STAT 151A_Midterm II

Huidi Wang

May 10, 2016

1.Introduction

The data used in this project are about Housing data for 506 census tracts of Boston from the 1970 census collected by Harrison and Rubinfeld (1979). It shows the relationship between housing price and other thirteen related varibles, such as crime rate, lower status of the population, property-tax rate, etc.. Our goal is to figure out how these measurements influence the reference viarible through linear regression methods.

A basic concept of linear regression is to calculate parameter beta before each explnatory variable, like

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$$

There are two main objectives in a regression problem. Like in this case, through comapring estimated betas, we can conclude with how each variable affect the housing price, and which has bigger influence.

In this project report, I will do a brief description about this dataset first, and then analyze data through detecting and eliminating unusual observations from the data, using variable selection methods to get the most powerful model, using t-test and linear model to estimate beta. Finally, I will diagnose the result model by checking if the data fit for the three assumptions underlying the linear model.

2. Brief Description of Data

1) Understanding all the variables given in the data:

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

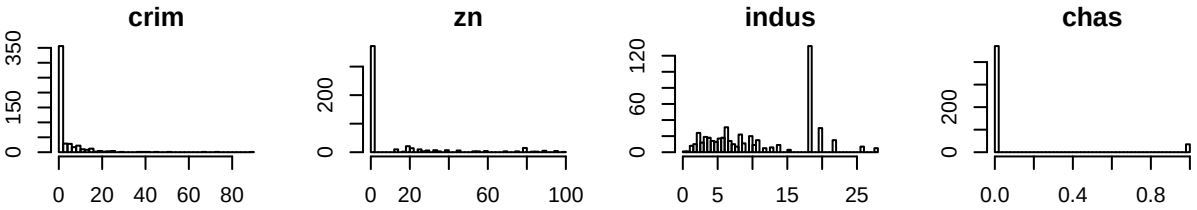
age: proportion of owner-occupied units built prior to 1940 **dis:** weighted distances to five Boston employment centres

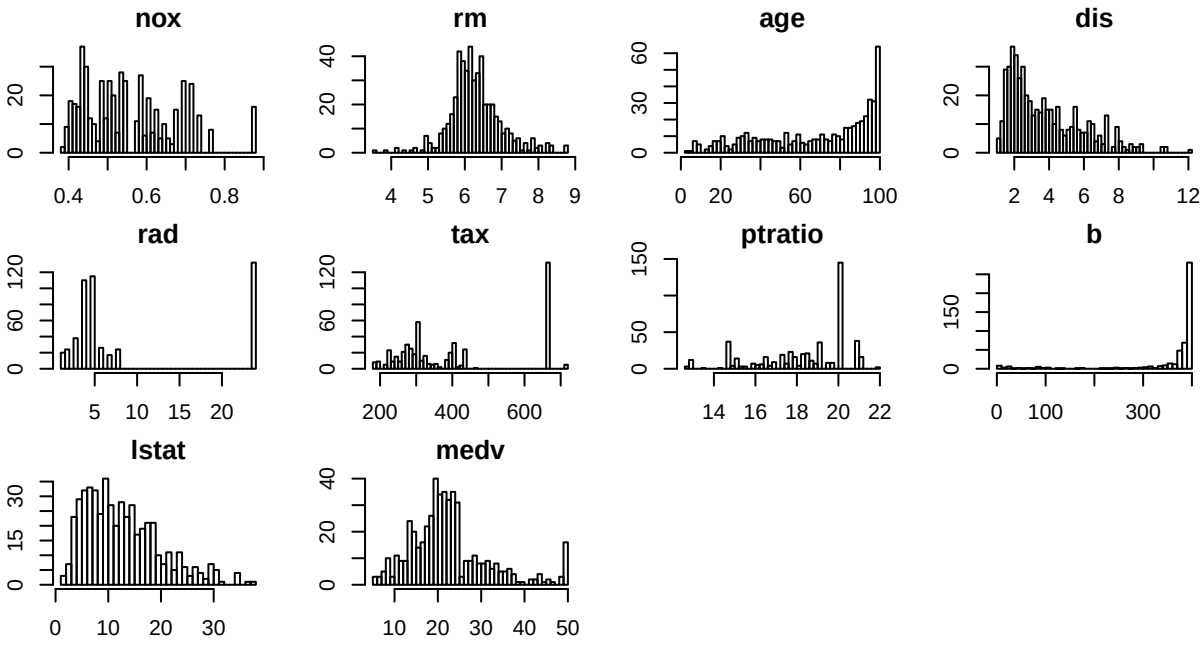
rad: index of accessibility to radial highways **tax:** full-value property-tax rate per USD 10,000

ptratio: pupil-teacher ratio by town **b:** 1000(B - 0.63)^2 where B is the proportion of blacks by town

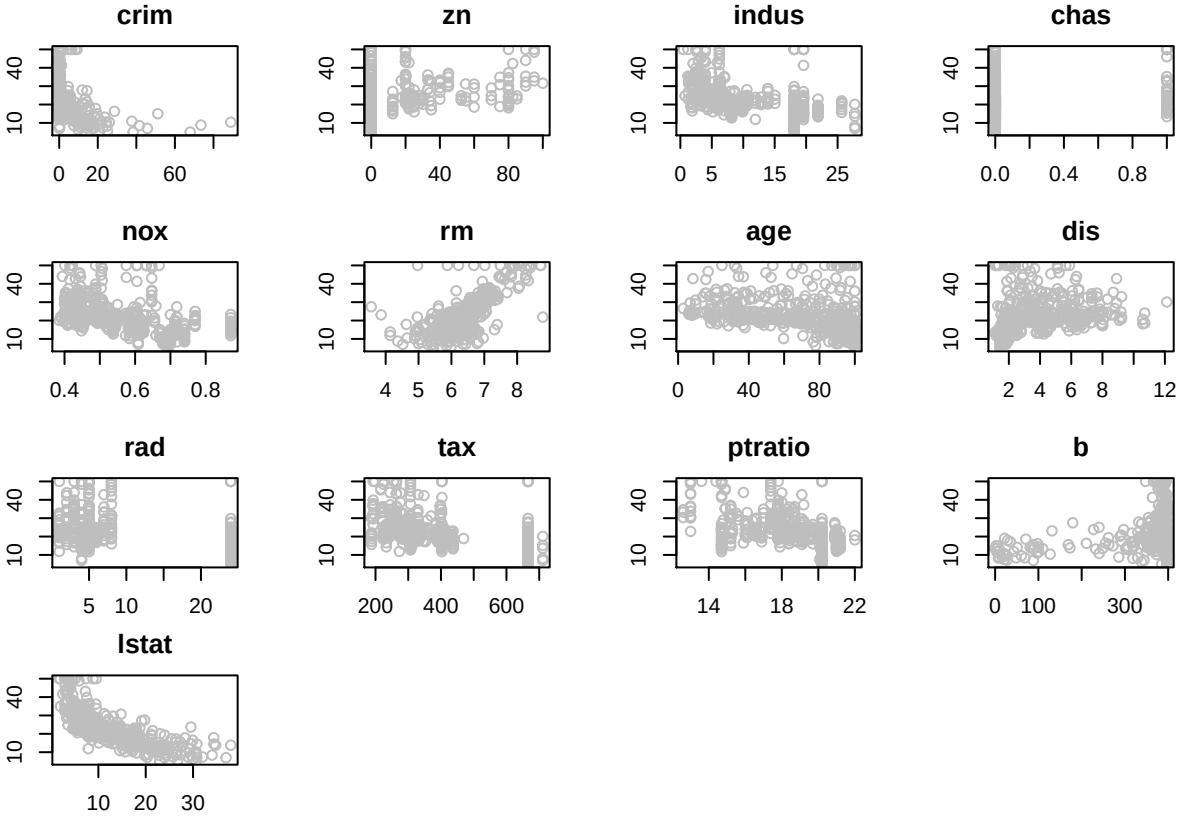
lstat: percentage of lower status of the population **medv:** median value of owner-occupied homes in USD 1000's

2) Summary the whole data first. Variable **chas** is defined as *level* rather than *numeric*. Therefore, written as 0/1. All other varibles' data are considered as *numeric*. The following graphs show each variable's histgram.





These histgrams represent each variable's distribution. **Crim,zn,chas,rad,tax,ptratio,b** have highly density in a small region. For example,most of observations have 0 crimes, and 0 proportion of residential land zoned for lots over 25,000 sq.ft. As for **nox, rm, age, dis, lstat**, Their values are more scattered in the domain. **rm** even looks like a normal distribution.



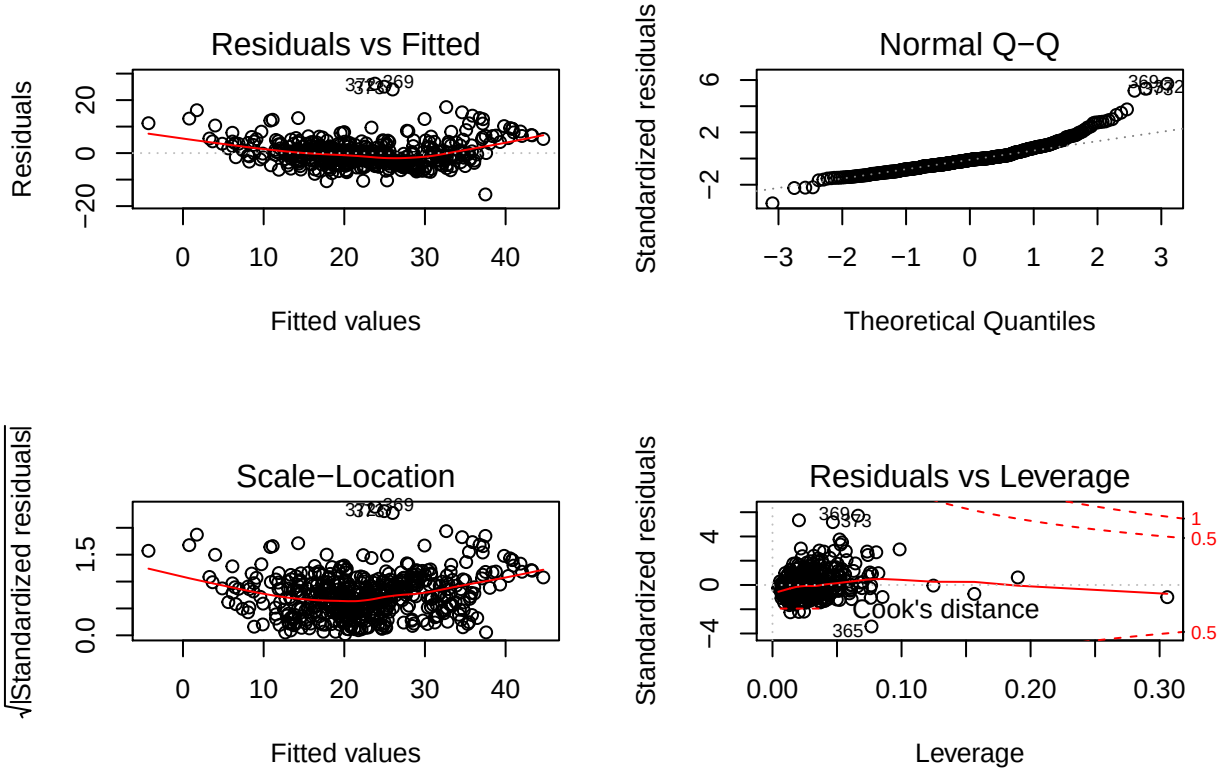
3)The relationship between **medv** with each explanatory variable is shown in the above graphs. By observation, **medv** has the most linear relationship with **rm** and **lstat**.

4) Outliers:Roughly estimating outliers. For R code please see appendix.

a) After taking a look at previous figures, I found some reasonable outliers which are ## 354 381 406 419 (See Appendix)

b) In addition, the plots from linear model of medv with all other variables give other possible outliers.(#369,#372,#373)

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'col' will be disregarded
```



To give a short background about these plots:#1)The first plot (residuals vs.fitted values) is a simple scatterplot between residuals and predicted values. It looks randomly more or less around horizontal line equals 0.It means the residual error is 0. #2)Normal QQ_plot (standardized residuals vs. theoretical quantiles):If the residual errors are distributed normally, the points should be on a straight line.#3) The scale-location plot shows that the square root of the standardized residuals as a function of the fitted values. The trend in the plot should be randomly like #1). #4) The forth plot shows Cook's distance that tells us which points have the biggest influence on the regression by leverage.

We can compare the plots' shape with removing outliers (Seen in the R-code). If removing ouliers, we could find out that these outliers we observed make sense since qq-plot bocomes more normal after removing outliers, residuals become more constant and scattered, and leverages become smaller.

From this part, we have get some basic understanding about all the variables, predicted that the most linear relationship belongs to **medv** against **rm** and **lstat**, and roughly detected couple of outliers.

3. Model Analysis

In this part, I will use several moethods to select the most predictive powerful variables as a model to generate housing price, and diagonalize the selected model's data to see if they fit for the three assumptions underlying the linear model.

To make e_i have equal variance and linearaity in the linear model, I will take log of **medv**, and use these values to do variable selection. (I'll further explain how log(medv) could make the model more accurate in the (3) part.) So now, we have initial model

$$log(medv) = \beta_0 + \beta_1 crim + \beta_2 zn + ... + \beta_{13} lsts$$

Once we summary this model through r, we will see the result like this:

```
#Residuals:
#    Min       1Q   Median       3Q      Max
#-0.73361 -0.09747 -0.01657  0.09629  0.86435
#              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
#chas1        0.1008876  0.0344859   2.925 0.003598 **
#...
#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
```

This chart is saying the smaller P-value is, the stronger influence the variable has. It also gives us residual standard error is 0.1899 and residual region (-0.73361,0.86435).

(1) Model Selection From here, I will use six methods to do variable selection. (For more details, please read through R code.)

#1. Backward elimination using individual p-value. At each stage we remove the predictor with the largest p-value over 0.05. As a result, variables **crim,zn,chas,nox,rm,dis,rad,tax,ptratio,b,lstat** have P-value less than 0.05, which means they have more predictive power, should be selected. (Eleven variables total)

#2. Forward Selection: using p-value at each stage, we add the predictor with the smallest p-value less than 0.05. As a result, eleven variables are selected as predictors. They are **crim,zn,chas,nox,rm,dis,rad,tax,ptratio,b,lstat**. Particularly, **lstat and rm** are firstly selected.

#3. Adjusted R^2 method: get 12 variables

#4. AIC method: get 11 variables

#5. BIC method: get 10 variables

#6. Mallow's Cp method

Conclusion: Through cross-validation, the smallest cv.scores belong to **crim, zn, chas, nox, rm, dis, rad, tax, ptratio, b, lstat** eleven variables.

(2) Regression Diagnostics

As we know, the estimates for β and their confidence intervals in the linear model depend on the assumptions underlying the linear model. In particular, we have three assumptions:

–The errors $e_1, ..., e_n$ are independent, have equal variance σ^2 and are normally distributed.

–We have assumed that the expected value of the reponse vector Y equals $X * \beta$.

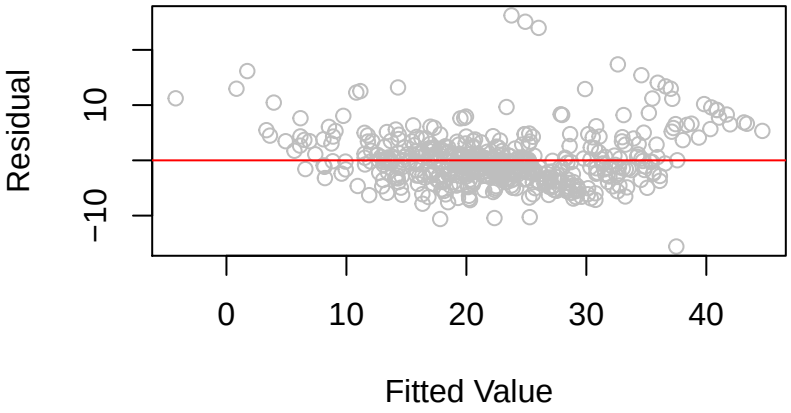
–We have assumed that all the subjects obey the same linear model. In practice, it may happen that a few subjects do not obey the model. These few observations might change the choice and fit of the model.

In this part, I will check the three assumptions one by one under selected model. If they are qualified, we could say the data used in the selected model could result with more accurate β and Confidence Interval through linear regression.

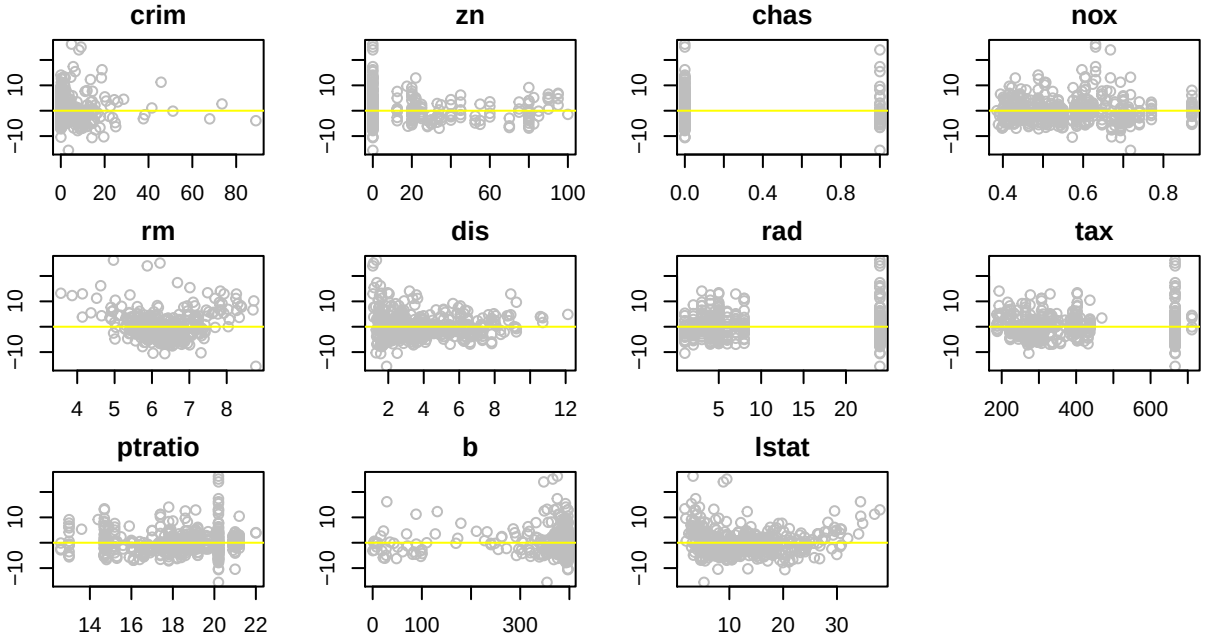
#Step1: Check constant variance, normality, and correlated errors:

1. **constant variance:** There are two ways to observe if the data have constant variance.

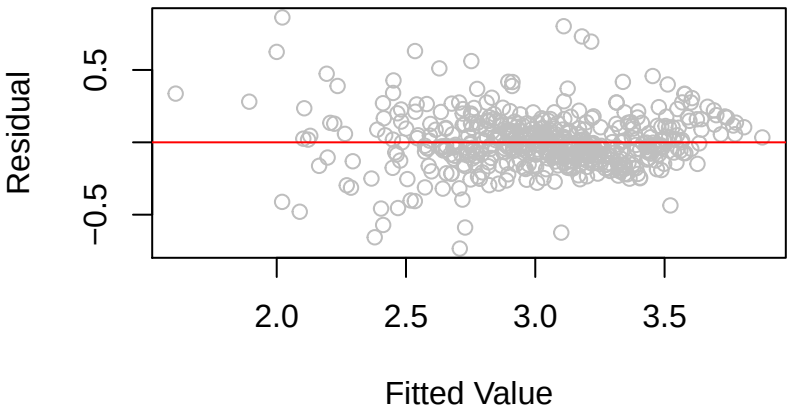
%%plot of residuals against fitted values



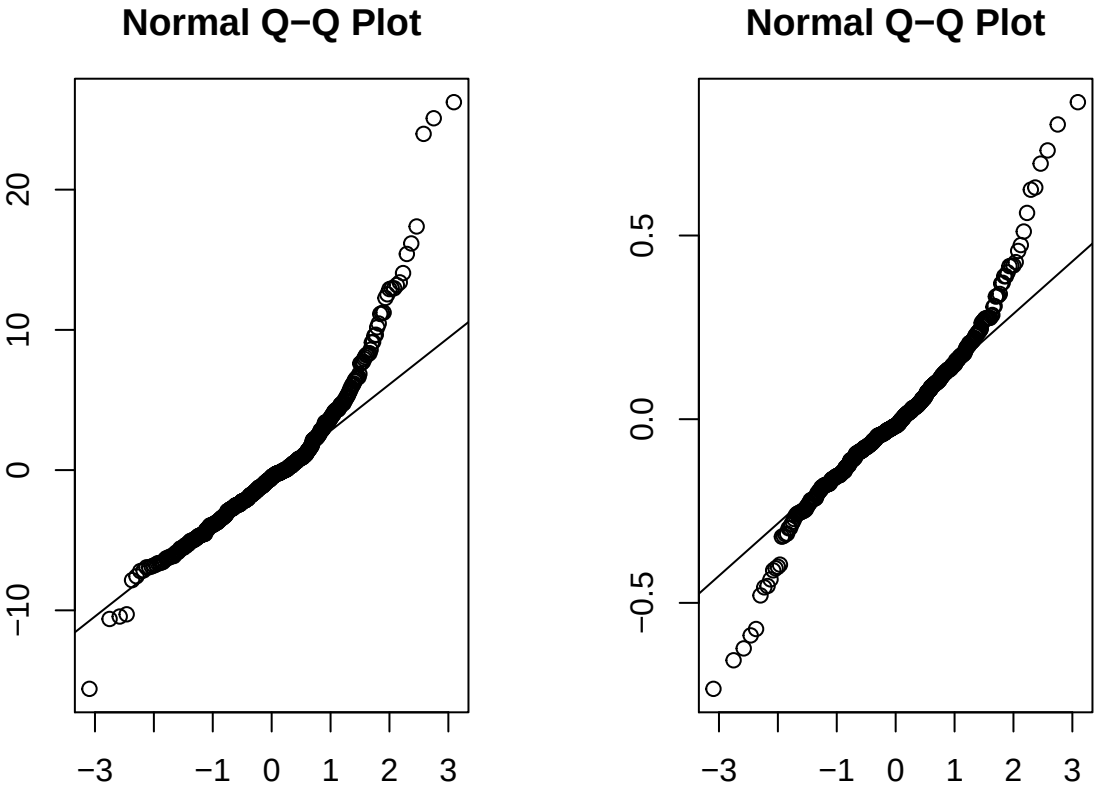
%%plot of residuals against explanatory variables



The first plot has no evidence of non-constant variance or non-linearity since symetric by 0, except some errors. In the second couple of plots, you could say constant variance as shown in **nox, rm, lstat**. If we take log of **medv**, the residuals will become more randomly and symmetric by 0 in both first and second graphs. Like below (comparing to first plot):



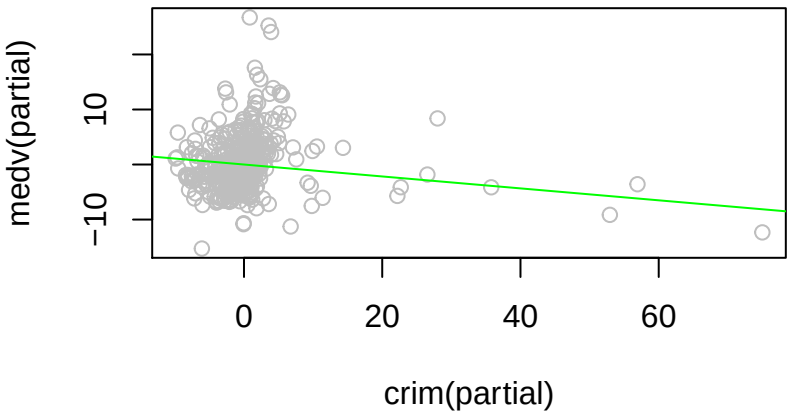
2. Normality: Use qqnorm to check if the data owns normality. The left one w/o taking log, and the right one w/ taking log. Obviously, the right one have more points in the normal line, which means more normal.



3. Correlated Error: plot the autocorrelation function of the residuals. From the plot, we can find that most of residuals are around blue line, which means the errors are not correlated. (Please see graph in Appendix)

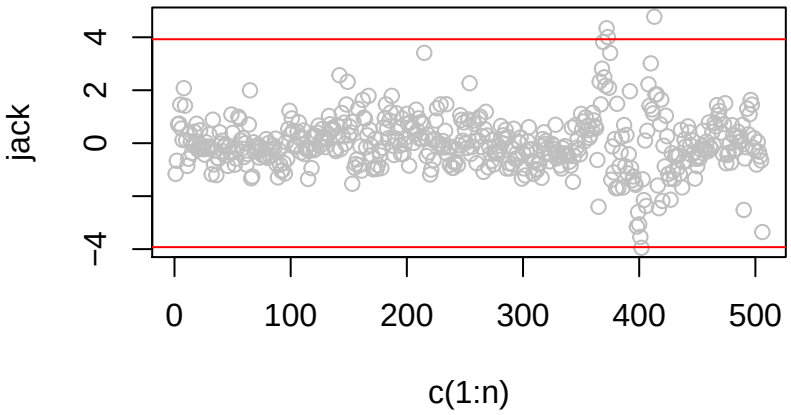
#Step2: detect linearity: partial regression

Take first variable crim as an example. After plotting its partial regression, and comparing its slope to crime's estimated β, we can find they are exactly same. Actually, all variables have this property. (Check R code for more plots about this.) It means linerality assumption is qualified.

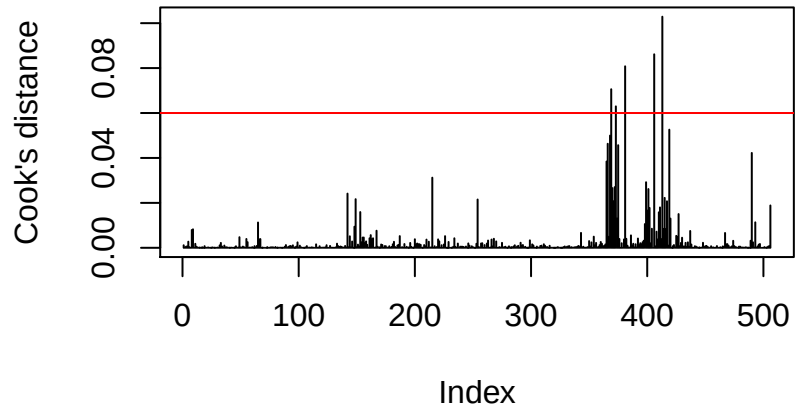


#Step3: Use three methods to detect unusual obervations

1) predicted residuals: there are four outliers to be found by this way, which are #372, #373, #413, #402



2) Cooks' distance: Over than criteria 0.06 (improve r/n to 0.04 to narrow down the amount of outliers) are counted as ouliers: #369 373 381 406 413



#3) Check for leverges: compare if outliers all have high leverages. Through calculation in r code, 84.6% outliers belong to high leverage, which means these outliers are reasonable. Besides, if we check with coefficient function which gives how $/beta_i$ changes with eliminating ith observation. The intercept values seem to change quite a bit in all high leverage cases.

Conclusion: After analyzing the dataset, we can conclude with a) **age** and **indus** should be eliminated from linear model. b) outliers until now would be #381 406 419 354 372 373 413 402 369 373 381 406 413. They are proved with high leverages. c)The dataset w/ log(medv) are qualified to linear model's three consumptions.

In the next part, I will use the selected model with emlinating ouliers to do linear regression, and compare the result with initial model.

4. General Conclusions/Discussion

(1) Comparison and Conclusion

Let's summarize the new model, and compare it with the initial model shown in the beginning of Part 3. First, residuals have the most obvious change to smaller from region (-0.73361,0.86435) to (-0.65631,0.64094). For the p-value, almost all variables are less than 0.05 except zn, but zn is 0.066 which is not above 0.05 a lot, still acceptable. Residual standard error decreases as well from 0.7896 to 0.1718. All of these changes show that analyzing data has generated a much better model than before. Remeber we predicted **lstat** and **rm** have the most linear relationship with **medv** in Part 2. Here we can prove it since they have smallest P-value.

```
#Residuals:
#     Min       1Q   Median       3Q      Max
#-0.65631 -0.08939 -0.01445  0.09164  0.64094
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)  3.7877629  0.1897035  19.967  < 2e-16 ***
#crim        -0.0111000  0.0018972  -5.851 9.02e-09 ***
#zn           0.0009070  0.0004939   1.836  0.06693 .
#chas         0.0819159  0.0314310   2.606  0.00944 **
#nox         -0.6233285  0.1306061  -4.773 2.41e-06 ***
#rm           0.1148987  0.0151796   7.569 1.91e-13 ***
#...
#lstat       -0.0260544  0.0017851 -14.595  < 2e-16 ***
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#Residual standard error: 0.1718 on 485 degrees of freedom
```

All in all, Boston Housing dataset include 506 observations and 14 variables. It shows the relationship between median housing price and other related thirteen variables. Through data discription in part 2, we understand every variable's distribution, potential outliers from dataset, and inferred that **lstat** and **rm** have the most linear relationship with **medv**. In Part 3, we use six methods to select out eleven variables which have the most predictive power. Use regression diagnostics to prove the given data are qualified to use in linear model; especially, the data become more accurate if take log of medv. Finally, we conclude with a more powerful model to predict housing price under eleven variables.

(2) Interesting Questions and Discussion

(1) Which variables mainly determine the median housing price in a tract? As mentioned in the previous paragrah, the answer should be **lstat** and **rm**, because they look mostly like linear relationship in the graphs between medv with each explanatory variable. Besides, in the summary of linear model, they have smallest p-value. During variable selection, they are always picked up first.

(2)What is the willingness to pay for clean air? Go back to relationship of **medv** and **nox**, people would like to pay more if Nitric Oxides concentrates around 0.4 to 0.5. They may like to pay from 15 to 35. Some of them are even ok to pay over 40.

(3)How does lstat afftect medv? and why? As we can see from the gragh, a higher percentage of lower status of the population will drag down the median value of housing price. Because this part of people who have relative low income will have a higher demand on the low price of housing, and futher make medv decrease.

(4)Since rm and lstat both have strong linear relationoship with medv, then only consider them, which one is more powerful to predict medv? To answer this question, I need to do a new linear regression like medv=β_0+β_1.rm+β_2.lstat. (see more details in R code) The estimated β of **rm** is 5.09479, much bigger than **lstat** is -0.64236. Thus, **rm** is more predictive than **lstat**.