A Multiple Linear Regression into Housing Prices

Nicolas Thomas

Abstract

Regression analysis is commonly used in statistics to discover the relationship between variables and a specific responding variable. This can be conducted using a simple linear regression with one independent variable, or a multiple linear regression with multiple independent variables. For this study, a multiple linear regression analysis was conducted with five independent variables and one response variable, finding the relationship between them, and how they impact the response variable. This study specifically looks at the prices of houses, and how different criteria impact their price.

1. Objectives

The main questions that were aimed to answer in the regression analysis are which factors have the greatest impact on housing prices? Specifically main features like space in the house measured by square feet, the number of bedrooms and bathrooms it has, what type of neighborhood the house is in, and the year the house was built. It also can be utilized to find the amount that houses will likely be listed for by those looking to sell their home, and potential buyers can use the analysis to see if the prices are fair market value.

2. Methods

2.1. Research Data

The dataset was from Kaggle, and it's called "Housing Price Prediction Dataset", containing 6 columns for variables and 50,000 rows for each observation. One of the variables was used as the response variable (Price), and the other five variables as the independent variables. The environment where the linear regression was conducted was R-studio, creating an R-markdown file with all of the code and results.

2.2. Data Analysis

To begin the data analysis, the dataset was first reviewed and analyzed for the main parameters within it. More specifically, the variable ranges and any missing values. Then, if anything was incorrect, the data would be changed to receive better results.

After the initial tests, it was tested to find the frequencies of the six main variables, along with their general graphs to depict the information within each.

When initial tests were complete, the multiple linear regression analysis could begin. To verify this, the five main regression assumptions were tested. It was shown that there were no outliers by looking at the residuals vs leverage graph with cook's distance, making there no need for further data massaging.

Once the regression assumptions were proved, the collinearity and interactions between variables were found within the dataset, and any resulting changes to fix this were implemented.

Once complete with everything in the initial tests and looking at the variables, the main model was found to explain the data, and then different models to represent the regression were researched.

After testing multiple models, the model with the best criteria was selected as best explaining the data in the context of Housing Prices.

3. Analysis

To begin the linear regression, our data must be verified and can be utilized properly to explain the response variable of price for a house.

3.1. Frequency

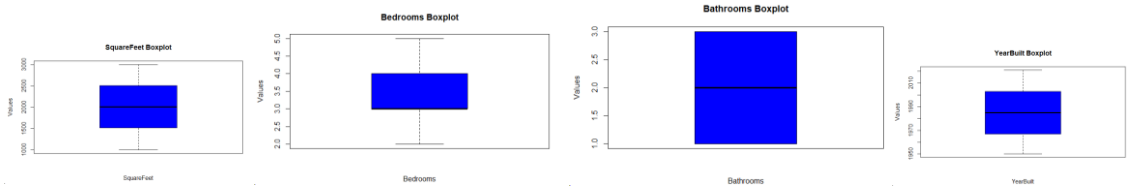
SquareFeet	Bedrooms	Bathrooms	Neighborhood	YearBuilt	Price
Min. :1000	Min. :2.000	Min. :1.000	Length:50000	Min. :1950	Min. :-36588
1st Qu.:1513	1st Qu.:3.000	1st Qu.:1.000	Class :character	1st Qu.:1967	1st Qu.:169956
Median :2007	Median :3.000	Median :2.000	Mode :character	Median :1985	Median :225052
Mean :2006	Mean :3.499	Mean :1.995		Mean :1985	Mean :224827
3rd Qu.:2506	3rd Qu.:4.000	3rd Qu.:3.000		3rd Qu.:2003	3rd Qu.:279374
Max. :2999	Max. :5.000	Max. :3.000		Max. :2021	Max. :492195

Looking at the dataset summary, it can be seen that the Square Footage has a range of 2000sqft, and has a mean of about 2006, showing that most homes being sold are within 1000 and 3000sqft. Bedrooms range from 2 to 5 and have a mean of about 3.5 and a median of 3, meaning that the bedrooms are skewed to the right from larger values. Bathrooms range from 1 to 3 and have a mean and median of 2, displaying a normal distribution. Neighborhood is a character variable that is Urban or Suburban, the main dummy variable, and the YearBuilt ranges from 1950 to 2003 with a mean and median of 1985, showing a normal distribution.

As seen in the summary, there is an error for price, as it isn't possible to list a price for a negative value. To fix this, a coding line was implemented to put all values of negative to positive, as it may have been a user mis-input within the dataset. There were no missing values, but just in case there were, a coding line was also implemented to make all missing values appear as zero.

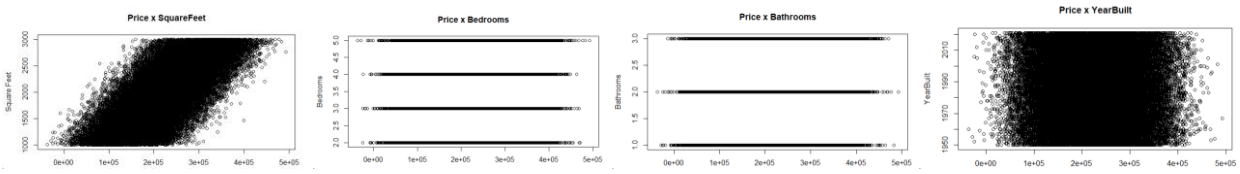
3.2. General Graphs

Looking at the boxplots for the independent variables, they all are approximately within the given range, and there are no outliers in any of the results. Square feet are centered at 2000, with an interquartile range from 1500 to 2500. Bedrooms have a mean of 3.5, and an interquartile range from 3 to 4, meaning most are within these bounds. The bathrooms boxplot displays an interquartile range from 1 to 3 which is the entire graph, showing that values are only 1, 2, or 3. The year built boxplot is from 1950 to 2015, and has an average at about 1987.



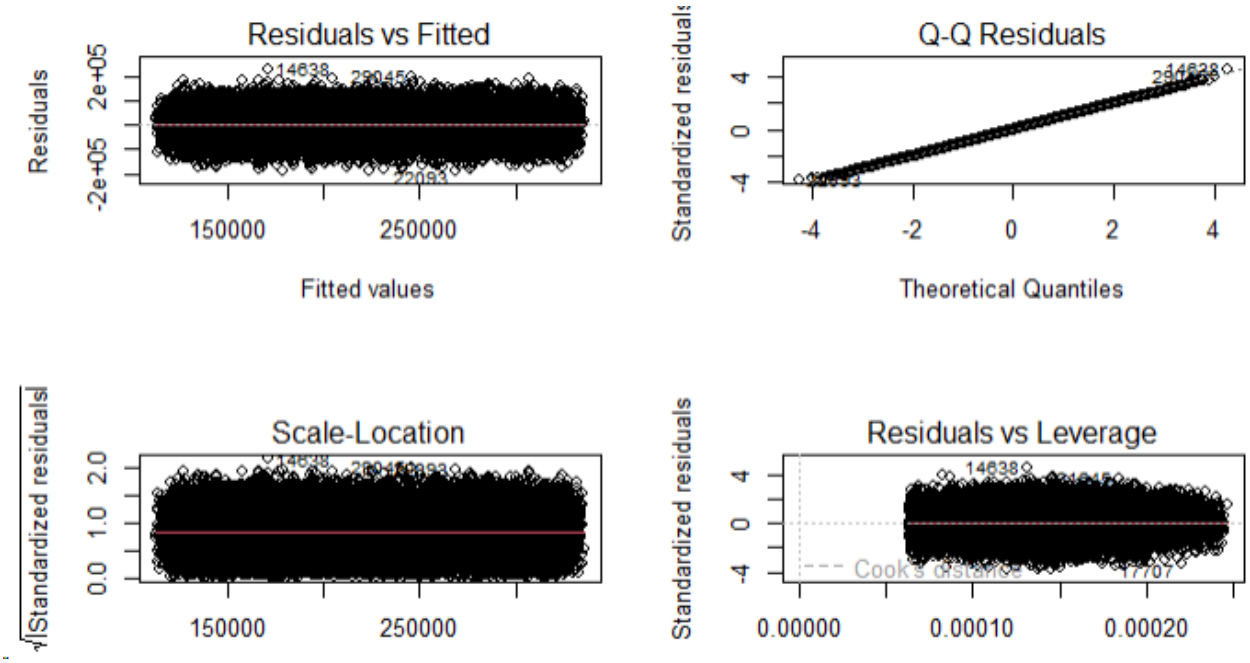
The main scatterplots can depict the relationship between the Prices and independent variables. Looking at the Price x SquareFeet scatterplot, the square feet rises, and the price rises. The Price and Bedrooms, along with the Bathrooms scatterplot both show that there are 4 different inputs for bedrooms and 3 different inputs for bathrooms. The Price and YearBuilt scatterplot is unusual as the

price doesn't rise for the more recent years, as the data is still normally distributed. It should look more like the Price x SquareFeet scatterplot.



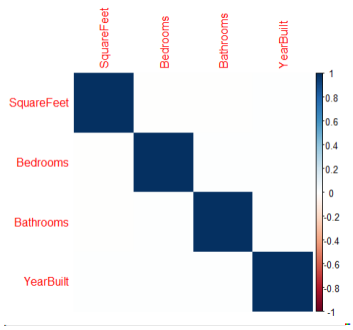
3.3. Regression Assumptions

Next, the five main regression assumptions of linearity, existence, normality, homoscedasticity, and independence must be proved to make sure that the data is reliable and can be accurately represented in a regression analysis. For linearity and existence, view the "Residuals vs Fitted" plot to evaluate the graph. Looking at the line, it approximately stays on the x-axis, proving linearity among the dataset. For normality, you can use the "Q-Q Residuals" plot or a histogram of the residuals. Since the QQ plot is approximately linear, and slightly tails off at both ends, the dataset is almost perfectly normal. Viewing the "Scale-Location" graph can help to understand if the data is homoscedastic or heteroscedastic. The line slightly tails off on the right, but it isn't very significant, and the data is homoscedastic. Since data points are placed under categories based on their neighborhood, the data isn't completely independent.



3.4. Multicollinearity

Multicollinearity must be tested next, as multiple independent variables that are related to a dataset can cause inaccurate results and errors in the best-fit model for a dataset. The collinearity graph for the data appears to be 0 for all different numerical variables, showing that multicollinearity doesn't exist. This is proved by looking at the VIF (Variance Inflation Factor) scores, shown below. The square root of the VIF scores are approximately 1, proving that multicollinearity does not exist.



	GVIF	Df	GVIF^(1/(2*Df))
housing$SquareFeet	1.000202	1	1.000101
housing$Bedrooms	1.000099	1	1.000049
housing$Bathrooms	1.000103	1	1.000051
housing$Neighborhood	1.000249	2	1.000062
housing$YearBuilt	1.000038	1	1.000019

None of the resulting variables had any significant reactions between one another, and

4. Different Models

The main goal of the multiple linear regression analysis is to find the best model to represent the given data. This was conducted by first finding the main model for the data, then testing a backwards elimination model and stepwise regression model. The criteria for each model would then be reviewed, looking at the R-squared, MSE, AIC, and Mallow's Cp. The r-squared value is a measure of how the regression data represents the actual data from the housing prices. MSE, or mean squared error, is the square of results errors in the data, shown in the ANOVA table. AIC, or Akaike Information Criterion is a value for finding how well a regression model fits to the housing prices data. Mallow's Cp compares the full model and different subsets of the predictor variables. Whichever model was the best based on the given criteria, was chosen as the final model and best fit for predicting the data.

4.1. Main Model

The main model consists of the variables for square feet, bedrooms, and bathrooms, neighborhoodsuburban, neighborhoodurban, and yearbuilt. The r-squared for the main model is approximately 0.5702, meaning that about 57% of the data can be explained by the model. The MSE is 49920 and the AIC is 1223720.

Our main model is:

Housing Price = $\hat{B}(0) + \hat{B}(1)SquareFeet + \hat{B}(2)Bedrooms + \hat{B}(3)Bathrooms + \hat{B}(4)NeighborhoodSuburb + \hat{B}(5)NeighborhoodUrban + \hat{B}(6)YearBuilt.$

4.2. Backwards Elimination

In the backwards elimination, start again with the main model. Then, remove the variable with the greatest significance, until none remain that have a p-value over 0.05. Doing this to the model, the Year Built and Neighborhoods variables are removed and now are left with the variables for square feet, bedrooms, and bathrooms.

Our backwards elimination model is:

Housing Price = $\hat{B}(0) + \hat{B}(1)SquareFeet + \hat{B}(2)Bedrooms + \hat{B}(3)Bathrooms.$

4.3. Stepwise Regression

In this study, the stepwise regression model was created through coding in r, looking at the statistical significance of variables and finding where is it maximized. After testing occurred, the only removed

variable was the Year Built variable, and all other variables remained. Through this testing, the model does have a similar R-squared to the main model at 0.5702 (slightly larger when looking at more decimals) and a lower MSE and AIC than the main model, making it a better choice for the representation of the regression.

Our stepwise regression model is:

Housing Price = $\hat{B}(0) + \hat{B}(1)SquareFeet + \hat{B}(2)Bedrooms + \hat{B}(3)Bathrooms + \hat{B}(4)NeighborhoodSuburb + \hat{B}(5)NeighborhoodUrban$

4.4. Conclusions

There is a small variation in R-squared, MSE, and AIC, but not unexpected, as the dataset does have 50,000 points. Looking at the combined factors for all models, it is seen that the stepwise regression does have the best fit for the model, as it contains the highest R-squared value, equal Mean Squared Error (MSE), lowest AIC score, and lowest Mallow's Cp score. Using these results, the stepwise regression was chosen as the final model to represent the data on housing price prediction, based on the five main independent variables.

	Main	Backwards	Stepwise
R-Squared	0.5702	0.5701	0.5702
MSE	2.4920e+09	2.4920e+09	2.4920e+09
AIC	1223720	1223732	1223719
Mallow's Cp	5	19.73	4.02

5. Discussion:

This study delves into whether the given five variables were overall significant in predicting the housing price in the data points, based on linear regression statistics. Starting with the initial data, the main dataset for housing prices was representative of six main variables, with the price as the dependent variable that was investigated in this study. The independent variables had four numerical, in square feet, bedrooms, bathrooms, and year built. The final variable was categorical and converted into a dummy variable in the data representation. There were 50,000 observations, more than enough to provide significance.

During initial tests, it was proven that there were no missing variables and negative data was changed to positive to not skew the results. To continue this study, the five main regression assumptions were all separately tested and satisfied, leading to the continuation of the data to be analyzed in a linear regression. Potential errors between numerical variables in the form of multicollinearity were found to be non-existent, as charts generated with r code and VIF values from the data proved that all numerical variables were independent.

Analyzing the data for its prediction of housing prices, the main model and two other methods, backwards elimination and stepwise regression, were then tested to find the model of best fit. The models all showed different variables in their representations, but the stepwise regression resulted in the most effective criteria and satisfied the dataset more than the other two, making it the final model. The final model of the linear regression for housing price uses the square feet, bedrooms, bathrooms,

and neighborhood variables to best explain the data.

For future studies, it may be ideal to cluster the data in specific locations or cities/states, so that the housing prices can be determined for a certain area. This way, there isn't a broad generalization in the data. Also, the data doesn't include where the housing is and when it was listed for sale based on the year built, which could be improved in future research. Some errors could've appeared in user mis-input for house price and the independent variables, along with not having half values for things like bathrooms and bedrooms. Overall, the multiple linear regression analysis is ideal and useful for the dataset, but more specifications from the data in future studies could be utilized for better results.

References

Dave, Priyanka. "Linear Regression Models." *Medium*, The Startup, 30 Oct. 2020, medium.com/swlh/linear-regression-models-dc81a955bd39.

Imran, Muhammad Bin. "Housing Price Prediction Data." *Housing Price Prediction Data*, Kaggle, 21 Nov. 2023, www.kaggle.com/datasets/muhammadbinimran/housing-price-prediction-data.

Kassambara. "Linear Regression Assumptions and Diagnostics in R: Essentials." *Linear Regression Assumptions and Diagnostics in R: Essentials*, STHDA, 11 Mar. 2018, www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.

Kassambara. "Stepwise Regression Essentials in R." *Stepwise Regression Essentials in R* , STHDA, 11 Mar. 2018, www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/.

Uyanik, Gulden Kata and Guler, Nese. "A study on multiple linear regression analysis" *Sakarya University, Sakayra, Turkey*