House Price Analysis and Prediction using Multiple Regression Model

Dathleen Bituin M. Satur
BS Applied Mathematics
University of the Philippines Mindanao
June 2021

1 INTRODUCTION

In this globalization era, investments are very vital for people who are interested in business. There are so many objects that can be used for investment, and specifically, property investment has shown a significant increase every year (Alfiyatin, 2017). It has been seen to be a good investment option for it can generate ongoing passive income as well as can be a good long-term investment especially that mostly, the value of it increases over time (Caldwell, 2020). However, starting up with this investment could be very expensive. Buying a house, apartment, or piece of land is costly (Majaski, 2021). Hence, for an average buyer, planning and decision-making are very important in choosing which house to buy. Moreover, with the rise of technology, online platforms for house retail have become very popular and may also help the buyers in knowing beforehand the price of the house.

With the advent of machine learning, Multiple Regression has been widely used in research to predict the outcome of a response variable using explanatory variables (Hayes, 2021). Moreover, in a study entitled, Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization, it has performed several tests using linear regression and particle swarm optimization methods to perform house price prediction (Alfiyatin et. al, 2017). In line with this, the primary objective of this study was to perform Multiple Regression Analysis of the house prices on different factors. Specifically, this study aimed to determine the factors affecting the price of houses, fit a regression model and determine the best regression model in predicting the price of houses.

2 DATA

The data were gathered from House Sales in King County, USA and was retrieved from King County GIS Open Data. The data are the home sales prices and characteristic for Seattle and King County from May 2014 to 2015. The total observations are 21,613. The description of the data is shown in the table below.

Variable	Description
Id	Identification
Price	Sale price
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Sqft_living	Size of living area in square feet
Floors	Number of floors
waterfront	‘1’ if the property has a waterfront, ‘0’ if not
view	An index from 0 to 4 of how good the view of the property was
condition	Condition of the house, ranked from 1 to 5
grade	Classification by construction quality which refers to the types of materials used and the quality of workmanship. Buildings of better quality (higher grade) cost more to build per unit of measure and command higher value.

Table 1. Data description

2.1 Data Cleaning

The data were then imported to R studio and data cleaning took place. After seeing the data, unnecessary data were dropped from the data frame which are id, waterfront, and view. We now have 7 variables in our data set which are Price, Bedrooms, Bathrooms, Sqft_living, Floors, Condition and Grade.

3 TECHNIQUES

3.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. In this study, EDA was performed before doing the modelling for us to investigate the data, have an overview of our variables, how they are related and try to seek the patterns distributed in the data. We analyzed for the correlation matrix and multiple scatterplots of the data to check if we are in the presence of any linear or non-linear relations.

3.2 Model Building

Multiple regression was used in building our model. According to Hyndman & Athanasopoulos (2018), building a multiple linear regression model can potentially generate more accurate forecasts since we expect that our forecast variable to not only depend one factor but on other predictors as well. Multiple regression is given by the form, $y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t$ where y is the variable to be forecast and $x_1, \ldots, x_k$ are the k predictor variables. Each of the predictor variables must be numerical. The coefficients $\beta_1, \ldots, \beta_k$ measure the effect of each predictor after taking into account the effects of all the other predictors in the model. Thus, the coefficients measure the marginal effects of the predictor variables (Hyndman & Athanasopoulos, 2018).

3.3 Model training

In this study, the data were split into training and test set. Separating data into training and testing sets is an important part of evaluating data mining models. It will minimize the effects of data discrepancies and better understand the characteristics of the model. Diagnostic checking was also performed by investigating the plots of Residual vs Fitted, standardized residual vs Fitted, Normal Q-Q and Residual vs Leverage. Model Training took place using the training dataset only.

3.4 Model Assessment

The models were tested using the test dataset to estimate how well our model is performing given that it is facing new inputs. Breusch-Godfrey test is used in this study to test the autocorrelation in the residuals designed to take account for the regression model. It is also referred to as the LM (Lagrange Multiplier) test for serial correlation. It is used to test the joint hypothesis that there is no autocorrelation in the residuals up to a certain specified order. A small p-value indicates there is significant autocorrelation remaining in the residuals. The Breusch-Godfrey test is similar to the Ljung-Box test, but it is specifically designed for use with regression models.

3.5 Model Comparison and Validation

To identify the best fit model, the study applied the use of Adjusted R-Squared which measure takes into account the degrees of freedom, that is, it takes into account the extra variables in our model. Root Mean Square Error (RMSE) was also used in this study to measure the standard deviation of the predicted price from the actual price. Moreover, Mean Absolute Percentage Error (MAPE) was also considered to measure the average of the absolute percentage errors of forecasts. Lastly, Akaike's Information Criterion (AIC) was also employed in this study. The concept of AIC is to penalise the fit of the model with the number of parameters that need to be estimated. The model with the minimum value of the AIC is often the best model for forecasting.

4 RESULTS AND DISCUSSION

4.1 Exploratory Data analysis

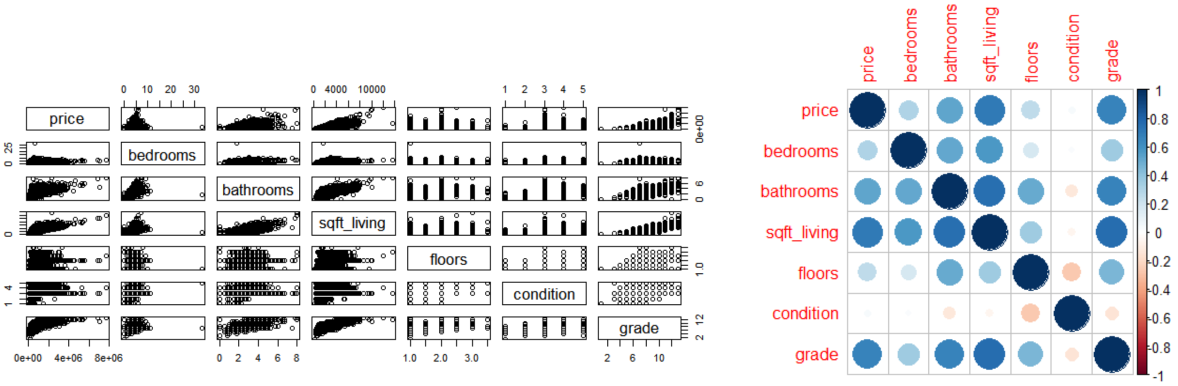
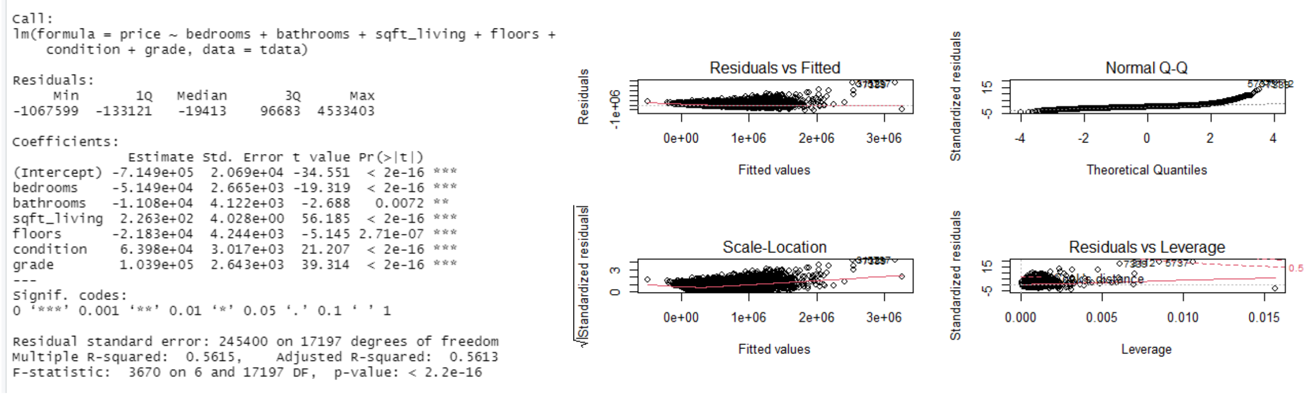


Figure 1. Scatterplot Matrix of the variables (left) and Correlation Matrix (right)

Figure 5.5 is a scatterplot matrix of seven variables. It shows the relationships between the forecast variable (Price) and each of the predictors. We see some non-linear relations between the target variable and some independent variables. Thus, we check for the correlation to see what variables/features have a linear relation with each other. We can notice that condition variable is less correlated to other predictors and showed negative correlation with bathrooms and floors. From the plots, we can see that there is multicollinearity in our data.

4.2 Model Assessment

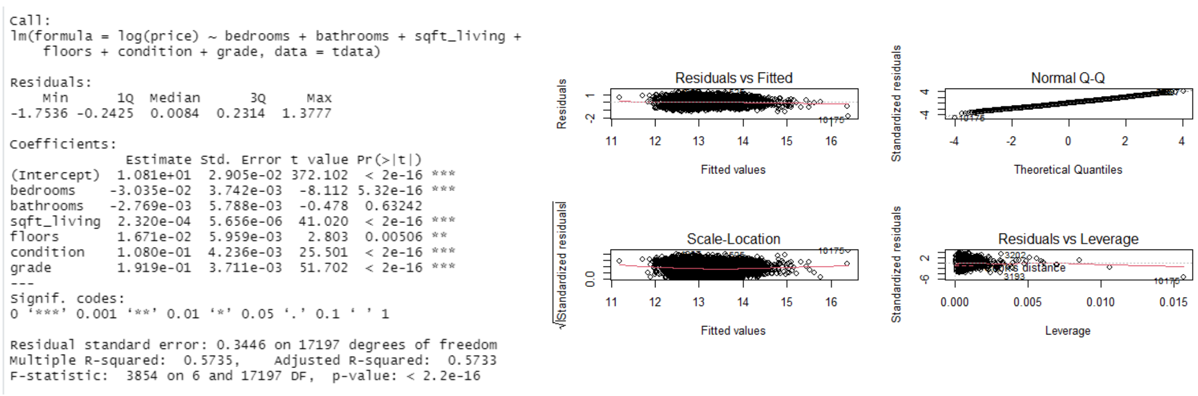
The variables are now fitted to create a regression model. In this study, R studio was used.



```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + floors +
    condition + grade, data = tdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1067599  -133121   -19413    96683  4533403

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.149e+05  2.069e+04 -34.551  < 2e-16 ***
bedrooms    -5.149e+04  2.665e+03 -19.319  < 2e-16 ***
bathrooms   -1.108e+04  4.122e+03  -2.688   0.0072 **
sqft_living  2.263e+02  4.028e+00  56.185  < 2e-16 ***
floors      -2.183e+04  4.244e+03  -5.145 2.71e-07 ***
condition    6.398e+04  3.017e+03  21.207  < 2e-16 ***
grade        1.039e+05  2.643e+03  39.314  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245400 on 17197 degrees of freedom
Multiple R-squared:  0.5615,	Adjusted R-squared:  0.5613
F-statistic:  3670 on 6 and 17197 DF,  p-value: < 2.2e-16
```

Figure 3. Model 1 Training Results

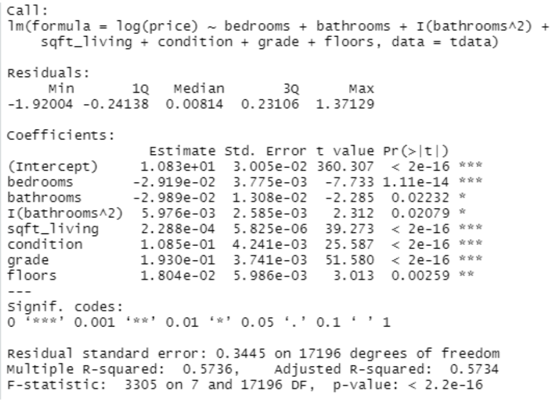
In Figure 3, it can be seen that there is a presence of heteroscedasticity, and the residuals are not normally distributed.



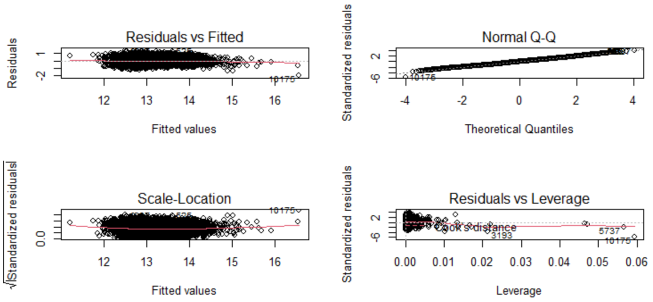
```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + sqft_living +
    floors + condition + grade, data = tdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7536 -0.2425  0.0084  0.2314  1.3777

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.081e+01  2.905e-02 372.102  < 2e-16 ***
bedrooms    -3.035e-02  3.742e-03  -8.112 5.32e-16 ***
bathrooms   -2.769e-03  5.788e-03  -0.478  0.63242
sqft_living  2.320e-04  5.656e-06  41.020  < 2e-16 ***
floors       1.671e-02  5.959e-03   2.803  0.00506 **
condition    1.080e-01  4.236e-03  25.501  < 2e-16 ***
grade        1.919e-01  3.711e-03  51.702  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3446 on 17197 degrees of freedom
Multiple R-squared:  0.5735,	Adjusted R-squared:  0.5733
F-statistic:  3854 on 6 and 17197 DF,  p-value: < 2.2e-16
```

Figure 4. Model 2 Training Results

In Figure 4, a log transformation has been performed in order to fix the heteroscedasticity. It can be observed in the Q-Q plot that the residuals are approximately normally distributed and heteroscedasticity has been removed. However, the adjusted R-squared is acceptable but it can be better. Figure 5.



```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + I(bathrooms^2) +
    sqft_living + condition + grade + floors, data = tdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92004 -0.24138  0.00814  0.23106  1.37129

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.083e+01  3.005e-02 360.307  < 2e-16 ***
bedrooms       -2.919e-02  3.775e-03  -7.733 1.11e-14 ***
bathrooms      -2.989e-02  1.308e-02  -2.285  0.02232 *
I(bathrooms^2)  5.976e-03  2.585e-03   2.312  0.02079 *
sqft_living     2.288e-04  5.825e-06  39.273  < 2e-16 ***
condition       1.085e-01  4.241e-03  25.587  < 2e-16 ***
grade           1.930e-01  3.741e-03  51.580  < 2e-16 ***
floors          1.804e-02  5.986e-03   3.013  0.00259 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3445 on 17196 degrees of freedom
Multiple R-squared:  0.5736,    Adjusted R-squared:  0.5734
F-statistic:  3305 on 7 and 17196 DF,  p-value: < 2.2e-16
```



Model 3 Training results

In model 3, log and polynomial transformation were performed between the variables to increase the R-squared. As seen in the Figure 5, the adjusted R-squared actually increased but only for 0.0001.

The table for the results of Breusch-Godfrey test was shown below.

Breusch-Godfrey test for serial correlation of order up to 1

	LM Test	df	p-value
Model 1	1.9312	1	0.1646
Model 2	5.0735	1	0.02429
Model 3	4.9004	1	0.02685

Table 2. Breusch-Godfrey test in all the models

The table above shows the results for the autocorrelation using Breusch-Godfrey test. The autocorrelation shows a significant spike at lag 1, but it is not quite enough for the Breusch-Godfrey to be significant at the 5% level. In any case, the autocorrelation is not particularly large, and at lag 1 it is unlikely to have any noticeable impact on the forecasts or the prediction intervals.

4.3 Model Comparison and Validation

Calculating the forecast accuracy measures is necessary in any forecasting method. The following table shows the results for the measures of predictive accuracy.

	MAPE	RMSE	AICs	Adj R-squared
Model 1	0.3303791	241434.4	475855.46	0.5613153
Model 2	0.2839139	229685.5	12170.03	0.5733394
Model 3	0.2840051	229642.1	12166.69	0.5734472

Table 3. Results for MAPE, RMSE, AICs, Adj R-squared

The table above shows that Model 3 is the best fit model in predicting the house price because it has the highest Adjusted R-squared and the lowest AICs. On the other hand, Model 1 scored the highest in terms of AIC and in MAPE.

4.4 Discussion

The prices of the house were analyzed using Multiple Regression Model on the following predictors: Bathrooms, Bedrooms, Sqft_living, Floors, Condition and Grade. It was shown that the variables are statistically significant by evaluating the individual p-values of the variables. From the results, the log and polynomial transformed model was the best fit having the lowest AIC of 12166.69 and highest adjusted R-squared which means that approximately 57% of the variation in response variable around its mean is represented in the model. Also, looking at the MAPE, it interprets that on average the we will fail to predict the actual price by approximately 28% and the RMSE tells us that the prediction of the model will be off by 229642.1 dollars.

5 CONCLUSION

The study aimed to perform Multiple Regression Analysis of the house prices on different predictors such as bedrooms, bathrooms, sqft_living, floors, condition and grade. The data were gathered from House Sales in King County, USA from May 2014 to 2015. Three Multiple Regression Models were generated, and Model 3 was determined to be the best fit having the lowest AIC and the highest Adjusted R-squared. It can be inferred that the model performance is not really high in terms of predictive accuracy. Predictors and more transformation of the data are needed in order to have a better prediction of the house price. Nonetheless, the model generated in this study was successful to fit a Multiple Regression model and explain the variations in the price of house using the predictors. For the other model, the error prediction values are still large. For future research, using different methods that will match the data are recommended to obtain smaller error prediction and use more data to get better results. Further study is recommended.

This paper would be a great help for anyone who is interested in purchasing a house as it will help them on having an initial prediction to see if they are getting a good deal on an undervalued house or if they are getting ripped off. Also, for the sellers, this will ensure that they are not under or overvaluing the house. Results for this study are beneficial for the preparation, budgeting, and management. Furthermore, this paper can help other researchers and guide them in performing factor analysis and prediction using the Multiple Regression Model.

6 REFERENCES

Alfiyatin, A. et. al (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization. (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 8, No. 10, 2017

Cladwell, M. (2020). Is Real Estate a Good Investment?. https://www.thebalance.com/is-real-estate-a-good-investment-2386365

Majaski, C. (2021). Renting vs. Owning a Home: What's the Difference?. https://www.investopedia.com/articles/personal-finance/083115/renting-vs-owning-home-pros-and-cons.asp

Onoprishvili, T. (2021). Tbilisi Housing Challenge 2020. https://www.kaggle.com/tornikeonoprishvili/tbilisi-housing-challenge-2020

Hayes, A. (2021). Multiple Linear Regression (MLR). https://www.investopedia.com/terms/m/mlr.asp

Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on June 19, 2021.

Codes can be accessed at : https://github.com/dathleenbituin/AMAT132-Forecasting-.git