Regression Analysis on Housing Prices

By: Joseph Choi, Gabriel Martinez, Ujjwal Sharma, Bindu Nithyananda
MSBA205 (Summer 2025)

Introduction

For our Regression Analysis Project, our group selected a dataset on housing prices because many factors influence real estate values, and understanding these drivers is essential for buyers, sellers, and investors. Our process involved identifying a relevant dataset, performing EDA, and checking regression assumptions to ensure the dataset is suitable for linear regression. After confirming these conditions, we built and refined our regression model by assessing variable significance and multicollinearity and removing predictors that did not contribute meaningfully.

This report addresses the research question: **Which factors most strongly influence house prices, and how accurately can a linear regression model predict them?**

Dataset

For this project, we used the House Price Regression data from Kaggle, which contains 1,000 records and 8 variables: *Square_Footage*, *Num_Bedrooms*, *Num_Bathrooms*, *Year_Built*, *Lot_Size*, *Garage_Size*, *Neighborhood_Quality*, and *House_Price* (target)

The dataset is clean and specifically designed for regression analysis, making it well-suited for both exploratory data analysis and linear modeling. It contains no missing values, all variables are numeric and easy to interpret, and no additional cleaning, preprocessing, or transformation was required.

Exploratory Data Analysis (EDA)

The summary statistics show that houses in the dataset range from 500 to 5,000 sq ft, 1 to 5 bedrooms, 1 to 3 bathrooms, and built between 1950 and 2022. Neighborhood quality scores range from 1-10 (avg. 5.6), and house prices span $110K to $1.1M (avg. $620K).

The bar charts of the categorical variables showed balanced distributions, with no single category dominating. This balance reduces the risk of bias from any one category in the regression model.

Regression Analysis

We fitted our first regression model using all predictors. Scatter plots showed a near-perfect linear relationship between square footage and price (correlation = 0.99), while other predictors displayed weak patterns (Fig. 1). Diagnostic checks confirmed that the regression assumptions were met as the strong linear trend with no curve patterns satisfied linearity, the residual plot showed no funneling and variance patterns which confirmed homoscedasticity (Fig. 2), and the Q-Q plot indicated that the errors were normally distributed.

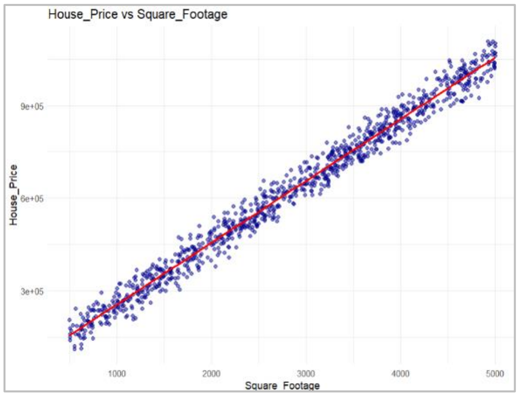

Fig. 1

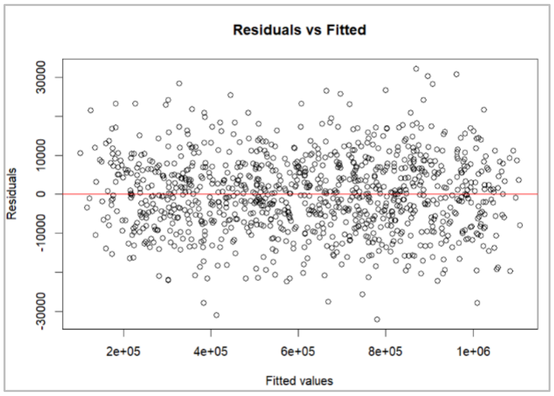

Fig. 2

Then, we evaluated variable significance and multicollinearity using p-values and VIF. All predictors were significant ($p < 0.05$) except for Neighborhood_Quality ($p = 0.454$). VIF values were near 1 for all predictors, indicating no multicollinearity. From this interpretation, we decided to fit a second model without Neighborhood_Quality to see if it would make a difference.

Here are the results:

- **Model 1 (all predictors)**: $R^2 = 0.9985$, Adjusted $R^2 = 0.9985$, Residual SE = 9,799
- **Model 2 (removed N_Q)**: $R^2 = 0.9985$, Adjusted $R^2 = 0.9985$, Residual SE = 9,797

The results were pretty much identical between the models. Ultimately, we determined that Model 2 is preferable because it is simpler and retains the same explanatory power. Bottom line, it achieves the same accuracy with fewer predictors.

Final Insights

Our analysis found that house prices are mainly driven by square footage, with smaller contributions from the number of bedrooms, bathrooms, year built, lot size, and garage size. Among these factors, square footage had by far the strongest linear relationship with price. After evaluating assumptions and refining our model, we found that a simple linear regression can predict housing prices with very high accuracy ($R^2 = 0.9985$).

The dataset indicates that increasing a home's size has the most significant impact on value, while other features such as lot size and garage capacity play a secondary role. Although neighborhood quality is essential in real-world decisions, in this dataset, it did not have a measurable statistical impact once other features were accounted for.