Housing in Ames*

Yanzun Jiang, Siyuan Lu, Yi Tang

September 30, 2024

Table of contents

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

1 Introduction

R Core Team (2023)

2 Data Description

Table 1: Preview of Data

sale_price	lot_area	overall_qual	year_built	roof_style	mas_vnr_area	total_bsmt_sf	central_air	garage_area	misc_val
215000	31770	6	1960	Hip	112	1080	Y	528	0
105000	11622	5	1961	Gable	0	882	Y	730	0
172000	14267	6	1958	Hip	108	1329	Y	312	12500
244000	11160	7	1968	Hip	0	2110	Y	522	0
189900	13830	5	1997	Gable	0	928	Y	482	0
195500	9978	6	1998	Gable	20	926	Y	470	0

The Ames Housing dataset (Table 1) 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 ((**ames?**)). 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 ((**tidyverse?**)). 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`.

#TODO: explain variables

These predictor variables all shows 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.

The variability in `lot_area` and the distribution of the `year_built` variable indicate interesting trends that warrant further exploration.# By analyzing these variables, this study aims to provide insights into how specific property characteristics affect housing prices in Ames, Iowa.

3 Ethics Discussion

4 Preliminary Results

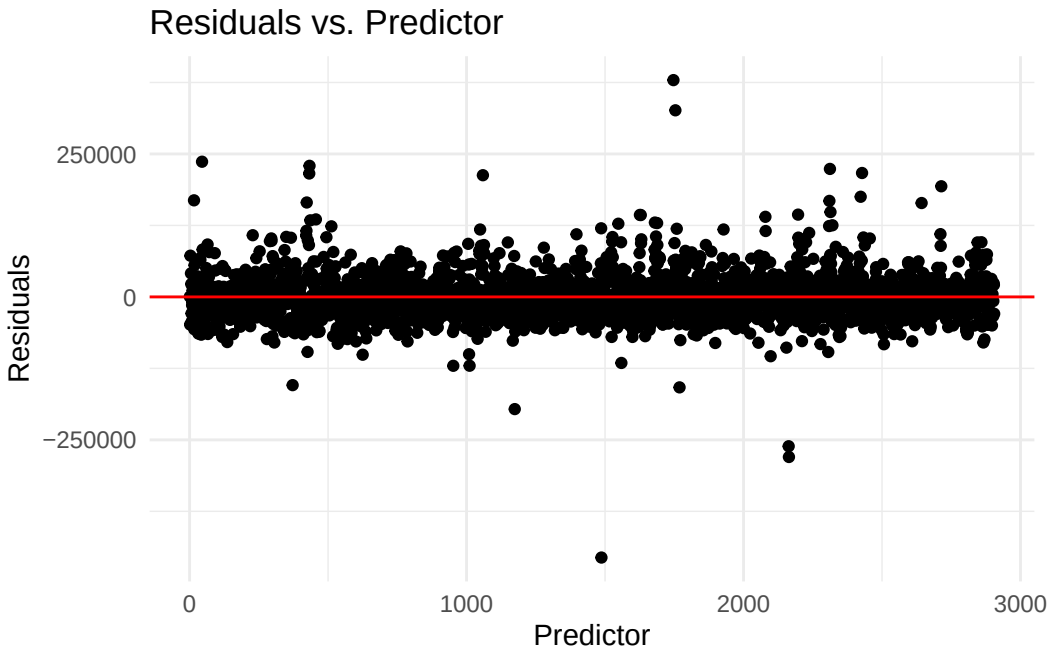

Figure 1: Uncorrelated errors testing

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.

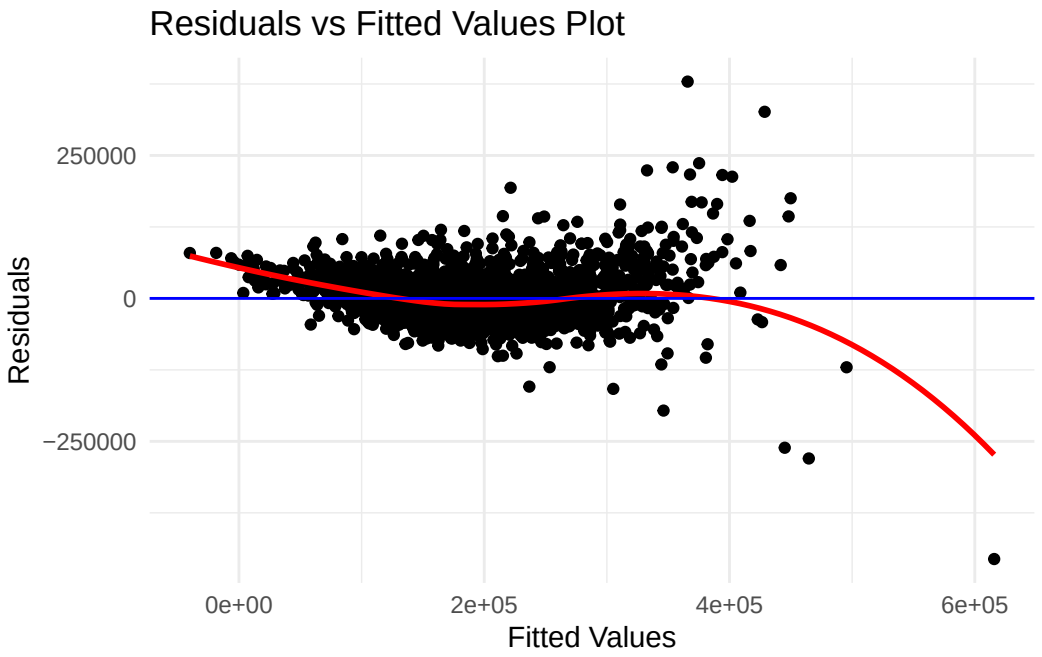

Figure 2: Constant variance testing

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 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.

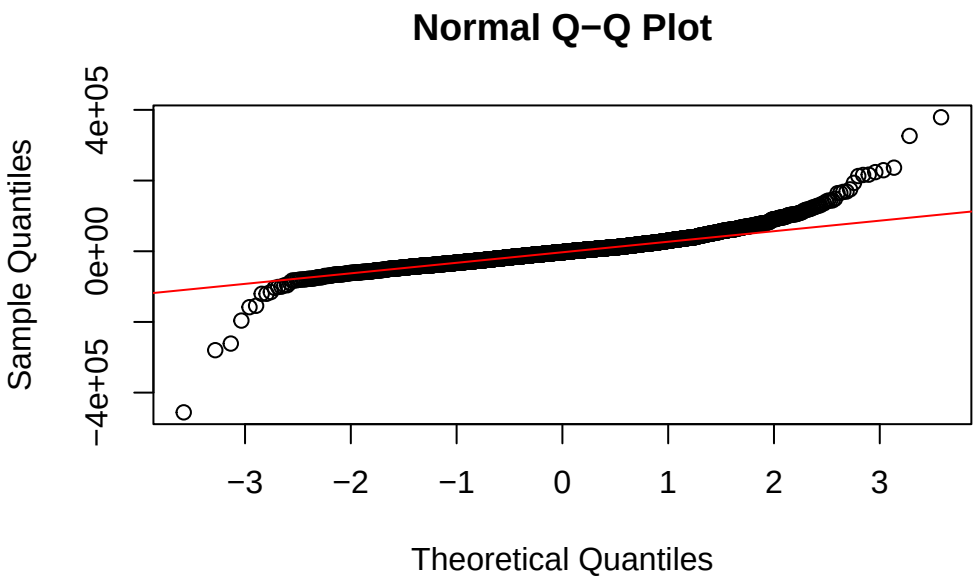

Figure 3: Q-Q Plot

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.

A Appendix

A.1 Contributions

Yanzun Jiang: Organized discussions and meetings; set up Github workspace for collaborating; downloaded data for setting up the linear regression model; cleaned data to make further analysis easier; completed Section 2 in the proposal; made the reference list; revised and combined group member's work together.

Siyuan Lu:

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

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/.