Linear Regression on Boston Housing Price

Jingyuan Liang, Kevin Yu, Haoni Zhan, Ida He

11/23/2019

1. Introduction

The purpose of this project is to learn the characteristics of the “Boston“ dataset and seek the relationships between different variables and the median house price in the Boston area. These variables including per capita crime rate by town, the proportion of owner-occupied units built prior to 1940, index of accessibility to radial highways and etc. For example, it is common sense that people prefer to live in a safe area so that the area with a lower crime rate usually associated with higher house prices. We also know that people tend to live in the area that closer to their workplaces or with greater accessibility to transportation and business area. We will apply the data science techniques we learned from the STA141A to this project. We would apply the basic statistics knowledge and ggplot package we learned to explore and visualize the characteristics of our dataset. We are going to use the linear regression model to seek the relationship between different variables of the housing (crime rate, accessibility to highways, distance to business areas and etc.) and the house prices in the Boston area and to generate a best-fit linear regression model to predict the price of houses in the Boston area.

Key questions about the dataset:

1. What are the top five variables affecting the Boston housing price most?
2. What is the relationship between the top five variables and the Boston housing price?
3. Find the linear regression model which predicts the relationship best.

2. Data description

We will be using the built-in Boston housing pricing dataset in R. The dataset contains 507 data points and each data point has 14 measures. We will be considering measures from various tests that attempt to quantify the price of a house.

This data contains the following variables:

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

3. Explotary Data Analysis

After getting the data, the first step we do is to process the data cleaning and explortary data analysis to study the characteristics of the data. According to our results, there are no missing value. All the values are numeric. Therefore, reguar data cleaning is not required here.

The second step we process is to see what variables we have in our data set. From the results we can see that there are variable such as “age”,“black”,“chas”,“crim”,“dis”,“indus”,“lstat”,“medv”,“nox”,“ptratio”,“rad”,“rm”,“tax”,“zn”. All of them are numeric variables. Additionally, the varibale “chas” is a dummy variable that indicate whether the house is tract bounds river.

```
##      crim        zn     indus      chas       nox        rm       age       dis 
## "numeric" "numeric" "numeric" "integer" "numeric" "numeric" "numeric" "numeric" 
##       rad       tax   ptratio     black     lstat      medv 
## "integer" "numeric" "numeric" "numeric" "numeric" "numeric"
```

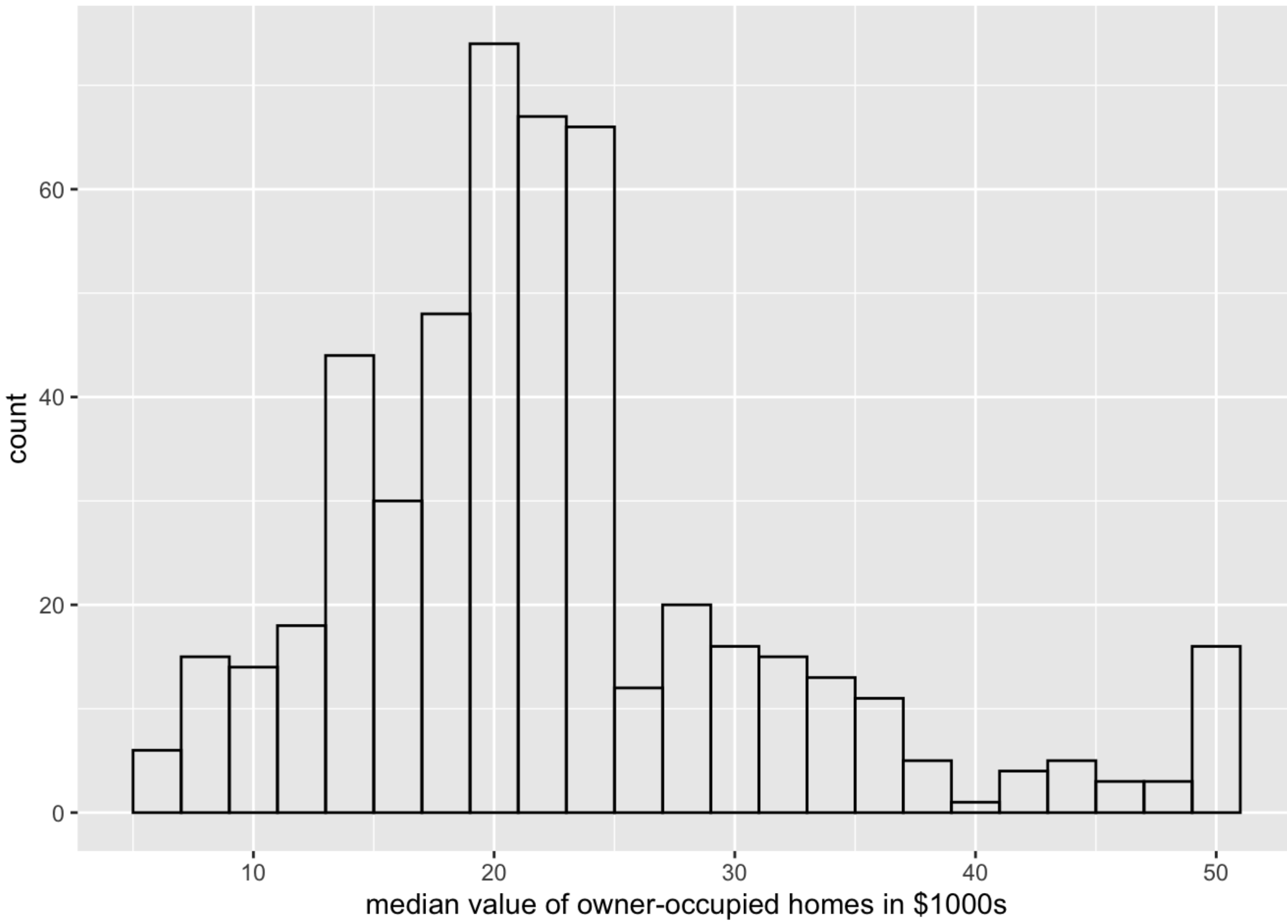
Variable Description

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

Then we are trying to learn the basic statistics of all the variables above. The most important variable to study here is the “medv”, the median value of owner-occupied home in $1000s. According to our result, the average medain value of homes in our data is 22.53, which is slightly higher than its median value 21.20. The cheapest home in our data set is 5000 dollars; the most expensive home is 50,000.

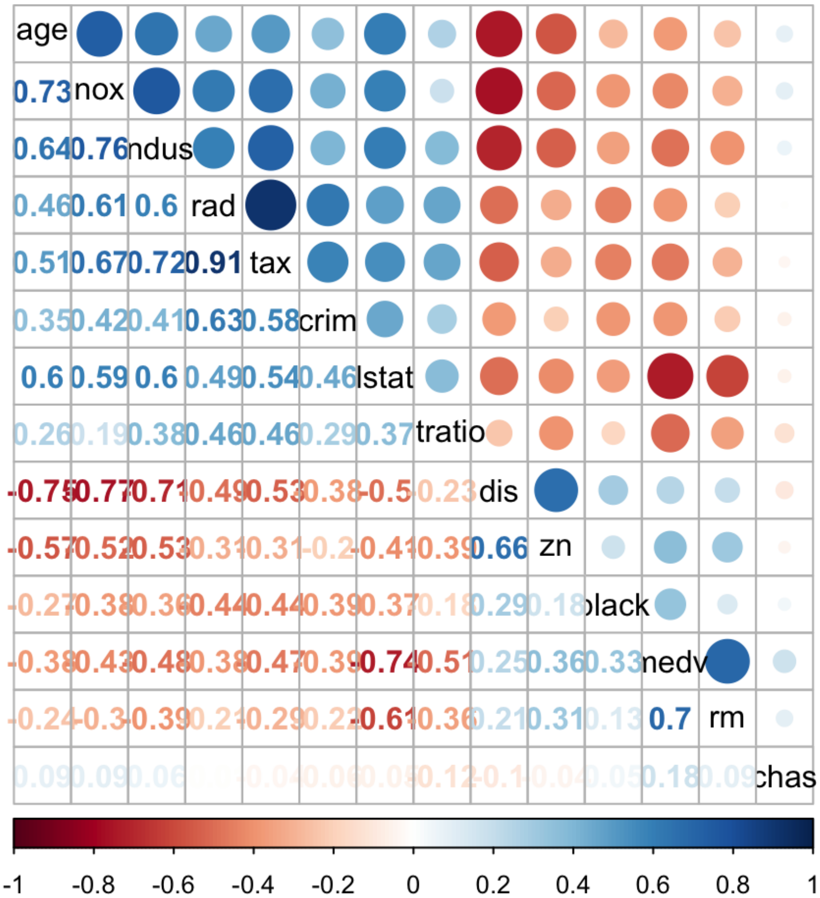
```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

To visualize our results for the median value of the home price, we graph a histrogram. From the result, we can see that the histrogram is a bell shape which is close to a normal distribution, but it is slightly skew to the right. Therefore, we apply the log transformation to the median value of owner-occupied homes in $1000s. It is also confirmed in the Box-Cox transformation in next Section.

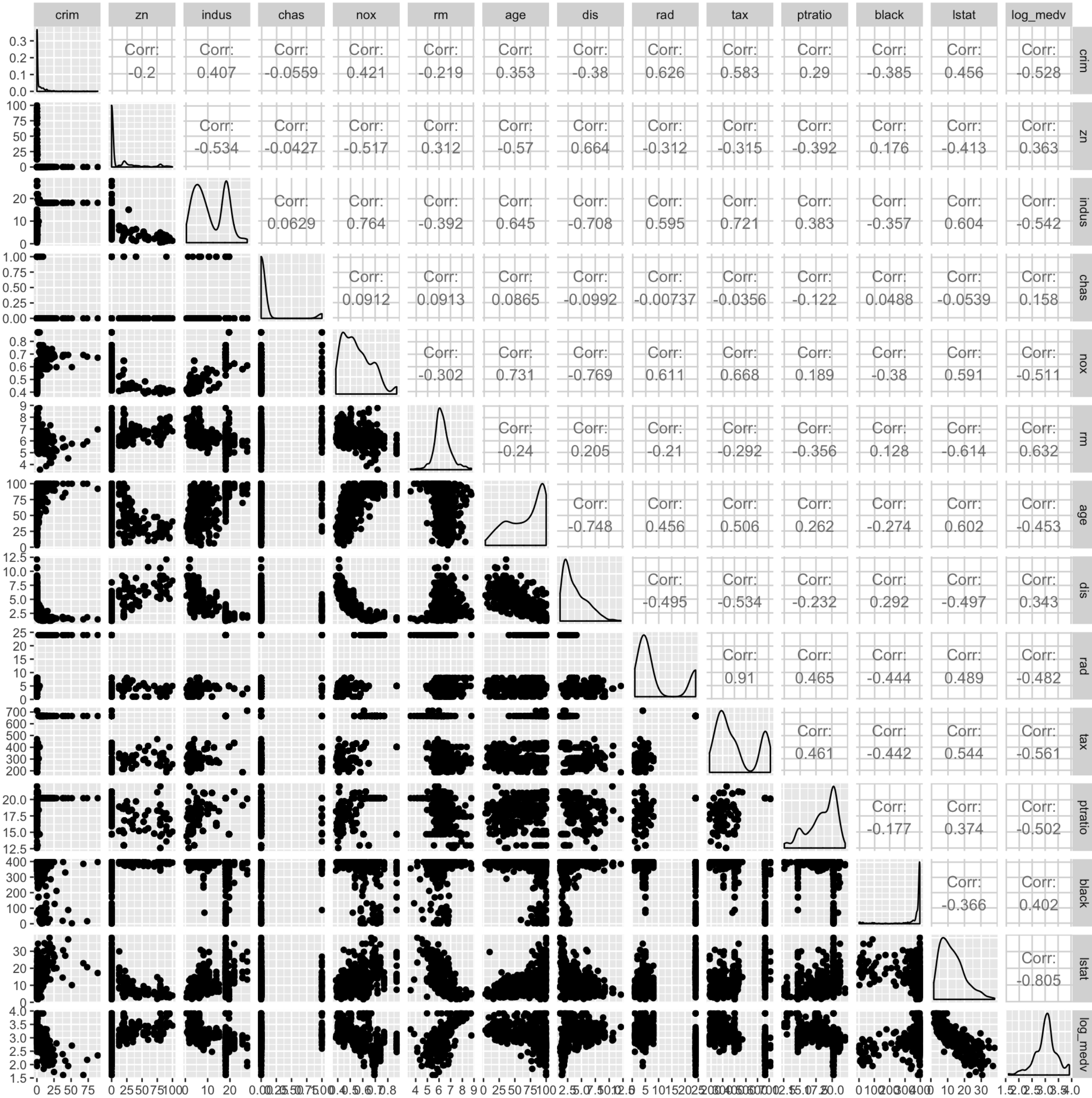


We continue our explortary data analysis to study if there are any correlation between each variables.The outcome variable “medv” is directly correlated with “rm” (number of rooms), “ptratio”(pupil-teacher ratio by town), and “lstat”(lower status of the population). These correlations themselves and the directions of these correlations makes perfect sense. As for the negative correlation between “medv” and “ptratio” might be due to the number of public schools is higher in the towns with low “medv”, and the educations in towns with high “medv” are better but fewer, according to the reality. These are the predictor variables that we need to concern with.

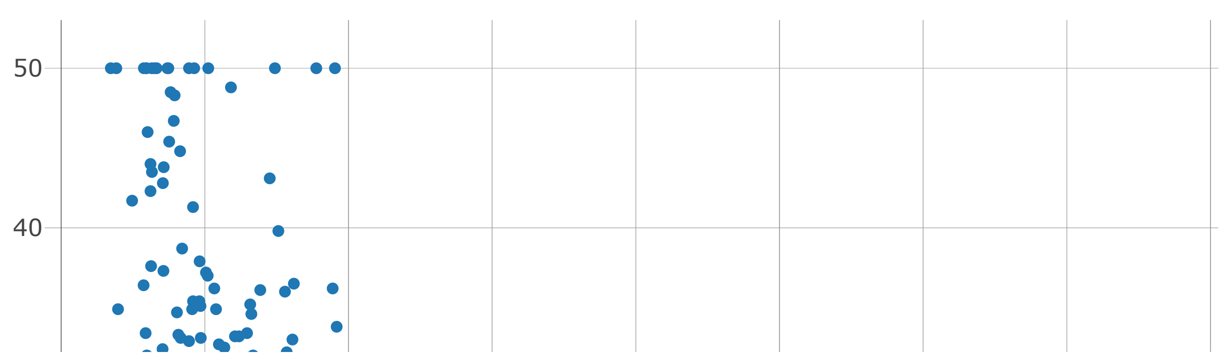
```
##          crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio
## crim     1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29
## zn      -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39
## indus    0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38
## chas    -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12
## nox      0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19
## rm      -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36
## age      0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26
## dis     -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23
## rad      0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46
## tax      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46
## ptratio  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00
## black   -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27  0.29 -0.44 -0.44   -0.18
## lstat    0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37
## medv    -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51
##         black lstat  medv
## crim    -0.39  0.46 -0.39
## zn       0.18 -0.41  0.36
## indus   -0.36  0.60 -0.48
## chas     0.05 -0.05  0.18
## nox     -0.38  0.59 -0.43
## rm       0.13 -0.61  0.70
## age     -0.27  0.60 -0.38
## dis      0.29 -0.50  0.25
## rad     -0.44  0.49 -0.38
## tax     -0.44  0.54 -0.47
## ptratio -0.18  0.37 -0.51
## black    1.00 -0.37  0.33
## lstat   -0.37  1.00 -0.74
## medv     0.33 -0.74  1.00
```

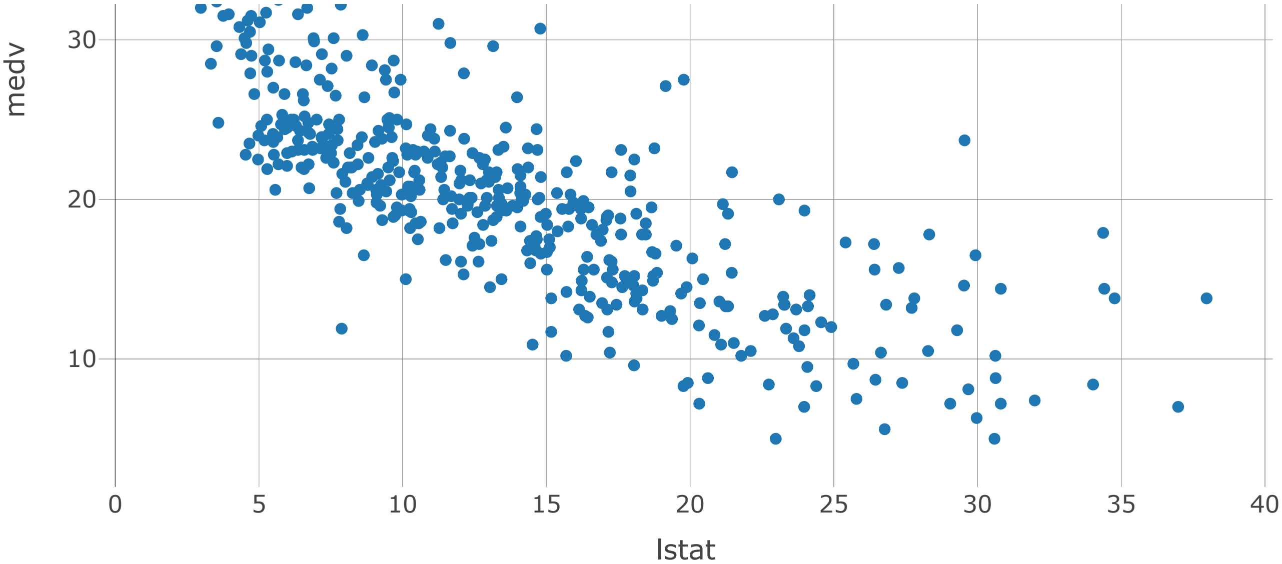


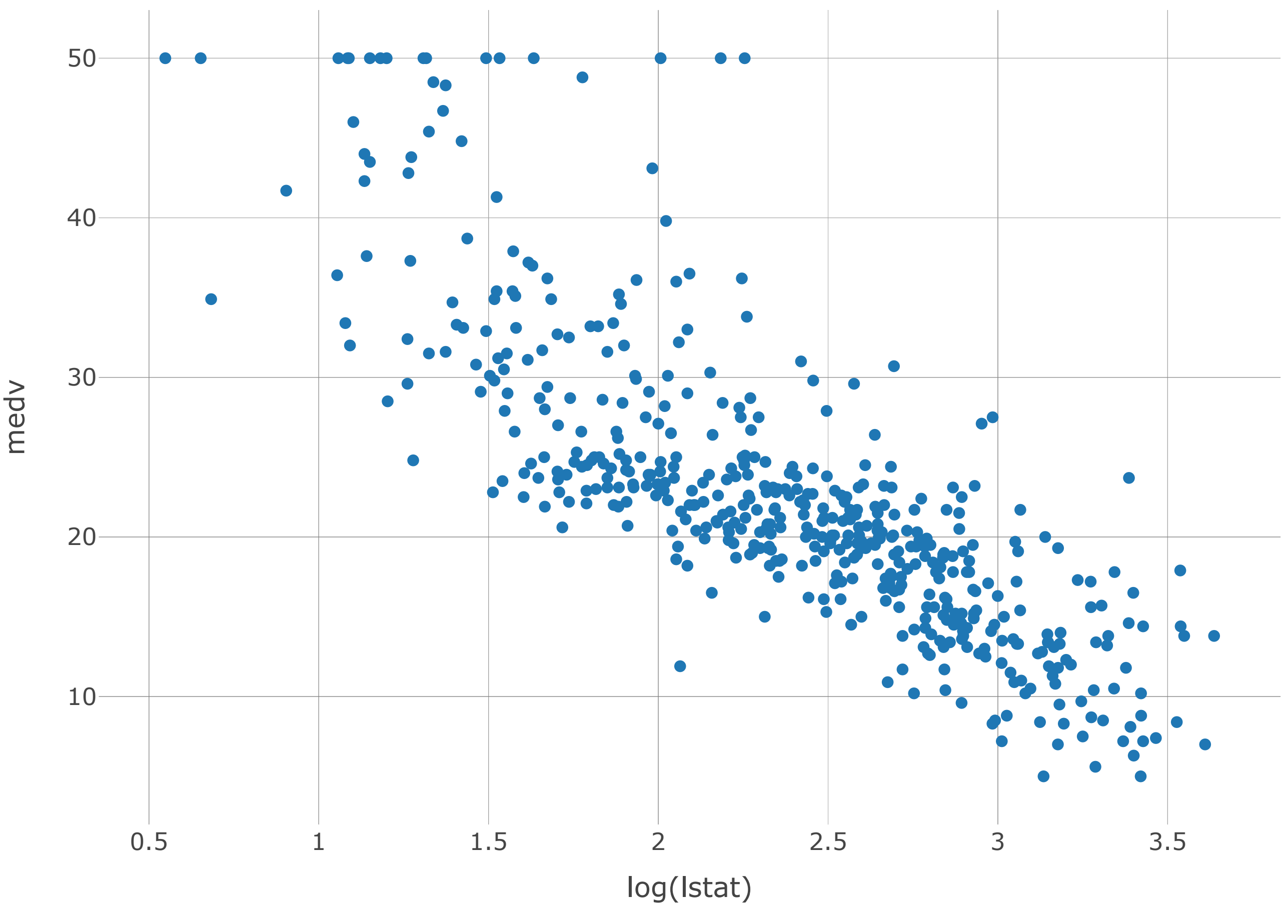
More clearly, we use scatter plot matrix to give an overview of relations among the predictors as well as between predictors and the response house price.



The scatter plot matrix shows that lots of the relations among the predictors and response are curved not linear indicating an appropriate transformation of the predictors is needed. The following two scatters could show that log-transformation could show a better linear relationship between predictors and the house price.







Also, the scatters among the predictors show that some of the predictors are highly correlated such as nox and indus, thus, it indicates model selection is needed to avoid multicollineary problems. As when we are using multiple linear regression, we should pay attention to variables with high correlations and consider dropping them to fit better multiple linear regression algorithms. Therefore, we use VIF function to check each variable's VIF. We would pay attention to variables with VIF is greater than 5, which corresponds to an R^2 of .80 with the other variables.

```
##     crim       zn    indus     chas      nox       rm      age      dis
## 1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945
##      rad      tax  ptratio    black    lstat
## 7.484496 9.008554 1.799084 1.348521 2.941491
```

We see variable “rad”, “tax” and “log(crim)” have really high VIF. This might be problematic since they might contribute to multicollinearity. We might need to drop these three variables first.

4. Fitting Model

Fitting Model

In this part, we are going to try to fit the dataset into different linear models. Here, we will take MEDV as the dependent variable and other remaining variables as independent variables.

##4.1 linear model 1

Linear model 1 here contains all the parameters in the Boston dataset. The coefficient and the significance for each parameter are found using the commend summary().

```
##                 coef
## (Intercept)  36.4595
## crim         -0.1080
## zn            0.0464
## indus         0.0206
## chas          2.6867
## nox         -17.7666
## rm            3.8099
## age           0.0007
## dis          -1.4756
## rad           0.3060
## tax          -0.0123
## ptratio      -0.9527
## black         0.0093
## lstat        -0.5248
```

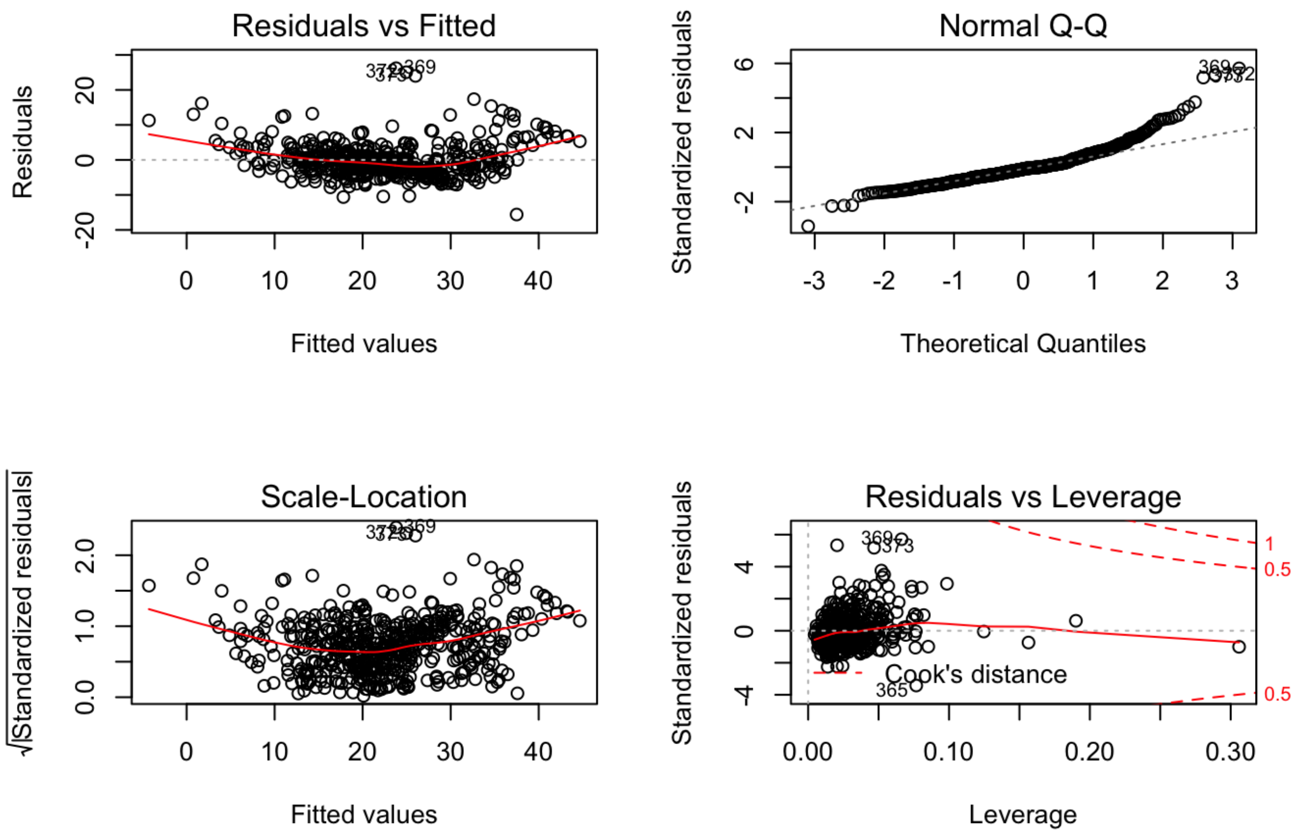
The coefficients of each variables in linear model 1 are shown above

```
##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

From above summary, we can see that the p-values of , “zn”, “nox”, “rm”, “dis”, “rad”, “ptratio”, “black”, “lstat” are small enough to be used to reject the null hypothesis of beta = 0. However, the p-value of the “indus”, and “age” are way larger then the regular alpha 0.05. The p-value of the variable “indus” is 0.7383, and the p-value of the variable “age” is 0.9582. Therefore, for variables “indus”and “age” we fail to reject the null hypothesis, and they are not statistically significant in this model.

Furthermore, for the variables “crim”, chas“, and”tax", their associated p-values are 0.001087, 0.001925, and 0.001112. Those variables would be considered as less significant variables here.

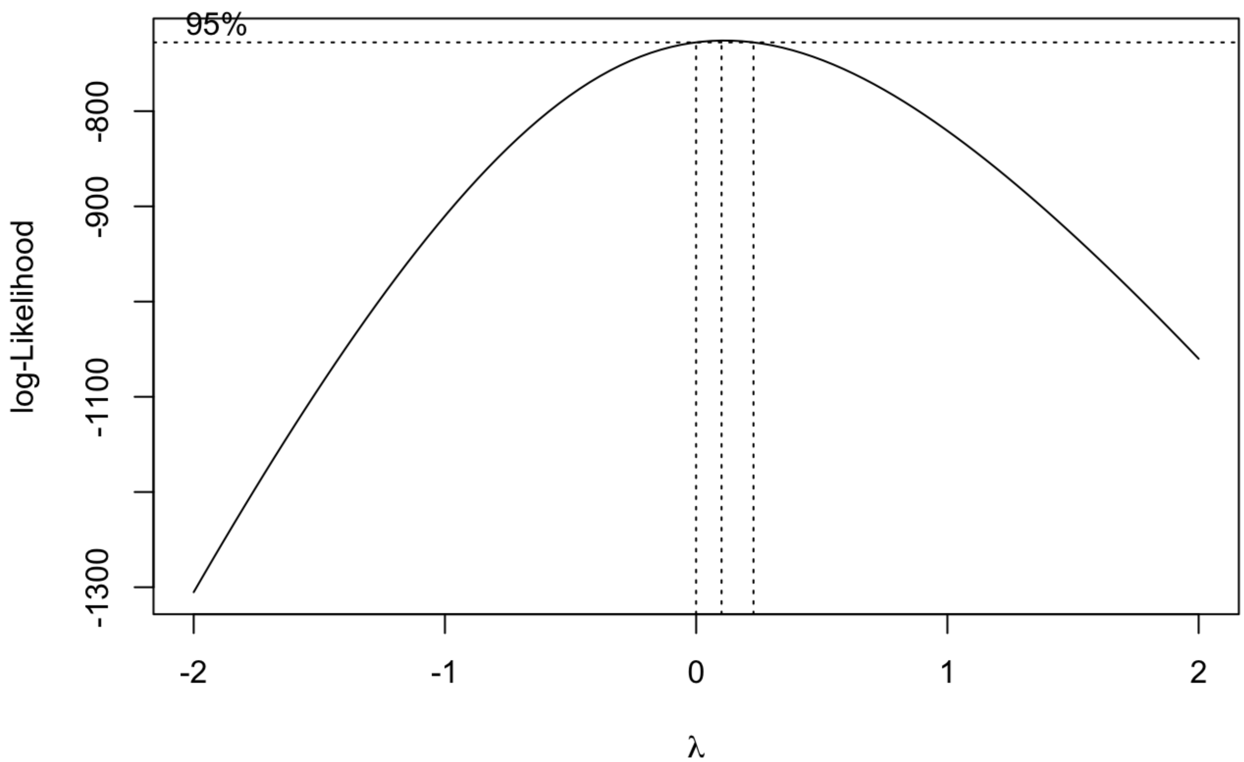
Also, according to the above results, the residual standard error here is 4.7450. The R-squared for this linear model is 0.7406 and the adjusted R-squared is 0.7338, which are both relatively high, indicating that there are approximately more than 70% of the observed variation can be explaind through the model’s inputs.



understanding the above diagnostic plots: First of all, we can see that the residual plot looks relatively u-shaped comparing to a straight line. This would indicate nonlinearity in the current linear model 1. From the qq plot, we can say that the data is approximately normally distributed, although there might be several possible outliers. In the scale-location plot, it is not spread equally along the the range of predictors. This might indicate that we should check the assumption of equal variance in this model. Lastly,the residuals vs. leverage plot shows that there’s no points higher than cook’s distance. So, there should be no influential points for this data.

##4.2 linear model 2: data transformation through log transformation

From model 1 we notice that the variable MEDV is not perfectly normally distributed and there is non-linear pattern, the spread of residuals also appear there is non-constant variance.



Also, the boxcox transformation indicates a best transformation includes 0 which means we would consider to use log transformation to transform the variable MEDV.

```
##                coef
## (Intercept)  4.1020
## crim        -0.0103
## zn           0.0012
## indus        0.0025
## chas         0.1009
## nox         -0.7784
## rm           0.0908
## age          0.0002
## dis         -0.0491
## rad          0.0143
## tax         -0.0006
## ptratio     -0.0383
## black        0.0004
## lstat       -0.0290
```

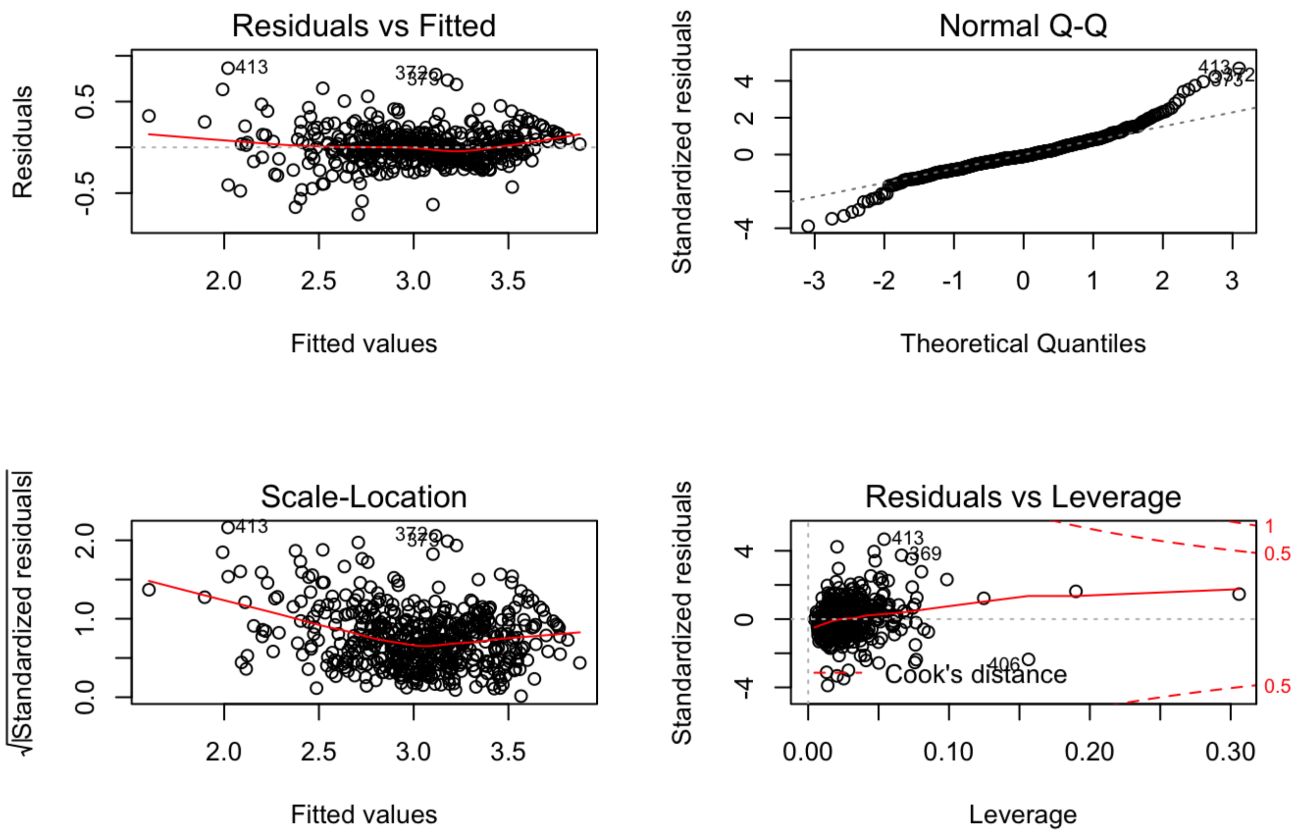
The coefficients of each variables in linear model 2 are shown above

```
##
## Call:
## lm(formula = log(medv) ~ crim + zn + indus + chas + nox + rm +
##     age + dis + rad + tax + ptratio + black + lstat, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73361 -0.09747 -0.01657  0.09629  0.86435
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.1020423  0.2042726  20.081  < 2e-16 ***
## crim        -0.0102715  0.0013155  -7.808 3.52e-14 ***
## zn           0.0011725  0.0005495   2.134 0.033349 *
## indus        0.0024668  0.0024614   1.002 0.316755
## chas         0.1008876  0.0344859   2.925 0.003598 **
## nox         -0.7783993  0.1528902  -5.091 5.07e-07 ***
## rm           0.0908331  0.0167280   5.430 8.87e-08 ***
## age          0.0002106  0.0005287   0.398 0.690567
## dis         -0.0490873  0.0079834  -6.149 1.62e-09 ***
## rad          0.0142673  0.0026556   5.373 1.20e-07 ***
## tax         -0.0006258  0.0001505  -4.157 3.80e-05 ***
## ptratio     -0.0382715  0.0052365  -7.309 1.10e-12 ***
## black        0.0004136  0.0001075   3.847 0.000135 ***
## lstat       -0.0290355  0.0020299 -14.304  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1899 on 492 degrees of freedom
## Multiple R-squared:  0.7896, Adjusted R-squared:  0.7841
## F-statistic: 142.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

From above summary, we can see that the p-values of “crim”, “nox”, “rm”, “dis”, “rad”, “tax”, “ptratio”, “black”, “lstat” are small enough to be used to reject the null hypothesis of beta = 0. However, the p-value of the “indus”, and “age” are way larger then the regular alpha 0.05. The p-value of the variable “indus” is 0.3168, and the p-value of the variable “age” is 0.6906. Therefore, for variables “indus”and “age” we fail to reject the null hypothesis, and they are not statistically significant in this model. What we can see here compared to model 1 is that the p-values for variable “indus” and “age” both decrease a little bit, although they are still large enough for supporting the null hypothesis.

Furthermore, for the variables “zn” and “chas”, their associated p-values are 0.0333 and 0.0036. Those variables would be considered as less significant variables here.

Also, according to the above results, the residual standard error here is 0.1899, which decreases significantly (from 4.4750 to 0.1899) comparing to linear model 1. The R-squared for this linear model is 0.7896 and the adjusted R-squared is 0.7841, which are both relatively high, indicating that there are approximately more than 70% of the observed variation can be explaind through the model’s inputs. Moreover, both R-squared and Adjusted R-squared increased by a small amount comparing to model 1.



understanding the above diagnostic plots: Lets compare the diagnostic plots here with the plots we got in model 1. we can see that the residual plot looks relatively less u-shaped now. The model 7 would have less possibilitiy of nonlinearity than model 1.

Also, from the qq plot, we can say that the data is more normally distributed, the data here are more fitted to the 45 degree line. In the scale-location plot, it is still not spread equally along the the range of predictors. This might indicate that we still need to check the assumption of equal variance in model 7. Also,the residuals vs. leverage plot shows that there are some influential points needed to be deal with.

To be noticing that we also tried to use a model that perform log transformation for “MEDV”(the y value), “crim”, and “lstat”. However, after performing this model, the R square decreased, and the diagnostic plots were not improved signicantly comparing to our model here. Therefore, we decide to continue using the model with transforming the “MEDV” only. For the further studying, we will use cross validation method to determine which method is better.

#5. Diagnostics: Ouliters, high leverage points and strong influential points

In this part, we decide to remove the ouliters, high leverage points and strong influential points in the data and then refit the model to check if a better model would be produced. The rules are:

1. For outliers, standardized residuals |r_i| > 2
2. For high leverage $h_{ii} > 2 \times (p+1)/n$
3. For strong influential points, cook’s distance $D_i > 4/(n-p-1)$.

The unusual points indice detected are as below:

```
##  [1]   8 142 149 215 254 365 366 368 369 370 371 372 373 374 375 398 399 400 401
## [20] 402 404 406 408 410 413 417 420 427 490 506 121 122 123 124 125 126 127 143
## [39] 146 153 155 156 157 163 164 284 354 355 356 381 411 415 419 428 489 491 492
## [58] 493  65 148 167
```

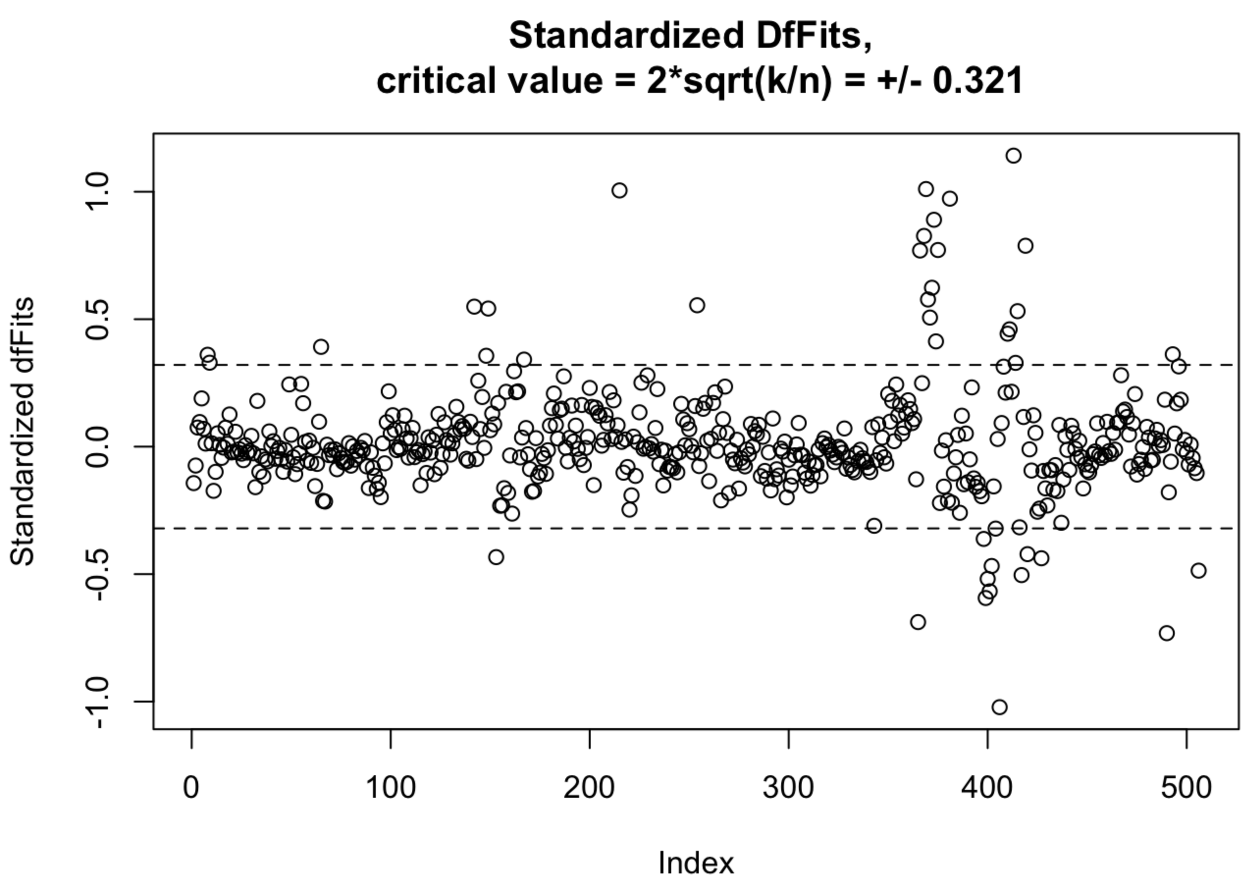
```
##         crim   zn indus chas    nox    rm   age     dis rad tax ptratio  black
## 8    0.14455 12.5  7.87    0 0.5240 6.172  96.1  5.9505   5 311    15.2 396.90
## 142  1.62864  0.0 21.89    0 0.6240 5.019 100.0  1.4394   4 437    21.2 396.90
## 149  2.33099  0.0 19.58    0 0.8710 5.186  93.8  1.5296   5 403    14.7 356.99
## 215  0.28955  0.0 10.59    0 0.4890 5.412   9.8  3.5875   4 277    18.6 348.93
## 254  0.36894 22.0  5.86    0 0.4310 8.259   8.4  8.9067   7 330    19.1 396.90
## 365  3.47428  0.0 18.10    1 0.7180 8.780  82.9  1.9047  24 666    20.2 354.55
## 366  4.55587  0.0 18.10    0 0.7180 3.561  87.9  1.6132  24 666    20.2 354.70
## 368 13.52220  0.0 18.10    0 0.6310 3.863 100.0  1.5106  24 666    20.2 131.42
## 369  4.89822  0.0 18.10    0 0.6310 4.970 100.0  1.3325  24 666    20.2 375.52
## 370  5.66998  0.0 18.10    1 0.6310 6.683  96.8  1.3567  24 666    20.2 375.33
## 371  6.53876  0.0 18.10    1 0.6310 7.016  97.5  1.2024  24 666    20.2 392.05
## 372  9.23230  0.0 18.10    0 0.6310 6.216 100.0  1.1691  24 666    20.2 366.15
## 373  8.26725  0.0 18.10    1 0.6680 5.875  89.6  1.1296  24 666    20.2 347.88
## 374 11.10810  0.0 18.10    0 0.6680 4.906 100.0  1.1742  24 666    20.2 396.90
## 375 18.49820  0.0 18.10    0 0.6680 4.138 100.0  1.1370  24 666    20.2 396.90
## 398  7.67202  0.0 18.10    0 0.6930 5.747  98.9  1.6334  24 666    20.2 393.10
## 399 38.35180  0.0 18.10    0 0.6930 5.453 100.0  1.4896  24 666    20.2 396.90
## 400  9.91655  0.0 18.10    0 0.6930 5.852  77.8  1.5004  24 666    20.2 338.16
## 401 25.04610  0.0 18.10    0 0.6930 5.987 100.0  1.5888  24 666    20.2 396.90
## 402 14.23620  0.0 18.10    0 0.6930 6.343 100.0  1.5741  24 666    20.2 396.90
## 404 24.80170  0.0 18.10    0 0.6930 5.349  96.0  1.7028  24 666    20.2 396.90
## 406 67.92080  0.0 18.10    0 0.6930 5.683 100.0  1.4254  24 666    20.2 384.97
## 408 11.95110  0.0 18.10    0 0.6590 5.608 100.0  1.2852  24 666    20.2 332.09
## 410 14.43830  0.0 18.10    0 0.5970 6.852 100.0  1.4655  24 666    20.2 179.36
## 413 18.81100  0.0 18.10    0 0.5970 4.628 100.0  1.5539  24 666    20.2  28.79
## 417 10.83420  0.0 18.10    0 0.6790 6.782  90.8  1.8195  24 666    20.2  21.57
## 420 11.81230  0.0 18.10    0 0.7180 6.824  76.5  1.7940  24 666    20.2  48.45
## 427 12.24720  0.0 18.10    0 0.5840 5.837  59.7  1.9976  24 666    20.2  24.65
## 490  0.18337  0.0 27.74    0 0.6090 5.414  98.3  1.7554   4 711    20.1 344.05
## 506  0.04741  0.0 11.93    0 0.5730 6.030  80.8  2.5050   1 273    21.0 396.90
## 121  0.06899  0.0 25.65    0 0.5810 5.870  69.7  2.2577   2 188    19.1 389.15
## 122  0.07165  0.0 25.65    0 0.5810 6.004  84.1  2.1974   2 188    19.1 377.67
## 123  0.09299  0.0 25.65    0 0.5810 5.961  92.9  2.0869   2 188    19.1 378.09
## 124  0.15038  0.0 25.65    0 0.5810 5.856  97.0  1.9444   2 188    19.1 370.31
## 125  0.09849  0.0 25.65    0 0.5810 5.879  95.8  2.0063   2 188    19.1 379.38
## 126  0.16902  0.0 25.65    0 0.5810 5.986  88.4  1.9929   2 188    19.1 385.02
## 127  0.38735  0.0 25.65    0 0.5810 5.613  95.6  1.7572   2 188    19.1 359.29
## 143  3.32105  0.0 19.58    1 0.8710 5.403 100.0  1.3216   5 403    14.7 396.90
## 146  2.37934  0.0 19.58    0 0.8710 6.130 100.0  1.4191   5 403    14.7 172.91
## 153  1.12658  0.0 19.58    1 0.8710 5.012  88.0  1.6102   5 403    14.7 343.28
## 155  1.41385  0.0 19.58    1 0.8710 6.129  96.0  1.7494   5 403    14.7 321.02
## 156  3.53501  0.0 19.58    1 0.8710 6.152  82.6  1.7455   5 403    14.7  88.01
## 157  2.44668  0.0 19.58    0 0.8710 5.272  94.0  1.7364   5 403    14.7  88.63
## 163  1.83377  0.0 19.58    1 0.6050 7.802  98.2  2.0407   5 403    14.7 389.61
## 164  1.51902  0.0 19.58    1 0.6050 8.375  93.9  2.1620   5 403    14.7 388.45
## 284  0.01501 90.0  1.21    1 0.4010 7.923  24.8  5.8850   1 198    13.6 395.52
## 354  0.01709 90.0  2.02    0 0.4100 6.728  36.1 12.1265   5 187    17.0 384.46
## 355  0.04301 80.0  1.91    0 0.4130 5.663  21.9 10.5857   4 334    22.0 382.80
## 356  0.10659 80.0  1.91    0 0.4130 5.936  19.5 10.5857   4 334    22.0 376.04
## 381 88.97620  0.0 18.10    0 0.6710 6.968  91.9  1.4165  24 666    20.2 396.90
## 411 51.13580  0.0 18.10    0 0.5970 5.757 100.0  1.4130  24 666    20.2   2.60
## 415 45.74610  0.0 18.10    0 0.6930 4.519 100.0  1.6582  24 666    20.2  88.27
## 419 73.53410  0.0 18.10    0 0.6790 5.957 100.0  1.8026  24 666    20.2  16.45
## 428 37.66190  0.0 18.10    0 0.6790 6.202  78.7  1.8629  24 666    20.2  18.82
```

```
## 489  0.15086  0.0 27.74    0 0.6090 5.454  92.7  1.8209   4 711    20.1 395.09
## 491  0.20746  0.0 27.74    0 0.6090 5.093  98.0  1.8226   4 711    20.1 318.43
## 492  0.10574  0.0 27.74    0 0.6090 5.983  98.8  1.8681   4 711    20.1 390.11
## 493  0.11132  0.0 27.74    0 0.6090 5.983  83.5  2.1099   4 711    20.1 396.90
## 65   0.01951 17.5  1.38    0 0.4161 7.104  59.5  9.2229   3 216    18.6 393.24
## 148  2.36862  0.0 19.58    0 0.8710 4.926  95.7  1.4608   5 403    14.7 391.71
## 167  2.01019  0.0 19.58    0 0.6050 7.929  96.2  2.0459   5 403    14.7 369.30
##     lstat medv
## 8   19.15 27.1
## 142 34.41 14.4
## 149 28.32 17.8
## 215 29.55 23.7
## 254  3.54 42.8
## 365  5.29 21.9
## 366  7.12 27.5
## 368 13.33 23.1
## 369  3.26 50.0
## 370  3.73 50.0
## 371  2.96 50.0
## 372  9.53 50.0
## 373  8.88 50.0
## 374 34.77 13.8
## 375 37.97 13.8
## 398 19.92  8.5
## 399 30.59  5.0
## 400 29.97  6.3
## 401 26.77  5.6
## 402 20.32  7.2
## 404 19.77  8.3
## 406 22.98  5.0
## 408 12.13 27.9
## 410 19.78 27.5
## 413 34.37 17.9
## 417 25.79  7.5
## 420 22.74  8.4
## 427 15.69 10.2
## 490 23.97  7.0
## 506  7.88 11.9
## 121 14.37 22.0
## 122 14.27 20.3
## 123 17.93 20.5
## 124 25.41 17.3
## 125 17.58 18.8
## 126 14.81 21.4
## 127 27.26 15.7
## 143 26.82 13.4
## 146 27.80 13.8
## 153 12.12 15.3
## 155 15.12 17.0
## 156 15.02 15.6
## 157 16.14 13.1
## 163  1.92 50.0
## 164  3.32 50.0
## 284  3.16 50.0
## 354  4.50 30.1
```

```
## 355  8.05 18.2
## 356  5.57 20.6
## 381 17.21 10.4
## 411 10.11 15.0
## 415 36.98  7.0
## 419 20.62  8.8
## 428 14.52 10.9
## 489 18.06 15.2
## 491 29.68  8.1
## 492 18.07 13.6
## 493 13.35 20.1
## 65   8.05 33.0
## 148 29.53 14.6
## 167  3.70 50.0
```

Another way to detact outliers and influential points is to see the dffits:

```
## [1] 0.3205726
```

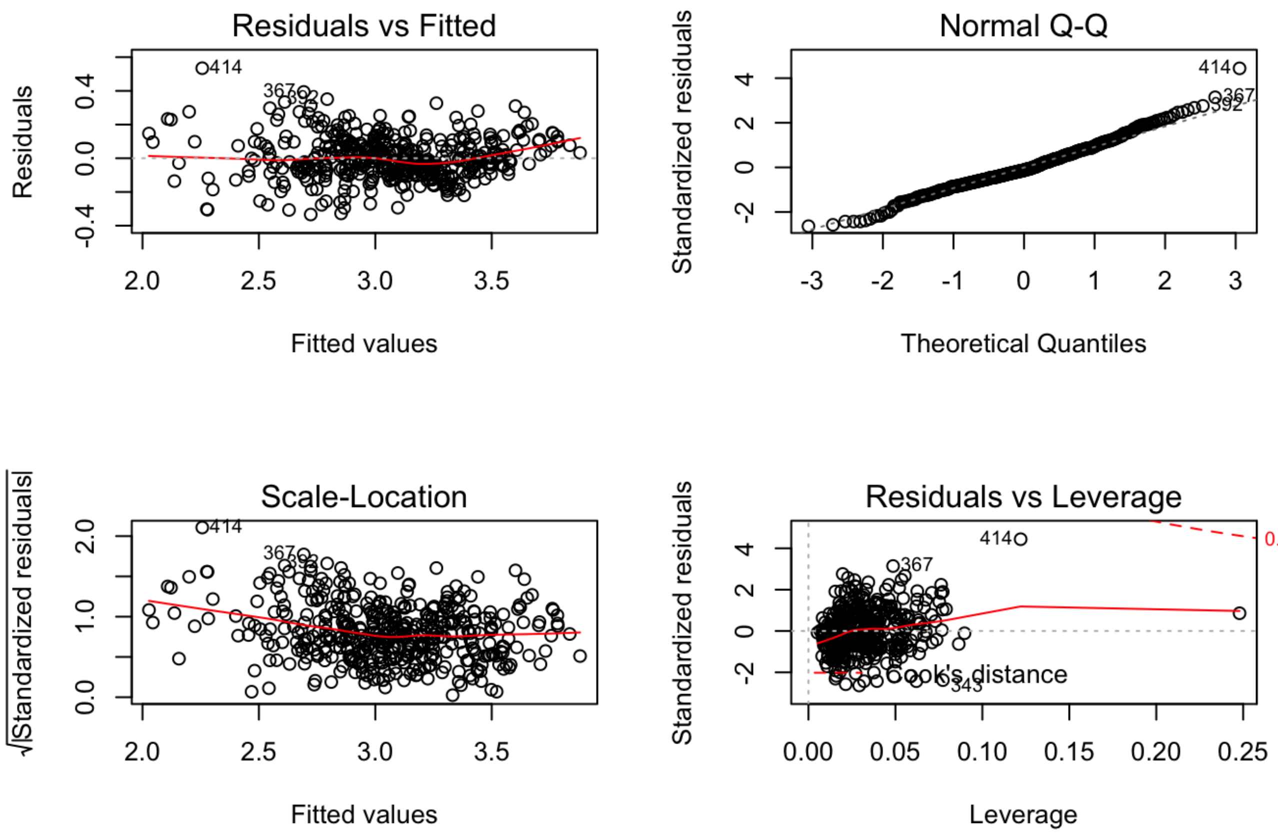


```
##         crim   zn indus chas    nox    rm   age    dis rad tax ptratio  black
## 8    0.14455 12.5  7.87    0 0.5240 6.172  96.1 5.9505   5 311    15.2 396.90
## 9    0.21124 12.5  7.87    0 0.5240 5.631 100.0 6.0821   5 311    15.2 386.63
## 65   0.01951 17.5  1.38    0 0.4161 7.104  59.5 9.2229   3 216    18.6 393.24
## 142  1.62864  0.0 21.89    0 0.6240 5.019 100.0 1.4394   4 437    21.2 396.90
## 148  2.36862  0.0 19.58    0 0.8710 4.926  95.7 1.4608   5 403    14.7 391.71
## 149  2.33099  0.0 19.58    0 0.8710 5.186  93.8 1.5296   5 403    14.7 356.99
## 153  1.12658  0.0 19.58    1 0.8710 5.012  88.0 1.6102   5 403    14.7 343.28
## 167  2.01019  0.0 19.58    0 0.6050 7.929  96.2 2.0459   5 403    14.7 369.30
## 215  0.28955  0.0 10.59    0 0.4890 5.412   9.8 3.5875   4 277    18.6 348.93
## 254  0.36894 22.0  5.86    0 0.4310 8.259   8.4 8.9067   7 330    19.1 396.90
## 365  3.47428  0.0 18.10    1 0.7180 8.780  82.9 1.9047  24 666    20.2 354.55
## 366  4.55587  0.0 18.10    0 0.7180 3.561  87.9 1.6132  24 666    20.2 354.70
## 368 13.52220  0.0 18.10    0 0.6310 3.863 100.0 1.5106  24 666    20.2 131.42
## 369  4.89822  0.0 18.10    0 0.6310 4.970 100.0 1.3325  24 666    20.2 375.52
## 370  5.66998  0.0 18.10    1 0.6310 6.683  96.8 1.3567  24 666    20.2 375.33
## 371  6.53876  0.0 18.10    1 0.6310 7.016  97.5 1.2024  24 666    20.2 392.05
## 372  9.23230  0.0 18.10    0 0.6310 6.216 100.0 1.1691  24 666    20.2 366.15
## 373  8.26725  0.0 18.10    1 0.6680 5.875  89.6 1.1296  24 666    20.2 347.88
## 374 11.10810  0.0 18.10    0 0.6680 4.906 100.0 1.1742  24 666    20.2 396.90
## 375 18.49820  0.0 18.10    0 0.6680 4.138 100.0 1.1370  24 666    20.2 396.90
## 381 88.97620  0.0 18.10    0 0.6710 6.968  91.9 1.4165  24 666    20.2 396.90
## 398  7.67202  0.0 18.10    0 0.6930 5.747  98.9 1.6334  24 666    20.2 393.10
## 399 38.35180  0.0 18.10    0 0.6930 5.453 100.0 1.4896  24 666    20.2 396.90
## 400  9.91655  0.0 18.10    0 0.6930 5.852  77.8 1.5004  24 666    20.2 338.16
## 401 25.04610  0.0 18.10    0 0.6930 5.987 100.0 1.5888  24 666    20.2 396.90
## 402 14.23620  0.0 18.10    0 0.6930 6.343 100.0 1.5741  24 666    20.2 396.90
## 404 24.80170  0.0 18.10    0 0.6930 5.349  96.0 1.7028  24 666    20.2 396.90
## 406 67.92080  0.0 18.10    0 0.6930 5.683 100.0 1.4254  24 666    20.2 384.97
## 410 14.43830  0.0 18.10    0 0.5970 6.852 100.0 1.4655  24 666    20.2 179.36
## 411 51.13580  0.0 18.10    0 0.5970 5.757 100.0 1.4130  24 666    20.2   2.60
## 413 18.81100  0.0 18.10    0 0.5970 4.628 100.0 1.5539  24 666    20.2  28.79
## 414 28.65580  0.0 18.10    0 0.5970 5.155 100.0 1.5894  24 666    20.2 210.97
## 415 45.74610  0.0 18.10    0 0.6930 4.519 100.0 1.6582  24 666    20.2  88.27
## 417 10.83420  0.0 18.10    0 0.6790 6.782  90.8 1.8195  24 666    20.2  21.57
## 419 73.53410  0.0 18.10    0 0.6790 5.957 100.0 1.8026  24 666    20.2  16.45
## 420 11.81230  0.0 18.10    0 0.7180 6.824  76.5 1.7940  24 666    20.2  48.45
## 427 12.24720  0.0 18.10    0 0.5840 5.837  59.7 1.9976  24 666    20.2  24.65
## 490  0.18337  0.0 27.74    0 0.6090 5.414  98.3 1.7554   4 711    20.1 344.05
## 493  0.11132  0.0 27.74    0 0.6090 5.983  83.5 2.1099   4 711    20.1 396.90
## 506  0.04741  0.0 11.93    0 0.5730 6.030  80.8 2.5050   1 273    21.0 396.90
##     lstat medv
## 8   19.15 27.1
## 9   29.93 16.5
## 65   8.05 33.0
## 142 34.41 14.4
## 148 29.53 14.6
## 149 28.32 17.8
## 153 12.12 15.3
## 167  3.70 50.0
## 215 29.55 23.7
## 254  3.54 42.8
## 365  5.29 21.9
## 366  7.12 27.5
## 368 13.33 23.1
```

```
## 369  3.26 50.0
## 370  3.73 50.0
## 371  2.96 50.0
## 372  9.53 50.0
## 373  8.88 50.0
## 374 34.77 13.8
## 375 37.97 13.8
## 381 17.21 10.4
## 398 19.92  8.5
## 399 30.59  5.0
## 400 29.97  6.3
## 401 26.77  5.6
## 402 20.32  7.2
## 404 19.77  8.3
## 406 22.98  5.0
## 410 19.78 27.5
## 411 10.11 15.0
## 413 34.37 17.9
## 414 20.08 16.3
## 415 36.98  7.0
## 417 25.79  7.5
## 419 20.62  8.8
## 420 22.74  8.4
## 427 15.69 10.2
## 490 23.97  7.0
## 493 13.35 20.1
## 506  7.88 11.9
```

```
##
## Call:
## lm(formula = log(medv) ~ crim + zn + indus + chas + nox + rm +
##     age + dis + rad + tax + ptratio + black + lstat, data = Boston[-ids,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33381 -0.07890 -0.01630  0.07803  0.53451
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.193e+00  1.647e-01  19.388  < 2e-16 ***
## crim        -1.305e-02  2.097e-03  -6.222 1.17e-09 ***
## zn           6.791e-04  4.009e-04   1.694 0.090998 .
## indus        8.602e-04  2.103e-03   0.409 0.682665
## chas         7.676e-02  2.800e-02   2.741 0.006378 **
## nox         -3.994e-01  1.219e-01  -3.276 0.001137 **
## rm           1.723e-01  1.454e-02  11.848  < 2e-16 ***
## age         -1.163e-03  4.022e-04  -2.893 0.004011 **
## dis         -4.413e-02  6.018e-03  -7.333 1.12e-12 ***
## rad          1.024e-02  2.378e-03   4.303 2.08e-05 ***
## tax         -5.063e-04  1.365e-04  -3.709 0.000235 ***
## ptratio     -3.193e-02  3.749e-03  -8.515 2.80e-16 ***
## black        5.717e-04  8.597e-05   6.649 8.93e-11 ***
## lstat       -2.119e-02  1.928e-03 -10.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1285 on 431 degrees of freedom
## Multiple R-squared:  0.8739, Adjusted R-squared:  0.8701
## F-statistic: 229.7 on 13 and 431 DF,  p-value: < 2.2e-16
```

Obviously, from the ouput of the new model, it can be seen that the adjusted R-squared is about 0.8701 which improves a lot from the model without remove unusual points.



More importantly, the diagnostic plots show that the situation is much better, the residuals plot shows linearity, constant variance assumptions are satisfied, and the normal qq plot shows the residuals points fit the straight line well enough now, there are no obvious points far from the line at the ends, thus, the normaliy assumption is true too. The model is indeed improved a lot.

##6. Model Selection ##6.1 Backward Selection(manual)

After log transformation, we will use the transformed linear regression models to see if the data fits. For the part below we are going to use backward selection method, which means we will begin with the full model containing all all predictors and all variables, and gradually remove or remodify the insignificant variables and correlated variable to increase the accuracy of the model, one at a time.

From the summary of model in section 5, we can see that the p-values of “crim”, “rm”, “dis”, “rad”, “tax”, “ptratio”, “black”, “lstat” are small enough to be used to reject the null hypothesis of beta = 0. However, the p-value of the “indus” are way larger then the regular alpha 0.05. The p-value of the variable “indus” is 0.6827. Therefore, for variables “indus”, we fail to reject the null hypothesis with significant level = 0.1, and it is not statistically significant in this model. Moreover, the p value of variable “zn” is 0.0909 which will fail to reject the null hyphothesis with significant level = 0.05. Therefore, we can conclude that variable “zn” is much less significant.

Furthermore, for the variables “chas”,“nox”,and “age”, their associated p-values are 0.0064, 0.0011 and 0.0040. Those variables would be considered as less significant variables here.

model1

```
##
## Call:
## lm(formula = log(medv) ~ crim + zn + chas + nox + rm + age +
##     dis + rad + tax + ptratio + black + lstat, data = Boston[-ids,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33328 -0.07993 -0.01622  0.07790  0.53489
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.192e+00  1.645e-01  19.403  < 2e-16 ***
## crim        -1.307e-02  2.094e-03  -6.243 1.03e-09 ***
## zn           6.508e-04  3.945e-04   1.650  0.09974 .
## chas         7.820e-02  2.775e-02   2.817  0.00506 **
## nox         -3.868e-01  1.179e-01  -3.282  0.00111 **
## rm           1.714e-01  1.438e-02  11.923  < 2e-16 ***
## age         -1.165e-03  4.018e-04  -2.900  0.00392 **
## dis         -4.455e-02  5.924e-03  -7.520 3.19e-13 ***
## rad          1.000e-02  2.307e-03   4.336 1.81e-05 ***
## tax         -4.792e-04  1.193e-04  -4.017 6.96e-05 ***
## ptratio     -3.179e-02  3.732e-03  -8.520 2.69e-16 ***
## black        5.702e-04  8.582e-05   6.645 9.17e-11 ***
## lstat       -2.120e-02  1.926e-03 -11.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1284 on 432 degrees of freedom
## Multiple R-squared:  0.8738, Adjusted R-squared:  0.8703
## F-statistic: 249.3 on 12 and 432 DF,  p-value: < 2.2e-16
```

We will gradually remove these insignificant and less significant variables from our model, the first step is removing variable “indus”, which is totally insignificant.

Compared to the thrid model, we can see R-squared decreased a little bit from 0.8739 to 0.8738, however, adjusted r-squared here increased a liitle bit from 0.8701 to 0,8703. From this result, we can conclude that removing variable “indus” help improving the model a bit.

Variable “zn” here is still much less significant with p value that is 0.0997, and “chas”,“nox”,and “age” are still less significan. model2

```
##
## Call:
## lm(formula = log(medv) ~ crim + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat, data = Boston[-ids,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33538 -0.08030 -0.01559  0.07771  0.53471
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.179e+00  1.646e-01  19.307  < 2e-16 ***
## crim        -1.261e-02  2.079e-03  -6.064 2.90e-09 ***
## chas         7.891e-02  2.781e-02   2.838 0.004753 **
## nox         -3.977e-01  1.179e-01  -3.373 0.000810 ***
## rm           1.772e-01  1.397e-02  12.681  < 2e-16 ***
## age         -1.251e-03  3.992e-04  -3.134 0.001843 **
## dis         -4.041e-02  5.377e-03  -7.515 3.30e-13 ***
## rad          9.404e-03  2.283e-03   4.120 4.54e-05 ***
## tax         -4.345e-04  1.164e-04  -3.732 0.000215 ***
## ptratio     -3.380e-02  3.535e-03  -9.560  < 2e-16 ***
## black        5.741e-04  8.595e-05   6.679 7.40e-11 ***
## lstat       -2.104e-02  1.928e-03 -10.915  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1286 on 433 degrees of freedom
## Multiple R-squared:  0.873,  Adjusted R-squared:  0.8698
## F-statistic: 270.6 on 11 and 433 DF,  p-value: < 2.2e-16
```

In the second step, We removed variable “zn” which is much less significant, and will be rejected at significant level with alpha = 1. From the summary here, we can see that R-squared decreased slightly from 0.8738 to 0.873, and adjusted r-squared here also decreased a bit from 0.8703 to 0,8698. From this result, we can conclude that removing variable “indus” will not help improving the model, but makes it worse instead.

Furthermore, in model here, variable “chas” and “age” are still less significant, but the p value of variable “nox” decreased from 0.0011 to0.0008 here, which means it can reject the null hypothesis at any significant level.

model3

```
##
## Call:
## lm(formula = log(medv) ~ crim + nox + rm + age + dis + rad +
##     tax + ptratio + black + lstat, data = Boston[-ids, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33812 -0.08287 -0.01415  0.08073  0.54082
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.211e+00  1.656e-01  19.395  < 2e-16 ***
## crim        -1.321e-02  2.085e-03  -6.334 5.99e-10 ***
## nox         -4.104e-01  1.188e-01  -3.456 0.000603 ***
## rm           1.767e-01  1.408e-02  12.542  < 2e-16 ***
## age         -1.203e-03  4.020e-04  -2.993 0.002917 **
## dis         -4.144e-02  5.409e-03  -7.661 1.22e-13 ***
## rad          1.052e-02  2.267e-03   4.641 4.60e-06 ***
## tax         -4.878e-04  1.158e-04  -4.212 3.08e-05 ***
## ptratio     -3.452e-02  3.555e-03  -9.710  < 2e-16 ***
## black        5.879e-04  8.651e-05   6.796 3.56e-11 ***
## lstat       -2.073e-02  1.940e-03 -10.687  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1297 on 434 degrees of freedom
## Multiple R-squared:  0.8707, Adjusted R-squared:  0.8677
## F-statistic: 292.1 on 10 and 434 DF,  p-value: < 2.2e-16
```

In this model, we removed variable “chas”. From removing variable “chas”, we got Rsquared = 0.8707 and Adjusted R-squared= 0.8677. In this case, we can see that R-squared and Adjusted R-squared still decreased a bit compared to them in the last model.

Furthermore, in this model, variable “Age” here is still less significant.

model4

```
##
## Call:
## lm(formula = log(medv) ~ crim + nox + rm + dis + rad + tax +
##     ptratio + black + lstat, data = Boston[-ids, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33630 -0.08010 -0.01208  0.08385  0.50227
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.296e+00  1.646e-01  20.020  < 2e-16 ***
## crim        -1.257e-02  2.093e-03  -6.004 4.07e-09 ***
## nox         -5.181e-01  1.142e-01  -4.536 7.44e-06 ***
## rm           1.647e-01  1.363e-02  12.084  < 2e-16 ***
## dis         -3.559e-02  5.090e-03  -6.993 1.02e-11 ***
## rad          1.098e-02  2.282e-03   4.811 2.08e-06 ***
## tax         -4.847e-04  1.169e-04  -4.147 4.05e-05 ***
## ptratio     -3.562e-02  3.568e-03  -9.986  < 2e-16 ***
## black        5.672e-04  8.702e-05   6.518 1.97e-10 ***
## lstat       -2.345e-02  1.732e-03 -13.540  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1308 on 435 degrees of freedom
## Multiple R-squared:  0.868,  Adjusted R-squared:  0.8652
## F-statistic: 317.8 on 9 and 435 DF,  p-value: < 2.2e-16
```

In this model, we removed variable “age” which is less significant from last model, and we got R-squared=0.868 and adjusted R-sqaured=0.8652, which are both decreased slightly.It means removing these variables making this model worse.

Through doing backward selection, we find that the model without variable “indus” has the highest adjusted r-squared, which is 0.8703. In this case, we can assume this model fit the data best.

##6.2 Other Selection Method & Comparsion Summary (Forward stepwise,Backward stepwise,best subset)

Now, we are trying to use forward-stepwise selection.

```
##
## Call:
## lm(formula = log(medv) ~ lstat + ptratio + crim + rm + dis +
##     nox + black + rad + tax + chas + zn, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73400 -0.09460 -0.01771  0.09782  0.86290
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0836823  0.2030491  20.112  < 2e-16 ***
## lstat       -0.0286039  0.0019002 -15.053  < 2e-16 ***
## ptratio     -0.0374259  0.0051715  -7.237 1.77e-12 ***
## crim        -0.0103187  0.0013134  -7.856 2.49e-14 ***
## rm           0.0906728  0.0162807   5.569 4.20e-08 ***
## dis         -0.0517059  0.0074420  -6.948 1.18e-11 ***
## nox         -0.7217440  0.1416535  -5.095 4.97e-07 ***
## black        0.0004127  0.0001071   3.852 0.000133 ***
## rad          0.0134457  0.0025405   5.293 1.82e-07 ***
## tax         -0.0005579  0.0001351  -4.129 4.28e-05 ***
## chas         0.1051484  0.0342285   3.072 0.002244 **
## zn           0.0010874  0.0005418   2.007 0.045308 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1898 on 494 degrees of freedom
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7844
## F-statistic: 168.1 on 11 and 494 DF,  p-value: < 2.2e-16
```

Now, we are trying to use backward-stepwise selection.

```
##
## Call:
## lm(formula = log(medv) ~ crim + zn + chas + nox + rm + dis +
##     rad + tax + ptratio + black + lstat, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73400 -0.09460 -0.01771  0.09782  0.86290
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0836823  0.2030491  20.112  < 2e-16 ***
## crim        -0.0103187  0.0013134  -7.856 2.49e-14 ***
## zn           0.0010874  0.0005418   2.007 0.045308 *
## chas         0.1051484  0.0342285   3.072 0.002244 **
## nox         -0.7217440  0.1416535  -5.095 4.97e-07 ***
## rm           0.0906728  0.0162807   5.569 4.20e-08 ***
## dis         -0.0517059  0.0074420  -6.948 1.18e-11 ***
## rad          0.0134457  0.0025405   5.293 1.82e-07 ***
## tax         -0.0005579  0.0001351  -4.129 4.28e-05 ***
## ptratio     -0.0374259  0.0051715  -7.237 1.77e-12 ***
## black        0.0004127  0.0001071   3.852 0.000133 ***
## lstat       -0.0286039  0.0019002 -15.053  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1898 on 494 degrees of freedom
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7844
## F-statistic: 168.1 on 11 and 494 DF,  p-value: < 2.2e-16
```

We also want to try the best subset selection.

```
## Subset selection object
## Call: regsubsets.formula(log(medv) ~ ., data = Boston, nvmax = 14)
## 13 Variables  (and intercept)
##         Forced in Forced out
## crim        FALSE      FALSE
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas        FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## black       FALSE      FALSE
## lstat       FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: exhaustive
##           crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
## 1  ( 1 )  " "  " " " "   " "  " " " " " " " " " " " " " "     " "   "*"  
## 2  ( 1 )  " "  " " " "   " "  " " " " " " " " " " " " "*"     " "   "*"  
## 3  ( 1 )  "*"  " " " "   " "  " " " " " " " " " " " " "*"     " "   "*"  
## 4  ( 1 )  "*"  " " " "   " "  " " "*" " " " " " " " " "*"     " "   "*"  
## 5  ( 1 )  "*"  " " " "   " "  " " "*" " " "*" " " " " "*"     " "   "*"  
## 6  ( 1 )  "*"  " " " "   " "  "*" "*" " " "*" " " " " "*"     " "   "*"  
## 7  ( 1 )  "*"  " " " "   " "  "*" "*" " " "*" " " " " "*"     "*"   "*"  
## 8  ( 1 )  "*"  " " " "   " "  "*" "*" " " "*" "*" "*" "*"     " "   "*"  
## 9  ( 1 )  "*"  " " " "   " "  "*" "*" " " "*" "*" "*" "*"     "*"   "*"  
## 10  ( 1 ) "*"  " " " "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"  
## 11  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"  
## 12  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"  
## 13  ( 1 ) "*"  "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"  
```

```
##   Adj.R2 CP BIC
## 1     12 11  10
```

```
## log_medv ~ crim + zn + indus + chas + nox + rm + dis + rad + 
##     tax + ptratio + black + lstat
## <environment: 0x7f8a2c867a48>
```

The best subset model includes predictors “crim”, “ptratio” and “lstat”. Now, we try to fit this best model.

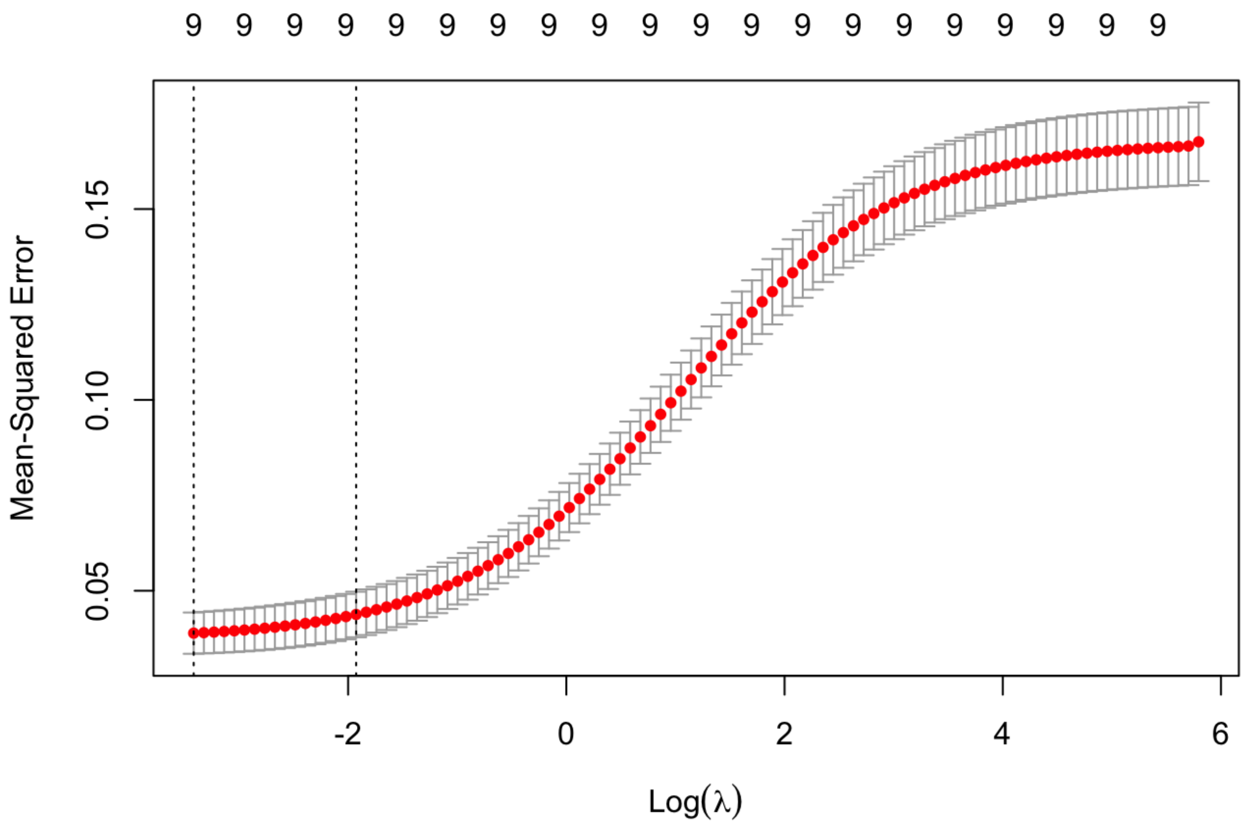
```
##
## Call:
## lm(formula = log(medv) ~ crim + zn + indus + chas + nox + rm +
##     dis + rad + tax + ptratio + black + lstat, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73345 -0.09809 -0.01744  0.09653  0.86552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0951779  0.2033706  20.137  < 2e-16 ***
## crim        -0.0102698  0.0013143  -7.814 3.38e-14 ***
## zn           0.0011461  0.0005450   2.103 0.035978 *
## indus        0.0024679  0.0024593   1.003 0.316129
## chas         0.1015851  0.0344120   2.952 0.003307 **
## nox         -0.7622525  0.1472924  -5.175 3.32e-07 ***
## rm           0.0922108  0.0163525   5.639 2.89e-08 ***
## dis         -0.0500137  0.0076307  -6.554 1.41e-10 ***
## rad          0.0141871  0.0026457   5.362 1.27e-07 ***
## tax         -0.0006240  0.0001503  -4.151 3.90e-05 ***
## ptratio     -0.0381084  0.0052160  -7.306 1.12e-12 ***
## black        0.0004163  0.0001072   3.883 0.000117 ***
## lstat       -0.0287597  0.0019066 -15.085  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1898 on 493 degrees of freedom
## Multiple R-squared:  0.7896, Adjusted R-squared:  0.7845
## F-statistic: 154.2 on 12 and 493 DF,  p-value: < 2.2e-16
```

Using forward selection or best subset selection, the selected models have lower adjusted R-square than the one we selected manually.

##7. Regularization models, Lasso, Ridge and Elastic Net

These regularization methods work by penalizing the magnitude of the coefficients of features and at the same time minimizing the error between the predicted value and actual observed values. This minimization becomes a balance between the error (the difference between the predicted value and observed value) and the size of the coefficients. The only difference between Ridge and Lasso is the way they penalize the coefficients. Elastic Net is the combination of these two. Elastic Net adds both the sum of the squares errors and the absolute value of the squared error.

First, we apply ridge method, alpha = 0, using 10-folds cross validation to tune the best lambda, the plot and the result shows the best lambda which minimize the error is about 0.0329:

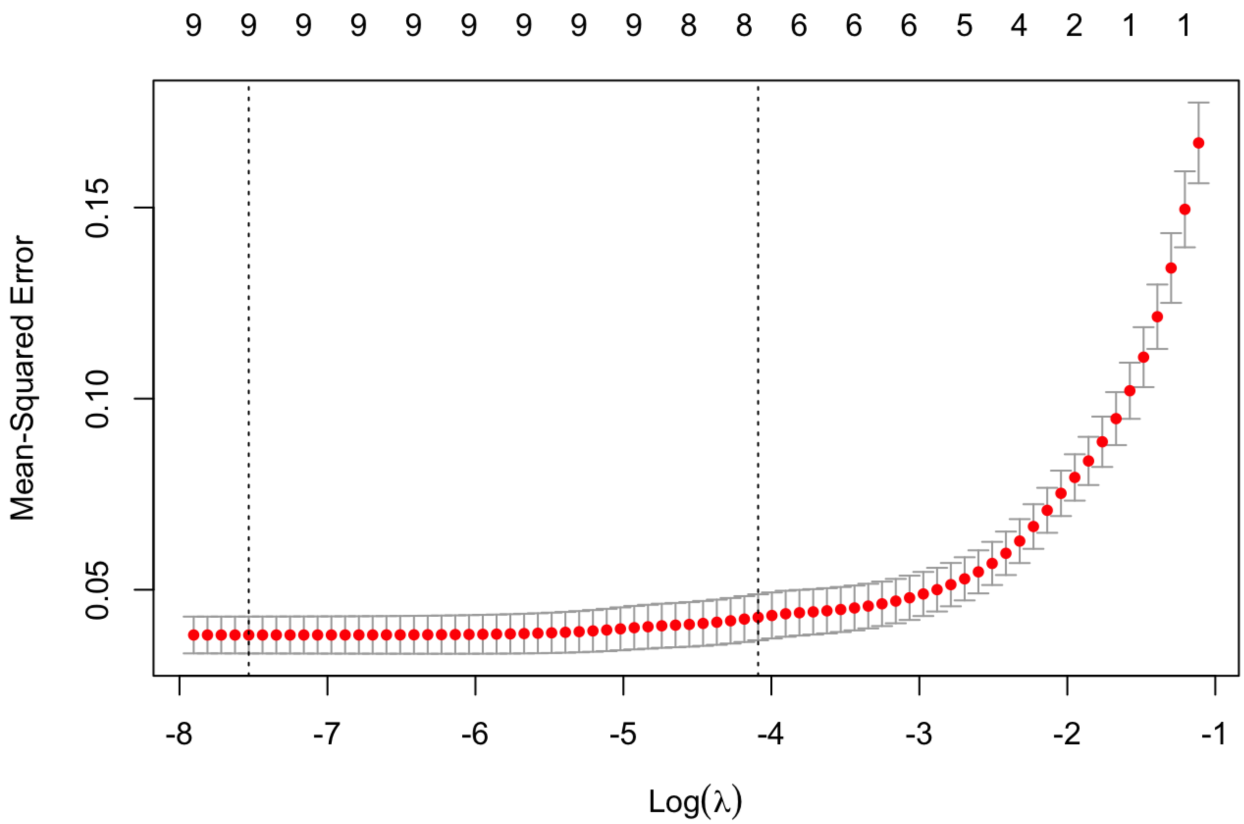



```
## [1] 0.03287379
```

The model founded by Ridge is:

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          1
## (Intercept)  3.6567054233
## crim        -0.0088617318
## nox         -0.5361625773
## rm           0.1167400487
## dis         -0.0319610367
## rad          0.0066582035
## tax         -0.0003011559
## ptratio     -0.0367520952
## black        0.0004365500
## lstat       -0.0255435491
```

Then, we apply lasso method, alpha = 1, using 10-folds cross validation to tune the best lambda, the plot and the result shows the best lambda which minimize the error is about 0.0005:

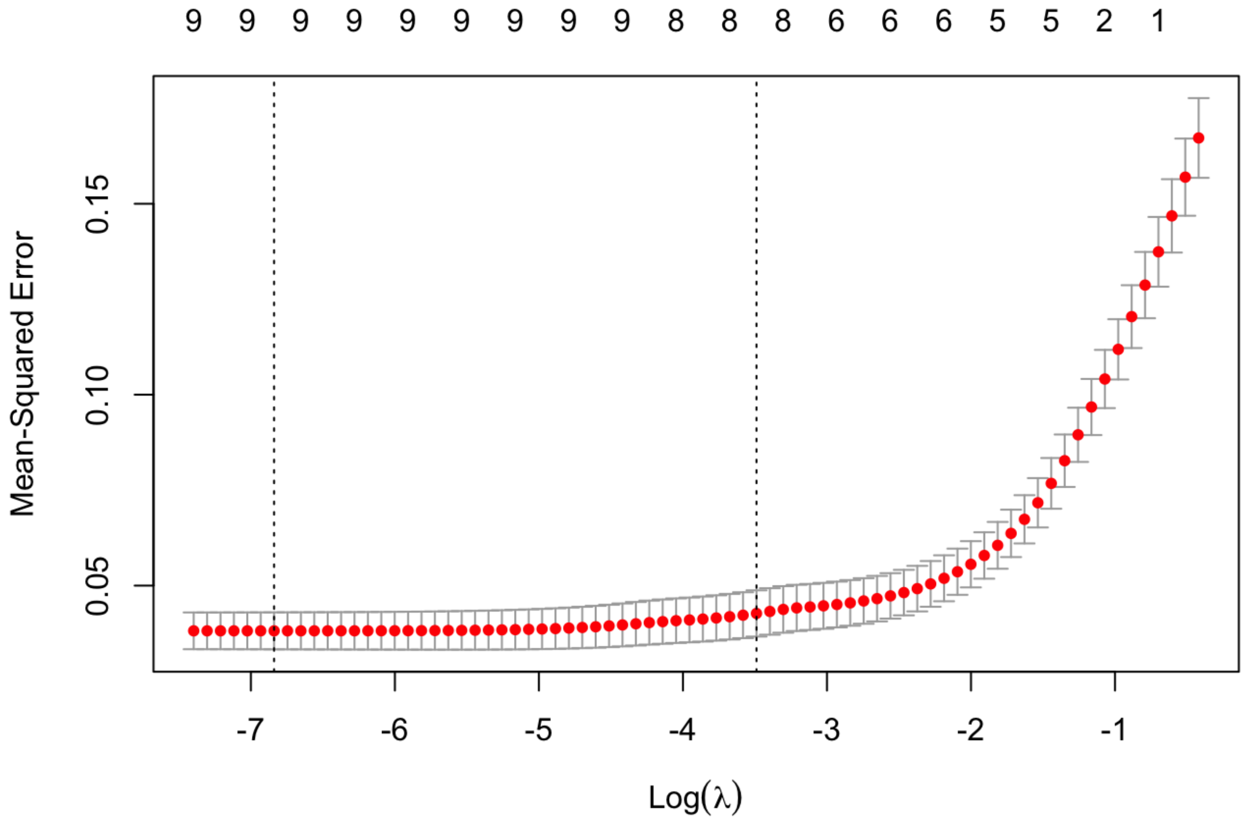



```
## [1] 0.0005357609
```

The model founded by Lasso is:

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          1
## (Intercept)  4.0725508193
## crim        -0.0101245677
## nox         -0.6940600918
## rm           0.0979479166
## dis         -0.0442748667
## rad          0.0127874622
## tax         -0.0005040470
## ptratio     -0.0415932952
## black        0.0004255423
## lstat       -0.0288071052
```

At last, we use Elastic Net, alpha = 0.5, using 10-folds cross validation to tune the best lambda, the plot and the result shows the best lambda which minimize the error is about 0.0011:




```
## [1] 0.001071522
```

The model founded by Elastic Net is:

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  4.0614574971
## crim        -0.0100902741
## nox         -0.6895943339
## rm           0.0984649630
## dis         -0.0439786477
## rad          0.0126110887
## tax         -0.0004978419
## ptratio     -0.0414635598
## black        0.0004257784
## lstat       -0.0287445748
```

8. Conclusion

Finally, with the best model founded, we can answer the key questions about the dataset:

8.1 What are the top five variables affecting the Boston housing price most?

The top five variables affecting the Boston housing price most are rm,lstat, ptratio, dis and black which have smallest p values.

8.2 What is the relationship between the top five variables and the Boston housing price?

Fixed other factors, average number of rooms per dwelling(rm) increases 1 unit, the house price would increase 17.23%. Fixed other factors, lower status of the population (lstat) increase 1 unit, the house price would decrease 2.12%. Fixed other factors, pupil-teacher ratio by town. (pratio) increase 1 unit, the house price would decrease 3.19%. Fixed other factors, weighted mean of distances to five Boston employment centres (dis) increase 1 unit, the house price would decrease 4.41%. Fixed other factors, the proportion of blacks by town(black) increase 1 unit, the house price would decrease 0.06%.

8.3 Find the linear regression model which predicts the relationship best.

The model founded is:

$$\hat{medv} = e^{3.193-0.013crim+0.001zn+0.078chas-0.387nox+0.171rm-0.001age-0.045dis+0.01rad-0.001tax+-0.032ptratio+0.001black-0.021lstat}$$

#Further Study Data Science is an interesting discipline that allows us to discover the world of data in creative ways. The further study of this data set for our group including:

a. Further comparsions between different data transformation models via cross validation
b. Further comparsions between model selection beyond comparing through adjusted R squares. This could including comparing the diagnositic table, AIC and etc.