**Analysis of Real Estate Price Prediction**

Nancy Yang, Peiyu Zhao, Ray Zhu

**Abstract**

In the linear regression model, we will explore how the six factors in the data affect house prices. First of all, we need to determine which can be used as reasonable predictors. Among these six factors, the difference between latitude and longitude is not obvious, so we consider whether to delete them. Then, we use Normal Q-Q plot, Residual vs. Leverage, scale-Location and Residuals vs. Fitted to determine whether there is a linear relationship. Besides, we set null hypothesis and alternative hypothesis and choose to use ANOVA in our model. After outputting the F-value, we find whether we reject the null hypothesis by comparing the P-value and F-value. Finally, we assign data into the model to find the distance to MRT station, number of convenience stores and house age are more relevant to the price.

**Introduction**

We are going to use “R” to analyze the dataset 5: real estate price prediction. There are a few factors: transaction, date house, distance to the nearest MRT or convenience stores as well as longitude and latitude. For example, in England that Cardiff Bus network has positive impact on improving prices of nearby properties (Wang et al.,2014). We are going to find whether linear relationships exist between two or some factors. We are also curious how much influences of each or a group predictor to real estate price.

Basically, normal perspectives, the real estate price could be decided because of complex reasons form national economic, society and so on. Besides, today we do not consider specific individuals’ inquiry.

**Methods**

We firstly except longitude and latitude because these two factors look very stable when other factors are changing a lot from chart. We set hypothesis as: beta0=beta n. We put six factors to different group and build several models: Full models consist of all factors, error only, odd named factors, and even numbers factors. We want to know compared to full models, other three models’ coefficients are significant different or not. By using the Anova test to check the F-test value to find which factor is significant or having effect on our model. Also, in order for the reader to have a intuitive understanding, we have made several charts to let readers know how does they correlated to the model that we created.

X1.transaction.date

X2.house.age

X3.distance.to.the.nearest.MRT.station

X4.number.of.convenience.stores

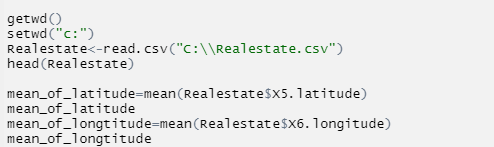
X5.latitude

X6.longitude

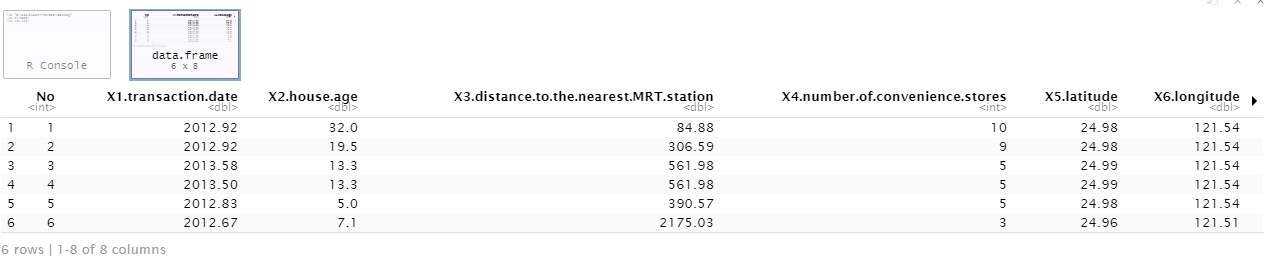
**Data Description**

Since this data set doesn’t give units of each charactors, we changed the unit of house price into dollar rather than one hundred thousand as well as the unit of house age into days rather than year. All of the changes of units are for the seek of a good understanding of the linear regression model created.

1.The first thing we want to do is to be clear which factor is reasonable to be a predictor.

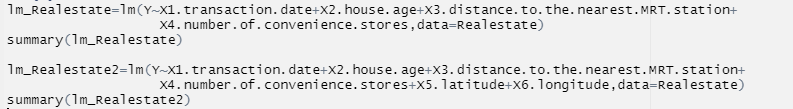


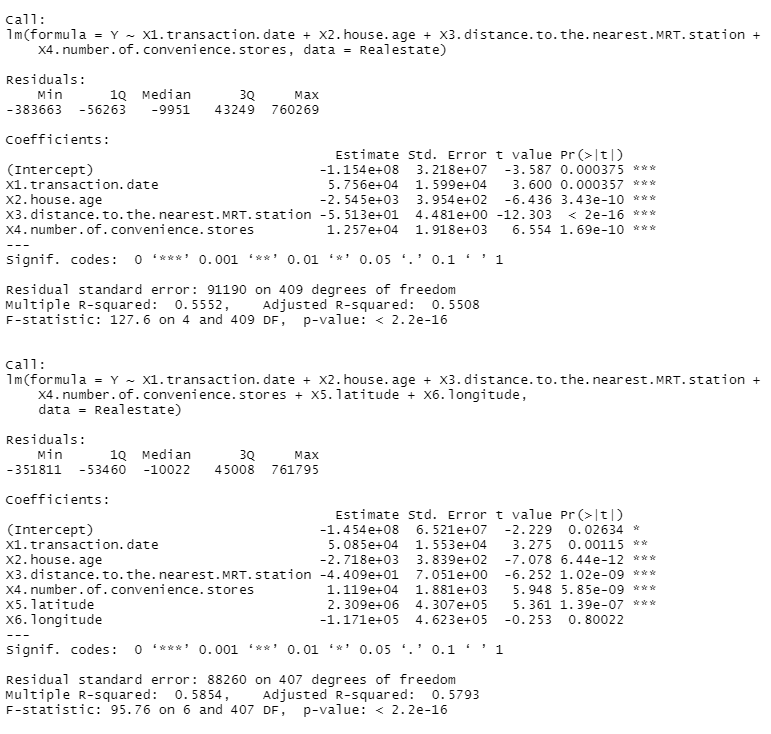
We can easily find that "latitude" and "longitude" these two contributors do not vary much over. (below)



As the X5 and X6 columns showed, we are going to ignore both characters first and it is been said that all the houses having roughly same latitude and longitude.

2. Seeing from the summary of the simple regression model that we created. The most essential thing we need to check is the correlation coefficient.



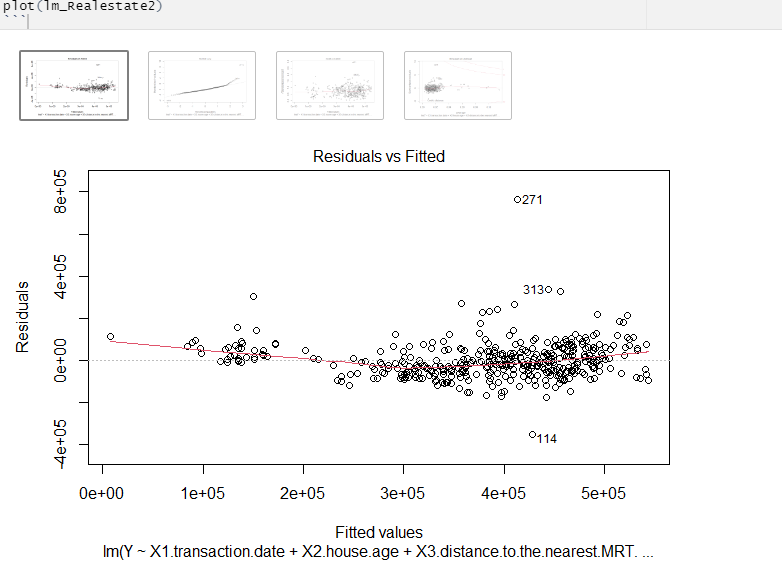


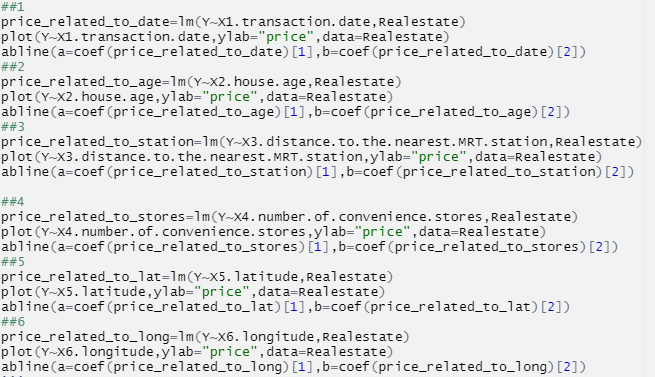
We find that with 4 predictors, The R^2 is 0.5553. Now we can see what happening if we put the remaining two predictors into model.

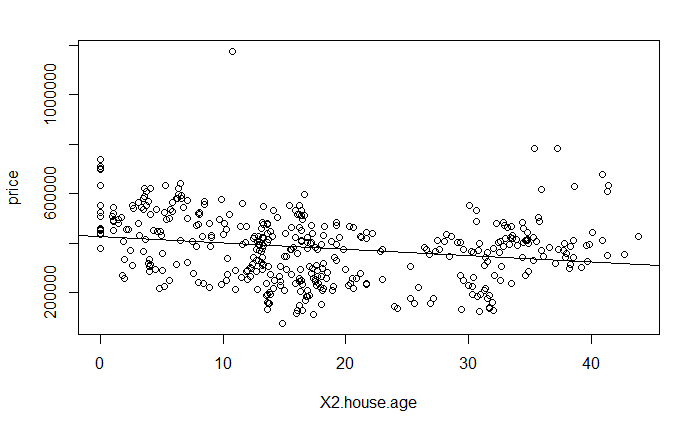
We surprisingly find that after putting these two predictors, the R^2 increased. So, it states that the conclusion we made before is wrong. As a result, we should take all factors into model!

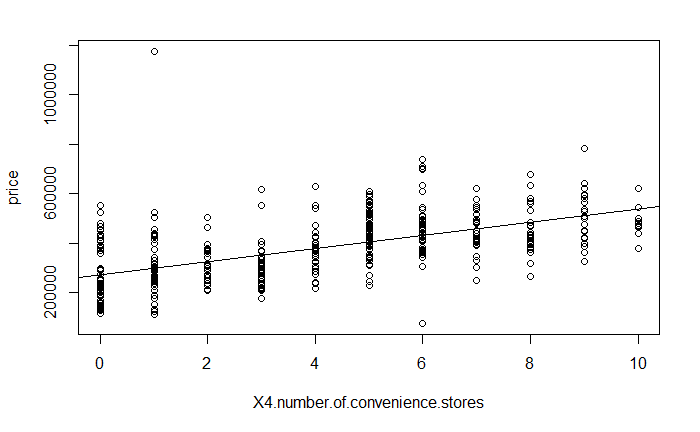
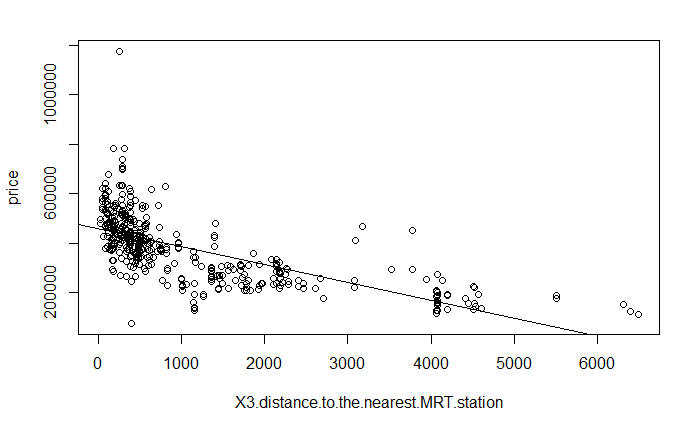
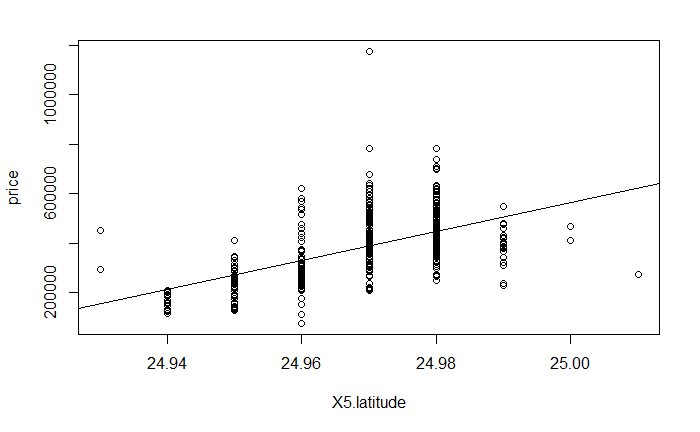
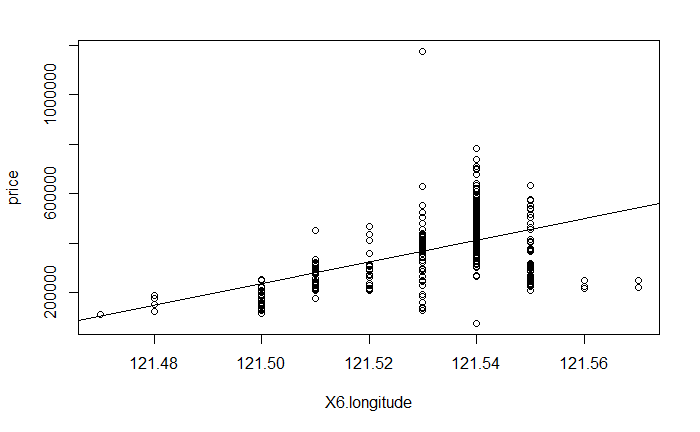
From the R^2, we know that almost 58% of variation in these predictors can be explained by our model.

3. We can check the regression diagnostic plots for this model. Followed by checking the fit-level of each predictors of six factors by using plot and albino command.





Six graphs clearly show that six factors are linear relate to real estate price.

We finished the process of prove the correlation coefficients.

4.Now we are going to the model selection to detach those unnecessary elements.

We are going to only analyze four models. Here we choosing to use the Anova to see the F test and find out which model is non significant compared to Full model.

If the model is nonsignificant, then we can not reject the NULL hypothesis(H0) and embrace with this model.

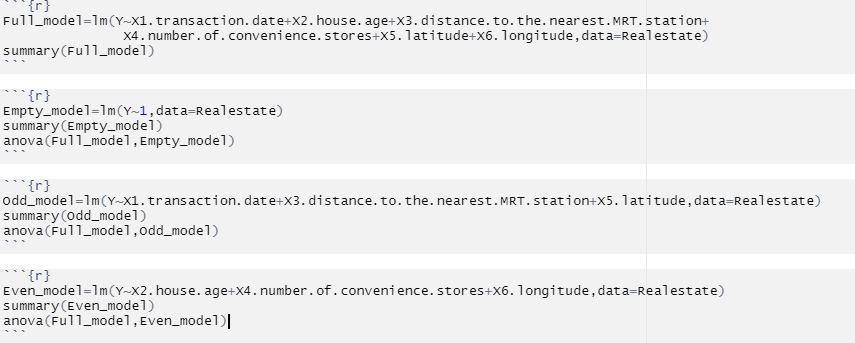
(1) Group1: contains all predictors, named full model.

Y= beta1\*X1+...+beta6\*X6+ERROR

(2) Group 2: just have an ERROR, called empty model.

(3) Group 3: The model has predictors from odd position which contains X1, X3 and X5

(4) Group 4: The model that contains predictors from even position: X2, X4, X6.



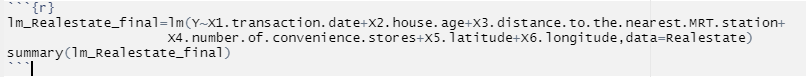
Seeing from the f test we find that these three models are all fail to reject the null hypothesis. So, we are continuing to use the full model.

So far, we have already provided the coefficient correlation relationship between factors and dependent variable (house price). Besides we are fully demonstrated the relationship of each factor’s contribution and effects.

5. Final model and Test.

Since we have already provided the coefficient correlation relationship between factors and dependent variable (house price). Besides we are fully demonstrated the relationship of each factor contribution and effects

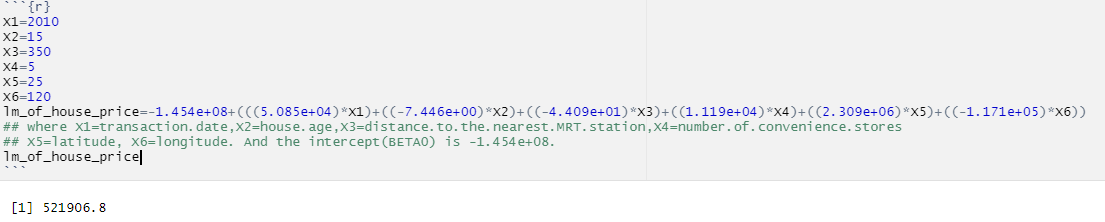
To conclude the final model that we created is



For the seek of an intuitive understanding we extract the BETA0 and BETA1 to BETA6.

Now, we can assign any value to the factors and it would return a satisfied result.

We estimate the price of NO.19 row with around value. BETA0(intercept) is -1.454e+08.



Compared to the data showed on csv file.



We are successfully estimate the price within acceptable error.

**Conclusion**

According to our linear regression analysis, we can conclude that these six factors all have a linear relationship with the house prices. It is worth mentioning that the age of the house, distance from the MRT station, and the number of nearby convenience stores have the greatest impact on the housing price, because in real life, convenient, fast travel mode and living environment save people time costs. In spite of the influence of latitude and longitude on housing prices is not particularly obvious, we cannot remove them. To sum up, convenient living locations for transportation and life tend to have high prices (Boyle et al., 2014).

References

Boyle, A., Barrilleaux, C., & Scheller, D. (2014). “Does Walkability Influence Housing Prices?” Social Science Quarterly.95(3), 852–867. Retrieved from https://doi.org/10.1111/ssqu.12065

Wang, Y., Potoglou, D., Scott, O., &Gong, Y. (2014). "Bus Stop, Property Price and Land Value Tax: A Multilevel Hedonic Analysis with Quantile Calibration." Land Use Policy: The International Journal Covering All Aspects of Land Use 42 (2015): 381-91. doi.org/10.1016/j.landusepol.2014.07.017