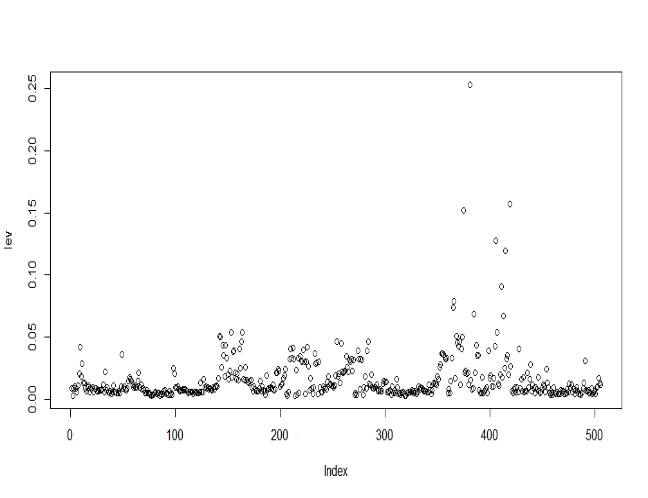


*Correlation matrix between the variables*

By looking at the correlation matrix between the variables we can see:-

* some of the variables are highly corelated with each other such as lstat and rm , nox and dis, dis and indus, dis and age .
* we will have to check for the interaction of these variables in our prediction model so that we can effectively remove them from the model because of redundancy or to include them in interaction terms with other variables.
* On testing it within the initial model we can find out that rm and lstat have a significant interaction term which is also increasing the Adjusted R-sq value. So we will keep this term in our final model.

**Data Cleaning(removing outliers and high leverage points)**



*Studentized residuals vs fitted* *values leverage values of the fitted model*

* A total of 9 outliers were identified whose studentized Residuals(rstudent) values were higher than 3
* 5 high leverage points were identified and removed from the data.
* This was done to prevent our RSS value from increasing much due to any anomaly.

**Final Fitted Model**.

***Log(medv)= 2.48 – 0.011 Crim + 0.09 chas -0.02 dis -0.022 ptratio +0.01 rad + 0.244 rm+(0.04 – 0.01 rm)\*lstat***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Log medv* | *intercept* | *crim* | *chas* | *ptratio* | *rad* | *rm* | *lstat* |
| *coefficient* | *2.48* | *0.011* | *0.09* | *0.022* | *0.01* | *0.244* | *(0.04-0.01rm)* |

This model has a R. sq statistic value of 80% and F statistic of 272.5.

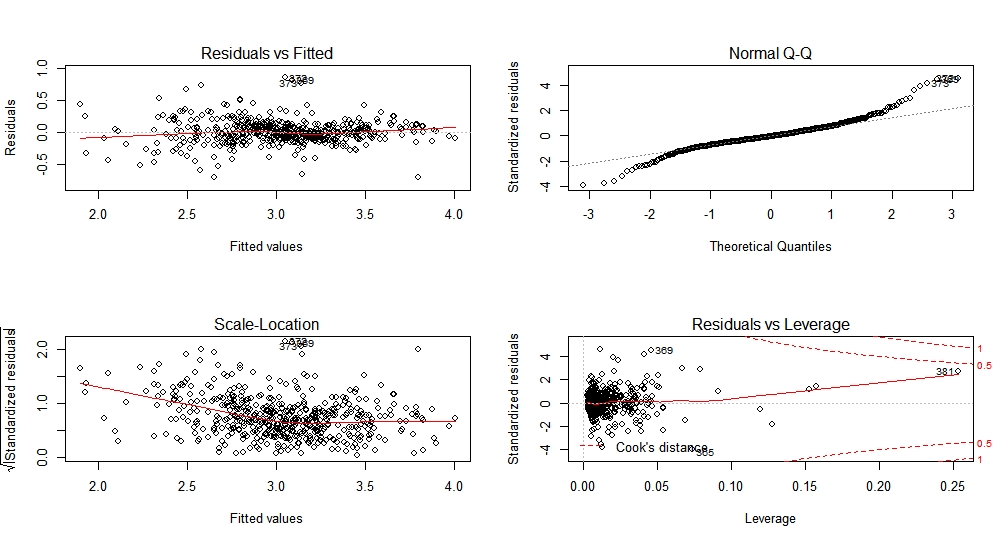
This model highlights the following relationship with each parameter

* **Intercept:-** The base price of house will be 1000\*(exp(3.9)) $ considering all other parameters to be zero.
* **Crime rate**:- house prices are negatively influenced by increasing crime rates in the locality. unit increase in crime rate per capita will decrease the house prices in that town by 1.1%
* **Vicinity of Charles river**:- If the property is located near chas river its price increases by 9%
* **Distance from Boston employment centres**:- unit increase in mean distance from 5 employment centres of Boston will decrease its price by 2%.
* **People to teachers ratio** : unit increase in this ratio will decrease the value of house by 2.2% . number of people per teacher increasing means lesser access to quality education.
* **No of rooms in the house**:- The value of house will increase 24.4% if we increase the number of rooms by 1 room.
* **Lower economic status of population**- The lstat parameter works in synergy with the rm parameter. If the percentage of people considered in lower economic status increase by 1 and and then the houses having different no of rooms will be effected differently. For example if a room has only 4 rooms its price will remain same if it’s a 5 room house its value will decrease by 1% .
* **Access to highway**- Unit increase in distance from highway decreases the value of house by 1%

Parameters like indus, nox, age, black , zn were removed because they were not having a significant relationship with the house pricing as they had very large P values(p> 0.05).

Some of the variables were also removed because of the strong correlation with other variables which implies redundancy and collinearity.

**Summary of the Diagnostic Plots**



* + **Residual vs fitted** – It shows that there is very slight non-linearity between the response and predictors regressed relationship. This plot is showing that the final model is an overall good fit.
  + **Normal Q**-Q plot- It shows that the residuals are normally distributed along Zero. Which should be the case with a good model.
  + **Scale Location Plot**- This plot shows that there is no non linearity in the final model.
  + **Residuals Vs leverage**- This graph tells us about any influential outlier if it is beyond cook’s distance. It can negatively influence our regression analysis. Point 375,381,406,415,419 from the data had been already omitted because of their high leverage. The final model doesn’t have any high leverage point as every point is inside cooks distance line so not much significant.