**House Price Prediction**

Shimpli Shashikant Borkar

1. **INTRODUCTION**

Property price forecasting is getting more crucial and advantageous in recent times. Housing prices are a reliable predictor of a state's financial health and general market state. By using provided features in the dataset, the goal is to estimate housing prices. Someone who might use such data on a regular basis, such as a real estate agent, would benefit from this model.

The objective of the project is:

1. To study the effect of the variables present in the Boston dataset on the median value of owner-occupied homes (MEDV).

2. To formulate a Linear Regression Model with:

a. Dependent Variable as MEDV

b. Independent Variables as remaining variables other than MEDV by selecting the most effective variables and removing the unwanted ones.

3. perform data analysis, predictive modeling, and diagnostics on the Boston housing dataset to predict MEDV (median value of owner-occupied houses in $1000's)

By examining the AIC and BIC values of the training model, the prediction's accuracy is assessed. The necessary pre-processing techniques are used, and the data is divided into two portions before the test is run. Both parts will be used, however, only one will be used during training and the other during testing. Additionally, the provided dataset is processed to improve performance. This is done by identifying the necessary features and using one of the selection methods to remove the unwanted variables because each house has certain attributes that contribute to estimating its MEDV.

1. **METHODS**
   1. **DATA GATHERING AND INFORMATION**

The Boston Housing Dataset is derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA [1]. The dataset has been selected from the Kaggle website for analysis. There are 506 observations and 13 variables in the dataset.

The following describes the dataset columns:

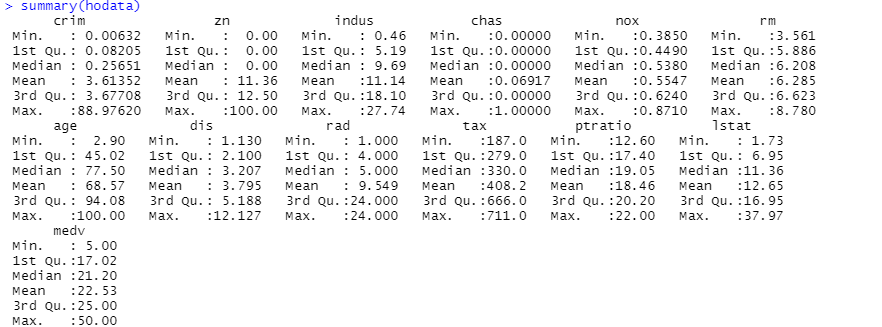
* CRIM - per capita crime rate by town
* ZN – the proportion of residential land zoned for lots over 25,000 sq. ft.
* INDUS – the proportion of non-retail business acres per town.
* CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
* NOX - nitric oxides concentration (parts per 10 million)
* RM – the average number of rooms per dwelling
* AGE – the proportion of owner-occupied units built prior to 1940
* DIS - weighted distances to five Boston employment centers
* 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 - Median value of owner-occupied homes in $1000's

The types of variables respectively are shown below in table 1. a.

|  |  |
| --- | --- |
| Variable | Type |
| CRIM | Numeric |
| ZN | Numeric |
| INDUS | Numeric |
| CHAS | Numeric |
| NOX | Numeric |
| RM | Numeric |
| AGE | Numeric |
| DIS | Numeric |
| RAD | Numeric |
| TAX | Numeric |
| PTRATIO | Numeric |
| LSTAT | Numeric |
| MEDV | Numeric |

* 1. **DESCRIPTIVE ANALYSIS**

Below is the descriptive summary of the values contained in each of the columns of the dataset under respective variables.

****

Our model's outcome or dependent or response variable is the 'median value of homes in $1000’s, represented by MEDV. The remaining 13 variables would be treated as independent or explanatory variables and will be used to predict the MEDV.

* 1. **INITIAL MODELING** 
     1. **DATA CLEANING**

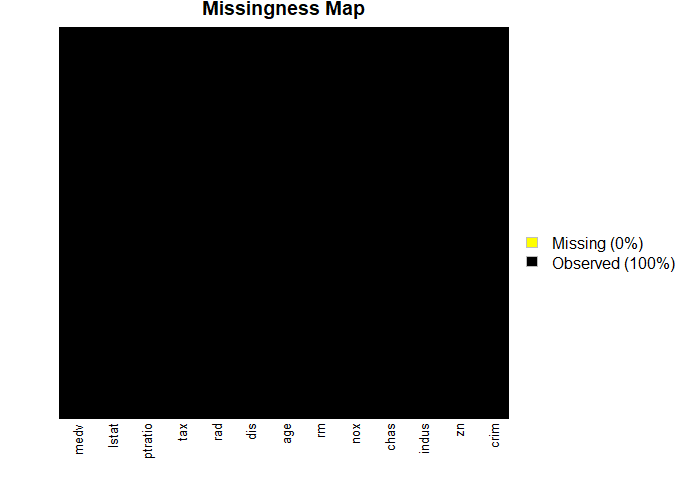
The problem of missing data in applied data analytics is extremely challenging. The dataset's missing values cannot just be overlooked. The very practical reason that most algorithms do not accept missing values means that it must be handled in some way. Therefore, before creating the prediction model, it is a good idea to find and replace any missing values for each column in the input data.

The process of calculating a statistical value for each column (such as a mean) and substituting that statistic for all missing values for that column is known as data imputation. It is a commonly used method since the statistic is simple to compute using the training dataset and it frequently yields good performance.

In the below result, we see that there are no N/A values present in the dataset under any column.



Then we go ahead and check if there are any missing values using the missmap function which plots the missing and observed values in a missingness map for each respective column. The black-shaded region shows that there are no missing values which should be indicated with yellow. Hence, we conclude that the dataset is fit for further analysis.



* + 1. **SPLITTING DATA INTO TRAINING AND TEST DATASETS**

Statistical methods are used to evaluate the accuracy of the model that we develop using untested data in order to assess how accurate it is. To put it another way, some information that the algorithms won't see will be held back and used to obtain a second, independent assessment of the model's potential accuracy. In other words, the model is fitted and tuned using training sets. To assess the model, test sets are set aside as "unseen" data. The imported dataset will now be divided into two parts: 20% will be kept back as a testing dataset, while the remaining 80% will be used as a training dataset to train our models.



The dimensions of the training and test data set are as follows:

|  |  |
| --- | --- |
| Dataset | Dimensions |
| Training set (80%) | 434 obervations,13 variables |
| Test set (20%) | 1. obervations,13 variables |

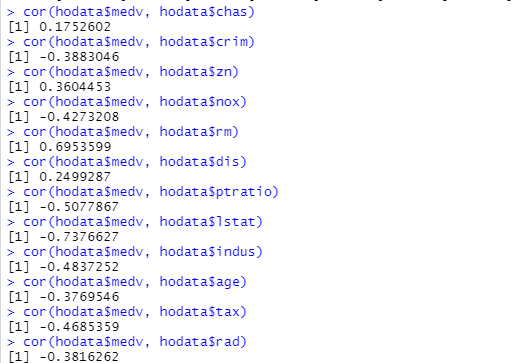
* 1. **DIAGNOSTICS**

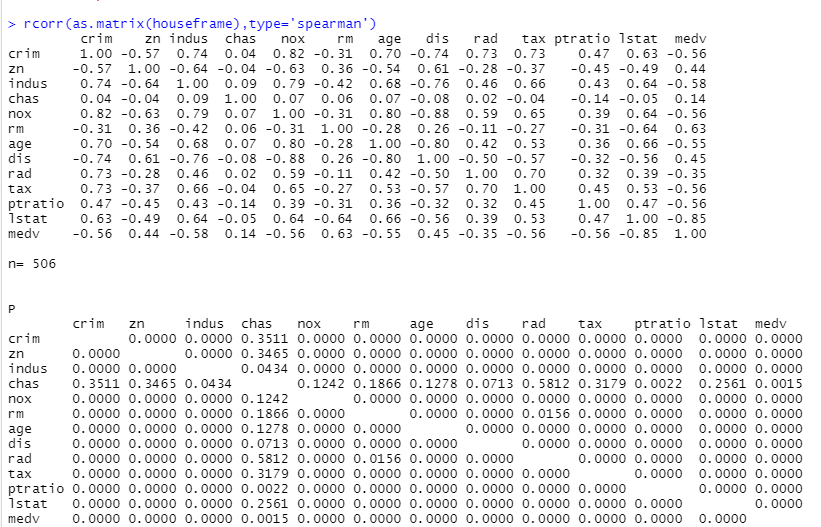
Since regression analysis is based on a number of assumptions, model fitting is only the beginning of the narrative. Regression diagnostics are employed in order to assess the model assumptions and look for observations that have a significant, unjustified impact on the analysis. Assumptions of our linear regression model:

1. Homogeneity of variance (homoscedasticity): Across the range of values for the independent variable, the magnitude of the inaccuracy in our prediction does not show significant changes.
2. Independence of observations: For the observations in the training and test dataset there are no hidden relationships among variables.
3. Correlation: In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated, then only one of them should be used in the regression model.
4. Normality: The data used for our model follows a normal distribution.
5. Linearity: the line of best fit through the data points is a straight line.
   * 1. **CORRELATIONS AND AUTOCORRELATIONS**

|  |
| --- |
| Positive correlation between medv and chas. |
| Negative correlation between medv and crim. |
| Positive correlation between medv and zn. |
| Negative correlation between medv and nox. |
| Positive correlation between medv and rm. |
| Positive correlation between medv and dis. |
| Negative correlation between medv and ptratio. |
| Negative correlation between medv and lstast. |
| Negative correlation between medv and indus. |
| Negative correlation between medv and age. |
| Negative correlation between medv and tax. |
| Negative correlation between medv and rad. |

A statistical measure called correlation shows how much two or more variables vary together, whether or not there is a causal relationship. When two variables rise or decrease simultaneously, there is a positive correlation; when there is a negative correlation, one variable increases as the other falls.

****

****

The probability that you'd have discovered the current result if the correlation coefficient were actually 0 is known as the P-value (null hypothesis). The correlation coefficient is referred to as statistically significant if this likelihood is less than the usual 5% (P0.05).

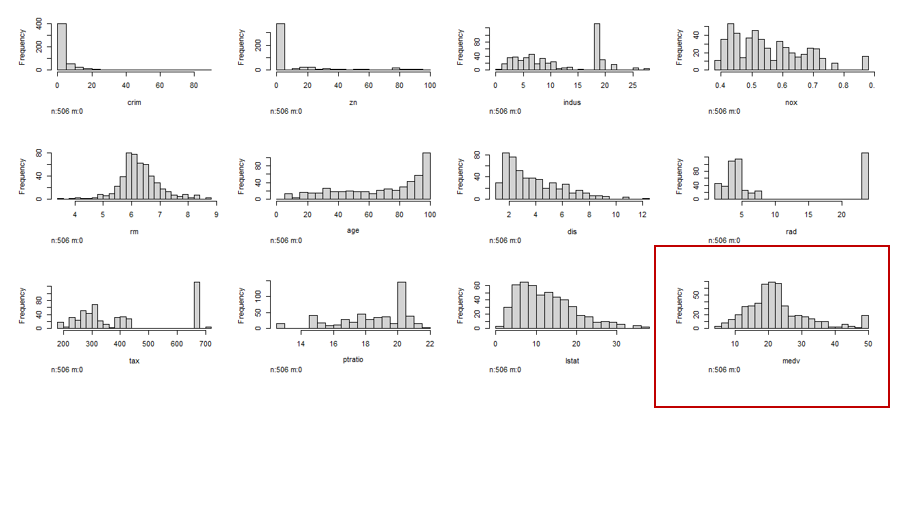
Each independent variable must have a strong association with the target variable. However, there shouldn't be a lot of correlation between the two independent variables. When independent variables in a regression model are correlated, multicollinearity emerges. Because explanatory variables should be independent, this association is problematic. When fitting the model and interpreting the findings, it may be difficult if there is a high enough degree of correlation between the variables.

|  |  |
| --- | --- |
| Correlation Matrix: Interpretation of the strength of correlation results | |
| Correlation coefficient range | Correlation Strength |
| 0.00-0.30 | Weak |
| 0.31-0.50 | Moderate |
| 0.51-0.80 | Strong |
| 0.81-1.00 | Very Strong |

Through the correlation result from the above, we understand that the P-value between independent variables with the target variable are all < 0.05, basically the correlation between the independent variables and the target variable MEDV are statistically significant. However, by looking at the correlation coefficient of the independent variable CHAS with the target variable MEDV (0.14) are weak correlations.

* + 1. **NORMALITY**

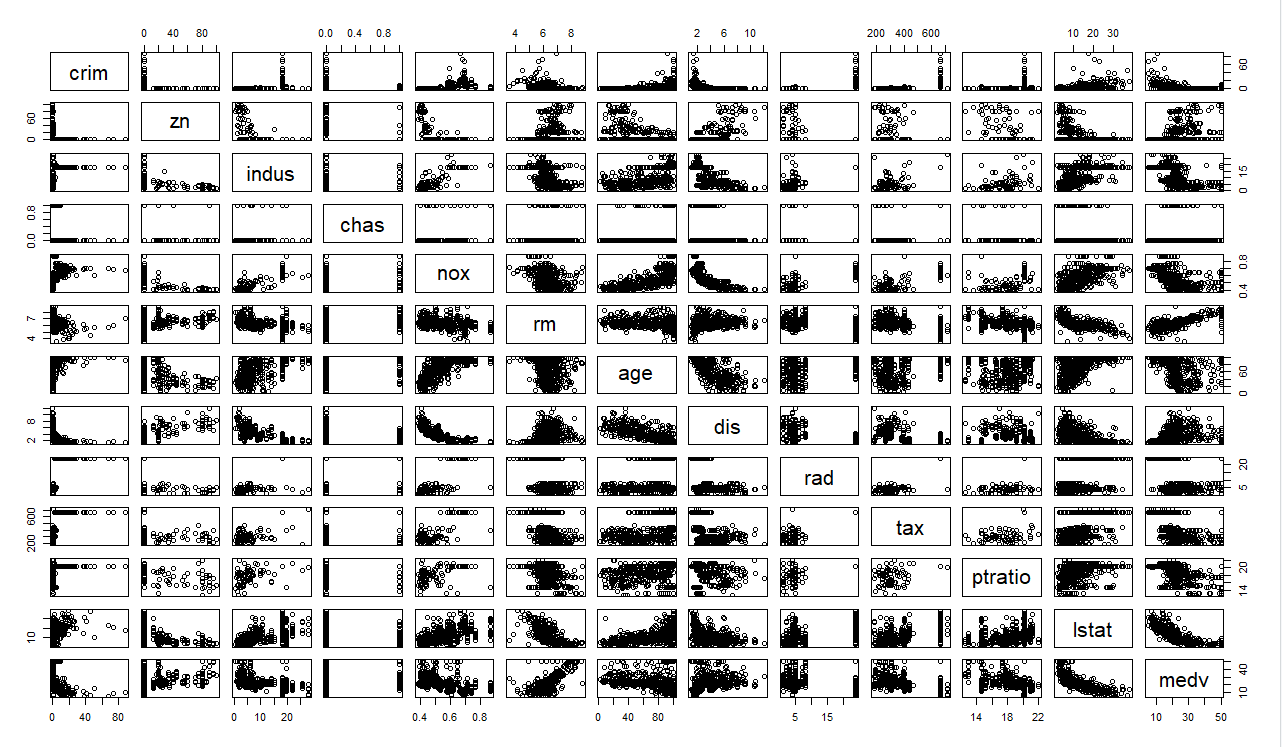
The hist.data.frame() function tests whether the dependent variable follows a normal distribution. The observations distribution of the dependent variable i.e. MEDV is roughly bell-shaped. Therefore, we can proceed with the linear regression.



* + 1. **LINEARITY**

There must be a linear relationship between the independent and dependent variables. With a scatter plot below from the last column, we can visually check if the distribution of the data points appears to be linear.

Despite having two distinct sections, the relationship between RAD and TAX with the MEDV still seems linear in the lower area. We can carry out linear regression now.



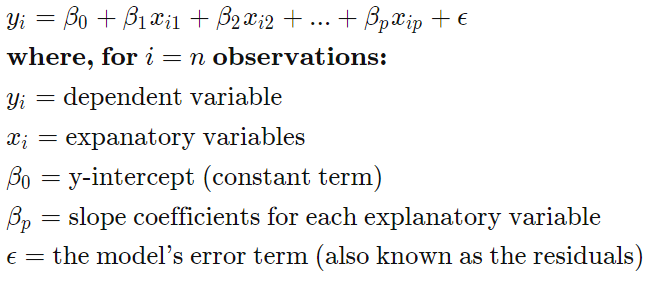
* + 1. **HOMOSCCEDASTICITY**

The foundation of linear regression models is the homoscedasticity assumption, which literally means "same variance". When the error term—the "noise" or random disturbance in the relationship between the independent variables and the dependent variable—is the same at all independent variable values, the condition is referred to as homoscedastic. When the amount of the error term varies throughout values of an independent variable, heteroscedasticity (the violation of homoscedasticity) is present. The degree to which the homoscedasticity assumption is violated affects the consequences, which rise with increasing heteroscedasticity.

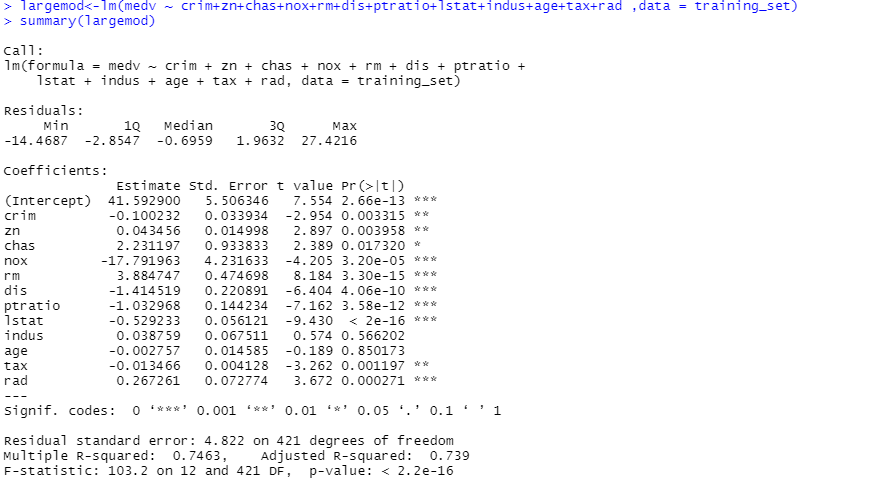
* 1. **BUILDING AN INITIAL MODEL**

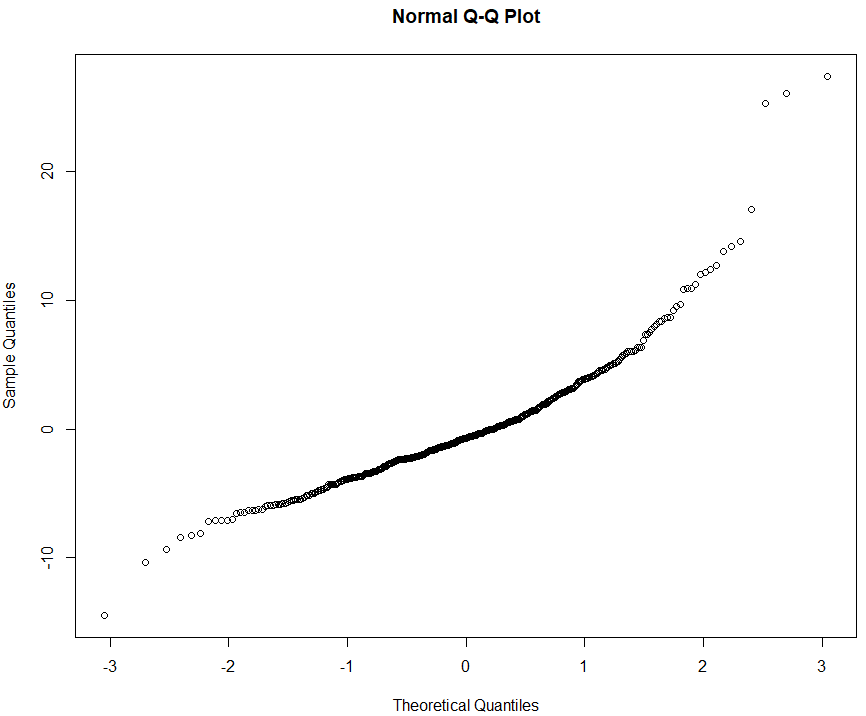
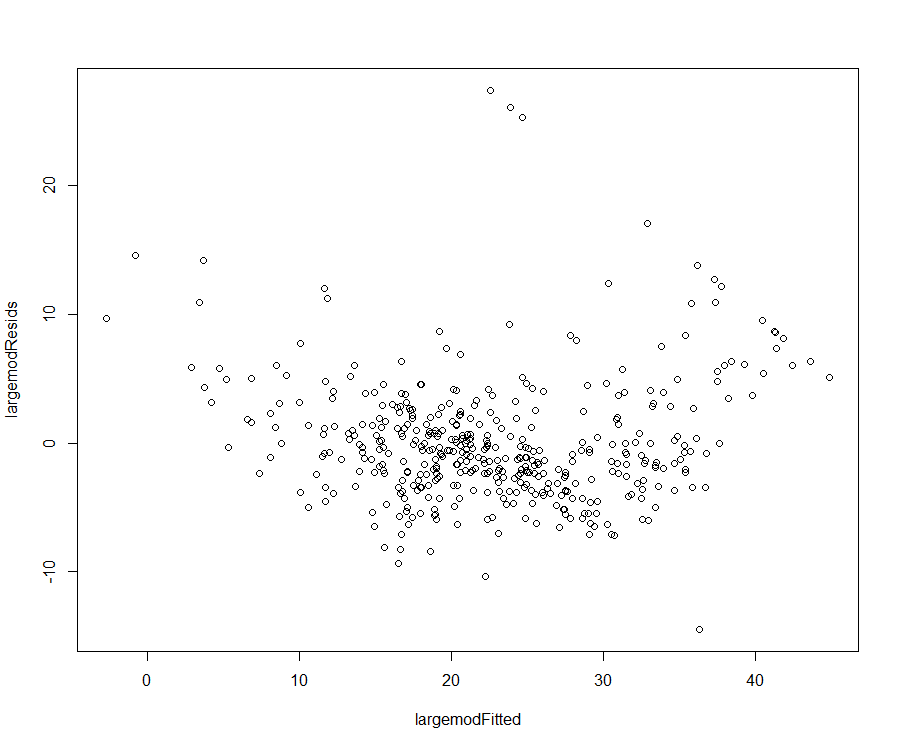
Numerous predictors are used in multiple linear regression. To select the optimal model for predicting the MEDV in this study, many predictors in the data were considered. When evaluating the model's performance, we will consider the p-value and residual error.

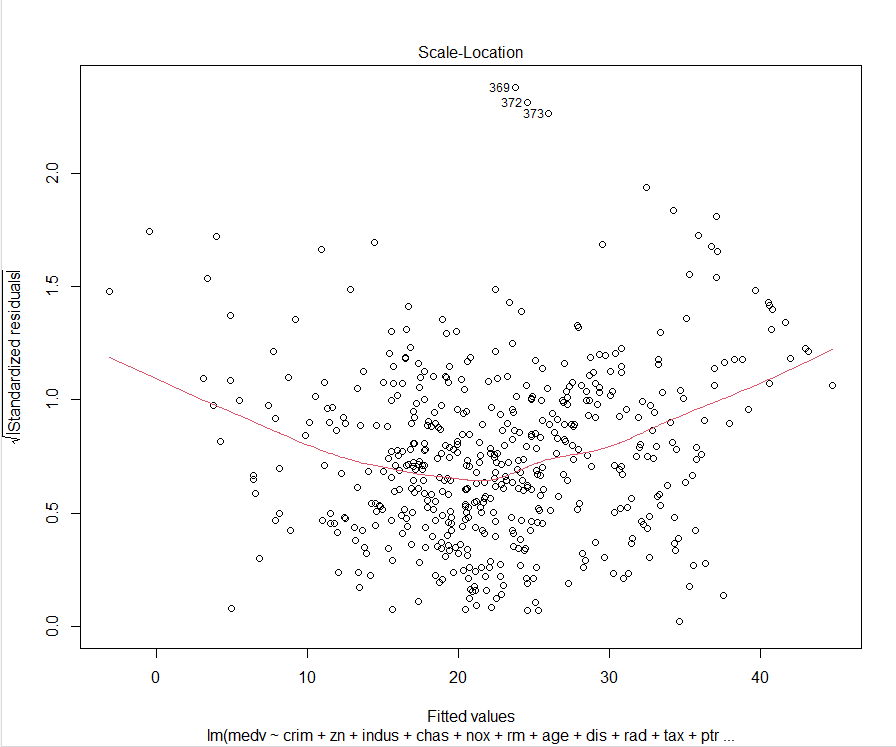
Multiple linear regression's equation is as follows [2]:



The first model is called largemod which has the response variable MEDV and all the other variables as the explanatory variables to predict medv.





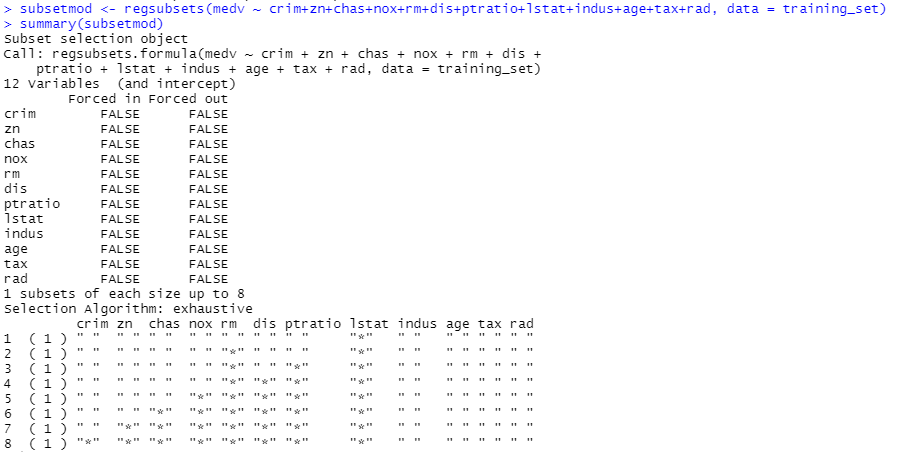


The model follows the assumption of homoscedasticity as seen in the above image.

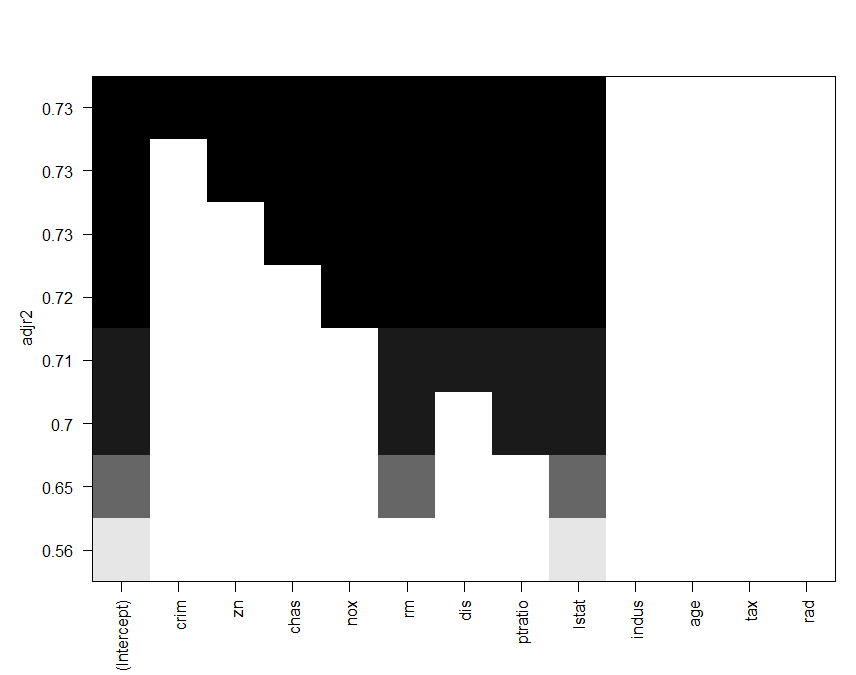
1. **MODEL SELECTION**

In order to determine the subset of variables in the data set that result in the best performing model, that is, a model that decreases prediction error, stepwise regression entails successively adding and removing predictors from the predictive model. Here, we've also tried to employ a technique called backward selection (or backward elimination), which begins with all of the predictors in the model, eliminates the least useful predictors one at a time, and terminates when all of the predictors are statistically significant.[4]

The R function regsubsets() can be used to identify different best models of different sizes.

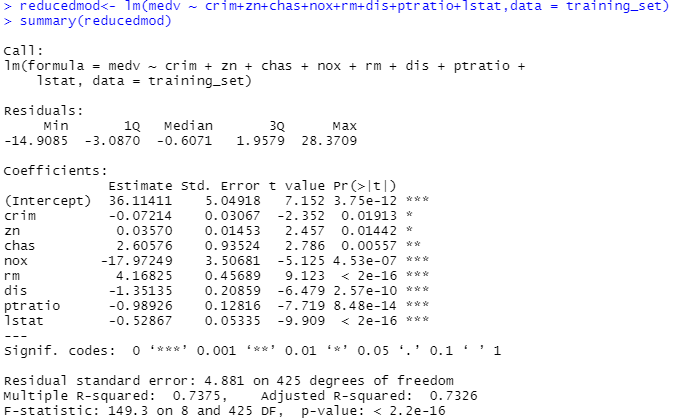


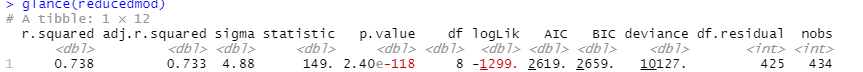
Plotting subset model



By looking at the largemod Signif. code, the following attributes are not significant to the fitting. Hence, we exclude those from the model: INDUS, AGE, RAD, TAX. Fitting the data again with the remaining attributes: CRIM, ZN, CHAS, NOX, DIS, PTRATIO, LSTAT since these are highly significant values as observed from the above plot and signif values of largemod.

* Creating model 2 called reducedmod based on subset selection by removing unwanted variables.

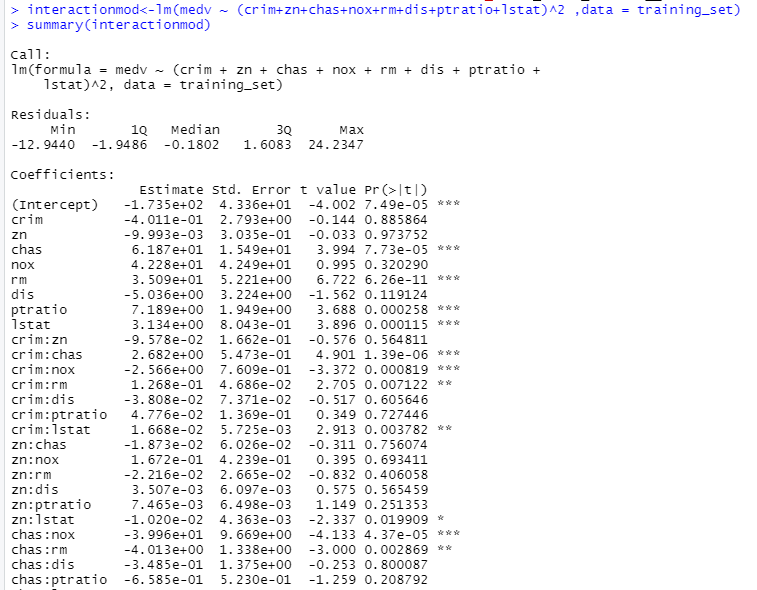


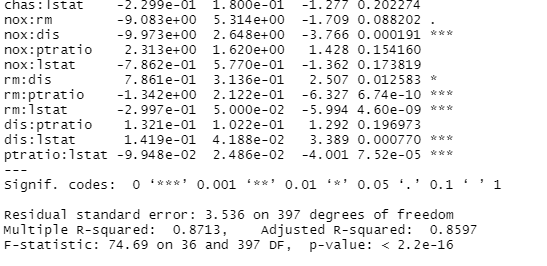


By removing the above-mentioned variables and creating a model, we can see that the AIC value of the model has increased which is not appropriate for improving our model. Hence, we move forward to the next steps to creating a better model.

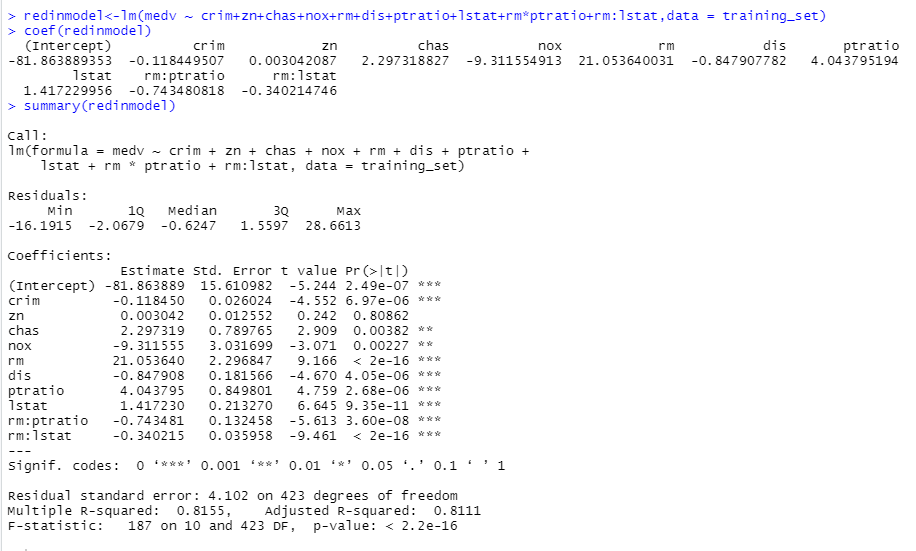
* Introducing interaction terms:

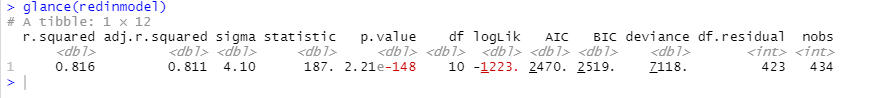
We choose significant interaction terms from the interactmod to improve our model. Interaction means how the effect on Y of a change in an independent variable depends on the value of another independent variable [3]. We square all the terms to find the interaction between them and check if their interactions are significant to the model for improvement.



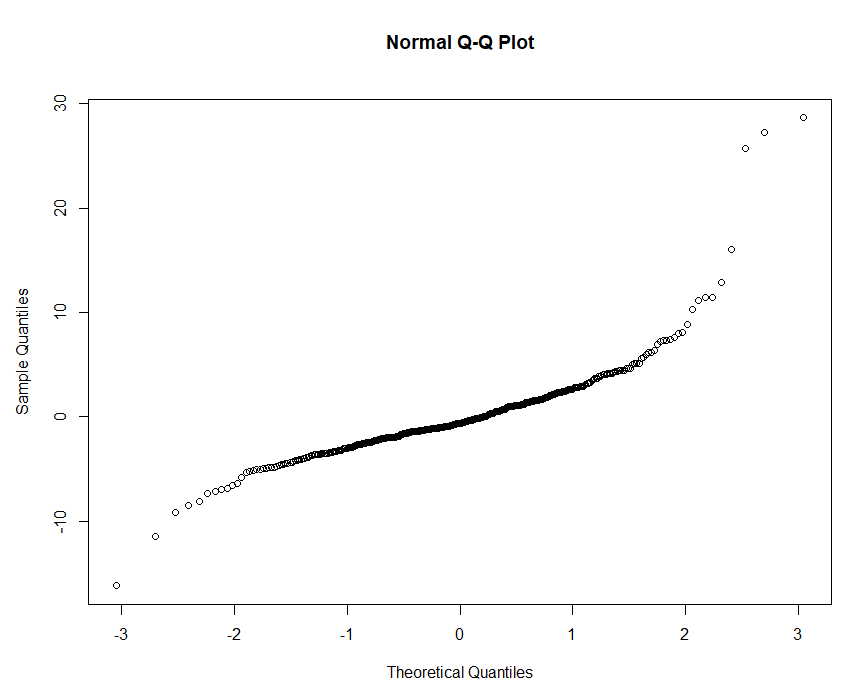


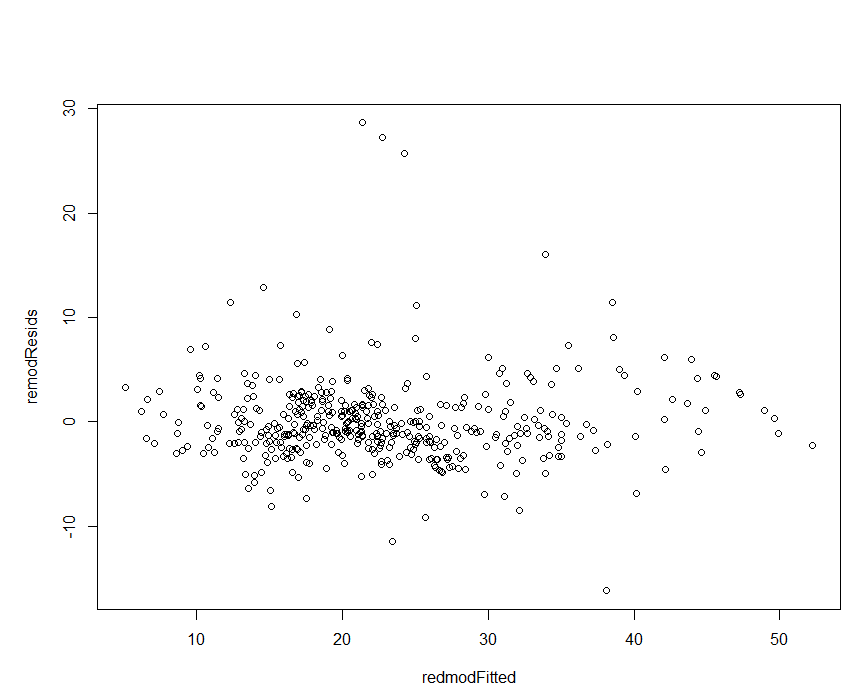
Out of the above, we choose interaction terms rm:lstat and rm:ptratio to be introduced to our model further improve it as their signif code shows that they are highly significant to the model. We make model redinmodel with these values.





As we can see from the above image, the AIC value has significantly decreased for the redinmodel and the adjusted r-squared value has increased dramatically to 0.8111. this shows that the model is the best model out of the 3. The output of this model redinmodel shows that the F-statistics is increased to 187. The predictor has highly significant p values. This model is better than previous models. This is our final model.





The residuals of this model are roughly centered around the zero line and with similar spread on both sides. Therefore, the model probably fits the assumption of heteroscedasticity. The qq-normal plot also justifies normality. The final regression equation for the model is: redinmodel<-lm (medv ~ crim+zn+chas+nox+rm+dis+ptratio+lstat+rm\*ptratio+ rm\*lstat

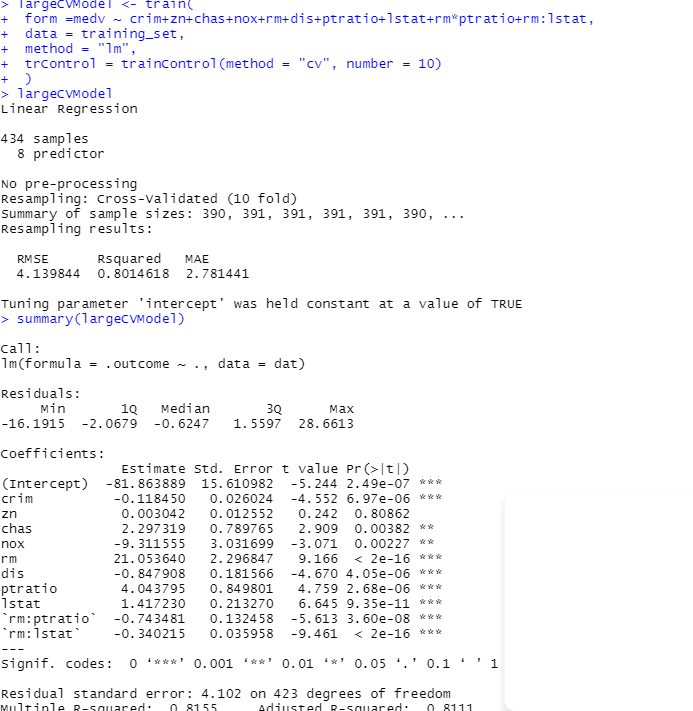
Coefficients are as follows:

|  |  |
| --- | --- |
| variable | coefficient |
| (Intercept) | -81.863889353 |
| Crim | -0.118449507 |
| zn | 0.003042087 |
| chas | 2.297318827 |
| nox | -9.311554913 |
| rm | 21.053640031 |
| dis | -0.847907782 |
| ptratio | 4.043795194 |
| lstat | 1.417229956 |
| rm:ptratio | 0.743480818 |
| rm:lstat | -0.340214746 |

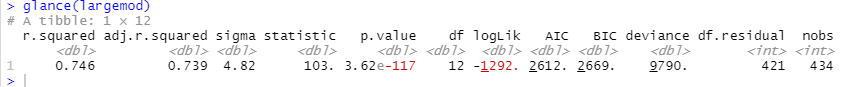
1. **PREDICTION AND SUMMARY**

**Cross Validation**

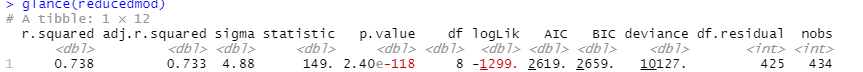
K-fold cross-validation is the process of dividing a dataset into a K number of folds and is utilized to assess the model's performance when presented with new data. K is the total number of groups into which the data sample is divided. As an example, if the k-value is 10, we can refer to this as 10-fold cross-validation. At some stage during the procedure, each fold serves as a testing set. It aids in preventing overfitting. We are aware that the best performance accuracy is achieved when a model is trained using all of the data in a single brief run. We can design a generalized model by avoiding this k-fold cross-validation.

****

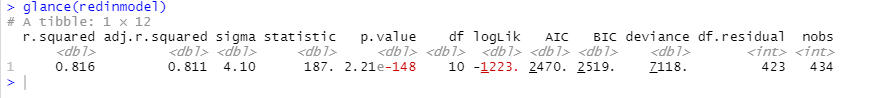
**ADJUSTED R-SQUARED AIC and BIC FOR MODEL1: largemod**

****

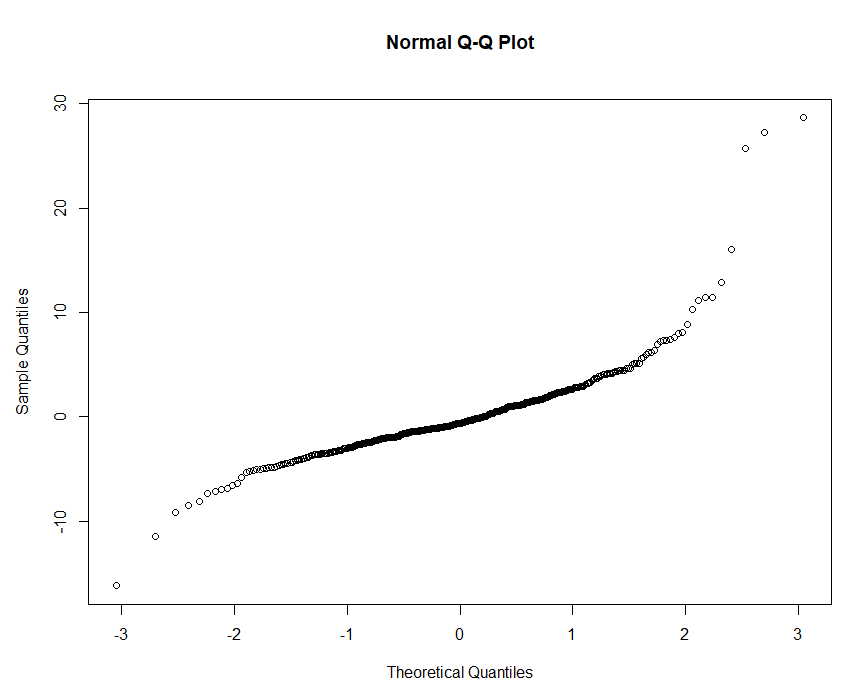
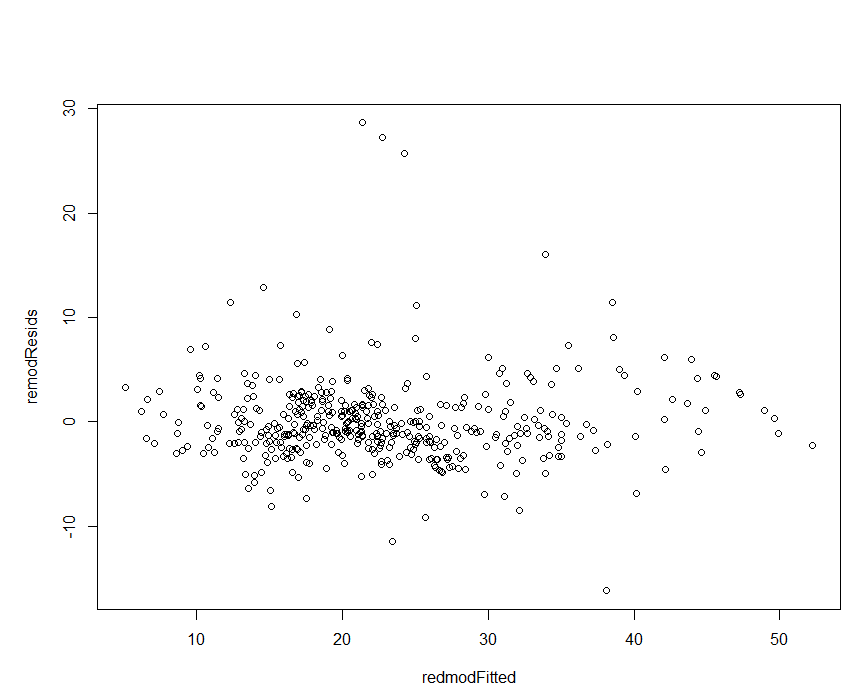
**ADJUSTED R-SQUARED AIC and BIC FOR MODEL1: reducedmod**



**ADJUSTED R-SQUARED AIC and BIC FOR MODEL1: redinmodel**



As we see in the above images, the redinmodel has the lowest value of AIC= 2470 and BIC=2519 and the highest value of adjusted r squared=0.8111 making it the best model among the three.



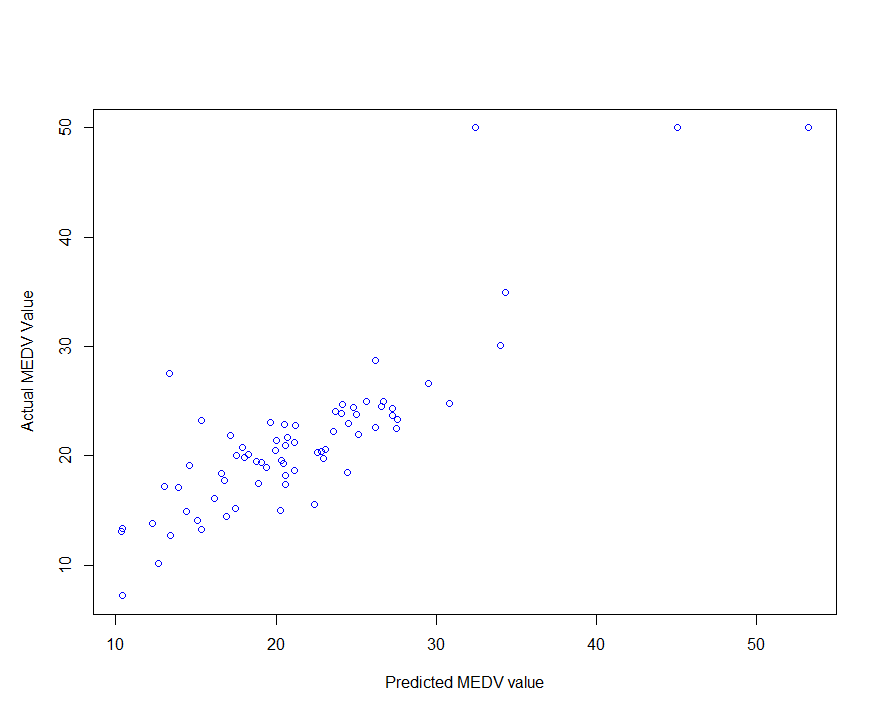
As the residuals show no bias, so we can say our model fits the assumption of homoscedasticity.

**CHECKING ACCURACY**

****

The correlation of the models shows that the accuracy is around 85 %. Let’s check the actual MEDV and predicted MEDV for the test data.

**Plotting Actual vs Predicted Medv values**

****

*The predicted MEDV values majorly lie between [$10,000,$40,000].*

**SUMMARY**

This report set out to identify the factors that contributed most to the variability in housing prices. To remove predictors and irrelevant observations, several statistical techniques were utilized.  When looking at the final model, it makes sense that housing costs are greater in places with lower student-teacher ratios. Homes with additional rooms cost more money. According to the fitted regression model, housing prices decrease as the share of the population with lower socioeconomic class increases. This raises the intriguing possibility that people would want to reside in areas with smaller concentrations of persons of low socioeconomic standing. It is crucial to remember that the data for this report was gathered decades ago. There is no denying that population numbers have changed over the years that follow, thus it would be fascinating to look at how this has affected Boston housing prices in the present.

**REFERENCES**

[1] <https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset/data>.

[2] <https://medium.com/analytics-vidhya/multiple-linear-regression-explained-215f2683cd5a>

[3] <https://www.econometrics-with-r.org/8-2-nfoasiv.html>

[4]<http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/>

[5] <https://www.kdnuggets.com/2022/07/kfold-cross-validation.html>

[6] <https://docs.tibco.com/pub/spotfire/6.5.2/doc/html/prd/prd_available_diagnostic_visualizations.htm>

**APPENDIX**

*Code for the Project:*

#Project1

#Shimpli Borkar

#house price prediction

library(tidyverse)

library(ISLR2)

library(stargazer)

library(caret)

library(leaps)

library(Amelia)

library(ggplot2)

library(lmtest)

library(Hmisc)

library(tseries)

library(caTools)

library(broom)

hodata<-read\_csv("C:\\Users\\Lenovo\\Downloads\\house.csv")

summary(hodata)

view(hodata)

head(hodata)

summary(hodata)

dim(hodata)

sapply(hodata,class)

houseframe<- data.frame(hodata)

missmap(hodata,col=c('yellow','black'),y.at=1,y.labels='',legend=TRUE)

colSums(is.na(houseframe))

cor(hodata$medv, hodata$chas)

cor(hodata$medv, hodata$crim)

cor(hodata$medv, hodata$zn)

cor(hodata$medv, hodata$nox)

cor(hodata$medv, hodata$rm)

cor(hodata$medv, hodata$dis)

cor(hodata$medv, hodata$ptratio)

cor(hodata$medv, hodata$lstat)

cor(hodata$medv, hodata$indus)

cor(hodata$medv, hodata$age)

cor(hodata$medv, hodata$tax)

cor(hodata$medv, hodata$rad)

rcorr(as.matrix(houseframe),type='spearman')

#checking normality

hist.data.frame(hodata)

#checking linearity

plot(hodata)

split = sample.split(hodata$medv, SplitRatio = 0.8)

training\_set = subset(hodata, split == TRUE)

test\_set = subset(hodata, split == FALSE)

view(training\_set)

view(test\_set)

dim((training\_set))

dim(test\_set)

largemod<-lm(medv ~ crim+zn+chas+nox+rm+dis+ptratio+lstat+indus+age+tax+rad ,data = training\_set)

summary(largemod)

glance(largemod)

largemodResids <- largemod$residuals

largemodFitted <- largemod$fitted.values

plot(largemodFitted,largemodResids)

hist(largemodResids)

qqnorm(largemodResids)

stargazer(largemod, type="text")

subsetmod <- regsubsets(medv ~ crim+zn+chas+nox+rm+dis+ptratio+lstat+indus+age+tax+rad, data = training\_set)

summary(subsetmod)

plot(subsetmod, scale = "adjr2" )

reducedmod<- lm(medv ~ crim+zn+chas+nox+rm+dis+ptratio+lstat,data = training\_set)

summary(reducedmod)

glance(reducedmod)

anova(reducedmod,largemod)

#pvalue should be less than 0.5 here

interactionmod<-lm(medv ~ (crim+zn+chas+nox+rm+dis+ptratio+lstat)^2 ,data = training\_set)

summary(interactionmod)

redinmodel<-lm(medv ~ crim+zn+chas+nox+rm+dis+ptratio+lstat+rm\*ptratio+rm\*lstat,data = training\_set)

coef(redinmodel)

summary(redinmodel)

anova(redinmodel,interactionmod)

remodResids <- redinmodel$residuals

redmodFitted <- redinmodel$fitted.values

plot(redmodFitted,remodResids)

hist(remodResids)

qqnorm(remodResids)

glance(redinmodel)

largeCVModel <- train(

form =medv ~ crim+zn+chas+nox+rm+dis+ptratio+lstat+rm\*ptratio+rm:lstat,

data = training\_set,

method = "lm",

trControl = trainControl(method = "cv", number = 10)

)

largeCVModel

summary(largeCVModel)

predictionvalue<-predict(largeCVModel,test\_set)

plot(predictionvalue,test\_set$medv,xlab="Predicted MEDV value", ylab="Actual MEDV Value",col="Blue")

test\_set$predictionvalue<-predict(largeCVModel,test\_set)

acpred<-data.frame(test\_set$medv,test\_set$predictionvalue)

names(acpred)<-c("medv","predictionvalue")

correlation\_accuracy<-cor(acpred)

correlation\_accuracy

predict(largeCVModel, test\_set, interval="predict")