CSC 423 Final Report: Statistical Modeling and Analysis Results for the Factors Affecting Home Values in Boston

Submitted to:
Prof. Bill Qualls

Report Prepared By:
Sriram Yarlagadda
Sarah Cummings
Haifa Alsunaid

November 16, 2015

Table of Contents

Executive Summary

This report will demonstrate the journey of searching for a useful regression model for Boston Housing Data from 1978, which has been obtained from the UCI Machine Learning Repository. This dataset contained one dependent, the median $ value of owner occupied homes in 10,000 (MEDV), and 13 independent variables, of which one is qualitative and 13 are quantitative. This dataset has 506 instances, with a satisfactory number of observations for each variable in order to yield an accurate result for our analysis. Throughout this journey, R Studio was used as our statistical tool. The aim of this report was to find which variables have the greatest effect on median house value of towns surrounding Boston, and how exactly those variables affect (MEDV) of those homes.

In this report, the dataset has been refined and the analysis of underlying assumptions has been conducted to achieve reliable estimates of the coefficients. Pairwise plots between the variables showed we had no need for interaction terms, however they did reveal that five variables that were in need of transformation. We also found five influential points through calculating Hat Values, Cook'S Distance, and Studentized Residuals. These points have been removed from our data to avoid the improper data fitting.

With R studio, the Stepwise model selection method with AIC criterion was used create a linear model for our data. The following variables were significant predictors of median house values in the suburbs of Boston with 95% estimate's accuracy and 5% margin error:

Percent lower status of population, average number of rooms per dwelling, pupil teacher ratio, distance to employment centers, percentage of African Americans, index of accessibility to radial highways, full value property tax rate per $10,000, nitric oxides concentration (parts per 10 million), Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), proportion of residential land zoned for lots over 25,000 sq.ft, and per capita crime rate by town.

Our model also had a statistically significant F test, and an adj-r squared value of 0.87.

1.0 Introduction and Data

As young people living in Chicago, we are constantly reminded of just how expensive it is to live in a big city with so much to offer. Chicago, however, is not the only city with people who face a high cost of living. While searching for data, we were drawn to a dataset that aims to identify the median values of homes in different areas in the suburbs of Boston. Our file, Boston Housing Data from 1978, was obtained from the UCI Machine Learning Repository. The data consists of 14 attributes and 506 cases. The dependent variable is the median $ value of owner occupied homes (MEDV) and the rest are independent variables, of which one is qualitative and eleven are quantitative. The variables are listed below:

```
1. CRIM      per capita crime rate by town\
2. ZN        proportion of residential land zoned for lots over
             25,000 sq.ft.\
3. INDUS     proportion of non-retail business acres per town\
4. CHAS      Charles River dummy variable (= 1 if tract bounds
             river; 0 otherwise)\
5. NOX       nitric oxides concentration (parts per 10 million)\
6. RM        average number of rooms per dwelling\
7. AGE       proportion of owner-occupied units built prior to 1940\
8. DIS       weighted distances to five Boston employment centres\
9. RAD       index of accessibility to radial highways\
10. TAX      full-value property-tax rate per $10,000\
11. PTRATIO  pupil-teacher ratio by town\
12. B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks
             by town\
13. LSTAT    % lower status of the population\
14. MEDV     Median value of owner-occupied homes in $1000's (dependent variable)
```

1.1 Research Question and Hypotheses

Our research questions are : How do the variables provided in the dataset affect the median value of homes in Boston towns? Which variables affect the median value of homes the most? We aim to create a linear model that will answer these research questions. Given the variables provided in the dataset, here are some of our hypotheses:

- Per capita crime rate and median house value would have a negative relationship
- A lower pupil teacher ratio would be associated with higher median house value.
- A high proportion of nonretail business acres per town will have a negative effect on the median value of the surrounding residential areas.
- Percentage of the population with lower status will have negative relationship with the median house value.

- An increase in nitric oxides concentration will have adverse effect on the median house value.
- A high proportion of residential land zoned for lots over 25,000 sq.ft will increase the median value of homes.
- The average number of rooms per dwelling will have a positibe relationship with the median house value.
- A high full value property tax rate per $10,000 will decrease the median house value.

Please refer to the appendix for loading the data and creating the dummy variables.

2.0 Correlation Check

After cleansing, loading in our data, and creating dummy variables for our categorical variable- CHAS, we examine the independent variables. We calculated the VIF values for each of the predictors to see if any of them are strongly correlated to each other. The presence of a strong correlation could result in multicollinearity, which could cause unreliable estimates of the coefficients. See the appendix for this code. The VIF values:

```
##     CRIM       ZN    INDUS      NOX       RM      AGE      DIS      TAX
## 1.663648 2.272992 3.660714 4.294324 1.880883 3.077311 3.953729 3.403205
##  PTRATIO        B    LSTAT
## 1.725085 1.338875 2.928554
```

Since all of the predictors have a VIF values less than 5, there is no significant correlation between the predictors and we do not expect any multicollinearity in our model.

2.1 Scatterplots and Necessary Variable Transformations

In order to ascertain if any of the variables need to be transformed, we create pairwise plots comparing each of the independent variables with the dependent variable. Figure A below shows the variables that have a non-linear relationship with MEDV. We perform the necessary transformations below, and plot these transformed variables with MEDV in FIgure B.

- CRIM_inv = 1/CRIM
- INDUS_inv = 1/INDUS
- LSTAT_ln = log(LSTAT)
- NOX_inv = 1/NOX
- DIS_ln = log(DIS)

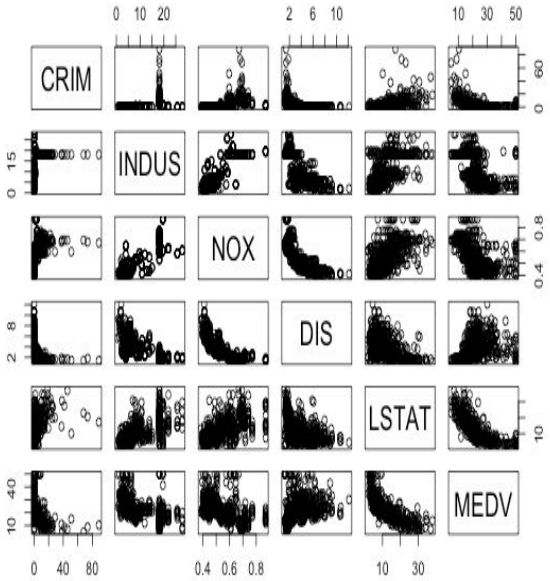


Figure A: variables vs. MEDV before transformation

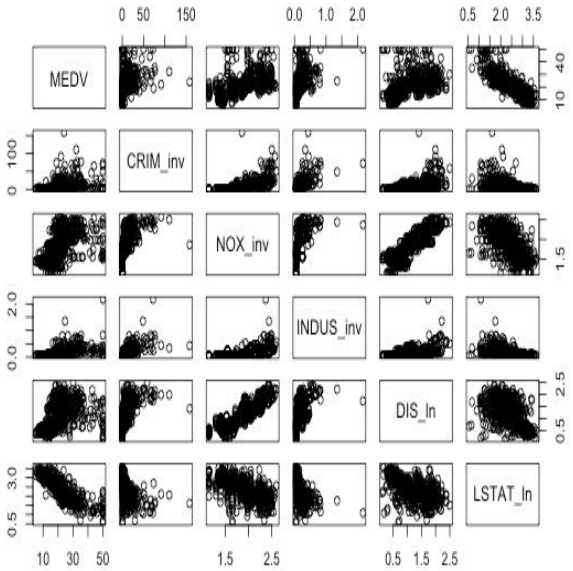


Figure B: variables vs. MEDV after transformation

Figure B shows that the transformed variables have a linear relationship with the dependent variables, hence the transformations proved effective.

2.2 Influential Points Detection

We next examine our data for influential points and remove such points to prevents inappropriate data-fitting by the linear model. We calculate the Studentized Residuals, Hat Values, and Cook's Distance for each point, looking for points with Hat Values >0.5, Cook'S Distance >1, or deleted Studentized Residuals >3 or <-3. We ascertain the following five influential points, and remove them from our data:

```
InfluPoints

##     StudentizedResids    HatVals cooks.distance.fit.
## 215          3.220684 0.05619811          0.04329258
## 365         -3.765055 0.07489994          0.07984176
## 369          4.413143 0.10061729          0.14999819
## 372          5.722165 0.02516752          0.05672177
## 373          5.412893 0.05055668          0.10537850
```

3.0 Linear Model Selection

Now that we have fully cleaned our data and transformed our variables, use our software to select a final model. We use the stepwise model selection with the AIC criterion to form a model. The final step is shown below:

```
## Step:  AIC=1345.91
## MEDV ~ LSTAT_ln + RM + PTRATIO + DIS_ln + B + NOX_inv + CHAS +
##     TAX + RAD + INDUS_inv + CRIM_inv
##
##             Df Sum of Sq     RSS    AIC
## <none>                    7010.5 1345.9
## + AGE        1      11.1  6999.4 1347.1
## + ZN         1       5.8  7004.7 1347.5
## - CRIM_inv   1     112.7  7123.2 1351.9
## - INDUS_inv  1     166.4  7176.9 1355.7
## - CHAS       1     168.4  7178.9 1355.8
## - RAD        1     176.9  7187.3 1356.4
## - NOX_inv    1     192.6  7203.1 1357.5
## - B          1     268.9  7279.4 1362.8
## - TAX        1     294.8  7305.3 1364.5
## - PTRATIO    1     741.7  7752.1 1394.3
## - DIS_ln     1     781.8  7792.3 1396.9
## - RM         1    1546.9  8557.4 1443.8
## - LSTAT_ln   1    3831.9 10842.3 1562.4
```

```
## Call:
## lm(formula = MEDV ~ LSTAT_ln + RM + PTRATIO + DIS_ln + B + NOX_inv +
##     CHAS + TAX + RAD + INDUS_inv + CRIM_inv, data = mydata)
##
## Coefficients:
## (Intercept)     LSTAT_ln           RM      PTRATIO       DIS_ln
##   29.235053    -8.046390     3.638702    -0.697466    -4.942906
##           B      NOX_inv        CHAS1          TAX          RAD
##    0.009171     4.048446     2.403950    -0.012047     0.172202
##   INDUS_inv     CRIM_inv
##    4.651693    -0.040870
```

As seen in our attached R output for stepwise selection method, our final model is:

y = 29.235 - 8.05 **LSTAT_ln +** 3.639 **RM** - 0.697 **PTRATIO** - 4.943 **DIS_ln +** 0.009 **B +** 0.172 **RAD** - 0.012 **TAX +** 4.048 **NOX_inv +** 2.404 **CHAS +** 4.652 **INDUS_inv -** 0.040 **CRIM_inv.**

Only the AGE and ZN variables have been left out.

From the model, we derive the following interpretations:

- For every unit of lower status population, we predict the median house value to decrease approximately 8.05 thousand dollars, holding all else constant.
- For every additional average room per dwelling, we predict the median house value will increase 3.639 thousand dollars, holding all else constant.
- For unit of pupil teacher ratio unit, we predict the median house value will decrease 0.697 thousand dollars, holding all else constant.
- For each additional weighted log distance unit to employment centers, we predict the median house value will decrease 0.697 thousand dollars, holding all else constant.
- For every additional proportion unit of blacks per town, we predict the median house value will increase 0.009 thousand dollars, holding all else constant.
- For each additional distance unit to the highways, we predict the median house value will increase 0.172 thousand dollars, holding all else constant.

- For every additional unit of tax rate, we predict the median house value will decrease 0.012 thousand dollars, holding all else constant.
- For every additional parts per 10 million of Nitric oxides concentration, we predict the median house value will increase 4.048 thousand dollars, holding all else constant.
- We predict towns with tracks that bound the river to have a median house value 2.404 thousand dollars more than towns that don't, holding all else constant.
- For every additional inverse acre of non-retail business per town, we predict the median house value will increase 4.652 thousand dollars, holding all else constant.
- For every additional unit of inverse per capita crime rate, we predict the median house value will decrease 0.04 thousand dollars, holding all else constant.

3.1 Coefficient Estimate Confidence Intervals

We are 95% confident that the true value of each coefficient estimate (the Median $ value of owner occupied homes (MEDV)) falls within the corresponding intervals listed to the right:

```
##                    2.5 %      97.5 %
## (Intercept) 20.948429226 37.52167719
## LSTAT_ln    -9.013419129 -7.07936082
## RM           2.950439368  4.32696480
## PTRATIO     -0.887993896 -0.50693738
## DIS_ln      -6.258060551 -3.62775131
## B            0.005010521  0.01333230
## RAD          0.075877285  0.26852641
## TAX         -0.017266834 -0.00682748
## NOX_inv      1.878181765  6.21871027
## CHAS1        1.025912517  3.78198770
## INDUS_inv    1.968916413  7.33446948
## CRIM_inv    -0.069507904 -0.01223192
```

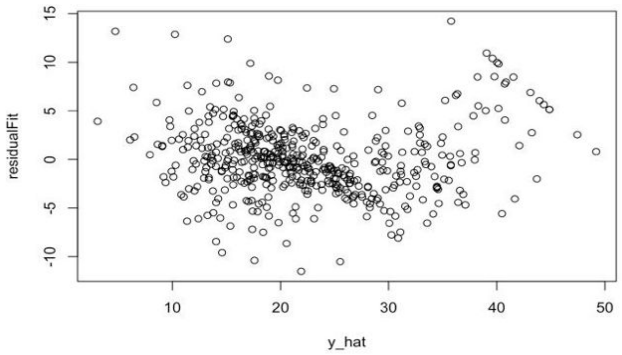
3.2 Model Evaluation

Next, we evaluate this model by verifying its adj-*R-sq* value, t-tests and F-test results-- please refer to the appendix for the code and output.

As seen in the output provided, our F statistic (212), has a significant p-value < 2.2 x10^-16 . This means we can reject the null hypothesis that $B_1 = B_2 = \ldots B_{12} = B_{13} = 0$ and conclude that our model is a useful predictor of median house values. We also find an adjusted R-squared of 0.827, which means that 82.7% of the variability in our data is described by our model. Additionally, all of our coefficients have a significant t-test p-value. Furthermore, the model has an MSE of 3.786 so we conclude that this model itself is useful in predicting the median value of owner occupied home.

3.3 Residual Analysis and Checking Assumptions:

We plot and examine a residual plot of this model to verify that the residuals are homoscedastic, are showing no inherent trends, and have no outliers.



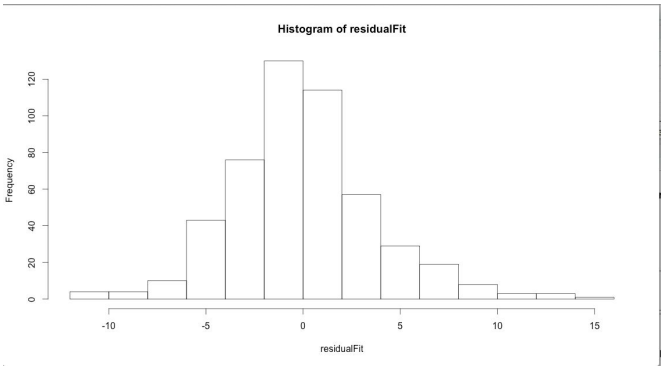


Figure C: Residual Plot

Figure D: Residual Histogram

As we can see in the residual plot above, the variance is relatively stable and homoscedastic. In addition, there doesn't appear to any noticeable pattern. Partial residual plots have also been plotted for this model to verify this. Regarding the outliers, we have an MSE = 3.8 for this data. Looking at the residual plot, we find that majority (>95%) of the residuals are between +/-2*MSE (-7.6 and +7.6). And very few points are beyond the +/-3*MSE mark (-11.4 and +11.4). As seen in the histogram above, our residuals are also normally distributed with mean equal to zero.

4.0 Conclusion

In conclusion, we refined our data, created five transformed variables, removed five influential points, and checked all necessary conditions for regression analysis. We were able to find a useful linear model to predict median house values of towns surrounding boston by using the stepwise selection method with AIC criteria in RStudio. The linear model we have constructed has a satisfactory residual plot without any certain tendency and seems to satisfy all the underlying assumptions as well as the regression model assumptions.

With our analysis we were able to answer our research questions. We found that the following variables were significant predictors of median house values in the suburbs of Boston with 95% estimate's accuracy and 5% margin error:

Percent lower status of population, average number of rooms per dwelling, pupil teacher ratio, distance to employment centers, percentage of African Americans, index of accessibility to radial highways, full value property tax rate per $10,000, nitric oxides concentration (parts per 10 million), Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), proportion of residential land zoned for lots over 25,000 sq.ft, and per capita crime rate by town.

Appendix:

1. Loading and Cleaning Data:

Loading the data:

```
setwd("C:/Users/syarlag1/Desktop/RRepo")
mydata = read.table("./housing.data.txt") #load data
colnames(mydata) = c("CRIM","ZN","INDUS","CHAS","NOX","RM","AGE","DIS","RAD",
"TAX","PTRATIO","B","LSTAT","MEDV")#add colnames
```

Next based on the description of the data variables, we create dummy variables for the categorical variable - CHAS.

```
mydata$CHAS = factor(mydata$CHAS)
```

Here is a quick summary of all the variables in our data:

```
str(mydata)

## 'data.frame':    506 obs. of  14 variables:
##  $ CRIM   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ ZN     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ INDUS  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ CHAS   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ NOX    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.5
24 ...
##  $ RM     : num  6.58 6.42 7.18 7 7.15 ...
##  $ AGE    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ DIS    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ RAD    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ TAX    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ PTRATIO: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ B      : num  397 397 393 395 397 ...
##  $ LSTAT  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ MEDV   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

2. Correlation check:

```
library(car)

## Warning: package 'car' was built under R version 3.2.2

fit = lm(MEDV~., data = mydata[,-c(4,9)])#not including the factor variables
vif(fit) #non of the vars have VIF>6, hence should be no issue of multicollinearity

##     CRIM       ZN    INDUS      NOX       RM      AGE      DIS      TAX
## 1.663648 2.272992 3.660714 4.294324 1.880883 3.077311 3.953729 3.403205
##  PTRATIO        B    LSTAT
## 1.725085 1.338875 2.928554
```

3. Scatter plots and Variable Transformations

To create our initial scatter plots for all the variables versus MEDV:

```
pairs(mydata[,c(1:4,14)])
```

```
pairs(mydata[,c(5:9,14)])
```

```
pairs(mydata[,c(10:14)])
```

To transform the necessary variables we used the below codes:

```
mydata$CRIM_inv = 1/mydata$CRIM
mydata$NOX_inv = 1/mydata$NOX
mydata$INDUS_inv = 1/mydata$INDUS
mydata$DIS_ln = log(mydata$DIS)
mydata$LSTAT_ln = log(mydata$LSTAT)

mydata = mydata[,-c(1,3,5,8,13)] #replacing existing variables with their functions
```

To create scatterplots of our new transformed variables versus MEDV:

```
pairs(mydata[,9:14])
```

4. Influential Points Detection- Code:

```
fit = lm(MEDV~.,data=mydata)
InfluTable = data.frame(StudentizedResids = rstudent(fit), HatVals = hatvalues(fit),
cooks.distance(fit))
InfluTable
```

```
##     StudentizedResids     HatVals cooks.distance.fit.
## 1       -0.7225325507 0.328804895        1.828514e-02
## 2       -0.4747587086 0.025159310        4.161666e-04
## 3        0.4267045137 0.027606182        3.698394e-04
## 4       -0.5926516380 0.032339223        8.395575e-04
## 5        1.1010538401 0.021528855        1.904472e-03
## 6       -0.1063015489 0.028135251        2.341368e-05
## 7        0.6936552945 0.018598067        6.519864e-04
## 8        2.5255798287 0.034547903        1.612736e-02
## 9        1.3358372708 0.044022392        5.860209e-03
## 10       0.5979145239 0.033030636        8.734175e-04
## 11      -0.2920421653 0.038309846        2.431345e-04
## 12      -0.0337813875 0.028249101        2.374429e-06
```

This code gave us the five influential which are in the report-- section 2.2

```
InfluPoints = data.frame()
for(i in 1:506){
  if(InfluTable[i,1] > 3 | InfluTable[i,1]< -3 | InfluTable[i,2] > 0.5 | InfluTable[i
,3] > 1){
    InfluPoints <- rbind(InfluPoints, InfluTable[i,])
  }
}
InfluPoints
```

```
##     StudentizedResids    HatVals cooks.distance.fit.
## 215          3.220684 0.05619811          0.04329258
## 365         -3.765055 0.07489994          0.07984176
## 369          4.413143 0.10061729          0.14999819
## 372          5.722165 0.02516752          0.05672177
## 373          5.412893 0.05055668          0.10537850
```

We remove these points from our dataset

```
mydata = mydata[-as.numeric(rownames(InfluPoints)),]
```

5. **Model Selection-** Stepwise Selection Code

```
full = lm(MEDV~., data=mydata)
null = lm(MEDV~1, data=mydata)
step(null, scope = list(lower = null, upper = full), direction = "both")
```

```
## Start:  AIC=2201.84
## MEDV ~ 1
##
##              Df Sum of Sq   RSS    AIC
## + LSTAT_ln    1   27882.6 12555 1617.8
## + RM          1   22408.5 18029 1799.2
## + PTRATIO     1   11705.1 28732 2032.6
## + TAX         1   10698.1 29739 2049.9
## + INDUS_inv   1    8686.5 31751 2082.7
## + NOX_inv     1    8477.1 31960 2086.0
## + RAD         1    7380.1 33057 2102.9
## + AGE         1    6734.5 33703 2112.6
## + ZN          1    5837.5 34600 2125.7
## + B           1    4715.0 35722 2141.7
## + DIS_ln      1    4623.5 35814 2143.0
## + CRIM_inv    1    4307.8 36130 2147.4
## + CHAS        1    1114.8 39323 2189.8
## <none>                    40437 2201.8
```

The final step in the stepwise selection method

```
## Step:  AIC=1345.91
## MEDV ~ LSTAT_ln + RM + PTRATIO + DIS_ln + B + NOX_inv + CHAS +
##     TAX + RAD + INDUS_inv + CRIM_inv
##
##              Df Sum of Sq     RSS    AIC
## <none>                     7010.5 1345.9
## + AGE         1      11.1  6999.4 1347.1
## + ZN          1       5.8  7004.7 1347.5
## - CRIM_inv    1     112.7  7123.2 1351.9
## - INDUS_inv   1     166.4  7176.9 1355.7
## - CHAS        1     168.4  7178.9 1355.8
## - RAD         1     176.9  7187.3 1356.4
```

```
## Call:
## lm(formula = MEDV ~ LSTAT_ln + RM + PTRATIO + DIS_ln + B + NOX_inv +
##     CHAS + TAX + RAD + INDUS_inv + CRIM_inv, data = mydata)
##
## Coefficients:
## (Intercept)     LSTAT_ln           RM      PTRATIO       DIS_ln
##   29.235053    -8.046390     3.638702    -0.697466    -4.942906
##           B      NOX_inv        CHAS1          TAX          RAD
##    0.009171     4.048446     2.403950    -0.012047     0.172202
##   INDUS_inv     CRIM_inv
##    4.651693    -0.040870
```

6. **Model Evaluation-** Code and Output

```
fit = lm(MEDV ~ LSTAT_ln + RM + PTRATIO + DIS_ln + B + RAD + TAX + NOX_inv + CHAS + I
NDUS_inv + CRIM_inv, data=mydata)
summary(fit)

##
## Call:
## lm(formula = MEDV ~ LSTAT_ln + RM + PTRATIO + DIS_ln + B + RAD +
##     TAX + NOX_inv + CHAS + INDUS_inv + CRIM_inv, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.523  -2.268  -0.290   1.879  14.015
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.235053   4.217483   6.932 1.32e-11 ***
## LSTAT_ln    -8.046390   0.492170 -16.349  < 2e-16 ***
## RM           3.638702   0.350292  10.388  < 2e-16 ***
## PTRATIO     -0.697466   0.096969  -7.193 2.40e-12 ***
## DIS_ln      -4.942906   0.669349  -7.385 6.64e-13 ***
## B            0.009171   0.002118   4.331 1.80e-05 ***
```

```
## RAD          0.172202   0.049024   3.513 0.000485 ***
## TAX         -0.012047   0.002657  -4.535 7.26e-06 ***
## NOX_inv      4.048446   1.104557   3.665 0.000274 ***
## CHAS1        2.403950   0.701353   3.428 0.000660 ***
## INDUS_inv    4.651693   1.365401   3.407 0.000711 ***
## CRIM_inv    -0.040870   0.014575  -2.804 0.005247 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.786 on 489 degrees of freedom
## Multiple R-squared:  0.8266, Adjusted R-squared:  0.8227
## F-statistic:   212 on 11 and 489 DF,  p-value: < 2.2e-16
```

7. **Residual Analysis-** Code

```
residualFit = resid(fit)
y_hat = predict(fit, newdata = data.frame(mydata))
plot(y_hat, residualFit)
```

8. **Coefficient Confidence Intervals -**95% confidence interval code and output

```
confint(fit)
```

```
##                    2.5 %      97.5 %
## (Intercept) 20.948429226 37.52167719
## LSTAT_ln    -9.013419129 -7.07936082
## RM           2.950439368  4.32696480
## PTRATIO     -0.887993896 -0.50693738
## DIS_ln      -6.258060551 -3.62775131
## B            0.005010521  0.01333230
## RAD          0.075877285  0.26852641
## TAX         -0.017266834 -0.00682748
## NOX_inv      1.878181765  6.21871027
## CHAS1        1.025912517  3.78198770
## INDUS_inv    1.968916413  7.33446948
## CRIM_inv    -0.069507904 -0.01223192
```