EXPLORING FACTORS THAT INFLUENCING HOUSING PRICES OF BEIJING

ZIXIANG ZHANG
Department Of Statistics, University Of Toronto.

RESEARCH MOTIVATION

This study aims to identify the key factors influencing housing prices in Beijing, a crucial economic and political hub in China. Instabilities in Beijing's housing market affect local living standards, affordability and even national monetary policies. Understanding these factors is critical for designing effective policy interventions as well as investment decisions.

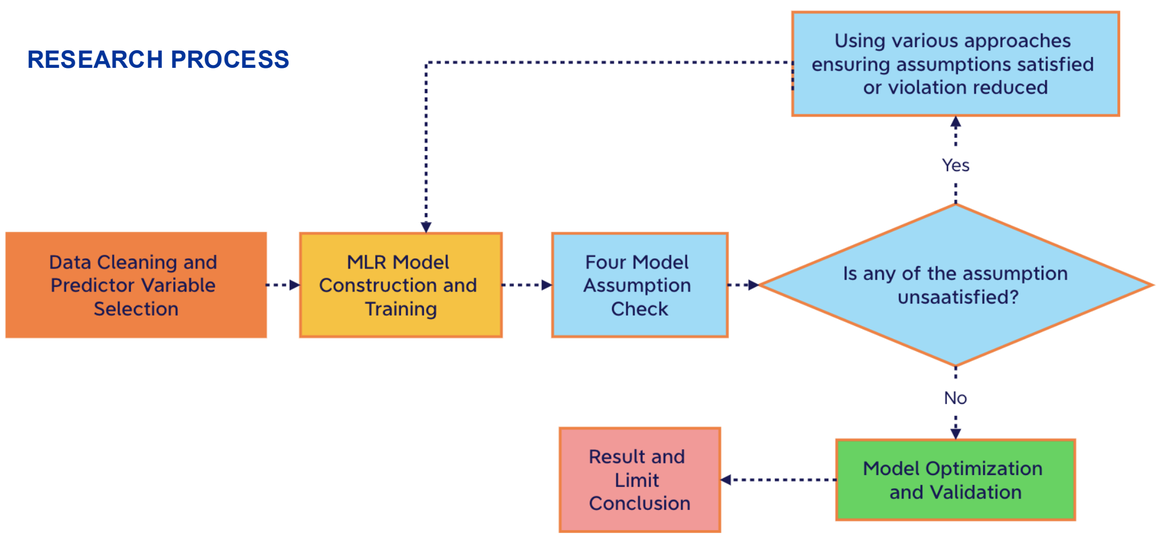
DATASET BACKGROUND & OVERVIEW

The dataset for this analysis is sourced from the Lianjia website, a leading real estate platform in China, recording real transaction data. It includes key variables such as metro proximity, square footage, and year of construction, focusing on Beijing's housing market from 2011 to 2017. With over 300,000 observations and 26 variables, the dataset ensures robust analysis and minimizes the impact of outliers, making it ideal for studying housing price trends in Beijing. We have selected some numerical data and summerized them as follow:

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
totalprice	60.5	207.5	298	350.7	430	1797
constructiontime	1990	1998	2003	2002	2006	2016
communityAverage	20483	45307	56901	60591	71761	129983

METHODS OVERVIEW

A multiple linear regression model was constructed using the cleaned dataset to analyze factors influencing housing prices in Beijing. The model prioritizes interpretability, estimating regression coefficients to determine the magnitude and direction of each factor's impact. Key assumptions of linearity, normality, constant variance, and independence were tested and validated using diagnostic plots and statistical tests (e.g., T-tests, F-tests, Partial F-tests). The model was further optimized and validated to ensure reliability, and results were compared with relevant literature.



MODEL CONSTRUCTION

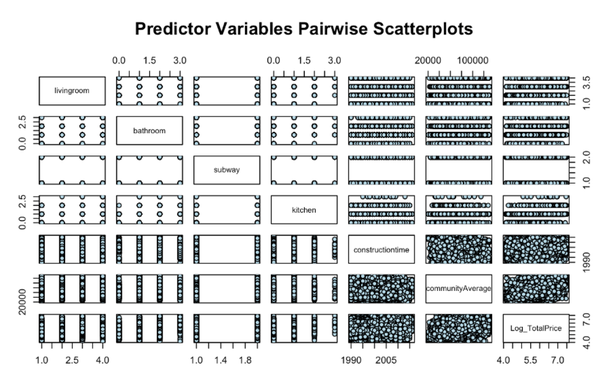
The analysis process involved both model selection and validation. By examining the Variance Inflation Factors (VIFs) for each predictor variable, we identified and excluded variables that introduced excessive multicollinearity or instability to the model. Subsequently, T-tests were performed on each constructed model, leading to the removal of predictors with no significant contribution to the response variable.

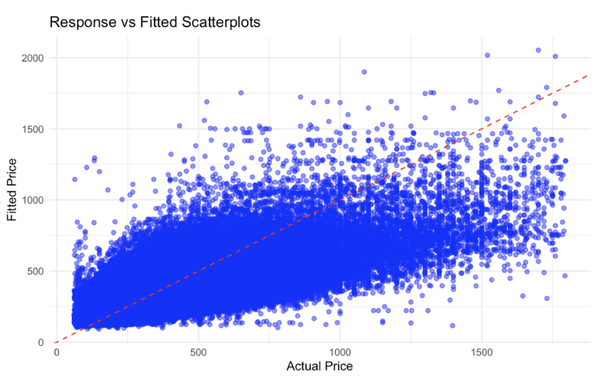
Once all remaining variables were statistically significant (i.e., with p-values less than 2e-16), we evaluated whether further optimization could be achieved by eliminating any redundant predictors. Partial F-tests was used to drop potentially redundant predictors and the All-Subsets Selection method was employed to furthur validate these adjustments and ensure the robustness of our final model. The finalized model is as follows:

$$\log(\text{TotalPrice}) = -60.38 + 0.3002 \cdot \text{LivingRoom} + 0.1749 \cdot \text{Bathroom} + 0.02503 \cdot \text{Subway1} + 0.007259 \cdot \text{CommunityAverage} + 1.416 \cdot \text{ConstructionTime} + 0.1373 \cdot \text{Kitchen}$$

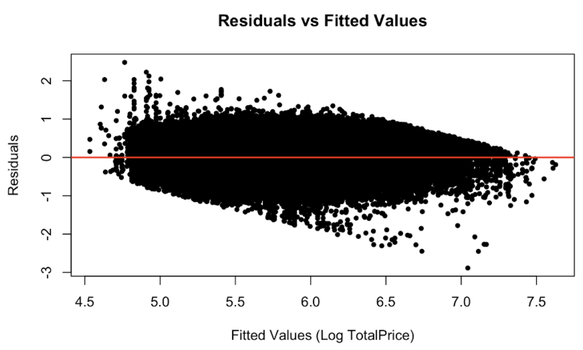
ANALYSIS & RESULTS

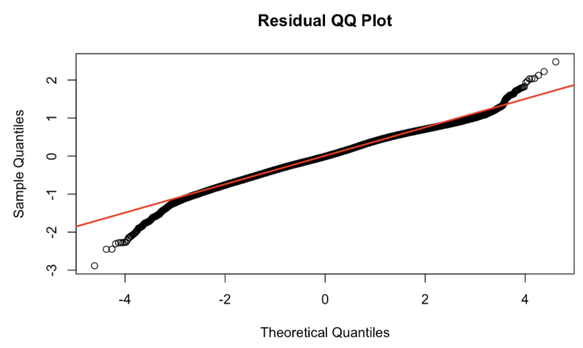
We proceeded to verify that our final model satisfies two preconditions and all four model assumptions: linearity, constant variance, normality, and independence. From the two plots below, we observed random diagonal scatter with no identifiable non-linear trends, as well as no curves or non-linear patterns in the pairwise scatterplots of the predictors. These findings suggest that the model satisfies the conditional mean and response relationship, enabling us to perform assumption checks through visual diagnostics.





Further, by using the Residual QQ Plot and the Residual vs. Fitted Scatter Plot, we found that the QQ plot closely aligns with the diagonal, with acceptable deviations at the extremes(head and tail). In terms of constant variance, our residual versus fitted plot shown no extreme patterns of fanning or outward spread, with all data points attached tightly to each other. We conclude these two asssumptions has been satisfied as well.



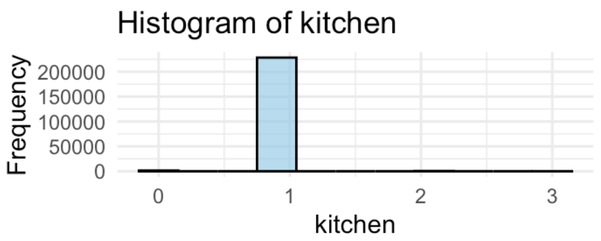


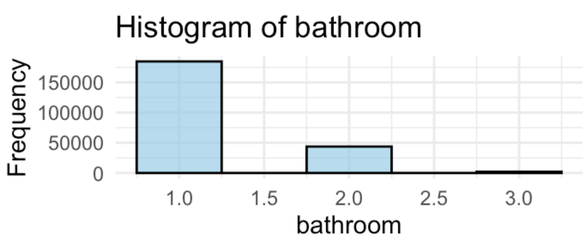
In the end, no cluster patterns has been found in any of the plots we drawed. We conclude independence assumption has been satisfied, and thus, we have a model that ready to go.

CONCLUSION & LIMITATIONS

By analyzing our final linear model, we conclude that the most influential factors in room configuration that affecting Beijing's housing prices is the number of living rooms. Properties with more living rooms tend to be more expensive than those with fewer. Similarly, the number of bathrooms positively impacts housing prices, more so than the presence of kitchens. Additionally, newer properties and those located near subway stations tends to be more expensive compared to those farther away.

Several limitations must be acknowledged. Since the model was constructed using mid-range housing prices and common room types, its performance may vary when applied to datasets with higher-priced properties or unusual room configurations. Additionally, many predictor variables are highly skewed. For instance, according to the histogram below, the kitchen variable shows almost all values equal to 1, with only a small fraction differing.





This extreme distribution may significantly impacting model performance. Sufficient data about these minor observations is needed to enhance the reliability of the model. What's more, the range of construction time is only between 1990 and 2016. As a result, the model may become unreliable as the time goes by. Further studies and approaches are needed in order to better understand the housing prices in Beijing in the view point of statistician.

ACKNOWLEDGEMENTS

1. Zhang, Z (2024). Exploring Influential Factors to Housing Prices in Beijing via MLR. Unpublished manuscript.