- **Problem Statement and Data Selection**

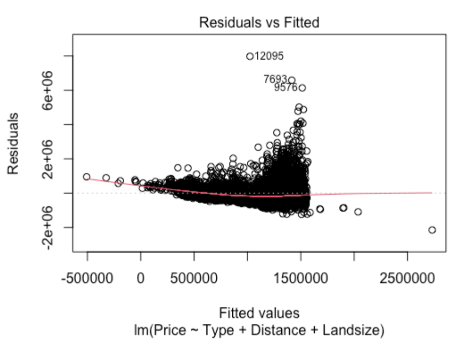
I want to compare how different variables influence the housing price of Melbourne. Initially, the chosen variables are Price, Type, Rooms, Distance, Bedroom2, Bathroom, Car, Landsize and Building Area; however, after the test of multicollinearity and autocorrelation[1], the variables Bedroom2 and Rooms have been excluded. I also cleaned NA values so that total obs dropped from 13580 to 6840.In terms of Type, I used dummy variable in my regression models since it is categorical. I used hieratical model selection to compare different models, and I also excluded BuildingArea when I did my regression because of the issue of homoscedasticity which I would talk about in detail in the next section. After getting the R Squared values for three different regression models, I have found that R Squared values have increased as I added more variables. Then, I did ANOVA test for checking the significance of them.

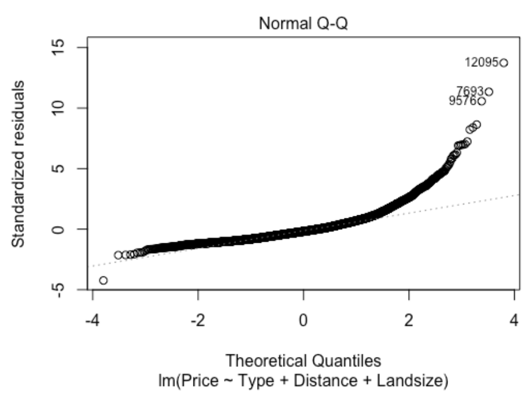
- **Planning**

Firstly, I was wondering if there were any possible outliers and influential cases in my selected variables because these variables include all the variables used in my hieratical regression later.[2] I calculated the standardized residuals based on the regression model which I already did the assumption check, finding out that there were total 231 possible outliers out of 6840 observations, which was about 3 percent, meaning that we do not need to worry too much about outliers. By calculating and plotting Cooks' distance[3], we can see that there were only few influential cases which we do not need to concern too much because it naturally existed. Then, I did hieratical regression and ANOVA test using data excluding BuildingArea because of disturbing the assumption of homoscedasticity.

- **Analysis**

My base model includes the variables of Price, Type, Distance and Landsize. I visualized it to test assumptions of normality, linearity, and homoscedasticity.



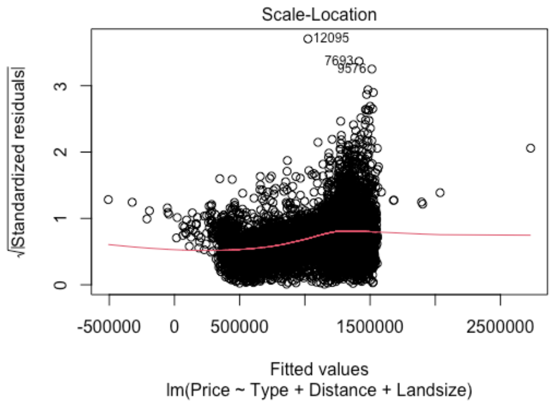


---

[1] The Statistic of Durbin Watson test is 1.436407. It is fine to accept, but the statistics of VIF for Rooms and Bedroom2 are 12.250642 and 11.961701 separately, recommending dropping.

[2] The chosen variables I used here are Price, Type, Distance, Bathroom, Car, Landsize and Building Area.

[3] See the plot in Appendix.



By summary, the house types, h and t[4], by comparing with u, are all positively affected price. H is more influential than other types according to the coefficient. Distance is negatively correlated with price, meaning that increasing the distance from CBD would decrease price. Landsize is the least one affecting price with the coefficient 55.761. All coefficients are statically significant by checking p values which are less than 0.05. R Squared value is 0.2564, meaning that it explains 25% of the variability in the model. Obviously, the price could be correlated with more variables. Then, I tested my second model, by adding additional variable Bathroom.[5] The assumptions have been tested fine.[6] The additional variable Bathroom generated positive correlation with price. Notably, the R Squared Values have increased to around 42 %, meaning that this regression model explains more variability of factors impacting price. Thirdly, I tested additional regression model by adding a new variable Car.[7] The assumptions are fine.[8] It generated positive significant correlation with price as well. More importantly, R Squared Values have increased to 43%, meaning that this model with additional new variable better explained more variability than the previous ones. Lastly, I did ANOVA test[9] to check if the variances are statistically significant. According to p values which are all less than 0.05, they suggest that the increased R Squared values are statistically significant.

- **Conclusion**

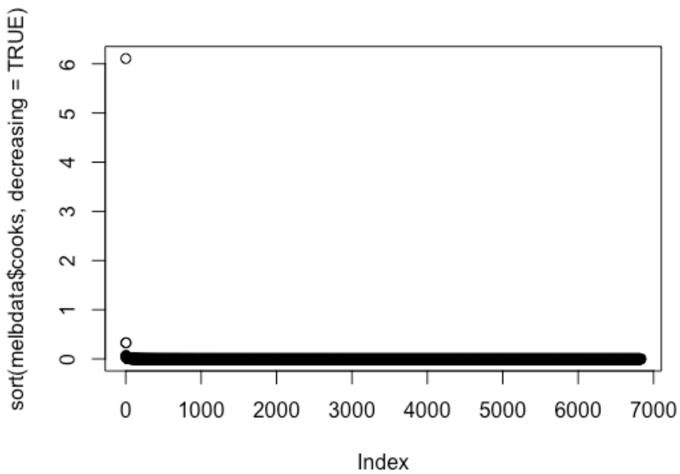
In this analysis, I successfully tested my model selection method and discovered several variables which could affect my dependent variable price. By adding more variables separately, R Squared values have increased and all the coefficients were statistically significant. By doing regression analysis, all the selected variables were correlated with price. Among these variables, I have found that house types could be the most important factor which impact price by comparing the coefficients in the three models.[10]

---

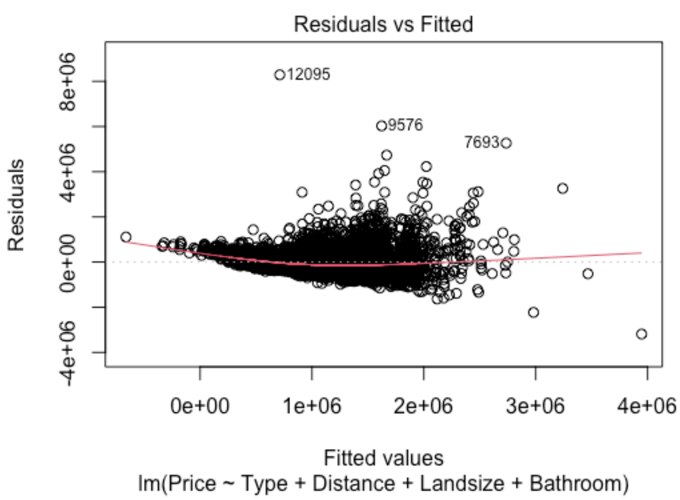
[4] The letter h represents house, cottage, villa, semi and terrace, u represents unit, duplex, t represents townhouse.
[5] Which represents the number of bathrooms.
[6] See Appendix.
[7] Car represents the number of car spots.
[8] See Appendix.
[9] See Appendix.
[10] The changes in price for h versus unit have been the most in the three models. See appendix.

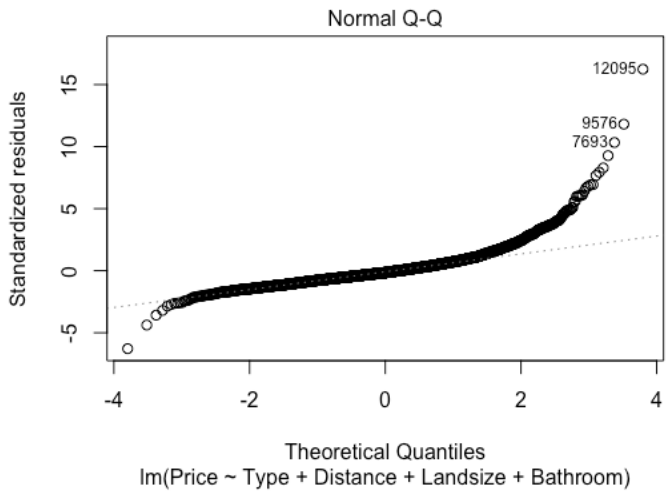
Appendix

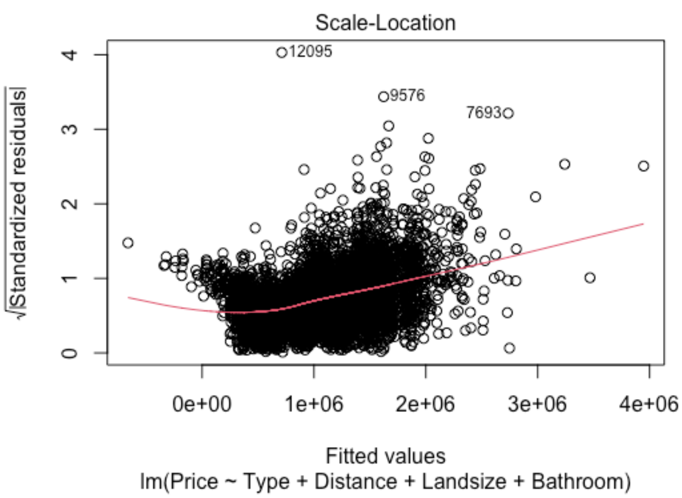
Cook's Distance



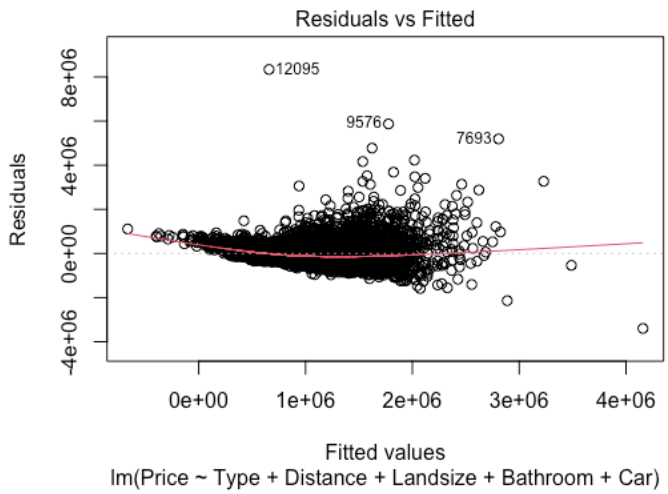
Second Model Assumption Check

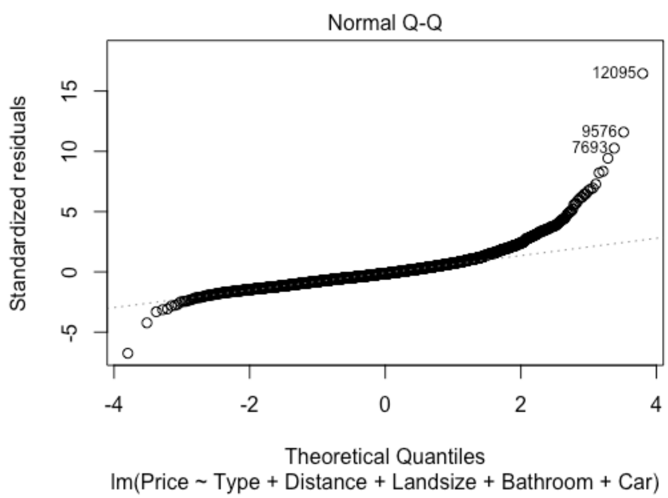


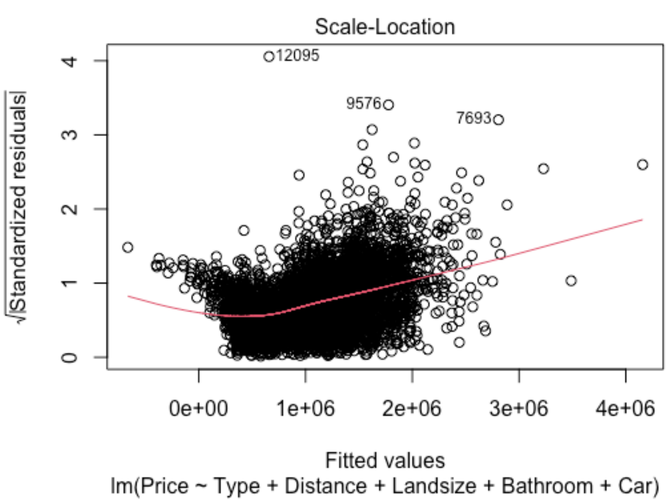




Third Model Assumption Check


Residuals vs Fitted
Residuals
-4e+06
0e+00
4e+06
8e+06
12095
9576
7693
0e+00
1e+06
2e+06
3e+06
4e+06
Fitted values
lm(Price ~ Type + Distance + Landsize + Bathroom + Car)


Normal Q-Q
Standardized residuals
-5
0
5
10
15
12095
9576
7693
-4
-2
0
2
4
Theoretical Quantiles
lm(Price ~ Type + Distance + Landsize + Bathroom + Car)


Scale-Location
√|Standardized residuals|
0
1
2
3
4
12095
9576
7693
0e+00
1e+06
2e+06
3e+06
4e+06
Fitted values
lm(Price ~ Type + Distance + Landsize + Bathroom + Car)

Regression for base model:

```
lm(formula = Price ~ Type + Distance + Landsize, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-2146175  -361047  -106083   208947  7976339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 807243.964  17654.926  45.723  < 2e-16 ***
Typeh_v_u   783632.643  17645.196  44.411  < 2e-16 ***
Typet_v_u   394874.478  27504.761  14.357  < 2e-16 ***
Distance    -32377.748   1212.497 -26.703  < 2e-16 ***
Landsize       55.761      7.765   7.181 7.66e-13 ***
```

Regression for the second model:

```
lm(formula = Price ~ Type + Distance + Landsize + Bathroom, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3186959  -291776   -78218   203796  8287793

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 351870.120  18545.763  18.973  < 2e-16 ***
Typeh_v_u   583441.848  16139.051  36.151  < 2e-16 ***
Typet_v_u   140387.492  24832.532   5.653 1.64e-08 ***
Distance    -34909.842   1067.215 -32.711  < 2e-16 ***
Landsize       31.236      6.847   4.562 5.16e-06 ***
Bathroom    409960.434   9143.210  44.838  < 2e-16 ***
```

Regression for the third model:

```
lm(formula = Price ~ Type + Distance + Landsize + Bathroom +
   Car, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3398277  -289285   -69144   202166  8341809

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 320030.95   18715.81  17.099  < 2e-16 ***
```

```
Typeh_v_u   558536.24   16238.26  34.396  < 2e-16 ***
Typet_v_u   131655.21   24683.56   5.334 9.93e-08 ***
Distance    -36977.96    1081.59 -34.189  < 2e-16 ***
Landsize        26.34       6.82   3.861 0.000114 ***
Bathroom    385578.59    9427.82  40.898  < 2e-16 ***
Car          69637.58    7223.68   9.640  < 2e-16 ***
```

ANOVA test:

```
Analysis of Variance Table

Model 1: Price ~ Type + Distance + Landsize
Model 2: Price ~ Type + Distance + Landsize + Bathroom
Model 3: Price ~ Type + Distance + Landsize + Bathroom + Car
  Res.Df        RSS Df  Sum of Sq        F    Pr(>F)
1   6825 2.3021e+15
2   6824 1.7782e+15  1 5.2387e+14 2037.503 < 2.2e-16 ***
3   6823 1.7543e+15  1 2.3894e+13   92.933 < 2.2e-16 ***
```