STA3010
2019-2020 term 2
Course Project: California Housing Price Prediction

Professor：Feng YIN

Student ID： 117010075 Student Name: Guo Jiahui

Part1.

1)Data Division

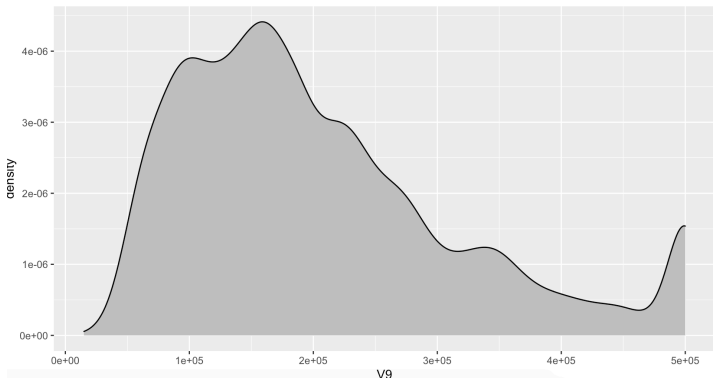
Randomly divide the data into a training data set with 60 percent of the samples and a test data set with 40 percent of the samples by using the ‘train_test_split’ from ‘sklearn’.

2)Data Detail

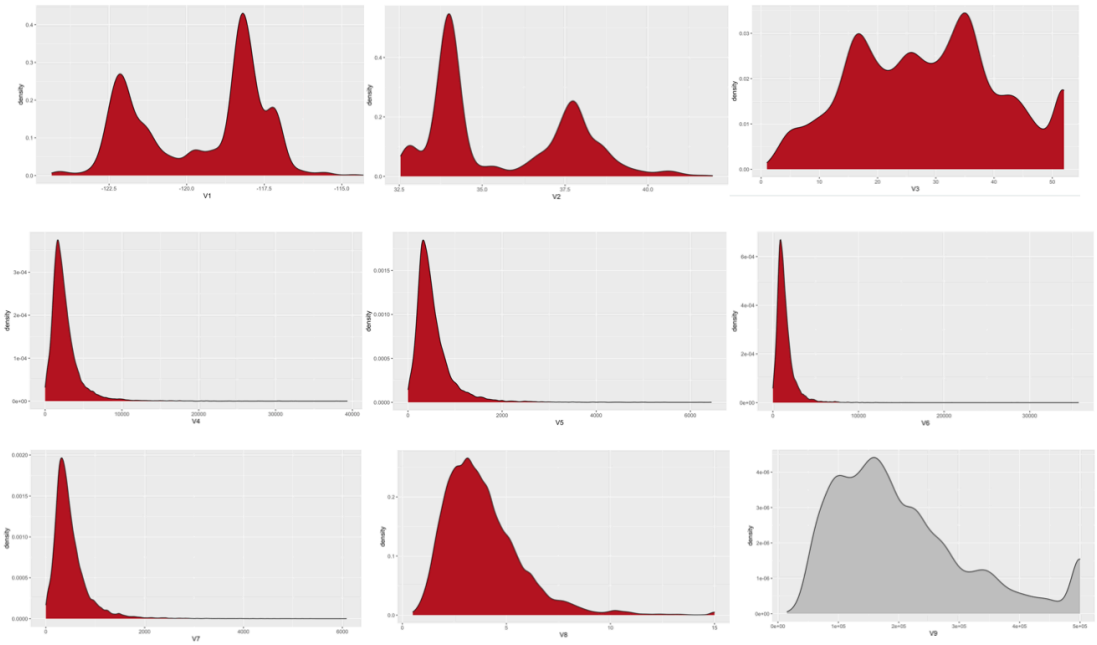
1\The main characters of the complete data set are:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.81
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.61
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.00
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.00
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.00
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.00
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.00
Correlatio	-0.04572469	-0.1444075	0.1106223	0.1106223	0.1320659	0.04775299	-0.02443643	0.06267293	0.6919736
SST	2.74832E+14								

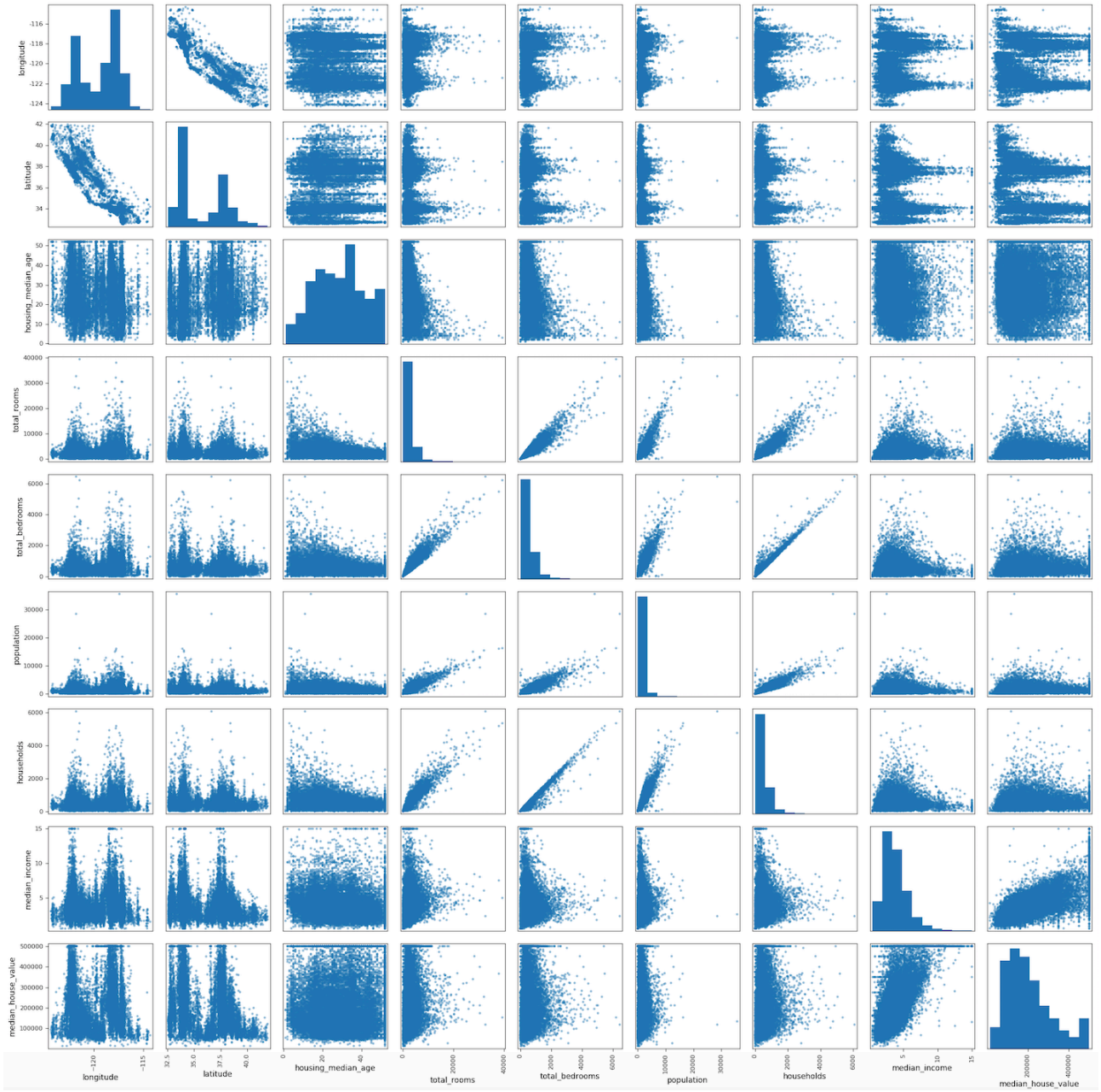
2\This is the density function plot of median house value(we most interest in). This is very similar to chi-square distribution plot.



3\These are all regressors’ density function:

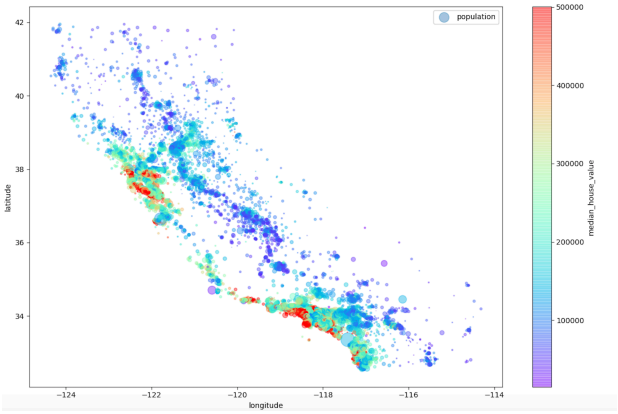


4\Here are the scatter plots for every factor:

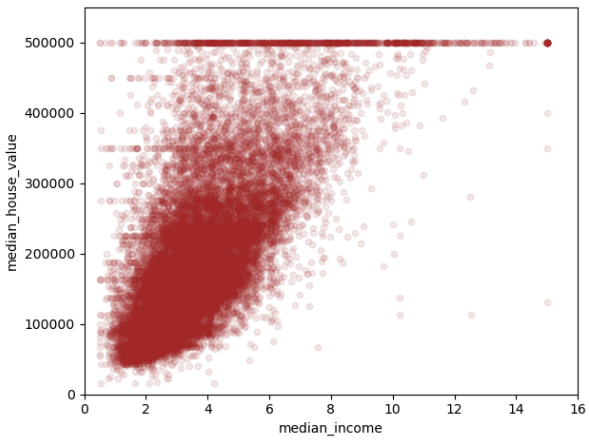


From this figure, we can clearly find that the household, total rooms and total bedrooms has obvious positive correlation among themselves.

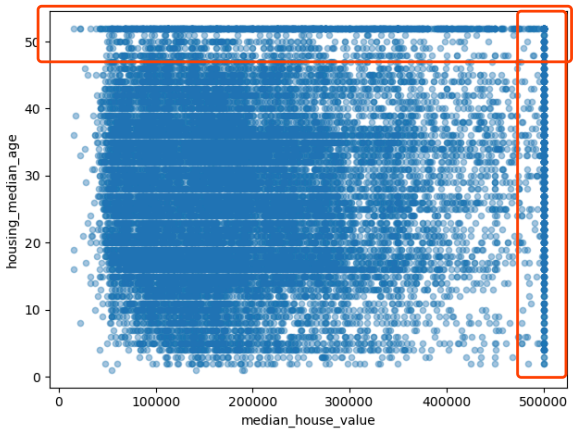
Besides, we can find some interesting plots which show some useful information:



The shape of this plot is just the shape of California. We found that the scatters in original plot are mess up. Therefore, we can use heat map to show the information. Then we can find that the houses located close to San Francisco, Los Angeles and the coastline tends to be more expensive.



This plot shows that the median income and the median house price has a strong positive correlation.



Meanwhile, from many plots we are aware of that some factors for example, the age have up limits, which can be a potential problem for data analysis.

Part2. Multi-collinearity

1) If we use VIF analysis without standarization, we find that there might be some problem at total_bedrooms and households. That makes sense because that the total_bedrooms must be related to total_rooms, and the same for household.

Furthermore, except x5 and x7 regressors, variance decomposition proportions π_{91}, π_{92}, π_{43}, and π_{64} exceed 0.5, which indicates the multicollinear relationship. But that also makes sense because that the geographical positions in general are very similar (all in California), and the house age cannot be too much different.

2) Then using Eigensystem analysis. After performing standarization, we obtain the eigenvalues= 3.91243965 1.92267742 1.69686478 0.91022813 0.29332659 0.14252168 0.06264480 0.04454859 0.01474837, and the condition number=265.2794, which is smaller than 1000. This indicates that **the multi-collinearity problem is not severe**.

Conclusion: We can assume that the **multi-collinearity problem is not significant.**

Part3. Multiple linear regression

After using training data set to it a multiple linear regression model (y=b0+b1x1+b2x2+b3x3+ b4x4+b5x5+b6x6+b7x7+b8x8), and the least-squares (LS) fitting we obtain the following indicators:

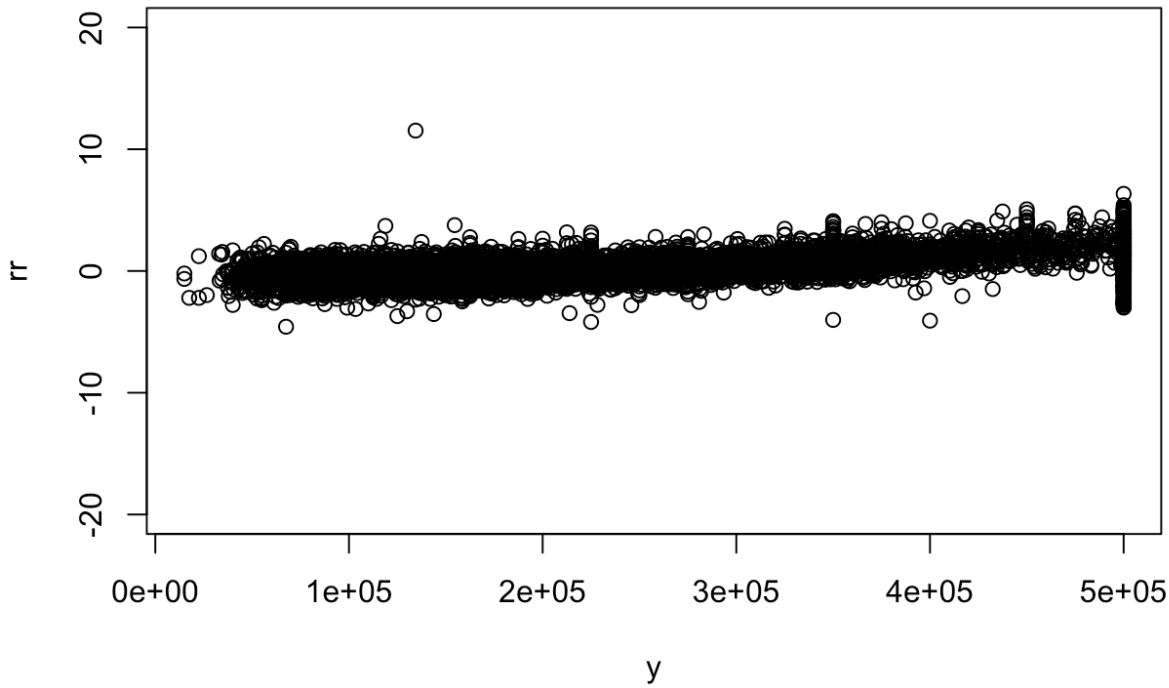
R^2 =SSR/SST=0.6380

R^2adj =1-(SSres/(n-p))/(SST/(n-1))=0.6378

test-MSE =4815181579 (if standardized, test-MSE=0.3619517)

Part4.Zero mean Gaussian i.i.d. assumption

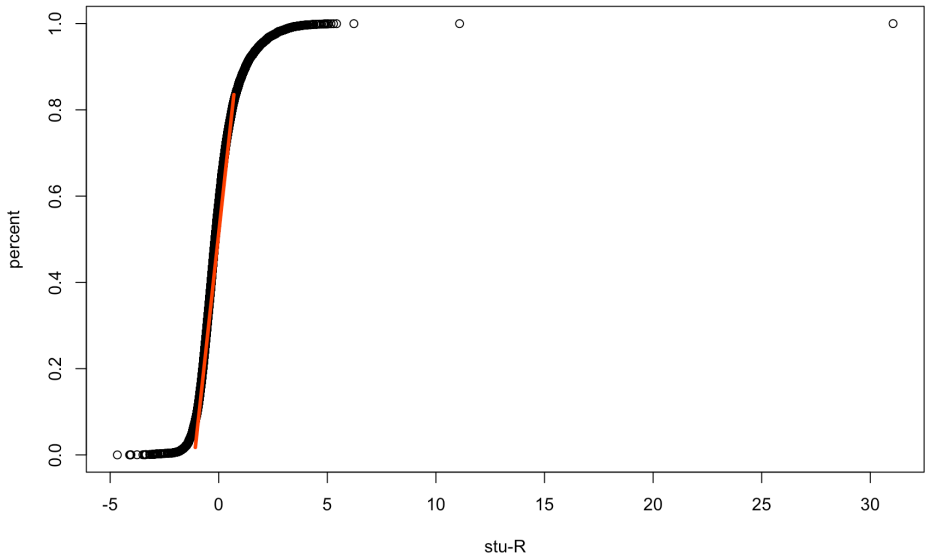
1)R-studentized residuals plot:



Almost every point is around 0. It seems that the tail seems to have some problem. I think it is because there are many up- limits in the original data(mentioned before), which makes the tail heavier. It should not be worried about.

Thus, we can assume that the errors are normally distributed.

2)normal probability plot:



It is obvious that the points are approximately on a straight line between around 0.33 to 0.67 points.

Conclusion:

Combing the results form 1) and 2), we can say that the multiple linear regression model is subject to zero mean Gaussian i.i.d. assumption.

Part5. Polynomial regression model

1/ For polynomial regression model, I choose K=2 through forward selection(the t test for a higher order term is non-significant). Meanwhile, after comparison, I choose interaction term x1*x2, x4*x5, x4*x6, x4*x8 poor their contribution to make R^2 larger.

2/Merits of polynomial regression model:

1. Polynomial regression model is better at dealing with nonlinear data compared with multiple linear regression. It is more flexible.
2. This model is fast to build and easy to comprehend. If there is some problem, we can easily find out what happened and adjust it.
3. Polynomial regression model is easy to visualize.

Demerits of polynomial regression model:

1. Polynomial regression can have overfitting problem or under-fitting problem.
2. It may be difficult to choose a proper K. We need some information when we decide the K.
3. This model can be hard to deal with highly complex data.

3/After the least-squares (LS) fitting, we obtain the following indicators(the regression function:y=b0+b1x1+b2x2+b3x3+b4x4+b5x5+b6x6+b7x7+b8x8+b12x1^2+b22x2^2+b32x3^2+ b42x4^2+b52x5^2+b62x6^2+b72x7^2+b82x8^2+ b012x1x2, b045x4x5,b046x4x6, b048x4x8):

R^2 =SSR/SST=0.6736

R^2adj =1-(SSres/(n-p))/(SST/(n-1))=0.6730

test-MSE =4342269565 (if standardized, test-MSE=0.3264034)

4/If we add an L_2 regularization term, $(\lambda/N)*\theta T\theta$ to cost function:

According to the Ridge regression, we obtain the function: θR=(XTX+λIp)^-1*XTy

In this model, the λ needs to be large. For example, λ=10000. If λ is too small(i.e. equals to1), there will be no significant impact. If λ is too big, the effect will be weaker.

L_2 regularization term successfully reduces the test-MSE.

Part6.NN model

1) key function:

1/ *J*=max(sc−*s*+Δ,0) #we want to minimize this loss function. Δ should be lager than 0 and we can simply let it equal to 1.

2/ model = Sequential()

add(Dense(30, input_dim=9, activation="relu"))

model.add(Dense(15, activation="relu"))

It is a stack of many layers.Using add() to pile up the model; Using an activation function "relu", which is a nonlinear function.

It is a nonlinear model because the activation function is needed to be nonlinear in order to get the closer result, as the model cannot be strictly linear. And the reason why the activation function is nonlinear is that we can only get linear results if we only use linear activation function.

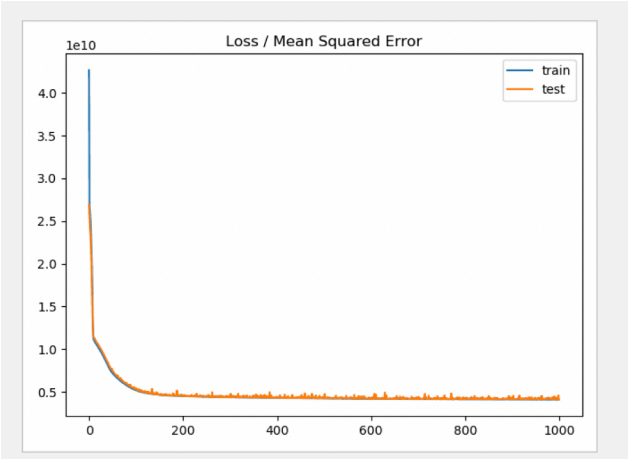
3/ θt+1=θt-α∇(θt)J

#The model parameters are trained to make the loss function reach the global minimum. A recursive way called gradient descent is used to reach the global minimum.

2) Results and Indicators:

```
 - 1s - loss: 4088497729.9845 - val_loss: 4587229819.0388
Epoch 1000/1000
 - 1s - loss: 4099421170.6047 - val_loss: 4132652597.5814
0.6932013524100009
4132652613.8647604
```

R^2adj =0.6961528059026643
test-MSE =4132652613.86 (if standardized, test-MSE=0.310349)



Compared with multiple linear regression model and polynomial regression model. NN model has better indicator performance. As the test-MSE are the best among these three model.

This is the plot of loss function. We can see the model is will fitted.

3) Merits of NN:

1. NN is much more flexible and has a strong ability to fit nonlinear model.
2. NN can be trained to study deeper, which can fit the model better.
3. NN can adjust itself when more data is added, and we do not need to care about the data structure.

Demerits of NN:

1. NN's speed is not fast as the linear regression model. It may take a longer time if we pile up more layers or training for more time(1000 times for me to do this task). The computing power is needed to be good enough.
2. If NN's result has some problem, it is very difficult to go back and find out what is wrong. For example, I once calculated R^2 equal to a large negative number which is impossible and i did know why.
3. Compared with the linear regression model, NN needs a lot of data to run. If the data is not big, the result will be weak.
4. NN may lose some data information and we usually can not be aware of the loss.

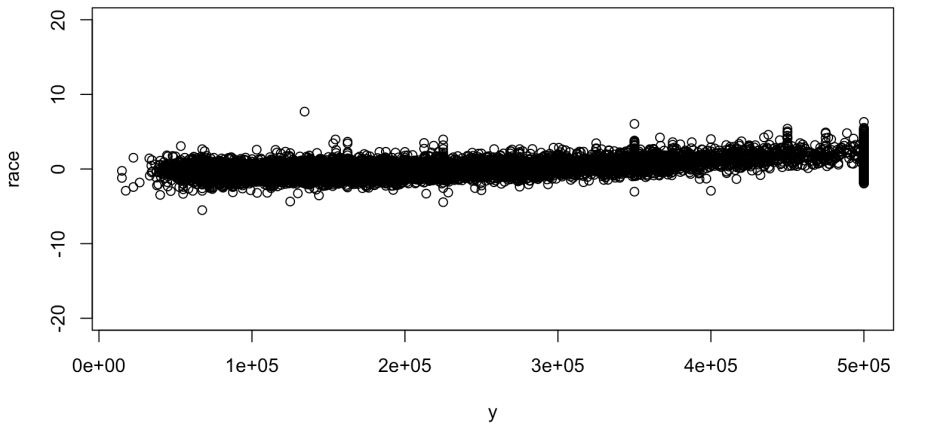
Part7. ACE model

1)After the least-squares (LS) fitting, we obtain the following indicators:

R^2 =SSR/SST= 0.6871

R^2adj =1-(SSres/(n-p))/(SST/(n-1))=0.6869
test-MSE =4132652613.8647604 (if standardized, test-MSE=0.3128461)

The R^2adj is better than those in multiple linear regression model and polynomial regression model, but is little worse than NN model. This is a good result. Meanwhile, the test-MSE is smaller than multiple linear regression model and polynomial regression model.



The residual plot above is almost the same with the Ordinary LS, but we can see that the points are more concentrate than the OLS. What's more, we can find the points with large absolute value of residual are more close to zero compared with the OLS.

As the ACE has better R^2adj, test-MSE and more concentrated data, we can say ACE is superior to the original multiple linear regression model.

2)Merits of ACE:

1. ACE is a non-parametric multivariate regression tool, uses an iteration with a variable-span scatterplot smoother to figure out the structure of your data, and is easy to evaluate surrogate models of data.
2. ACE provides a method for estimating maximal correlation between random variables, which means it can expose the relations between predictors and responses from complicated data sets.
3. This algorithm has good computer efficient.
4. ACE can incorporate variables of quite different type in terms of the set of values they can assume.
5. ACE can deal with the mixed-type variables.
6. It provides graphical output to indicate a need for transformation as well as to guide in their choice.
7. When there is a huge degree of association between predictor variables, ACE can help assist in assessing the variability

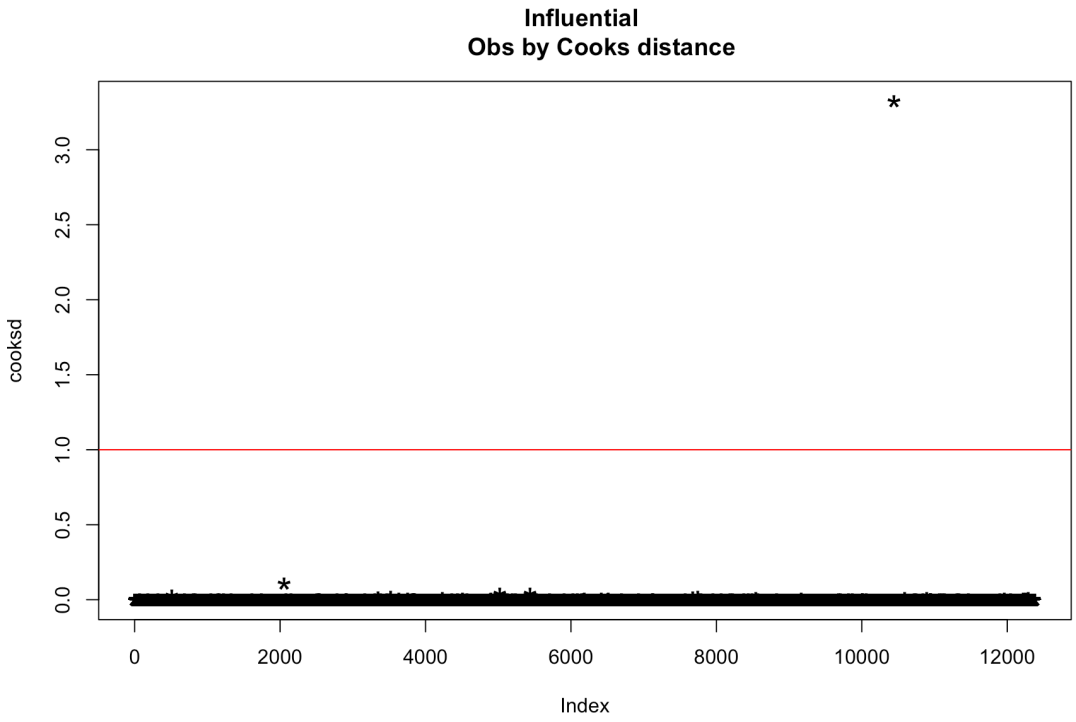
Demerits of ACE:

1. ACE is not intuitive like the OLS model.
2. ACE may lose some data information.

Part8. Outlier

From the r-student residual plot of the training set and the 9x9 plot in Data detail part, it can be found that there are several points which has relatively large r-student residual. Thus our first hypothesis is there may have outliers, but not much.

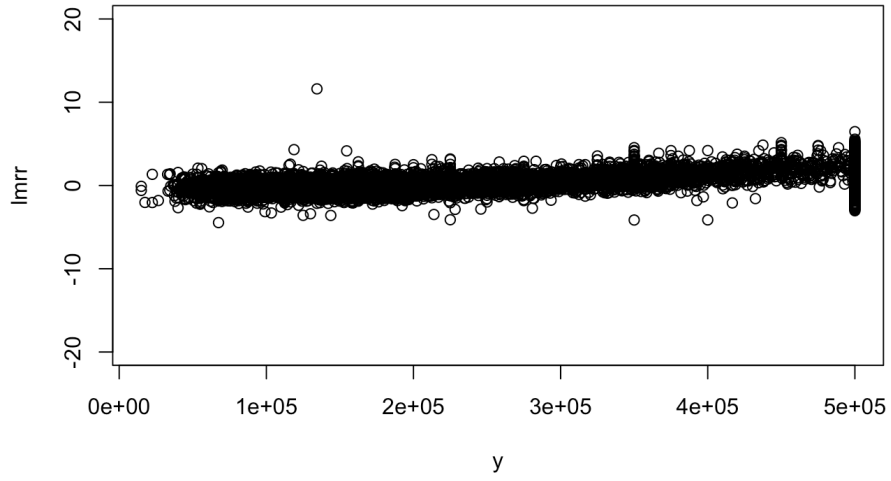
Then we use Cook's measure, and the obtain this plot:



The red line's h=1 because F0.5,p,n-p is approximately equal to 1 for moderate n and not too small p. This plot suggests that only one point is an influential point and is possibly to be an outlier. Assume the point whose Di>1 is an outlier, there will be only one outlier. Compared with the total training set which contains 12384 data, the fraction is very small(1/12384) and it can be regarded as no fraction.

Thus, I think it is not necessary to do robust regression, simply remove the outlier will be better as there is no group of outliers.

If we still do the robust regression(using Huber's t function here), we obtain this plot:



There is almost no difference between this residual plot and the one of the multiple linear regression model.

Part9. Summary

California Housing Price is a nice data set as it does not have significant multicollinearity problem and do not have many outliers. It gives us convenience to handle the data and complete the follow-up tasks.

From the data and the analysis, the median income has the highest correlation with the median housing price. Other regressors has relatively low correlation. Longitude and latitude together give us the location of the house.

Interestingly, the location is an abstract concept for the computer, and it is more likely to be logical indicators(whether the house is near the cities center, in big city, near the bay and so on) but not numbers. Although longitude and latitude are statistic significant, the correlation cannot be detected by simple linear model. As a common sense, people know that the location has great impact in the housing price. Therefore, a model which is more flexible and is good at handling nonlinear data can expose some information deeper and perform a better prediction result.

California Housing Price is a highly complex data set. There are too many factors that have great impacts on the housing median price. It cannot be perfectly explained through simple regression model. The previous models used are still too shallow and have serious error when compared with test set. Deeper learning should be done in order to obtain a better predict performance.

Part10. Comparison

On the aspect of training performance, the more simple the model is, the faster will the model finish training. Obviously, multiple linear regression model is the fastest. For the effect of training, the multiple linear regression model is relatively shallow. Polynomial regression model and ACE model are trained more(add more and higher order regressors, data transformation.etc) to fit the original data better. NN model uses the data information more fully and has better R^2 and R^2adj compared to previous models. Thus, NN model has the best training performance.

About prediction performance, the deeper the model training, the better will the prediction performance be. According to the R^2 , R^2adj and test-MSE from the previous tasks, we can rank from the best to the worst: NN, ACE, polynomial regression, and multiple linear regression. Therefore, regarding these four models, we can conclude that it is worthwhile to train more.

Multiple linear regression is very likely to meet underfit problem, as the data cannot be strictly linear. Thus, the error will be relatively larger than other models. For the polynomial regression model, it is possible to overfit if the K is not properly selected. Overfitting will meaninglessly make the model much more complicated. At the same time, the main feature of the data might be hidden.

If there is some problem in the model and we want to find out what is wrong and fix it, multiple linear regression model, polynomial regression model and ACE model is convenient to do so(we can adjust regressors, K, λ.etc very easily). Meanwhile, it is easy to check intermediate steps for these model. However, this can be very difficult for NN model. It is hard to explain why some problems occur and fix it.

The time complexity of NN model is quite different. The time complexity of linear model is $O(n^3)$, while for NN model, the expression of its time complexity contains many variables, which could significant affect the time complexity of NN model.

Compared with other three models, NN model needs more data. Otherwise, the performance would be bad. However, we cannot guarantee to gather the enough data. People have to choose the best model considering the quantity of data.