北京师范大学 香港浸会大学 联合国际学院
BEIJING NORMAL UNIVERSITY · HONG KONG BAPTIST UNIVERSITY
UNITED INTERNATIONAL COLLEGE

Introduction to Business Data Analytics

Boston House Pricing Model

Students:

Crystal 1930018090

Gretchen 1930024003

Lizzy 1930018036

Halie 1930018065

Lecturer:

Dr. Jingzhi ZHANG

May 15, 2022

Abstract

In our project, we tried to provide the house pricing method for real estate agencies in Boston. We used the data set collected online. After the variable interpretation and analysis, we developed the multiple-linear regression model and improve the variable selection. Thereafter, we diagnosed the model and did some correction. Thereafter, we got the final regression model. To deal with its limitation, we also constructed the conditional inference tree. Then we offered some business suggestions. At the end, we analyzed the limitation of our method and gave out some possible improved methods.

Contents

1 Background and Data Description

1.1 Background of the Problem

Our group is very interested in the influencing factors that affect housing prices.The real estate agents in the US have no basic salary and need time to build up a client base, so competition is fierce. In the US, the real estate agents are required to understand everything related to house property, from market trends, area judgements to house type, which we have referred in the regression. Needless to say, the market is crucial. If the general economic situation is not prosperous, the real estate industry will not be booming. For example, taxes like full value property tax rate are too high, which will limit the development of housing sector.

1.2 Objectives and Significance

1.3 Data Description

We find a set of data from heywhale.com.(Harrison et al.,1978)[1] This data set contains the data for census tracts in the Boston in 1970. With tracts containing no housing units or comprised entirely of institutions excluded, the Boston sample contains 506 census tracts.

We choose "median value of owner-occupied homes" as the dependent variable because the problem we studied is to predict Boston house prices .

And then, we study housing prices mainly using the characteristic price method,which means according to the various characteristic attributes that make up the real estate, identify the factors that influence the price of the house and build the model. Then we select eleven independent variables, which can be divided into the following characteristic attributes in a practical sense.

a.The first characteristic attribute is structural

- RM: Average number of rooms in owner units.RM presents spaciousness and,in a certain sense,quantity of housing.
- AGE: Proportion of owner units built prior to 1940.

b.The second characteristic attribute is Neighborhood

- CRIM: Crime rate by town. CRIM gauges the threat to well-being that households perceive in various neighborhoods of Boston. spaciousness and,in a certain sense,quantity of housing.
- ZN: Proportion of a town's residential land zoned for lots greater than 25000square feet. Such zoning restricts construction of small lot houses.
- CHAS: If tract bounds the Charles River.

- INDUS: Proportion non-retail business acres per town.INDUS serves as a proxy for the externalities associated with industry-noise, heavy traffic, and unpleasant visual effects.

- TAX: Full value property tax rate (/lO,OOO). Measures the cost of public services in each community. This measures the cost of public services in each community.

- LSTAT: Proportion of population that is lower status = 1/2 (proportion of adults without some high school education and proportion of male workers classified as laborers).

- PTRATIO: Pupil-teacher ratio by town school district.

c.The third characteristic attribute is Accessible

- DIS: Weighted distances to five employment centers in Boston region.

- RAD: Index of accessibility to radial highways.The highway access index was calculated on a town basis.Good road access variables are needed so that auto pollution variables do not capture the locational advantages of roadways.

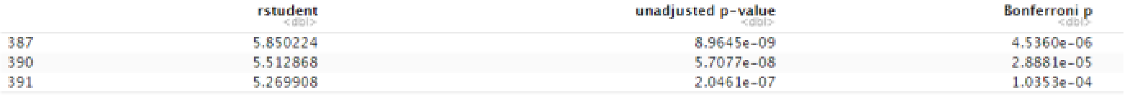
d.The forth characteristic attribute is Air Pollution

- NOX: Nitrogen oxide concentrations (annual average concentration in parts per hundred million).

2 Data Analysis and Model Building

2.1 Data Pre-processing

At the beginning, we do the data pre-processing. There is no missing value in the data set but there are three outliers, so we delete them and get a new data set.



	rstudent <dbl>	unadjusted p-value <dbl>	Bonferroni p <dbl>
387	5.850224	8.9645e-09	4.5360e-06
390	5.512868	5.7077e-08	2.8881e-05
391	5.269908	2.0461e-07	1.0353e-04

Figure 1: Outliesr

2.2 Uni-variate Statistical Data Analysis

In order to see a more general picture of the relationship between variables, we construct all the scatter plots of the paired data of each independent variable and Y, which can give us the first guess of their relationship.

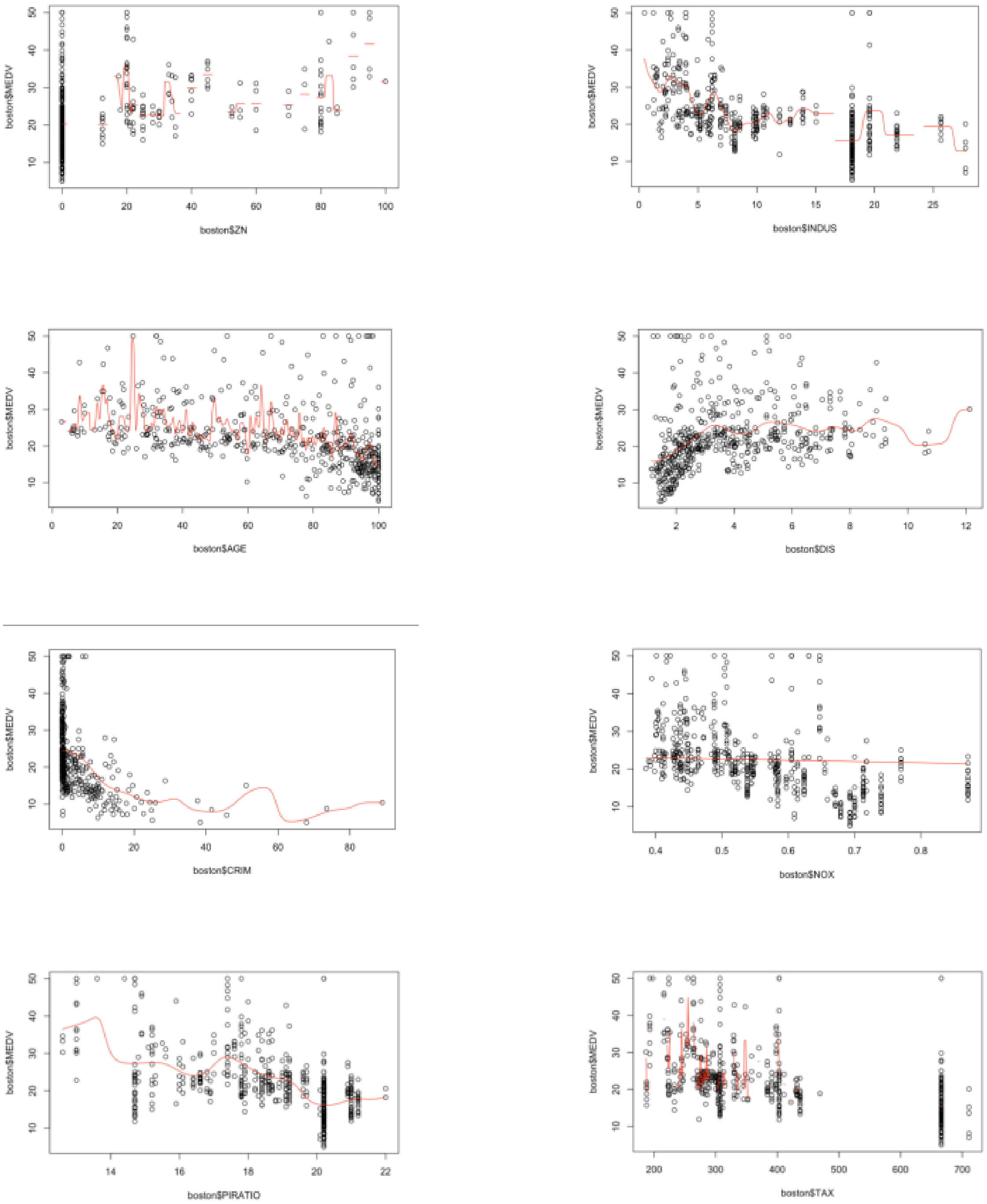


Figure 2: Scatter plots

According to the above plots, we can see that the variable CRIM, INDUS, NOX, AGE, TAX, PTRATIO have negative effects on MEDV while ZN and DIS have positive effects. And all of these variables are approximately linear with MEDV.

About the the rest of variables, we find that CHAS is a categorical variable because it just has zero and one.

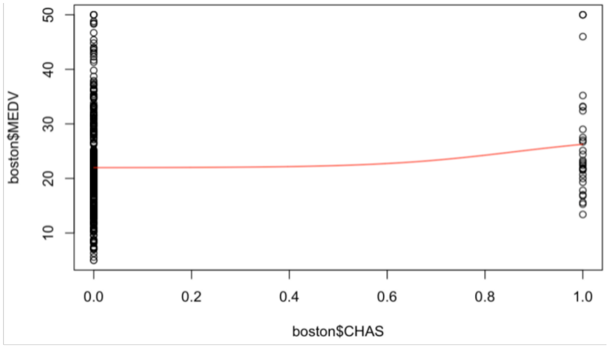


Figure 3: CHAS

About RM, it has positive effects on MEDV. It has small radian on the left but just a little points, so we can ignore them and regard it is linear with MEDV.

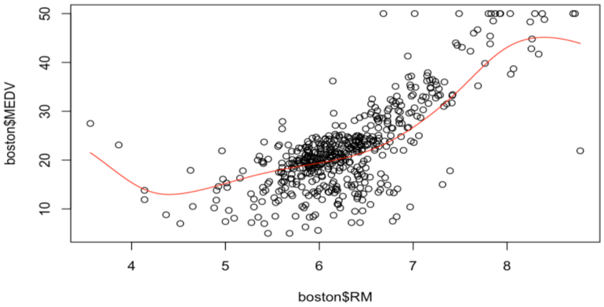


Figure 4: RM

Moreover, RAD is very special. It doesn't have a trend and all the points vertically distributed in nine columns. So we think it is an ordinary variable and we will further transform it.

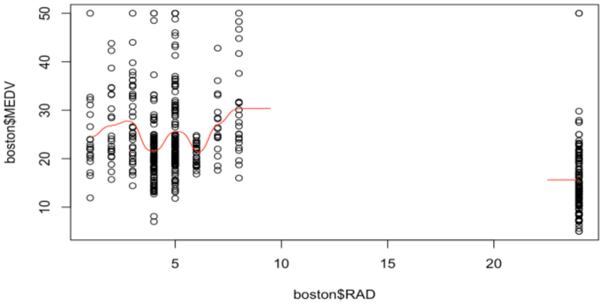


Figure 5: RAD

The last one is LSTAT. It is non - linear, so we will also transform it.

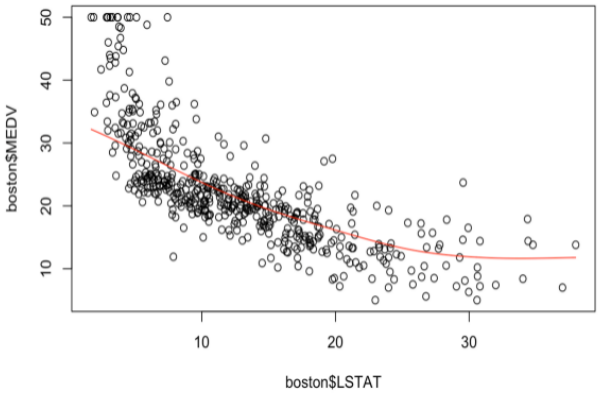


Figure 6: LSTAT

Because most variables are linear, so we can use multiple-linear regression to fit the model.

2.3 Variable Transformation

2.3.1 Transform RAD

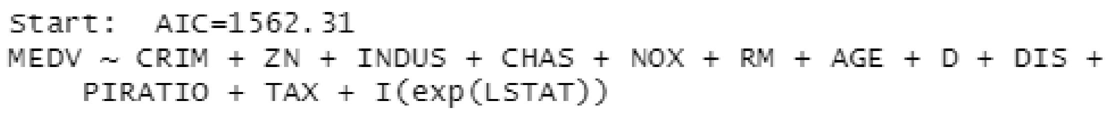
RAD is an ordinary variable. Because there are 9 levels ranging from 1 to 8 and the other one is 24. So we totally try three kinds of partition methods. One is cutting the half, one is dividing the data at level 8, the other one is dividing the data at level 7. Substituting them to the model, comparing the adjusted R square respectively, finally we decide to choose the third one which is the data larger or equal to 7 are assigned by 1 and the data less than 7 are assigned by 0. The expression is,

$$\text{RAD} \Rightarrow D := \begin{cases} 1 & \text{RAD} < 7\ (1,2,3,4,5,6) \\ 0 & \text{RAD} >= 7\ (7,8,24) \end{cases}$$

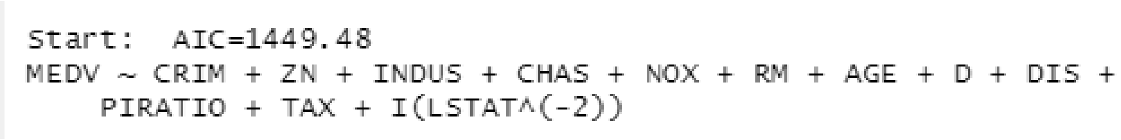
Therefore, we can generate a new variable D to represent RAD in the further analysis.

2.3.2 Transform LSTAT

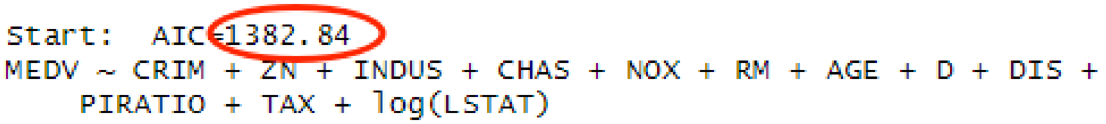
We made four assumptions that it maybe can be transformed to the polynomial 1st, 2nd, logarithmic and exponential form. We use AIC and adjusted R square as the criteria to select the best form. We find that the logarithmic has the smallest AIC and the largest adjusted R square. So we choose the log (LSTAT) as the final variable.



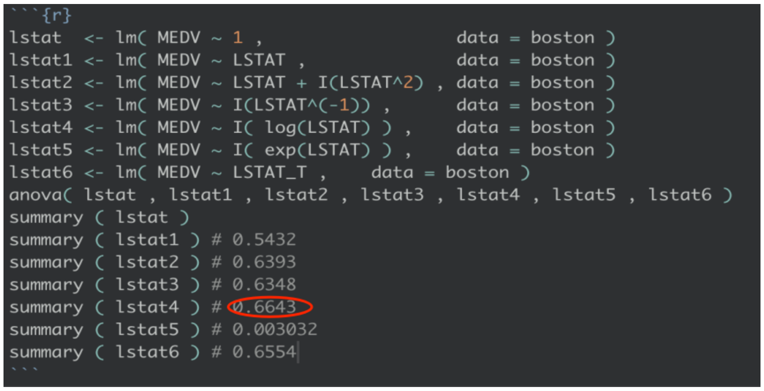
```
Start:  AIC=1562.31
MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + D + DIS +
    PIRATIO + TAX + I(exp(LSTAT))
```



```
Start:  AIC=1449.48
MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + D + DIS +
    PIRATIO + TAX + I(LSTAT^(-2))
```



```
Start:  AIC=1382.84
MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + D + DIS +
    PIRATIO + TAX + log(LSTAT)
```



~~~
```{r}
lstat  <- lm( MEDV ~ 1 ,                    data = boston )
lstat1 <- lm( MEDV ~ LSTAT ,                data = boston )
lstat2 <- lm( MEDV ~ LSTAT + I(LSTAT^2) ,   data = boston )
lstat3 <- lm( MEDV ~ I(LSTAT^(-1)) ,        data = boston )
lstat4 <- lm( MEDV ~ I( log(LSTAT) ) ,      data = boston )
lstat5 <- lm( MEDV ~ I( exp(LSTAT) ) ,      data = boston )
lstat6 <- lm( MEDV ~ LSTAT_T ,     data = boston )
anova( lstat , lstat1 , lstat2 , lstat3 , lstat4 , lstat5 , lstat6 )
summary ( lstat )
summary ( lstat1 ) # 0.5432
summary ( lstat2 ) # 0.6393
summary ( lstat3 ) # 0.6348
summary ( lstat4 ) # 0.6643
summary ( lstat5 ) # 0.003032
summary ( lstat6 ) # 0.6554
```
~~~

Figure 7: AIC & Ajusted R^2

Furthermore, we plot the scatter plot of log (LSTAT) with MEDV, and find that now they are in visually linear relationship.

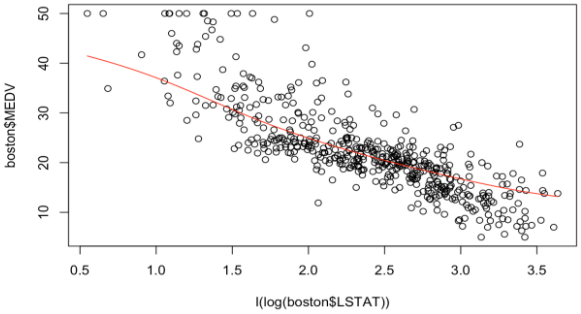


Figure 8: Logarithmic Transformation

2.4 Correlation

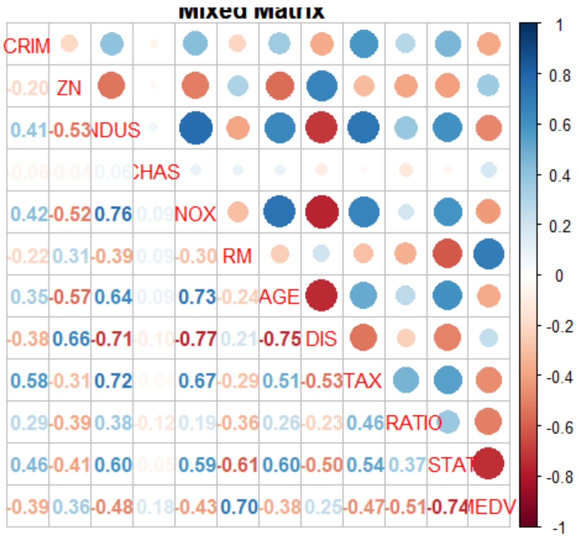


Figure 9: Thermodynamic Diagram

According to the thermodynamic diagram to find the relationship between all the variables.

2.4.1 Correlation Between X and Y

We can evaluate the different significance degree of the correlation coefficient of each independent variable. Ranking the correlation coefficient in a custom way, which is the numbers in the range of (0,0.3) are little correlated, in [0.3, 0.45) are weak related, in [0.45, 0.7) are intermediate related and in [0.7, 1] are strong related. Therefore, we conclude that there are two variables are little related, four are weak related, three are intermediate related, and also two are strong related.

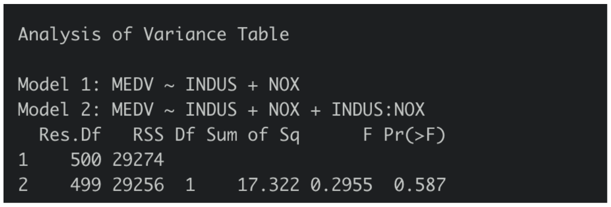
2.4.2 Correlation Between X and X and interaction terms

From the correlation matrix, the relationship between explanatory variables can also be observed. We can see that INDUS and DIS are relatively highly related with other variables, which means they may interplay with others. Therefore, they may not be good variables and there is a great chance that they will be deleted later.

Another thing we can conclude is that the some of the variables may depend on others. As a result, we try to introduce interaction terms to see whether we can improve the model. According to the matrix, we choose many possible terms, such as INDUS and NOX, INDUS and DIS, INDUS and TAX, NOX and AGE, NOX and DIS, AGE and DIS. And we pick out those terms which can be illustrated in reality. That are INDUS and NOX, INDUS and DIS, NOX and AGE, NOX and DIS.

Afterwards we use ANOVA to check the significance and to determine whether we should add these interaction terms into the model. But unfortunately, none of the interaction terms are significant. (The ANOVA tables for each interaction term are in appendix.) So in the end we do not add them in to the model.

For example, the DIS and NOX has significant correlation coefficient. NOX means nitrogen oxide concentrations (annual average concentration in parts per hundred million) and DIS represents the weighted distances to five employment centers in Boston region. NOX can reflect the polluted degree so that the closer some place is to the employment center, the larger the NOX is. According to figure 10, the p-value is 0.587 which means the interaction term INDUS:NOX is not significant. Therefore we don't add it into the model.



```
Analysis of Variance Table

Model 1: MEDV ~ INDUS + NOX
Model 2: MEDV ~ INDUS + NOX + INDUS:NOX
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    500 29274
2    499 29256  1    17.322 0.2955  0.587
```

Figure 10: ANOVA table for INDUS and NOX

Similarly, we do the analytic for the other interaction terms, the ANOVA table is shown in the appendix.

2.5 Multiple-linear Regression Model

2.5.1 Model Building

Now we have done variable analysis and determined to use multiple-linear regression model. The result is $\text{MEDV} = 54.0810 - 0.1123 \times \text{CRIM} + 0.0221 \times \text{ZN} - 0.0126 \times \text{INDUS} + 1.7756 \times \text{CHAS} - 13.7637 \times \text{NOX} + 2.9824 \times \text{RM} + 0.0125 \times \text{AGE} - 1.2311 \times \text{DIS} - 2.3562 \times \text{D} - 0.0063 \times \text{TAX} - 0.8415 \times \text{PIRATIO} - 8.0813 \times \log(\text{LSTAT})$.

2.5.2 Variable Selection

And then we do step-wise regression analysis to select significant variables and get a new model: $\text{MEDV} = 53.4471 - 0.1118 \times \text{CRIM} + 0.0215 \times \text{ZN} + 1.8084 \times \text{CHAS} - 13.0775 \times \text{NOX} + 3.0964 \times \text{RM} - 1.2807 \times \text{DIS} - 2.3802 \times \text{D} - 0.0067 \times \text{TAX} - 0.8384 \times \text{PIRATIO} - 7.8463 \times \log(\text{LSTAT})$.

The new model delete AGE and INDUS. As we assume before, the INDUS may be deleted. For AGE, we can know that this variable means the proportion of housing constructed before 1940 in this region. And the distribution shows that 50% of the data lies in the interval [77.3%,100%] and 8.1% of the AGE is 100%. Hence the distribution is too tense for the tail. We also provide a way to improve the data. If we make the boundary year 1940 earlier, and the distribution will be more disperse, the variables maybe more significant.

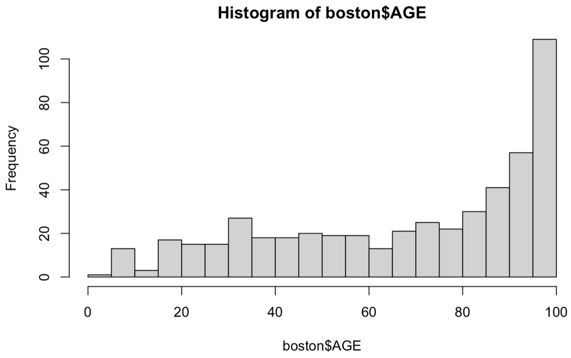


Figure 11: Distribution of AGE

3 Model Diagnosis

3.1 Assumptions Checking

Model diagnosis is aimed to ensure the assumptions for the assumptions of multiple linear regression, no outliers and no multicollinearity are satisfied. If one of them is not met, adjusting the model like transforming the variables is indispensable.

Checking four assumptions comes first. The linear hypothesis may not be valid, because the trend line in red is not basically coincided with y = 0 so there might be a trend relationship between the residual and fitted value. It is necessary to plot the partial residuals of all independent variables in the next step. The normality hypothesis is satisfied for the trend line of the scattered points is largely coincided with a straight line. Especially, the head and tail of this graph do not deviate from the line generally. The homoscedasticity assumption is met, since the trend line is essentially horizontal. We can assume the independence of errors is satisfied as the data set is cross-sectional. Checking outliers again, there is no point with a distance greater than 0.5 Cook distance.

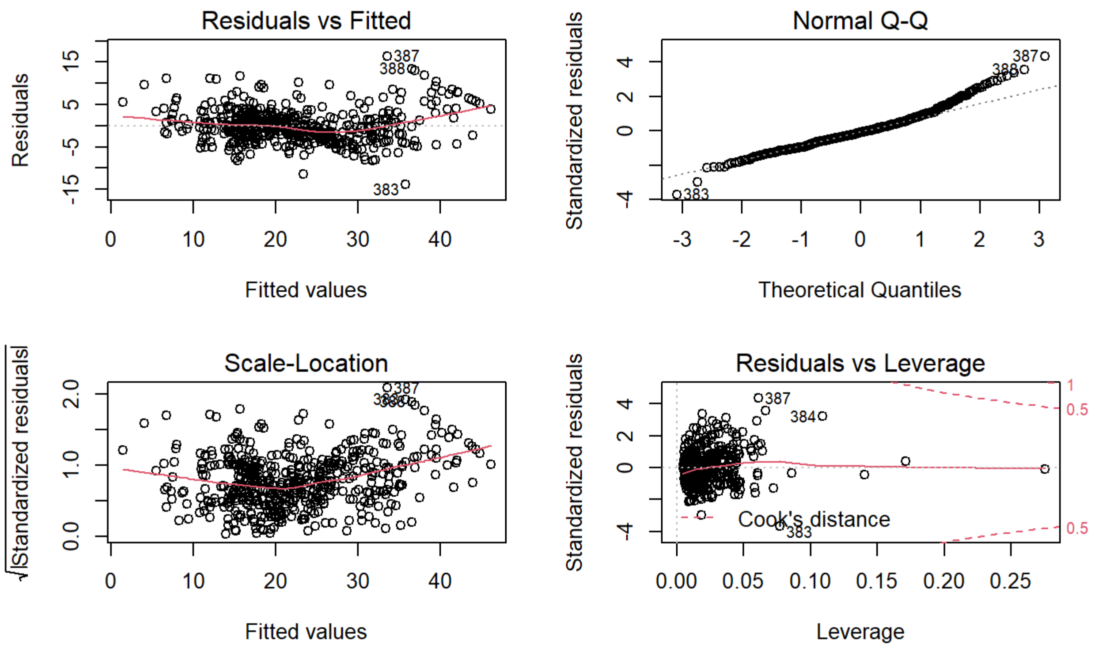
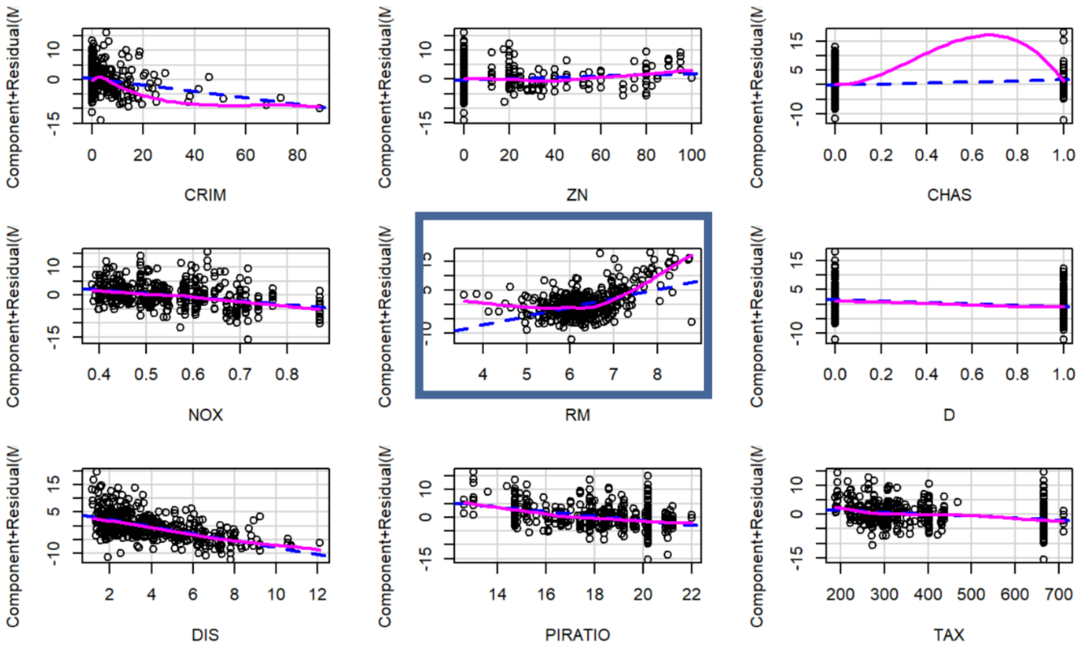
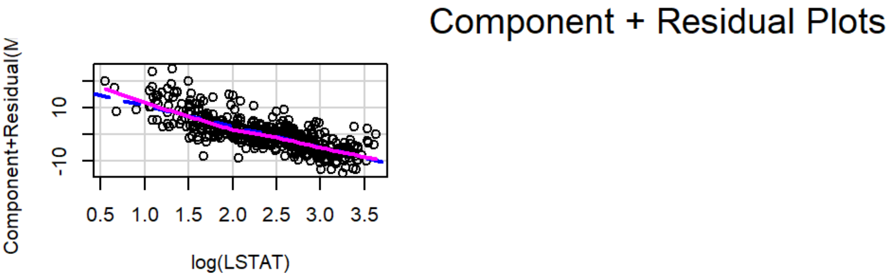


Figure 12: Residual plots

3.2 Component Residual Plots

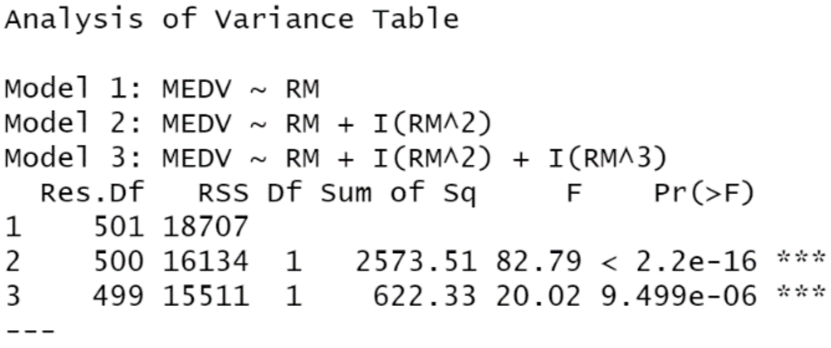
Except for the categorical variables and RM, the other independent variable is linear with the residual. From the trend line, we can find that RM is the quadratic or cubic relationship.





3.3 Compare Models using F-test

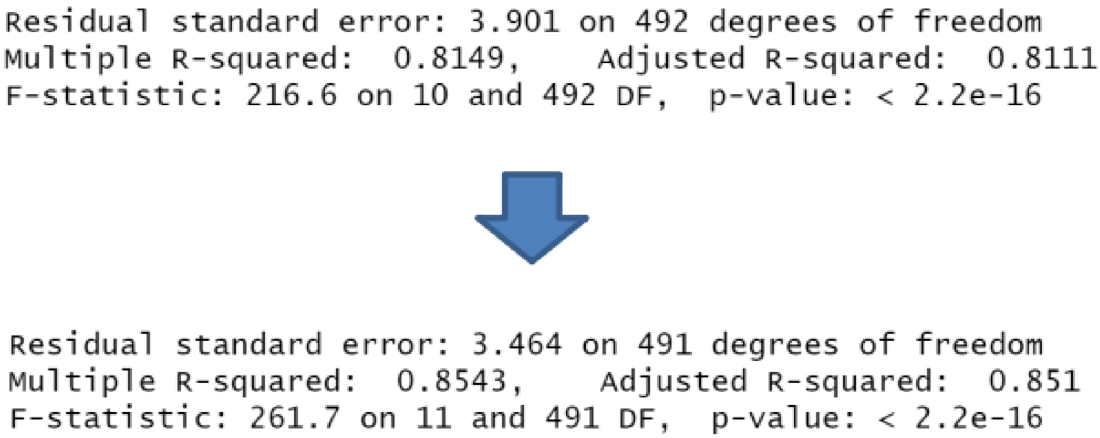
After comparing models using F-test to check which one makes the model more appropriate, we can find that the p-value of model with the quadratic term is smaller than model with the cubic term. Thus, it is better for a quadratic relationship. After adding RM^2 into the regression function, it is obvious the adjusted R-squared is higher from 0.8163 to 0.8559.



```
Analysis of Variance Table

Model 1: MEDV ~ RM
Model 2: MEDV ~ RM + I(RM^2)
Model 3: MEDV ~ RM + I(RM^2) + I(RM^3)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)    
1    501 18707                                 
2    500 16134  1   2573.51 82.79 < 2.2e-16 ***
3    499 15511  1    622.33 20.02 9.499e-06 ***
---
```

Figure 13: ANOVA table



```
Residual standard error: 3.901 on 492 degrees of freedom
Multiple R-squared:  0.8149,	Adjusted R-squared:  0.8111 
F-statistic: 216.6 on 10 and 492 DF,  p-value: < 2.2e-16
```

```
Residual standard error: 3.464 on 491 degrees of freedom
Multiple R-squared:  0.8543,	Adjusted R-squared:  0.851 
F-statistic: 261.7 on 11 and 491 DF,  p-value: < 2.2e-16
```

3.4 Linearity Revalidating

When revalidating the linearity, the trend line is more horizontal from the graph of Residuals vs Fitted and the relationships between RM and residual and RM^2 and residual are more linear from the graph of partial residual.

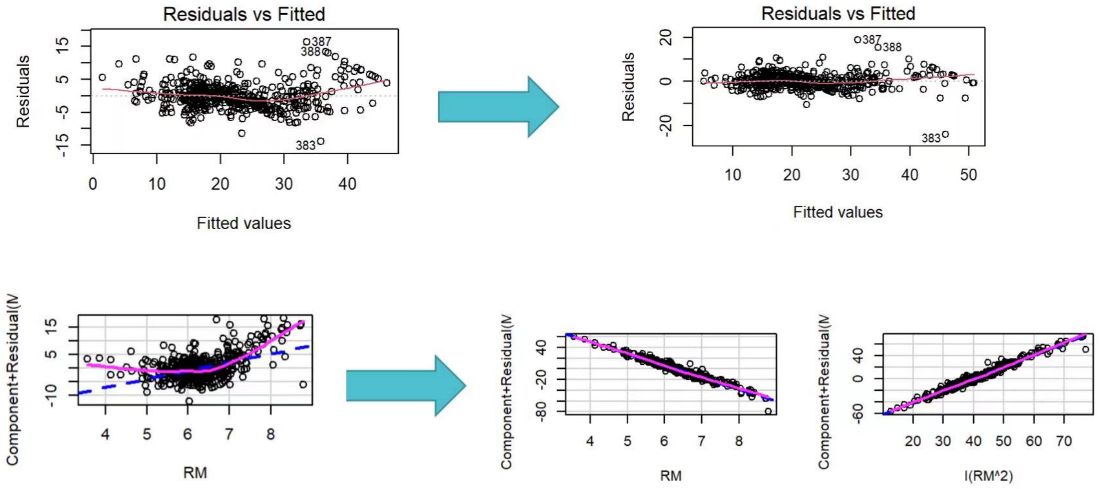
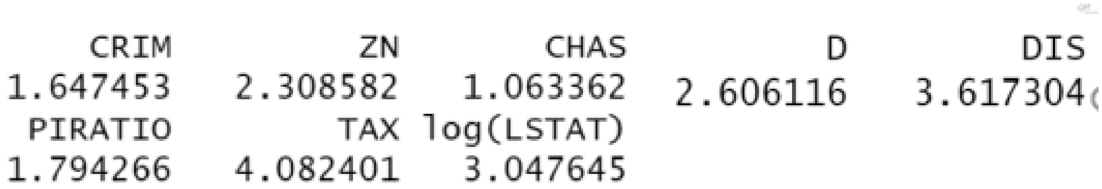


Figure 14: partial residual plot

3.5 Variance Inflation Factor

According to variance inflation factor (VIF), except for RM and RM^2, the VIFs of other independent variables are less than 10, which means they do not have multicollinearity with other variables.



```
    CRIM        ZN      CHAS         D        DIS
1.647453  2.308582  1.063362   2.606116   3.617304
 PIRATIO       TAX log(LSTAT)
1.794266  4.082401  3.047645
```

Figure 15: VIF

Eventually, after diagnosis and transformation, we obtain the final model as follows: MEDV $= 125.5857 - 0.1316 \times \text{CRIM} + 0.0118 \times \text{ZN} + 1.6790 \times \text{CHAS} - 13.8964 \times \text{NOX} - 0.9795 \times \text{DIS} - 1.3879 \times \text{D} - 0.0050 \times \text{TAX} - 0.6900 \times \text{PIRATIO} - 6.9594 \times \log(\text{LSTAT}) - 22.2997 \times \text{RM} + 2.0187 \times \text{RM}^2$.

Conditional Inference Tree

Also we construct the conditional inference tree.

First thing we need to do is data partion. In this step, we randomly select 70% of the data as the training set and 30% as testing set.

Second, we need to do the data classification since we need to transform the numerical data MEDV into categorical data in 4 different catelogs A, B, C and D according to quantiles. We take quantiles as the boundaries and divide data into 4 parts.

Quantile	MEDV	Classification
25%	[0 , 16.8)	A
50%	[16.8 , 21.1)	B
75%	[21.1 , 25)	C
100%	[25 , 50]	D

Table 1: Classification

And then we construct the tree.

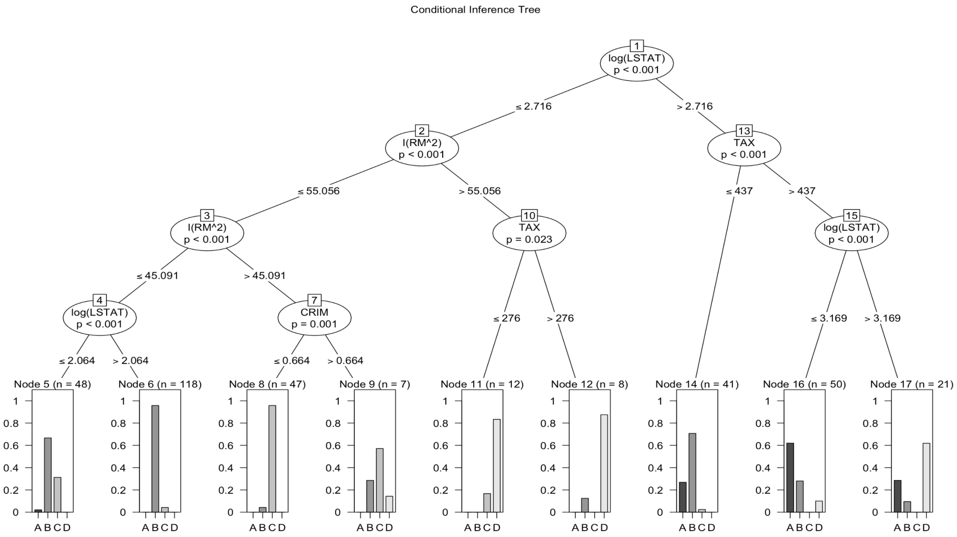


Figure 16: Conditional inference tree

Different from regression model using 10 variables, conditional inference tree only takes 4 variables into account. They are log(LSTAT), RM^2, TAX and CRIM. The 4 significant variables send the result to different branches and finally reach one of the 9 results.

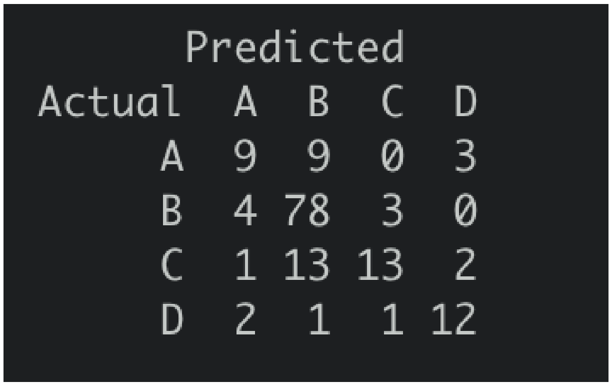
For example, if we got a new house with information in the table below.

log(LSTAT)	2.93
RM2	51.3
TAX	432
CRIM	0.54011

Following the path, the result finally reach node 8 which means there are around 95% possibility that the MEDV is in catelog C. That is, there is a 95% likelihood that the price is between 21.2 and 25.

And then we get the accuracy matrix from which we can know that

$$\text{accuraterate} = \frac{9 + 78 + 13 + 12}{506 \times 30\%} = 74.172\%$$



```
       Predicted
Actual  A  B  C  D
     A  9  9  0  3
     B  4 78  3  0
     C  1 13 13  2
     D  2  1  1 12
```

Figure 17:

The tree and the multiple-linear regression model have different advantages and disadvantages. The conditional inference tree only take 4 variables and don't need any calculation while regression model needs to calculate 10 variables. However, the tree can merely give a possibility interval instead of a specific number offered by regression. To summary, the tree is fast but inaccurate while regression is slower but accurate.

When it comes to practical application, if the real estate agency need to response quickly, they can choose the Conditional Inference Tree. If they have enough time or carrying a computer or calculator around, they can use the regression model to give a specific offer price.

Business Suggestions

With this model, the real estate agency can offer an efficient price. If the price is too high, customers wouldn't buy it and go to the rival agencies instead. There will be no profits.

On the other hands, if the price is too low, even though it is easy to sell. The agency would lose the potential profits that they could have.

Also they can use the model to adjust their price in time. For example, we can tell from the model that people do not like to live around highway, if you secretly know that a highway will be constructed near your to-be-sale house, you should reduce the price and sell it right away before the construction. Otherwise you may reduce the price much more later with worse revenue.

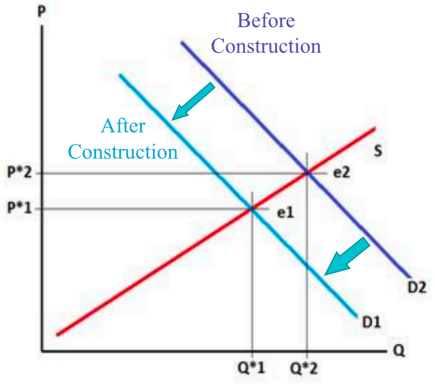


Figure 18: Predicted result

Limitations

a. The categorical variable(RAD) is processed by R as the continuous variable. It means that some undesired non-integer like decimals or the integers which are not given in the original dataset. The results of the regression may not be very precise and easy to be explained eventually.

b. Random forest aiming at high dimensional data is more proper if we want deal with dataset with many features. Moreover, it has the advantage of fast training and is convenient to make into a parallelization method (trees are independent of each other during training). Compared with the multiple linear regression, it is more straightforward to realize, more unbiased and can be applied flexibly.

However, for the small dataset(our original dataset), the linear regression is likely to be more precise than random forest.

Improvement

a. Aim at the categorical variable CHAS, we can use as.factor to process it in the further study. This code can make categories just be expressed to 0 and 1, which can avoid the data is represented as a number between 0 and 1.

b. About the house price prediction part, in the further study about data analysis, we think we can try the random forest method to predict if the number of data is enough large. This method has a lot of advantages rather than conditional inference tree.

· It constructs many trees to prevent the risk of overfitting problem.

· Random forest can provide an effective method to balance data set errors when there is classification imbalance.

· In the training process, the interaction between features can be detected.

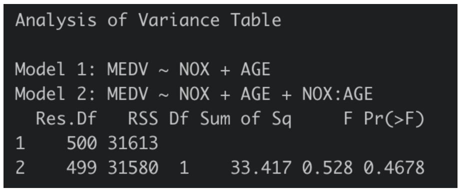
· It can process data with high dimensions(many features) without feature selection because feature subsets are randomly selected. While conditional inference tree just select several variables as the significant variables, it is not rigorous.

References

[1] Harrison, D. & Rubinfeld, D.L.Hedonic prices and the demand for clean air. $Economics\&Management$, 81-102.https://www.heywhale.com/mw/dataset/590bd595812ede32b73f55f2

Appendices

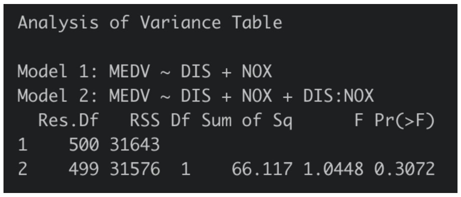
A ANOVA table for INDUS and AGE



```
Analysis of Variance Table

Model 1: MEDV ~ NOX + AGE
Model 2: MEDV ~ NOX + AGE + NOX:AGE
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1    500 31613
2    499 31580  1    33.417 0.528 0.4678
```

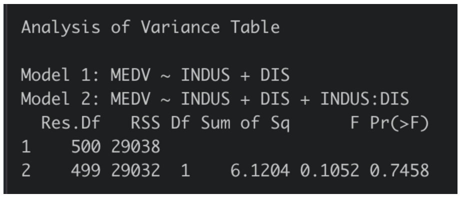
B ANOVA table for NOX and NOX



```
Analysis of Variance Table

Model 1: MEDV ~ DIS + NOX
Model 2: MEDV ~ DIS + NOX + DIS:NOX
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    500 31643
2    499 31576  1    66.117 1.0448 0.3072
```

C ANOVA table for DIS and DIS



```
Analysis of Variance Table

Model 1: MEDV ~ INDUS + DIS
Model 2: MEDV ~ INDUS + DIS + INDUS:DIS
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    500 29038
2    499 29032  1    6.1204 0.1052 0.7458
```