Regression Analysis on The Median Value of Owner-Occupied Homes

YICHAO, IVAN, DAI

Wenzhou-Kean University, ID 1098325, MATH 3700 W02

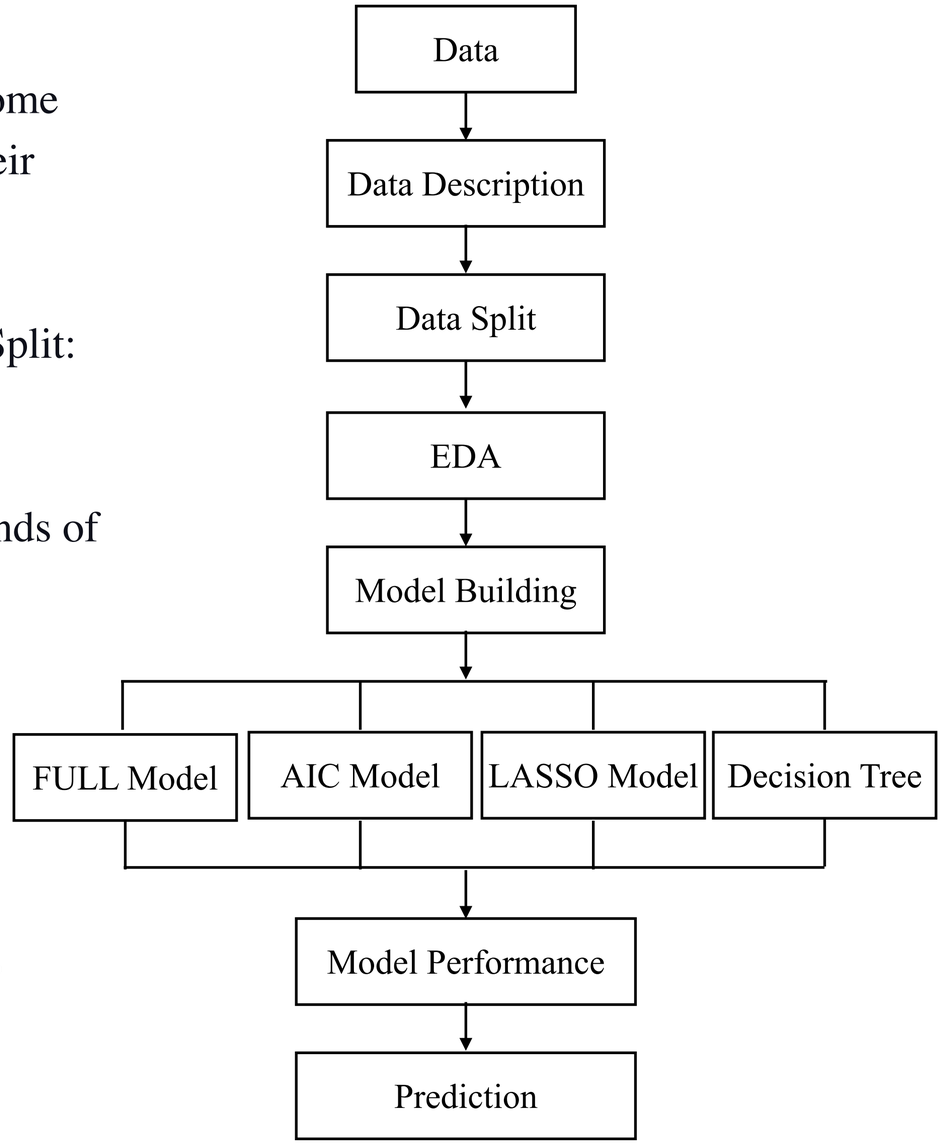
ABSTRACT

The report uses the housing prices database. In this report, owner-occupied homes' median value (in thousands of dollars) are predicted using several models. Firstly, report will start with the basic dataset description to describe the data type, dataset dimension, and descriptive statistics of each variable. The report will then split the data set into the train set and test set, use the train set to do the exploratory data analysis, and build several models. The report will tart with the OLS FULL Model, and then more advanced model including AIC backward Selection model, LASSO model, and the decision tree model based on the Classification and Regression Tree (CART) algorithm to predict the homes' median value. Finally, the report will use the root mean squared root (RMSE) and R square to evaluate model performance and choose the best model to predict on the test set. The final result shows that the Decision Tree model has a relatively good model performance on the train set with RMSE equal to 3.119253 and R square equal to 0.8000974.

1 INTRODUCTION

House price prediction is one of the hot topics in recent years. People try to find some relevant explanatory variables to predict the house price in the future and make their own decisions whether to buy or sell the house right now. The report structure is organized as a flow chart showing in the figure.

1. Data Description: Variables description, data type, descriptive statistics. Data Split: Train set and test set
2. Exploratory Data Analysis on the train set: Explore the relationship between explanatory variables and the owner-occupied homes' median value (in thousands of dollars) to see whether there is a significant difference or trend.
3. Model Building Based on the Train Set:
   - OLS Full Model
   - AIC Backward Selection Model
   - LASSO Model
   - Decision Tree (CART)
4. Model Performance: R square (measure the how much the model can explain), RMSE (measure how the accuracy the model is)
5. Prediction on Test Set



2 VARIABLES DESCRIPTION

The explanatory variables are mainly talking about some social or environmental factors, including the per capita crime rate, percent of the low-status population, index of accessibility, and other different variables. Here is a table showing a brief description of variables.

Variables	Data Type	Description
medv	Numeric	the median value of owner-occupied homes (in thousands of dollars)
rm	Numeric	average number of rooms per dwelling
age	Numeric	proportion of owner-occupied units built prior to 1940
crim	Numeric	per capita crime rate by town/suburbs
zn	Numeric	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	Numeric	proportion of non-retail business acres per town
chas	Numeric	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
tax	Numeric	cost of public services in each community
ptratio	Numeric	pupil-teacher ratio by town
black	Numeric	variable $1000(Bk - 0.63)^2$, where Bk is the proportion of black population
lstat	Numeric	percent of lower status of the population.
dis	Numeric	weighted mean of distances to five Boston employment centers
rad	Numeric	index of accessibility to radial highways
nox	Numeric	the annual concentration of nitrogen oxide (in parts per ten million)

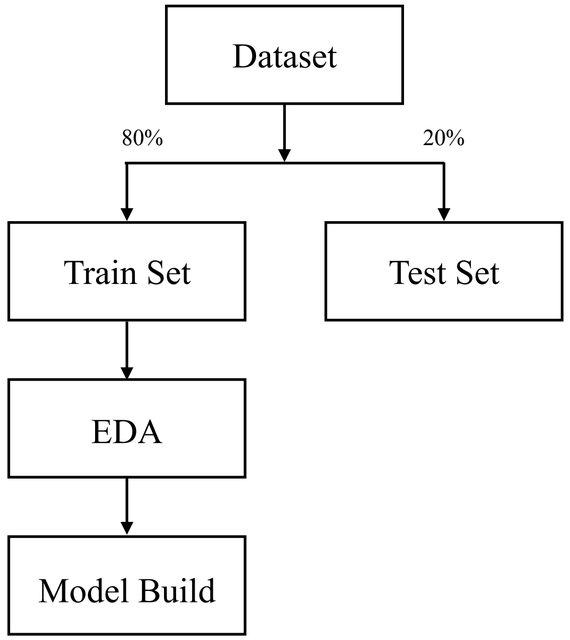
3 DESCRIPTIVE STATISTIS

The entire data set has 506 observations and 14 variables. All the variables are numeric. The following table shows the descriptive statistic of the variables, including the 5-point statistics and means value. Although all the variables are numeric, however, variables rad and chas are more like factor variables. Moreover, all 506 observations do not contain any missing values.

Variables	Min	Q1	Median	Mean	Q3	Max
medv	5	17.02	21.2	22.53	25	50
rm	3.561	5.886	6.208	6.285	6.623	8.78
age	2.9	45.02	77.5	68.57	94.08	100
crim	0.00632	0.08204	0.25651	3.61352	3.67708	88.9762
zn	0	0	0	11.36	12.5	100
indus	0.46	5.19	9.69	11.14	18.1	27.74
chas	0	0	0	0.06917	0	1
tax	187	279	330	408.2	666	711
ptratio	12.6	17.4	19.05	18.46	20.2	22
black	0.32	375.38	391.44	356.67	396.23	396.9
lstat	1.73	6.95	11.36	12.65	16.95	37.97
dis	1.13	2.1	3.207	3.795	5.188	12.127
rad	1	4	5	9.549	24	24
nox	0.385	0.449	0.538	0.5547	0.624	0.871

4 SPLIT DATA SET

In this section, the report will split the whole data set into the training set and test set. 80% of the observation would go to the train set, and 20% of the observations would go to the test set. The result of section Exploratory Data Analysis and Model Building are all based on the train set. For the test set, we would use it to compare different model performance. As a result, there are 407 observations in the training set and 99 observations in the test set.



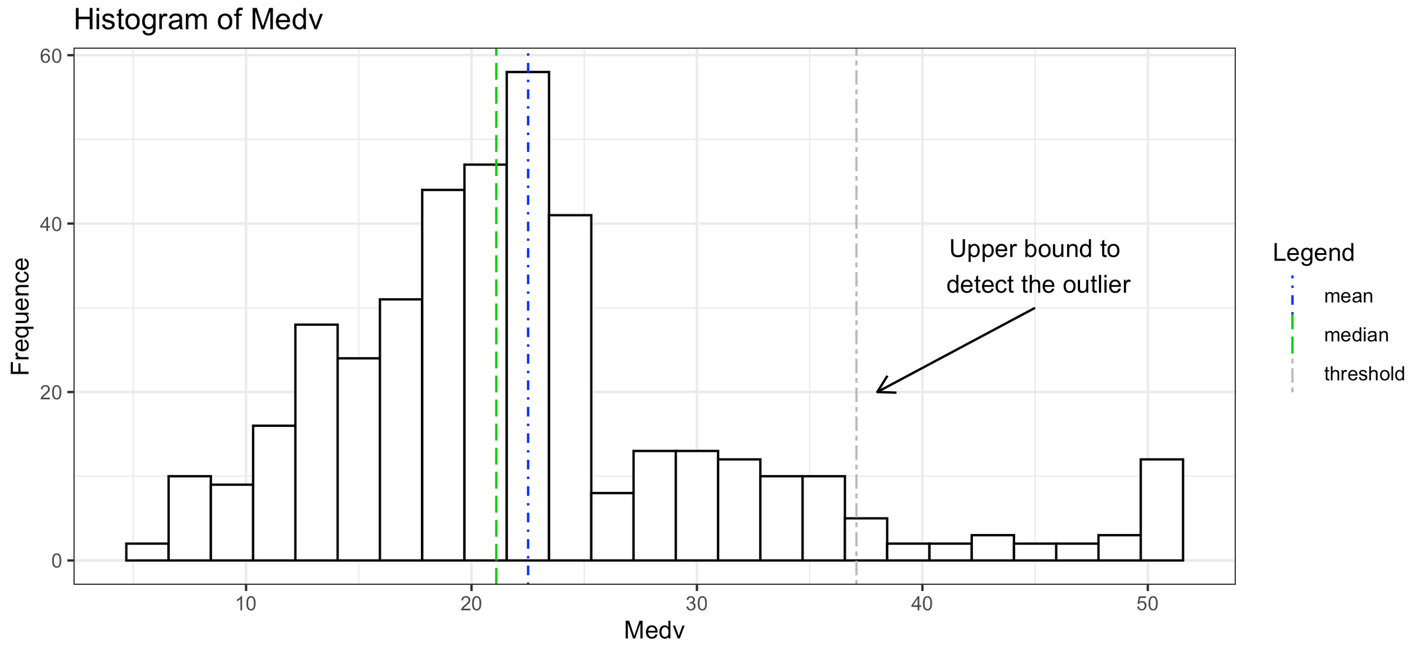
Set	Observations	Number of Variables
Train Set	497	14
Test Set	99	14

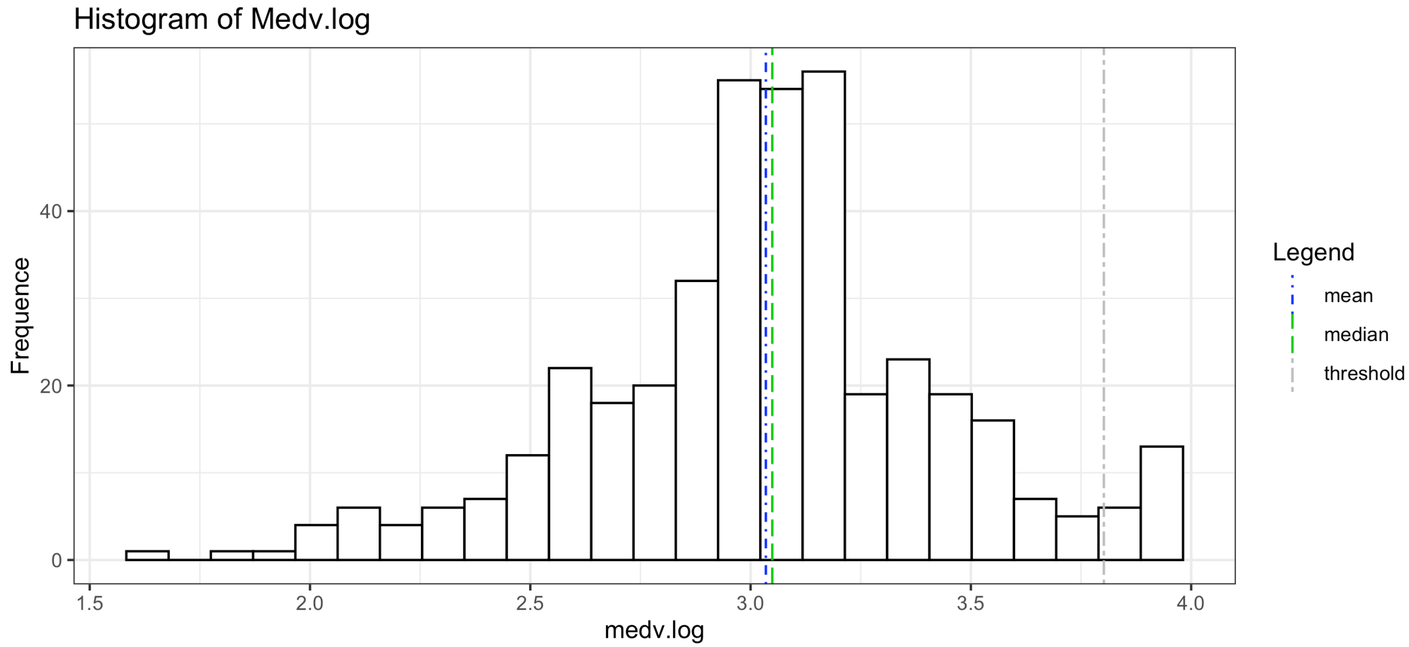
5 EXPLORATORY DATA ANALYSIS

In this section, the report aims to explore some relationships between the explanatory variable and response variable and determine whether there are some needs to do the data transformation or some adjustment to the initial model. The section mainly cover 7 parts:

- The histogram of the median value of house price, determine its skewness and other statistics.
- Scatter plot between the median value of house price and the proportion of low-status population, scaled by per capital crime rate.
- Scatter plot between the median value of house price and weighted mean of distances to five Boston employment centers, scaled by the index of accessibility.
- Box plot of median value of house price against whether there is river limitation.
- Box plot and density distribution of median value of house price against the index of accessibility.
- Scatter plot between the median value of house price and average number of rooms per dwelling, scaled by proportion of owner-occupied units built prior to 1940.
- Correlation matrix among the explanatory variables.

5.1 Histogram of median value of house price



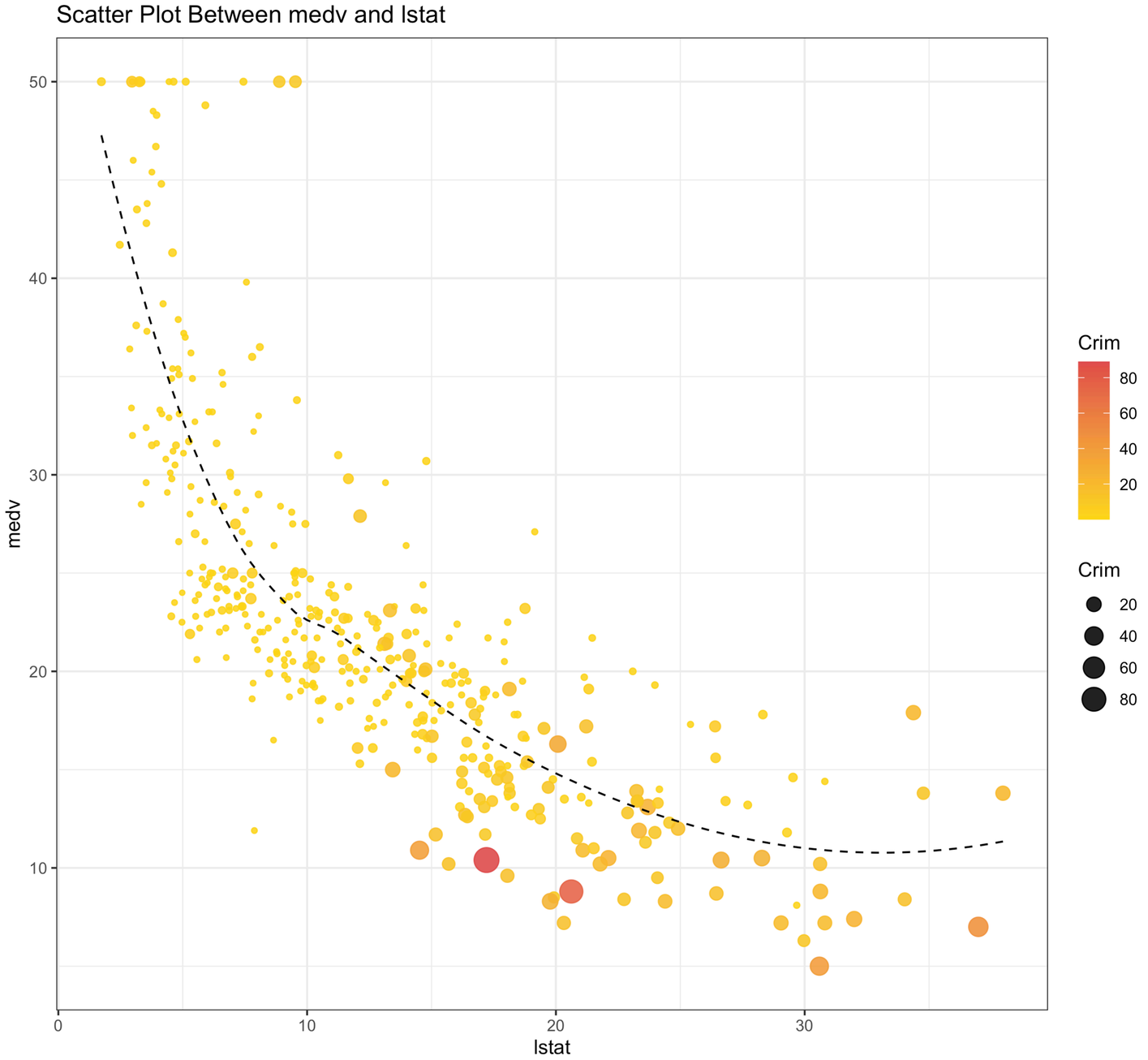


The histogram of the median value of house price (medv) is showing on the top left. We can see that the mean value is slightly higher than the median value, which indicates that the distribution of medv is right skew. The grey line shows the x-intercept, calculated by the third quantile plus the 1.5 times IQR, a threshold value to detect the upper outlier. We can see that there is an extensive range of medv outside the threshold, which indicates that the response variable is extremely right skew. The extremely right-skew response variable may increase the model's error and break the rule of normality of error for the Ordinary Least Square method (OLS). As a result, it is better to do a logarithm transformation to get a more normalized response variable.

	medv	**medv.log**
Skewness	1.106148	0.1008842

The table shows the skewness of the response variable before and after the logarithm transformation. If skewness is higher than 1 or lower than -1, the distribution is highly skewed. We can see the value of skewness changing from really high to approximately 0. Also, we can see from the bottom figure that the median and mean values are almost the same, which suggests a more normal distribution.

5.2 Scatter plot between the medv and lstat, scaled by crim



The scatter plot shows the relationship between the medv and the proportion of the low-status population. We can see a negative relationship between these two variables. The linear relationship seems to be relatively strong, with a correlation coefficient equal to -0.7430133.

Variables	Coef	Pr(>|t|)	Sig.
(Intercept)	34.6527	<2e-16	***
lstat	-0.9561	<2e-16	***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1,*
Multiple R-squared: 0.5521, Adjusted R-squared: 0.551

The table shows the simple regression between the medv and lstat. Medv would decrease by 0.96 when there is one unit increasing in lstat. By looking at the p-value of coefficients of lstat, there is strong evidence against the claim that lstat does not affect the mean level (intercept) of medv, and R square is equal to 0.55, which suggest lstat can explain 55% of the variation in medv.

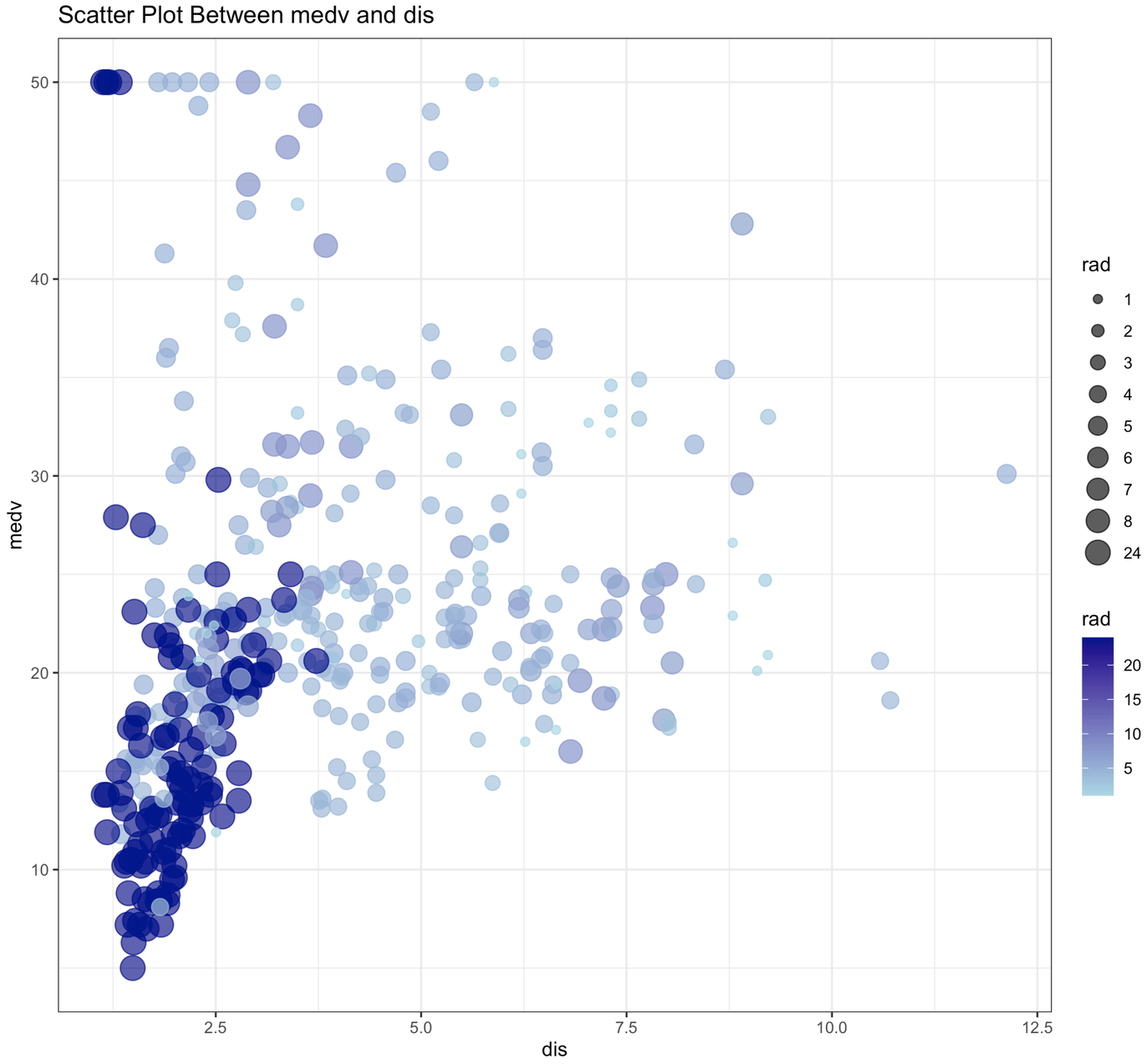
Moreover, we can also observe the non-linear relationship; the dotted line shows the polynomial relationship.

The value of crim determines the point size and the depth of color. We can see the more prominent and darker point is gathering at the right bottom of the figure, representing the area with lower medv and a higher proportion of the low-status population. The spots are also getting bigger and darker along the regression line, which suggests a negative relationship between the medv and crim and a positive relationship between lstat and crim. By using a t-test to inference whether there is a significant difference between crim < 50 and crime > 50, we get the following result:

	Test statistics	P-value
$\boldsymbol{H_0}: \mu_{>50} - \mu_{>50} = \mathbf{0}$ $\boldsymbol{H_a}: \mu_{>50} - \mu_{>50} \neq \mathbf{0}$	14.102	0.0083

As a result, we have enough evidence to reject the null hypothesis and state there is a significant difference in medv between crim < 50 and crime > 50

5.3 Scatter plot between the medv and dis, scaled by rad

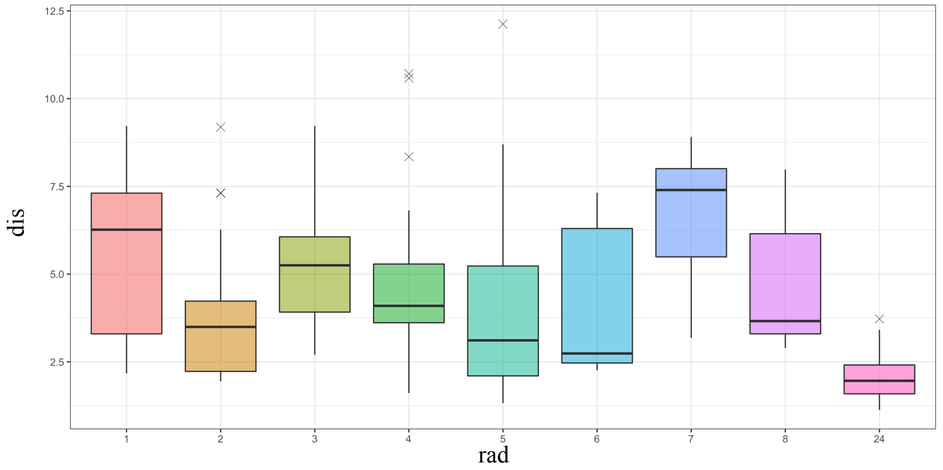


The scatter plot shows the relationship between the medv and the weighted mean of distances to five Boston employment centers. The plot shows a relatively weak positive relationship, with a correlation coefficient equal to 0.2595908.

Variables	Coef	Pr(>|t|)	Sig.
(Intercept)	18.1972	<2e-16	***
dis	1.1425	1.08e-07	***

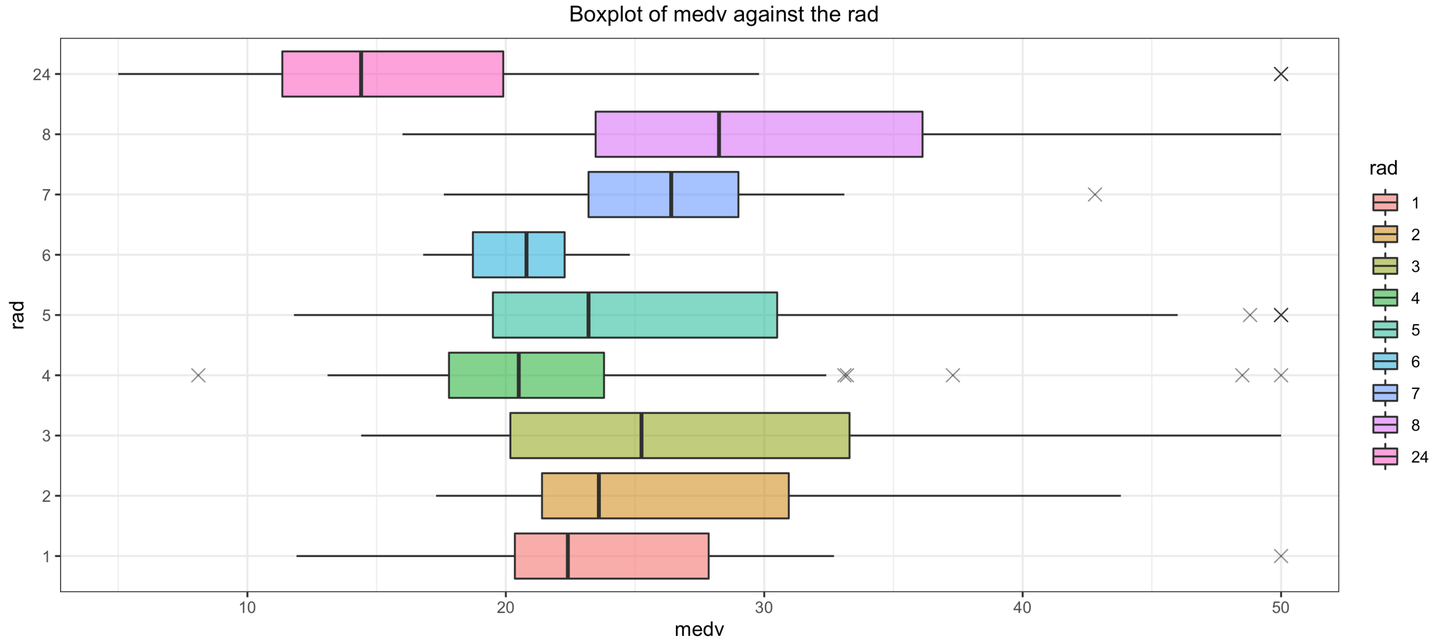
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1,*
Multiple R-squared: 0.06739, Adjusted R-squared: 0.06508

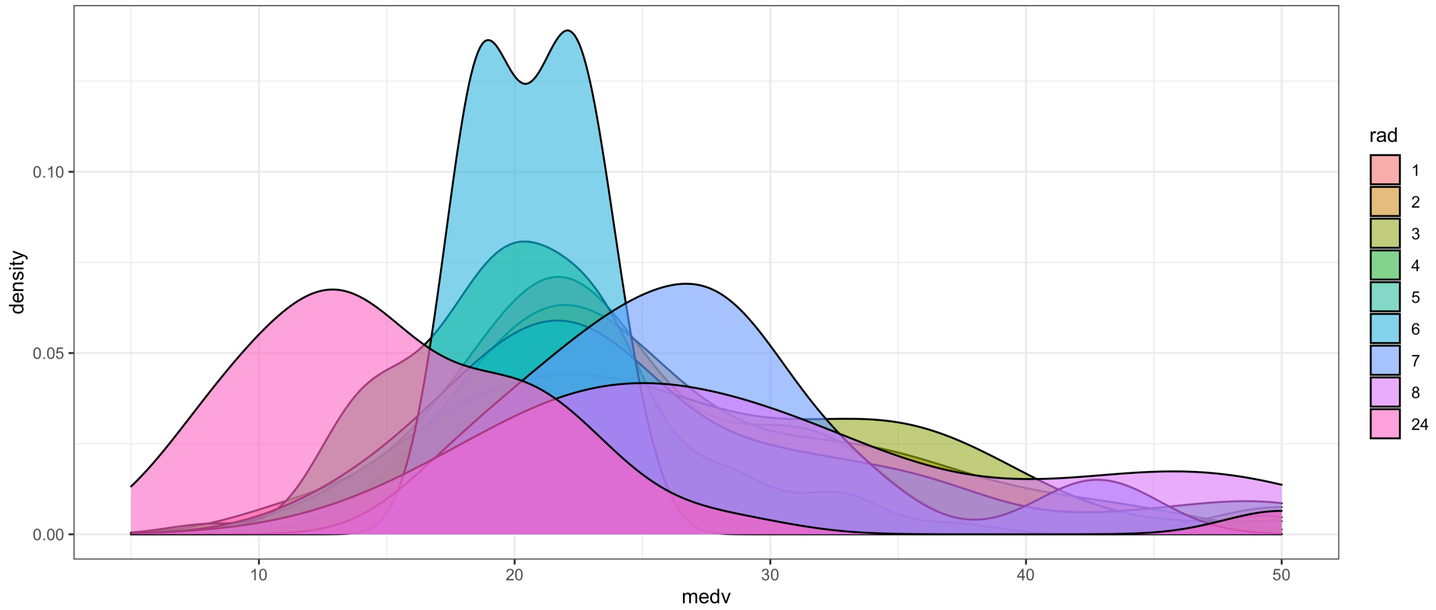
The table shows the simple regression output. For each additional unit increase in dis, the average increase in medv is 1.14. Although the correlation is not strong, the p-value of dis coefficient provides strong evidence to state dis has a significant effect on medv. Also, R square equal to 0.067 suggests dis can explain 6.7% of the variation in medv. Moreover, The point size is scaled by the index of accessibility to radial highway, and the better accessibility, the darker the color is. From the box plot below, we can see that the distribution of dis with the best accessibility (rad = 24) is relatively small. However, the corresponding medv fluctuates.



Most of the medv with rad = 24 are below 30 (in thousand dollars), but there are several points equal to 50. We want to visualize the distribution of medv with different rad by using a box plot and density distribution.

5.3.1 Box plot of medv against different rad and density distribution





The box plot and density plot of medv against different rad is showing in the figure. The medv with rad = 24 has the smallest median; however, the medv with second largest rad = 8 has the largest median, which may indicate that the medv has a positive relationship with the rad in a specific range of rad; however, when rad is too large, medv may have a negative relation with the rad. We can also see that the medv with rad = 6 has the most narrow distribution. By using one-way ANOVA, we can use inference whether medv with different rad has a significant difference.

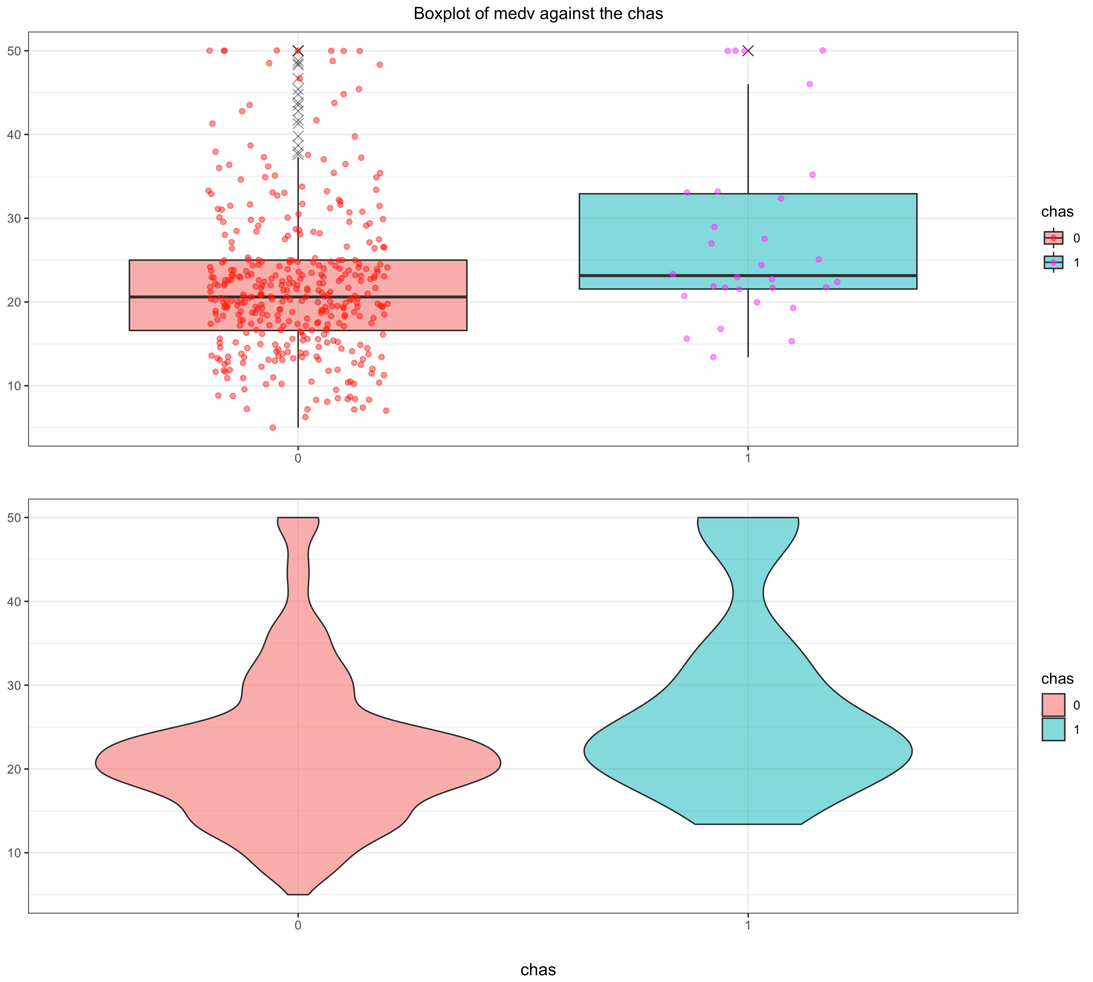
Null hypothesis: there is no difference in medv among different rad levels.

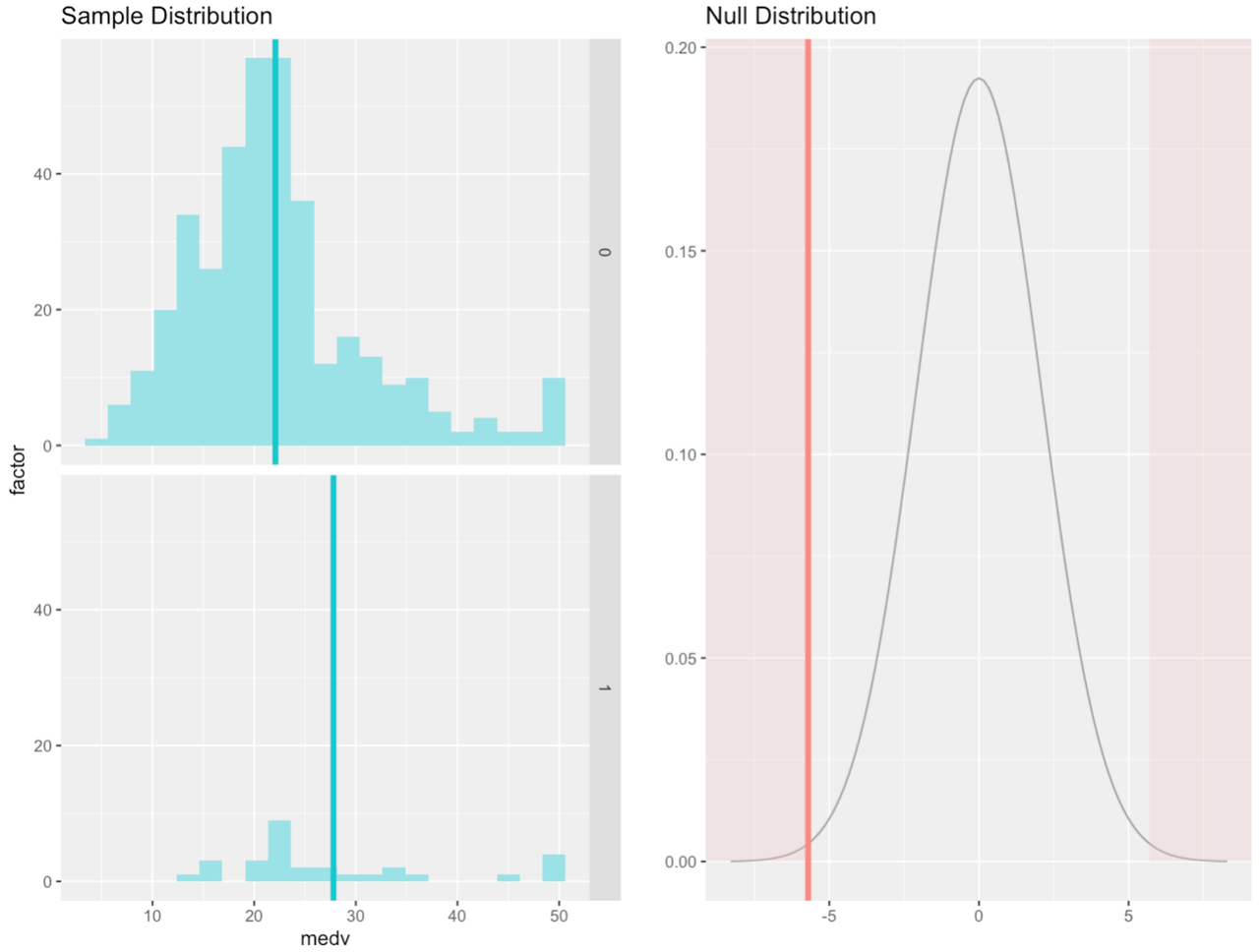
Alternative hypothesis: there is at least one pair of medv that are significantly different due to the different rad levels.

	Df	SSE	MSE	F value	Pr(>F)	
rad	8	8031	1003.9	15.31	<2e-16	***
Residuals	398	26094	65.6			

From the table above, we can see the p-value is almost 0, which indicate there is enough evidence that to reject the null hypothesis and state at least one pair of medv are significantly different due to different rad level.

5.4 Box plot of medv against chas



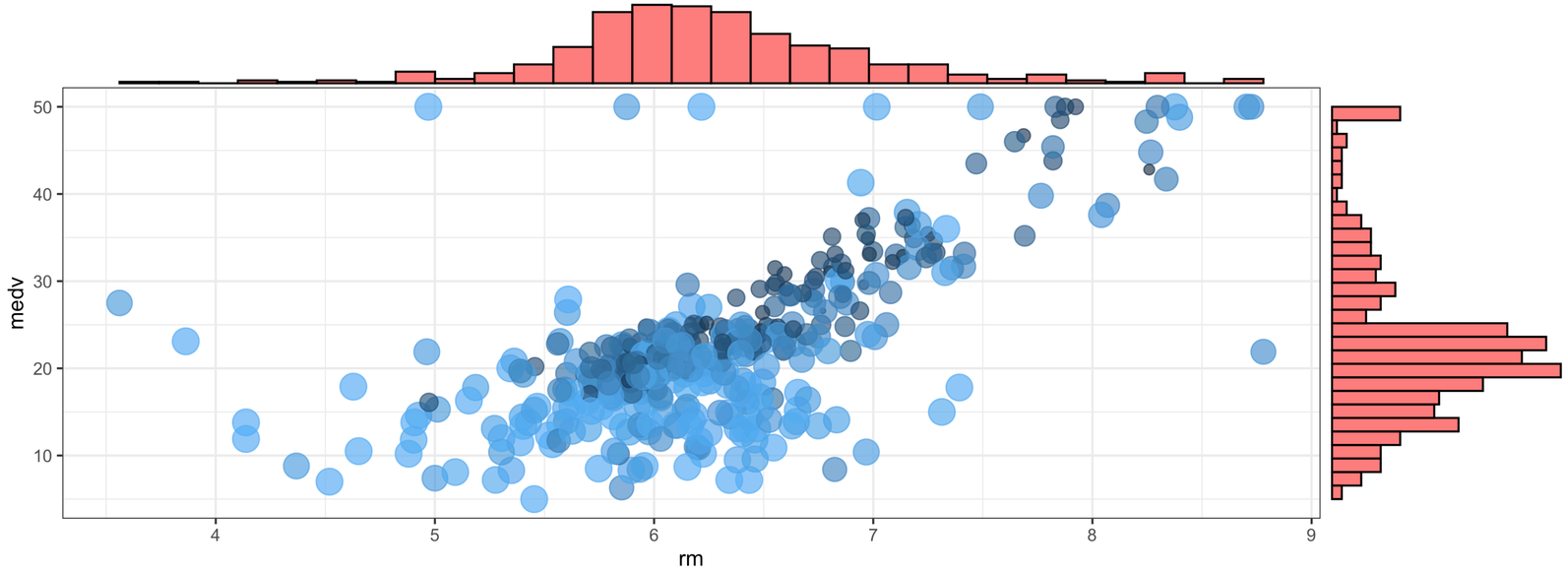


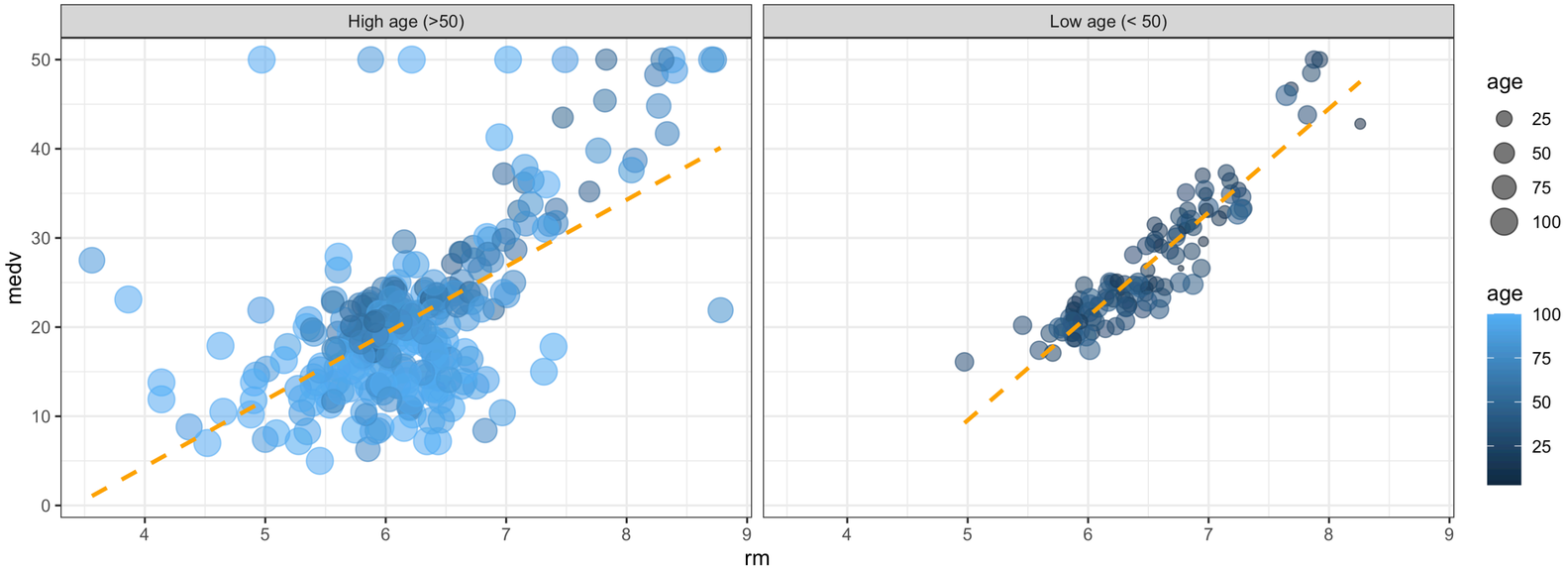
The left figure shows the box plot and violin plot of the medv against whether there is a river limitation (chas). First, we can see more observations have river bounds; however, for both types, medv are gathering around 20 (in thousands of dollars). We want to use the t-test to check whether there is a significant difference if there is a river limitation.

Because the P-value is smaller than the significant level (5%); as a result, we should have enough evidence to reject the null and state there is a significant difference.

Test statistics	P-value
-2.7523	0.0101

5.5 Scatter Plot between medv and rm, scaled by age





The figure shows the scatter plot between medv and the average number of rooms per dwelling (rm), scaled by the proportion of units built before 1940 (age). The overall relationship is positive; however, when rm < 4.5, the scatter plot shows a strong negative relationship between the rm and the medv where the correlation coefficient is equal to -0.978819. However, when rm > 4.5, there is a strong and positive relationship between rm and medv where the correlation coefficient is equal to 0.7154657.

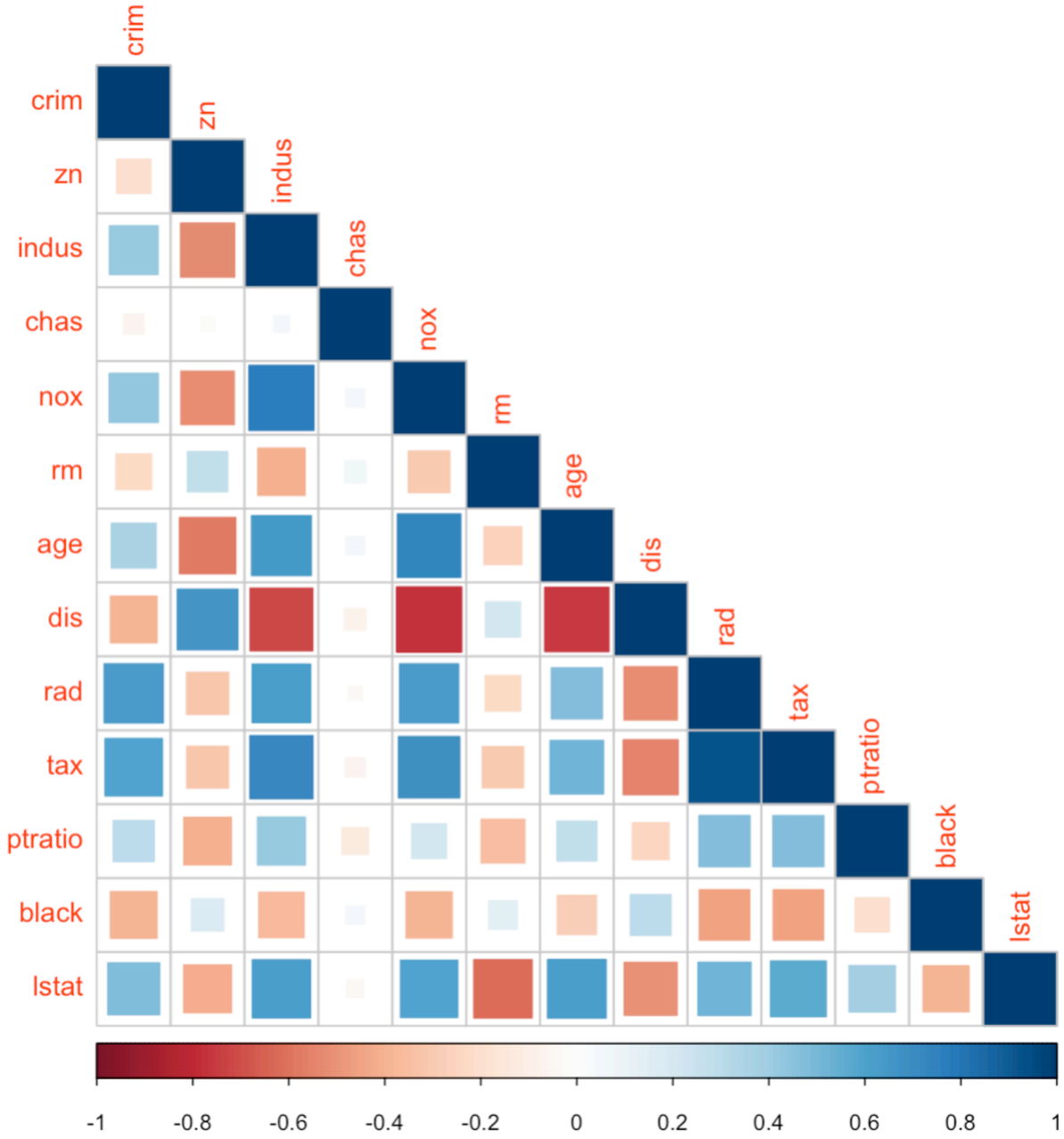
Variables	Coef	Pr(>|t|)	Sig.
(Intercept)	-31.6283	<2e-16	***
rm	8.6190	<2e-16	***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1,*
Multiple R-squared: 0.4706, Adjusted R-squared: 0.4693

The table shows the simple regression result between medv and rm. For each additional unit increase in rm, the average increase in medv is 8.62. P-value also provides strong evidence for their relationship. Moreover, R square equal to 0.47, which means rm can explain 47% variation in medv.

Moreover, we can see that the linear relationship is more evident for the house age < 50. The calculated correlation coefficient between rm and medv is 0.9233481 for the observation age < 50, however, the calculated correlation coefficient between rm and medv is only 0.6185904 for the observation where age > 50

5.6 Correlation Matrix



The figure shows the correlation matrix between each pair of explanatory variables. The area occupied by the colored square in the white square reflects the strength of the correlation. We can see that some variables are highly correlated. We want to avoid the multicollinearity problem in the Ordinary Least Square (OLS) multi-regression model. This section will use the variance inflation factor (VIF) to see whether there is a severe multicollinearity problem by building the full model, which uses all the explanatory variables to predict the medv. The following table shows the VIF for each variable:

Variables	**VIF**	**Variables**	**VIF**
rm	1.911905	ptratio	1.851645
age	3.213786	black	1.364539
crim	1.844585	lstat	3.131835
zn	2.323334	dis	4.065420
indus	3.956419	rad	8.152643
chas	1.066866	nox	4.447852
tax	9.660128		

tax and rad have VIF greater than 5, which indicates serious multicollinearity problems

6 MODEL BUILDING

This section aims to build four different models, starting with OLS full model and then including AIC backward selection model, LASSO model, and Decision Tree based on the CART algorithm. In the previous section, we have found that medv is a highly right-skewed variable; thus, we transform it by applying a natural logarithm. As a result, the response variable for all the models is log(medv). The report would calculate its R square and root mean squared error for each model to show the model performance.

- OLS full model
- AIC backward selection model, avoid overfitting, eliminate multicollinearity problem
- LASSO model, shrinkage method, avoid overfitting, eliminate multicollinearity problem
- Decision Tree based on CART algorithm

One of the performance formula is showing as following, where $\bar{y}$ bar is the predicted value by using the model, and y_i is the observed value in the dataset.

$$RMSE = \sqrt{\sum \frac{(\bar{y}_i - y_i)^2}{n}}$$

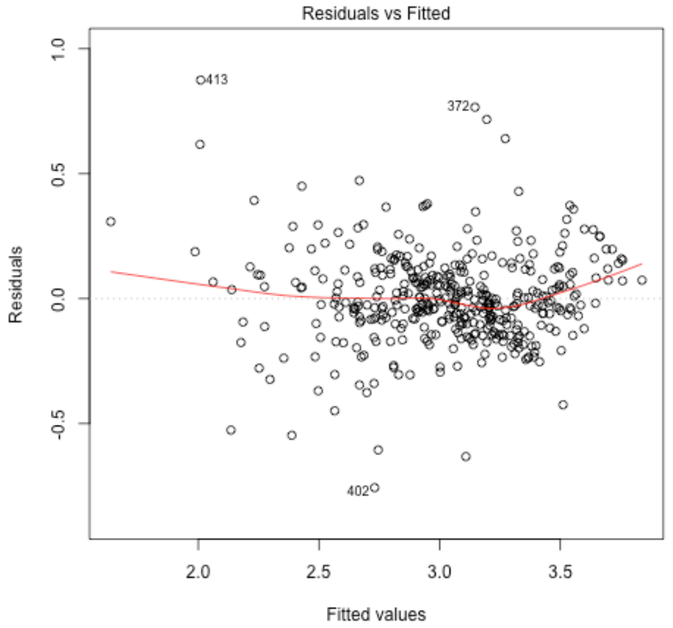
Because the response variable is medv.log, as a result, for every model, we should transform the predicted result by using exponential, and then calculate the R square and the RMSE.

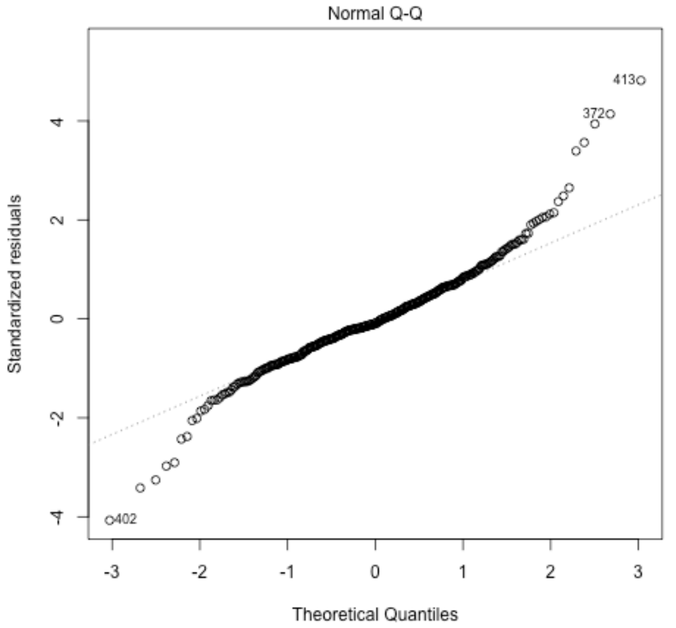
6.1 OLS FULL Model

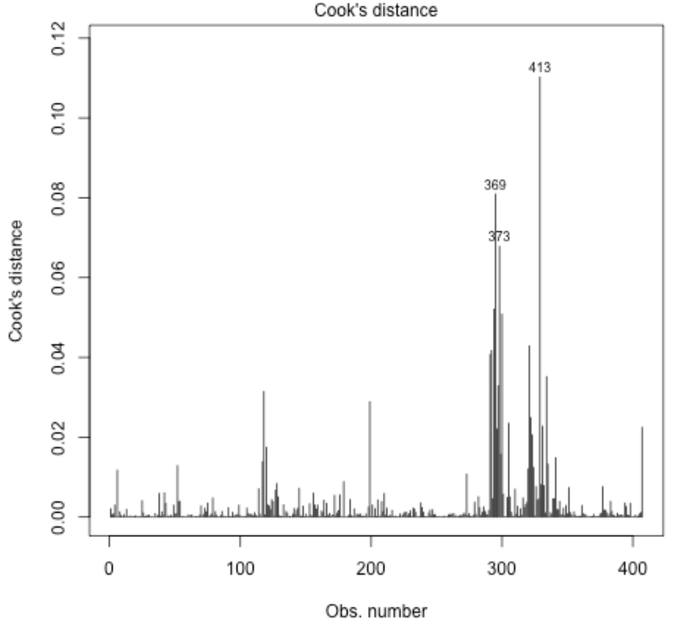
The table showing on the right shows the OLS full model. The OLS multi-regression should first follow the three main assumptions: normality of errors, correlation and multicollinearity, and homoscedasticity. We can see the residuals are approximately normal distribution through the diagnostic QQ-plot and residual plot on the right. There is no influential point (no points in the Cook's distance; however, the model has a multicollinearity problem according to the VIF in the EDA section. We can see that most of the coefficients have a significant effect on medv.log. For example, holding other variables constant, for every additional unit in nox, the average decrease in medv.log is 0.77. The R square of the model is equal to 0.79, which means the model can explain 79% of the variation in medv.log.

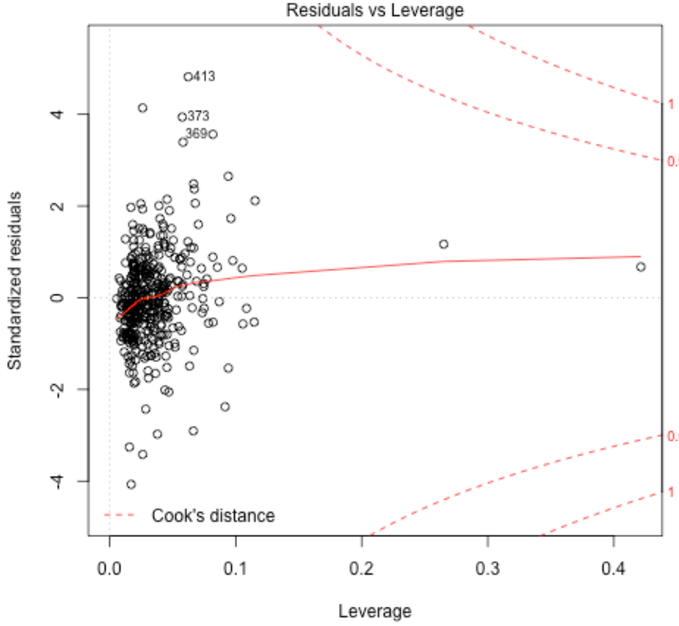
Variables	Coef	Pr(>|t|)	Sig.
(Intercept)	4.0742251	< 2e-16	***
crim	-0.0086881	2.59E-08	***
zn	0.0010046	0.108363	
indus	0.0016972	0.532606	
chas	0.0922525	0.01241	*
nox	-0.7689606	7.65E-06	***
rm	0.0822323	4.23E-06	***
age	0.0007802	0.198424	
dis	-0.0443855	1.22E-06	***
rad	0.0141637	3.56E-06	***
tax	-0.0005809	0.000676	***
ptratio	-0.0381756	2.37E-10	***
black	0.0004942	2.04E-05	***
lstat	-0.0307852	< 2e-16	***

*Regression Summary, Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1, Multiple R-squared: 0.7918, Adjusted R-squared: 0.7849*









6.2 AIC Backward Model

Variables	Coef	Pr(>|t|)	Sig.
(Intercept)	4.0417400	< 2e-16	***
crim	-0.0087764	1.83e-08	***
zn	0.0008860	0.15287	
chas	0.0963215	0.00860	**
nox	-0.6834618	1.69e-05	***
rm	0.0859062	7.86e-07	***
dis	-0.0491355	8.56e-09	***
rad	0.0134374	5.63e-06	***
tax	-0.0005342	0.00073	***
ptratio	-0.0369238	4.11e-10	***
black	0.0005033	1.35e-05	***
lstat	-0.0295791	< 2e-16	***

	R Square	Adj R Square	RMSE
AIC Model	0.775429	0.7691752	4.3393

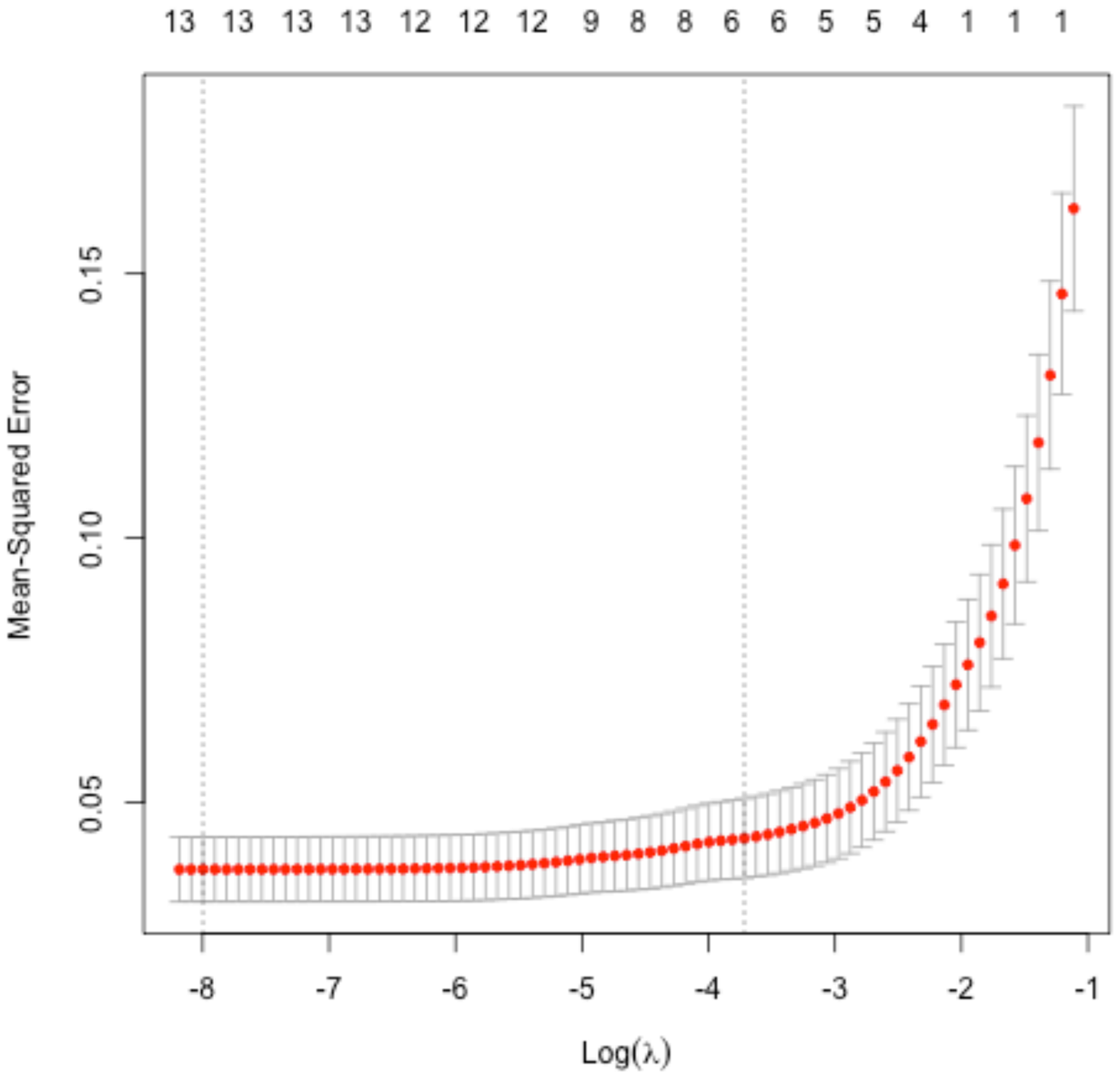
The model starts from the full model and delete one variable from the full model at one time to achieve a better model with a lower AIC, which is calculated as:

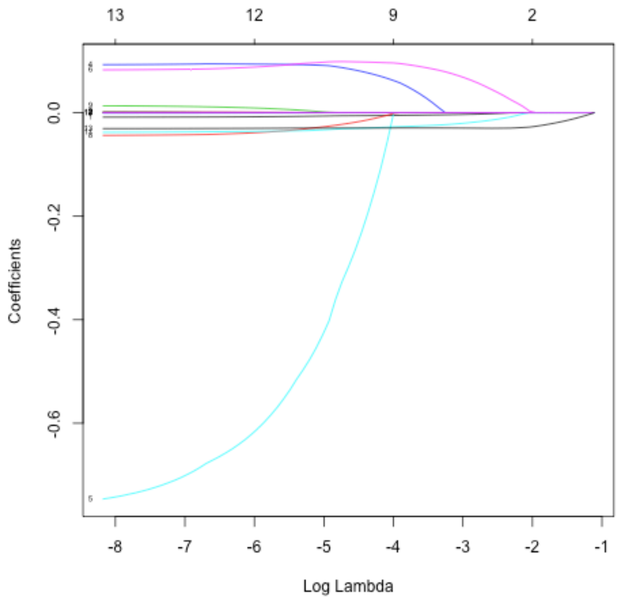
$$AIC = 2k - \ln(\hat{L})$$

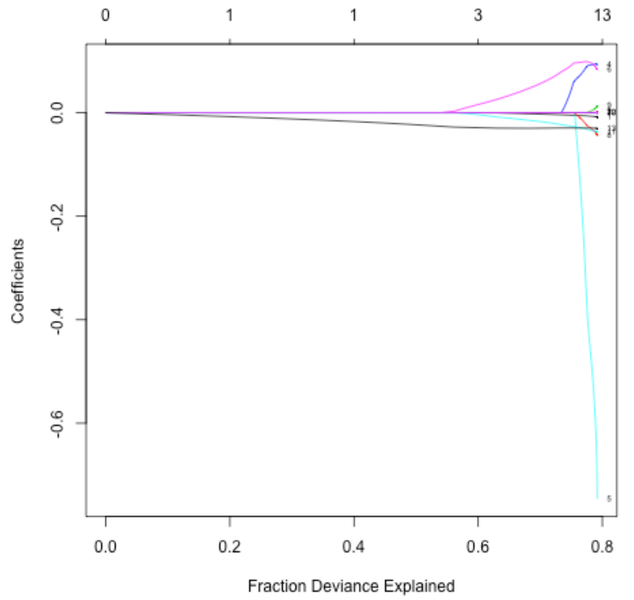
k: number of estimated parameters in the model.
L: maximum value of the likelihood function for the model.
The lower AIC is calculated, the better the model is.

A backward AIC final model through the stepwise-backward method would eliminate multicollinearity problems and overfitting problems. The result showing in the top table has removed two explanatory variables, which are indus and age. We can see that almost all the variables have a significant effect on medv.log.
By transforming the predicted log value back to the typical value, we calculated the model performance on the training set, which is shown in the bottom table.

6.3 LASSO Model







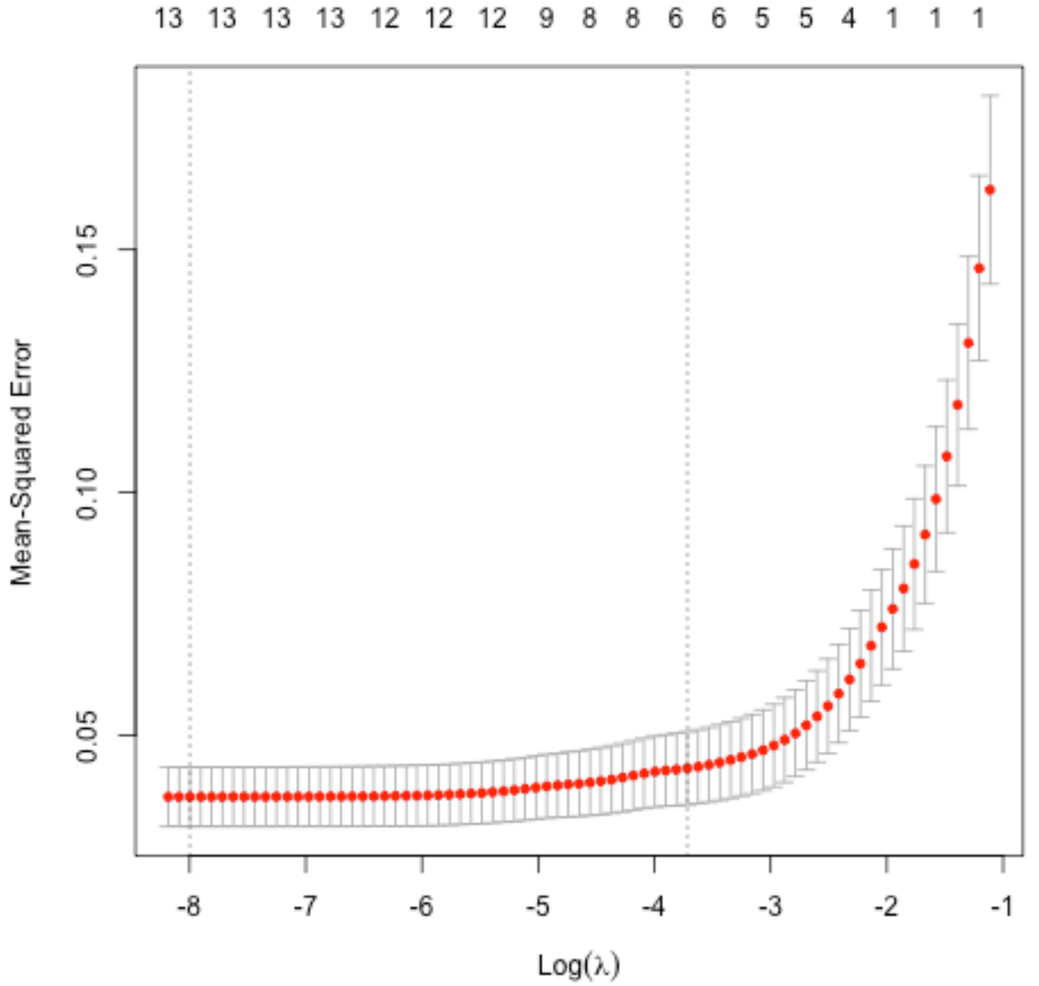
When lambda is equal to 0, no parameters are eliminated. The estimation of the coefficient is precisely identical to the previous one with OLS linear regression. However, As lambda increases, more and more coefficients are influenced and tend to be zero; when lambda tends to infinite, all coefficients will tend to 0. LASSO method selection process chooses the best lambda with its corresponding model. In this model, the best lambda is 0.000336631.

6.3 LASSO Model

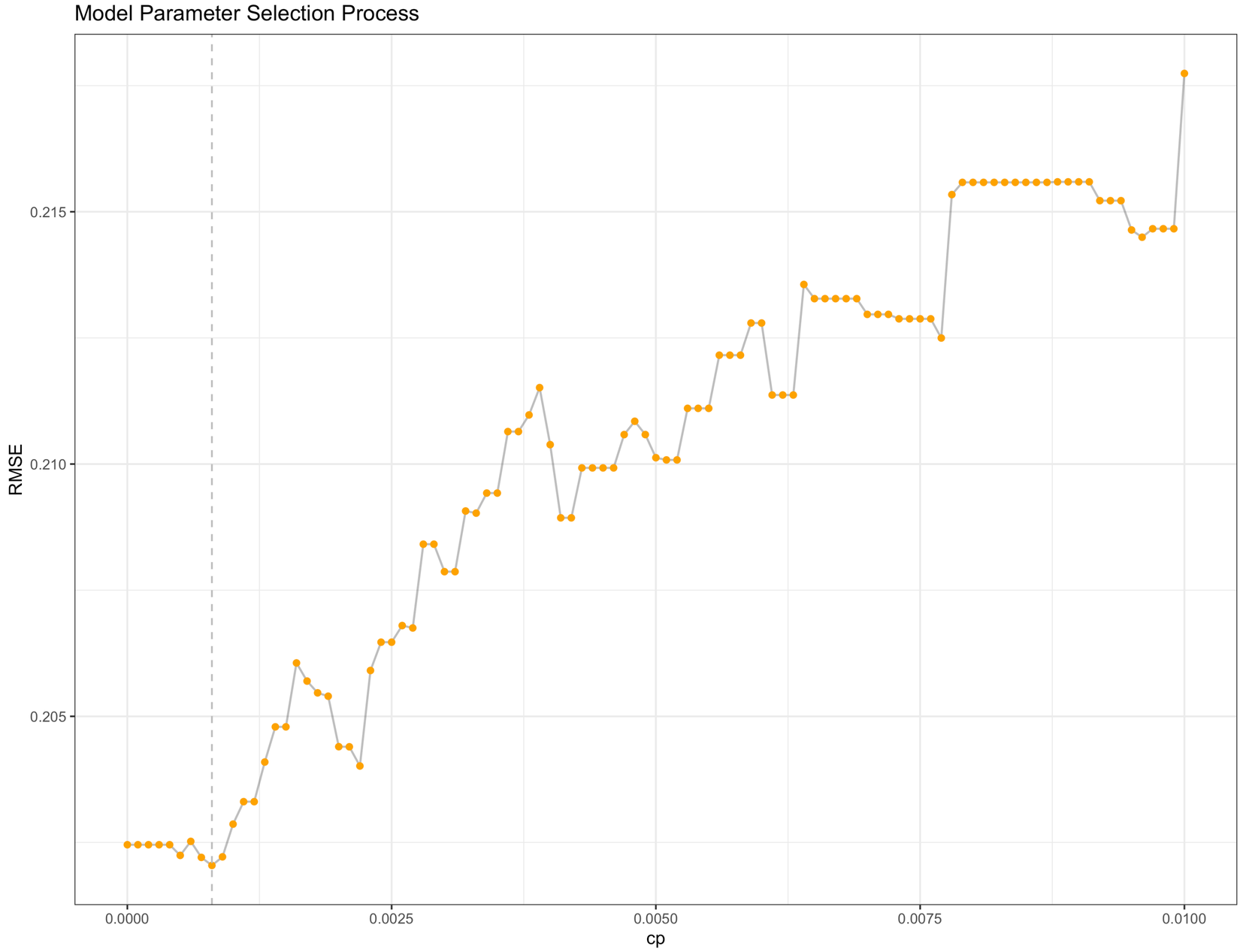
Variables	Coef
(Intercept)	4.0742251
crim	-0.0085736823
zn	0.0009285056
indus	0.0012299462
chas	0.0929906556
nox	-0.7470500709
rm	0.0831255962
age	0.0007155919
dis	-0.0438638987
rad	0.0133598297
tax	-0.0005398714
ptratio	-0.0378202880
black	0.0004911908
lstat	-0.0306514035

	R Square	Adj R Square	RMSE
LASSO Model	0.7758343	0.7684191	4.335383

The table on the top shows the variable coefficients corresponding to the lambda equal to 0.000336631, where log(0.000336631) = -7.996523.
The table on the bottom shows the model performance on the training set.



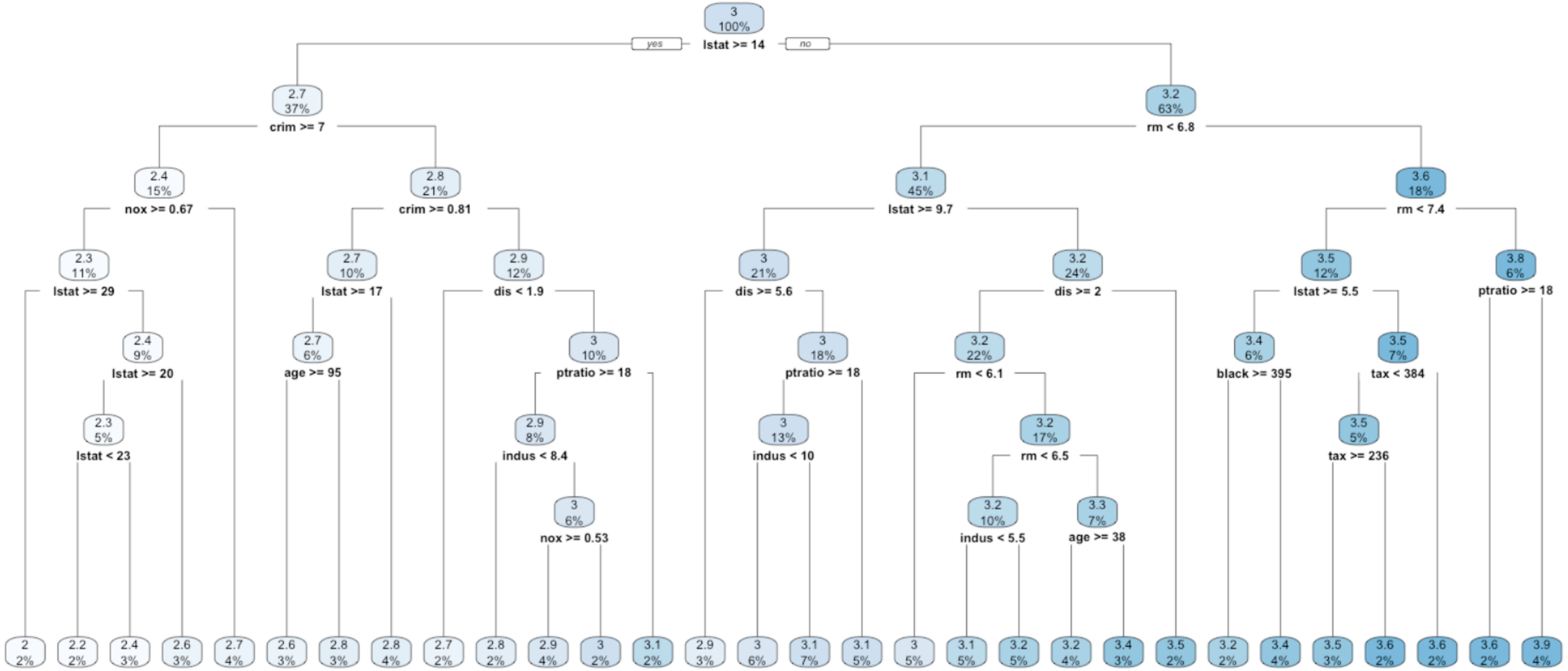
6.4 Decision Tree - CART



cp stands for complexity parameter. When cp = 0, the tree will be the most complex one with lots of splits; however, cp = 0 may not be the best one. The figure shows the tuning tree selection process to choose the best cp with the lowest MSE (RMSE). The following table shows the first 10 of the tuning process. We can see that the best choice is when cp = 8e-04.

cp	RMSE
0e+00	0.2024554
1e-04	0.2024554
2e-04	0.2024554
3e-04	0.2024554
4e-04	0.2024554
5e-04	0.2022443
6e-04	0.2025216
7e-04	0.2022043
8e-04	0.2020450
9e-04	0.2022152

6.4 Decision Tree - CART



	R Square	Adj R Square	RMSE
Decision Tree	0.8839579	0.8801194	3.119253

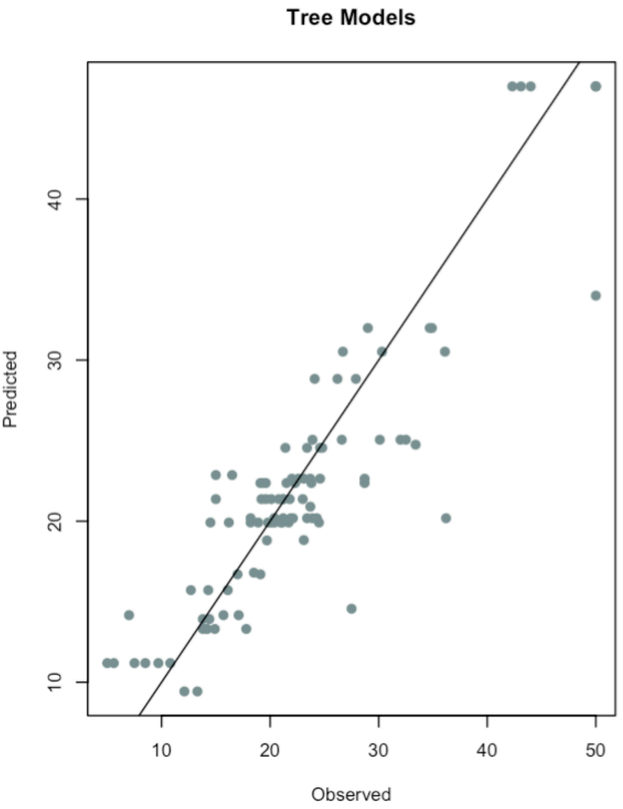
The decision tree model is showing as the figure, by using different variables as the classifiers. The table on the left shows the model performance on the training set.

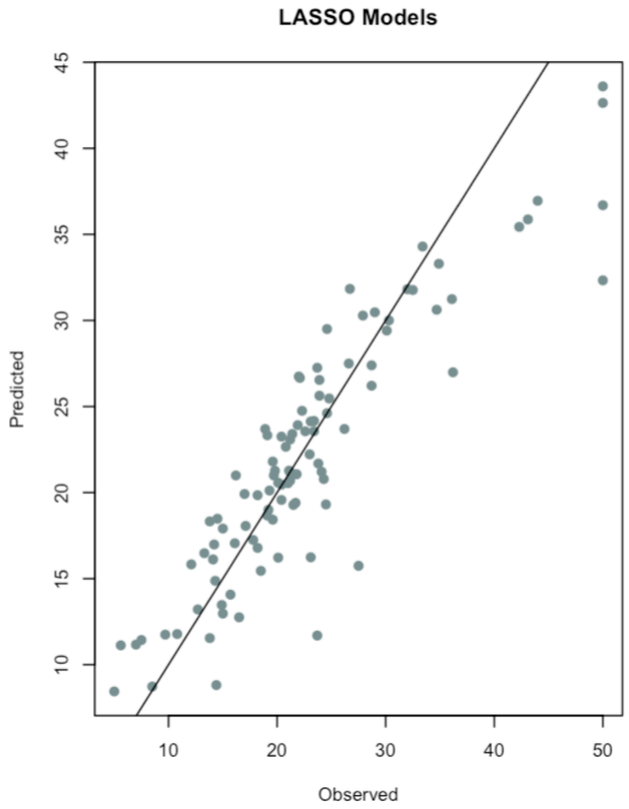
6.5 Comparison among the models

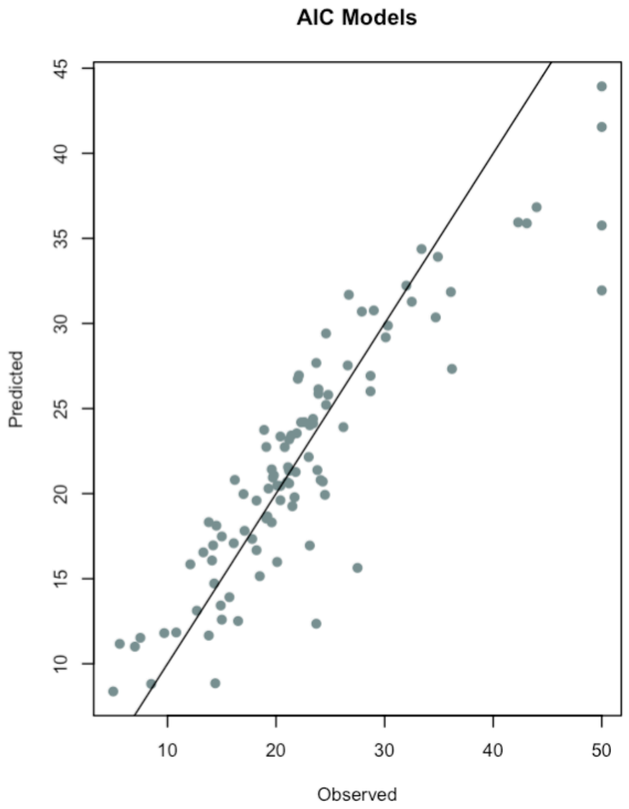
By comparing the three models, we can observe that the tree model has an R square equal to 88% and RMSE equal to 3.12, which has the highest R square (explain more variation in medv) and the lowest RMSE (the model are more accurate). As a result, the tree model has the best model performance in the training set.

	R Square	**Adj R Square**	**RMSE**
AIC Model	0.775429	0.7691752	4.3393
LASSO Model	0.7758343	0.7684191	4.335383
Decision Tree	0.8839579	0.8801194	3.119253

6.6 Prediction on the Test Set







The figure shows the model performance on the test set. The figure showing at the top shows the relationship between the observed value and the predicted value. If the model is 100% accurate, then all the spots should be on the line exactly. We can see the tree model still has the best model performance; it can explain about 80% of the variation of medv in the test set with the highest accuracy.

	R Square	Adj R Square	RMSE
AIC Model	0.7929228	0.7612522	4.238743
LASSO Model	0.7946502	0.7632437	4.221027
Decision Tree	0.8000974	0.7695241	4.164666

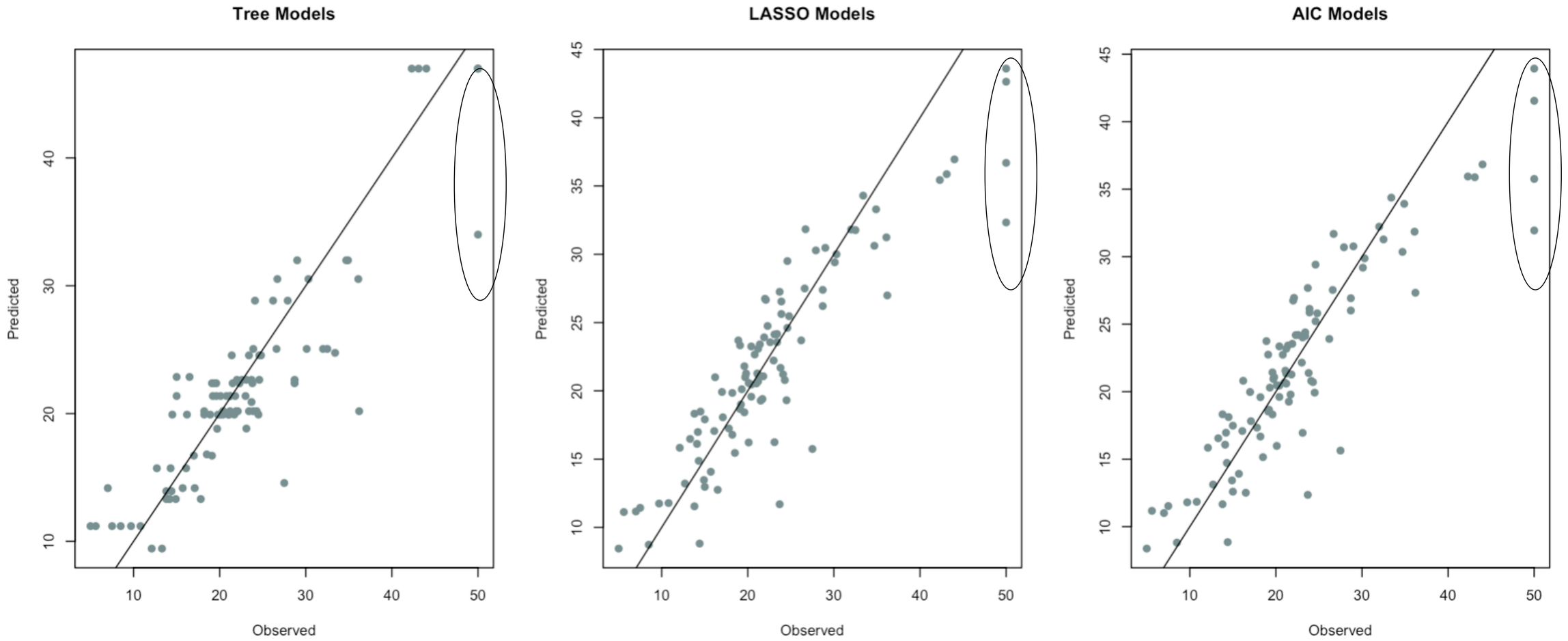
7 CONCLUSION

The tree model performs the best compared to the other two models on both the training set and test set through the previous model building and model analysis. The model comparison on the train set and test set are showing in the tables below. The report does not add the full model into contrast because it is overfitting and have multicollinearity problems. As a result, the report would choose the decision tree with a cp equal to 8e-04 as the best prediction model to predict the Median Value of Owner-Occupied Homes (medv).

Train Set	R Square	Adj R Square	RMSE
AIC Model	0.775429	0.7691752	4.3393
LASSO Model	0.7758343	0.7684191	4.335383
Decision Tree	0.8839579	0.8801194	3.119253

Test Set	R Square	Adj R Square	RMSE
AIC Model	0.7959083	0.7646942	4.208077
LASSO Model	0.7946502	0.7632437	4.221027
Decision Tree	0.8000974	0.7695241	4.164666

8 FURTHER THINKING



When the observed medv is equal to 50, all the models show a relatively large bias; by removing these values, RMSE would decrease significantly and increase the R square (showing in the R script). Further research can focus on the observation with medv equal to 50 and explore why they are different and hard to predict.