Multiple_regression

Nivedita

2024-04-06

Objective

To perform multiple linear regression on Boston dataset in Mass package.

Analysis

Boston dataset contains information collected by the U.S Census Service concerning real estate information in the area of Boston Mass. The information was obtained from the StatLib archive and has been used extensively throughout the literature to bench algorithms.

```
## Warning: package 'MASS' was built under R version 4.3.3
```

```
# Show number of rows and columns of the Boston dataset
dim(Boston)
```

```
## [1] 506  14
```

In the context of R, the Boston dataset is found in the MASS library and has 506 rows and 14 columns.

```
# Show first 6 rows of data
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

```
## Print out the column names of Boston dataset using names function
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

The variables give in the dataset are the following:

crim —per capita crime rate by town.

zn — proportion of residential land zoned for lots over 25,000 sq.ft.

indus — proportion of non-retail business acres per town.

chas — Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox — nitrogen oxides concentration (parts per 10 million).

rm — average number of rooms per dwelling.

age — proportion of owner-occupied units built prior to 1940.

dis — weighted mean of distances to five Boston employment centres.

rad — index of accessibility to radial highways.

tax — full-value property-tax rate per $10,000.

ptratio — pupil-teacher ratio by town.

black — 1000(Bk−0.63)^2 where Bk is the proportion of blacks by town.

lstat — lower status of the population (percent).

medv — median value of owner-occupied homes in $1000s.

First we will check for missing observations in the dataset.

```
sapply(Boston, anyNA)
```

```
##    crim      zn   indus    chas     nox      rm     age     dis     rad     tax
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## ptratio   black   lstat    medv
##   FALSE   FALSE   FALSE   FALSE
```

We can see that all values are 'FALSE' that means there is no missing value in the dataset.

```
summary(Boston)
```

```
##       crim                zn              indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm              age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

To fit a linear regression model, we select those features which have a high correlation with our target variable MEDV. For finding the variables with the strongest correlation with medv we plot correlation matrix.

```
ggcorrplot(cor(Boston), hc.order = TRUE, type = "upper", lab = TRUE, lab_size = 3, insig = "blank")
```

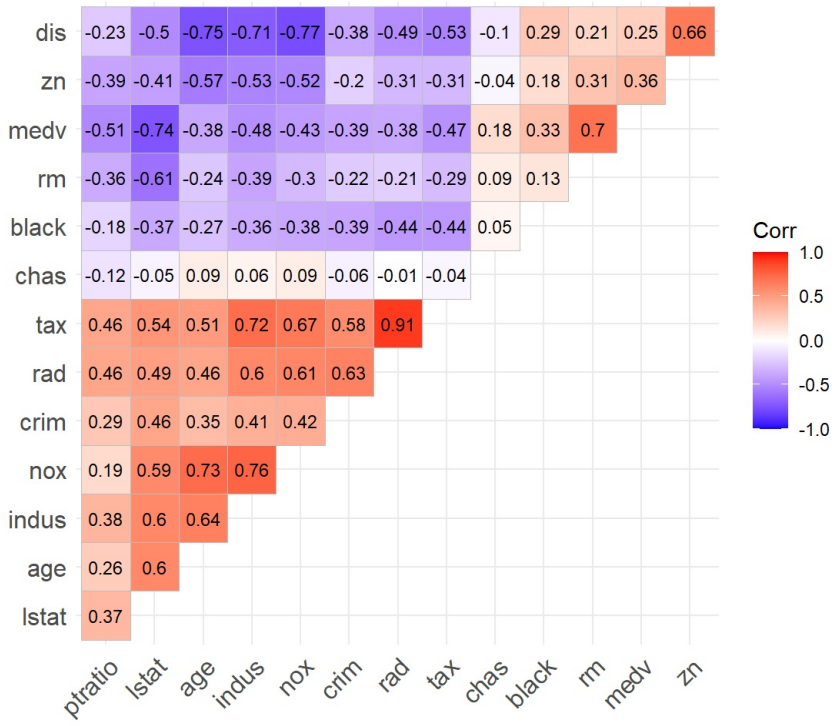

Fig. 1 Correlation matrix between each variable in Boston Dataset

The correlation coefficient ranges from -1 to 1. If the value is close to 1, it means that there is a strong positive correlation between the two variables. When it is close to -1, the variables have a strong negative correlation.

By looking at the correlation matrix in Fig. 1 we can see that:

To fit a linear regression model, we select those features which have a high correlation with our target variable MEDV. By looking at the correlation matrix we can see that RM has a strong positive correlation with MEDV (0.7) where as LSTAT has a high negative correlation with MEDV(-0.74). Also PTRATIO has moderate negative correlation with MEDV(-0.51).

An important point in selecting features for a linear regression model is to check for multicolinearity. The features RAD, TAX have a correlation of 0.91. These feature pairs are strongly correlated to each other. We should not select both these features together for training the model. Same goes for the features DIS and AGE which have a correlation of -0.75.

So out dependent variable is MEDV and independent variable are LSTAT, RM and PTRATIO.

Now we will use scatter plot to visualize the relationship between dependent variable and independent variable

```
## medv against lstat scatterplot
plot(Boston$lstat,Boston$medv,xlab="lower status of the population (LSTAT)",ylab ='median value of owner-occupied
homes in $1000s (MEDV)',main = 'Scatter plot between LSTAT and MEDV',col='skyblue',pch=19)
```

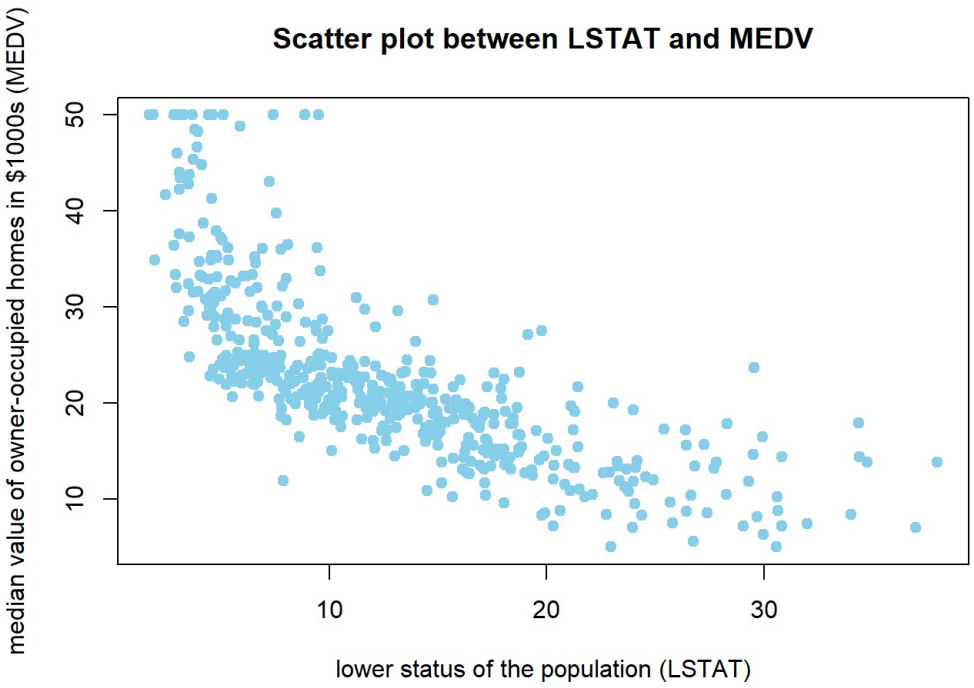

Fig. 2 Scatter plot between lower status of the population (percent) and median value of owner-occupied homes in $1000s

The scatterplot in Fig. 2 demonstrates a slightly curved linear relationship between MEDV and LSTAT and the prices tend to decrease with an increase in LSTAT.

```
## medv against lstat scatterplot
plot(Boston$rm,Boston$medv,xlab="average number of rooms per dwelling (RM)",ylab ='median value of owner-occupied
homes in $1000s (MEDV)',main = 'Scatter plot between RM and MEDV',col='skyblue',pch=19)
```

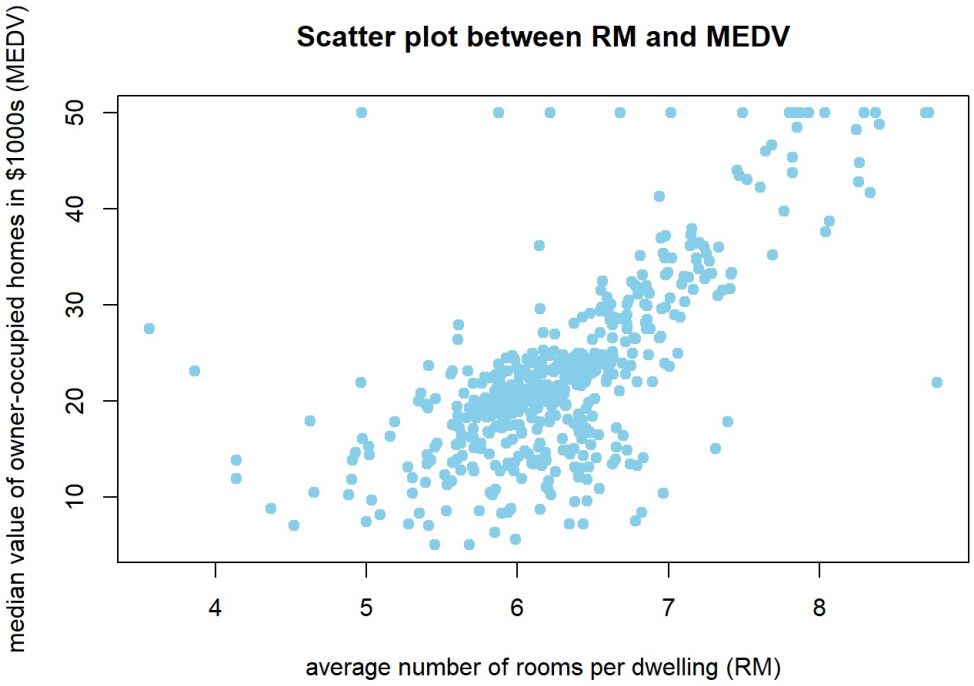

Fig. 3 Scatter plot between average number of rooms per dwelling and median value of owner-occupied homes in $1000s

From Fig. 3 we can see that the prices increase as the value of RM increases linearly. The data seems to be capped at 50.

```
## medv against lstat scatterplot
plot(Boston$ptratio,Boston$medv,xlab="pupil-teacher ratio by town (PTRATIO)",ylab ='median value of owner-occupie
d homes (MEDV)',main = 'Scatter plot between PTRATIO and MEDV',col='skyblue',pch=19)
```

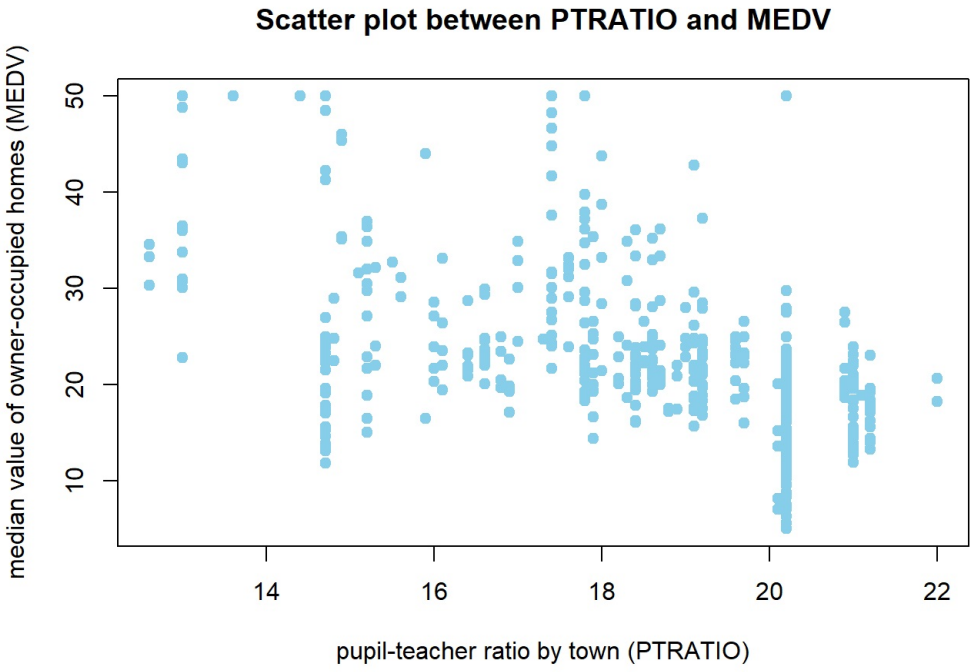

Fig. 4 Scatter plot between pupil-teacher ratio by town and median value of owner-occupied homes in $1000s

Suppose, Y = MEDV (dependent variable) X1 = LSTAT, X2 = RM and X3 = PTRATIO

So, Multiple linear regression model is given by-

Y = b0 + b1 X1 + b2 X2 + b3 X3 + e

where, b1,b2,b3 are slope coefficient , b0 is intercept term, e is disturbance term.

Now we will find the value of estimates of regression coefficients b0,b1,b2,b3.

```
#Fitting linear regression model
lm.fit<-lm(medv~lstat+rm+ptratio,data=Boston)
lm.fit
```

```
##
## Call:
## lm(formula = medv ~ lstat + rm + ptratio, data = Boston)
##
## Coefficients:
## (Intercept)        lstat           rm      ptratio
##     18.5671      -0.5718       4.5154      -0.9307
```

so we get,

Y_hat = 18.5671 - 0.5718 X1 + 4.5154 X2 -0.9307 X3

we can see that coefficient estimate of lstat = -0.5718. This means that if we increase 1% in lower status of the population (lstat) then on an average $571.8 median value of owner-occupied homes (medv) will decrease when effect of average number of rooms per dwelling(rm) and pupil-teacher ratio by town (ptratio) is kept constant.

Coefficient estimate of lstat = 4.5154. This means that if we increase 1 unit of average number of rooms per dwelling(rm) then on an average $4515.4 median value of owner-occupied homes (medv) will increase when effect of lower status of the population (lstat) and pupil-teacher ratio by town (ptratio) is kept constant.

Coefficient estimate of lstat = -0.9307. This means that if we increase 1 unit of pupil-teacher ratio by town (ptratio) then on an average $930.7 median value of owner-occupied homes (medv) will decrease when effect of lower status of the population (lstat) and average number of rooms per dwelling(rm).

Now our hypotheses will be,

Ho1: LSTAT has no significant effect on MEDV

H11: LSTAT has significant effect on MEDV

Ho2: RM has no significant effect on MEDV

H12: RM has significant effect on MEDV

Ho3: PTRATIO has no significant effect on MEDV

H13: PTRATIO has significant effect on MEDV

```
summary(lm.fit)
```

```
##
## Call:
## lm(formula = medv ~ lstat + rm + ptratio, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4871  -3.1047  -0.7976   1.8129  29.6559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.56711    3.91320   4.745 2.73e-06 ***
## lstat       -0.57181    0.04223 -13.540  < 2e-16 ***
## rm           4.51542    0.42587  10.603  < 2e-16 ***
## ptratio     -0.93072    0.11765  -7.911 1.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.229 on 502 degrees of freedom
## Multiple R-squared:  0.6786, Adjusted R-squared:  0.6767
## F-statistic: 353.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

We can see that, p value corresponding to LSTAT is < 2e-16 which is < 0.05, so we have enough evidence to reject null hypothesis H01 at 5% level of significance that means LSTAT has significant effect on MEDV.

p value corresponding to RM is < 2e-16 which is < 0.05, so we have enough evidence to reject null hypothesis H02 at 5% level of significance that means RM has significant effect on MEDV.

p value corresponding to PTRATIO is 1.64e-14 which is < 0.05, so we have enough evidence to reject null hypothesis H03 at 5% level of significance that means PTRATIO has significant effect on MEDV.

R-squared = 0.6786 that means 67.86% of the variation in the dependent variable, y or medv can be explained by the independent variables.

```
#Create diagnostic plots
plot(lm.fit,which=1)
```

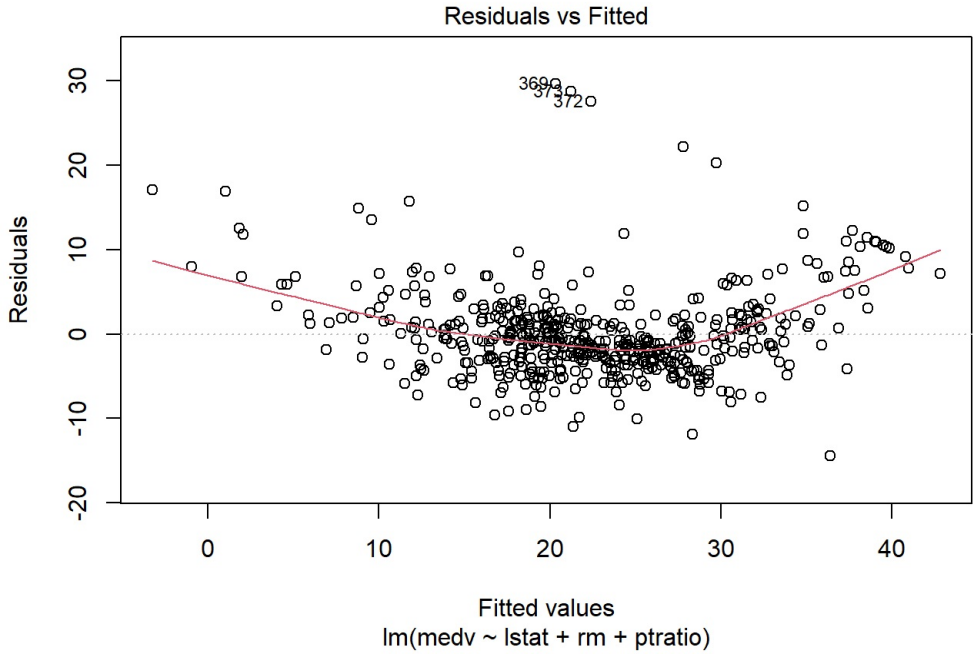

Fig. 5 Residuals vs Fitted Plot

The plot in Fig. 4 This plot helps to determine if the residuals exhibit non-linear patterns. If the red line across the center of the plot is roughly horizontal then we can assume that the residuals follow a linear pattern.

In our case, the red line deviates side from a perfect horizontal line. The residuals doesn't follow a linear pattern.

```
#Create diagnostic plots
plot(lm.fit,which=2)
```

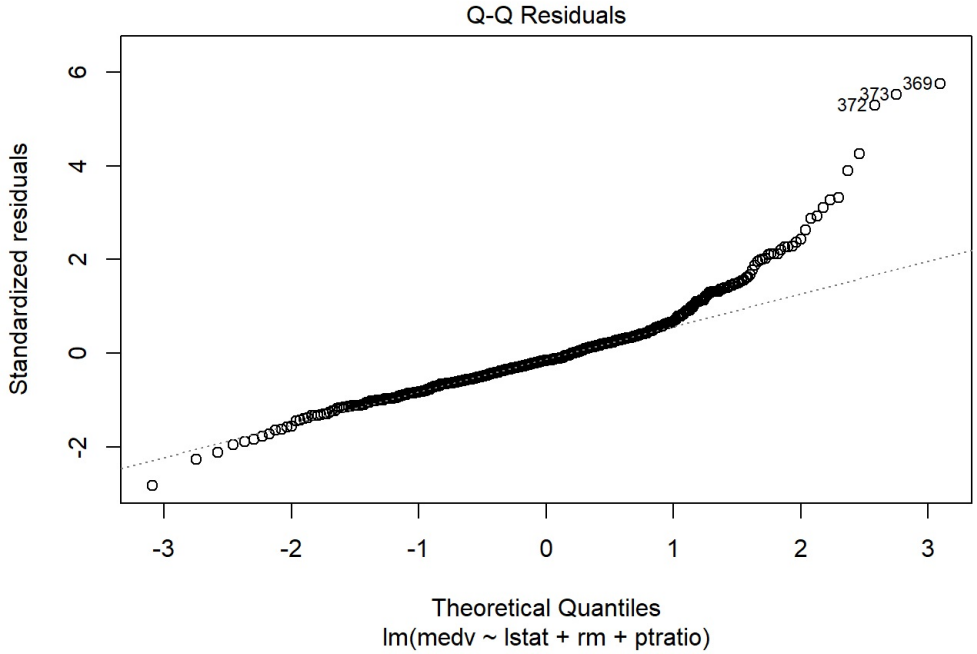

Fig. 5 Normal Q-Q Plot

The plot in Fig. 5 is used to determine if the residuals of the regression model are normally distributed. If the points in this plot fall roughly along a straight diagonal line, then we can assume the residuals are normally distributed.

The observations 369,373,372 deviate far from the line. Residuals are not normally distributed,a little skewed to the right. A different functional from may be required.

```
#Create diagnostic plots
plot(lm.fit,which=3)
```

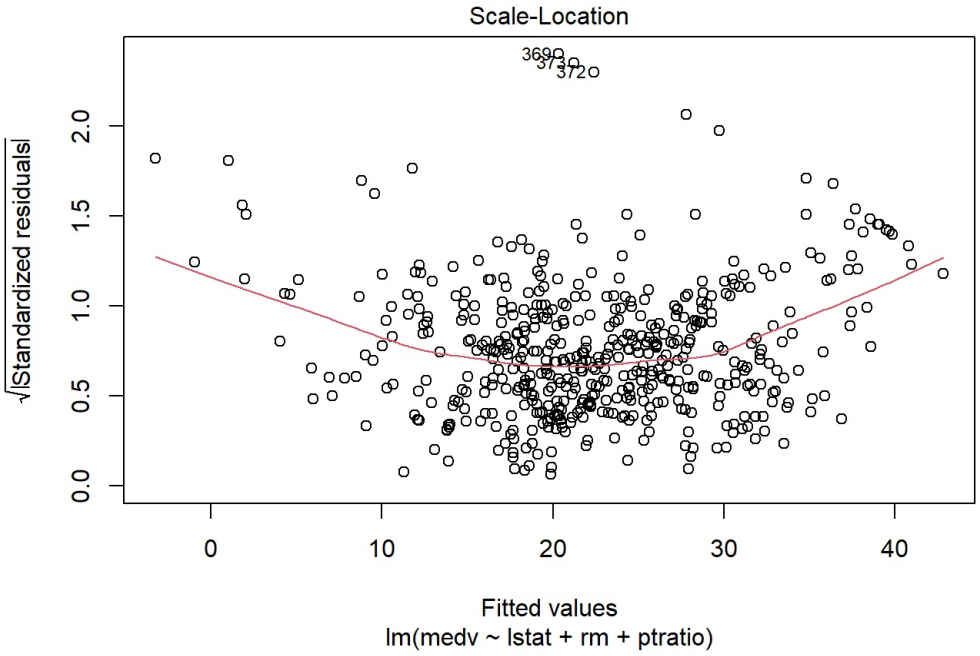

Fig. 6 Scale-Location Plot

The plot in Fig. 6 is used to check the assumption of equal variance (also called “homoscedasticity”) among the residuals in our regression model. If the red line is roughly horizontal across the plot, then the assumption of equal variance is likely met.

The red line displayed for our case is not exactly a straight horizontal line so we can perform Breusch-Pagan test for homoscedasticity.

```
bptest(lm.fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm.fit
## BP = 1.6223, df = 3, p-value = 0.6543
```

We can see that in Breusch-Pagan test p-value is 0.6543 which is greater than 0.05 that mean we have enough evidence to not regject the null hypothesis that means homoscedasticity is present (the residuals are distributed with equal variance).

```
#Create diagnostic plots
plot(lm.fit,which=5)
```

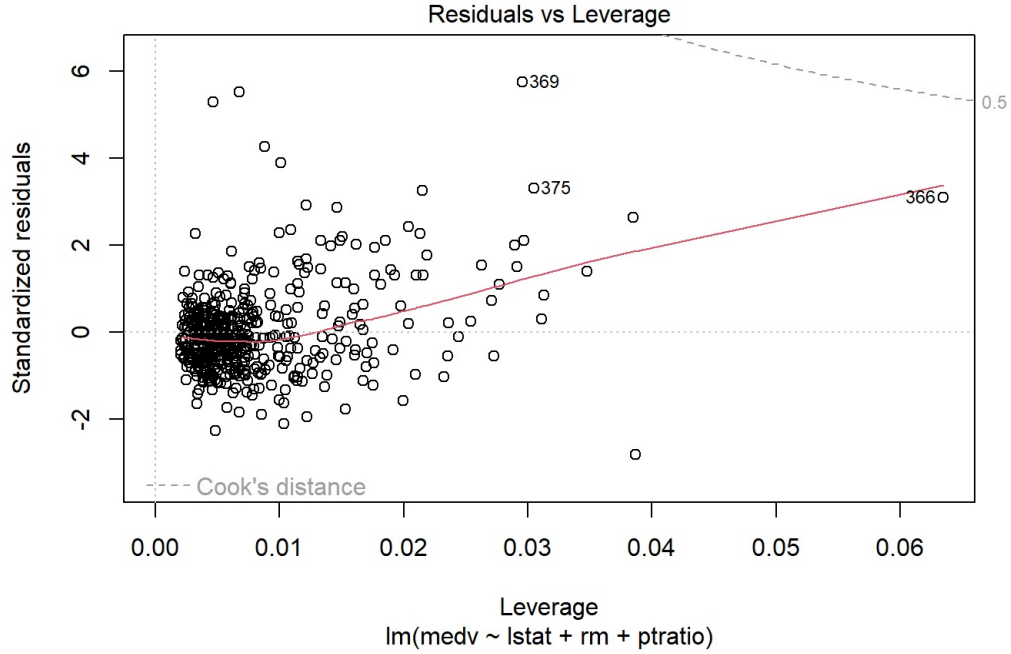

Fig. 7 Residuals vs Leverage Plot

The plot in Fig. 7 is used to identify influential observations. If any points in this plot fall outside of Cook's distance (the dashed lines) then it is an influential observation.

All observations are far within cook's distance and there are not many overly influential points within the data set.

As there is need to change the functional form of the model. So suppose the model is re-defind as follows:

Y = b0 + b1 X1 + b2 x1^2 + b3 X2 + b4 X3 + e

Now we will find the value of estimates of regression coefficients b0,b1,b2,b3.

```
y<-Boston$medv
x1<-Boston$lstat
x2<-Boston$ptratio
x3<-Boston$rm
x12<-x1^2
lm.fit2<-lm(y~x1+x12+x2+x3)
lm.fit2
```

```
##
## Call:
## lm(formula = y ~ x1 + x12 + x2 + x3)
##
## Coefficients:
## (Intercept)           x1          x12           x2           x3
##    25.78492     -1.64986      0.03202     -0.73084      3.87544
```

so we get,

Y_hat = 25.78492 - 1.64986 X1 + 0.03202 x1^2 - 0.73084 X2 + 3.87544 X3

```
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x12 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3166  -2.9506  -0.4864   2.2152  28.3357
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.784917   3.691814   6.984 9.14e-12 ***
## x1          -1.649855   0.121054 -13.629  < 2e-16 ***
## x12          0.032019   0.003404   9.406  < 2e-16 ***
## x2          -0.730838   0.110633  -6.606 1.01e-10 ***
## x3           3.875438   0.398850   9.717  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.826 on 501 degrees of freedom
## Multiple R-squared:  0.7269, Adjusted R-squared:  0.7247
## F-statistic: 333.3 on 4 and 501 DF,  p-value: < 2.2e-16
```

Conclusion

We have performed multiple linear regression on Boston dataset in Mass package in which we observed that lower status of the population (lstat), average number of rooms per dwelling (rm), pupil-teacher ratio by town (ptratio) have significant effect on MEDV. Also there was need to change the functional form of the model. So we re-defind the model.