BOSTON HOUSING DATA ANALYSIS

Goyal, Prakhar

Table of Contents

1. Linear regression on Boston housing data

1.1 Problem and Approach

The data was originally published by Harrison, D. and Rubinfeld, D.L. `*Hedonic prices and the demand for clean air*', J. Environ. Economics & Management, vol.5, 81-102, 1978. This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. The report examines the various methods we can use to generate a linear regression model on a sample data set (70%) of the original data. Based on the training data set (30%), we then developed the best possible linear regression model.
In order to determine the best linear regression model, the team first performed exploratory data analysis, and reviewed the original paper for the Boston Housing data to gain some context and background understanding. The team noted any significant outliers, correlations, and patterns in the data. Based on the team's preliminary analysis, the team then tested a linear regression model. The team then further performed analysis and variable selection to determine our best linear regression model.

1.2 Major Results

Based on the analysis of the training data set, we have concluded that the best linear model for the Boston Housing data set includes all variables except "age", and "indus". This model was generated using the Backward elimination stepwise model. The team decided on this model after comparing the results from Full Model, LASSO, and Stepwise Regression (Forward, Backward and Both).

1.3 Exploratory Data Analysis

1.3.1 About Boston Dataset

```
> str(Boston_train)
'data.frame':   354 obs. of  14 variables:
 $ crim   : num  10.233 0.0907 11.1081 0.9762 0.3183 ...
 $ zn     : num  0 45 0 0 0 0 0 0 0 0 ...
 $ indus  : num  18.1 3.44 18.1 21.89 9.9 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.614 0.437 0.668 0.624 0.544 0.532 0.448 0.544 0.538 0.671 ...
 $ rm     : num  6.18 6.95 4.91 5.76 5.91 ...
 $ age    : num  96.7 21.5 100 98.4 83.2 74.9 6.6 87.3 88.8 96.2 ...
 $ dis    : num  2.17 6.48 1.17 2.35 4 ...
 $ rad    : int  24 5 24 4 4 24 3 4 4 24 ...
 $ tax    : num  666 398 666 437 304 666 233 304 307 666 ...
 $ ptratio: num  20.2 15.2 20.2 21.2 18.4 20.2 17.9 18.4 21 20.2 ...
 $ black  : num  380 378 397 263 391 ...
 $ lstat  : num  18 5.1 34.8 17.3 18.3 ...
 $ medv   : num  14.6 37 13.8 15.6 17.8 23.7 25.3 23.8 14.8 13.1 ...
```

The Dataset consists of 506 observations under 14 variables. We have however, split the dataset into a training dataset (354 observations) and test dataset.

We can observe that there aren't any missing values in the dataset, so imputation is not required.

1.3.2 Summary of Boston data

1.3.2.1 Pairwise Correlation

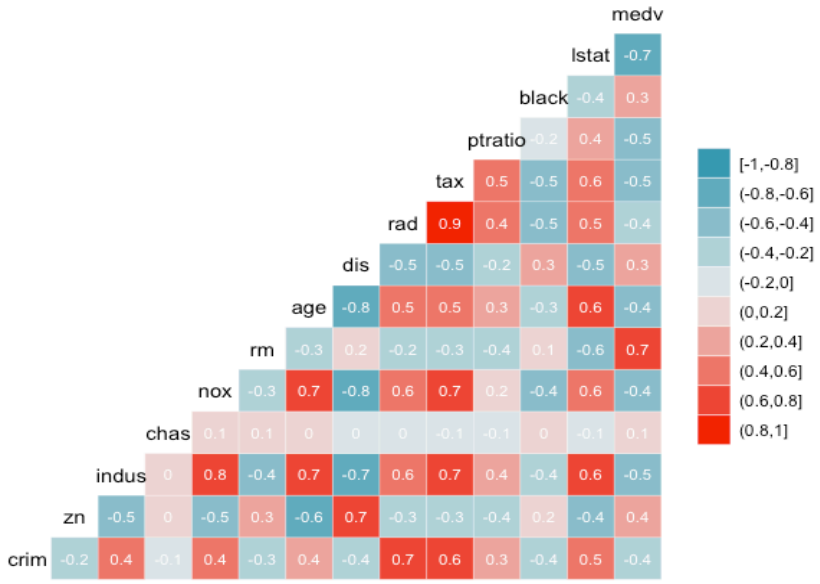


Figure 1:Pairwise Correlation Matrix

Key Insights:

- Status of People shows significant negative correlation with the prices, while "rm"(avg number of rooms) shows a positive correlation with the prices. Both the patterns are to be expected.

1.3.2.2 Outliers

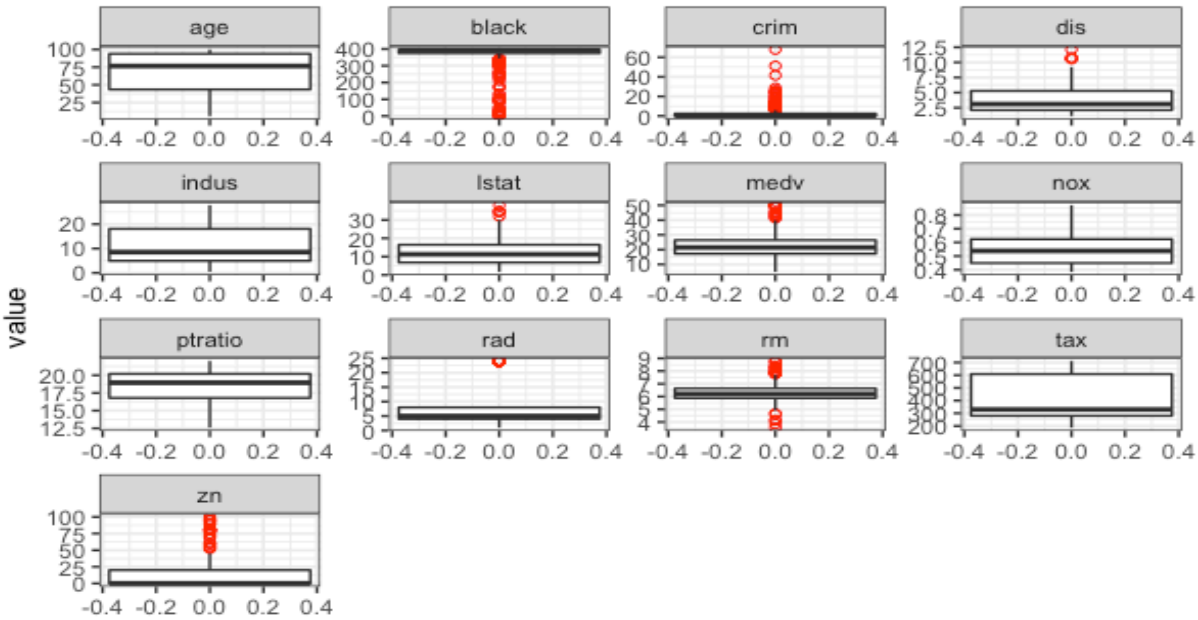


Figure 2:Boxplots

Key Insights:

- We can see that there is a lot of variations in the mean and median values of Crim,Zn and Indus hinting that these data are not following a normal distributions
- It is reasonable since there might be some dangerous and unsafe areas leading such outliers. In the same way, for the index of accessibility to radial highways (rad), no outlier seems to exist as the 3rdquartile (24) is equal to the maximum index (24) and the minimum one is 1 that closely to 0.
- The box plot of median value in $1000 presents a lot of outliers.There are 36 observations as the outliers accounted for 10.56% of 354 observations in the Boston training set. These outliers might be house constructed by special customs with the expensive value up to $50,000, similar to the outliers in the Status of People("lstat") column

1.4 LINEAR REGRESSION

Taking all variables and creating a model results in the following summary

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

Residuals:
     Min       1Q   Median       3Q      Max
-12.1118  -2.8043  -0.5757   2.0733  26.5902

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.538586   5.981134   6.443 4.00e-10 ***
crim         -0.116789   0.052379  -2.230 0.026420 *
zn            0.051668   0.016080   3.213 0.001438 **
indus         0.058024   0.073780   0.786 0.432157
chas         -0.096442   1.126683  -0.086 0.931836
nox         -18.379937   4.554231  -4.036 6.73e-05 ***
rm            3.712123   0.472045   7.864 4.99e-14 ***
age          -0.002857   0.016124  -0.177 0.859477
dis          -1.512728   0.225718  -6.702 8.56e-11 ***
rad           0.331069   0.078535   4.216 3.20e-05 ***
tax          -0.015916   0.004326  -3.680 0.000272 ***
ptratio      -0.964747   0.154066  -6.262 1.15e-09 ***
black         0.009542   0.003333   2.863 0.004454 **
lstat        -0.490190   0.060599  -8.089 1.07e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.768 on 340 degrees of freedom
Multiple R-squared:  0.7376,    Adjusted R-squared:  0.7276
F-statistic: 73.51 on 13 and 340 DF,  p-value: < 2.2e-16
```

Figure 3:Summary of Full Model Linear Regression

Key Insights:

- We can observe that columns "age", "indus", and "chas" have significant p-values and hence their null hypothesis cannot be rejected.
- Rsq is 73.76% and adj.Rsq is 72.76%
- P-value of the f-statistic is very low, hence we can reject a null model.
- The traditional linear regression uses ordinary least square for the estimation, but we know that OLS is very sensitive to outliers, and as there are some 10% outliers in the response variable, we can use robust linear regression as well.

1.5 VARIABLE SELECTION

1.5.1 Backward Elimination Using AIC

```
Step:  AIC=1114.23
medv ~ crim + zn + nox + rm + dis + rad + tax + ptratio + black +
    lstat

          Df Sum of Sq    RSS    AIC
<none>                  7744.1 1114.2
- crim     1    115.76 7859.9 1117.5
- black    1    181.07 7925.2 1120.4
- zn       1    239.65 7983.7 1123.0
- tax      1    321.29 8065.4 1126.6
- rad      1    401.13 8145.2 1130.1
- nox      1    402.35 8146.4 1130.2
- ptratio  1    884.40 8628.5 1150.5
- dis      1   1182.29 8926.4 1162.5
- rm       1   1427.67 9171.8 1172.1
- lstat    1   1643.79 9387.9 1180.4
```

Figure 4:Backward Selection using AIC

Key Insights:

- Best Subset claims the lowest BIC is associated with the model excluding “Indus”, “age”, and “zn”
- We observe that all the stepwise regression variable selection techniques have a very close AIV value, however the least AIC is in *Backward selection*
- Backward Selection model includes all variables except “Indus” and “age”.

Creating a model based on the Backward selection process reveals the following summary

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

Residuals:
     Min       1Q   Median       3Q      Max
-12.0480  -2.8015  -0.6089   2.1505  26.5592

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.170188   5.940277   6.426 4.41e-10 ***
crim         -0.117979   0.052251  -2.258 0.024581 *
zn            0.051052   0.015694   3.253 0.001256 **
chas         -0.021305   1.119409  -0.019 0.984827
nox         -17.538222   4.174497  -4.201 3.39e-05 ***
rm            3.662843   0.462227   7.924 3.26e-14 ***
dis          -1.536269   0.212692  -7.223 3.33e-12 ***
rad           0.314030   0.074869   4.194 3.49e-05 ***
tax          -0.014384   0.003835  -3.751 0.000207 ***
ptratio      -0.946776   0.151558  -6.247 1.24e-09 ***
black         0.009377   0.003316   2.827 0.004969 **
lstat        -0.489360   0.057436  -8.520 5.16e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.759 on 342 degrees of freedom
Multiple R-squared:  0.7371,    Adjusted R-squared:  0.7286
F-statistic: 87.17 on 11 and 342 DF,  p-value: < 2.2e-16
```

Figure 5:Model Summary using the Backward selection

	Parameters
Model type	Model.backward
Number of Predictor Variables	11
MSE	22.643554491225
R-Squared	0.737090115909129
Adjusted R-Squared	0.728633950046558
Test MSPE	23.9414974434185
AIC	2122.83621067337
BIC	2173.13707054411

Figure 6:Backward Selection Model Estimates

1.5.2.4 LASSO Variable Selection

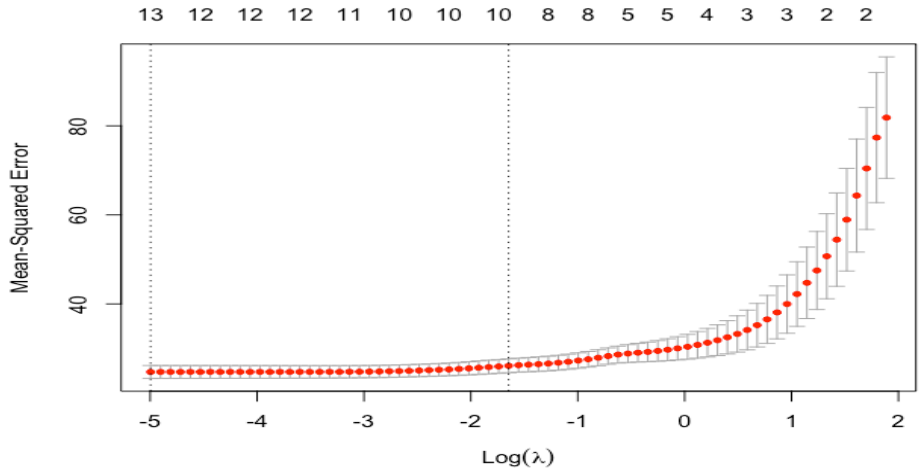


Figure 7:Cross Validation LASSO(5-fold)

Taking λ.1se as λ, coefficients of lasso.fit is

```
14 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept)  25.842502638
crim         -0.009603491
zn            0.024499441
indus         .
chas          .
nox         -10.133213420
rm            4.092144779
age           .
dis          -0.905492239
rad           0.014465611
tax          -0.003166377
ptratio      -0.834246507
black         0.007464374
lstat        -0.496643349
```

Figure 8:Coefficients of lasso.1se

Using the above coefficients, Lasso model leads to the below estimates

	Parameters
Model type	model.lasso.1se
Number of Predictor Variables	10
MSE	24.1837240806855
R-Squared	0.718386477250339
Adjusted R-Squared	0.710176170464634
Test MSPE	25.6266025776958

Figure 9:LASSO Model Estimates(1se)

We have 3 models(Full, Backward and LASSO) and below is the comparison of the estimates to select the best model

			Parameters
Model type	"model_Full"	"Model.backward"	"model.lasso.1se"
Number of Predictor Variables	"13"	"11"	"10"
MSE	"22.7337993364683"	"22.643554491225"	"24.1837240806855"
R-Squared	"0.737585912931911"	"0.737090115909129"	"0.718386477250339"
Adjusted R-Squared	"0.727552433132248"	"0.728633950046558"	"0.710176170464634"
Test MSPE	"24.0862798591897"	"23.9414974434185"	"25.6266025776958"

Figure 10:Estimate Comparison of 3 models

Key Insights:

- Based on the three tables, we can clearly see that Backward technique is producing best results in terms of Adj R-sq and Test Mean Square Predicted error. Hence, we go with the ***Model.Backward*** as our final model.
- An argument can be made to consider the LASSO generated model as the final model as there are least number of variables and there is not much increase in the Test prediction error and Adj R-sq.
- Our Final Model consists of all the variables as predictor variables except "*Age*" and "*Indus*"

1.6 RESIDUAL DIAGNOSIS

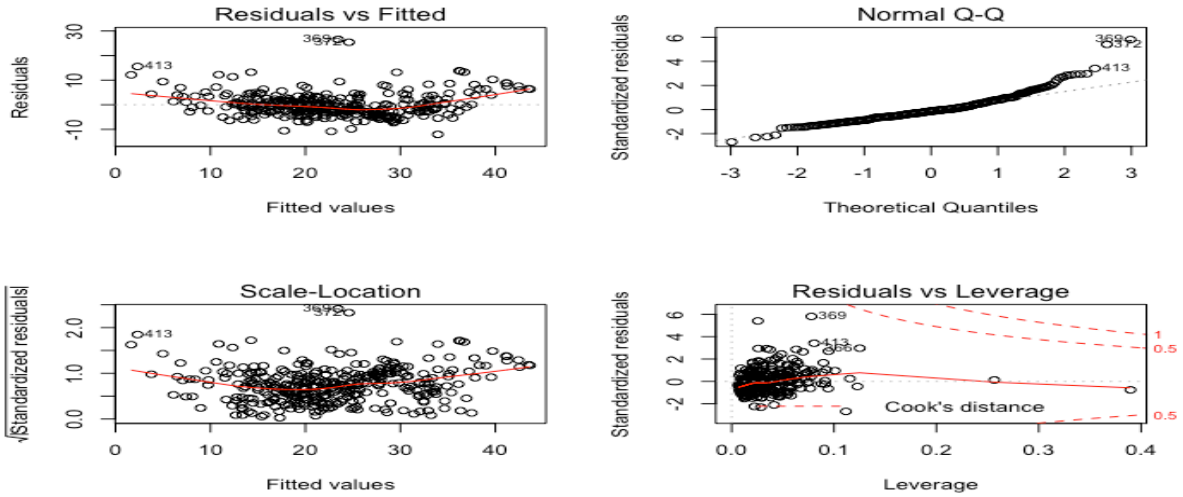


Figure 11:Residual Diagnosis for the Selected Final Model(Backward Selection)

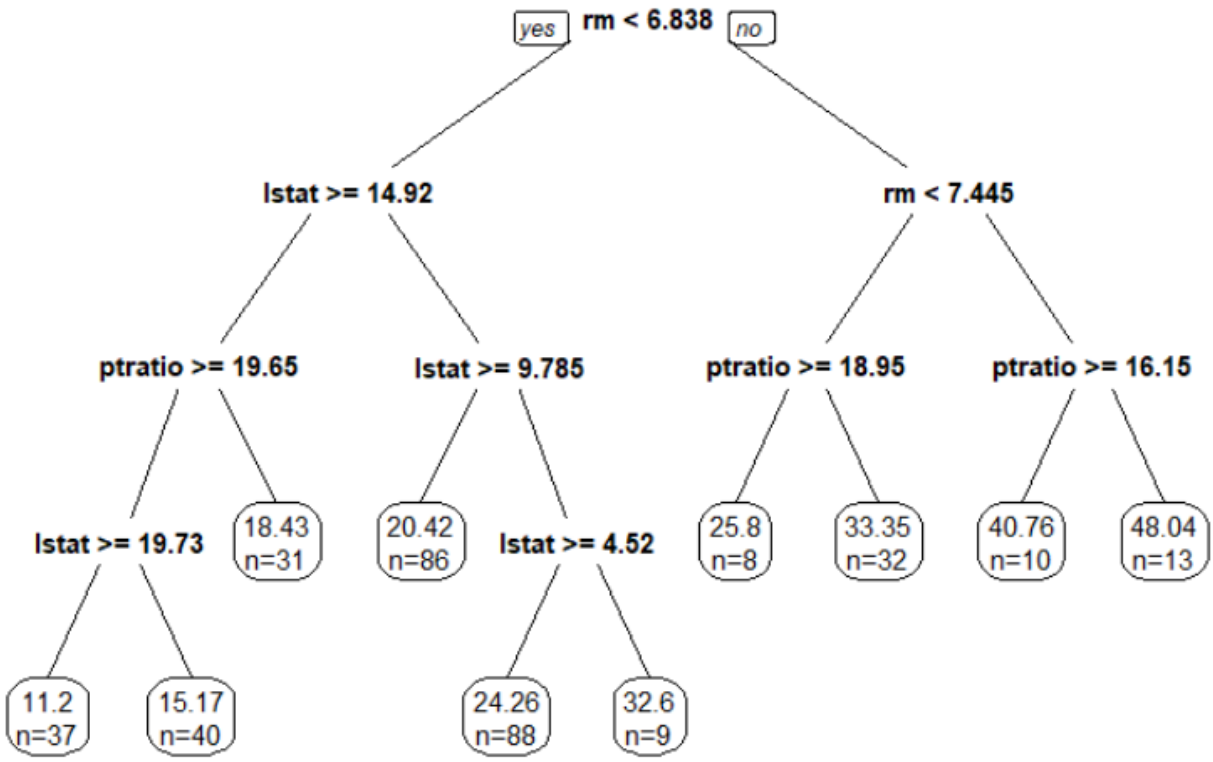
Key Insights:

- The variance is not completely constant and hence the assumption of constant variance is not totally satisfied
- From the q-q plot we see that it is not completely normal and a little skewed to the right.

1.7 CROSS VALIDATION

After doing cross validation using 3-fold, the average MSPE using CV comes out to be 24.42 vs 25.12 from step regression method. So, not a lot of difference there.

1.8 FITTING A DECISION TREE



MSPE is 24.55 - Almost same as that given by CV and step regression methods

1.9 COMPARISON OF TWO DATASETS

Model Type	MSPE - Set 1	MSPE - Set 2
Step (AIC)	25.12	22.77
CV	24.90	25.25
Regression Tree	24.55	23.63