Predicting House Prices Using Multiple Regression

By: Jatin Suri

1. Introduction

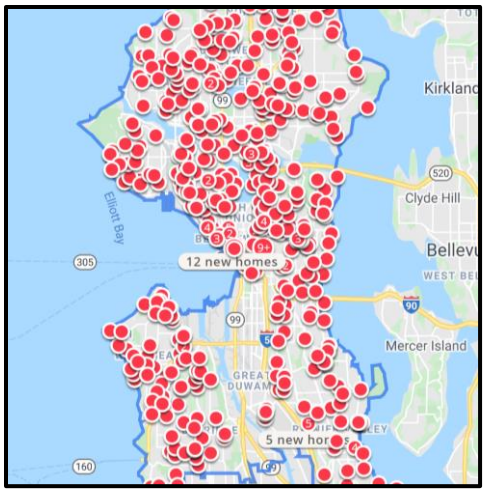
Estimating the value of real estate is a crucial problem for many stakeholders such as sellers, buyers, investors, realtors, etc. Since real estate is often a sizable asset, stakeholders need to list or buy properties at an optimal price; however, determining an ideal price for a property is challenging because there are a plethora of qualitative and quantitative factors that influence the price of a property. Nevertheless, models can predict the price of a house by using information such as the house size, number of bedrooms, bathrooms, etc. Predictive modeling is the approach of building a model from an observed data set and then using that model to estimate a dependent variable for a new data set. This paper will present a multiple regression model to estimate home values based on information about the house.

2. Methodology

Zillow is an American online real estate marketplace that lists millions of for-sale houses and rentals across the United States. Zillow's economic team often makes housing data publicly available for different cities in the U.S. The dataset used for this paper is acquired from Zillow's economic team and can be acquired from the site Kaggle. The data set includes 10,000 observations of house prices in Seattle, Washington in 2017 with 14 different variables. During data cleaning, the *dates* and *reviews* variables were dropped. These variables were dropped because the reviews were in text format and thus were not useful unless some sort of natural language processing was applied. Dates were dropped because it would be complicated to convert them into a usable format. Moreover, all the house listings were in the same year, so

there should not be any time-series issues with the data set. The data was downloaded as a CSV file from Kaggle, cleaned, and then loaded into R for analysis.

Figure 1: Zillow Housing Listings in Seattle, Washington



As shown Figure 1, the Zillow data set consists of random house listings within the Seattle city border. The data only includes houses listed for sale (not rent), so there are not variations in prices that could depend on stay duration. Additionally, there is an even distribution of house locations across the city, so the regression model will predict general house prices in the Seattle area; however, it will have less accuracy in predicting the prices of houses that may fall in more rural parts of Seattle (lower than average house prices) or the financial district (higher than average house prices).

A multiple linear regression model is used to predict house prices. Multiple linear regression is like simple linear regression, but it includes more than 1 explanatory variable. It can be represented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \cdots . \beta_n X_n + \epsilon$$

where Y is the response variable; β_0 is the intercept; n is the number of independent variables; β_n is the slope for X_n, X is the independent variable; ϵ is the error term. When using linear

regression, it is important to verify that the Guess-Markov assumptions of linearity are met to guarantee the validity of ordinary least squares. The parameters for the model are linear, and the data were randomly sampled; however, the independence assumption most likely will not hold because some variables are closely related. Therefore, in the analysis, the final model will be checked for the other linearity assumptions such as multi-collinearity and heteroskedasticity.

Since regression is being used to predict the price of the houses, the price is the response variable, while the other 12 variables are the explanatory variables. The descriptions of the variables are explained in Table 1.

Table 1: Response and Explanatory Variables

Response Variable	**Name**	**Description**
Y	Price	Listed price of the house
Explanatory Variables	**Name**	**Description**
X_1	bedrooms	Number of bedrooms
X_2	bathrooms	Number of bathrooms
X_3	sqft_living	Area of interior of house (in sq. feet)
X_4	condition	An index between 1 and 5 determining the condition of the house
X_5	grade	A house construction index between 1 and 13 where 13 has the highest quality design and 1 has the lowest
X_6	yr_built	Year the house was built
X_7	floors	Number of floors
X_8	sqft_living15	Average living space (in sq. feet) for the nearest 15 neighbors
X_9	sqft_lot15	Average lot space (in sq. feet) for the nearest 15 neighbors
X_{10}	sqft_above	Area of house above ground level (in sq. feet)

X_{11}	**sqft_basement**	**Area of basement (in sq. feet)**
X_{12}	sqft_lot	Area of land (in sq. feet)

Initial Regression Model (no adjustments):

$$Price = \beta_0 + \beta_1(bedrooms) + \beta_2(bathrooms) + \beta_3(sqft\_living) + \beta_4(condition) + \beta_5(grade) + \beta_6(yr\_built) + \beta_7(floors) + \beta_8(sqft\_living) + \beta_9(sqft\_lot15) + \beta_{10}(sqft\_above) + \beta_{11}(sqft\_basement) + \beta_{12}(sqft\_lot)$$

The initial model is the unadjusted linear regression model that will be used to estimate the housing prices in Seattle. In the next section, this model will be tweaked to construct a regression analysis and final model.

3. Analysis

Table 2: ANOVA of corrected regression model

Coefficient	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	20.77016	0.2950362	70.399	< 2e-16
bedrooms	-0.047465	0.004693	-10.114	< 2e-16
bathrooms	0.0903516	0.0073997	12.21	< 2e-16
log(sqft_living)	0.4162081	0.0155018	26.849	< 2e-16
condition	0.0410075	0.005152	7.959	1.92E-15
grade	0.2365556	0.0045021	52.543	< 2e-16
yr_built	-0.006569	0.0001417	-46.371	< 2e-16
floors	0.0586229	0.0076237	7.69	1.62E-14
---	---	---	---	---
Residual standard error	0.321		Degrees of Freedom	9992
Multiple R-squared	0.6297		Adjusted R-squared	0.6294
F-statistic	2427 on 7 and 9992 DF		p-value	< 2.2e-16

Table 2 presents an ANOVA for the corrected regression model. This model has dropped insignificant variables (p-values > .05) and variables that broke the multicollinearity assumption.

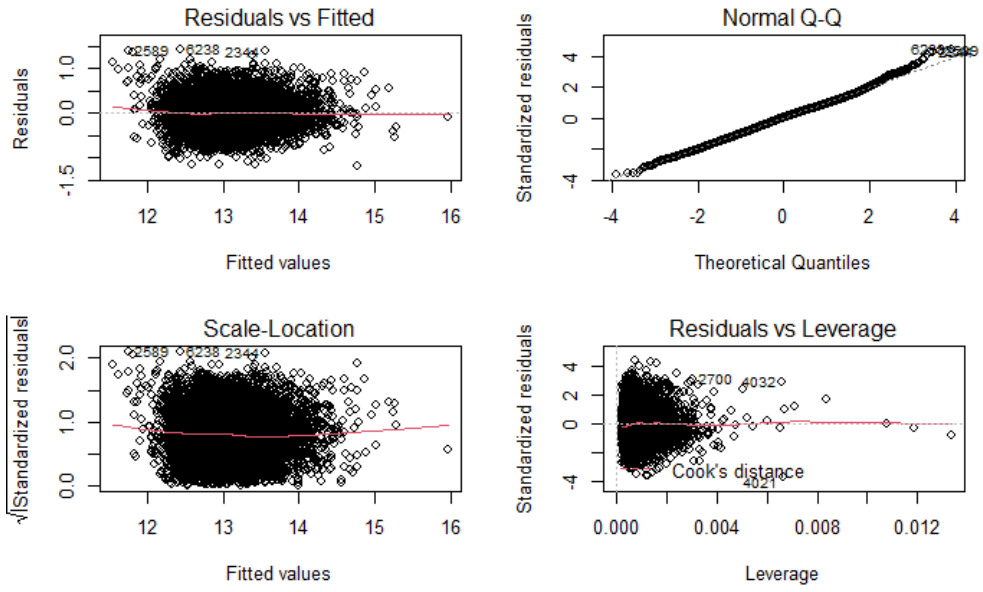
Additionally, box-cox was utilized to find an adequate transformation that would reduce the apparent heteroskedasticity. A log transformation of the predictor variable and sqft_living (X_3) provided a much better fit of the data. Table 2 shows that the p-values for all the variables are significant ($p < .0001$). Moreover, the difference between the multiple R-squared and the adjusted R-squared is quite low (.0003), which suggests that the model is not overfitting too many parameters.

The linearity assumptions of multicollinearity and heteroskedasticity can be tested for this model using the variance inflation factor test and visually inspecting the residual plots, respectively.

Table 3: Variance Inflation Factor Test (to 4 significant figures)

bedrooms	bathrooms	log(sqft_living)	condition	grade	yr_built	floors
1.771	3.111	4.148	1.143	2.661	1.530	1.472

Table 4: Residual Plots



Since the Variance Inflation Factor for all the explanatory variables does not exceed 5, there are no multicollinearity concerns for this model. Additionally, the residuals vs. fitted plot in Table 4 for the final model has a much flatter line indicating that heteroskedasticity has been rectified through the log transformation.

Final Model

$$\log(\text{price}) = 20.770160183 - (0.047464548)\mathrm{X}_1 + (0.090351559)\mathrm{X}_2 + \log(0.416208149)\mathrm{X}_3 + (0.041007547)\mathrm{X}_4 + (0.236555619)\mathrm{X}_5 - (0.006568783)\mathrm{X}_6 + (0.058622899)\mathrm{X}_7 + \epsilon$$

where X_1 is bedrooms; X_2 is bathrooms; X_3 is sqft_living, X_4 is condition, X_5 is grade, X_6 is yr_built; X_7 is floors; ϵ is the error.

4. Conclusion

Although the final model is not perfect, it can explain 63% of the variance of house prices through the 7 explanatory variables. Considering that house pricing is extremely complicated as consumers have diverse tastes, this model does a good job of predicting house prices with the current explanatory variables.

It was expected that most of the explanatory variables will have a positive relationship with price; however, the model consists of two variables with a negative relationship. The number of bedrooms (X_1) and the year built (X_6) may have a negative correlation due to two possible reasons.

1. The model shows that the number of bedrooms has a slightly negative relationship with price. Since this is the relationship when other variables are fixed, it can be concluded that more rooms for a fixed sq_feet are undesirable. This is most likely because the rooms will be smaller, and consumers may prefer bigger rooms over the number of rooms.

2. One reason why an older home may have a negative relationship with the price is that classic or vintage houses may have been built in popular locations. As a result, these places are short distances to urban centers, schools, businesses, or transportation centers. Therefore, these houses can be listed for a higher price and still have demand.

Although the final model is useful, it does have some limitations. Namely, there are three limitations that could be addressed to improve the prediction model.

1. Location data is needed to improve the model. Although all these houses are in Seattle, there is significant price variation depending on where a house is located. Houses closer to more urban areas may have a greater influence on price than houses in rural areas. Therefore, precise coordinates of the listed houses could improve the model fit.

2. An estimated 37% of the price is explained by other factors in the final model. Since buying a house is highly subjective, more factors may be needed to create a better model. Other factors to consider would include crime rates, distance to schools, distance to downtown, demographics, access to transportation, air quality levels, etc.

3. Grouping variables may also improve the model fit as there may be a large increase in price from 1 bedroom to 2 bedrooms; however, the rise in price from 4 bedrooms to 5 bedrooms is probably smaller.

5. Bibliography

Gohar, U. (2020, March 5). *How to use Residual Plots for regression model validation?* Medium. https://towardsdatascience.com/how-to-use-residual-plots-for-regression-model-validation-c3c70e8ab378.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

Zillow. (2018, January 24). *Zillow Economics Data*. Kaggle. https://www.kaggle.com/zillow/zecon.