Prediction of House Prices

A Regression Analysis on Boston Housing Data

Presented by:

Rashmi Prathigadapa

MS – BANA (Data Mining)

Executive Summary

In this analysis, we are going to build a linear regression model to predict the values of a dependent variable based on values of independent variables present in the dataset. The dataset that we have considered for our analysis is the **'Housing Values in Suburbs of Boston'**. The Boston housing dataset has median value of the house along with other parameters that could potentially be related to housing prices. These are the factors such as socio-economic conditions, environmental conditions, educational facilities and some other similar factors. There are 506 observations in the data for 14 variables including the median price of house in Boston. There are 12 numerical variables in our dataset and 1 categorical variable. The aim of this project is to build a linear regression model estimate the 'medv' (median price of owner-occupied homes in Boston).

We have started with the exploratory data analysis. The summary statistics of the dataset were first observed to understand the range of values and nature of the variables in the dataset. To understand the distribution of the values, boxplots were plotted, and it is observed that there are outliers. The correlation matrix was then observed, and we notice that housing value has a strong positive correlation with rm (number of rooms). It is expected, as a spacious house with more rooms would have a higher valuation. Also, 'medv' has a strong negative correlation with lstat (lower status of the population). It means that a house in an area with lower socioeconomic status naturally has a lower value and vice versa.

We have built an initial model without any transformation on the data and then standardized the values to scale them for the multivariate regression analysis. Using these variables without standardization in effect gives the variable with the larger range a higher weight in the analysis. Transforming the data to comparable scales can prevent this problem. These results were later compared with the results generated from other regression techniques to understand the accuracy of the predictions.

The different methods used for variable selection were best subset selection, forward selection, backward elimination, stepwise selection and LASSO regression. The results from the different variable selection methods were compared to come up with the best fit model to predict the median housing values in Boston. The models generated similar results and there was not a considerable difference between them. LASSO suggested a 10-variable model whereas all other techniques suggested 11 variable models. Below is the best fit model obtained:

Finally, residual diagnostics were performed on the best fit model and the assumptions of linear regression modeling were checked.

Exploratory Data Analysis

Data Summary:

The dataset 'Housing Values in Suburbs of Boston' consists of 506 observations and 14 variables. The dependent variable is 'medv', median price of houses in Boston according to 1974 census and rest of them are considered as the independent variables. Out of these 13 variables, 12 are numerical and 1 variable categorical variable.

Below is the data dictionary:

Variable Name	Description	Variable Type
age	proportion of owner-occupied units built prior to 1940.	Integer
black	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.	Integer
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).	Integer
crim	per capita crime rate by town.	Integer
dis	weighted mean of distances to five Boston employment centers.	Integer
indus	proportion of non-retail business acres per town.	Integer
lstat	lower status of the population (percent).	Integer
medv	median value of owner-occupied homes in \$1000s.	Integer
nox	nitrogen oxides concentration (parts per 10 million).	Integer
ptratio	pupil-teacher ratio by town.	Integer
rad	index of accessibility to radial highways.	Integer
rm	average number of rooms per dwelling.	Integer
tax	full-value property-tax rate per \$10,000.	Integer
zn	proportion of residential land zoned for lots over 25,000 sq.ft.	Integer

Summary Statistics:

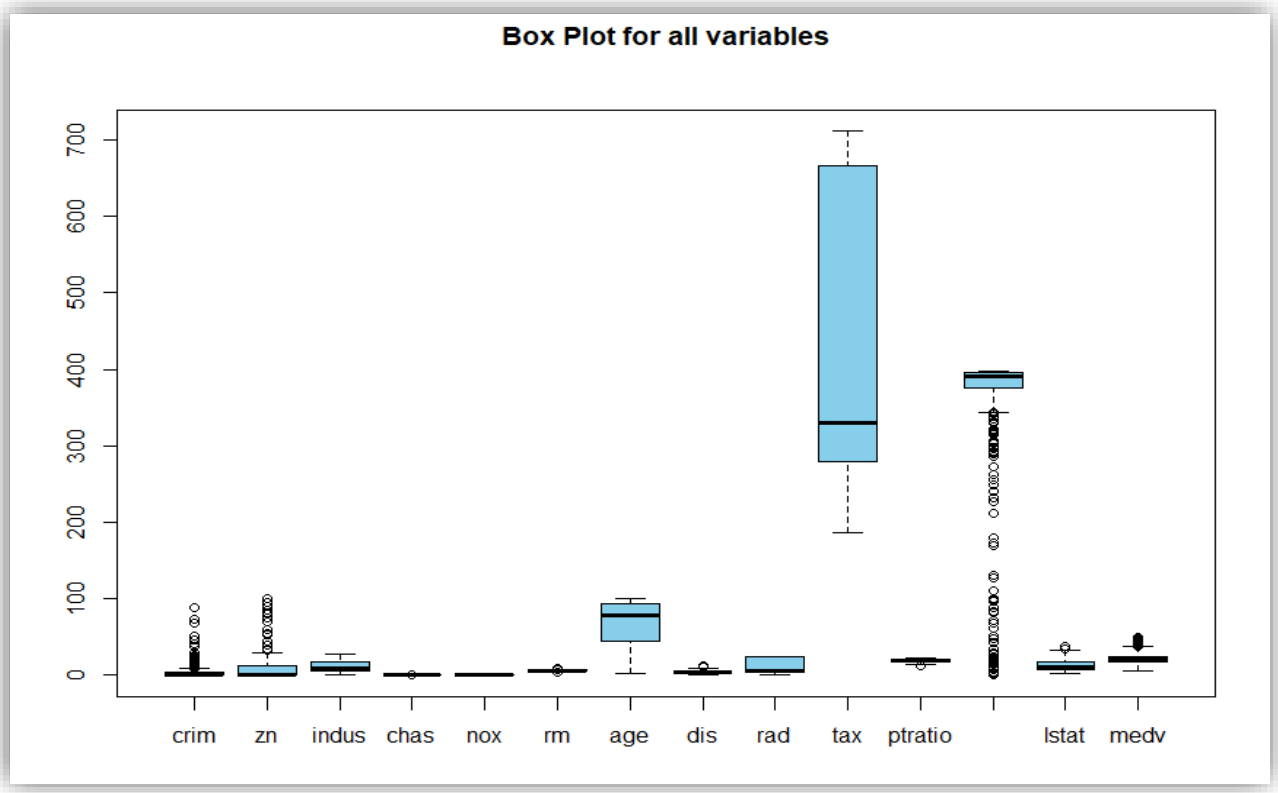
Below table shows the summary statistics and distribution of the observations in each variable. There are no missing values in the data. But, the summary statistics suggest that there might be outliers in some of the variables.

crim	zn	indus	chas	nox	rm	age
Min. : 0.00632	Min. : 0.00	Min. : 0.46	Min. :0.00000	Min. :0.3850	Min. :3.561	Min. : 2.90
1st Qu.: 0.08204	1st Qu.: 0.00	1st Qu.: 5.19	1st Qu.:0.00000	1st Qu.:0.4490	1st Qu.:5.886	1st Qu.: 45.02
Median : 0.25651	Median : 0.00	Median : 9.69	Median :0.00000	Median :0.5380	Median :6.208	Median : 77.50
Mean : 3.61352	Mean : 11.36	Mean :11.14	Mean :0.06917	Mean :0.5547	Mean :6.285	Mean : 68.57
3rd Qu.: 3.67708	3rd Qu.: 12.50	3rd Qu.:18.10	3rd Qu.:0.00000	3rd Qu.:0.6240	3rd Qu.:6.623	3rd Qu.: 94.08
Max. :88.97620	Max. :100.00	Max. :27.74	Max. :1.00000	Max. :0.8710	Max. :8.780	Max. :100.00

dis	rad	tax	ptratio	black	lstat	medv
Min. : 1.130	Min. : 1.000	Min. :187.0	Min. :12.60	Min. : 0.32	Min. : 1.73	Min. : 5.00
1st Qu.: 2.100	1st Qu.: 4.000	1st Qu.:279.0	1st Qu.:17.40	1st Qu.:375.38	1st Qu.: 6.95	1st Qu.:17.02
Median : 3.207	Median : 5.000	Median :330.0	Median :19.05	Median :391.44	Median :11.36	Median :21.20
Mean : 3.795	Mean : 9.549	Mean :408.2	Mean :18.46	Mean :356.67	Mean :12.65	Mean :22.53
3rd Qu.: 5.188	3rd Qu.:24.000	3rd Qu.:666.0	3rd Qu.:20.20	3rd Qu.:396.23	3rd Qu.:16.95	3rd Qu.:25.00
Max. :12.127	Max. :24.000	Max. :711.0	Max. :22.00	Max. :396.90	Max. :37.97	Max. :50.00

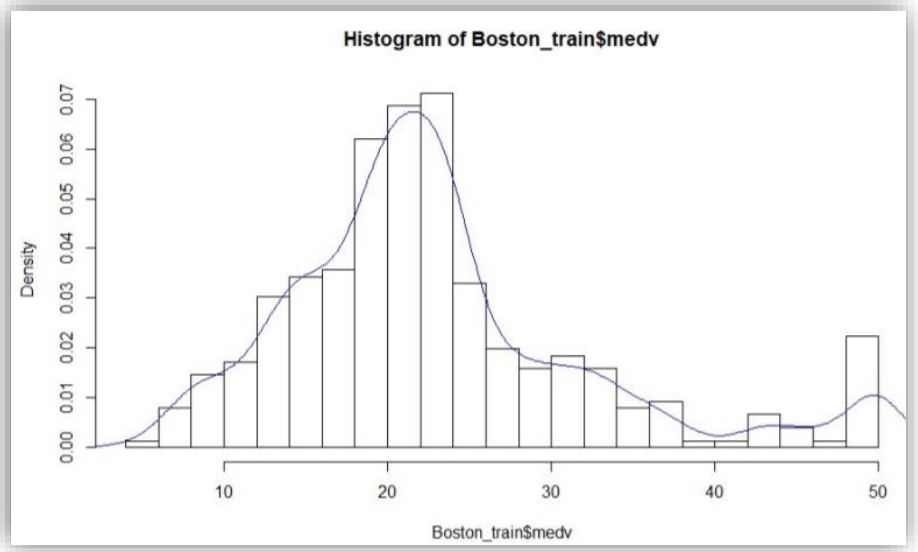
As a part of this problem, we would like to predict the response of variable "Medv against the other 13 variables. To begin, let's understand the distribution of this variable.

Box Plots: To further investigate about the outliers, let us look at the variation in the values by box plots for each of these variables.



It can be seen that, there are outliers in the column 'black' (proportion of blacks by town), 'crim' (per capita crime rate by town) and 'zn' (proportion of residential land zoned for lots over 25,000 sq.ft). The highest variability is seen in the column tax (full-value property-tax rate per $10,000).

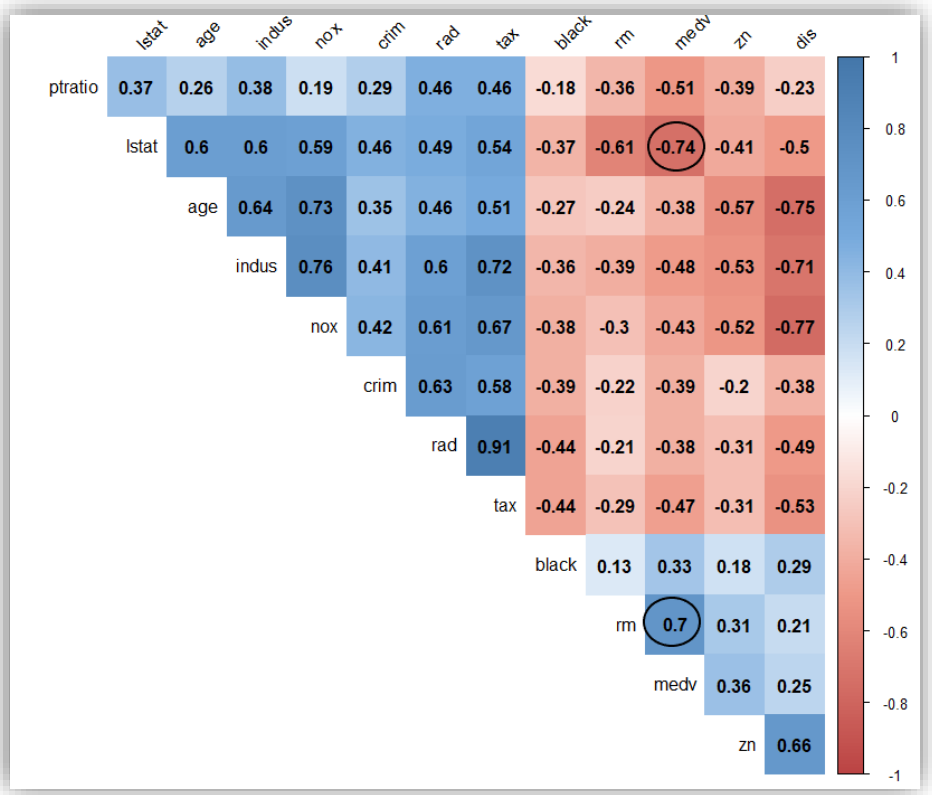
Distribution of the response variable 'medv':



The variable 'Medv' follows a normal distribution and is skewed to the right with mean value greater than the median value. This is also evident in the Boxplot that there are some outliners on the upper outlier limit and just one outliner on the lower outlier limit. With this understanding of the response variable, we conduct a pairwise correlation on the training data set.

Pairwise Correlations:

From the correlations below, we can see that our dependent variable 'medv' is highly correlated with the variables 'lstat' and 'rm'. Also, the accessibility to highways 'rad' is highly correlated to the 'tax' rates implying that the as the accessibility to highways increases the amount of taxes paid on the houses is positively correlated.



The highest correlation is observed between the variables "rad" and "tax" with a coefficient of 0.91. Also, the variable "chas" shows no significant correlation between any of the other variables. We also observe high correlation values of 0.77 (dis~nox), 0.75 (dis~age), 0.76 (nox~indus) & 0.73 (age~nox). It will be important to note these correlations before running the linear regression model.

Linear Regression

1. **Full Model (13 variables):** Let's perform linear regression on the dataset. First, we segregate the entire data into 75% training and 25% test data. We built model on the training data and check the performance of our model using the test data. The variable of our interest here is 'medv'. Our main objective is to predict the value of 'medv' based on other independent variables present in the dataset. We will first perform multiple linear regression using all the variables. The equation obtained for full model:

medv=42.67-0.97***crim**+1.27***zn**+0.02***indus**-2.1***nox**+2.05***rm**+0.23***age**-3.2***dis**+2.7***rad**-1.6***tax**-2.2***ptratio**+0.88***black**-4.07***lstat**+3.1***chas**

From this equation, we can say that by keeping all the variables constant and increasing crime rate by 1 unit, would reduce the median house price by 0.97 units. It is observed that, the p-value is not statistically significant for the variables 'indus' and 'age' in the full model, so we have eliminated them in our next model.

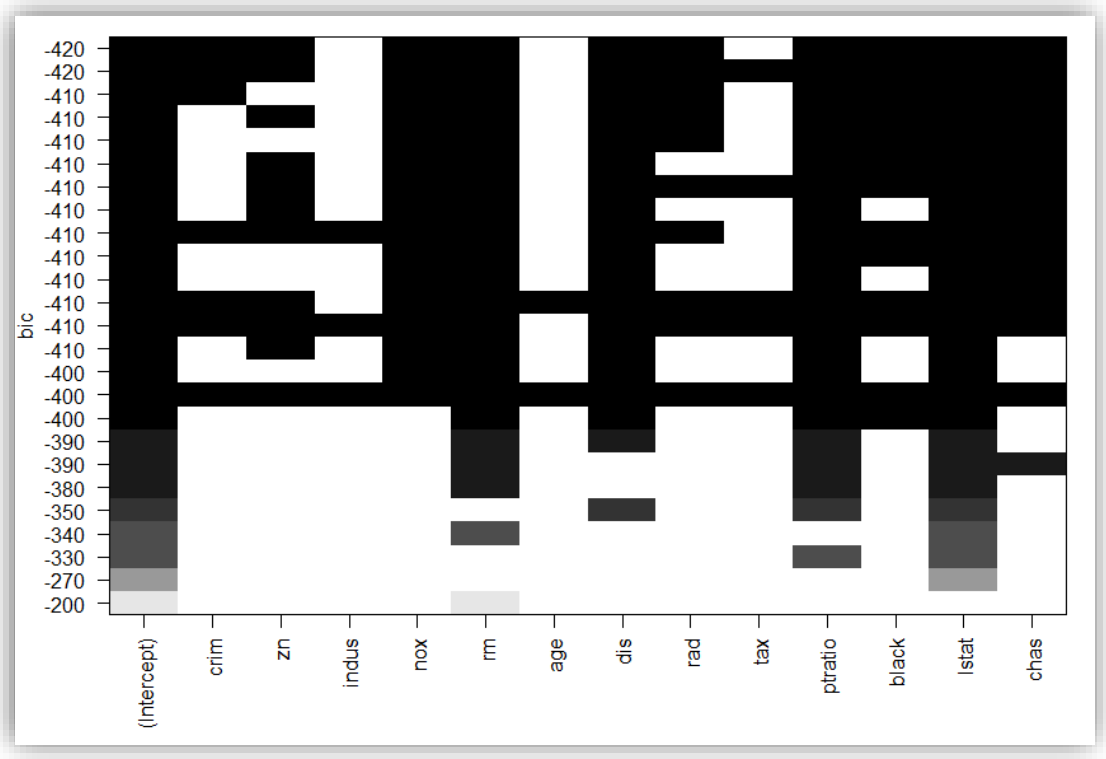
2. **P-value Model (11 variables):** By eliminating the insignificant variables 'indus' and 'age' from the full model, we now have 11 variables in our dependent variables. All of them are statistically significant. The adjusted R-squared, AIC and BIC values for this model have improved. The equation obtained for p-value model:

medv=42.35-0.97***crim**+1.24***zn**-2.02***nox**+2.08***rm**-3.28***dis**+2.7***rad**-1.6***tax**-2.2***ptratio**+0.89***black**-4***lstat**+3.14***chas**

Linear Regression using Variable Selection techniques

Now, we will now apply different techniques of variable selection to decide the best fit model.

3. **Best Subsets Model (10 variables):** Here we apply the best subset selection approach to the Boston housing data. We wish to predict the median value of a house based on the available predictors. We decide on the number of variables that are required in the model using AIC, BIC, r squared, and adjusted r squared.



From this plot, we can see that indus, age and tax are on the top row with the least BIC value. So, model is built using these 10 variables. The adjusted R squared value has reduced compared to previous 2 models. The BIC value of the model also reduced but Mean Square Error value slightly increased and Mean squared prediction error has increased. Below is the equation obtained:

medv=41-0.95***crim**+1.06***zn**-2.27***nox**+2.17***rm**-3.15***dis**+1.49***rad**-2.3***ptratio**+0.92***black**-4.04***lstat**+3.27***chas**

If there are 'n' independent variables, the number of possible nonempty subsets is 2^n - 1. If we try a best subset regression with more than 50 variables, we might need to wait for a lot of time, so this is not an efficient method.

4. **Forward Selection(11 variables):** Forward stepwise selection is a computationally efficient alternative to best subset selection. While the best subset selection procedure considers all 2p possible models containing subsets of the p predictors, forward stepwise considers a much smaller set of models.

Best model selected by forward selection method is given below:

medv=42.36-0.97***crim**+1.25***zn**-2.02***nox**+2.08***rm**-3.28***dis**+2.7***rad**-1.6***tax**-2.2***ptratio**+0.89***black**-4***lstat**+3.14***chas**

This is same as the p-value model, with 11 dependent variables.

5. **Backward Elimination(11 variables):**Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection. However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Best model selected by backward elimination method is given below.

medv=42.35-0.97***crim**+1.24***zn**-2.02***nox**+2.08***rm**-3.28***dis**+2.7***rad**-1.6***tax**-2.2***ptratio**+0.89***black**-4***lstat**+3.14***chas**

This is also same as the p-value model and Forward selection model with 11 dependent variables.

6. **Stepwise Selection:** Stepwise selection has the flexibility to both add/ remove variables in the process of selecting the best regression model. Best model selected by stepwise selection method is given below.

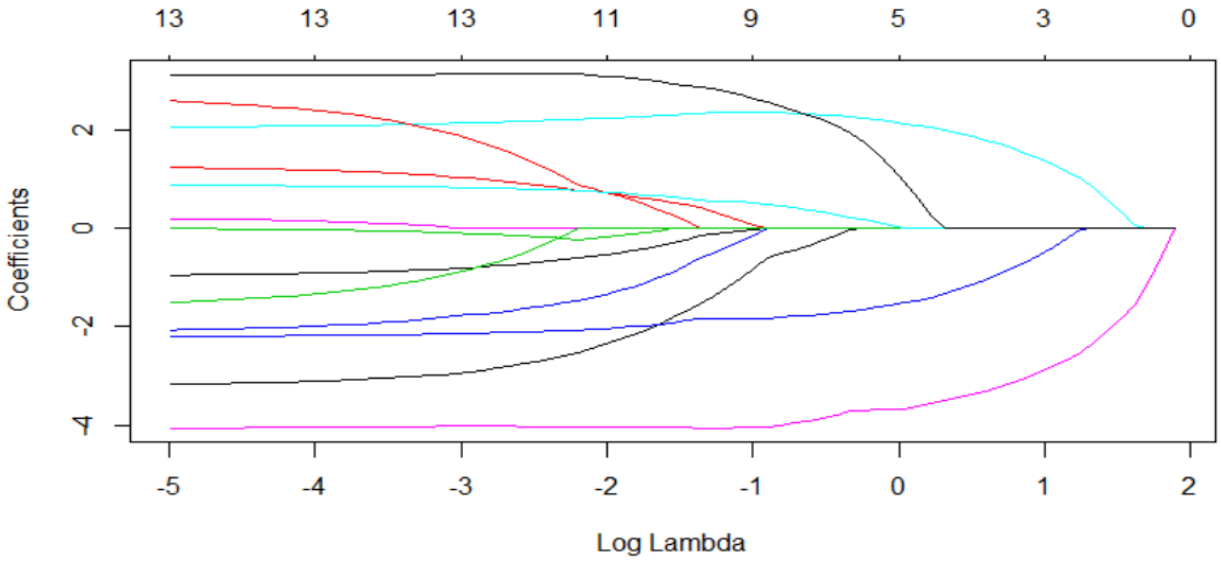
medv=42.35-0.97***crim**+1.24***zn**-2.02***nox**+2.08***rm**-3.28***dis**+2.7***rad**-1.6***tax**-2.2***ptratio**+0.89***black**-4***lstat**+3.14***chas**

Even the stepwise gives same 11 dependent variables model.

Lasso Regression

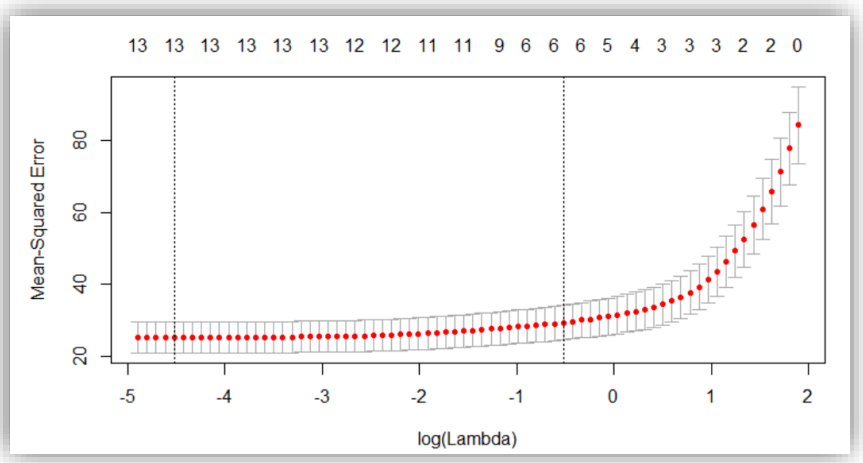
7. Lasso Regression:

Lasso regression is a type of **linear regression** that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. "LASSO" stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.



The above plot shows the coefficients as the log lambda value increases, we can see that as the value increase, number of on-zero coefficients decrease. Thus, leads to sparse solutions. In Lasso, we need to determine the model that gives best lambda value.

Using cross-validation we find appropriate lambda value using error versus lambda plot. We take the value with the least error as well as the error value which is one standard deviation away from the lowest error value. we then build models based on both. For the higher error value, the number of variables selected decreases.



In our Lasso model, we have considered the lambda which gives minimum cross validation error. Below is the obtained equation:

medv=41.16-0.9***crim**+1.17***zn**-0.03***indus**-1.97***nox**+2.09***rm**-0.14***age**-3.09***dis**+2.36***rad**-1.29***tax**-2.17***ptratio**+0.86***black**-4.05***lstat**+3.12***chas**

Model Comparison

We have compared the MSE, R squared, Adjusted R squared, MSPE and AIC BIC values for all the above models.

- The lasso model shows least MSE value compared to other models.
- The R-squared and Adjusted R squared values are almost the same for all models. Slightly less for the stepwise 11 variables model.
- The prediction error is least for 11 variables model, which was chosen in P-value and Stepwise selection models.
- So, there is a tie-off between Lasso (13 variables) and stepwise 11 variables model, we would like to chose 11 variables model because of the least prediction error.

Model Comparison	FULL Model	P-Value Model	Best subsets Model	Forward selection	Backward Elimination	Stepwise Selection	Lasso Model
MSE	24.14	24.03	24.30	24.03	24.03	24.03	23.28
R-Squared	0.72	0.72	0.72	0.72	0.72	0.72	0.72
Adjusted R-Sq	0.71	0.71	0.71	0.71	0.71	0.71	0.72
Test MSPE	19.41	19.33	20.45	19.33	19.33	19.33	19.44
AIC	2298.00	2294.31	2297.51	2294.31	2294.31	2294.31	N/A
BIC	2357.06	2345.50	2344.76	2345.50	2345.50	2345.50	N/A

Residual Plots

Residual diagnosis needs to be performed to check the validity of the assumptions made for building the regression model. We draw residual plots for the stepwise selection model.

From the residual plots, it can be inferred that:

i) There is a slight u shape and the variance doesn't completely look constant.
ii) Q-Q plot suggests normal data but with a right skewed tail.
iii) This plot contains equally spread points and therefore, homoscedastic in nature
iv) There are is an outlier but that doesn't tend to influence the model, so we can keep it.