MATH 336 Final Project

Analysis of House Price using Multiple Regression

Hirokatsu Suzuki

Introduction

The real-estate industry has been changing as the economy develops. The price of the house has been up and down, and it is important to understand the trend to purchase an ideal house. The prices of the houses vary depending on multiple factors including the size, location, interior, etc. The objective of this project was to construct a multiple regression model to predict the price of the house. The dataset included the price of the house, which is the dependent variable, and 18 other independent variables about the house. Some of the variables such as view, condition, grade were categorical variables which will be programmed to be treated as categorical variables during the model selection.

Scatter plots between some independent variables versus the price were used to explore relationships. Few of the obvious factors, such as number of bedrooms, number of bathrooms, size of the interior space, were chosen to plot against the price.

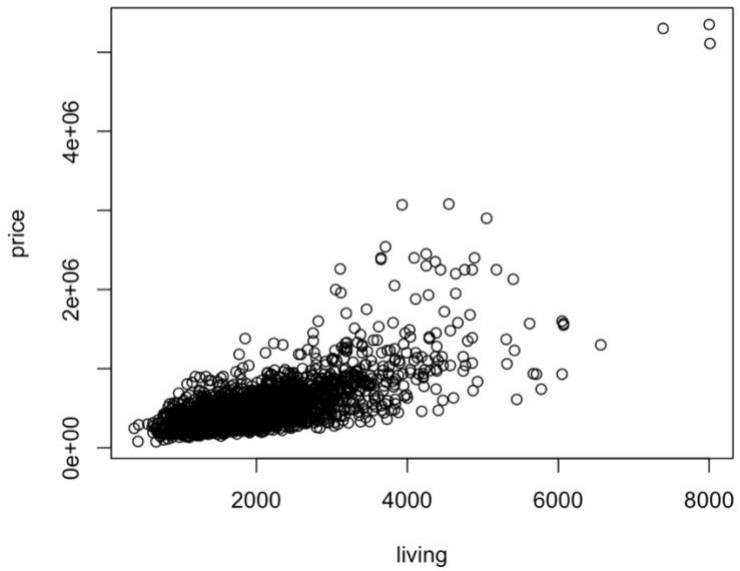

Figure.1 Scatter plot of size of the interior living space versus the price.

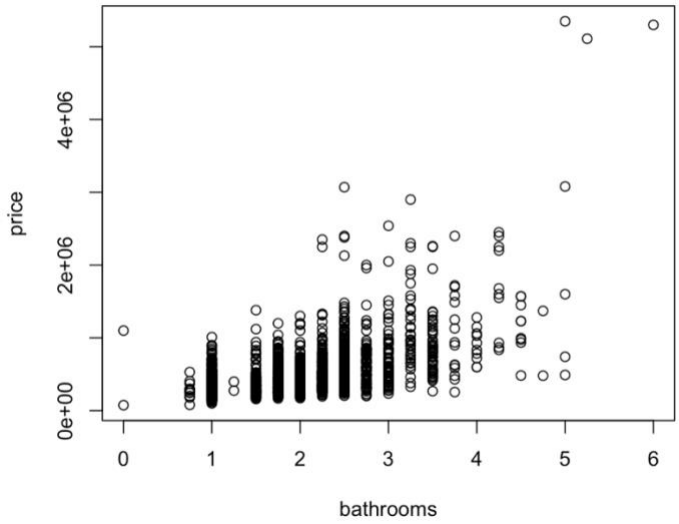

Figure.2 Scatter plot of number of bathrooms versus the price.

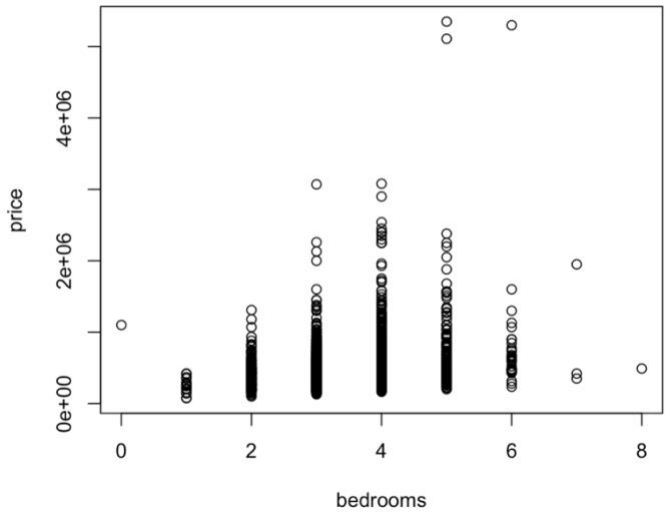

Figure.3 Scatter plot of number of bedrooms versus the price.

In general, three of the scatter plots explained the effects of the variables to the price. In figure 1, the price increased proportionally with the size of the interior living space. In figure 2 and figure 3, they were not as clear to observe how number of bathrooms and bedrooms affected the price compare to figure 1. Hence, these were a few examples of the exploratory data analysis to understand the data. In addition, zip code was excluded when constructing the regression model because they are not a representation of the geographic areas.

Regression Analysis

A full model was constructed to obtain the overall summary of the regression model. Then stepwise function was used to conduct both forward and backward BIC variable selection. Backward BIC variable selection is favored because it has an advantage of considering the effects of all variables at one time. Below is the final regression model and its summary.

$$\log(price) = bathrooms + sqft\_living + waterfront + view + grade + year + lat$$

Equation.1 Multiple regression model.

```
Residual standard error: 0.2558 on 1981 degrees of freedom
Multiple R-squared:  0.7627,    Adjusted R-squared:  0.7605
F-statistic: 353.6 on 18 and 1981 DF,  p-value: < 2.2e-16
```

Figure.4 Summary of the regression model.

Log transformation was used on the dependent variable, price, as the model produced better adjusted R squared value. The normality test was also conducted to assess the model before and after the log transformation. In figure 5, there were many data points lying way below or above the reference line. After the log transformation, those data points were lying much closer to the reference line, meaning the data fit the transformed model better.

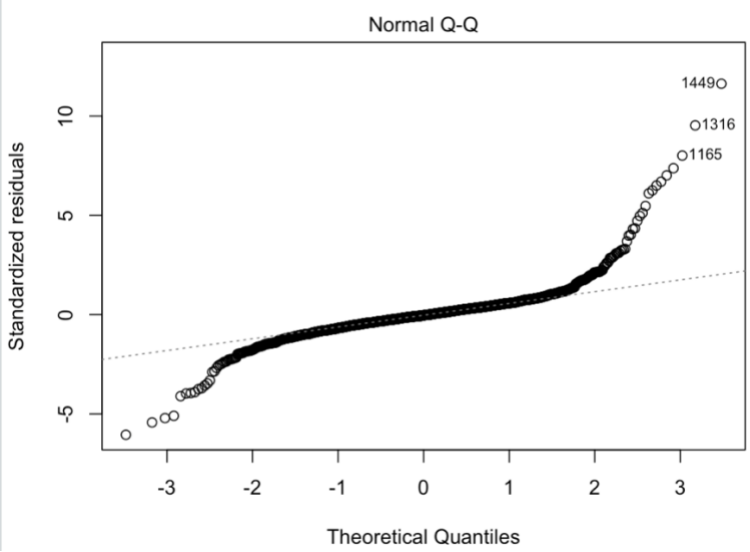

Figure.5 Normality test before the log transformation.

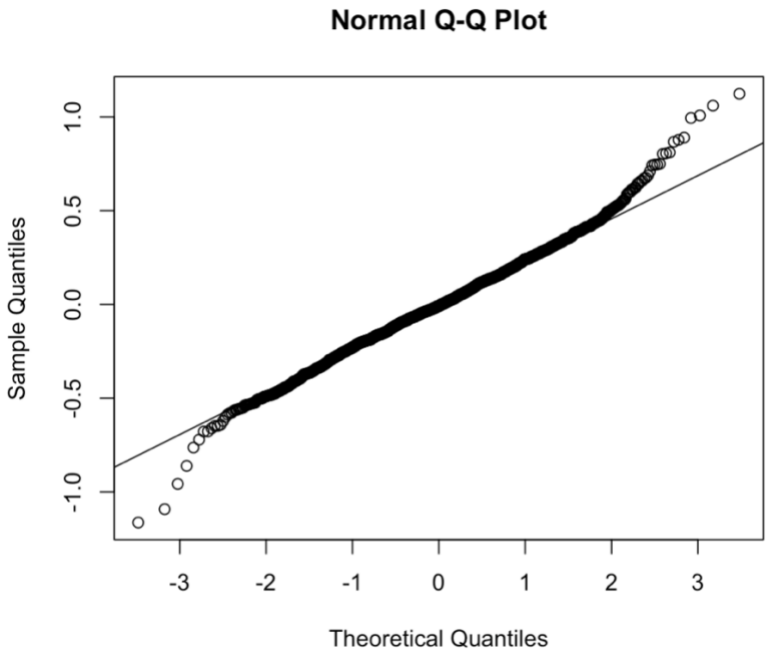

Figure.6 Normality test after the log transformation.

Then some statistical tests were conducted in figure 7 and 8 to obtain fit statistics of the model. In figure 7, the p-values of all variables used in this model was less than 0.001 indicating they are highly significant.

```
                  Df Sum Sq Mean Sq F value Pr(>F)
bathrooms          1 162.31  162.31 2481.24 <2e-16 ***
living             1 101.52  101.52 1551.89 <2e-16 ***
waterfront_factor  1   8.06    8.06  123.14 <2e-16 ***
view_factor        4  11.26    2.82   43.04 <2e-16 ***
grade_factor       9  33.97    3.77   57.69 <2e-16 ***
year               1  37.65   37.65  575.60 <2e-16 ***
lat                1  61.63   61.63  942.18 <2e-16 ***
Residuals       1981 129.59    0.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure.7 ANOVA test of the regression model.

Best Subsets Regression

Model Index	Predictors
1	living
2	living lat
3	living grade_factor lat
4	living view_factor grade_factor lat
5	living view_factor grade_factor year lat
6	bathrooms living view_factor grade_factor year lat
7	bathrooms living waterfront_factor view_factor grade_factor year lat

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.4825	0.4822	0.4813	2323.6193	1767.8371	-3910.6417	1784.6398	282.8578	0.1416	1e-04	0.5186
2	0.6510	0.6506	0.6498	919.2619	982.0493	-4695.7823	1004.4529	190.8621	0.0956	0.0000	0.3501
3	0.6956	0.6939	0.111	548.8455	726.4964	-4967.0512	799.3081	166.5468	0.0838	0.0000	0.3056
4	0.7317	0.7297	-Inf	249.3308	481.8567	-5216.9890	577.0721	146.8568	0.0740	0.0000	0.2696
5	0.7559	0.7539	-Inf	49.3500	294.8025	-5403.1723	395.6187	133.6778	0.0674	0.0000	0.2456
6	0.7602	0.7581	-Inf	15.5364	261.3328	-5436.4321	367.7500	131.3938	0.0663	0.0000	0.2415
7	0.7627	0.7605	-Inf	-3.0000	242.7061	-5454.8947	354.7242	130.1110	0.0657	0.0000	0.2393

Figure.8 More statistics of the regression model.

The test used in figure 8 measured the change in statistics as more variables are added to the model. R square is one of the most common values when assessing the linear regression model, but since this project is constructing the multiple regression model, it is not relative. Other values, such as adjusted R squared and Cp, are some important statistics to evaluate the model which will be discussed in the next section.

Discussion

- Adjusted R squared

In figure 8, adjusted R squared was a good measure for the model as it measured the change in R squared value as more variables were added into the model. This value was low when using three or less variables, but as more variables were added, this value increased. This concluded that the regression model worked better when more variables were added into the model.

- Mean Squared Error

Another value that was also important is the mean squared error (MSE) of the model. MSE was calculated as the mean of the residuals squared of the model which was determined to be 0.065. Such low MSE indicated high accuracy of the model.

- Mallow's Cp Statistic

Cp statistic has many disadvantages when it comes to "assessing" the model, because it was used to gather more information. In figure 8, as more variables were added, the Cp decreased which indicated the model was better fitted. But the Cp value eventually dropped to negative value which the Cp criteria for model selection implied that overfitting was not entirely avoided.

Overall, the evidence of increasing in adjusted R squared as more variables are added and low MSE indicated that this regression model could predict the house price with high accuracy.

Conclusion

In conclusion, the final model for predicting the house price was found to be equation 1. The reason to conduct log transformation was showed using a normality test in figure 5 and 6. The model before the log transformation showed some accuracy in the model; however, a transformed model had better fit in figure 6. The model was then assessed using various statistical tests. Adjusted R squared and MSE indicated that the model with full 7 variables had higher accuracy.

Due to some limitations in the software, only 2000 data was used for analysis and constructing multiple regression model. Generally, more data will produce a better model; hence using a larger dataset in the future work could end up with better model.