**BANA 7046**

**Homework 2**

Presented By:

Anjali Shalimar, Digvijay Kawale, Meenal Narsinghani

**Index:**

**Problem 1**

Executive Summary

1. Introduction
2. Exploratory Data Analysis
3. Correlation
4. Variable Selection
5. Residual Diagnosis

**Problem 2**

1. Problem Statement
2. Results and Inferences
3. Conclusion

**Problem 3**

1. Problem Statement
2. Results and Inferences
3. Conclusion

**Executive Summary:**

The Boston Housing data set contains the information collected by the U.S. Census Service concerning the housing in the area of Boston. It has been extensively used throughout the literature to benchmark the algorithms. The aim of this study was to build a best linear regression model (without variable transformation) using the variables in the data set to predict the housing prices.

The data set contains 14 variables, one of them is median prices of houses which was the response variable and rest of the variables are used as predictor variables. Exploratory data analysis showed that there were a lot of outliers in many variables and most of them had non-normal distributions.

To get the best model variable selection analysis was done using three methods – Best Subset method of variable selection, Stepwise method of variable selection and LASSO method of variable selection. Using the BIC and adjusted R square criteria Best subset method suggested to use 11 type of variables as predictors. Same results were observed using AIC criteria in the Stepwise method of variable selection. The LASSO method of variable selection suggested to use only 6 variables as predictors using the optimum lambda (Tuning Parameter) criterion. The model suggested by LASSO method was selected as the best model. Residual diagnosis was done for the selected model which showed that they were normally distributed.

1. **Introduction**\*

The Boston housing data set contains information collected by the U.S. Census Service concerning housing in the area of Boston. It has been used extensively throughout the literature to benchmark algorithms. The data set consists for 506 observations with 14 attributes. Below is the list of all the attributes:

|  |  |  |
| --- | --- | --- |
| **Variable Names** | **Description** | **Type** |
| crim | Per capita crime rate by town | Numeric |
| zn | Proportion of residential land zoned for lots over 25,000 sq. ft | Numeric |
| indus | Proportion of non-retail business acres per town | Numeric |
| chas | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) | Binary |
| nox | Nitrogen oxides concentration (parts per 10 million) | Numeric |
| Rm | Average number of rooms per dwelling | Numeric |
| age | Proportion of owner-occupied units built prior to 1940 | Numeric |
| dis | Weighted mean of distances to five Boston employment centres | Numeric |
| rad | Index of accessibility to radial highways | Integer |
| tax | Full-value property-tax rate per \$10,000 | Numeric |
| ptratio | Pupil-teacher ratio by town | Numeric |
| black | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town | Numeric |
| Lsat | Lower status of the population (percent) | Numeric |
| medv | Median value of owner-occupied homes in \$1000s | Numeric |

Using the random 70% of the Boston data set as our training data, we have done the Exploratory Data Analysis. Our final aim is to predict the Median value of homes using Linear Regression model with this training data. The linear regression is conducted in the data without any variable transformation. We have done the variable selection using methods like best subset, stepwise, and LASSO.

1. **Exploratory Data Analysis**

**Crim**: We observe that the mean of the ‘crim’ is greater than its median with a large difference. It is observed form Fig 2.2 that there are a lot of outliers in the quantitative variable crim. This is causing its distribution to be extremely positively skewed.

\*Source: <https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| crim | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0.00632 | 0.08232 | 0.23912 | 3.27295 | 3.68567 | 67.92080 |

A screenshot of a cell phone

Description automatically generated

**Zn:** We observe that the mean of the ‘Zn’ is greater than its median with a large difference. The median itself has a zero value which tells us that 50% of the values are 0. It is observed in Fig 2.4 that there are lot of outliers in the variable ‘Zn’. Fig 2.3 also shows that it has positively skewed distribution.

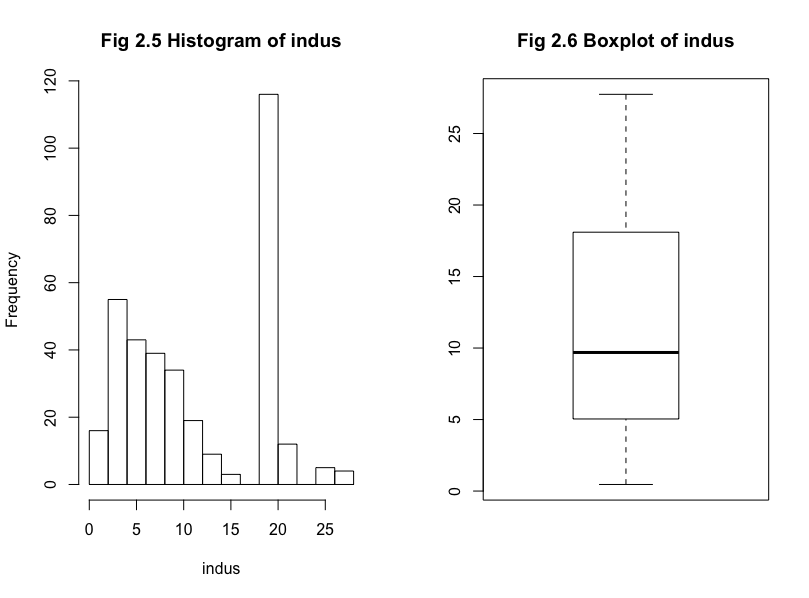
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Zn | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0 | 0 | 0 | 11.52 | 19 | 95 |

A screenshot of a cell phone

Description automatically generated

**Indus:** We observe that the mean of the ‘indus’ is greater than its median value. We find no outliers in the variable ‘indus’ which is observed in Fig 2.6.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| indus | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0.46 | 5.16 | 9.69 | 11.07 | 18.1 | 27.74 |



**Nox:** We observe that the mean of the ‘Nox’ is slightly greater than its media. We find no outliers in the variable ‘indus’ which is observed in Fig 2.8.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| nox | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0.385 | 0.448 | 0.538 | 0.5547 | 0.631 | 0.871 |

A screenshot of a social media post

Description automatically generated

**Rm:** We observe that the mean of the ‘rm’ is almost equal to its median. We observe from figures 2.9 and 2.10 that the variable is normally distributed but there are lot of outliers both the sides.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| rm | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 3.561 | 5.875 | 6.212 | 6.292 | 6.633 | 8.78 |

A screenshot of a cell phone

Description automatically generated

**Age:** We observe that the mean of the ‘Age’ is less than its median. Hence, it might have negatively skewed distribution. It is also evident from the Fig 2.11 that it has a negatively skewed distribution. From Fig 2.12 we do not see any outliers in this variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| age | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 6 | 44.05 | 77.7 | 68.23 | 93.6 | 100 |

A screenshot of a cell phone

Description automatically generated

**Dis:** We observe that the mean of the ‘dis’ is greater than its median with a large difference. Hence, it might have positively skewed distribution. We observe in Fig 2.14 that there are outliers in the data that are causing its distribution to be positively skewed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dis | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 1.130 | 2.088 | 3.112 | 3.826 | 5.213 | 12.127 |

A screenshot of a cell phone

Description automatically generated

**Rad:** We observe that the mean of the ‘rad’ is greater than its median with a large difference. Hence, it might have positively skewed distribution.From Fig 2.16there are no outliers observed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| rad | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 1 | 4 | 5 | 9.515 | 24 | 24 |

A screenshot of a cell phone

Description automatically generated

**Tax:** We observe that the mean of the ‘tax’ is greater than its median with a large difference. It is observed from Fig 2.18 that there are no outliers in the variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| tax | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 187 | 277 | 330 | 408.8 | 666 | 711 |

A screenshot of a cell phone

Description automatically generated

**Ptratio:** We observe that the mean of the ‘ptratio’ is less than its median. From Fig 2.20 we can observe that there are 2 outliers in the data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ptratio | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 12.60 | 17.40 | 19.1 | 18.49 | 20.2 | 22 |

A screenshot of a cell phone

Description automatically generated

**Black:** We observe that the mean of the ‘Black’ is greater than its median with a large difference. From Fig 2.22 we observe that there are lot of outliers in the variable observations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| black | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0.32 | 376.12 | 391.43 | 360.28 | 396.13 | 396.90 |

A screenshot of a cell phone

Description automatically generated

**Lsat:** We observe that the mean of the ‘Lsat’ is greater than its median. We can see from the Fig 2.23 that the distribution of variables is positively skewed and also from Fig 2.24 we can see that there are few outliers in the data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| lsat | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 1.920 | 6.715 | 11.25 | 12.555 | 17.105 | 36.980 |

A screenshot of a cell phone

Description automatically generated

**Medv**: We observe that the mean of the ‘medv’ is greater than its median. From Fig 2.25 it is observed that variable has a positively skewed distribution. Fig 2.26 tells us that there are lot of outliers in the variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| medv | | | | | |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 5 | 16.8 | 21.1 | 22.66 | 25.15 | 50 |

A screenshot of a cell phone

Description automatically generated

1. **Correlation Matrix**

There is good correlation observed between the pairs zn-dis, nox-indus, age-indus, age-nox, dis-indus, dis-nox, dis-age, rad-crim, tax-rad, tax-indus, tax-nox, lstat-indus, lstat-nox, lstat-rm, lstat-age, medv-rm, med-lstat. Table 2.1 shows the pairwise correlations of all the variables in our data set.

A close up of a piece of paper

Description automatically generated

**Table 2.1 Pairwise Correlation of all the variables.**

1. **Linear Regression**

We observed that most of the X variables had high values of their coefficients. The highest value of coefficient is for the variable nox. Lot of variables have lower values too. As per the values of the coefficients the most significant values are for nox, rm, dis, and lstat. We have already observed that the variables pair rad-tax and dis-age have very high correlation. So, to avoid multicollinearity they shouldn’t be together in the model.

The p-values also suggest that the most significant X variables are nox, rm, rad, tax, ptratio, dis and stat. The AIC and BIC values of the model are 2716 and 2778 respectively. The adjusted-r-square value of the model is 0.7354. After the observing the model summary we need to work on variables selection. We will be using the method like best subset, stepwise, and LASSO to find the best model to predict the prices of houses.

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| crim | -0.09 |
| zn | 0.04 |
| indus | -0.007 |
| chas | 1.22 |
| nox | -16.62 |
| rm | 3.84 |
| age | -0.003 |
| dis | -1.54 |
| rad | 0.29 |
| tax | -0.013 |
| ptratio | -0.98 |
| black | 0.008 |
| lsat | -0.493 |

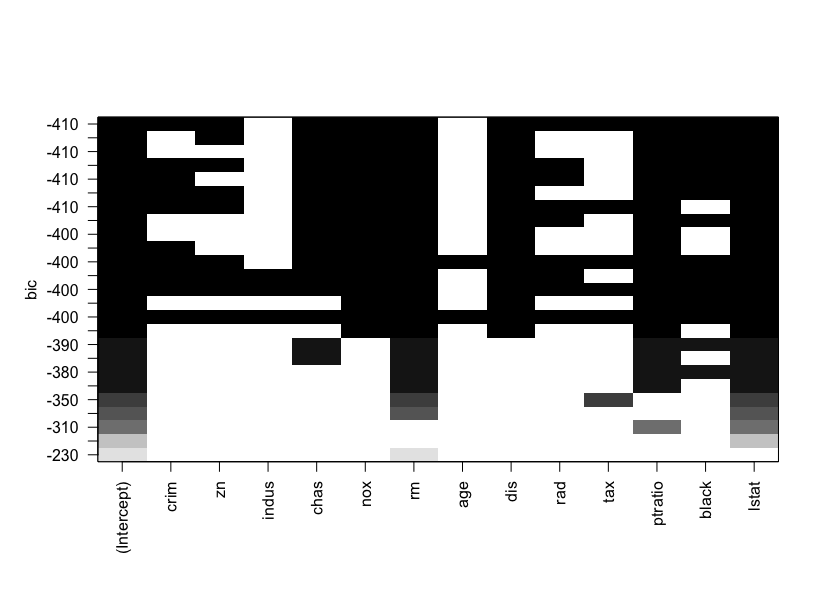
**Table 4.1 Variable and Coefficients for model with all variables**

1. **Variable Selection**:

**Best Subset Method for Variable Selection:** The best subsets regression is a model selection approach that consists of testing all possible combination of the predictor variables, and then selecting the best model according to some statistical criteria. After applying the best subset model for the given data set it gives the results that has number of variables ranging from 2 to 13. A total of 25 best subsets are given in the results by the best subset method.

We will compare the values such as adjusted r square values and BIC values to select the best model from these 25 models.

The minimum BIC value and maximum value of adjusted r square both are given by the model that has 11 variables which excludes indus and age.



**Fig 5.1 BIC values of Models using Best Subset Method**

**Stepwise Regression Method for Variable Selection:** The stepwise regression consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.

Using the stepwise regression method for model selection, the results suggest that the minimum AIC value is observed in the model which has 11 variables that exclude indus and age. The results are aligned with our Best Subset method of variable selection.

Table 5.1 shows the values of coefficients for all the selected variables.

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| Intercept | 33.95 |
| crim | -0.108 |
| zn | 0.048 |
| chas | 3.23 |
| nox | -17.43 |
| Rm | 4.08 |
| dis | -1.4 |
| rad | 0.28 |
| tax | -0.01 |
| ptratio | -0.99 |
| black | 0.009 |
| lsat | -0.46 |

**Table 5.1** **Variable and Coefficients for Selected Model**

**LASSO Method for Variable Selection:** In [statistics](https://en.wikipedia.org/wiki/Statistics) LASSO is a [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) method that performs both [variable selection](https://en.wikipedia.org/wiki/Variable_selection) and [regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics)) in order to enhance the prediction accuracy and interpretability of the [statistical model](https://en.wikipedia.org/wiki/Statistical_model) it produces. We will be using LASSO to select the variables to be used for prediction of housing prices.

By taking the minimum value of lambda as its optimal value the LASSO method of variable selection suggests taking 6 variables – crim, chas, rm, ptratio, black and lstat. Fig 5.2 shows the plot log lambda vs the coefficients. Fig 5.2 shows the plot of log lambda vs Mean Squared Error. The dotted lines represent the optimal value of the lambda.

A screenshot of a social media post

Description automatically generated

**Fig 5.2 Plot of Log Lambda vs Coefficients**

A screenshot of a social media post

Description automatically generated

**Fig 5.3 Plot of Log Lambda vs the Mean-squared Error of Models**

We go forward with choosing the model suggested by the LASSO method of variable selection. Below is the table of variable and coefficients.

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| Intercept | 13.35 |
| crim | -0.007 |
| chas | 0.39 |
| rm | 4.14 |
| Ptratio | -0.657 |
| Black | 0.002 |
| lstat | -0.451 |

**Table 5.2 Variables and Coefficients for the selected model**

1. **Residual Diagnosis:** The mean squared error of the model 26.4995. The values of R square and adjusted R square are 0.701 and 0.6958 respectively. Fig 6.1 shows that the residuals are randomly distributed. Fig 6.2 shows that the residuals are normally distributed.

**A close up of a mans face

Description automatically generated**

**Fig 6.1 Distribution of Residuals**

**A screenshot of a cell phone

Description automatically generated**

**Fig 6.2 QQ-plot of Residuals**

**2. A Simulation Study (Linear regression)**

**2.1 Problem and Approach**

**Different simulations** of a linear model –

y = 5 + 1.2\*x1 + 3\*x2 + ε ……… Eq. (I)

were generated by **varying** the **sample size** and the **variance of the error terms**.

Linear regression model was then fit, using **forward stepwise variable selection** approach, to this simulated data to identify the best line of fit.

Various metrics like **R2, adjusted R2 and MSE** were recorded for the different runs and inferences were drawn regarding in what cases linear regression performs a better job in estimating the line of fit.

**2.2 Results and Inferences**

The data was generated based on the linear model, given in eq.(I), by varying –

* sample size: 25,100, 200, 500, 5000
* std. dev. of error terms: 1, 0.5, 0.1

The below table summarizes the output of these different iterations –



From the above table we can observe –

* Increasing variance with sample size held constant:
* As we increase the variance, the model begins to perform poorly
* The coefficients deviate from their true values
* the model MSE also increases
* Increasing sample size for same variance:
* Coefficients converge closer to their original values in magnitude
* MSE begins to decrease
* Values of R2 and adjusted R2 also increase
* Overall the model performance increases
* Sample size and variance both are increased:

When both sample size and variance is increased, coefficients converge closer to their original values in magnitude although there was a slight tradeoff because of increase in variance

* MSE also increased substantially

**2.3 Conclusion**

From the above analysis we conclude that, a minimum number of data points are required for the linear model to fit the regression line, even if the data is generated from a linear model. As we observed, for sample size = 25, the model was not able to fit the regression line. Even we are able to contain he variance of the error terms, then the model may perform better even with low sample size.

Increasing the sample size would cause the coefficients to converge to their true values, and increasing the variance would increase the model MSE.

**3. Monte Carlo Simulation Study**

**3.1 Problem and Approach**

The Monte Carlo simulation is a computational algorithm that can be used to estimate parameters using repeated random sampling method. By the law of large numbers, with the increase in sampling iterations the simulation results approach closer to the true value of the parameters.

The below problem uses the method of Monte Carlo simulation to estimate the regression coefficients of the linear model whose **true mean response** is formulated as below:

*E(y|x) = 5 + 1.2X1 + 3X2*

Where,

*y* is the true response,

5 is the intercept of the true regression line ( ),

1.2 is the true coefficient of predictor *X1* (,

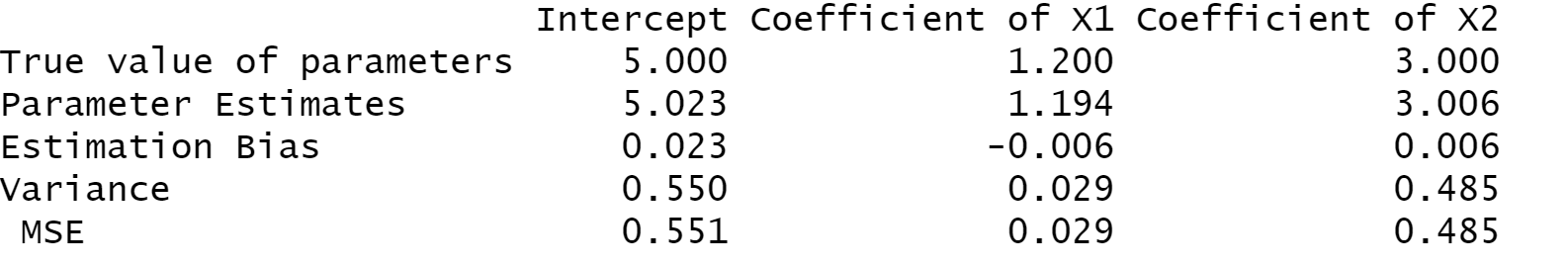
3 is the true coefficient of predictor *X2* )

In the model *X1 ~ N(2,0.42* ) and *X2 ~ N(-1 , 0.12)* and the sample size is set as 200.

The predictor terms are fixed, and the error term of the model is simulated 100 times. Each time a linear model is fit. The average of the model coefficients obtained from these 100 simulations are taken as the **estimated coefficient of the fitted linear model**. The bias, variance and MSE of these coefficients are listed. The accuracy of the estimated model coefficients (of the fitted regression line) is computed by comparing the same with the coefficients of the true regression line.

**3.2 Results and Inferences**

The simulation study gave the following results when seed is set as 13473173.



The average model MSE was observed to be **0.9841248**.

From the above results on MSE and bias, we can infer that simulation is a good technique to estimate the model parameters at a high iteration count (200 in this example). The MSE in the model is very close to 1 which is the error variance in the true regression line.

**3.3 Conclusion**

From the above study we can conclude that Monte Carlo simulation is a good approach to estimate the mean value of the response variable in a linear regression model. Incorporating higher iteration counts while simulating the results can help estimate model parameters. The parameters estimated using this approach is a good fit of the true regression coefficients and can hence be used to further reduce MSE and bias in predictions.