# Project 2

# Analysis of Boston Housing Prices Using Regression Modeling

**Submitted by**

**Md Salman Rahman**

**Graduate Student**

**University of Texas Rio Grande Valley**

**Submitted to**

Dr. George Yanev

December 6, 2021

# Contents

# 1 Introduction

In this study, I will explore some research question such as how the prices of the house in Boston related to other feature described in Table 1. The main objective is to develop a regression model for Boston housing data-set. I will estimate the coefficient of explanatory variable from Boston housing data set and confidence interval will also available for different regression model. Anova analysis is also done for different regression model. This are my primary research question which I will explore using statistical analysis and regression approach.

Regression modeling generally leads to a functional relationship between the response and a set of explanatory variable. In our study, MEDV(median price value of owner-occupied homes in $1000) is the response variable and other variable listed in Table 1 are the explanatory variable.

# 2 Material and Methods

## 2.1 Data Description

The data set is about Boston housing information and it contains total 506 cases (https://www.cs.toronto.edu/ delve/data/boston/bostonDetail.html). There are 13 attributes in each case of the data set. Exploratory data analysis is performed on the data set. I draw several diagram such as histogram, pie chat, box-plot along with some descriptive analysis and use them for visualization and better understanding of the relationship of the target variable with explanatory variable (Available in appendix).

First, five head of the data-set are shown in the figure 1 for better visualization about the data-sets.

Table 1: Description of Attribute of the Data sets

| Variable Name | Description |
|---|---|
| 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 |
| LSTAT | % lower status of the population |
| MEDV (regressor) | Median value of owner-occupied homes in $1000's |

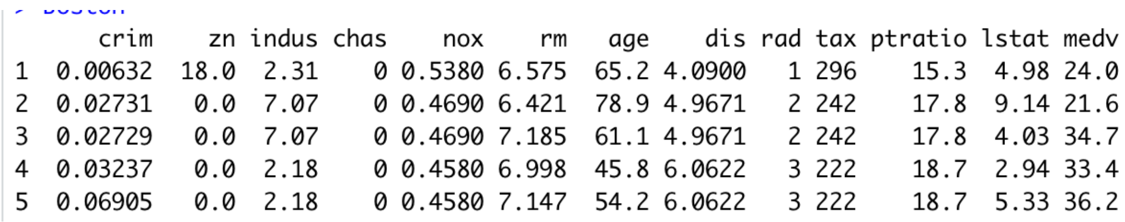
## 2.2 Regression Model

### 2.2.1 First Model

In this study Median value of owner-occupied homes in $1000's(MEDV) is our target/outcome variable which indicates the price of the house in Boston. And, other 12 are the predictor variable for the regression. I can simply represent them using the following equation:

$$\begin{aligned} MEDV = \beta_0 + \beta_1 * CRIM + \beta_2 * ZN + \beta_3 * INDUS + \beta_4 * CHAS \\ + \beta_5 * NOX + \beta_6 * RM + \beta_7 * AGE + \beta_8 * DIS + \beta_9 * RAD \\ + \beta_{10} * TAX + \beta_{11} * PTRATIO + \beta_{12} * LSTAT + \beta_{13} * BLACK + \epsilon \quad (1) \end{aligned}$$

Here I am developing my first multiple linear regression model with 12 predictor



```
> Boston
       crim   zn indus chas    nox    rm  age    dis rad tax ptratio lstat medv
1 0.00632 18.0  2.31    0 0.5380 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
2 0.02731  0.0  7.07    0 0.4690 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
3 0.02729  0.0  7.07    0 0.4690 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
4 0.03237  0.0  2.18    0 0.4580 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
5 0.06905  0.0  2.18    0 0.4580 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
```

Figure 1: First 5 value of the entire data-sets

variable and 1 response variable for Boston housing dataset. $\beta_1$ to $\beta_{12}$ are the coefficient/parameters of the predictors and $\epsilon$ indicates the error term of the model. I deploy the regression model using R software.

### 2.2.2 Second Model

In the second model, I use some interaction term between the variable such as ZN (proportion of residential land zoned for lots over 25,000 sq.ft) and TAX (full-value property-tax rate per \$10,000). Also, interaction between DIS(weighted distances to five Boston employment centres) and RAD(index of accessibility to radial highways) is also considered.

$$\begin{aligned} MEDV = \beta_0 + \beta_1 * CRIM + \beta_2 * ZN + \beta_3 * INDUS + \beta_4 * CHAS \\ + \beta_5 * NOX + \beta_6 * RM + \beta_7 * AGE + \beta_8 * DIS + \beta_9 * RAD \\ + \beta_{10} * TAX + \beta_{11} * PTRATIO + \beta_{12} * LSTAT + \beta_{13} * BLACK + \beta_{14} * ZN * TAX + \beta_{15} * DIS * RAD + \epsilon \end{aligned} \tag{2}$$

### 2.2.3 Third Model

After getting the results from the first and second model, I analyze the model considering the several important measure such as checking the assumption made for specifying the regression model, residual plot, and p value of the coefficient. Upon checking, I develop a new regression model excluding some explanatory variable such as (age and rad) which have negligible influence on the model. Details about the model diagnostic is described in the result section.

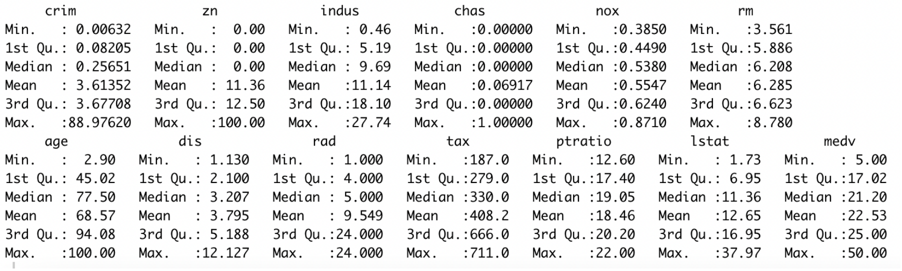
$$MEDV = \beta_0 + \beta_1 * CRIM + \beta_2 * ZN + \beta_3 * INDUS + \beta_4 * CHAS$$
$$+\beta_5 * NOX + \beta_6 * RM + \beta_7 * DIS + \beta_8 * TAX + \beta_9 * PTRATIO + \beta_{10} * LSTAT + \beta_{11} * BLACK + \epsilon \tag{3}$$

From this three model based on certain condition such as checking the assumption, examining the residual plot, and correlation, $R^2$, adjusted $R^2$, F Statistics, and residual standard error, I decided about the best regression model for Boston housing data-set.

# 3 Results

## 3.1 Descriptive Analysis

I start with some descriptive analysis to explore the dataset. Figure 2 represents the summary of the dataset indicating the the summary statistics such as minimum, 1st quartile, median, mean, 3rd quartile, and maximum value of the 12 explanatory and 1 outcome variable.



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

Figure 2: Summary of the Datasets

## 3.2 Correlation Analysis

Figure 3 shows the pairwise correlation between different variable. From figure 2, we saw that indus and nox, nox and dis, rad and tax are highly positively correlated. Correlation matrix showing the entire picture of the correlation between variable. In the Appendix, pairwise scatter plot also included in figure 25.

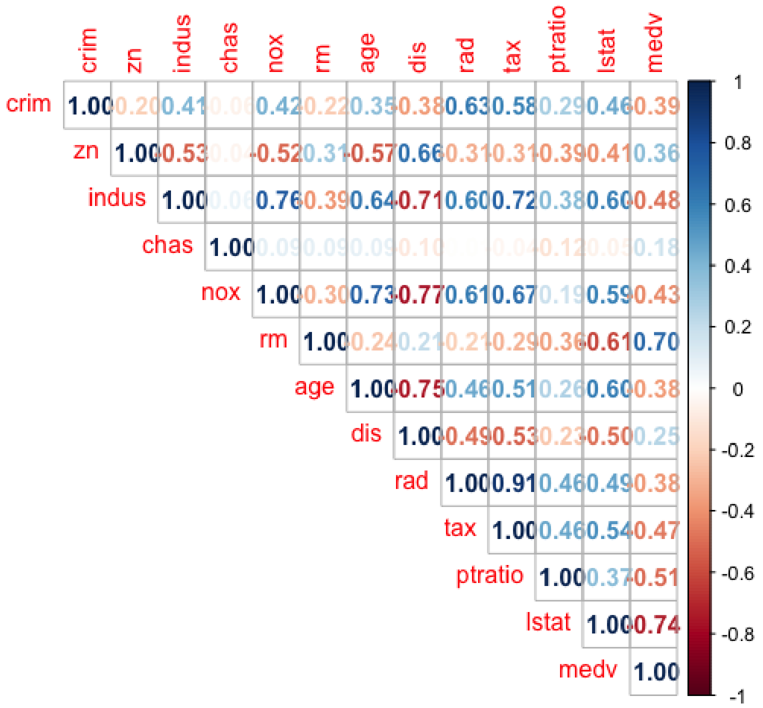


Figure 3: Pairwise Correlation Matrix

## 3.3 Variance Inflation Factor (VIF)

Variance Inflation Factor (VIF) is a statistical measure which find the amount of multicollinearity in a regression model. In an statistical model if there is a high multicollinearity then it will make it more difficult to estimate the relationship between each of the independent variable and dependent variable. In this case VIF is a good measure for testing multicollinearity.

### 3.3.1 VIF of Regression Model 1

Table 2, 3, and 4 shows the Variance Inflation Factor(VIF) of three regression model. And, we can observe from the first model that chas (1.074) has the lowest VIF value and tax (9.001) have the highest VIF value.

Table 2: Variance Inflation Factor(VIF) of Regression Model 1

| Predictor | VIF |
|---|---|
| crim | 1.792 |
| zn | 2.299 |
| indus | 3.991 |
| chas | 1.074 |
| nox | 4.393 |
| rm | 1.933 |
| age | 3.101 |
| dis | 3.956 |
| rad | 7.484 |
| tax | 9.001 |
| ptratio | 1.799 |
| black | 1.349 |
| lstat | 2.941 |

### 3.3.2 VIF of Regression Model 2

Table 3 shows the variance inflation factor of second model. From table 3, we can conclude that chas (1.082) has the lowest VIF value and interaction term of zn and tax (zn*tax) (25.192) have the highest VIF value.

Table 3: Variance Inflation Factor(VIF) of Regression Model 2

| Predictor | VIF |
|---|---|
| crim | 1.999 |
| zn | 23.826 |
| indus | 4.311 |
| chas | 1.082 |
| nox | 4.394 |
| rm | 1.959 |
| age | 3.105 |
| dis | 7.330 |
| rad | 20.470 |
| tax | 11.314 |
| ptratio | 1.807 |
| black | 1.349 |
| lstat | 2.966 |
| zn*tax | 25.192 |
| dis*rad | 7.361 |

### 3.3.3 VIF of Regression Model 3

Similary the variance inflation factor of model 3 are described in table 4. We can observe from the second model that chas (1.059) has the lowest VIF value and tax (7.272) have the highest VIF value.

All in all, we can see that the VIF is much smaller in model 3 for all the explanatory variable compare to other two model. And, in all case chase have the lowest VIF value whereas tax has the highest Variance inflation factor value.

Table 4: Variance Inflation Factor(VIF) of Regression Model 3

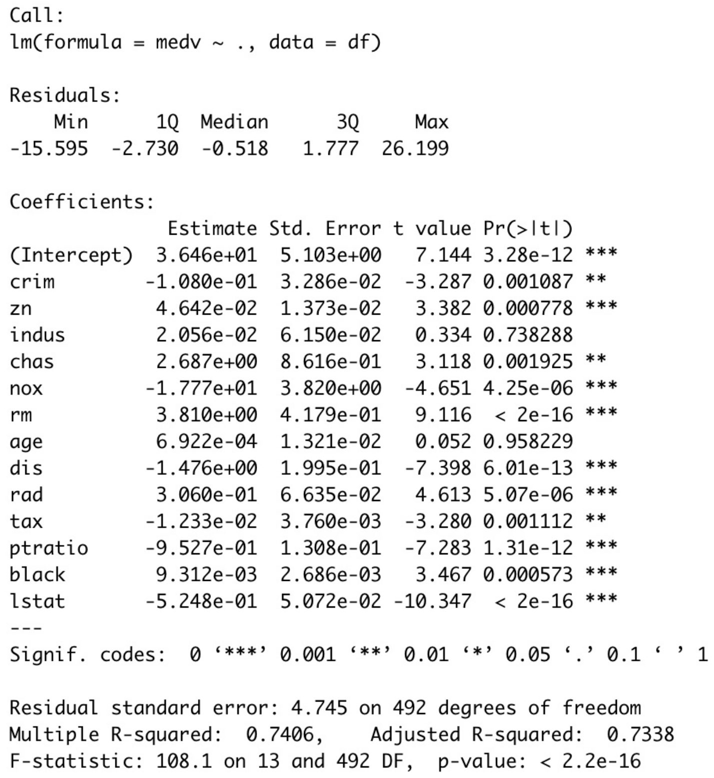
| Predictor | VIF |
|---|---|
| crim | 1.789 |
| zn | 2.239 |
| chas | 1.059 |
| nox | 3.778 |
| rm | 1.835 |
| dis | 3.443 |
| rad | 6.861 |
| tax | 7.272 |
| ptratio | 1.758 |
| black | 1.342 |
| lstat | 2.582 |

### 3.4 Regression Analysis of Three Model

I develop best three regression model for analyzing Boston housing prizes. Figure 4 shows the summary of regression analysis of first regression model and we see that p value of all the coefficient are significant except indus and age. And, the regression coefficient and standard error of regression are also shown in figure 4. Also, the $R^2$, adjusted $R^2$, and F value of regression model 1 are 0.7406, 0.7338, and 108.1.

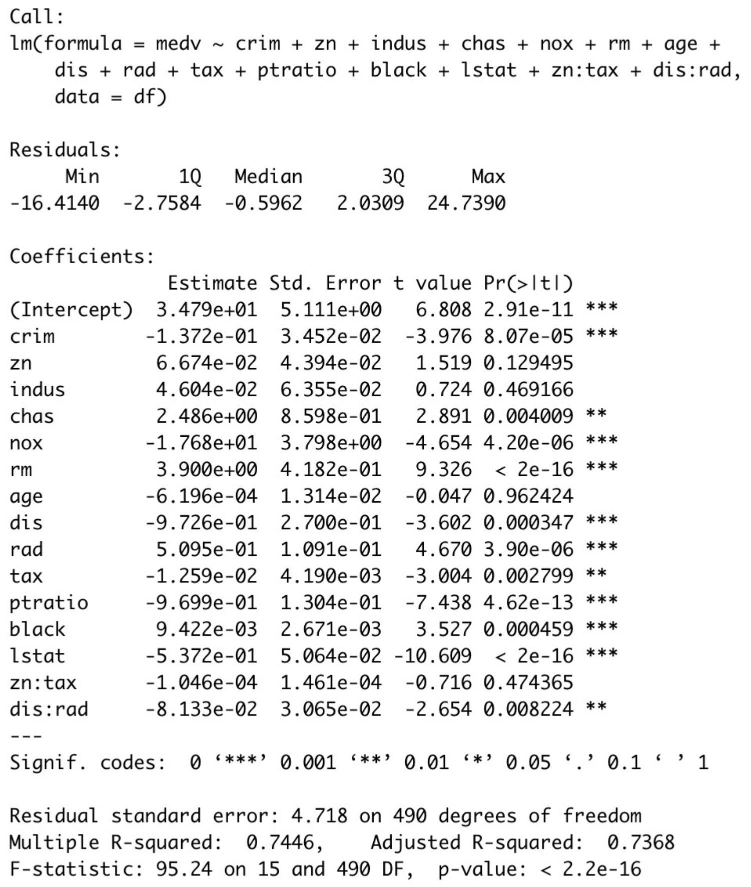
Similarly, figure 5 shows the summary of regression analysis of second regression model and we see that p value of zn, indus, age, and interaction term (zn*tax) are not significant. And, the regression coefficient and standard error of regression are also shown in figure 5. Also, the $R^2$, adjusted $R^2$, and F value of regression model 2 are 0.7446, 0.7368, and 95.24.

Figure 6 shows the summary of regression analysis of third regression model and we see that p value of all the coefficient are significant. And, the regression



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

Residuals:
    Min      1Q  Median      3Q     Max 
-15.595  -2.730  -0.518   1.777  26.199 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 ** 
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288    
chas         2.687e+00  8.616e-01   3.118 0.001925 ** 
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229    
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 ** 
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338 
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

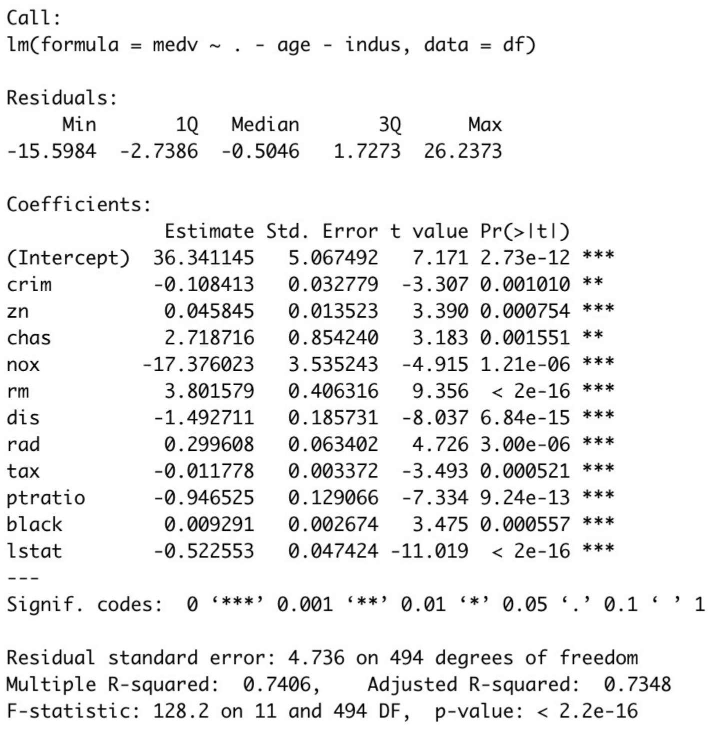
Figure 4: Summary of Regression Model 1



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

Residuals:
     Min       1Q   Median       3Q      Max
-16.4140  -2.7584  -0.5962   2.0309  24.7390

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.479e+01  5.111e+00   6.808 2.91e-11 ***
crim        -1.372e-01  3.452e-02  -3.976 8.07e-05 ***
zn           6.674e-02  4.394e-02   1.519 0.129495
indus        4.604e-02  6.355e-02   0.724 0.469166
chas         2.486e+00  8.598e-01   2.891 0.004009 **
nox         -1.768e+01  3.798e+00  -4.654 4.20e-06 ***
rm           3.900e+00  4.182e-01   9.326  < 2e-16 ***
age         -6.196e-04  1.314e-02  -0.047 0.962424
dis         -9.726e-01  2.700e-01  -3.602 0.000347 ***
rad          5.095e-01  1.091e-01   4.670 3.90e-06 ***
tax         -1.259e-02  4.190e-03  -3.004 0.002799 **
ptratio     -9.699e-01  1.304e-01  -7.438 4.62e-13 ***
black        9.422e-03  2.671e-03   3.527 0.000459 ***
lstat       -5.372e-01  5.064e-02 -10.609  < 2e-16 ***
zn:tax      -1.046e-04  1.461e-04  -0.716 0.474365
dis:rad     -8.133e-02  3.065e-02  -2.654 0.008224 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.718 on 490 degrees of freedom
Multiple R-squared:  0.7446,    Adjusted R-squared:  0.7368
F-statistic: 95.24 on 15 and 490 DF,  p-value: < 2.2e-16
```

Figure 5: Summary of Regression Model 2

coefficient and standard error of regression are also shown in figure 6. Also, the $R^2$, adjusted $R^2$, and F value of regression model 3 are 0.7406, 0.7348, and 128.2.



```
Call:
lm(formula = medv ~ . - age - indus, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.5984  -2.7386  -0.5046   1.7273  26.2373 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
crim         -0.108413   0.032779  -3.307 0.001010 ** 
zn            0.045845   0.013523   3.390 0.000754 ***
chas          2.718716   0.854240   3.183 0.001551 ** 
nox         -17.376023   3.535243  -4.915 1.21e-06 ***
rm            3.801579   0.406316   9.356  < 2e-16 ***
dis          -1.492711   0.185731  -8.037 6.84e-15 ***
rad           0.299608   0.063402   4.726 3.00e-06 ***
tax          -0.011778   0.003372  -3.493 0.000521 ***
ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
black         0.009291   0.002674   3.475 0.000557 ***
lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.736 on 494 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7348 
F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

Figure 6: Summary of Regression Model 3

### 3.5 Model Selection, Residual Plots, and Diagnostic Checking

The residual generally represent the deviation between the response and the fitted value and hence estimate a random component $\epsilon$ for the model. If there is any misspecification or departure from the underlying assumption in the model, then the residual plot will show up some pattern in the residuals. Hence, I can say that residual analysis is an effective way for discovering the model inadequacy.

Diagnostics checking involved checking the assumption of the model along with investigating any influential point. And, analysis of figure 7 to 18 will helps us for

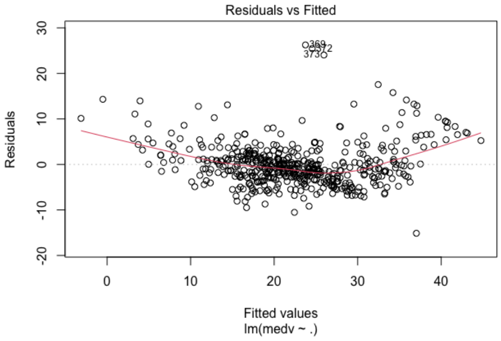


Figure 7: Resid. Vs Fitted (M-1)

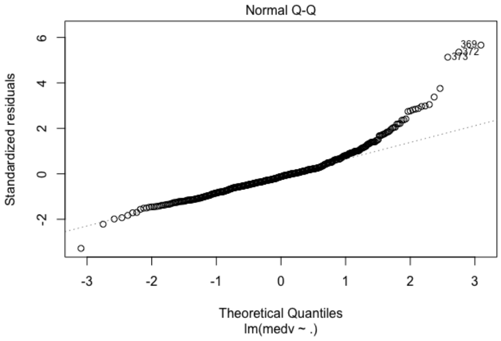


Figure 8: Normal Q-Q Plot (M-1)

making decision about diagnostic checking. In those figure M-1 means regression model 1, M-2 means regression model 2, and M-3 means regression model 3.

### 3.5.1 Residual Plot of Model 1

I construct several graphical representation of residual to investigate whether the residual follow some assumption such as normality and independent condition. These graphs will tell us whether the fitted model is an adequate representation or not. For example if the model is adequate then we will not find any systematic patterns in the residuals. In figure 7 fitted value vs residual plot, fitted value indicate that corresponding to the ith observation with $x_i$ as the value for the explanatory variable. It is the value which is implied by the fitted model. From the residual plot, we know that the positive value for the residual (on Y axis of the residual vs fitted value plot) means the prediction was too low, and negative means the prediction is too high. And, zero means he guess was exactly correct. Here, in the residual vs fitted value plot doesn't showing any systematic pattern through it is not fully perfect, we can say that the plot is good. The plot contains some outlier but we can see that almost all the points in the Y axis of residual vs fitted plot are close to zero which indicate the prediction is good.

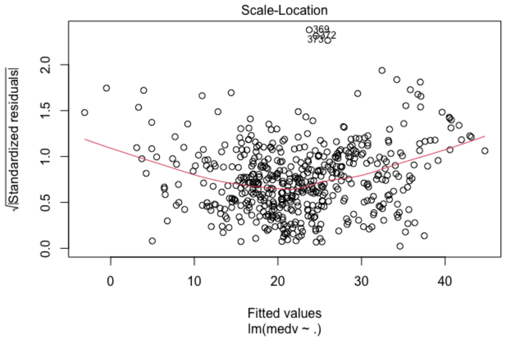


Figure 9: Scale Location Plot (M-1)

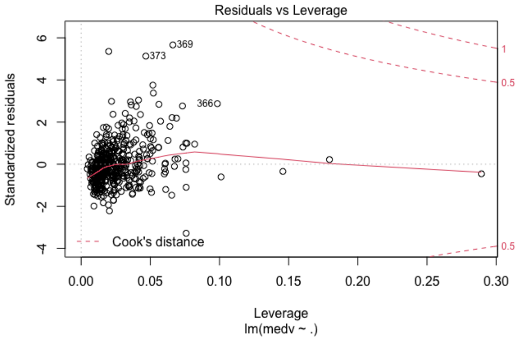


Figure 10: Residual Vs Leverage Plot (M-1)

### 3.5.2 Residual Plot of Model 2

Residual plot is a great way to check some assumption such as normality of the model. For the second model also, I construct some graphical representation of the residual plot and these graphs will tell us whether the fitted model is an adequate representation or not. For example if the model is adequate then we will not find any systematic patterns in the residuals.

Figure 11 to 14 indicate similar residual, normal Q-Q, and scale location plot of second regression model (M-2). In figure 11 fitted value vs residual plot, fitted value indicate that corresponding to the ith observation with $x_i$ as the value for the explanatory variable. From the residual plot, we know that the positive value for the residual (on Y axis of the residual vs fitted value plot) means the prediction was too low, and negative means the prediction is too high. And, zero means he guess was exactly correct. Here, in the residual vs fitted value plot is similar to regression model 1 having some outlier and it doesn't showing any systematic pattern through it is not fully perfect, we can say that the plot is good. Similar to model 1, we can see that almost all the points in the Y axis of residual vs fitted plot are close to zero which indicate the prediction is good.

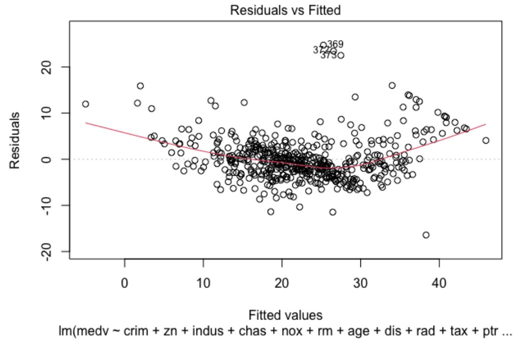


Figure 11: Resid. Vs Fitted (M-2)

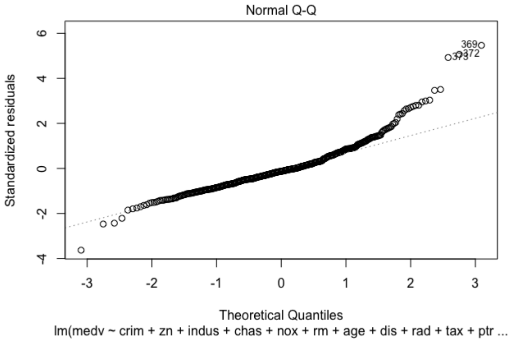


Figure 12: Normal Q-Q Plot (M-2)

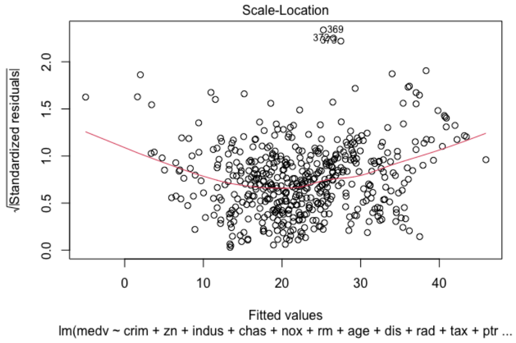


Figure 13: Scale Location Plot (M-2)

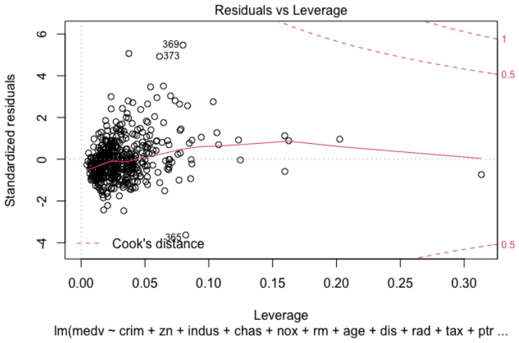


Figure 14: Residual Vs Leverage Plot (M-2)

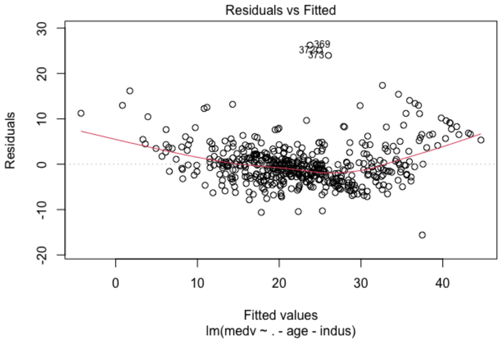


Figure 15: Resid. Vs Fitted (M-3)

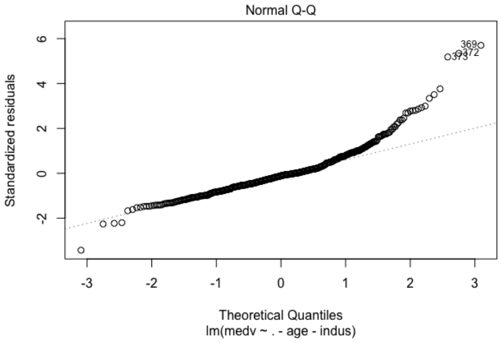


Figure 16: Normal Q-Q Plot (M-3)

### 3.5.3 Residual Plot of Model 3

In the similar approach like model 1 and 2, I construct some graphical representation of residual to investigate whether the residual follow some assumption such as normality and independent condition. These graphs will tell us whether the fitted model is an adequate representation or not. Figure 15 to 18 indicate the residual, normal Q-Q, and scale location plot of third regression model (M-3).
Here in figure 15, we can see that almost all the points in the Y axis of residual vs fitted plot are close to zero which indicate the prediction is good as like as model 1 and 2.

## 3.6 Confidence Interval(CI)

There are three classical approach for making inference of a model such as point estimation, confidence interval(most popular), and hypothesis testing[1]. I am going with the most popular one called estimating the confidence interval of three model. Generally, confidence interval is a range of value, defined by lower and upper bound for an unknown parameter that is likely to include a population value with a certain degree of confidence.

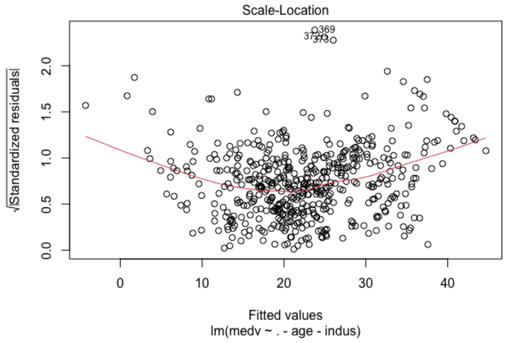


Figure 17: Scale Location Plot (M-3)

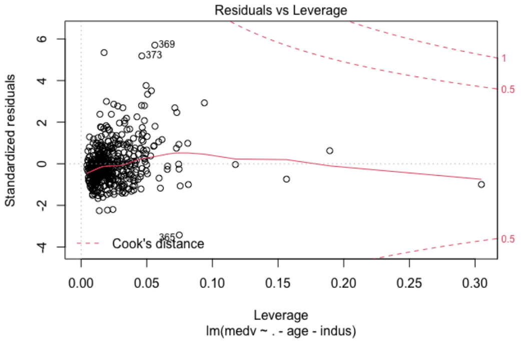


Figure 18: Residual Vs Leverage Plot (M-3)

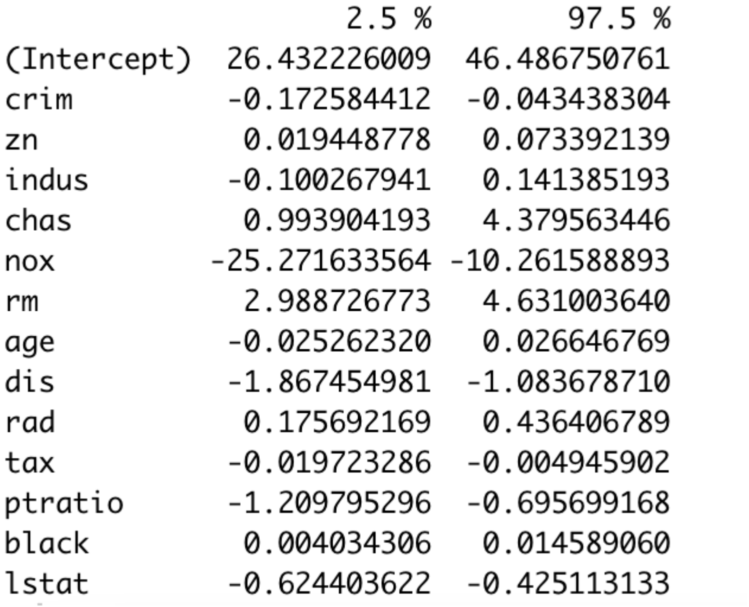
In this study, I consider 95% significance level for calculating the confidence interval. The 95% confidence interval indicates a range of value in which we are 95% confident that it contain the true mean of the population. And, figure 19, 20, and 21 indicate the lower limit(2.5%) and upper limit (97.5%) of confidence interval for the regression model 1, 2, and 3 respectively.

### 3.6.1 CI of Regression Model 1

Figure 19, shows the confidence interval of intercept and for all other coefficient of the explanatory variable consider in the regression model 1. From the figure 19, we can see that coefficient of indus and age contains 0 within the confidence interval. That indicate that we can able to test the hypothesis whether the coefficient of this two explanatory variable is 0 or not. And, there is high chance that the coefficient of indus and age will be 0 because in figure 4, we see that the p value of this two explanatory variable are not significant.

### 3.6.2 CI of Regression Model 2

Similarly, figure 20, shows the confidence interval of intercept and for all other coefficient of the explanatory variable consider in the regression model 2. From



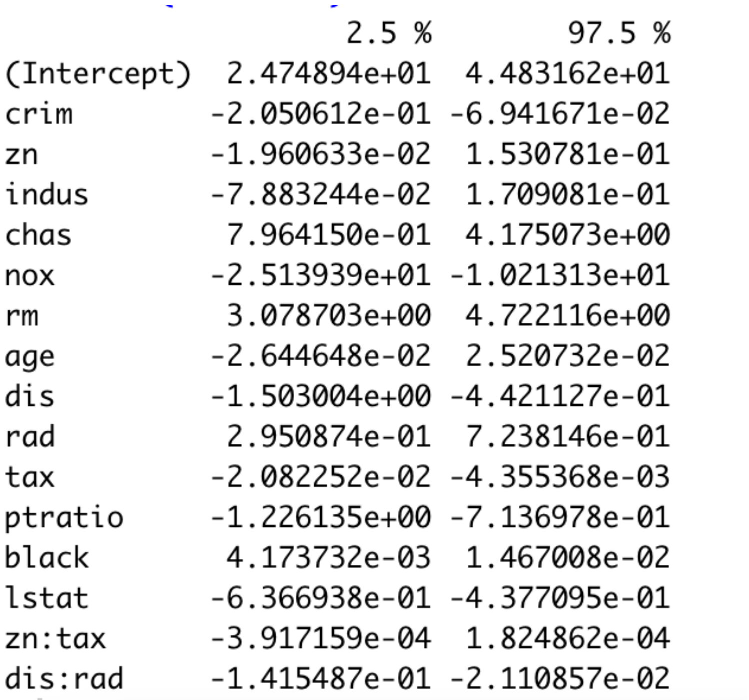
```
                   2.5 %        97.5 %
(Intercept)  26.432226009  46.486750761
crim         -0.172584412  -0.043438304
zn            0.019448778   0.073392139
indus        -0.100267941   0.141385193
chas          0.993904193   4.379563446
nox         -25.271633564 -10.261588893
rm            2.988726773   4.631003640
age          -0.025262320   0.026646769
dis          -1.867454981  -1.083678710
rad           0.175692169   0.436406789
tax          -0.019723286  -0.004945902
ptratio      -1.209795296  -0.695699168
black         0.004034306   0.014589060
lstat        -0.624403622  -0.425113133
```

Figure 19: Confidence Interval for Model 1

the figure 20, we can see that coefficient of zn, indus, age, and zn*tax contains 0 within the confidence interval. That indicate that we can able to test the hypothesis whether the coefficient of this three explanatory variable and the interaction term (zn*tax) are 0 or not. And, there is high chance that the coefficient of zn, indus, age, and the interaction term will be 0 because in figure 5, we see that the p value of this two explanatory variable are not significant.

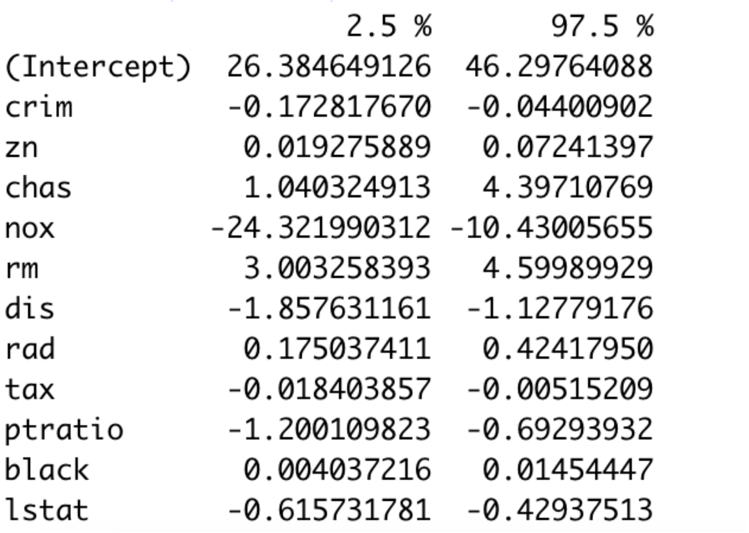
### 3.6.3 CI of Regression Model 3

As like as first two model, figure 21, shows the confidence interval of intercept and coefficient of the explanatory variable consider in the regression model 3. From the figure 21, we can see that there is no coefficient which contains 0 within the confidence interval and which is also possible to verify from figure 6 as we see that p value for the coefficient of third model is highly significant.



```
                     2.5 %        97.5 %
(Intercept)  2.474894e+01  4.483162e+01
crim        -2.050612e-01 -6.941671e-02
zn          -1.960633e-02  1.530781e-01
indus       -7.883244e-02  1.709081e-01
chas         7.964150e-01  4.175073e+00
nox         -2.513939e+01 -1.021313e+01
rm           3.078703e+00  4.722116e+00
age         -2.644648e-02  2.520732e-02
dis         -1.503004e+00 -4.421127e-01
rad          2.950874e-01  7.238146e-01
tax         -2.082252e-02 -4.355368e-03
ptratio     -1.226135e+00 -7.136978e-01
black        4.173732e-03  1.467008e-02
lstat       -6.366938e-01 -4.377095e-01
zn:tax      -3.917159e-04  1.824862e-04
dis:rad     -1.415487e-01 -2.110857e-02
```

Figure 20: Confidence Interval for Model 2



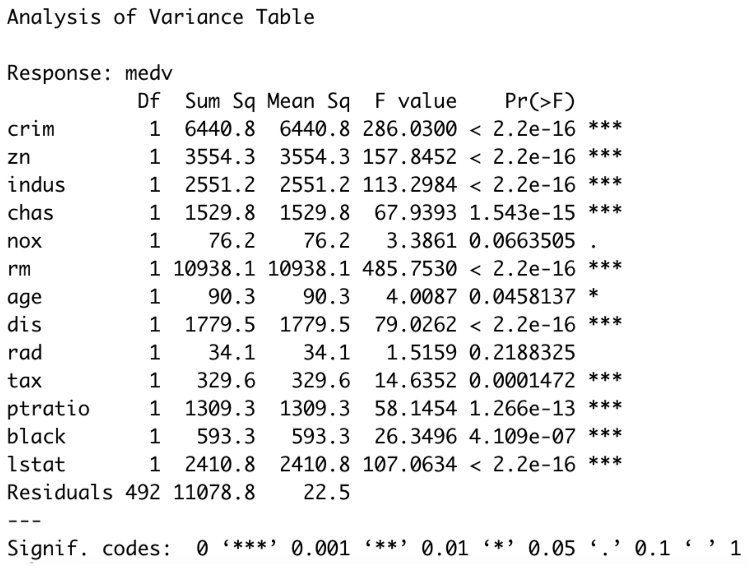
```
                    2.5 %       97.5 %
(Intercept)  26.384649126  46.29764088
crim         -0.172817670  -0.04400902
zn            0.019275889   0.07241397
chas          1.040324913   4.39710769
nox         -24.321990312 -10.43005655
rm            3.003258393   4.59989929
dis          -1.857631161  -1.12779176
rad           0.175037411   0.42417950
tax          -0.018403857  -0.00515209
ptratio      -1.200109823  -0.69293932
black         0.004037216   0.01454447
lstat        -0.615731781  -0.42937513
```

Figure 21: Confidence Interval for Model 3

## 3.7 ANOVA Analysis

### 3.7.1 ANOVA for Regression Model 1

Figure 22, indicates the ANOVA analysis of regression model 1, and in figure 22 ANOVA table includes the degree of freedom, sum of square, mean sum of square, Fisher ratio (F), and p value of the first regression model. ANOVA table indicate that the p value for coefficient of explanatory variable nox, rad, and age are not significant. The F (Fisher) test in the ANOVA table is also known as a test for overall significance of the regression.

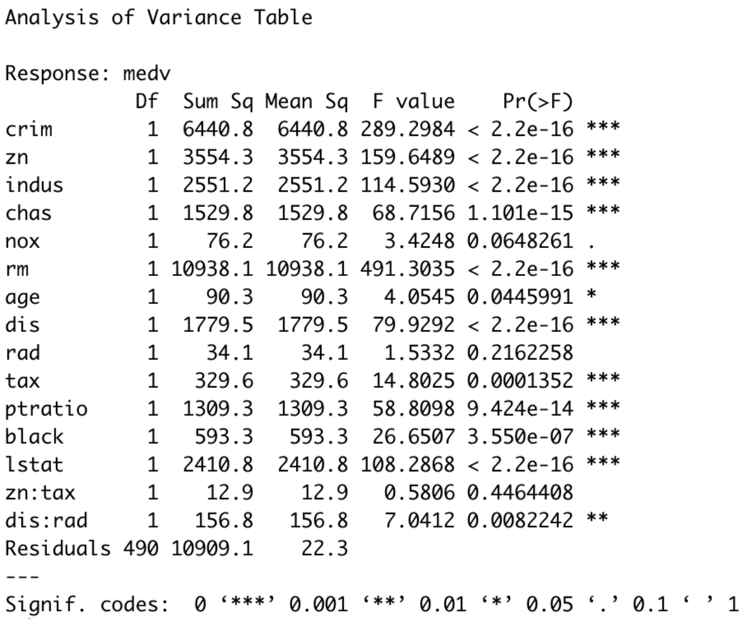


```
Analysis of Variance Table

Response: medv
           Df  Sum Sq Mean Sq  F value    Pr(>F)    
crim        1  6440.8  6440.8 286.0300 < 2.2e-16 ***
zn          1  3554.3  3554.3 157.8452 < 2.2e-16 ***
indus       1  2551.2  2551.2 113.2984 < 2.2e-16 ***
chas        1  1529.8  1529.8  67.9393 1.543e-15 ***
nox         1    76.2    76.2   3.3861 0.0663505 .  
rm          1 10938.1 10938.1 485.7530 < 2.2e-16 ***
age         1    90.3    90.3   4.0087 0.0458137 *  
dis         1  1779.5  1779.5  79.0262 < 2.2e-16 ***
rad         1    34.1    34.1   1.5159 0.2188325    
tax         1   329.6   329.6  14.6352 0.0001472 ***
ptratio     1  1309.3  1309.3  58.1454 1.266e-13 ***
black       1   593.3   593.3  26.3496 4.109e-07 ***
lstat       1  2410.8  2410.8 107.0634 < 2.2e-16 ***
Residuals 492 11078.8    22.5                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 22: Anova Table for Regression Model 1

### 3.7.2 ANOVA for Regression Model 2

Figure 23, indicates the ANOVA analysis of regression model 2, and in figure 23 ANOVA table includes all the necessary metric of anova analysis such as the degree of freedom, sum of square, mean sum of square, fisher ratio (F), and p value of the second regression model. ANOVA table indicate that the p value for coefficient of explanatory variable nox, rad, age, and interaction between age zn and tax (zn*tax) are not significant.



```
Analysis of Variance Table

Response: medv
           Df  Sum Sq Mean Sq  F value    Pr(>F)    
crim        1  6440.8  6440.8 289.2984 < 2.2e-16 ***
zn          1  3554.3  3554.3 159.6489 < 2.2e-16 ***
indus       1  2551.2  2551.2 114.5930 < 2.2e-16 ***
chas        1  1529.8  1529.8  68.7156 1.101e-15 ***
nox         1    76.2    76.2   3.4248 0.0648261 .  
rm          1 10938.1 10938.1 491.3035 < 2.2e-16 ***
age         1    90.3    90.3   4.0545 0.0445991 *  
dis         1  1779.5  1779.5  79.9292 < 2.2e-16 ***
rad         1    34.1    34.1   1.5332 0.2162258    
tax         1   329.6   329.6  14.8025 0.0001352 ***
ptratio     1  1309.3  1309.3  58.8098 9.424e-14 ***
black       1   593.3   593.3  26.6507 3.550e-07 ***
lstat       1  2410.8  2410.8 108.2868 < 2.2e-16 ***
zn:tax      1    12.9    12.9   0.5806 0.4464408    
dis:rad     1   156.8   156.8   7.0412 0.0082242 ** 
Residuals 490 10909.1    22.3                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

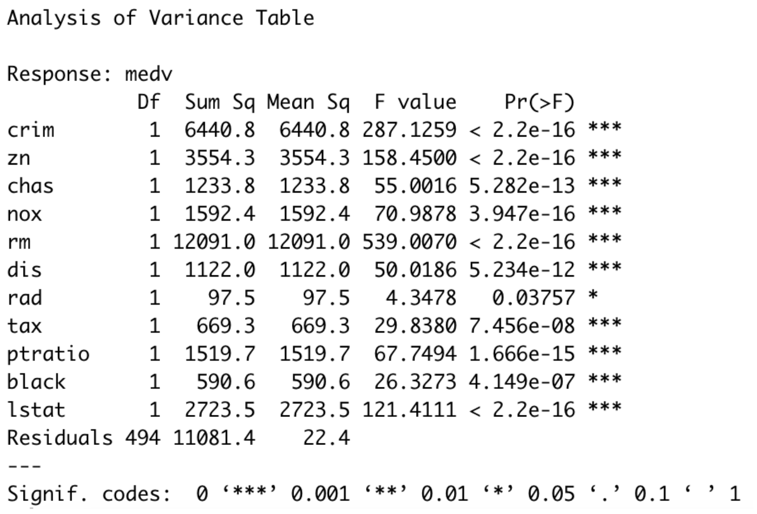
Figure 23: Anova Table for Regression Model 2

### 3.7.3 ANOVA for Regression Model 3

As like as model 1 and model 2, figure 24, indicates the ANOVA analysis of regression model 3, and in figure 24 ANOVA table includes all the necessary metric of anova analysis. ANOVA table indicate that the p value for coefficient of all the explanatory variable are significant except (dis) which is not fully significant (p value = 0.03).

# 4 Discussion and Conclusion

In this section, I will discuss the performance of the three model and decide which one will be best. For measuring the performance of the model, I will use some metric such as $R^2$, adjusted $R^2$, F statistics, and residual standard error. $R^2$ is a



```
Analysis of Variance Table

Response: medv
           Df  Sum Sq Mean Sq  F value    Pr(>F)    
crim        1  6440.8  6440.8 287.1259 < 2.2e-16 ***
zn          1  3554.3  3554.3 158.4500 < 2.2e-16 ***
chas        1  1233.8  1233.8  55.0016 5.282e-13 ***
nox         1  1592.4  1592.4  70.9878 3.947e-16 ***
rm          1 12091.0 12091.0 539.0070 < 2.2e-16 ***
dis         1  1122.0  1122.0  50.0186 5.234e-12 ***
rad         1    97.5    97.5   4.3478   0.03757 *  
tax         1   669.3   669.3  29.8380 7.456e-08 ***
ptratio     1  1519.7  1519.7  67.7494 1.666e-15 ***
black       1   590.6   590.6  26.3273 4.149e-07 ***
lstat       1  2723.5  2723.5 121.4111 < 2.2e-16 ***
Residuals 494 11081.4    22.4                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 24: Anova Table for Regression Model 3

descriptive measure for finding the association between response and explanatory variable. It provides a summary measure of how well the model fits the data. It can also give us some feedback on the importance of adding a variable to (or deleting a variable from) a model. In table 5, I summarize different performance metrics for three model. From the table 5, we saw that $R^2$, adjusted $R^2$, and residual standard error are almost same for all the three model but model 3 has the highest F Statistics

Table 5: Performance Evaluation of Different Regression Model

| Criteria | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| $R^2$ | 0.741 | 0.745 | 0.741 |
| Adjusted $R^2$ | 0.734 | 0.737 | 0.735 |
| F Statistics | 108.1 | 95.24 | 128.2 |
| Residual Standard Error | 4.745 | 4.718 | 4.736 |

(Fisher value). All in all, I can declare the third regression model as the best model.

$$\begin{aligned} MEDV = 36.341 - 0.108 * CRIM + 0.049 * ZN + 2.719 * CHAS \\ - 17.376 * NOX + 3.802 * RM - 1.493 * DIS + 0.299 * RAD - 0.0118 * TAX - \\ 0.947 * PTRATIO + 0.009 * BLACK - 0.523 * LSTAT \end{aligned} \quad (4)$$

In future, we can able to make prediction on new data using this equation 4 and our regression analysis will become machine learning model if we can make prediction using final regression model equation(4).

# References

[1] Johannes Ledolter Bovas Abraham. *Introduction to regression modeling.* Belmont, CA : Thomson Brooks/Cole, 2006.

# 5 Appendix

## 5.1 R Code

```
############################### Final Project  ####################################


# Author: Md Salman Rahman
# Course: MATH 6364 Statistical Methods
# Course Instructor: Dr. George Yanev
```

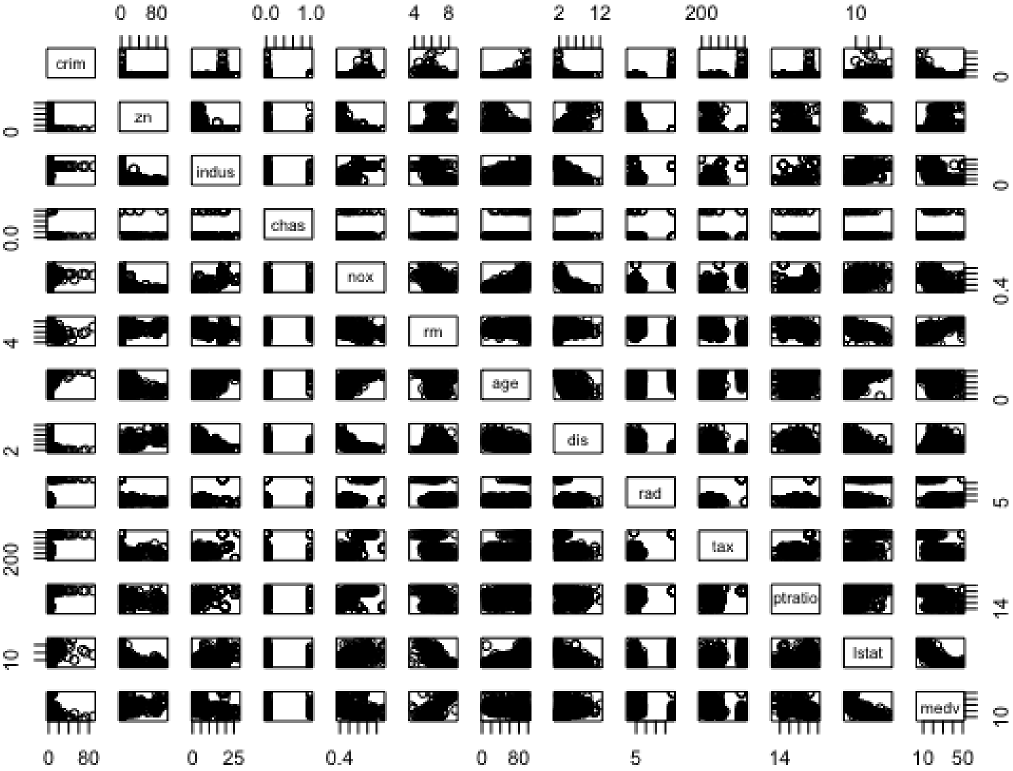


Figure 25: Pairwise Scatterplot

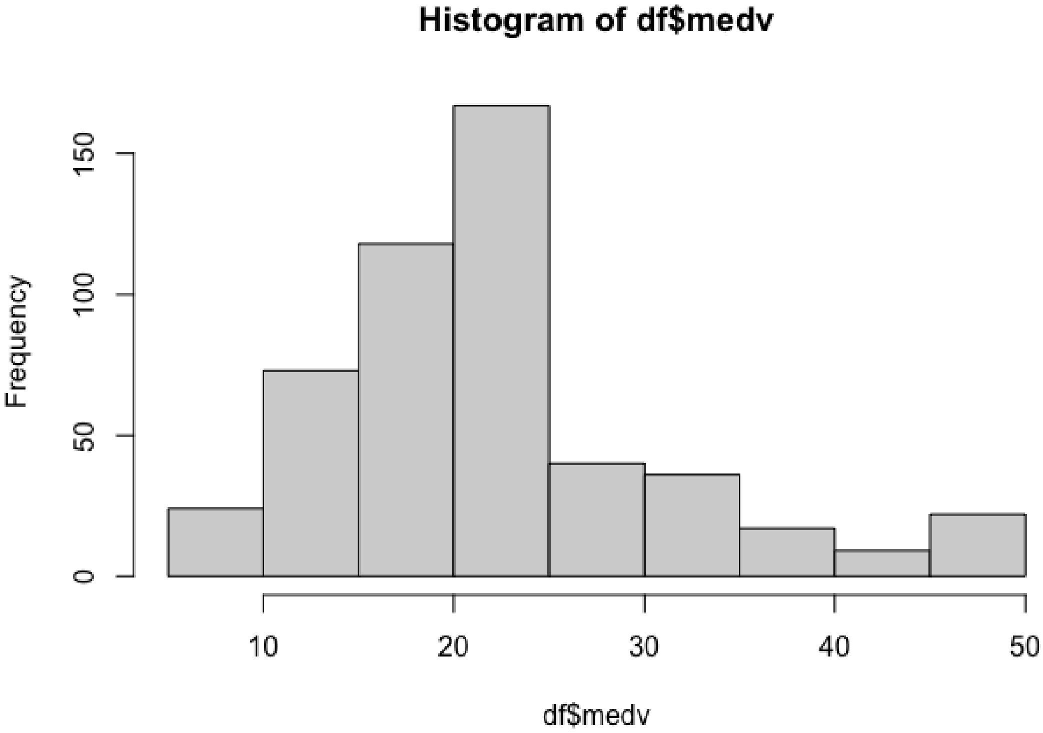


Figure 26: Histogram Plot of MEDV

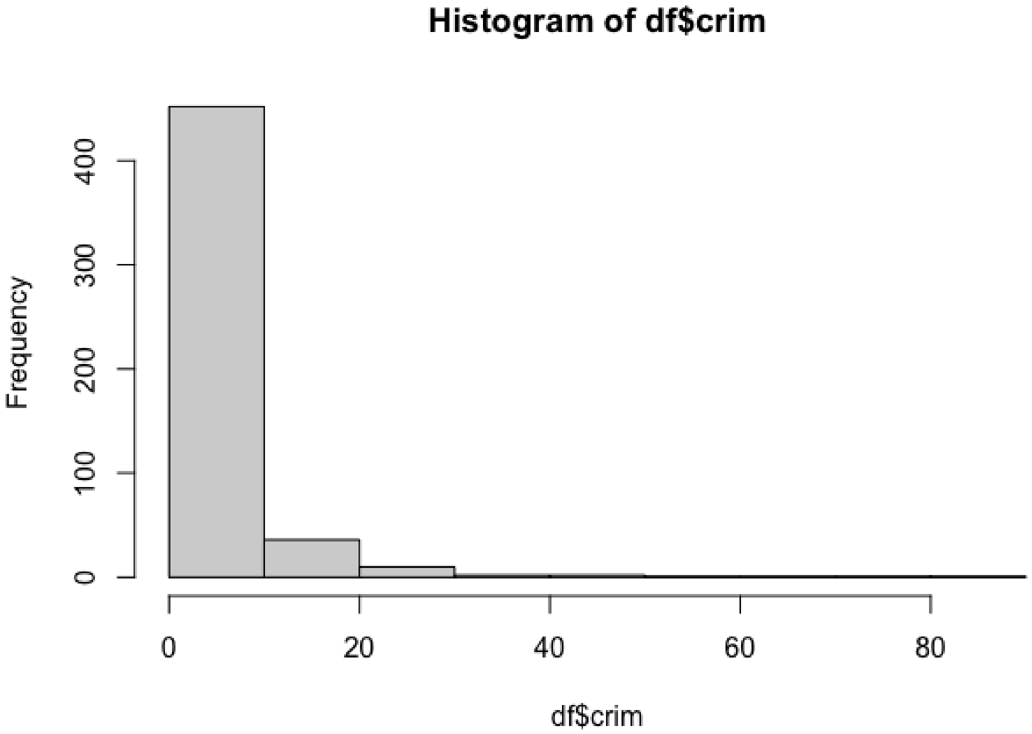


Figure 27: Histogram Plot of CRIM

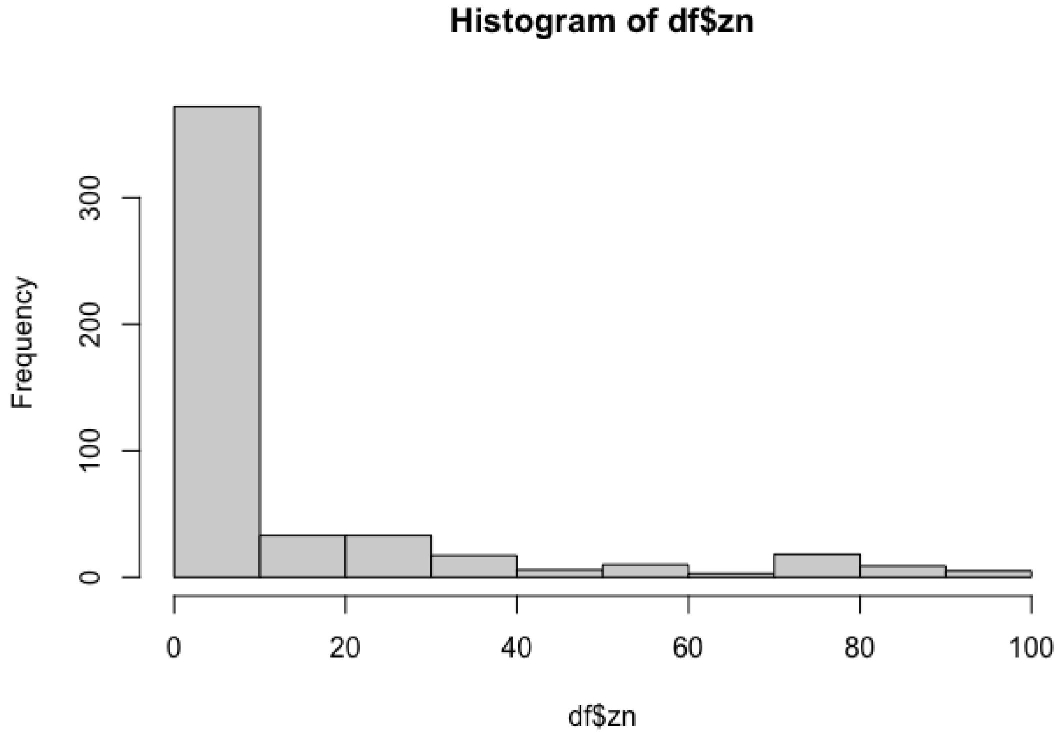


Figure 28: Histogram Plot of ZN

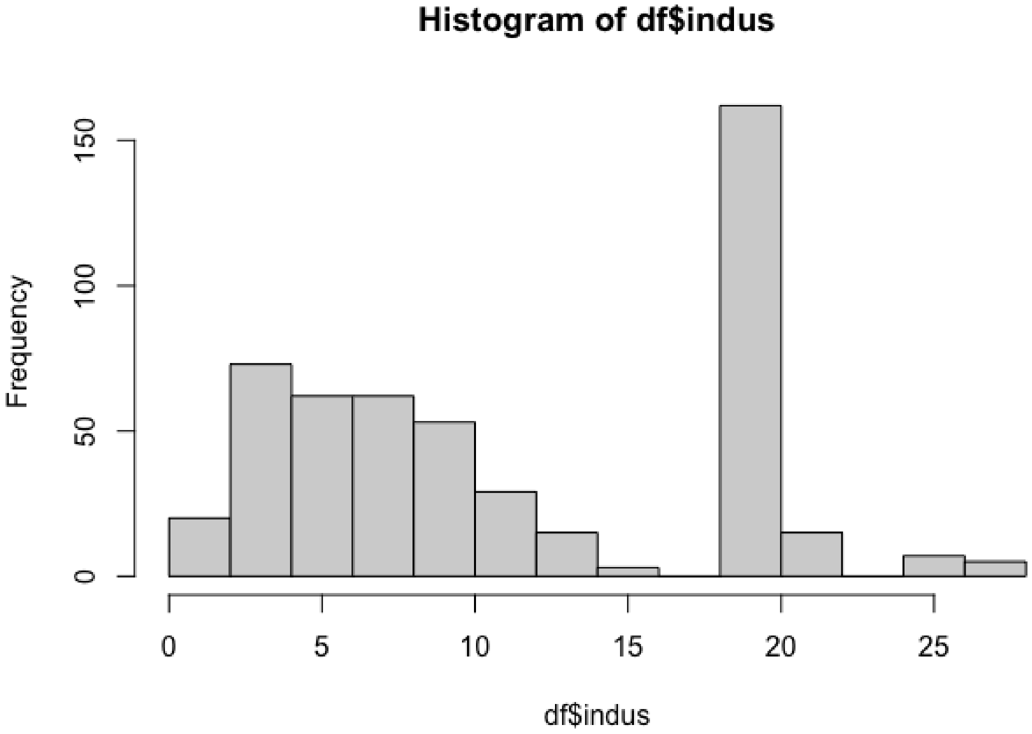


Figure 29: Histogram Plot of INDUS

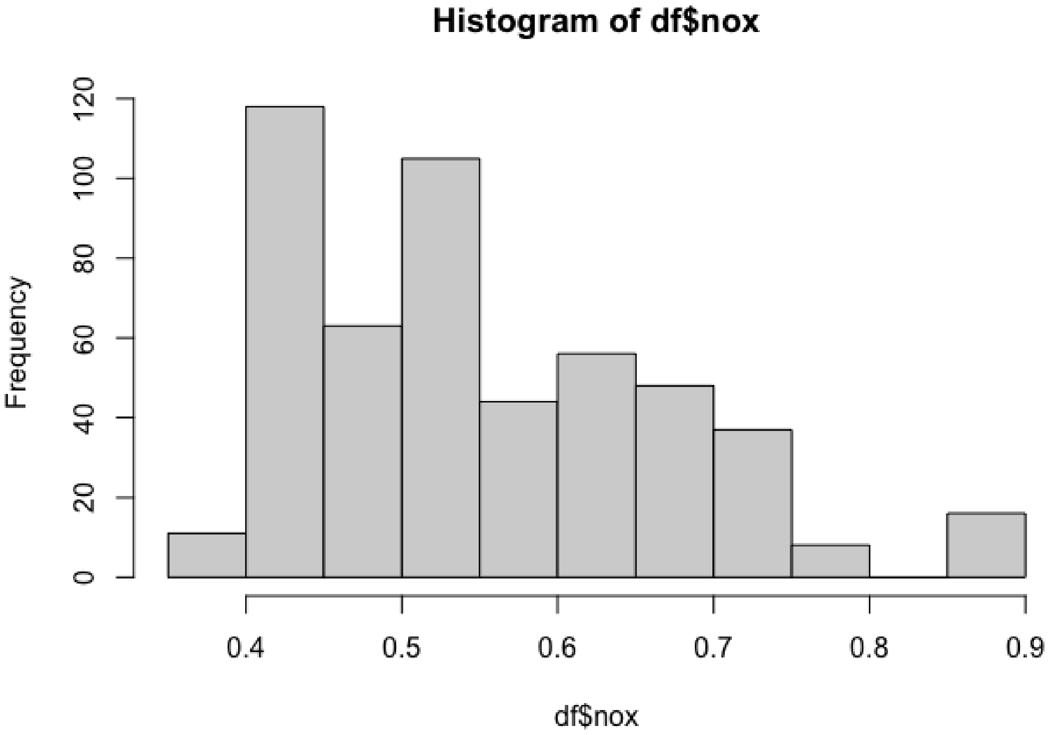


Figure 30: Histogram Plot of NOX

```
setwd("/Users/salman/OneDrive - The University of Texas-Rio Grande Valley/Course_vi

library("readxl")
library(ISLR2)
library(ggplot2)
data("Boston")

df<- Boston

summary(df)
str(df)

########################  Checking Missing Value #################
library(tidyr)
library(dplyr)
summary(df)
table(is.na(df))

##################### descriptive analysis  ###################
head(df)
# correlation analysis
library(corrplot)
# correlation matrix
corrplot(cor(df[,]),
         method = "number",
         type = "upper" # show only upper side
)
```

```
# scatter plot of several variable
pairs(df[,])


# histogram
hist(df$medv)
hist(df$crim)
hist(df$zn)
hist(df$indus)
hist(df$chas)
hist(df$nox)


#################### linear regression #########################

#model 1
lm.model <- lm(medv~., data = df)
names(lm.model)
coef(lm.model)
#confidence interval
confint(lm.model)
summary(lm.model)

# anova analysis
anova(lm.model)

#variance inflation factor
```

```
library(car)
vif(lm.model)


plot(lm.model)


#model 2

lm.model2 <- update(lm.model, . ˜. + zn:tax+dis:rad)

summary(lm.model2)

#confidence interval
confint(lm.model2)
#variance inflation factor

library(car)
vif(lm.model2)

# anova analysis
anova(lm.model2)


plot(lm.model2)

#model 3
```

```
lm.model3 <- lm(medv ˜. -age-indus, data=df)

summary(lm.model3)

#confidence interval
confint(lm.model3)
#variance inflation factor

library(car)
vif(lm.model3)

# anova analysis
anova(lm.model3)

plot(lm.model3)
```