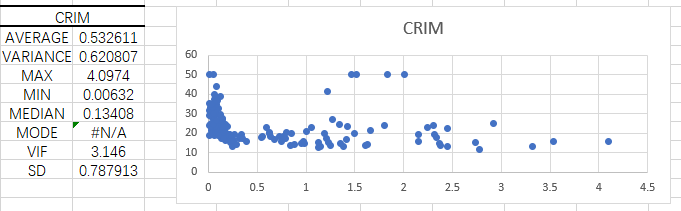
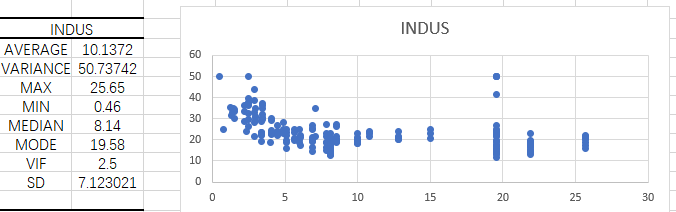
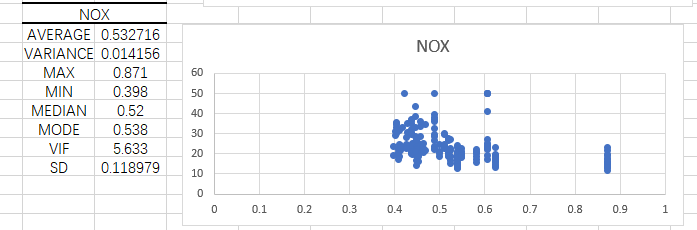
**Dataset Boston house price Regression Report**

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of [Boston MA](http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html). The following describes the dataset columns:

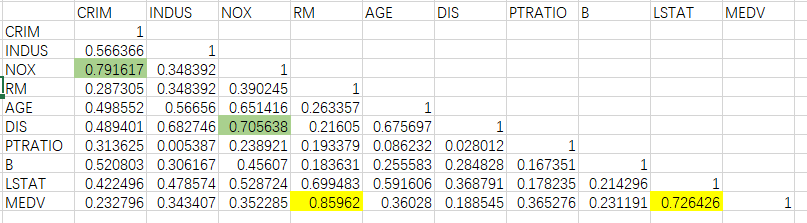
* CRIM - per capita crime rate by town
* INDUS - proportion of non-retail business acres per town.
* NOX - nitric oxides concentration (parts per 10 million)
* RM - average number of rooms per dwelling
* AGE - proportion of owner-occupied units built prior to 1940
* DIS - weighted distances to five Boston employment centres
* PTRATIO - pupil-teacher ratio by town
* B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
* LSTAT - % lower status of the population
* MEDV - Median value of owner-occupied homes in $1000's

First of all, I use average, variance, max, min, median, mode, graph and correlation to analyze the performance of all independent variables. The analysis of 3 features is shown here.

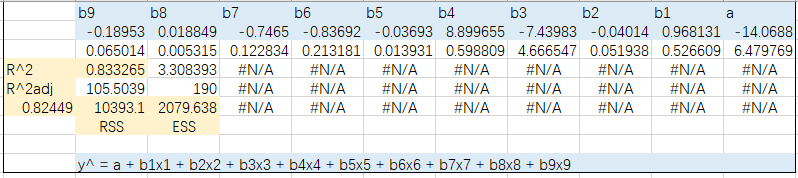




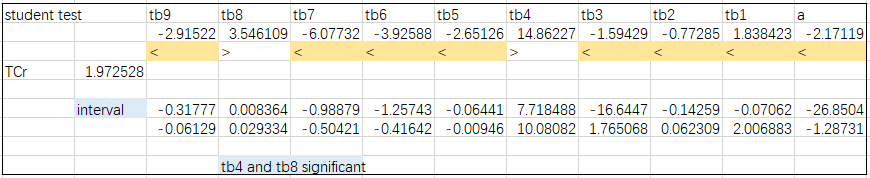
After that, I use the module correlation get the whole relation map between each other. I want to figure out that which x\_i have strong relation with y, which means that the correlation coefficient should be larger than 0,4, and I also need to check that between which two independent variables there is a strong tie which shows that the correlation coefficient is large than 0,7. In table 1, we can notice that CRIM and NOX have strong connection , also for NOX and DIS. And the LSTAT 、RM also have strong connection with MEDV.



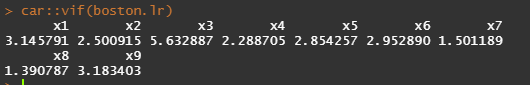
I get the coefficient, standard error of the variables with Function LINEST. At first glance, we can find that is 0,833265 which is very close to 1,this shows a strong correlation between the variable and the variable being predicted of the regression model. As to the coefficient of determination and the adjusted coefficient of determination, , while . Using can avoid being trapped into specifying an overly complex model.



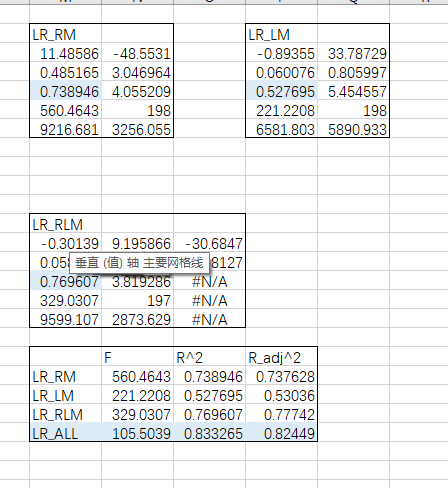
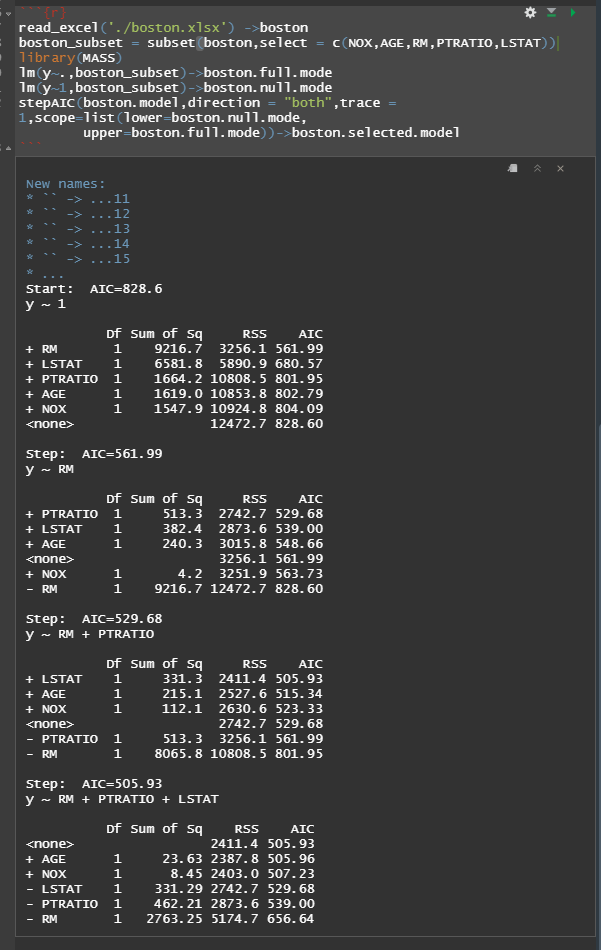
We can get the conclusion after finished the student test for every coefficient. that, tb4 and tb8 are significant in my case.



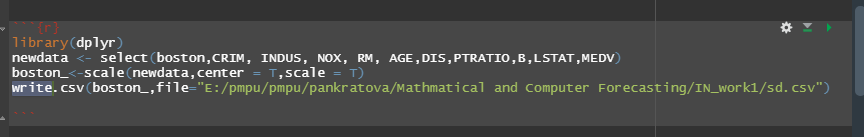
I use program R to calculate VIF for each feature. Because when 0<VIF≤5, we can say that there is no multicollinearity ; when 5<VIF≤10 there exists weak multicollinearity; when 10<VIF≤100 Have moderate or strong complex covariance. In my case , all VIF\_j is belong to (0,5], so from this result we can see there no exist the multicollinearity in my data case.



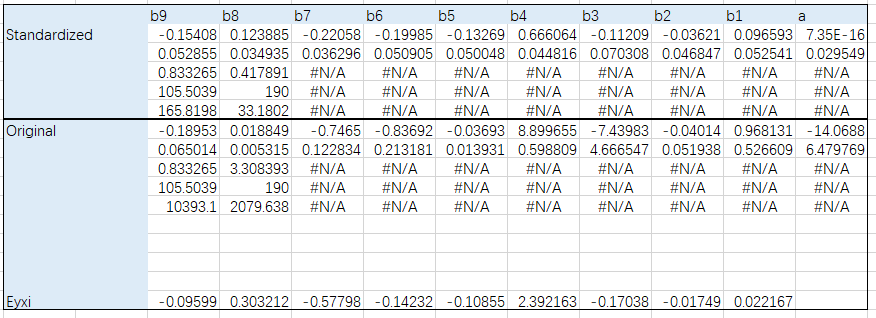
I choose the feature RM, LSTAT, to make a new LR model to predict the MEDV , and here is the comparison for each LR model. And in program R there exist a function based on an alternative assessment to check the best combination of all features. So I use it to get this result:

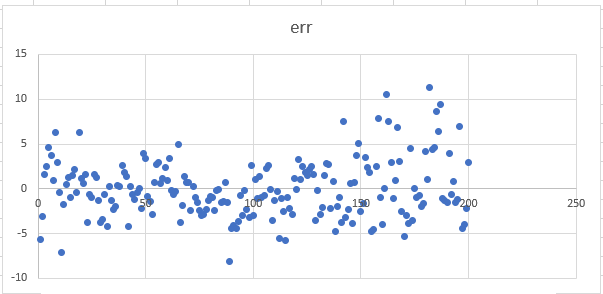
 

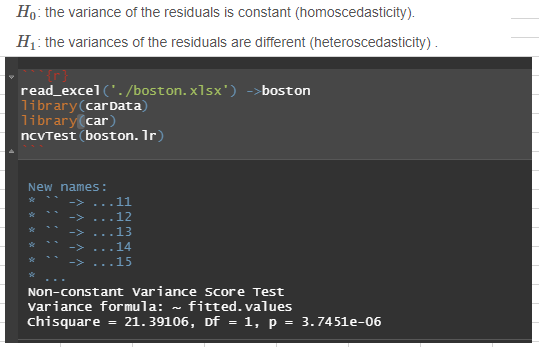
In order to calculate the Eyxi we must first regularise the data. I use R to regularize the data:



And after that, I use the I ran a linear regression on the already regularised data and got the following results. From this, the average elasticity coefficient was calculated.So in my case, with increasing factor xj at 1% of the avarage level the dependent variable increases be Eyxi% from its average level when the other factors are held constant.

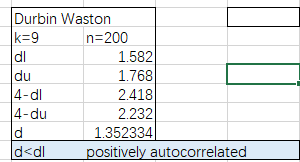


For the homoskedasticity in my case , I have applied different methods in R and EXCEL to obtain the following results, the conclusion is we have to reject the H0, that’s mean the variance of the residuals are different(heteroscedasticity). 



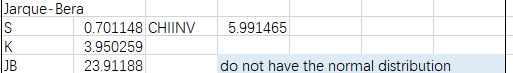
Durbin-Waston test.

Durbin-Waston statistic is . For the value of dL, dU, we can find the corresponding number in the Durbin-Waston table according to K and N. In my case, dL=1,582, dU=1,768. And we can notice d is located between [0, dL], it means that there is statistical evidence that the error terms are positively autocorrelated.



Jarque-Bera test.

With the usage of , , . And in my case, JB is larger than Chi-square distribution with degree of freedom is equal to 2 which represents that the residuals are not normally distributed.



New point forecast.

By using Function MMULT, MINVERSE, TRANSPOSE, with = , here A is the row vector of new data set while adding 1 on the first place and the first column of X is also should be 1 column. Then we can calculate the value of and also the confidence interval of new point as shown .

