Group members:
1. So Tsz Hang(55880708)
2. Chan Ying Tung(56206549)
3. Chan Hiu Ching (56214931)
4. Ng Tin Lai (56227570)
5. LEE Nap Tak (56207718)
6. Tang Tsz Wun (56214273)

Title: Regression model on Hong Kong housing price

Abstract

According to the current social situation, housing price is a hot issue. When the housing demand increases, it affects the housing price directly. Since some of the housing rental prices cannot be found on the property website. Therefore, this project will only focus on how different factors affect Hong Kong property selling prices.

In this research, it is expected that saleable floor area *SFA* is the most effective factor on housing prices. We also expect that the height level (Height of building * Floor Level) and the number of rooms in the house have a positive relationship to the price of a house, while the age has a negative relationship. Living in some districts like central or facing South may have higher prices due to its convenience and comfort respectively.

Objectives

In this project, we would like to investigate which and how factors will affect housing prices. In order to find out which variable has the largest estimated effect, we would like to collect the data from the internet and develop a multiple linear regression model. We also aim to predict the future of house prices through the model. Furthermore, we would like to practice some code skills and concepts from lectures and notes in this project. For predicting house price, we would collect the housing price from the internet and other information which may be a variable in the model. In addition, we would like to set up the hypothesis test and check whether the assumption is correct with reality.

Background

Housing price is one of the hottest social issues in Hong Kong. Most of the Hong Kong city dwellers can not bear the housing price. The trend of housing prices has been increasing during the past years. Although the Hong Kong Government proposes many housing policies, for instance, providing more housing supply to Hong Kong citizens and developing more public housing to increase the housing supply. However, the house prices are still increasing. Therefore, we want to investigate which factor has the most influential effect on house prices.

The housing price has continuously risen during the past year. Some predicted that the housing price will still increase in the future. Considering the phenomenon in Hong Kong, due to the current situation, COVID-19 has limited the migration movement between Hong Kong and Mainland China. If the government resumes normal traveler clearance between Hong Kong and the Mainland, it will point to a sharp rise in housing demand. However, the housing supply cannot afford the demand, it leads to the high level of housing prices.

Motivation

Considering the high housing prices as an alarming problem, figuring out the most influential effect on housing prices could address the main problem. Thus the government could implement appropriate measures, such that the government, the stakeholders and the buyer (citizens) could strike a balance in order to solve Hong Kong's housing affordability issues. Besides, the prediction of a housing price in this model can used to be a tool for buyers by determining whether the house is underestimated or overestimated.

Data Collection

By studying the cases in the market, we can determine the variables that are worth analysis in the model. For the case in Grand Yoho, one case is on a higher floor facing West with Sea view selling at $9,900,000, while another case is on a lower level facing North with building view selling at $9,500,000.

Hence, we focus on the independent variables including saleable floor area, gross floor area, direction and numbers of room, living room and bathroom. Then, the dependent variable is the selling price. The above data are collected from the property websites including Midland Realty, Centaline Property Agency Limited and Ricacorp Properties Limited. Moreover, the number of primary schools are collected from the Education Bureau.

There are less significant efforts spent in collecting data, since the above properties websites can provide accuracy data. However, not all the data can be found on the property websites, such as rental price and direction. In addition, the data about the number of schools are counted by the Education Bureau. The above data should be credible and the error should be small.

Methods

First, we decide some variables which are essential to the house price. Using these parameters, we can utilize R to make an analysis of the variance table and a multiple regression model. Before finalising the mode, we would determine leverage point and remove the outliers, and also adjust by using transformation, *R²* and *F* -statistics. Finally, we would analyze the results by comparing different variables.

In the data, *HL=Height*Level* means the approximation of the height of the house. For example, high level house in a 36 storey building has *HL=0.75*36=27*. The paired plot (see Figure 1) is used to discover the pattern between two variables. The correlation and variance inflation factor are calculated (see Figure 2). Note that *SFA* and *GFA vif* are both high due to *SFA* and *GFA* are highly correlated, *cor(SFA,GFA) = 0.9790*. We pick the parameters *SFA*, *HL* and *Bathroom* to produce the simple model. *anova(m2,m1)* (see Figure 3) shows that the *RSS* only reduces slightly, which means we obtain a similar model with lesser predictors.

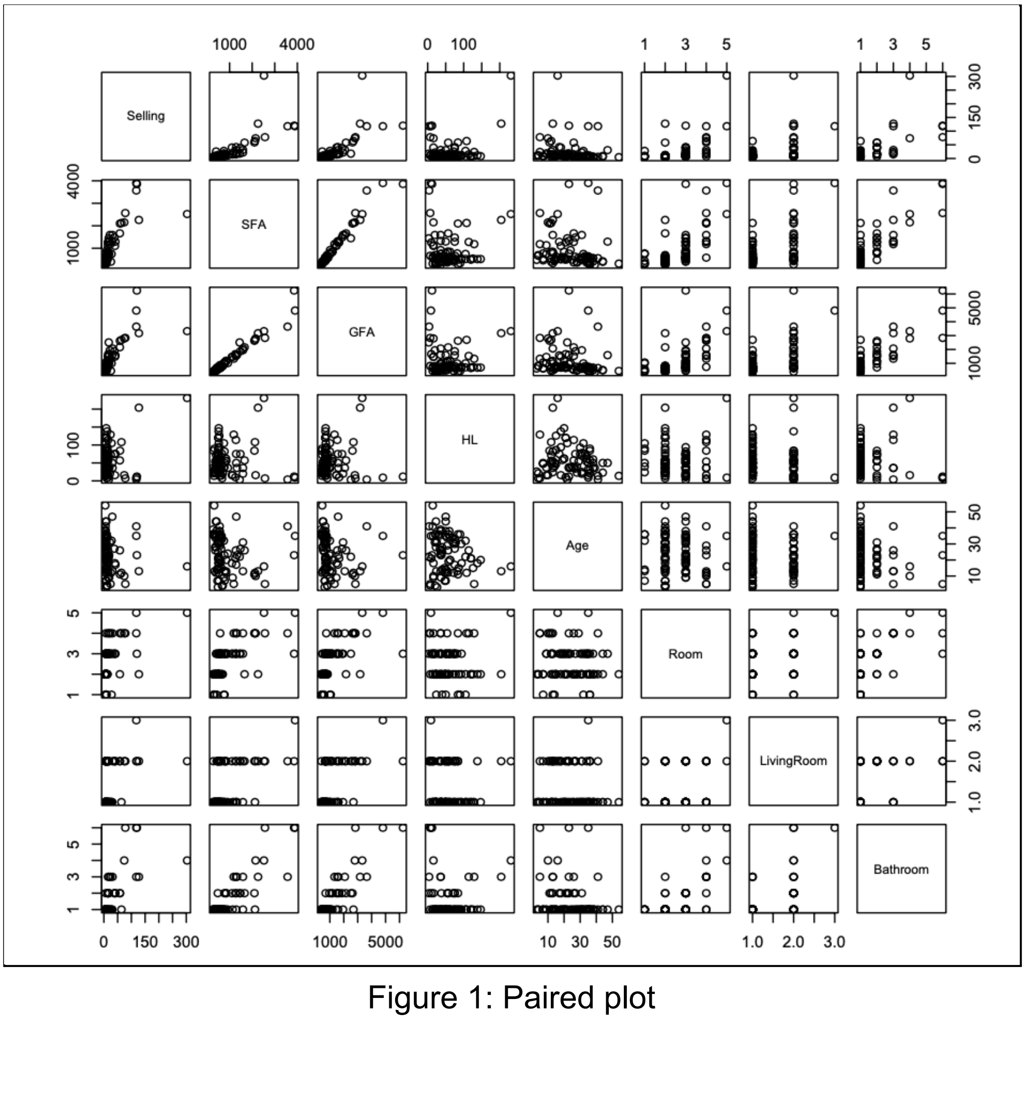
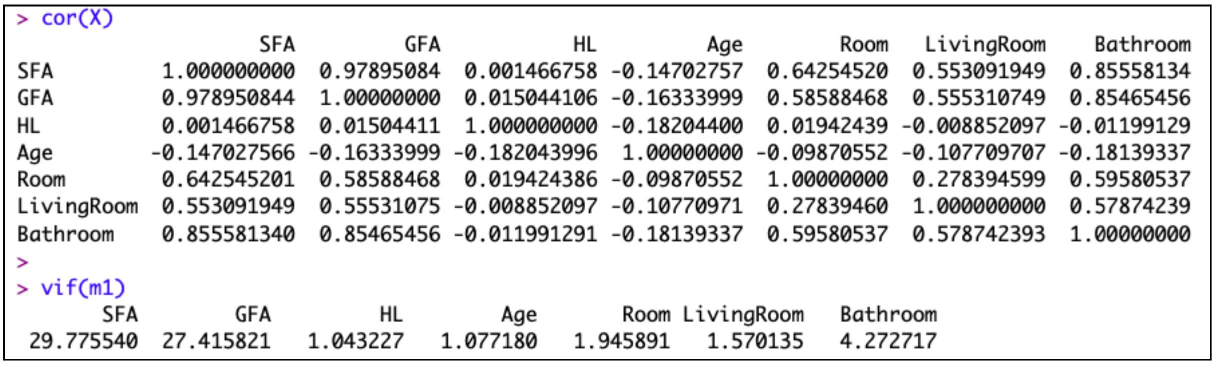
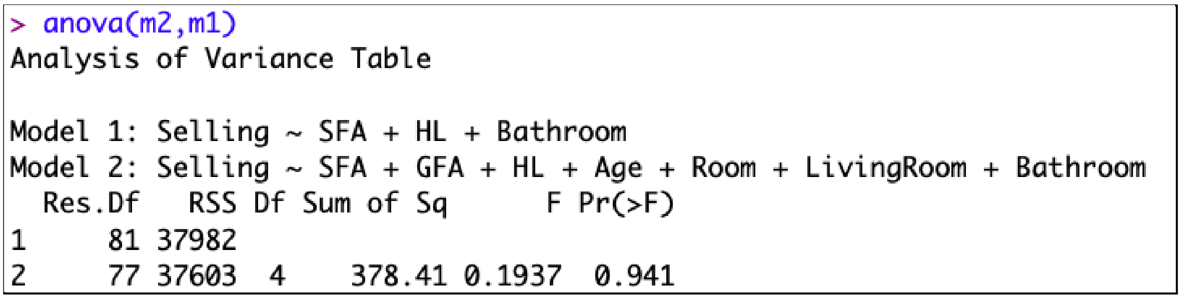


Figure 1: Paired plot



```
> cor(X)
                   SFA        GFA           HL         Age       Room   LivingRoom   Bathroom
SFA         1.000000000 0.97895084  0.001466758 -0.14702757 0.64254520  0.553091949 0.85558134
GFA         0.978950844 1.00000000  0.015044106 -0.16333999 0.58588468  0.555310749 0.85465456
HL          0.001466758 0.01504411  1.000000000 -0.18204400 0.01942439 -0.008852097 -0.01199129
Age        -0.147027566 -0.16333999 -0.182043996 1.00000000 -0.09870552 -0.107709707 -0.18139337
Room        0.642545201 0.58588468  0.019424386 -0.09870552 1.00000000  0.278394599 0.59580537
LivingRoom  0.553091949 0.55531075 -0.008852097 -0.10770971 0.27839460  1.000000000 0.57874239
Bathroom    0.855581340 0.85465456 -0.011991291 -0.18139337 0.59580537  0.578742393 1.00000000
>
> vif(m1)
      SFA       GFA        HL       Age      Room LivingRoom   Bathroom
 29.775540 27.415821  1.043227  1.077180  1.945891   1.570135   4.272717
```

Figure 2: *cor(X)* and *vif(m1)*



```
> anova(m2,m1)
Analysis of Variance Table

Model 1: Selling ~ SFA + HL + Bathroom
Model 2: Selling ~ SFA + GFA + HL + Age + Room + LivingRoom + Bathroom
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     81 37982
2     77 37603  4    378.41 0.1937  0.941
```

Figure 3: *anova(m2,m1)*

Note that there is some pattern in the residual plot of *m2* (see Figure 4), which indicates a wrong model has been fitted, thus we try transformation. InverseResponsePlot(m2) gives λ=*0.3760*, we pick λ=*1/3* and produce the transformed model m3=lm(formula = (Selling)^(1/3) ~ SFA + HL + Bathroom). The pattern of the residual plot solved a little bit in *m3* (see Figure 4).

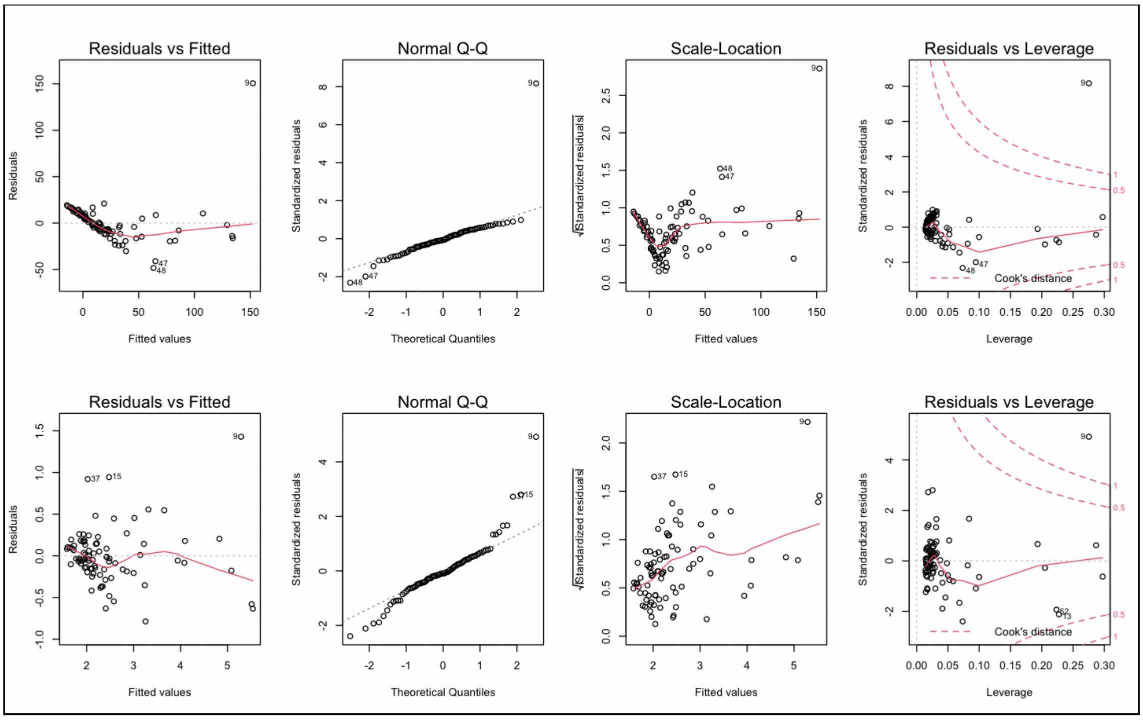


Figure 4: plot(m2) and plot(m3)

We create hat value vs standardised residual plot to identify the outlier (see Figure 5), find that case *6,9,12,13,19,24,47,51,52* are leverage points, where case *9* is a bad leverage point, thus we disregard case *9*. In model m4=lm(formula = (Selling)^(1/3) ~ SFA + HL + Bathroom, subset=(1:85)[-9]), the *Bathroom* parameter's *p*-value = *0.822* is now very high (see Figure 6). The reduced model m5=lm(formula = (Selling)^(1/3) ~ SFA + HL, subset=(1:85)[-9]) solved the issue (see Figure 7), with increased adjusted R square and F-statistic. The reason is the removed case has 4 bathrooms, which is abnormal.

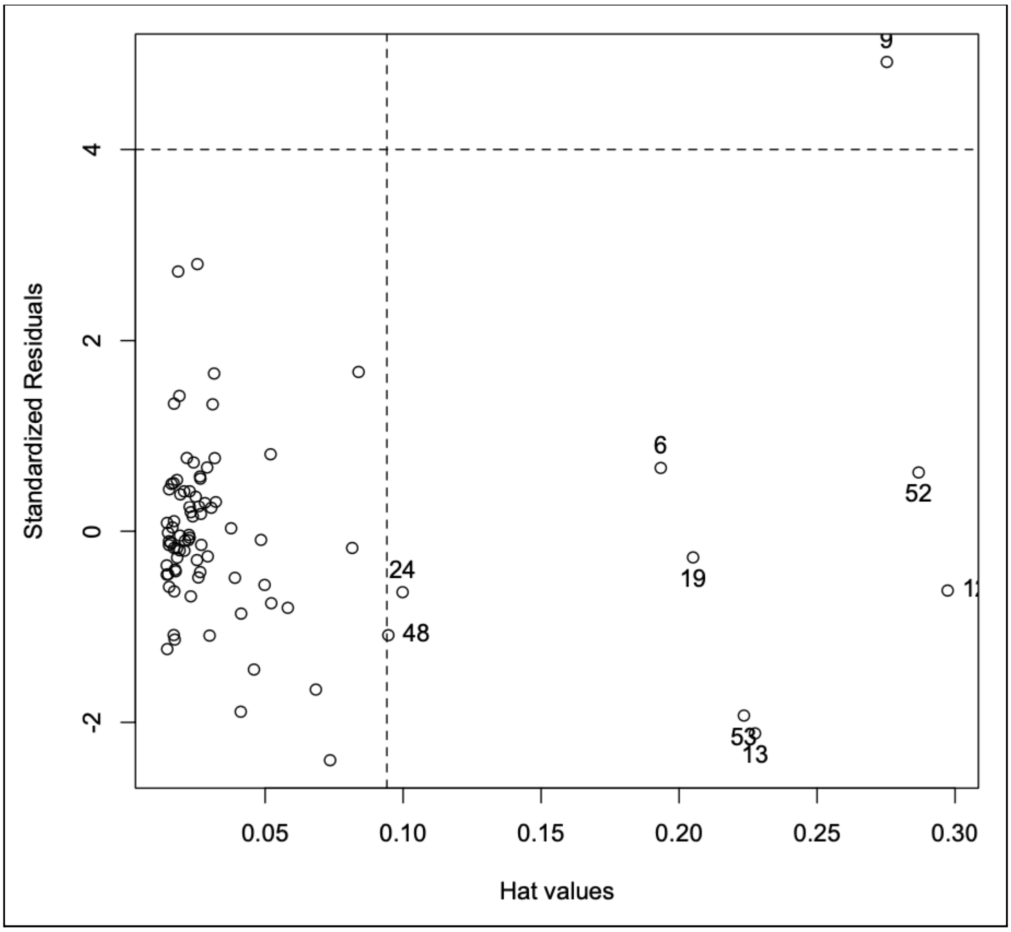
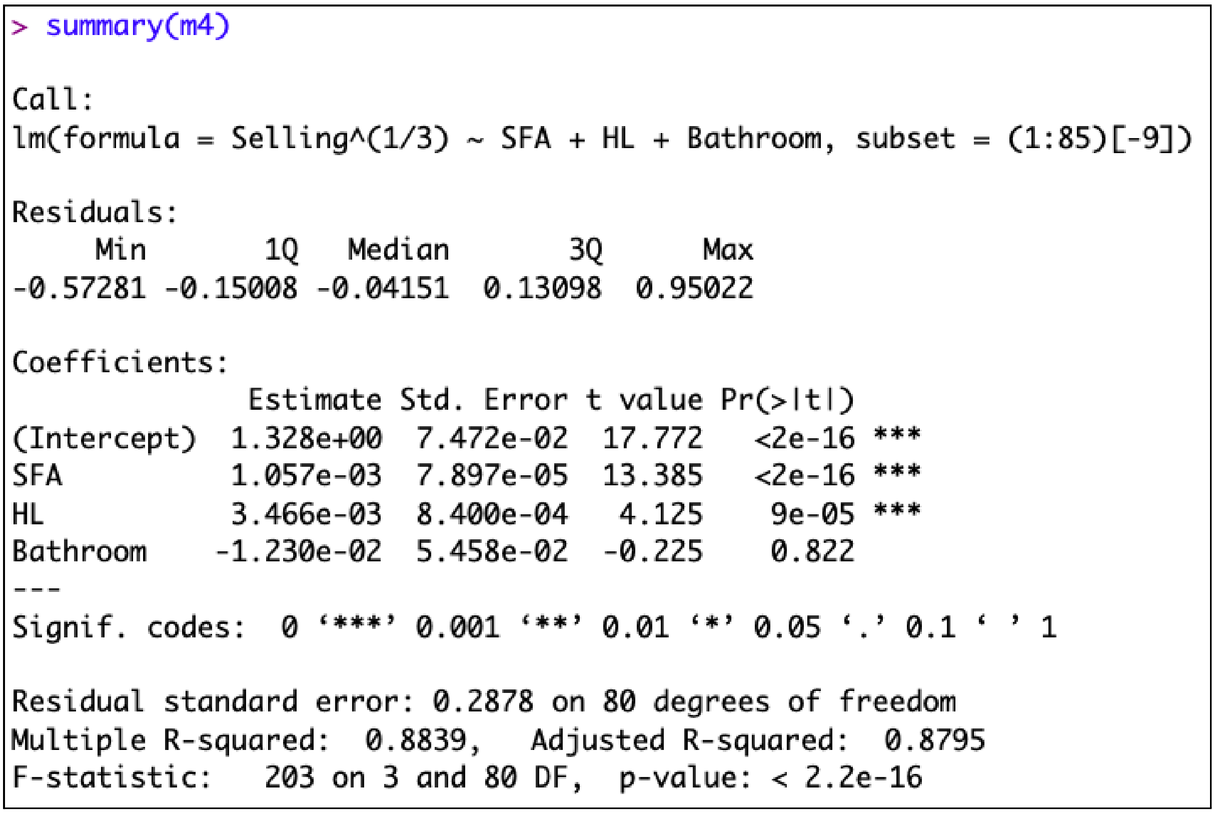
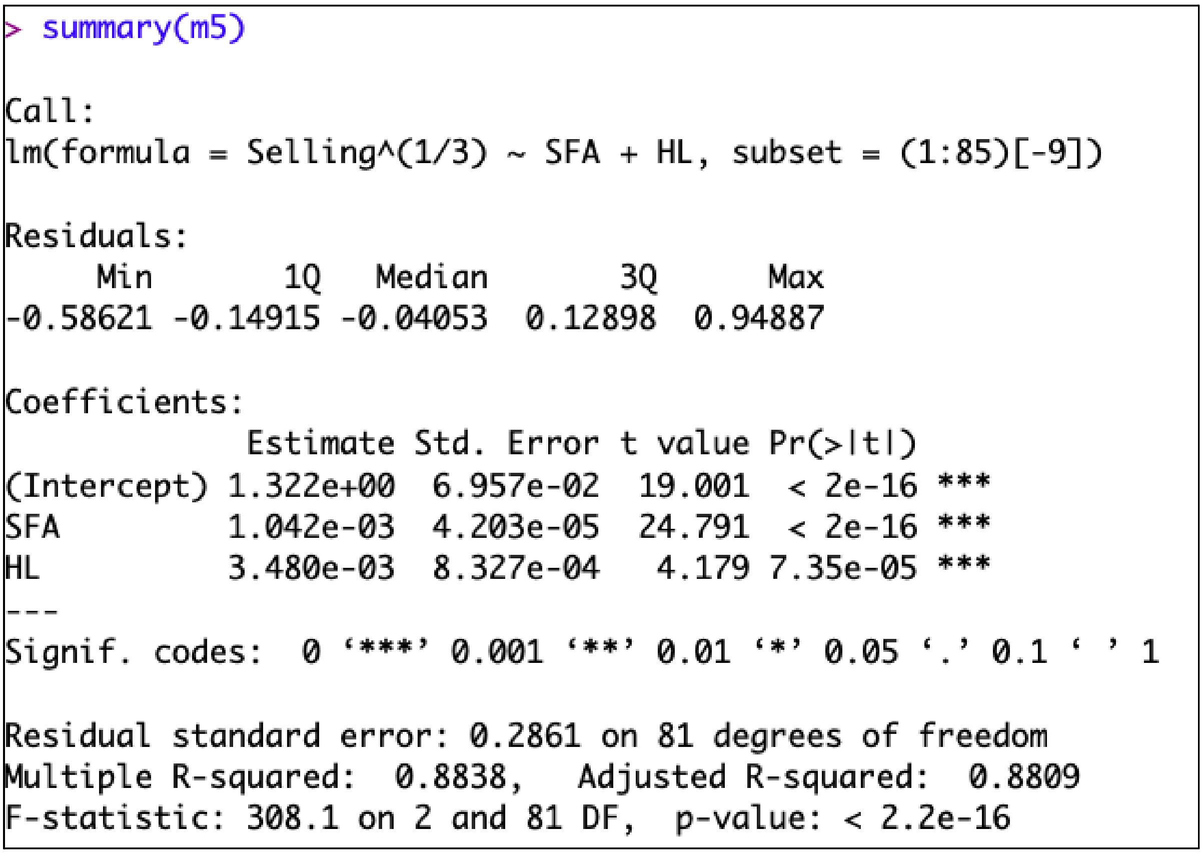


Figure 5: Hat value vs Standardised residual plot



```
> summary(m4)

Call:
lm(formula = Selling^(1/3) ~ SFA + HL + Bathroom, subset = (1:85)[-9])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.57281 -0.15008 -0.04151  0.13098  0.95022 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.328e+00  7.472e-02  17.772   <2e-16 ***
SFA          1.057e-03  7.897e-05  13.385   <2e-16 ***
HL           3.466e-03  8.400e-04   4.125    9e-05 ***
Bathroom    -1.230e-02  5.458e-02  -0.225    0.822    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2878 on 80 degrees of freedom
Multiple R-squared:  0.8839,	Adjusted R-squared:  0.8795 
F-statistic:   203 on 3 and 80 DF,  p-value: < 2.2e-16
```

Figure 6: summary(m4)



```
> summary(m5)

Call:
lm(formula = Selling^(1/3) ~ SFA + HL, subset = (1:85)[-9])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.58621 -0.14915 -0.04053  0.12898  0.94887 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.322e+00  6.957e-02  19.001  < 2e-16 ***
SFA         1.042e-03  4.203e-05  24.791  < 2e-16 ***
HL          3.480e-03  8.327e-04   4.179 7.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2861 on 81 degrees of freedom
Multiple R-squared:  0.8838,	Adjusted R-squared:  0.8809 
F-statistic: 308.1 on 2 and 81 DF,  p-value: < 2.2e-16
```

Figure 7: summary(m5)

In summary, the *SFA* and *HL* are significant (see Figure 8), Figure 9 shows the invention from *m1* to *m5*.

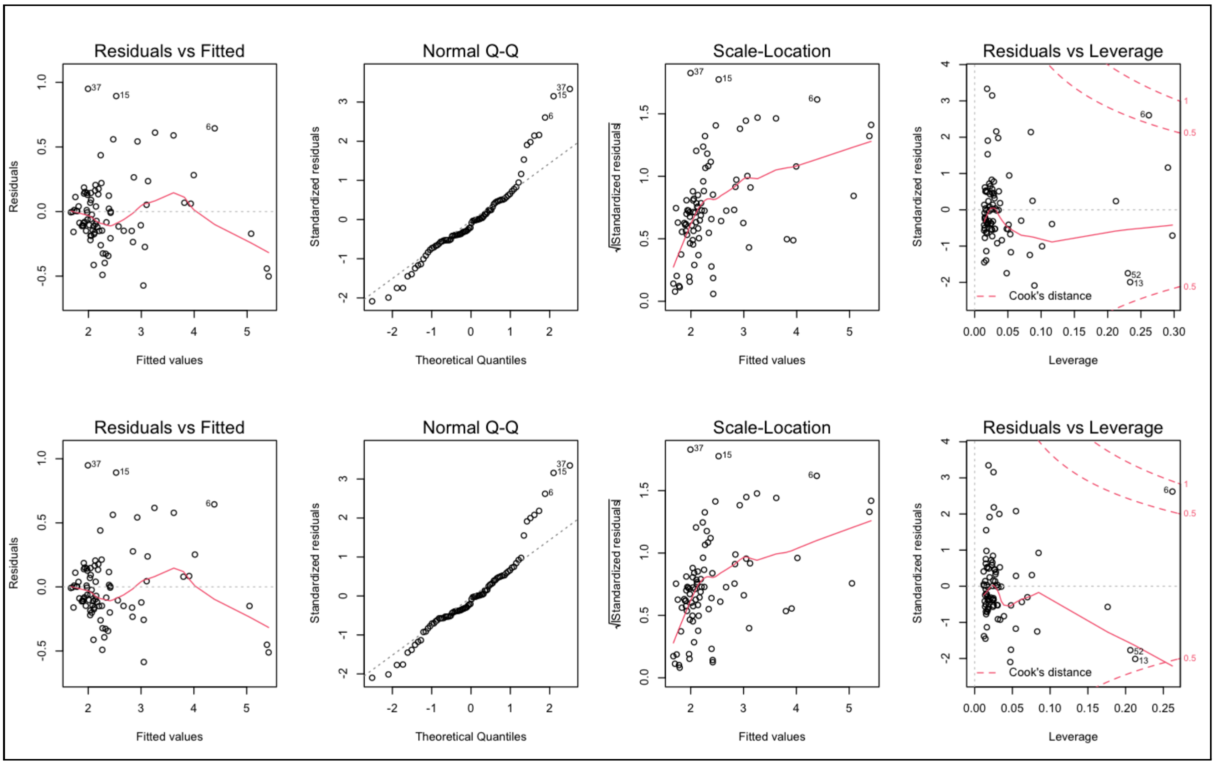


Figure 8: plot(m4) and plot(m5)

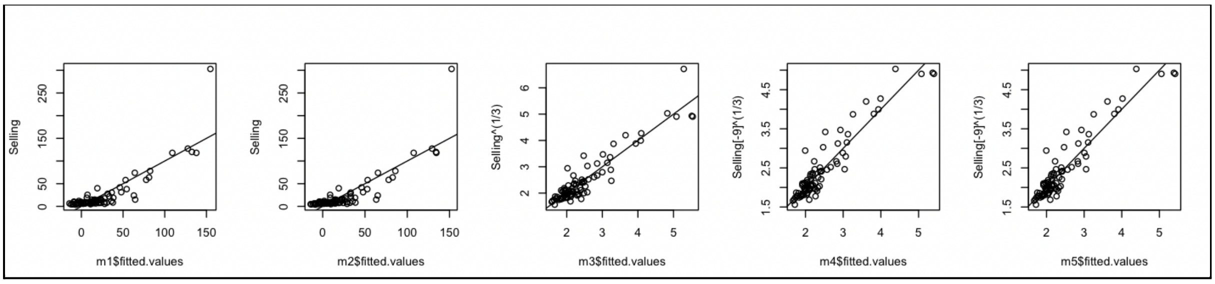


Figure 9: Selling vs Model's fitted value plot

Comparison

We created new models by adding dummy variables to the simple model *m0* one by one.

1. m0=lm(formula = Selling^(1/3) ~ SFA + HL ,subset=(1:85)[-9])
2. mDistrict=lm(formula = Selling^(1/3) ~ SFA + HL + HK + K, subset=(1:85)[-9])
3. mDirection=lm(formula = Selling^(1/3) ~ SFA + HL + N.S + E.W, subset=(1:85)[-9])
4. mHill=lm(formula = Selling^(1/3) ~ SFA + HL + Hill, subset=(1:85)[-9])
5. mSea=lm(formula = Selling^(1/3) ~ SFA + HL + Sea, subset=(1:85)[-9])

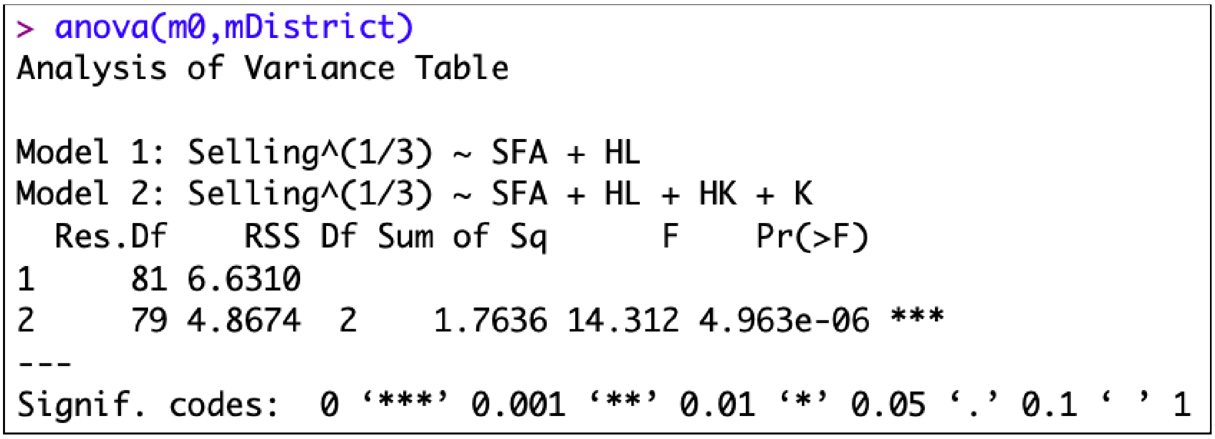
In the simple model *m0*, we have R^2=0.8838 and *adjusted* R^2=0.8809.

In model *mDistrict*, dummy variables *HK* and *K* represent whether the house is located at Hong Kong Island or Kowloon respectively, “*HK=0* and *K=0*“ means the house is located in New Territories. Note that the p-value of *HK* and *K* are very small, *1.17e-06* and *0.0136*, with estimation *3.675e-01* and *1.674e-01*. Indicate the parameters have a positive effect on the housing selling price. At the same time, the model *mDistrict* gives a better fitted model, with increased R^2=*0.9147* and *adjusted* R^2=0.9104.

In model mDirection, *N.S* represents the direction of the houses towards North (*N.S=1*) or South (*N.S=0*), similarly for *E.W*. The p-value of *N.S* and *E.W* are *0.192* and *0.512* respectively, which is high. Also the *F*-statistic drops to *155.1*, we conclude the direction of a housing does not have a significant effect on the selling price.

Similar situations appear in the models *mHill* and *mSea*. The *p*-value of the dummy variables are *0.161882* and *0.219028*. Models’ *adjusted* R^2 are close to the simple model *m0*, however result in significant drops in the *F*-statistics, *208.6* and *207.3*. Therefore we decide to not to use both *mHill* and *mSea* models.

In summary, we choose *mDistrict* as the final model, the anova(m0, mDistrict) agrees by giving a very small *p*-value, and a significantly dropped *RSS*.(see Figure 10)



```
> anova(m0,mDistrict)
Analysis of Variance Table

Model 1: Selling^(1/3) ~ SFA + HL
Model 2: Selling^(1/3) ~ SFA + HL + HK + K
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     81 6.6310                                  
2     79 4.8674  2    1.7636 14.312 4.963e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: anova(m0,mDistrict)

Result

The prediction of house prices is vital to driving real estate efficiency. Previously, house prices were merely calculated by the variable of selling prices in a locality. However, more factors will be preferred if we desire to estimate the payment precisely. Therefore, the house price prediction model is a great way to fulfill the information gap. With this model, we will be able to foresee the reasons for gaining the ultimate trend of high property prices as well as various difficulties encountered during our investigation.

In the procedures of data collection, one of the obstacles is the consistency of the data archive. Sometimes, each column of information such as the renting price may not be obtained regarding the selling houses. Under this circumstance, we may try to avoid using this variable to form a graph. Furthermore, linear regression is sensitive to outliers. According to Figure 5, we discovered that case 9 is a bad leverage point. Hence, we need to remove it. In this case, a few errors will happen if we consider the standardized residual which indicates that the point will be rejected if it falls outside the interval form -4 to 4. This situation cannot exclude other errors occurring during our research, causing the wrong decision of rejecting non-outlier data points. Consequently, the linear regression will be skewed away, affecting the true underlying relationship.

Another issue is about the estimation of primary school distance. For the distance data, we obtained all of them from real estate websites. However, all of the numbers are just the approximation rather than the exact value. Thus, this may also bring out the problem of underperformance. Additionally, it requires all variables to be multivariable normal. In our project, we have transformed the normal data to normal by adopting the method of log transformation. It may also lead to the effect of multicollinearity.

Conclusion

To sum up, before generating the regression model, this project assumes the saleable floor area is the most influential factor that affects housing prices. Although some leverage point and outlier exist, after removing the leverage point and the conduct the R process, it proves that saleable floor area *SFA* and *HL=Height*Leve*l are significant in this model. It is because the difference between multiple R^2 and *Adjusted* R^2 is small.