STAT 3340

Final Project Report

Regression Analysis of real estate price prediction

2020, Dec 11th

Group 27

Nick Zuber B00807437

Zhe Sun B00819043

Section 1: Abstract

The goal of the report is to provide readers with a regression analysis of real estate properties and their costs. It includes many predictors inside the given dataset. We used the skills we learned to perform multiple linear regression processes to fit the data as close together as possible starting with the pairs function “pairs()” in R, which gives the relationship between numeric variables to indicate approximate relations. Using r and r-squared we found the alkaline information criterion around 3000 for all the datasets, indicating the data may be far off the hypothesis. Then we used backwards selection to obtain the best model and used the vif and graphs, which indicated that the dataset fits our regression model not too badly.

Commented [YF1]: Should this be R and R^2 or r and r^2?

Commented [YF2]: Is this correct?

Section 2: Introduction

The housing market is affected by economic ups and downs, as people may become hesitant about making larger purchases during times of uncertainty. As a result, the true cost of housing is important information for the millions of people trying to decide if putting their hard-earned money into a home purchase is a good idea. Most people would want to know how much housing should cost based on the main factors that go into a house and its living conditions. In the report we apply a linear regression model to a dataset of real estate information to predict prices and trends in real estate. Most people know there are multiple factors that go into deciding the best prices for a house. The main purpose of our work is to illustrate the best model to fit out entire set of data. In the remainder of the report, we discuss the data source, the work needed to obtain the best fitting model, and finally we describe the results along with our conclusions. The data itself and R markdown file can be found in the GitHub.

Commented [YF3]: r or R?

Section 3: Data Description

The dataset focuses on the real estate industry costs from the website Kaggle. The dataset contained information for 415 different properties and, as required, one extra data point was added to the work. The dataset provided the stats behind the date of transactions, the age of the house, distance to the nearest MRT (transit) station, the number of convenience stores in the area, latitude, longitude, and the average price of units in the area. Added to these stats was a house with transaction date of 2012.668, age of 20.4, with the distance to nearest MRT of 307.38, the number of convenience stores within walking distance at 4, a latitude of 24.7044, longitude of 121.6000, and a unit price for the area of 40.57 per square foot. These numbers were randomly selected within similar ranges of the other data points.

Commented [YF4]: feet? Inches? Miles?

Commented [YF5]: Units of what? $ per square foot?

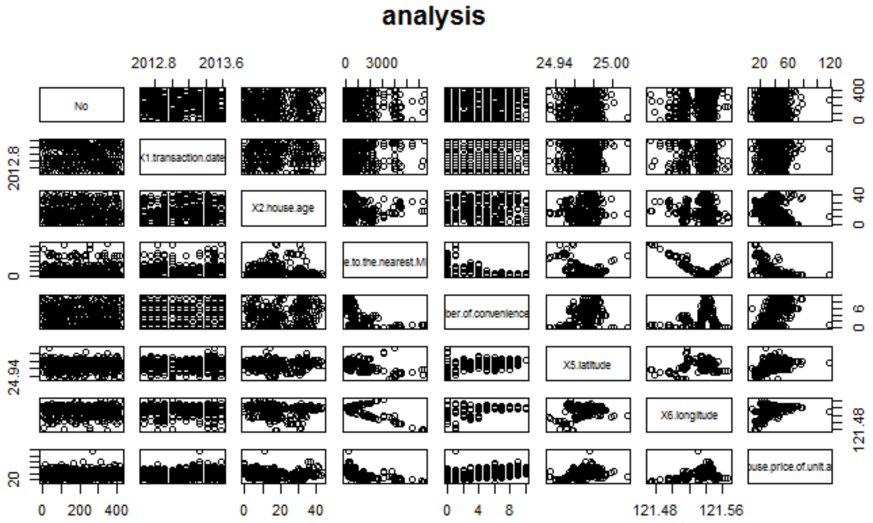
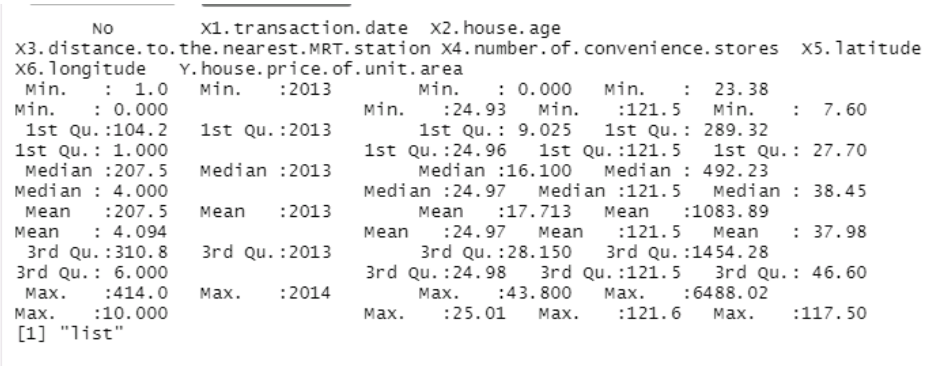


Figure 1. Correlation between variables

Figure one and two show the dataset displayed in scatterplots and the correlations between all the values in our dataset. It appears that the strongest correlations are for latitude and the unit price in the area.



```
       No        X1.transaction.date  X2.house.age
X3.distance.to.the.nearest.MRT.station X4.number.of.convenience.stores  X5.latitude
X6.longitude   Y.house.price.of.unit.area
 Min.   :  1.0   Min.   :2013       Min.   : 0.000   Min.   :  23.38
Min.   : 0.000                     Min.   :24.93   Min.   :121.5   Min.   :  7.60
 1st Qu.:104.2   1st Qu.:2013       1st Qu.: 9.025   1st Qu.: 289.32
1st Qu.: 1.000                     1st Qu.:24.96   1st Qu.:121.5   1st Qu.: 27.70
 Median :207.5   Median :2013       Median :16.100   Median : 492.23
Median : 4.000                     Median :24.97   Median :121.5   Median : 38.45
 Mean   :207.5   Mean   :2013       Mean   :17.713   Mean   :1083.89
Mean   : 4.094                     Mean   :24.97   Mean   :121.5   Mean   : 37.98
 3rd Qu.:310.8   3rd Qu.:2013       3rd Qu.:28.150   3rd Qu.:1454.28
3rd Qu.: 6.000                     3rd Qu.:24.98   3rd Qu.:121.5   3rd Qu.: 46.60
 Max.   :414.0   Max.   :2014       Max.   :43.800   Max.   :6488.02
Max.   :10.000                     Max.   :25.01   Max.   :121.6   Max.   :117.50
[1] "list"
```

Figure 2. Correlation matrix between variables

To give a clearer understanding of the relationships between variables we used box plots to display the scatterplots a bit more clearly.

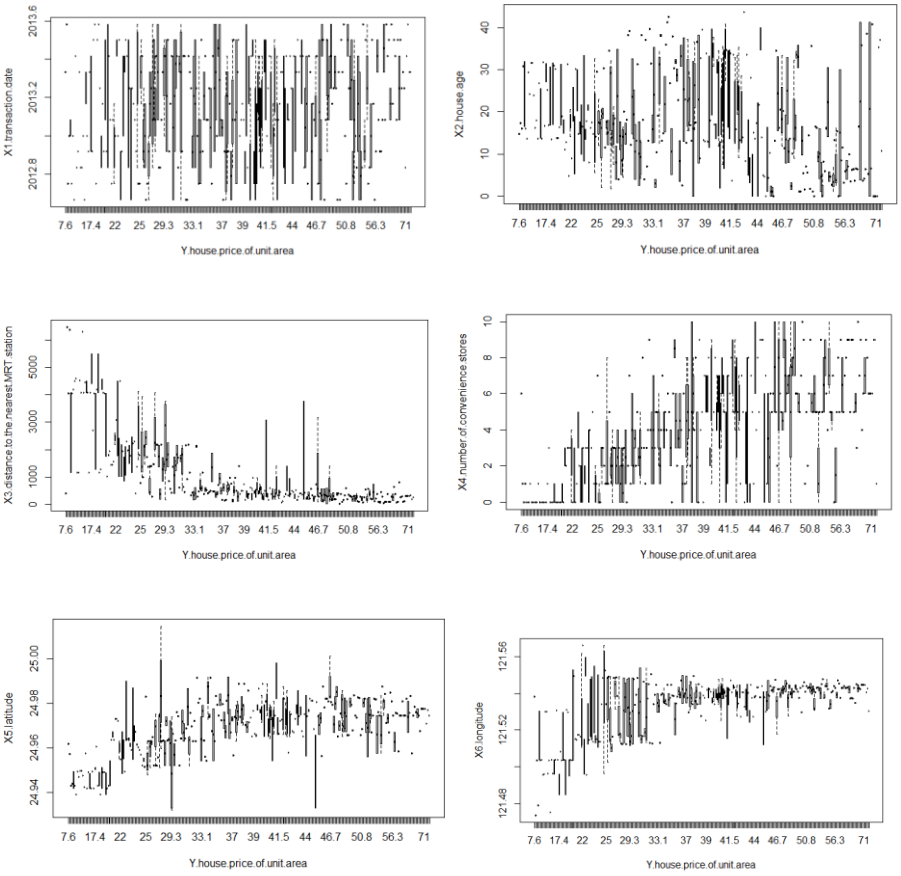


Figure 3. Boxplot for analysis

From the boxplots it appears that the variables are normally distributed.

Section 4. Methods

Multiple Linear Regression Model is in the form

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + \in$$

Where y is the dependent variable in the study (unit price in area), xj is the independent variable present. Whereas b is the coefficient of the independent variable. The linear regression model is then applied using the lm() function with R and the model can easily be viewed with the summary() function. The p-value is <0.05 for significance of the model and it will be assumed to be significant at 95%. Also, the p-value should be >0.05 for individual significance. Lastly the regression diagnostics were done using the plot() function which gives the plots common variance, normality and outlier values. Additionally, the residuals were plotted against the leverage to find outliers

Commented [YF6]: Make sure these are all correct. I have no idea.

Commented [YF7]: this

Commented [YF8]: this

Commented [YF9]: this

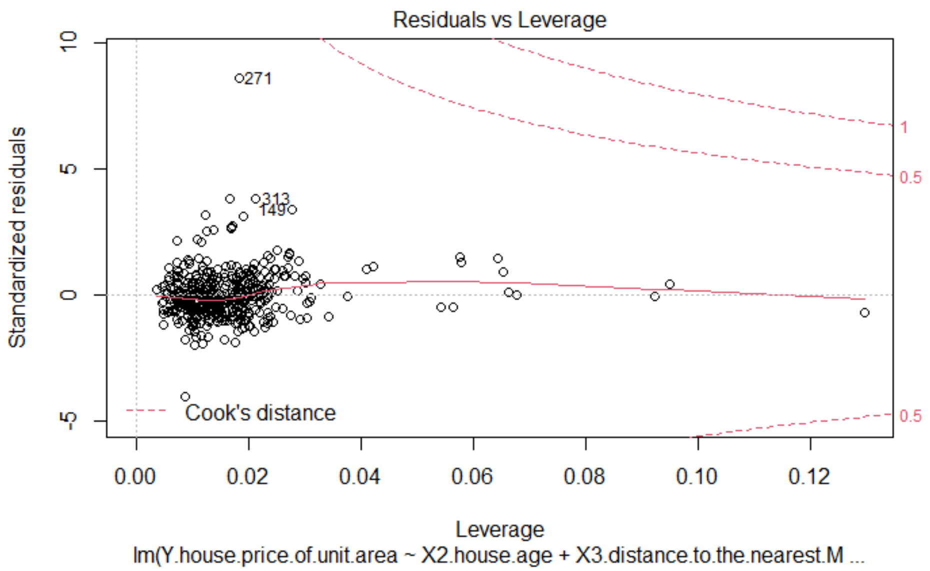


Figure 4. Residuals vs Leverage

A visual analysis of figure 4 shows some outliers such as data points: 271, 313 and 149. However, the red line appears to hover around 0 indicating that the mean of error is approximately 0.

Commented [YF10]: mean of error or mean error?

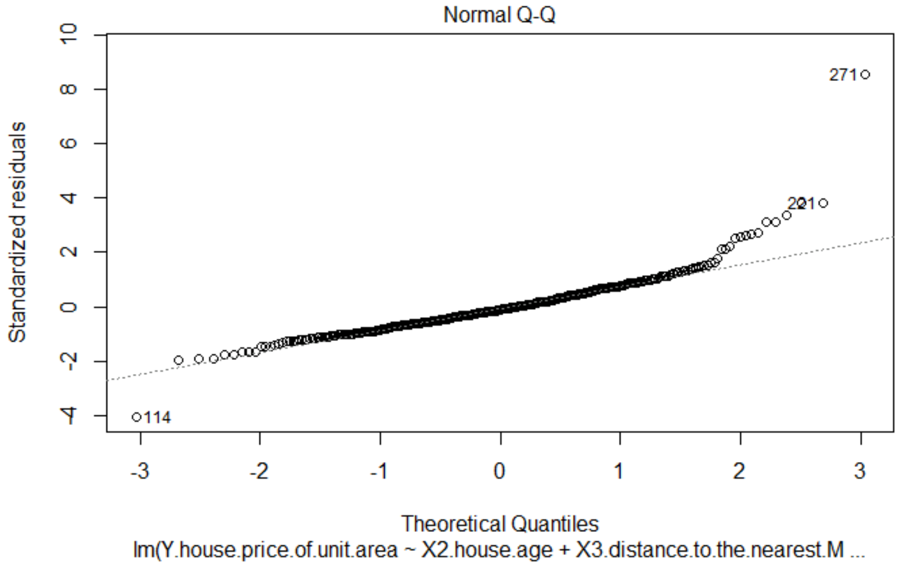
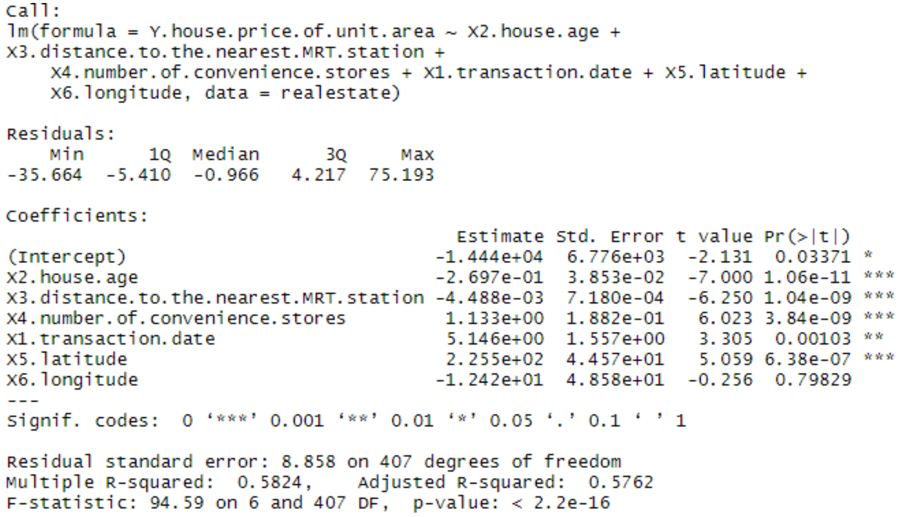


Figure 5. Normal Q-Q plot

The normal Q-Q plot indicates that the residuals are normally distributed as both top and bottom tails have larger values, which is what we expect with the Gauss-Markov assumptions.

Section 5. Results



```
Call:
lm(formula = Y.house.price.of.unit.area ~ X2.house.age + 
X3.distance.to.the.nearest.MRT.station + 
    X4.number.of.convenience.stores + X1.transaction.date + X5.latitude + 
    X6.longitude, data = realestate)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.664  -5.410  -0.966   4.217  75.193 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            -1.444e+04  6.776e+03  -2.131  0.03371 *  
X2.house.age                           -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
X3.distance.to.the.nearest.MRT.station -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
X4.number.of.convenience.stores         1.133e+00  1.882e-01   6.023 3.84e-09 ***
X1.transaction.date                     5.146e+00  1.557e+00   3.305  0.00103 ** 
X5.latitude                             2.255e+02  4.457e+01   5.059 6.38e-07 ***
X6.longitude                           -1.242e+01  4.858e+01  -0.256  0.79829    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.858 on 407 degrees of freedom
Multiple R-squared:  0.5824,	Adjusted R-squared:  0.5762 
F-statistic: 94.59 on 6 and 407 DF,  p-value: < 2.2e-16
```

Figure 6. Summary of best model

In figure 6 we see the information for the best model through our method. The estimated coefficient of latitude is 225.5 which indicates it has the most positive relationship with it and the unit price in the area. The r squared and adjusted r squared are 0.5824 and 0.5762 respectively. This indicates that although it is not an extremally good model there is some significance in it as it was close to a 0.7 which is ideal for a good model.

Commented [YF11]: With itself and the unit price?

Commented [YF12]: R^2 or r^2

Section 6. Conclusion

The study of the dataset using linear regression provided reasonably accurate results for predicting the price of a house, but the housing market can be extremally complex and there will always be more factors that may affect those assumptions than one can account for in a single dataset. From a personal perspective the project provided an opportunity to review R, learn how to use GitHub, design a regression analysis, and learn from the results we obtained from the model.

Section 7. Appendix

GitHub File

https://github.com/Zuber841/Stat-final

References:

Lionbridge AI. 2020. *10 Open Datasets For Linear Regression | Lionbridge AI*. [online] Available at: <https://lionbridge.ai/datasets/10-open-datasets-for-linear-regression/> [Accessed 12 December 2020].

Kaggle.com. 2020. *Real Estate Price Prediction*. [online] Available at: <https://www.kaggle.com/quantbruce/real-estate-price-prediction> [Accessed 12 December 2020].

Guides.github.com. 2020. *Hello World · Github Guides*. [online] Available at: <https://guides.github.com/activities/hello-world/> [Accessed 12 December 2020].