**Exploring IRIS Data set with EDA**

**DATA MINING**

CASE: Boston Housing

Table of Contents

[Executive Summary for Boston Housing Data 3](#_Toc536494639)

[Exploratory Data Analysis 3](#_Toc536494640)

[Linear Regression 6](#_Toc536494641)

[Variable Selection 7](#_Toc536494642)

[Best Subset 7](#_Toc536494643)

[Stepwise Regression 7](#_Toc536494644)

[LASSO Regression 10](#_Toc536494645)

[Conclusion: 11](#_Toc536494646)

[Executive Summary for Q2 Simulation Study 11](#_Toc536494647)

[Problem & Approach 11](#_Toc536494648)

[Major Results 11](#_Toc536494649)

[Executive Summary for Monte Carlo Simulation Study 13](#_Toc536494650)

[Primary Setup 13](#_Toc536494651)

[Monte Carlo Simulation Study 13](#_Toc536494652)

# Executive Summary for Boston Housing Data

**Problem and Approach**

We conducted our analysis on the Boston housing dataset using randomly distributed 75% of the original dataset to arrive at the best fitted model using linear regression for the response variable MEDV. Exploratory data analysis was conducted on the dataset which involved nature of variables, box plot was used to determine any outliers and correlation matrix was used to identify any correlation. Linear regression was performed on the model and different variable selection techniques were used like Subset selection, Stepwise regression and LASSO regression.

**Major findings**

The best model was obtained using stepwise regression as below:

**Model <-** medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim + rad + tax

# Exploratory Data Analysis

The dataset consists of 506 observations of 14 attributes. The median value of house price, represented by MEDV, is the the dependent variable. The data gives values for various features of different suburbs of Boston as well as the median-value for homes in each suburb. The features were chosen to reflect different factors influencing the price of houses including the structure of the house (age and spaciousness), the quality of the neighborhood, transportation access to employment centers and highways, and pollution.

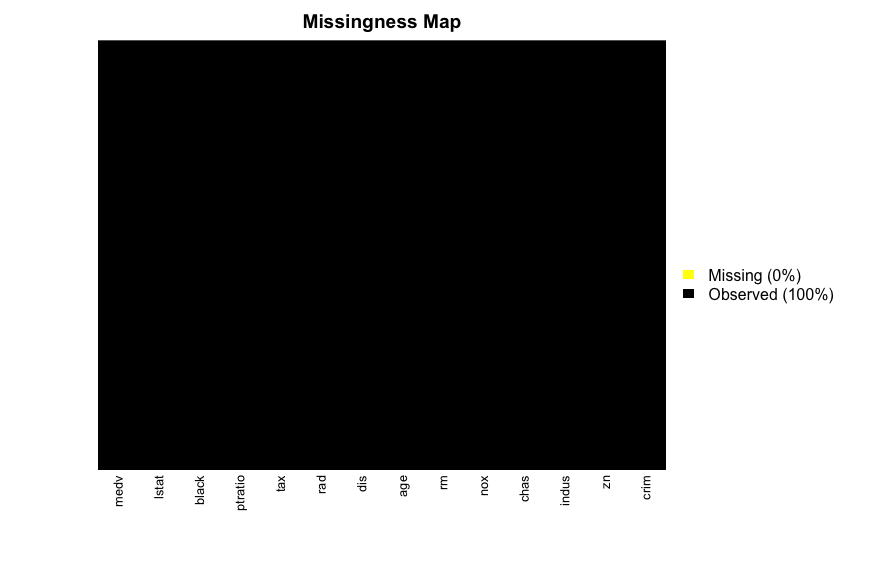
Below is a brief description of the variables in the dataset:

**Table 1: Variable Description**

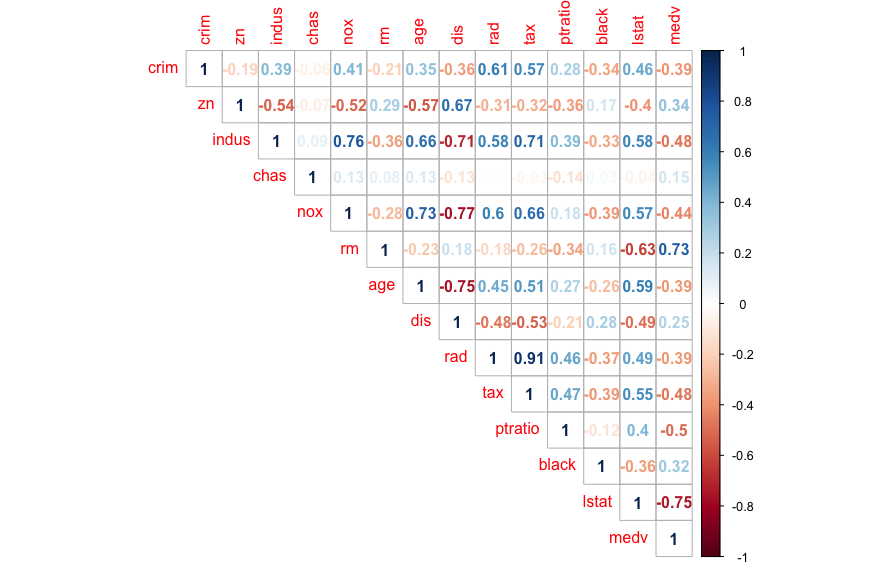
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Nature of variable** | **Type** | **Notes** |
| medv | Housing value - Response variable/ variable of interest | Numerical | Median value of owner-occupied homes |
| rm | Housing structure related | Numerical | Average number of rooms in owner units |
| age | Housing structure related | Numerical | Proportion of units built prior to 1940 |
| black | Neighborhood related | Numerical | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town |
| lstat | Neighborhood related | Numerical | Proportion of neighboring population that is of lower status |
| crim | Neighborhood related | Numerical | Crime rate by town |
| zn | Neighborhood related | Numerical | Proportion of land zoned for commercial construction |
| indus | Neighborhood related | Numerical | Proportion of industrial area acres |
| tax | Neighborhood related | Numerical | Full value property tax rate |
| ptratio | Neighborhood related | Numerical | Pupil-teacher ratio by town district |
| chas | Neighborhood related | Integer (Binary) | 1 if land tract bounds Charles River; 0 otherwise |
| dis | Accessibility related | Numerical | Weighted distances to employment centres |
| rad | Accessibility related | Integer | Index of highway accessibility |
| nox | Air pollution related | Numerical | Nitrogen oxide concentrations in pphm |

There are no missing or duplicate values which can be seen from the diagram below:

**Figure 1: Missing Value Map**



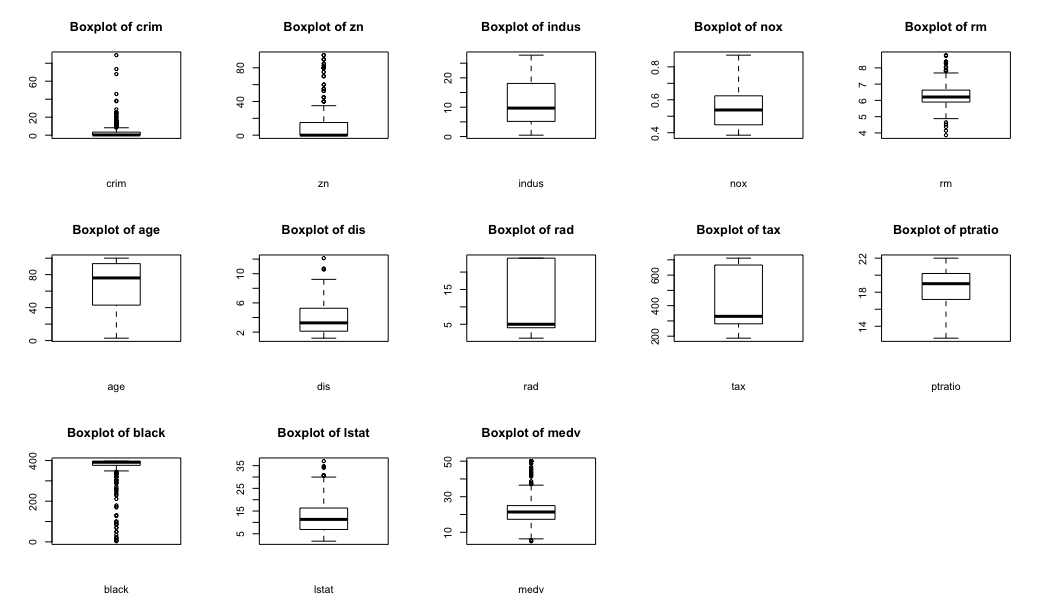
**Figure 2: Correlation Matrix**



From correlation matrix, some of the observations made are as follows:

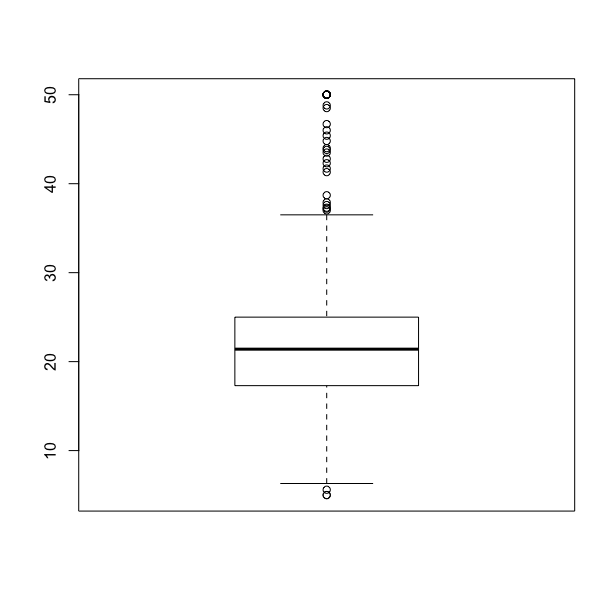
1. Median value of owner-occupied homes (in 1000$) increases as average number of rooms per dwelling increases and it decreases if percent of lower status population in the area increases
2. nox or nitrogen oxides concentration (ppm) increases with increase in proportion of non-retail business acres per town and proportion of owner-occupied units built prior to 1940.
3. rad and tax have a strong positive correlation of 0.91 which implies that as accessibility of radial highways increases, the full value property-tax rate per $10,000 also increases.
4. crim is strongly associated with variables rad and tax which implies as accessibility to radial highways increases, per capita crime rate increases.
5. indus has strong positive correlation with nox, which supports the notion that nitrogen oxides concentration is high in industrial areas.
6. We notice that the housing value has a strong positive correlation with rm as expected, as more spacious house with more rooms would have a higher valuation. medv has a strong negative correlation with lstat meaning an area with lower socioeconomic status naturally has a lower value.
7. The feature with the least correlation to MEDV is the proximity to Charles River, CHAS.

**Figure 3: Boxplot of all variables**



From the boxplot above, it can be inferred that the variables crim,zn,rm,dis,black,lstat and medv have outliers present.

**Figure 4: Boxplot of MEDV response variable**

****

It can be observed from the boxplot above that the response variable MEDV has many outliers outside the upper limit and few outliers below the lower limit.

# Linear Regression

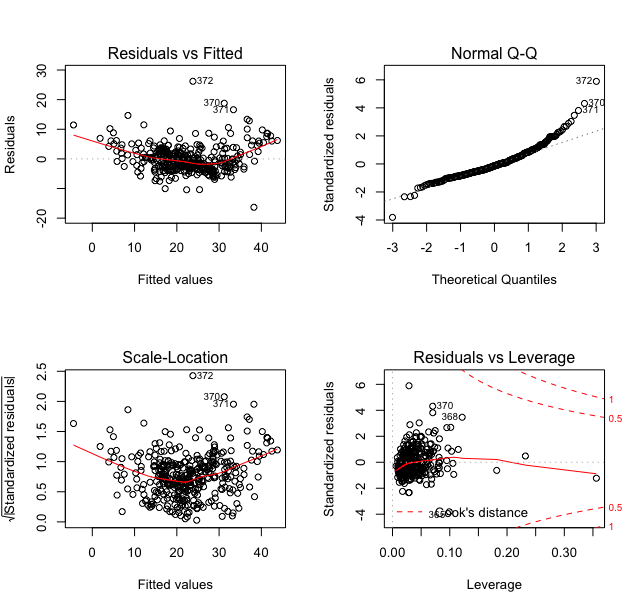
We conduct a linear regression on the data set with all the variables without any variable transformation in order to understand the relation among the independent variable and the dependent variable MEDV.

We notice that Indus and age have very high p-value and seem to be non-significant.

**Table 2: Summary Statistics for Model1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model1** | **AIC** | **BIC** | **MSE** | **Ajusted R2** | **R2** |
| Medv ~ . --> All Covariates | 2233.687 | 2292.75 | 19.62 | 0.7544 | 0.7629 |

**Figure 5: Residual Diagnosis of Model1**

****

1. The Q-Q plot suggests potential outliers present in the dataset.
2. The residual density is mostly around zero and any specific pattern cannot be interpreted
3. This model can explain the variability of around 75% in the dependent variable MEDV
4. Few observations have a large crook’s distance.
5. The AIC and the BIC value are pretty close to each other.
6. Model 1 is not suitable for the prediction of the MEDV variable.

# Variable Selection

The following variable selection methods have been performed on the above model:

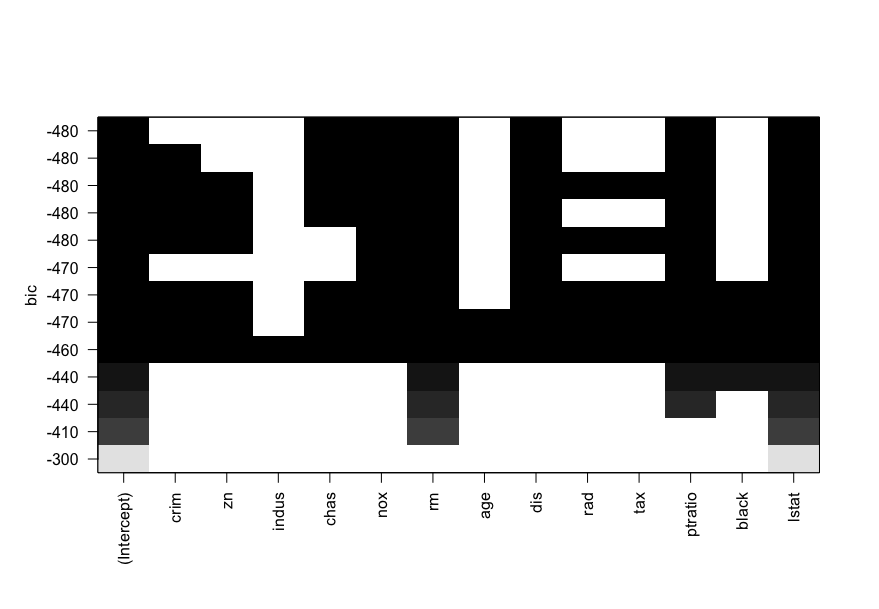
1. Best Subset
2. Stepwise
3. LASSO

## Best Subset

Best subset (13 variable) and stepwise (forward, backward, both) techniques of variable selection were used to come up with the best linear regression model for the dependent variable medv.

The output using the Best subset has been captured and displayed below:

**Figure 6: Best Subset BIC**

****

**Observations:**

**1. Variable included in the variable selection process is “lstat” followed by “rm”**

**2. Variable not included in the variable selection process is “age” followed by “indus”**

**3. “nox” is the second-best variable to be included in the selection process followed by “indus”**

The bic and adjusted r squared values in the above table suggests that the model with 11 variables is the best model. We arrive at the model **medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + black + lstat** as our best model, with 11 predictors.

**4. As we have a high number of predictors here, we should proceed with stepwise method for variable selection**

## Stepwise Regression

The results of the stepwise regression using the forward method have been captured in the table below.

It can be observed that with each step adding a covariate to the model resulted in decrease in the AIC value. The final model resulting through this process has been highlighted below.

We will be conducting the stepwise regression using the backward methodology to verify the same.

**Table 3: Stepwise Regression using Forward**

|  |  |
| --- | --- |
| **Model** | **AIC** |
| medv ~ lstat | 1851.01 |
| medv ~ lstat + rm | 1735.58 |
| medv ~ lstat + rm + ptratio | 1678.13 |
| medv ~ lstat + rm + ptratio + dis | 1661.39 |
| medv ~ lstat + rm + ptratio + dis + nox | 1663.47 |
| medv ~ lstat + rm + ptratio + dis + nox + chas | 1621.97 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black | 1612.47 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn | 1606.31 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim | 1604.19 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim + rad | 1596.1 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim + rad + tax | 1585.76 |

The results of the stepwise regression using the backward method have been captured in the table below.

**Using the backward selection method, as the covariates are removed in each step resulting in lower AIC value, we get the same model as obtained using the forward step methodology. We conduct the stepwise regression using both in the next case.**

Model <- medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + black + lstat

**Table 4: Stepwise Regression using Backward**

|  |  |
| --- | --- |
| **Model** | **AIC** |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim + rad + tax + indus + age | 1156.13 |
| medv ~ crim + zn + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat | 1154.35 |
| medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + black + lstat | 1152.74 |

**The below table captures the results of the stepwise regression using the both option:**

**Table 5: Stepwise Regression using Both**

|  |  |
| --- | --- |
| **Model** | **AIC** |
| medv ~ lstat | 1851.009 |
| medv ~ lstat + rm | 1735.58 |
| medv ~ lstat + rm + ptratio | 1678.13 |
| medv ~ lstat + rm + ptratio + dis | 1661.39 |
| medv ~ lstat + rm + ptratio + dis + nox | 1633.47 |
| medv ~ lstat + rm + ptratio + dis + nox + chas | 1621.97 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black | 1612.47 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn | 1606.31 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim | 1604.19 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim + rad | 1596.1 |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim + rad + tax | 1585.76 |

**Using the stepwise regression with both option results in the same model with 11 covariates (excluding indus and age) as obtained in the previous steps:**

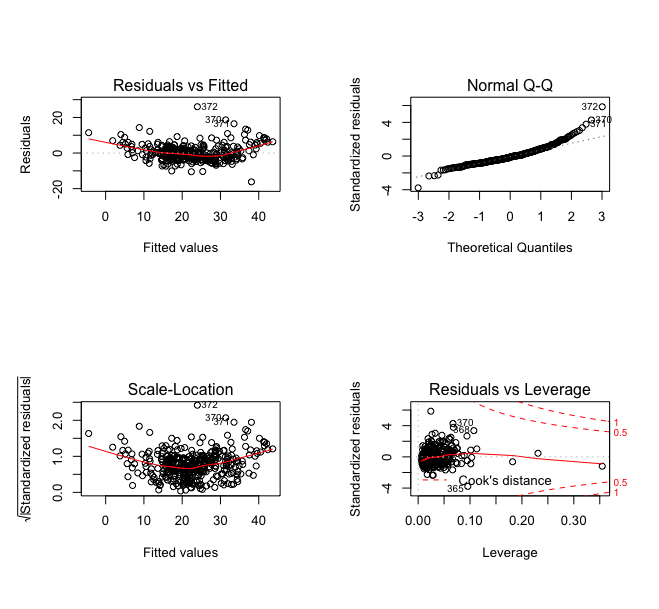
**Model <-** medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim + rad + tax

The summary statistics for the final selected model is given below:

**Table 6: Summary statistics for model**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model1** | **AIC** | **BIC** | **MSE** | **Ajusted R2** | **R2** |
| medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim+rad+tax | 2230.279 | 2281.467 | 19.65 | 0.7554 | 0.7625 |

**Figure 7: Residual plots for model1**

****

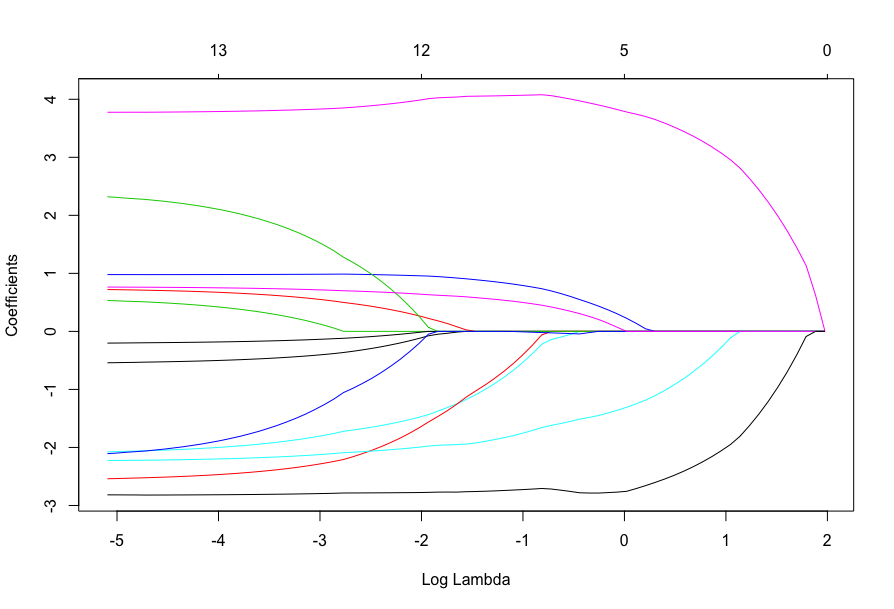
**The residual plots suggest similarity with the plots obtained using the linear regression, we would be performing LASSO regression on the model in order to check whether a better model can be obtained than the current one.**

## LASSO Regression

We will try the LASSO regression on the dataset.The objective is to minimize that coefficients of non-significant covariates.  
Here lambda is the penalty factor which helps in variable selection and so higher the lambda, lesser will be the significant variables included in the model.

We fit the LASSO model to our data. From the plot below, we see that as the value of lambda keeps on increasing, the coefficients for the variables tend to 0.

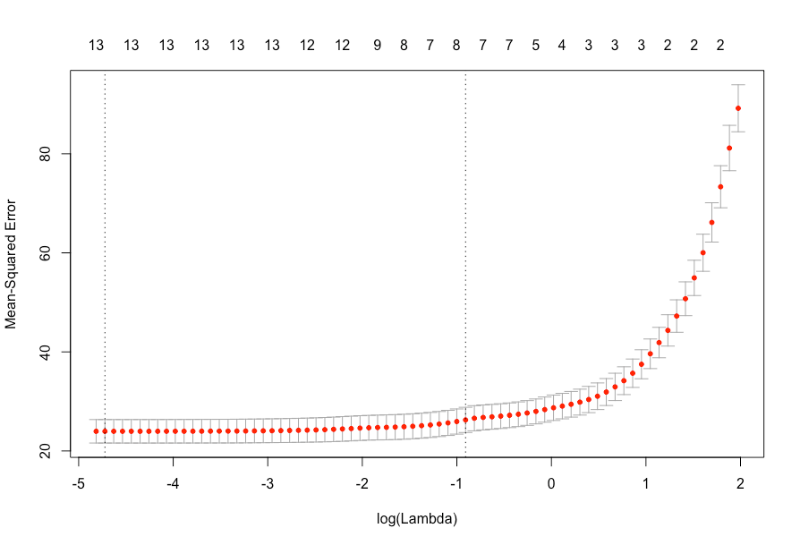
**Figure 8: Coefficients Vs Log Lambda**



Using cross-validation we now find the appropriate lambda value using error versus lambda plot.  
We plot the MSE vs Log(Lambda) below. For the higher error value , the number of variables selected decreases.

For model with lambda=min, coefficients of age and Indus get reduced to zero. For model with lambda=1se, coefficients of indus, age, rad and tax get reduced to zero

**Figure 9:MSE vs Log(Lambda)**



**Table 7: LASSO Comparison**

|  |  |  |
| --- | --- | --- |
|  | **LAMBDA.MIN** | **LAMBDA.1SE** |
| Value | 0.008 | 0.403 |
| MSE | 21.04 | 23.611 |
| MSPE | 26.48 | 31.32 |
| R-Squared | 0.758 | 0.726 |
| Adj Rsquared | 0.75 | 0.71 |

# Conclusion:

**We can see that the LAMBDA.MIN has a lower MSE and MSPE than the LAMBDA.1SE but it is higher than the model obtained through the stepwise selection.**

**So, we select the final model as per the stepwise regression as below:**

**Model <-** medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn + crim + rad + tax

Executive Summary for Q2 Simulation Study (Linear Regression)

# Problem & Approach

The main purpose of this simulation study was to understand the impact of sample size U variance of error term. The true linear equation was set as **Y=4+0.9∗x1+3∗x2+ ε** and **x1~N(2,**  ); **x2~N(-1,**  ); **ε ~N(0,**  ), where σ=1.

Using rnorm() function, x1 and x2 variables were created. The multiplication of those two variables was x3. Error term was also created using rnorm() with zero mean and standard deviation of 1.Finally, a data set including y, x1, x2, x3 variables was created (200 rows and 4 variables. A best subset was selected using stepwise regression.

As it is shown in Table 1., under condition of sample size n=200 and standard deviation of error term, ε, was set to 1, explanatory variables x1 and x2 were selected. After running linear regression, the fitted model was Y=4.72+0.77\*x1+3.52\*x2. Model MSE was 0.23; was 0.1985; and was 0.1903. The distance between model coefficients and true coefficients was 0.895 ().

**Table1.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| σ | n | Stepwise Model Selection | Model coefficients | Distance from true coefficients | MSE |  | adjusted |
| 1 | 200 | x1+ x2 | (4.7182, 0.7662, 3.5175) | 0.895 | 0.23 | 0.1985 | 0.1903 |

Four different sample sizes (25, 100, 500, and 5000) and 3 different standard deviations (0.1, 0.5, and 1) were introduced to simulate. Providing MSE (, , and adjusted , model performance of each scenario was displayed in Table 2.

# Major Results

As noise level increases, MSE and Distance between model coefficients and true coefficients also increase, but and adjusted severely decrease. In terms of model selection, incorrect variables were selected with small sample size cases (n=25) along with relatively larger noise levels (σ=0.5 and σ=1). Obviously, the most ideal case was smallest variance of noise and largest sample size case. When σ=0.1 and n=5000, the case indicated smallest distance and MSE, and largest and adjusted . Contrarily, When σ=1 and n=25, the case indicated largest distance and MSE, and smallest and adjusted

Within a same value of σ scenario, distance between model coefficient and true coefficient decreases by increasing size of n. For example, cases with σ=1 as n changes from 25, 100, 200 (from table 1), 500, to 5000, the distance decreases from 5.1276, 2.3867, 0.895 (from table 1), 0.3736, to 0.1667 respectively. Within a same value of σ scenario, largest sample size always linked with smallest MSE, but not always linked with largest and adjusted

**Table2.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| σ | n | Stepwise Model Selection | Model coefficients | Distance from true coefficients | MSE |  | adjusted |
| 0.1 | 25 | x1+x2 | (4.33726, 0.84918, 3.23066) | 0.4117 | 0.0142 | 0.9531 | 0.9488 |
| 100 | x1+x2 | (4.1666, 0.9032, 3.1709) | 0.2387 | 0.0109 | 0.9677 | 0.967 |
| 500 | x1+x2 | (4.010666, 0.879129, 2.970910) | 0.0374 | 0.0113 | 0.9592 | 0.959 |
| 5000 | x1+x2 | (4.007824, 0.903862, 3.014205) | 0.0167 | 0.0104 | 0.9651 | 0.9651 |
| 0.5 | 25 | X2+x3 | (6.7963, 0, 5.2286,  -0.6280) | 3.7404 | 0.3530 | 0.426 | 0.3738 |
| 100 | x1+x2 | (4.8330, 0.9159, 3.8543) | 1.1933 | 0.2731 | 0.5824 | 0.5738 |
| 500 | x1+x2 | (4.05333, 0.79564, 2.85455) | 0.1868 | 0.2814 | 0.4449 | 0.4427 |
| 5000 | x1+x2 | (4.03912, 0.91931, 3.07103) | 0.0834 | 0.2599 | 0.5338 | 0.5336 |
| 1 | 25 | X2 | (8.343, 0,  5.573) | 5.1276 | 1.3948 | 0.1351 | 0.09754 |
| 100 | x1+x2 | (5.6660, 0.9317, 4.7087) | 2.3867 | 1.092 | 0.2969 | 0.2824 |
| 500 | x1+x2 | (4.10666, 0.69129, 2.7091) | 0.3736 | 1.1257 | 0.1391 | 0.1356 |
| 5000 | x1+x2 | (4.07824, 0.93862, 3.14205) | 0.1667 | 1.0404 | 0.23 | 0.2297 |

Executive Summary for Monte Carlo Simulation Study**:**

The goal of Monte Carlo Simulation Study was to simulate the mean linear function E(y|x) = 4 + 0.9\*x1 + 3\*x2 and determine if the estimates of the function has bias. The study was conducted with a sample size of 200 and 100 simulation iterations. The true values of the parameters were compared to the simulated mean values to determine the bias.

Reference the table below for details.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **coef1** | **coef2** | **coef3** | **MSE** |
| **True Values** | 4.000000 | 0.900000 | 3.000000 | 0.0000000 |
| **Mean from Simulation** | 4.020429 | 0.881577 | 2.998727 | 0.9307307 |
| **Variance from Simulation** | 0.50578928 | 0.01834017 | 0.42140792 | 0.2309000 |

The true value and the simulated values in the table above are comparable. Their parameter values for coef1, coef2 and coef3 are very close. This indicates that there is no bias. Also the value of the Mean from the Simulated model is very close to 1. This is what is expected given that sigma squared is 1.

# Primary Setup

The predictor variable x1 and x2, and error were generated using the normal distribution random function rnorm(). The mean of 2 and the standard deviation of 0.5 was used for the predictor variable x1. The mean of -1 and the standard deviation of 1 was used for the predictor variable x2. The mean of 0 and the standard deviation of 1 was used for the error. This study was conducted with a sample size of 200 and 100 simulation iterations. Coefficients were initialized as coef1 = 4, coef2 = 0.9 and coef3 = 3.

# Monte Carlo Simulation Study

The Monte Carlo Simulation linear model study was generated with a sample size of 200 and 100 simulation iterations. The distribution normal random predictor variables x1, x2 and the error were used to create a dataset for the parameters. The plotted linear models are below.

**Figure 10: Residual Diagnosis Plot**

