Housing in Ames*

Yanzun Jiang, Siyuan Lu, Yi Tang

October 1, 2024

Table of contents

*Code and data supporting this proposal is available at: https://github.com/Stary54264/Housing-in-Ames

1 Introduction

Understanding the relationships between variables that influence house prices can provide deeper insights for various stakeholders, including homeowners, potential buyers, real estate agents, and policymakers, to make better decisions. Accurate prediction of house prices helps to make informed decisions about buying, selling and investing in real estate. In addition, it helps to assess the economic health of an area and plan for future developments.

The primary research question we aim to answer is: what are the key factors that significantly influence house prices in Ames from 2006 to 2010? Sale price of the house is the response variable. Predictor variables include lot size, overall quality, year of construction, roof style, masonry veneer area, total basement area, central air conditioning, garage size, and value of miscellaneous feature. The hypothesis we state is that there is a statistically significant linear relationship between certain property characteristics (predictors) and the sale price of houses (response). This hypothesis will be tested using linear regression, which would be appropriate in this case. We use residual plots and a Q-Q plot to check the assumption.

Linear regression provides coefficients that quantify the relationship between each predictor and the response variable, making it easier to interpret the impact of each factor on house prices. Specifically, it will provide estimates of how much the sale price is expected to change with a one-unit change in each predictor variable, holding other variables constant. Our primary goal is to understand the impact of each predictor on house prices, so the focus should be on interpretability instead of precision or accuracy.

We found several peer-reviewed articles that focus on similar problems with this paper. Wang (2013) analyzed the influence factors for market supply and demand of commodity housing, and established price influence factors analysis model based on multiple linear regression. The article studies influencing factors of house price using the same method, linear regression, as we do, providing an example of setting a linear regression on factors and house price.

In Zhou et al. (2021), authors develop a novel fuzzy linear regression framework using symmetric and asymmetric trapezoidal fuzzy numbers for determining the relationship of particular (non-) policy factors with the house prices. The article shows the relationship between policy factors and house price, which demonstrates a different aspect of our research question.

In the real estate industry, the price of property plays a crucial role in economic growth. The research Zaki et al. (2022) attempts to predict the price of a house using MLTs. Here, the price of the property is predicted using Extreme Gradient (XG) boosting algorithm and hedonic regression pricing. Both XGBoost and hedonic pricing models use 13 variables as inputs to predict house prices. The article uses an alternative regression method to study the relationship between multiple factors and house price.

In Section 2, we introduce the source and overview of our dataset. Variables used in regression is explained in detail. Then, in Section 3, we talk about the data ethics. Finally, in

Section 4, we address the result of the regression model, plots that verifies the assumptions, and descriptions.

2 Data Description

Table 1: Preview of Data (First Half)

sale_price	lot_area	overall_qual	year_built	roof_style
215000	31770	6	1960	Hip
105000	11622	5	1961	Gable
172000	14267	6	1958	Hip
244000	11160	7	1968	Hip
189900	13830	5	1997	Gable
195500	9978	6	1998	Gable

Table 2: Preview of Data (Second Half)

mas_vnr_area	total_bsmt_sf	central_air	garage_area	misc_val
112	1080	Y	528	0
0	882	Y	730	0
108	1329	Y	312	12500
0	2110	Y	522	0
0	928	Y	482	0
20	926	Y	470	0

The Ames Housing dataset (Table 1, Table 2) was sourced from the `AmesHousing` package (Kuhn (2020)) in R (R Core Team (2023)). It was originally compiled by the Ames City Assessor's Office through a comprehensive data dump of property tax records from 2006 to 2010, and it aimed to document residential property sales (De Cock (2011)). The dataset was initially designed for property tax assessments and general valuation, focusing on property characteristics such as lot area, the year built, and sale price. In contrast, this research aims to analyze how various property features influence house prices in Ames.

The dataset consists of 2930 observations and 82 variables relevant to understanding housing market dynamics. It was cleaned using `tidyverse` package (Wickham et al. (2019)). After cleaning, we selected 1 response variable, `sale_price`, and 9 predictor variables: `lot_area`, `overall_qual`, `year_built`, `roof_style`, `mas_vnr_area`, `total_bsmt_sf`, `central_air`, `garage_area`, and `misc_val`.

- `sale_price`: Priceof the house in dollars
- `lot_area`: Lot size in square feet
- `overall_qual`: Rates the overall material and finish of the house
- `year_built`: Original construction date
- `roof_style`: Type of roof
- `mas_vnr_area`: Masonry veneer area in square feet
- `total_bsmt_sf`: Total square feet of basement area
- `central_air`: Central air conditioning
- `garage_area`: Size of garage in square feet
- `misc_val`: Value of miscellaneous feature in dollars

These predictor variables all show the quality of the house, which will affect the price of the house directly. So, we believe there is a linear relationship between these predictor variables and the response variable.

Table 3: Summarize Table of Numerical Data

	Mean	Standard_Deviation	Median
sale_price	180425.31	79811.03	160000
lot_area	10143.13	7898.24	9434
overall_qual	6.09	1.41	6
year_built	1971.13	30.22	1973
mas_vnr_area	101.97	179.15	0
total_bsmt_sf	1050.52	440.66	990
garage_area	472.34	215.23	479
misc_val	51.07	568.76	0

From the summary table (Table 3), we can easily see that `mas_vnr_area` and `misc_val` might be right-skewed since their mean is a lot greater than their median. An interesting point is that the standard deviation of `misc_val` is quite large, which indicate that houses in Ames might differs significantly in miscellaneous features. By analyzing these variables, we aim to provide insights into how specific property characteristics affect housing prices in Ames, Iowa.

3 Ethics Discussion

Our data is collected from Ames City Assessor's Office (De Cock (2011)), then we cleaned the data to keep some key factors that is highly relavant to house prices. Raw and processed versions of the data from De Cock is published on Journal of Statistics Education in 2011. More detailed information about source of data is described in Section 2. Stakeholders of our analysis include homeowners, potential buyers, real estate agents, and policymakers.

4 Preliminary Results

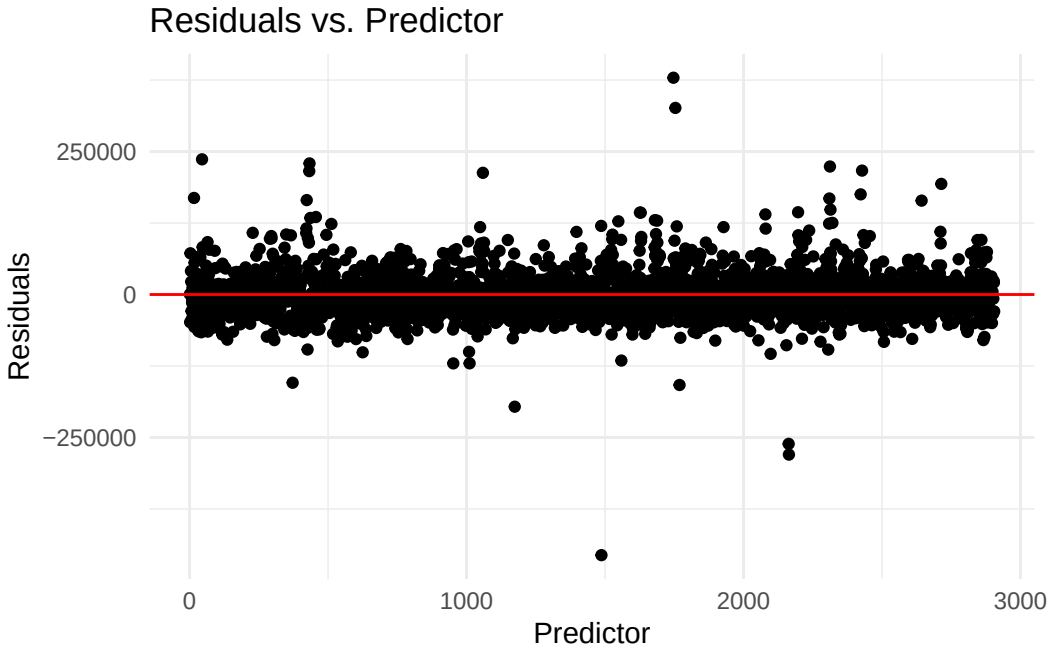

Figure 1: Uncorrelated errors testing

I plotted residuals against observation order to test uncorrelated errors. If errors are uncorrelated, residuals should scatter randomly around the horizontal axis without any clear pattern. From Figure 1, we can see that residuals appear to be randomly distributed across observations, which satisfy the assumption of independence in the linear regression model. But there still have some larger residuals, we can regard these points as potential outliers.

The constant variance assumption means we assume that the variance of the marginal distribution of responses does not change with the value of the predictor. The red loess curve added to the plot highlights this trend, showing a wider spread of residuals as the fitted values rise. From Figure 2, all data points lead to a fanning effect which against the assumption

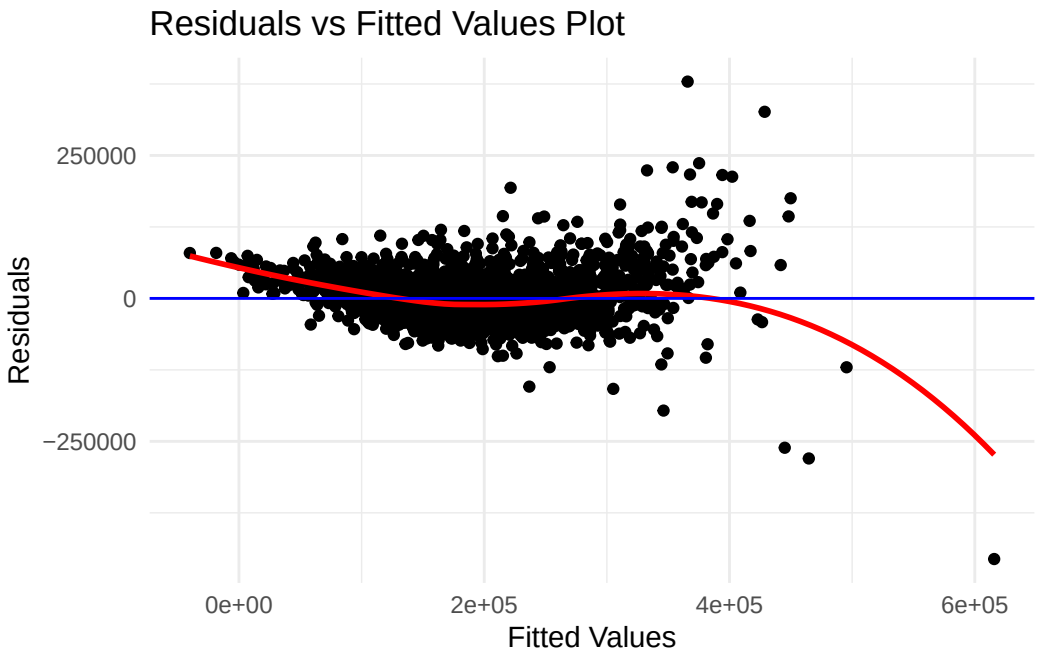

Figure 2: Constant variance testing

of constant variance(also known as homoskedasticity). The red loess curve added to the plot highlights this trend, showing a wider spread of residuals as the fitted values rise.

For linearity, the residuals are scattered relatively randomly around the horizontal line at zero, without showing any distinct or systematic patterns, such as curves or trends. Based on the Figure 2 plot, most of the residuals are centered around the horizontal line and no obvious non-linear patterns like U-shaped or inverted U-shaped trends in the plot, which satisfies the assumption of linearity, although there are a few outliers. This suggests that the relationship between the predictors (like lot_area, overall_qual, and year_built) and the response variable (sale price) is mostly linear.

In a well-behaved model that follows a normal distribution, the points in the Q-Q plot should align closely with the red reference line. Based on the Figure 3, most of the residuals are near the red line, which suggests that the central part of the residual distribution generally follows a normal distribution. Although there are some deviations at the tails, these are unlikely to have a major impact on the model's overall performance. Therefore, the Q-Q plot indicates that the data mostly satisfies the normality assumption.

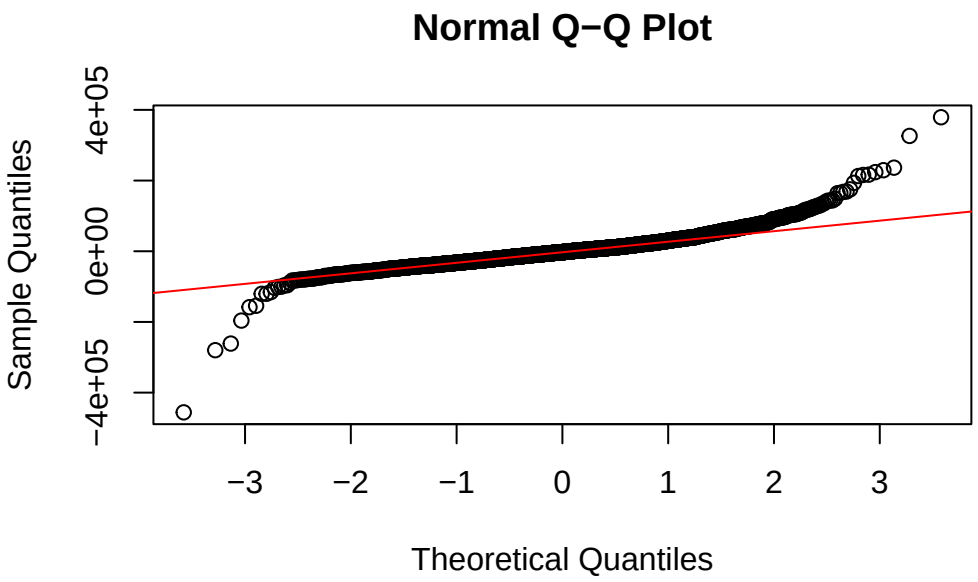

Figure 3: Q-Q Plot

A Appendix

A.1 Contributions

Group contribution is available at https://github.com/Stary54264/Housing-in-Ames/graphs/contributors. Below is a more specific version of group contribution.

- Yanzun Jiang: Organized discussions and meetings; assigned tasks to group members; set up Github workspace for collaborating; downloaded data for setting up the linear regression model; cleaned data to make further analysis easier; introduced the dataset; made the reference list; revised and combined group member's work together.
- Siyuan Lu: Set research question and hypothesis; searched and read peer-reviewed articles; introduced the project; checked data ethics.
- Yi Tang: Built linear regression model to predict house sale prices by using five key predictors in cleaned data. It assisted to understand the relationship between variables and ensure data meets key assumptions for statistical validity.

References

De Cock, Dean. 2011. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." *Journal of Statistics Education* 19 (3).

Kuhn, Max. 2020. *AmesHousing: The Ames Iowa Housing Data*. https://CRAN.R-project.org/package=AmesHousing.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wang, Cheng. 2013. "Influencing Factors Analysis of House Prices Based on Multiple Linear Regression." *Applied Mechanics and Materials* 415: 722–25.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zaki, John, Anand Nayyar, Surjeet Dalal, and Zainab H Ali. 2022. "House Price Prediction Using Hedonic Pricing Model and Machine Learning Techniques." *Concurrency and Computation: Practice and Experience* 34 (27): e7342.

Zhou, Jian, Yixuan Shen, Athanasios A Pantelous, and Hui Zhang. 2021. "The Range of Uncertainty on the Property Market Pricing: The Case of the City of Shanghai." *Finance Research Letters* 40: 101720.