Sara Atbaei
Peirong Ji
Yuang Tian

Math3340 Final Project
Dec.11th/2020

Abstract

In this project, our focus is to evaluate the relationship between variables given in the dataset and examine if certain variables are directly and largely affecting the house price of the unit area. Firstly, we visualized the dataset that contains a total of 415 observations (414 given variables plus 1 new data point). Followed by the summary of methods we used to analyse individual variables. The methods that were used were Stepwise forward selection, backward selection and "both" directions, and Subset Regression. Results in both Backward and Stepwise methods are essentially identical, which suggests the longitude does not affect house prices as significantly as the other variables do. Meanwhile, in order to ensure the viability of our final model, it was tested in several ways under Gauss-Markov Assumption. Finally, a list of results, conclusions and limitations will be discussed in detail.

Introduction

The price movement of houses is much more popular in today's time. House prices are not only the economic reflection of a certain country, but also something very important in everyone's day to day life. A precise prediction on the house prices can benefit both buyers and sellers. Therefore, a house price prediction model is helpful for people to be educated on the house prices that are accurate to their market of houses. According to the research in *House price prediction: Hedonic price model,* by V. Limsombunchai, C. Gan, & M. Lee (1), the results suggest houses with more bedrooms and bathrooms are more expensive. A new house has a relatively higher price compared to an old house. Many recent studies from *Modeling spatial and temporal house price patterns: A comparison of four models* (2*), House selling price assessment using two different adaptive neuro-fuzzy techniques*(3), and *Housing price ´ prediction: parametric versus semi-parametric spatial hedonic models*(4) also justify that different locations of the houses could be a determining factor of the prices of the houses. For example, the houses are usually more expensive if they are surrounded by bus stops and convenience stores. The research by P. Linneman. in *An empirical test of the efficiency of the housing market* (5) found empirical evidence to prove the house price is closely related to a wide variety of economic variables such as people's income, cost of construction etc. From the results of all the research above, it is obvious to see there are numerous different factors that contribute to different house prices. Thus, we need to do a detailed analysis on the given dataset on real estate to determine which variables have the greatest correlation with the house price of the unit area.

Data Description

For our given Real estate data, we have 414 data points, for our new data point (data point 415) we took the average of the outliers. To do this we first found the Deleted Studentized Residual Vs Predicted Value plot, reference (figure 1.), which gives us a plot for our normal points and our outlier/influential points. After this we wanted to check the outliers

once again with the Cooks D bar plot because this also gave us our outlier/influential points in a bar plot which is seen in Figure 2. The threshold for this data in the cooks D bar plot is 0.01 meaning if the cooks distance for any of the data points are greater than .01 then it is an outlier. We then put in the rcode, cooks.distance(model), to get all of our outlier numbers which satisfy the threshold. These are the data points 36,48, 106, 114, 117, 127, 129,149, 165, 167, 221, 229, 271, 313, 345, 362, 383, 387, 390. To get the average of the outliers in each column, X1,…,X6, we have done so in Excel using the AVERAGE= command. This gave us our new data point which is X1= 2013.23668, X2=22.1210526, X3=1480.21462, X4=3.52631579, X5=24.9718626, X6=55.48947368. After entering this in the excel file we can get our pairs plot for the data with the new data point, reference Figure 3.

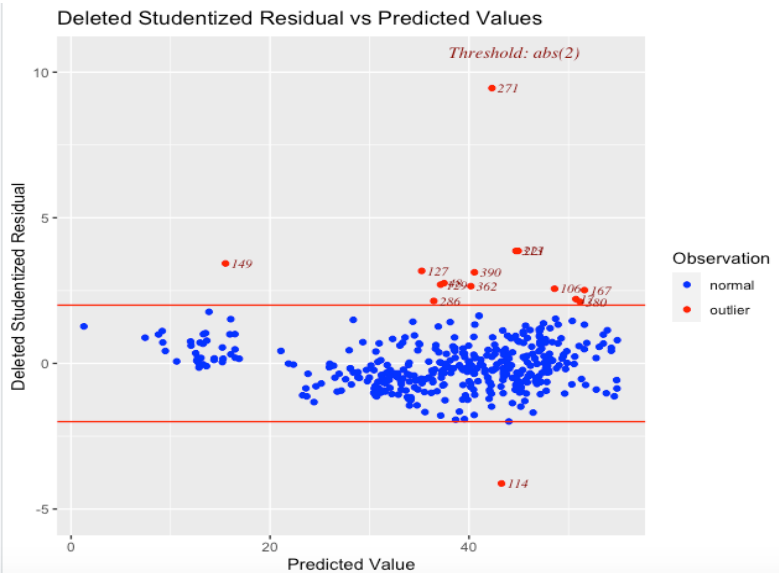

Figure 1: Deleted Studentized Residual Vs Predicted Value plot

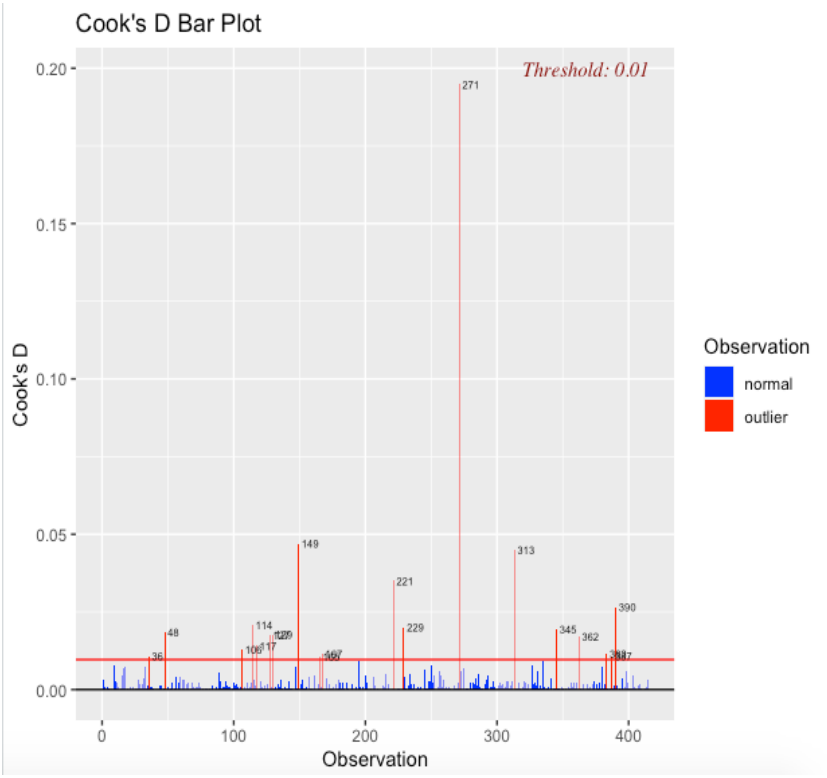

Figure 2: Cooks D bar plot

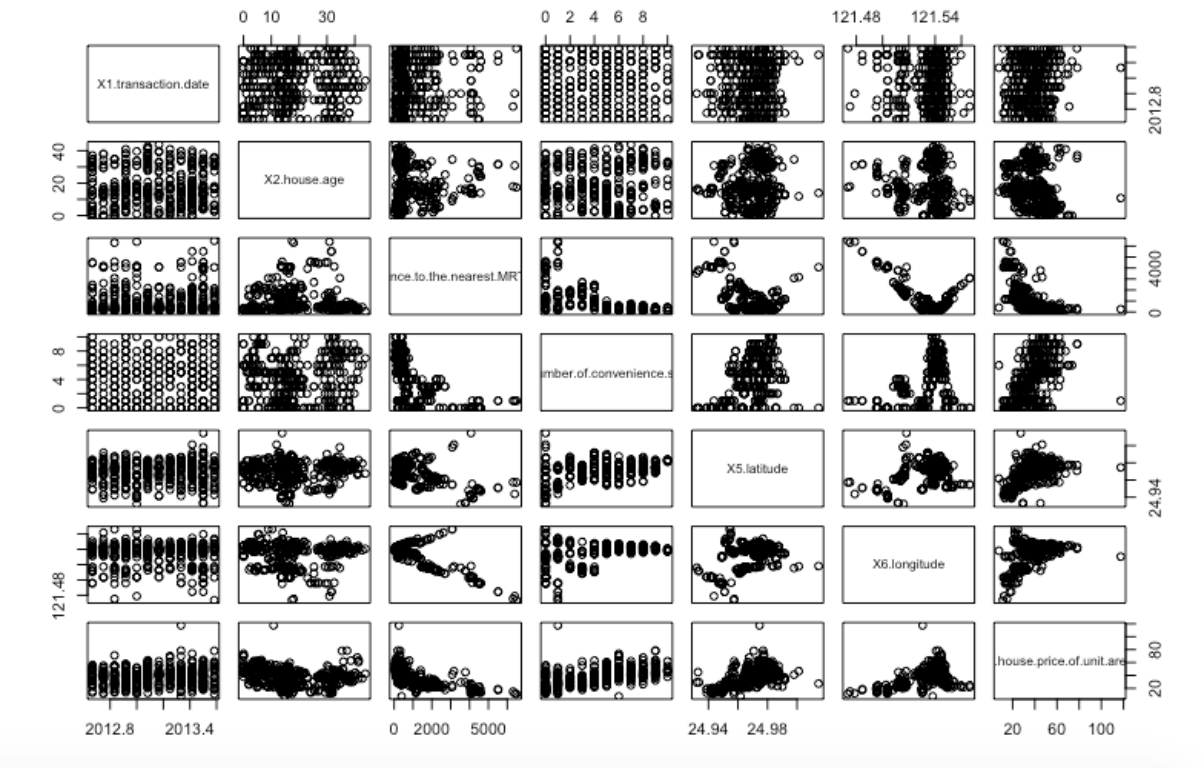

Figure 3: Pairs plot

Summary of Methods

Subset Regression Method

The first method we did for our fitted model is the subset regression method with leaps command in R. We used library("leaps") in Rstudio to perform all subsets regression. We set data=as.matrix(data[,-1]) and then perform the leaps rcode to find the adjusted R^2. This was performed with the following rcode, leaps(x=data[,-1],y=data[,1],method="adjr2"). We then get our output for *$adjr2* and pick the largest adjusted R^2. This would be number [17] which equals 0.0311625915. We then go back to the *$which* output from our rcode that we did and go down to the 17th line to find what our final model is. We do this method because, "According to the adjusted R^2 Criterion, the model chosen is the one with the largest adjusted R^2,"(Dr. Joanna Mills Flemming, Modelfitting, Dec.2nd,2020). For our Real Estate data, this is the model in the first row numbered 3,(T T F F F T), this is the model X1,X2,X6. After this we check to see if the model is satisfactory because this is not a reliable method to find the final model.

```
$adjr2
 [1]  0.0053892978  0.0013211241 -0.0007141037 -0.0011776491 -0.0021024780
 [6] -0.0023310475  0.0294634563  0.0134909088  0.0067490239  0.0054447359
[11]  0.0043577826  0.0032157397  0.0022802198  0.0001641967 -0.0008372914
[16] -0.0009189602  0.0311625915  0.0292624494  0.0275651900  0.0271993659
[21]  0.0130930695  0.0118818794  0.0110991343  0.0064989457  0.0053989303
[26]  0.0049898691  0.0301467981  0.0294489360  0.0288000458  0.0276588926
[31]  0.0269375121  0.0253226042  0.0121172810  0.0107220809  0.0095287961
[36]  0.0048910490  0.0286794931  0.0277813054  0.0270766112  0.0253488337
[41]  0.0097034199  0.0024978746  0.0262994347
```

```
$which
      1     2     3     4     5     6
1 FALSE FALSE FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE FALSE  TRUE FALSE
1 FALSE FALSE FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE FALSE FALSE
2 FALSE  TRUE FALSE FALSE FALSE  TRUE
2 FALSE FALSE FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE FALSE  TRUE FALSE FALSE
2 FALSE FALSE  TRUE FALSE FALSE  TRUE
2  TRUE FALSE FALSE FALSE FALSE  TRUE
2 FALSE FALSE FALSE  TRUE FALSE  TRUE
2 FALSE  TRUE  TRUE FALSE FALSE FALSE
2 FALSE FALSE FALSE  TRUE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE FALSE FALSE
2 FALSE  TRUE FALSE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE FALSE FALSE  TRUE
3 FALSE  TRUE FALSE  TRUE FALSE  TRUE
3 FALSE  TRUE FALSE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE FALSE FALSE  TRUE
3  TRUE FALSE FALSE FALSE  TRUE  TRUE
```

Stepwise Method

For the stepwise procedure/method we began with forward selection. This method we start with none of our variables, X1, X2, X3, X4, X5, and X6, and add on a variable every step until we get our final model, we add the variable with the smallest AIC. We do not add the variables with an AIC that is larger than the Step: AIC. First step in R for this was we identified what each variable equal: y=data [,7]; x1=data [,1]; x2=data [,2]; x3=data [,3]; x4=data [,4]; x5=data [,5]; x6=data [,6]. After this we set up our null and full models which are the commands, nullmodel=lm(y~1), and fullmodel=(lm(y~x1+x2+x3+x4+x5+x6)). The last two steps are setting up the stepwise procedure for forward selection and the summary output to give us our results. These are the commands: step1=step (nullmodel,

scope=list(lower=nullmodel), direction="forward"), and summary(step1).

The results for the stepwise forward selection

The first result for the stepwise procedure for forward selection was for the model y ~ 1, AIC=2168.41. The variable that was added for the next step/model was x3 with an AIC of 1921.9 which is less than 2168.41 and the smallest AIC out of all the variables.

The second result result for the stepwise procedure for forward selection was for the new model y ~ x3, AIC=1921.9. The variable that was added for the next step/model was x4 with an AIC of 1890.4 which is less than 1921.9 and the smallest AIC out of all the variables.

The third result for the stepwise procedure for forward selection was for the new model y ~ x3 + x4, AIC=1890.38. The variable that was added for the next step/model was x2 with an AIC of 1854.9 which is less than 1890.38 and the smallest AIC out of all the variables.

The fourth result for the stepwise procedure for forward selection was for the new model y ~ x3 + x4 +x2, AIC=1854.94. The variable that was added for the next step/model was x5 with an AIC of 1828.6 which is less than 1854.94 and the smallest AIC out of all the variables.

The fifth result for the stepwise procedure for forward selection was for the new model y ~ x3 + x4 +x2 + x5, AIC=1828.59. The variable that was added for the next step/model was x1 with an AIC of 1819.6 which is less than 1828.59 and the smallest AIC out of all the variables.

The final result for the stepwise procedure for forward selection was for the final model y ~ x3 + x4 +x2 + x5 + x1, AIC=1819.59. We do not add the last variable (variable 6) because it has an AIC of 1821.5 which is greater than 1819.59 so we have reached the cut off and cannot add anymore variables. In this case with our data for real estate, the only variable that wasn't included was x6 which is longitude.

The results for the stepwise forward and backward selection, direction "both"

The first result for the stepwise procedure for forward and backward selection was for the model y ~ 1, AIC=2168.41. The variable that was added for the next step/model was x3 with an AIC of 1921.9 which is less than 2168.41 and the smallest AIC out of all the variables.

The second result for the stepwise procedure forward and backward selection was for the new model y ~ x3, AIC=1921.9, The variable that was added for the next step/model was x4 with an AIC of 1890.4. which is less than 1921.9 and the smallest AIC out of all the variables.

The third result for the stepwise procedure for forward selection was for the new model y ~ x3 + x4, AIC=1890.38. The variable that was added for the next step/model was x2 with an AIC of 1854.9 which is less than 1890.38 and the smallest AIC out of all the variables.

The fourth result for the stepwise procedure forward and backward selection was for the new model y ~ x3 + x4+ x2, AIC= 1854.94, The variable that was added for the next step/model was x5 with an AIC of 1828.6 which is less than 1854.94 and the smallest AIC out of all the variables.

The fifth result result for the stepwise procedure forward and backward selection was for the new model y ~ x3 + x4+ x2 + x5, AIC= 1828.59, The variable that was added for the next step/model was x3:x4 with an AIC of 1783.9 which is less than 1828.59 and the smallest AIC out of all the variables.

The sixth result result for the stepwise procedure forward and backward selection was for the new model y ~ x3 + x4+ x2 + x5+ x3:x4, AIC= 1783.91, The variable that was added for the next step/model was x1 with an AIC of 17772.3 which is less than 1783.91 and the smallest AIC out of all the variables.

The seventh result for the stepwise procedure forward and backward selection was for the new model y ~ x3 + x4+ x2 + x5+ x3:x4 + x1, AIC= 1772.28, The variable that was added for the next step/model was x3:x5 with an AIC of 1763.5 which is less than 1772.28 and the smallest AIC out of all the variables.

The eighth result for the stepwise procedure forward and backward selection was for the new model y ~ x3 + x4+ x2 + x5+ x1+ x3:x4 + x3:x5, AIC= 1763.52, The variable that was added for the next step/model was x4:x5 with an AIC of 1728.3 which is less than 1763.52 and the smallest AIC out of all the variables.

The ninth result for the stepwise procedure forward and backward selection was for the new model y ~ x3 + x4+ x2 + x5+ x1+ x3:x4 + x3:x5 + x4:x5, AIC= 1728.31, The variable that was added for the next step/model was x3:x4:x5 with an AIC of 1719.3 which is less than 1728.31 and the smallest AIC out of all the variables.

The tenth and final result for the stepwise procedure forward and backward selection was for the new model y ~ x3 + x4+ x2 + x5+ x1+ x3:x4 + x3:x5 + x3:x4:x5, AIC= 1719.29, We stop here and have our final model y ~ x3 + x4+ x2 + x5+ x1+ x3:x4 + x3:x5 + x3:x4:x5, AIC= 1719.29 because x1:x2 with an AIC of 1719.3 is greater than 1763.52 and it is the smallest AIC out of all the variables.

Backward Method

The third method we used is Backward, and we performed it in two different ways with the same outcome. We use the function Summary() to compute the P-value of individual variables. The results of the first try-out are as follows, except for the P-value of Longitude, other P-values are extremely small. Thus, in this case, there is no need to determine a significance level used to compare with each P-value, because we can directly exclude the variable “longitude” since it’s very big.

In the second tryout, we use the function Summary() to compute the P-values without variable “longitude”, the result (House Price ~ Date, Age, MRT, Store, Latitude) suggests no more variables need to be excluded because all the P-values are smaller than a reasonable significance level. The Model we get is (House Price ~ Date, Age, MRT, Store, Latitude).

```{r}
summary(lm(HP ~ date+age+MRT+store+latitude+longitude, data=Real_estate_3))
```

```
Call:
lm(formula = HP ~ date + age + MRT + store + latitude + longitude,
    data = Real_estate_3)

Residuals:
    Min      1Q  Median      3Q     Max
-35.695  -5.452  -0.923   4.166  75.173

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.463e+04  6.808e+03  -2.149 0.032213 *
date         5.190e+00  1.565e+00   3.317 0.000992 ***
age         -2.682e-01  3.871e-02  -6.929 1.66e-11 ***
MRT         -4.459e-03  7.214e-04  -6.181 1.55e-09 ***
store        1.136e+00  1.891e-01   6.007 4.19e-09 ***
latitude     2.268e+02  4.478e+01   5.064 6.22e-07 ***
longitude   -1.181e+01  4.882e+01  -0.242 0.809034
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.901 on 408 degrees of freedom
Multiple R-squared:  0.5789,    Adjusted R-squared:  0.5727
F-statistic: 93.48 on 6 and 408 DF,  p-value: < 2.2e-16
```

```r
summary(lm(HP ~ date+age+MRT+store+latitude, data=Real_estate_3))
```

```
Call:
lm(formula = HP ~ date + age + MRT + store + latitude, data = Real_estate_3)

Residuals:
    Min      1Q  Median      3Q     Max
-35.655  -5.385  -1.035   4.183  75.318

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.608e+04  3.249e+03  -4.949 1.09e-06 ***
date         5.179e+00  1.562e+00   3.315 0.000998 ***
age         -2.679e-01  3.865e-02  -6.932 1.62e-11 ***
MRT         -4.332e-03  4.922e-04  -8.801  < 2e-16 ***
store        1.139e+00  1.885e-01   6.039 3.48e-09 ***
latitude     2.281e+02  4.439e+01   5.139 4.28e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.891 on 409 degrees of freedom
Multiple R-squared:  0.5788,    Adjusted R-squared:  0.5737
F-statistic: 112.4 on 5 and 409 DF,  p-value: < 2.2e-16
```

Results

The results for stepwise forward selection showed us that longitude does not affect / was not significant enough in our findings for real estate house selling. It was not an important factor to change the data, so it is not needed in the data based off of our forward selection findings. The variables that are significant to keep in the data for Y house price of unit area, are X1 transaction date, X2 house age, X3 distance to the nearest MRT station, X4 number of convenience stores, and X5 latitude. The final model for this was y ~ x3 + x4 +x2 + x5 + x1, AIC=1819.59. Meanwhile, our result in Backward also suggest that longitude isn’t affect house price significantly.

Our model suggests that in this specific dataset:

1. The closer to a nearby MRT station, the higher house price will be. The relationship is significantly demonstrated in the model.
2. The higher number of nearby convenience stores, the higher house price will be. There also are extreme cases that 0 nearby convenience stores with a high house price, but the majority of data point are confirming the positive association between the number of convivence stores and house price. It could be a mansion located in a relatively isolated area.
3. The majority of data points suggesting that new houses (low house age) typically has a higher house price, but there are extreme cases as well. It could be a 40 years old house but located in a very favourable downtown area.

4. The effect of latitude and transaction date on house price is less significant than factors discussed above.
5. The longitude is barely affecting the house price, because the longitude of all 415 data points is extremely close.

Test for Gauss-Markov Assumption

We used the function plot (lm (Model)) to generate 4 important graphs for the final model. In Residuals vs Fitted graph (figure 4), the redline and most of the points (residuals) are close to 0, thus the assumption of "expected value of the error term is zero" is met. Referring to the Normal Q-Q graph (figure 5), the model is very close to normally distributed, not a perfect normal distribution because of the existence of outliers and influential points. Scale-Location graph (figure 6) indicates the standard deviation of residuals are fitting the assumption, since when fitted value increases, the standard deviation of residuals are moving horizontally. By definition, leverage point has a value that bigger than (4/n), in this case is bigger than (4/415=0.009), and in Y-axis, there are several points is bigger than 4, so we identified several influential points such as 271, 313,149, etc. Details about these influential points will be discussed in the following section. By analysing the VIFs for the new model, the results suggest that there is no multicollinearity in the model.

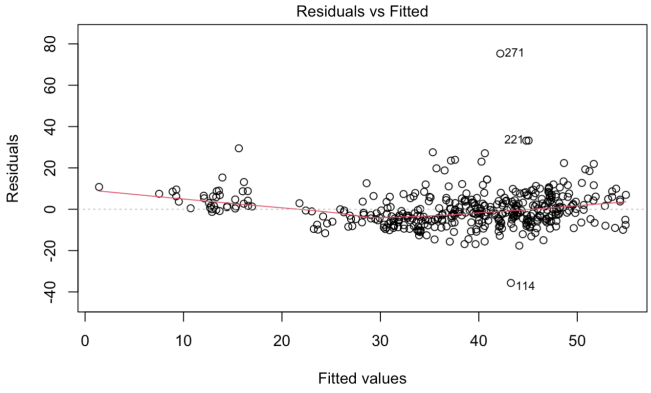

Figure 4: Residuals vs Fitted plot

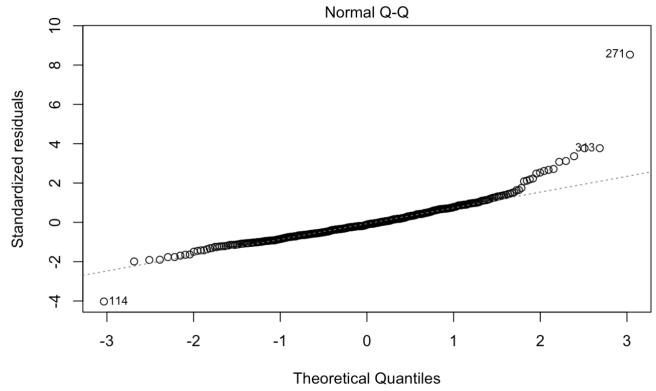

Figure 5: Normal Q-Q graph

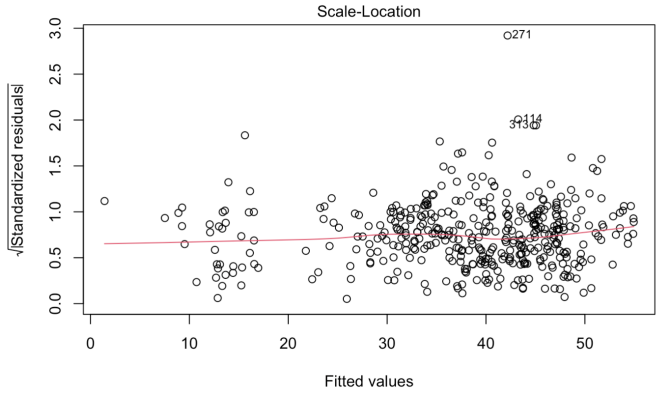

Figure 6: Scale-Location graph

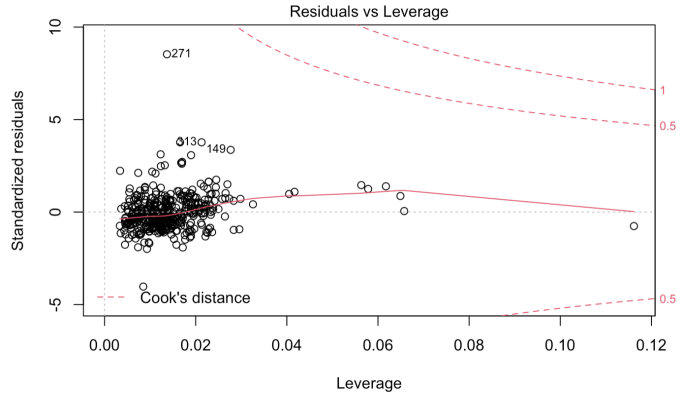

Figure 7: Residual Vs. Leverage

Influential Points

We have identified several influential points such as 271, 221, 114, 313, they will be discussed individually below.

Data	X1	X2	X3	X4	X5	X6	Y
271	2013.333	10.8	252.5822	1	24.9746	121.53046	117.5
221	2013.333	37.2	186.5101	8	24.98102	121.5365	40.2
114	2013.333	14.8	393.2606	6	24.96172	121.58	7.6

313	2013.583	35.4	318.5292	9	24.97071	121.54	78

Data 271: Everything else being equal, it has only 1 convenience store nearby while the price is high, could be a mansion located in a relative isolated area.
Data 221 and 313 are relatively old houses with higher prices, based on the number of nearby convenience stores they have, they most likely located in a city central area.
Data 114: A relatively young house has multiple convenience stores nearby, but have a very low price.

Conclusion

In our analysis, house age, distance to MRT stations and the number of nearby convenience stores are the top 3 factors that affect the house price of unit area. Indeed, in real life, a newly built house in a busy downtown is supposed to cost the most. However, we also identified several limitations of this model:

1. This model cannot be universally applied to the real estate market of different cities, states, countries. Because the effects of individual variables on house price may differ in other city, for example, in a geographically special area, longitude and latitude may become significant factors. Or in area that public transportation is poorly developed, the price will less likely impacted by the distance to nearby transportation stations or bus stops.
2. The longitude of all 415 data points is extremely close.
3. The time interval of transaction date is only a year, that is the reason why price upward movement is not obvious. In reality, if we apply a 10-year time interval, the effect of transaction date on house price will become more significant.

Additionally, there also many other important variables that will affect the house price:

1. The actual area of a house
2. Employment. Real estate expenses (rent or mortgage) rely on income, when employment fall and the house value fall too since purchasers are less financially capable to afford a house. For example, house prices spike in San Francisco since Silicon Valley offers most premium jobs in technology industry.
3. Interest rate. People who use mortgage to purchase a house will concern interest rate changes, increase in interest rate will decrease the amount that a purchaser can borrow when purchasing a house, and then affect purchasing decision.
4. Supply and demand also play an important role in property pricing, shortage plus a high demand will create high prices.

To sum up, there barely are “perfect models” that works for any situation, anywhere, anytime. In analysis, all variables are supposed to be independent, and needs for modifying the model often arises when we put the model in different scenarios.

Appendix

1. (V. Limsombunchai, C. Gan, & M. Lee. “House price prediction: Hedonic price model”. American Journal of Applied Sciences, 1 (3), 193-20, 2004)
2. (B. Case, J. Clapp, R. Dubin, and M. Rodriguez. “Modeling spatial and temporal house price patterns: A comparison of four models.” The Journal of Real Estate Finance and Economics, vol. 29, no. 2, pp. 167– 191, 2004.)
3. (I. H. Gerek. “House selling price assessment using two different adaptive neuro-fuzzy techniques.” Automation in Construction, vol. 41, pp. 33–39, 2014.)
4. (J.-M. Montero, R. M´ınguez, and G. Fernandez-Avil ´ es. “Housing price ´ prediction: parametric versus semi-parametric spatial hedonic models,” Journal of Geographical Systems, vol. 20, no. 1, pp. 27–55, Jan 2018.)
5. (P. Linneman. “An empirical test of the efficiency of the housing market”. Journal of Urban Economics 20(1986): 140-154, 1986.)