Stat 151 a
Midterm 2
Hongzhe Li
24474061

Introduction:

This project is mainly to analyze the Boston Housing data in the mlbench library of R, and try to build a model that can be used to model the relationship between housing price and the explanatory variables based on given data.
After seen those explanatory variables, my intuition says that crim must be an important explanatory variable in building up the model since based on common sense that areas which are dangerous will lead to a lower price for housing.

Description of the data:

As I mentioned earlier, the data is in the R library, which are housing data for 506 census tracts of Boston from the 1970 census. The dataframe BostonHousing contains the original data by Harrison and Rubinfeld (1979).
There are 506 observations on 14 variables, and the description of each variable is defined below:

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

Description of the analyses and results:

Clean the data:

First of all, I plotted the scatterplot matrices using pairs (figure 1) in order to have a rough view of the data and check if I can find something strange. Clearly, for rad VS medv and tax VS medv, there are two groups. As a result I think the data can be splited into two parts and I'll try to clean the data first and then start my analyses.

Figure 1

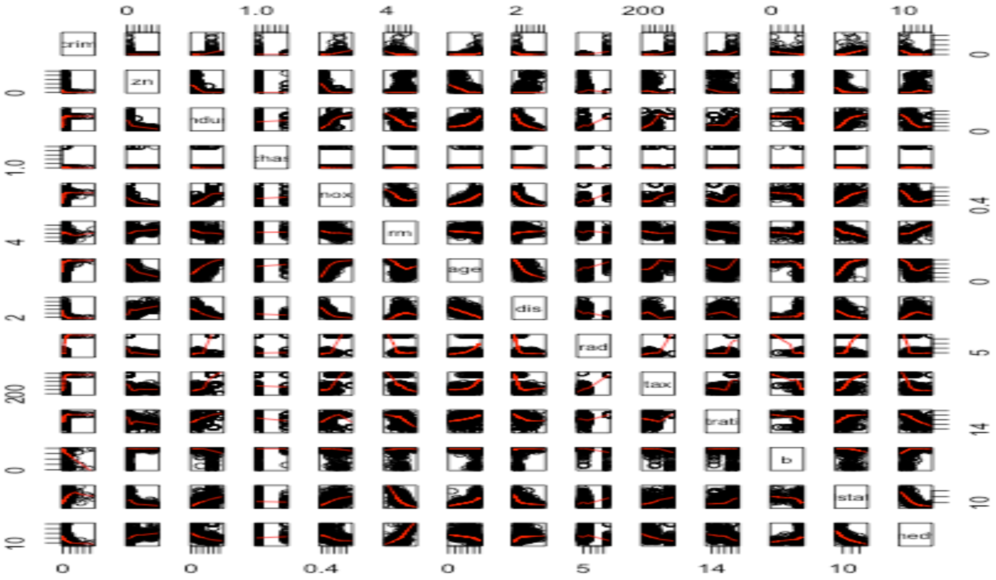
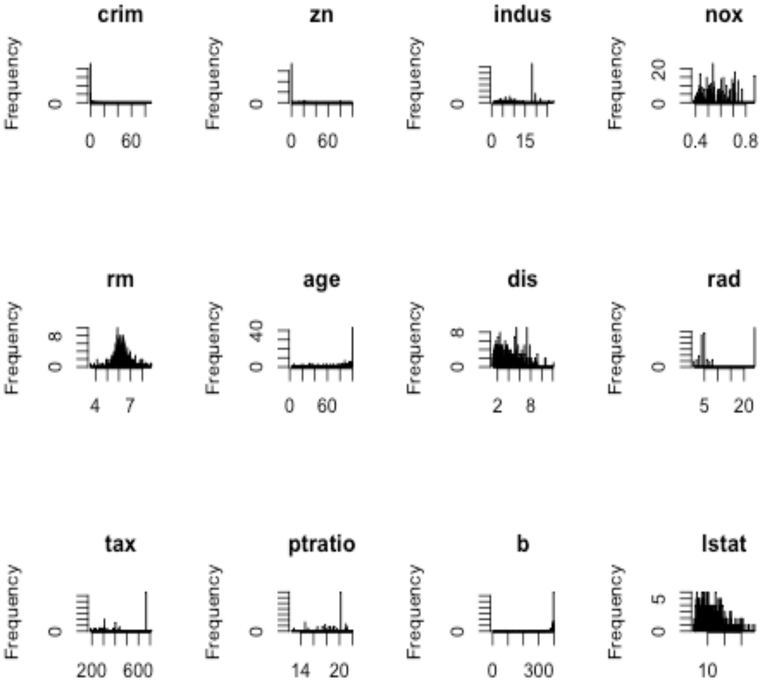


Figure 2

Second, I plotted the histogram (figure 2) for each explanatory variable VS medv (not for chas which is a binary variable), and I found several strange high frequencies (suche as rad = 24, tax = 666). I will extract all those data with high frequencies and treat the overlapping points as one group and the rest as another.



Since data from 357 to 488 has very high frequency and are also overlapping for almost every explanatory variables. I decided to regard data of number 357 to 488(high frequency) as one group and the rest as another.

I'll only investigate the low frequency group since these are the normal cases, which can be used to build a better model.

Also I realized that there are many data having medv equals to 50 no matter how the other explanatory changes. I'll also remove those points since they are not helpful in building up models.

In addition, by plotting scatterplot matrices again which reminds me that I need to do some variable transformation in order to make the data more symmetric. Then I took the log of crim, indus, b, and lstat, eventually the residuals VS fitted value plot is the best I have seen so far.

Then, I will investigate if there are any possible outliers in my selected data using histogram, leverage, standardized/studendized residuals, cook's distance and treat the overlapping points as possible outliers.
First, I used the same histograms which I have plotted before and observe if there are any points which are weigh different away from most data.(crim > 3, nox > 0.8, tax > 600, b <= 200)
The overlapping points are 143, 146, 147, 156, 157, 158 which are the probably outliers.
Then for the leverage, I found 103 143 146 147 153 155 156 157 are having high
For the standardized and studendized residual, I got 158 178 363.
For the cooks distance, the resulting points are 156 158 178 208 332.
Since cooks distance are the strongest way of determining outliers, I chose the possible outliers are 156, 158 and 178, which will be used to compared with my later models and then finalize if those are truly outliers.

Variable selections:

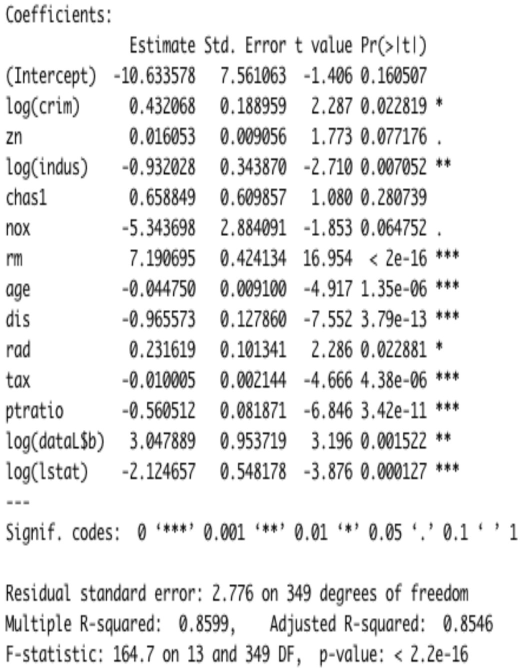
I'll use step, regsubsets in the leap library and cross-validation for my variable selection. The criteria I selected are adjusted R^2, AIC, BIC, and Mallow's Cp.
By using both direction step function based on AIC, and Mallows' CP I got the model with tweleve variables :
M2:medv ~ log(crim) + zn + log(indus) + chas + nox + rm + age + dis + tax + ptratio + log(dataL$b) + log(lstat)

By spotting the largest adjusted R^2, I got the full model(M3) with 13 variables

Smallest BIC gives me the nine variables model:
M1:medv ~ log(indus) + rm + age + dis + rad + tax + ptratio + log(dataL$b) + log(lstat)

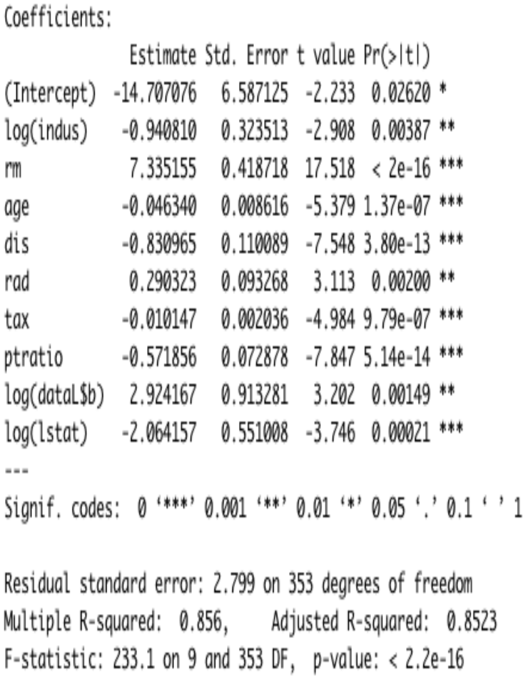
Thereafter, I applied cross-validation on the selected three models and the results are as follows:
2960.530 2988.313 2958.182

Since M1 and M3 are very close I'll look at the summary of each model and decide the suitable model.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.633578   7.561063  -1.406 0.160507
log(crim)     0.432068   0.188959   2.287 0.022819 *
zn            0.016053   0.009056   1.773 0.077176 .
log(indus)   -0.932028   0.343870  -2.710 0.007052 **
chas1         0.658849   0.609857   1.080 0.280739
nox          -5.343698   2.884091  -1.853 0.064752 .
rm            7.190695   0.424134  16.954  < 2e-16 ***
age          -0.044750   0.009100  -4.917 1.35e-06 ***
dis          -0.965573   0.127860  -7.552 3.79e-13 ***
rad           0.231619   0.101341   2.286 0.022881 *
tax          -0.010005   0.002144  -4.666 4.38e-06 ***
ptratio      -0.560512   0.081871  -6.846 3.42e-11 ***
log(dataL$b)  3.047889   0.953719   3.196 0.001522 **
log(lstat)   -2.124657   0.548178  -3.876 0.000127 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.776 on 349 degrees of freedom
Multiple R-squared:  0.8599,	Adjusted R-squared:  0.8546
F-statistic: 164.7 on 13 and 349 DF,  p-value: < 2.2e-16
```



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.707076   6.587125  -2.233  0.02620 *
log(indus)   -0.940810   0.323513  -2.908  0.00387 **
rm            7.335155   0.418718  17.518  < 2e-16 ***
age          -0.046340   0.008616  -5.379 1.37e-07 ***
dis          -0.830965   0.110089  -7.548 3.80e-13 ***
rad           0.290323   0.093268   3.113  0.00200 **
tax          -0.010147   0.002036  -4.984 9.79e-07 ***
ptratio      -0.571856   0.072878  -7.847 5.14e-14 ***
log(dataL$b)  2.924167   0.913281   3.202  0.00149 **
log(lstat)   -2.064157   0.551008  -3.746  0.00021 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.799 on 353 degrees of freedom
Multiple R-squared:  0.856,	Adjusted R-squared:  0.8523
F-statistic: 233.1 on 9 and 353 DF,  p-value: < 2.2e-16
```

Even though cross-validation score of the full model is slightly better than BIC, all the explanatory variables in the BIC model are statistically significant which implies a better model.

As a result, I'll use the result given by BIC as my selected model.

Then, I'll plug in my possible outliers in my selected model (As below) in order to decide whether I should include the as outliers:

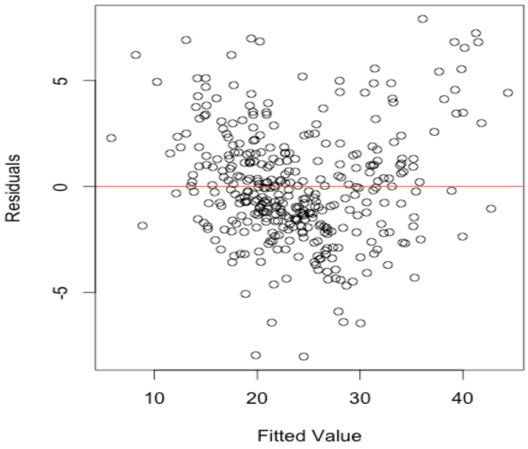
medv ~ log(indus) + rm + age + dis + rad + tax + ptratio + log(dataL$b) + log(lstat)

By comparing adjusted R^2 which improves 0.0089(0.8523 to 0.8612), I'll regard 158 and 178 as outliers and delete those outliers from my selected data.

Diagnostic:

After deleting outliers and fit the model again, I decided to run residual diagnostic in order to check how well my model works by checking error assumptions using plot of residuals VS fitted value, normality using QQplot, and correlated errors using ACF.

Figure 3



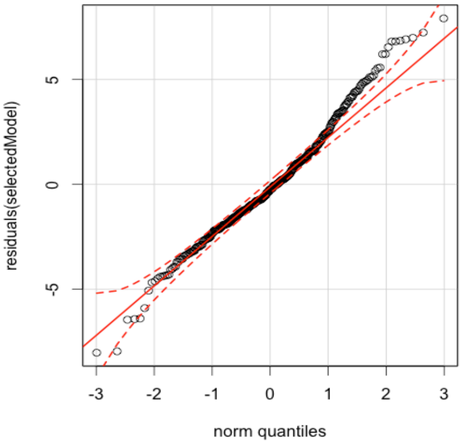


Figure 4

For residuals VS fitted Values(figure 3), this plot does not suggest any non-constant variance and non-linearity.

For the QQplot(figure4), there are some outliers in the end, but it overall performs normal.

For the ACF(figure5), there are 5 cutoffs in the beginning, but it overall perform non-correlated.

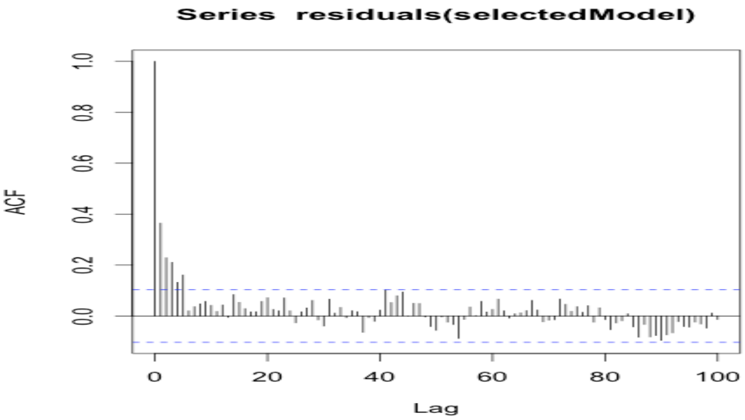


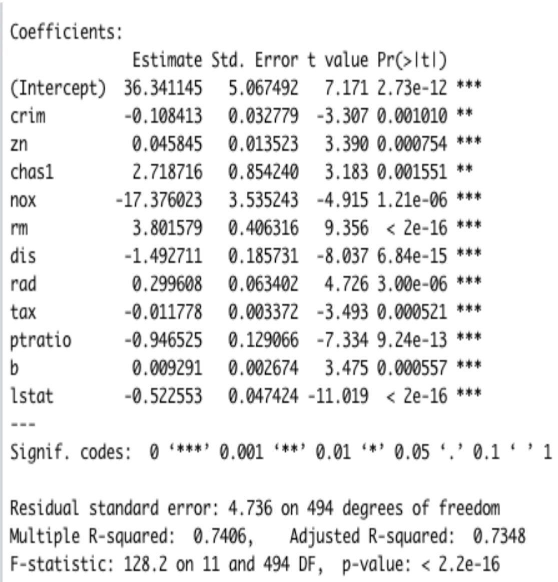
Figure 5

Finally I'll run the same procedure to check if there are possible outliers after fitting models, and the final result suggests one possible outlier (343), but including it as an outlier does not improve adjusted R^2 much(from 0.8612 to 0.8617). I will not treat it as an outlier.

Compare with model using whole data:

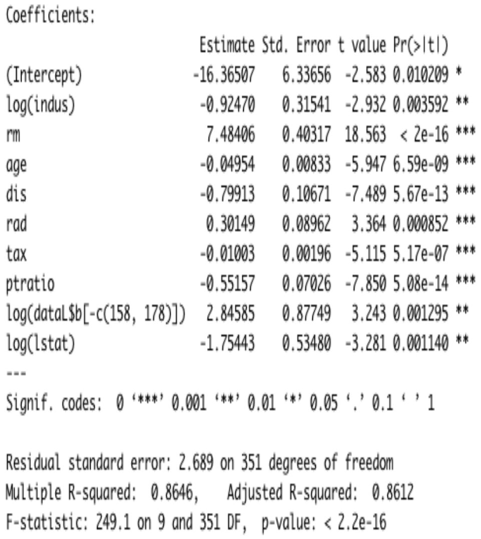
Finally I will using the full data to build an model and compare with my selected model.
Full model gives me the following:
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + b + lstat

By comparing the summary of both models, the adjusted R^2 for my selectedModel is much better than the full model.
Hence my finalized model is medv ~ log(indus) + rm + age + dis + rad + tax + ptratio + log(dataL$b) + log(lstat)



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
crim         -0.108413   0.032779  -3.307 0.001010 ** 
zn            0.045845   0.013523   3.390 0.000754 ***
chas1         2.718716   0.854240   3.183 0.001551 ** 
nox         -17.376023   3.535243  -4.915 1.21e-06 ***
rm            3.801579   0.406316   9.356  < 2e-16 ***
dis          -1.492711   0.185731  -8.037 6.84e-15 ***
rad           0.299608   0.063402   4.726 3.00e-06 ***
tax          -0.011778   0.003372  -3.493 0.000521 ***
ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
b             0.009291   0.002674   3.475 0.000557 ***
lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.736 on 494 degrees of freedom
Multiple R-squared:  0.7406,	Adjusted R-squared:  0.7348 
F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -16.36507    6.33656  -2.583 0.010209 *  
log(indus)                -0.92470    0.31541  -2.932 0.003592 ** 
rm                         7.48406    0.40317  18.563  < 2e-16 ***
age                       -0.04954    0.00833  -5.947 6.59e-09 ***
dis                       -0.79913    0.10671  -7.489 5.67e-13 ***
rad                        0.30149    0.08962   3.364 0.000852 ***
tax                       -0.01003    0.00196  -5.115 5.17e-07 ***
ptratio                   -0.55157    0.07026  -7.850 5.08e-14 ***
log(dataL$b[-c(158, 178)])  2.84585    0.87749   3.243 0.001295 ** 
log(lstat)                -1.75443    0.53480  -3.281 0.001140 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.689 on 351 degrees of freedom
Multiple R-squared:  0.8646,	Adjusted R-squared:  0.8612 
F-statistic: 249.1 on 9 and 351 DF,  p-value: < 2.2e-16
```

General discussion and conclusions:

In the beginning of this project, **I expect that the crim to be an important explanatory variable in deterring the housing price (medv), whereas it does not even included in my model.** I think I will probably invest this explanatory more detailed(such as see correlation with other variables and so on) in the future and trying to figure out why it was not included in my model.

To sum up, My idea of this project is to first clean the data which involves splitting the data, drop unnecessary terms, take variable transformation, find possible outliers using histogram, leverage, standardized and studendized residuals. Second, I'll use AIC, BIC, Forward and Backward Selection, Adjusted R^2, Mallows' Cp as my criteria to select models and using Cross-validation and summary to check my models. Third I'll check whether I should include the possible outliers I got from first step by comparing how well adjusted R^2 improves. Fourth I'll check again possible outliers and decided whether I should include it or not. Finally I compare the selected model from previous steps and compare with the selected model using the full data, which finally says that by cleaning data, I got a better model.