ANEKANT EDUCATION SOCIETY'S

TULJARAM CHATURCHAND COLLEGE OF ARTS, COMMERCE AND SCIENCE, BARAMATI

DEPARTMENT OF STATISTICS

2022-2023

PROJECT ON TOPIC

"Regression Analysis On Boston Housing"

SUBMITTED BY

Student Name	Roll No:	Exam Seat No:
Beldar Prasad Dattatray	**16953**	**6404**
Zambare Samarth kanifnath	**16955**	**6406**
Mundlik Sagar Rajendra	**16984**	**6414**

UNDER THE GUIDENCE OF

Mrs. P.M.Mohite

ANEKANT EDUCATION SOCIETY'S

TULJARAM CHATURCHAND COLLEGE of Arts, Commerce and Science, BARAMATI

CERTIFICATE

DEPARTMENT OF STATISTICS

Date: 10 /04/2022

This is to certify that **Beldar Prasad Dattatray, Zambare Samarth kanifnath, Mundlik Sagar Rajendra** of Class M.Sc I (Statistics) has completed assigned project of the **"Regression analysis On Boston Housing"** As laid down for the academic year 2022-23.

Prof Mrs.P.M.Mohite
Project Guide

Dr. A. S. Jagtap
HOD of Statistics

INDEX

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude towards Principal Dr C.V. Murumkar and Head of Department of Statistics Dr. A. S. Jagtap for their valuable guidance.

I would like to thank Prof. Miss P.M. Mohite for the help provided during the course of completion of the project. I am also thankful to the Department of Statistics for providing all the necessary facilities, cooperation their valuable guidance.

I am also thankful to all those whose has directly or indirectly helped me during the completition of my project.

INTRODUCTION

The Boston Housing Dataset is a classic dataset that contains information about different houses in Boston. It was originally a part of UCI Machine Learning Repository and has been removed now. It can also be accessed from the scikit-learn library[1]. The dataset contains 506 observations and 14 variables[2]. Each of the 506 rows in the dataset describes a Boston suburb or town, and it has 14 columns with information such as average number of rooms per dwelling, pupil-teacher ratio, and per capita crime rate. The last row describes the median price of owner-occupied homes[3].

OBJECTIVES

A) To study the factor is depends on Median Price Oresponse.

B) To study the testing of hypothesis for significance of regression model.

C) To Predict the values using forward and backward elimination.

SOURCE of the Data

WWW.kaggle.com

STATISTICAL SOFTWARE

R Software

Statistical terms used:

1)Multiple Linear Regression:

Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line. the objective of multiple

regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.

2)Testing Of Hypothesis:

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters. Hypothesis testing is a systematic procedure for deciding whether the results of a research study support a particular theory which applies to a population. Hypothesis testing uses sample data to evaluate a hypothesis about a population.

3)Confidence Interval:

A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence. Confidence, in statistics, is another way to describe probability

4)Forward Selection Method:

Forward selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model

5)Backward Elimination Method:

Backward selection (or backward elimination), which starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.Backward stepwise selection (or backward elimination) is a variable selection method which: Begins with a model that contains all variables under consideration (called the Full Model) Then starts removing the least significant variables one after the other.

DATA ANALYSIS

Call:

```
model=lm(medv~.,data=Boston1_csv);model

Call:
lm(formula = medv ~ ., data = Boston1_csv)

Coefficients:
(Intercept)        ...1        crim          zn       indus
  36.461352   -0.002526   -0.108762    0.048031    0.019932
       chas         nox          rm         age         dis
   2.705245  -17.541602    3.839225   -0.001938   -1.493304
         rad         tax     ptratio       black       lstat
   0.324925   -0.011598   -0.947985    0.009357   -0.526184
```

Interpretation:

The multiple linear regression model is,

Y=36.461352-0.108762X1+0.048031X2+0.019932X3+2.705245X4-17.541602X5+3.839225X6-0.001938X7-1.493304X8+0.324925X9-0.011598X10-0.947985X11+0.009357X12-0.526184X13.

```
> summary(model)

Call:
lm(formula = medv ~ ., data = Boston1_csv)

Residuals:
    Min      1Q  Median      3Q     Max
-15.8948  -2.7585  -0.4663   1.7963  26.0911

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.461352   5.100994   7.148 3.21e-12 ***
...1         -0.002526   0.002080  -1.215 0.225046
crim         -0.108762   0.032855  -3.310 0.001000 **
zn            0.048031   0.013785   3.484 0.000538 ***
indus         0.019932   0.061468   0.324 0.745871
chas          2.705245   0.861298   3.141 0.001786 **
nox         -17.541602   3.822390  -4.589 5.66e-06 ***
rm            3.839225   0.418422   9.175  < 2e-16 ***
age          -0.001938   0.013380  -0.145 0.884866
dis          -1.493304   0.199892  -7.471 3.68e-13 ***
rad           0.324925   0.068111   4.771 2.43e-06 ***
```

```
tax         -0.011598   0.003807  -3.046 0.002443 **
ptratio     -0.947985   0.130822  -7.246 1.67e-12 ***
black        0.009357   0.002685   3.485 0.000536 ***
lstat       -0.526184   0.050704 -10.377  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.743 on 491 degrees of freedom
Multiple R-squared:  0.7414,  Adjusted R-squared:  0.734
F-statistic: 100.6 on 14 and 491 DF,  p-value: < 2.2e-16
```

Interpretation:

Residual standard error: 4.743 on 491 df
Multiple R-squared: 0.7414
Adjusted R-squared: 0.734
F-statistic: 100.6 on 14 and 491 DF, p-value: < 2.2e-16

anova(model)
Analysis of Variance Table

```
Response: medv
          Df  Sum Sq Mean Sq  F value    Pr(>F)
...1       1  2193.4  2193.4  97.5034 < 2.2e-16 ***
crim       1  4487.0  4487.0 199.4566 < 2.2e-16 ***
zn         1  3511.7  3511.7 156.1041 < 2.2e-16 ***
indus      1  2357.2  2357.2 104.7821 < 2.2e-16 ***
chas       1  1530.5  1530.5  68.0336 1.486e-15 ***
nox        1    82.9    82.9   3.6857 0.0554616 .
rm         1 10978.3 10978.3 488.0069 < 2.2e-16 ***
age        1   116.8   116.8   5.1932 0.0231043 *
dis        1  1802.5  1802.5  80.1228 < 2.2e-16 ***
rad        1     0.2     0.2   0.0077 0.9300875
tax        1   291.2   291.2  12.9438 0.0003535 ***
ptratio    1  1299.3  1299.3  57.7551 1.517e-13 ***
black      1   597.1   597.1  26.5433 3.739e-07 ***
lstat      1  2422.7  2422.7 107.6922 < 2.2e-16 ***
Residuals 491 11045.6    22.5
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Test For Significance Of Regression:
H0: Regression model is not significant. Vs
H1: Regression model is significant.

Interpretation:
Here, F value = 100.6 > p-value = 2.2e-16
Here F value is greater than p-value then we may fail to accept H0.
Therefore, Regression model is significant.

```
> confint(model,level=0.95)
                 2.5 %        97.5 %
(Intercept)  26.438881859  46.483821883
...1         -0.006612380   0.001559862
crim         -0.173315955  -0.044208718
zn            0.020946446   0.075115079
indus        -0.100840750   0.140705377
chas          1.012959947   4.397530727
nox         -25.051860585 -10.031343530
rm            3.017106426   4.661343703
age          -0.028227220   0.024350328
dis          -1.886053922  -1.100553855
rad           0.191101186   0.458749752
tax          -0.019078297  -0.004116764
ptratio      -1.205025818  -0.690944556
black         0.004081285   0.014631996
lstat        -0.625808130  -0.426559531
```

Interpretation:

The confidence interval for the intercept () is (26.438881859 46.483821883**).**

The confidence interval for the crim is (-0.173315955 -0.044208718 **).**

The confidence interval for the zn is (0.020946446 0.075115079**).**

The confidence interval for the indus is (-0.100840750 0.140705377**).**

The confidence interval for the chas is (1.012959947 4.397530727**)**

The confidence interval for the nox is (-25.051860585 -10.031343530 **)**

The confidence interval for the rm is (3.017106426 4.661343703 **)**

The confidence interval for the age is (-0.028227220 0.024350328**)**

The confidence interval for the dis is (-1.886053922 -1.100553855**)**

The confidence interval for the rad is (0.191101186 0.458749752 **)**

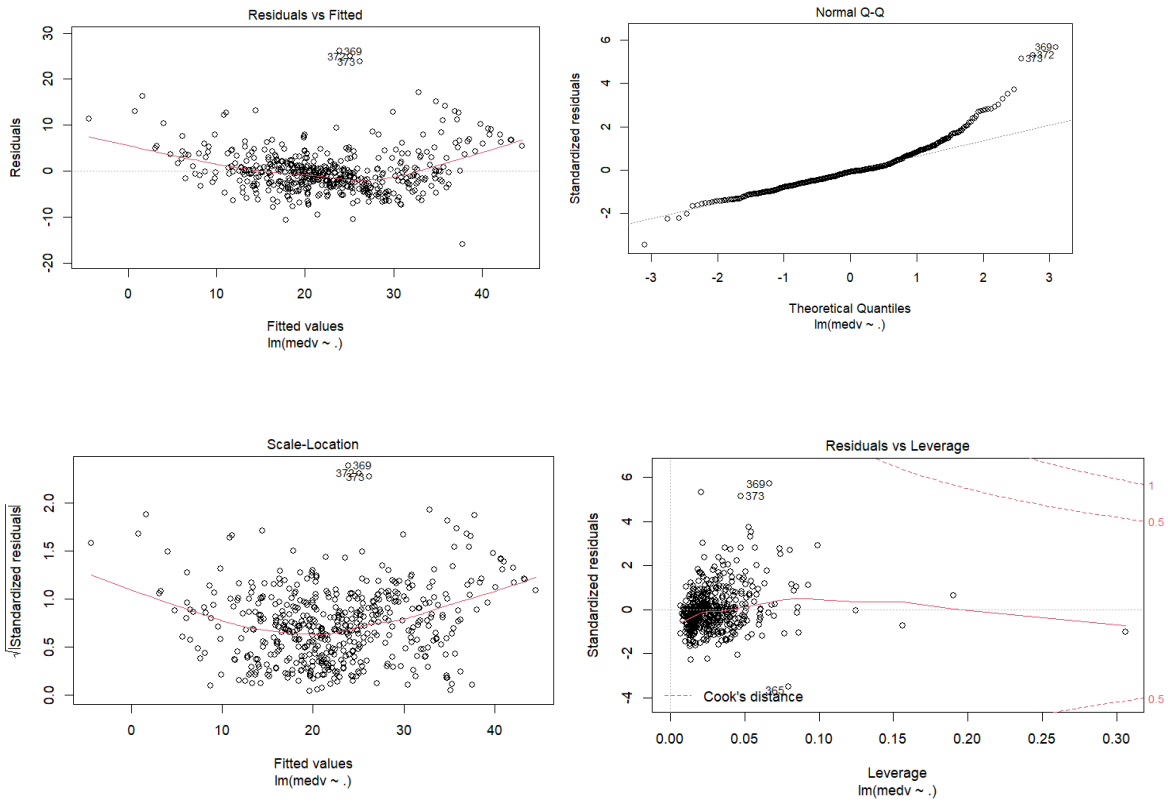
The confidence interval for the tax is (-0.019078297 -0.004116764**)**

The confidence interval for the ptratio is (-1.205025818 -0.690944556 **)**

The confidence interval for the black is (0.004081285 0.014631996 **)**

The confidence interval for the lstat is (-0.625808130 -0.426559531**)**

>Plot(model)



> forward=stepAIC(model,method=forward);forward
Start: AIC=1590.12
medv ~ ...1 + crim + zn + indus + chas + nox + rm + age + dis +
    rad + tax + ptratio + black + lstat

          Df Sum of Sq   RSS   AIC

```
- age      1      0.47 11046 1588.2
- indus    1      2.37 11048 1588.2
- ...1     1     33.20 11079 1589.6
<none>                 11046 1590.1
- tax      1    208.73 11254 1597.6
- chas     1    221.93 11268 1598.2
- crim     1    246.53 11292 1599.3
- zn       1    273.12 11319 1600.5
- black    1    273.20 11319 1600.5
- nox      1    473.78 11519 1609.4
- rad      1    511.97 11558 1611.0
- ptratio  1   1181.26 12227 1639.5
- dis      1   1255.48 12301 1642.6
- rm       1   1893.94 12940 1668.2
- lstat    1   2422.66 13468 1688.5

Step:  AIC=1588.15
medv ~ ...1 + crim + zn + indus + chas + nox + rm + dis + rad +
    tax + ptratio + black + lstat

          Df Sum of Sq    RSS    AIC
- indus    1      2.37 11048 1586.2
- ...1     1     32.79 11079 1587.7
<none>                 11046 1588.2
- tax      1    210.46 11256 1595.7
- chas     1    221.47 11268 1596.2
- crim     1    246.53 11293 1597.3
- black    1    272.88 11319 1598.5
- zn       1    278.41 11324 1598.7
- rad      1    513.75 11560 1609.2
- nox      1    519.51 11566 1609.4
- ptratio  1   1193.16 12239 1638.0
- dis      1   1362.40 12408 1645.0
- rm       1   1970.67 13017 1669.2
- lstat    1   2749.73 13796 1698.6

Step:  AIC=1586.25
medv ~ ...1 + crim + zn + chas + nox + rm + dis + rad + tax +
    ptratio + black + lstat

          Df Sum of Sq    RSS    AIC
- ...1     1     32.94 11081 1585.8
<none>                 11048 1586.2
- chas     1    228.66 11277 1594.6
- tax      1    236.05 11284 1595.0
- crim     1    248.68 11297 1595.5
```

```
- black    1    271.50 11320 1596.5
- zn       1    276.10 11324 1596.7
- rad      1    533.86 11582 1608.1
- nox      1    541.12 11590 1608.5
- ptratio  1   1199.77 12248 1636.4
- dis      1   1458.98 12507 1647.0
- rm       1   1975.19 13024 1667.5
- lstat    1   2754.60 13803 1696.9

Step:  AIC=1585.76
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    black + lstat

         Df Sum of Sq   RSS    AIC
<none>                 11081 1585.8
- chas     1    227.21 11309 1594.0
- crim     1    245.37 11327 1594.8
- zn       1    257.82 11339 1595.4
- black    1    270.82 11352 1596.0
- tax      1    273.62 11355 1596.1
- rad      1    500.92 11582 1606.1
- nox      1    541.91 11623 1607.9
- ptratio  1   1206.45 12288 1636.0
- dis      1   1448.94 12530 1645.9
- rm       1   1963.66 13045 1666.3
- lstat    1   2723.48 13805 1695.0

Call:
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
    tax + ptratio + black + lstat, data = Boston1_csv)

Coefficients:
(Intercept)         crim           zn         chas          nox
  36.341145    -0.108413     0.045845     2.718716   -17.376023
        rm          dis          rad          tax      ptratio
   3.801579    -1.492711     0.299608    -0.011778    -0.946525
     black        lstat
   0.009291    -0.522553
```

The Forward Selection linear regression model is,

Y=36.341145-0.108413X1+0.045845X2+2.718716X3-17.376023X4-3.801579X5-1.492711X6-0.299608X7-0.011778X8-0.946525X9-0.009291X10-0.522553X11

```
> library(MASS)
> backward=stepAIC(model,method=backward);backward
Start:  AIC=1590.12
medv ~ ...1 + crim + zn + indus + chas + nox + rm + age + dis +
    rad + tax + ptratio + black + lstat

          Df Sum of Sq   RSS    AIC
- age      1      0.47 11046 1588.2
- indus    1      2.37 11048 1588.2
- ...1     1     33.20 11079 1589.6
<none>                 11046 1590.1
- tax      1    208.73 11254 1597.6
- chas     1    221.93 11268 1598.2
- crim     1    246.53 11292 1599.3
- zn       1    273.12 11319 1600.5
- black    1    273.20 11319 1600.5
- nox      1    473.78 11519 1609.4
- rad      1    511.97 11558 1611.0
- ptratio  1   1181.26 12227 1639.5
- dis      1   1255.48 12301 1642.6
- rm       1   1893.94 12940 1668.2
- lstat    1   2422.66 13468 1688.5

Step:  AIC=1588.15
medv ~ ...1 + crim + zn + indus + chas + nox + rm + dis + rad +
    tax + ptratio + black + lstat

          Df Sum of Sq   RSS    AIC
- indus    1      2.37 11048 1586.2
- ...1     1     32.79 11079 1587.7
<none>                 11046 1588.2
- tax      1    210.46 11256 1595.7
- chas     1    221.47 11268 1596.2
- crim     1    246.53 11293 1597.3
- black    1    272.88 11319 1598.5
- zn       1    278.41 11324 1598.7
- rad      1    513.75 11560 1609.2
- nox      1    519.51 11566 1609.4
- ptratio  1   1193.16 12239 1638.0
- dis      1   1362.40 12408 1645.0
- rm       1   1970.67 13017 1669.2
- lstat    1   2749.73 13796 1698.6

Step:  AIC=1586.25
medv ~ ...1 + crim + zn + chas + nox + rm + dis + rad + tax +
    ptratio + black + lstat

          Df Sum of Sq   RSS    AIC
```

```
- ...1     1     32.94 11081 1585.8
<none>                 11048 1586.2
- chas     1    228.66 11277 1594.6
- tax      1    236.05 11284 1595.0
- crim     1    248.68 11297 1595.5
- black    1    271.50 11320 1596.5
- zn       1    276.10 11324 1596.7
- rad      1    533.86 11582 1608.1
- nox      1    541.12 11590 1608.5
- ptratio  1   1199.77 12248 1636.4
- dis      1   1458.98 12507 1647.0
- rm       1   1975.19 13024 1667.5
- lstat    1   2754.60 13803 1696.9

Step:  AIC=1585.76
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    black + lstat

         Df Sum of Sq   RSS    AIC
<none>                 11081 1585.8
- chas     1    227.21 11309 1594.0
- crim     1    245.37 11327 1594.8
- zn       1    257.82 11339 1595.4
- black    1    270.82 11352 1596.0
- tax      1    273.62 11355 1596.1
- rad      1    500.92 11582 1606.1
- nox      1    541.91 11623 1607.9
- ptratio  1   1206.45 12288 1636.0
- dis      1   1448.94 12530 1645.9
- rm       1   1963.66 13045 1666.3
- lstat    1   2723.48 13805 1695.0

Call:
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
    tax + ptratio + black + lstat, data = Boston1_csv)

Coefficients:
(Intercept)         crim           zn         chas
  36.341145    -0.108413     0.045845     2.718716
       nox           rm          dis          rad
 -17.376023     3.801579    -1.492711     0.299608
       tax      ptratio        black        lstat
  -0.011778    -0.946525     0.009291    -0.522553
```

The Backward Elimination linear regression model is,

Y=36.341145-0.108413X1+0.045845X2+2.718716X3-17.376023X4+3.801579X5-1.492711X6-0.299608X7-0.011778X8-0.946525X9+0.009291X10-0.522553X1.

CONCLUSIONS

1) We conclude that all factors are depend on medv except Age And Indus
2) We conclude that the regression model is significant.
3) We Did Forward Selection and backward Elimination Method And we Got A Model

REFERENCE

A)Montogomery

B)Google