*Science and Technology and Social Development Proceedings Series*

*Article*

*2024 International Conference on Business Economics, Education, Arts and Social Sciences (EASS 2024)*

# An Empirical Study on the Determinants of Housing Prices in Beijing and Model Optimization

**Xiang Lai** [1,*]

[1] University of Toronto, 27 King's College Circle, Toronto, Ontario M5S 1A1, Canada

[*] Correspondence: Xiang Lai, University of Toronto, 27 King's College Circle, Toronto, Ontario M5S 1A1, Canada

**Abstract:** This paper aims to investigate the potential factors influencing housing prices in Beijing and optimize the prediction model. Using the Kaggle dataset "Housing Price in Beijing", multiple factors affecting housing prices were examined, including square footage, bedroom count, bathroom count, follower count, the presence of an elevator, and subway proximity. The study initially established a multiple linear regression model (MLR) and then optimized the model through variable selection and hypothesis testing. The final model indicates that square footage, bathroom count, follower count, elevator presence, and subway accessibility significantly impact housing prices. Additionally, subway proximity positively correlates with housing prices, while increased square footage is negatively associated with price, possibly due to lower unit costs for larger properties and the higher market demand for smaller homes. By validating the model's performance using a test dataset, the final model demonstrates effective predictive capability and offers insights for future model improvements.

**Keywords:** housing prices; Beijing real estate; multiple linear regression; model optimization; subway accessibility

Published: 09 November 2024

## 1. Introduction

The primary objective is to investigate and identify potential factors that could impact housing prices in Beijing. As China's capital and economic epicenter, Beijing's real estate market holds substantial societal significance. Variations in housing prices directly impact the well-being of numerous households, the stability of the broader economic framework, and the sustainable growth of society. Understanding the fluctuation patterns of Beijing's housing prices is crucial for mitigating associated risks and making informed decisions in the real estate market.

Research into Beijing's housing market investigated various factors influencing housing prices. One study focused on the influence of Beijing's rail transfer stations on surrounding housing prices, revealing a direct relationship between home prices and transit accessibility (Dai, 2016). Greater subway availability and better transit connectivity correspond to higher home prices. Another article identified plot ratio as a significant determinant affecting housing transaction prices and established a positive correlation between them (Chen & Wu, 2016). Furthermore, an additional consideration is mixed land use. Utilizing the POI model combined with the GWR model, one article emphasized a distinct linear relationship between mixed land use and housing prices, indicating a specific impact on housing price fluctuations (Yang, 2021).

However, merely relying on the variables mentioned above from the studies is far from sufficient to grasp the entire picture. With the Kaggle dataset "Housing Price in Beijing" consisting of over 30,000 observations, we have the opportunity to delve deeper and gain a comprehensive understanding of the potential factors impacting house prices in Beijing. Given that most research articles employed a linear model (MLR), we have also chosen to adopt this approach. The primary numerical variables encompass square footage, bedroom count, drawing room count, bathroom count, and followers count. It is worth noting that the variable name of the bedroom count is "LivingRoom" in the data source. Categorical variables include ownership duration, elevator presence, building type, kitchen count, and subway proximity. Among these, the building type includes four levels (tower, bungalow, combination of plate and tower, and plate); and two levels for all others.

## 2. Methods

We first load the data into R studio and clean the data by filtering out the useless observations. After cleaning, we randomly split the resulting observations into two equal halves. One set becomes the training dataset used for the analysis process, and the other becomes the test dataset reserved for validation.

Utilizing common sense, academic research articles, and exploratory data analysis (EDA), 10 variables from the training dataset that might influence the response variable are chosen to construct the "full model." Conduct t-tests on the coefficients and select the most significant variables from the training dataset to form "model2."

Before analyzing the four assumptions, we need to check whether "model2" satisfies two conditions. Condition 1 is the conditional mean response condition. We create a scatterplot of the response variable versus the "fitted model2" to examine whether the points predominantly fall close to or align along a line. Condition 2 is the conditional mean predictor condition. We draw pairwise scatterplots of all numerical variables to ensure that their relationships do not exhibit more complexity than linearity.

If both conditions are met, the residual plots can indicate which assumption has been breached. For example, clusters suggest uncorrelated errors violation, curved patterns indicate linearity challenges, and fanning patterns signal constant variance violation. A QQ plot with minimal deviation from the 45-degree diagonal line indicates that the normality assumption is met. Variance stabilizing transformation addresses constant variance violation, while power transformations handle linearity and normality issues. If uncorrelated errors are violated, reconsider the model selection.

Upon proper transformation, "model3" is derived. Utilizing a partial F-test involves assessing the null hypothesis (H0), which posits that the beta coefficients of the removed predictors equal zero, against the alternative hypothesis (H1), where at least one beta coefficient of the removed predictors does not equal zero. Should the calculated F-value exceed the critical value, we reject H0, signifying a significant linear relationship between the response and at least one of the removed predictors. Conversely, if the calculated F-value is less than the critical value, we fail to reject H0, suggesting no significant linear relationship between the response and any of the removed predictors. Consequently, these predictors could be omitted from the model.

We subsequently examine problematic observations: leverage points distant from the center of the X-space, outliers significantly deviating from the conditional means, and three categories of influential points. Instances of problematic observationsought to be included in the limitations section. Proceed to calculate VIF: VIFs below 5 signify no severe multicollinearity concerns.

The final step is validation. We model all the variables from the previously fitted final model using the test dataset and repeat the above steps, observing whether there are big discrepancies in the data or graphs to determine whether our finalized model is optimal.

## 3. Results

### 3.1. Data Summary

The dataset is randomly divided into two parts: a training dataset and a testing dataset, each containing 50% of the total data. The training dataset is used to fit a model and the test dataset is used to assess the model validation.

**Table 1.** Summary of the Response and Numerical Variables.

| | Minimum | 1st Quantile | Median | Mean | 3$^{rd}$ Quantile | Maximm | Standard Deviation |
|---|---|---|---|---|---|---|---|
| | | | Training Dataset | | | | |
| Price | 1,096 | 30,763 | 41,795 | 46,910 | 58,067 | 145,174 | 21,695.96 |
| Square | 10.60 | 57.74 | 72.95 | 82.10 | 97.57 | 318.12 | 35.57 |
| LivingRoom | 1 | 1 | 2 | 1.97 | 2 | 7 | 0.79 |
| DrawingRoom | 0 | 1 | 1 | 1.16 | 1 | 3 | 0.52 |
| BathRoom | 0 | 1 | 1 | 1.17 | 1 | 4 | 0.41 |
| Followers | 0 | 3 | 7 | 8.96 | 14 | 24 | 6.98 |
| | | | Test Dataset | | | | |
| Price | 13,013 | 31,194 | 41,180 | 46,966 | 57,942 | 149,900 | 21,791.24 |
| Square | 18.06 | 57.50 | 73.21 | 82.40 | 97.53 | 353.44 | 36.13 |
| LivingRoom | 1 | 1 | 2 | 1.99 | 2 | 5 | 0.78 |
| DrawingRoom | 0 | 1 | 1 | 1.13 | 1 | 3 | 0.53 |
| BathRoom | 0 | 1 | 1 | 1.18 | 1 | 4 | 0.42 |
| Followers | 0 | 3 | 8 | 9.50 | 15 | 24 | 7.18 |

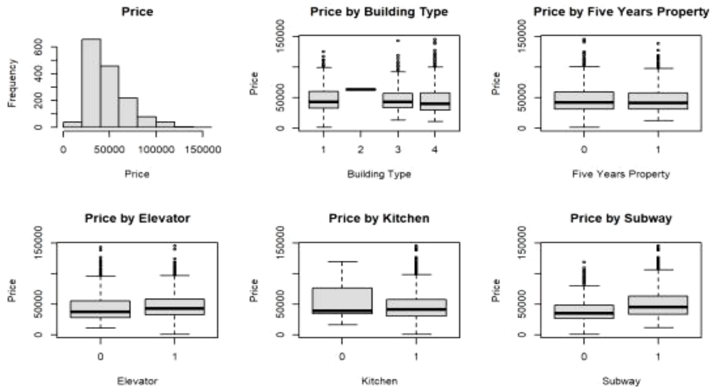

**Figure 1.** Histogram of Response and Boxplots of Categorical Variables for the Training dataset.

The mean, median, standard deviation and quantiles of response and numerical variables are summarized in Table 1. In Figure 1, the histogram indicates the distribution of observations: the response variable "price" is roughly normal. The boxplots facilitate the comparison of distributions across different levels within the categorical variables.

### 3.2. Model Selection

Using common sense and relevant articles, we form our first **full model**.

Price=42818.7-168.84[square]+1699.6[livingRoom]+383.39[drawingRoom]+6986.97[bathroom]+480.3[followers]+23423.68[buildingType2]+53.83[buildingType3]+2852.93[buildingType4]-1341.71[fiveYearsProperty]+5476.05[elevator]-8878.74[kitchen]+10751.98[subway]

The model's full F-test is significant, and individual t-tests show that 5 of the predictors are significant. The non-significant variables are then removed to fit a second model ("**model2**").

Price=37039.46-131.71[square]+7038.33[bathroom]+475.44[followers]+2743.55 [elevator]+10668.24[subway]

Model Assessments:

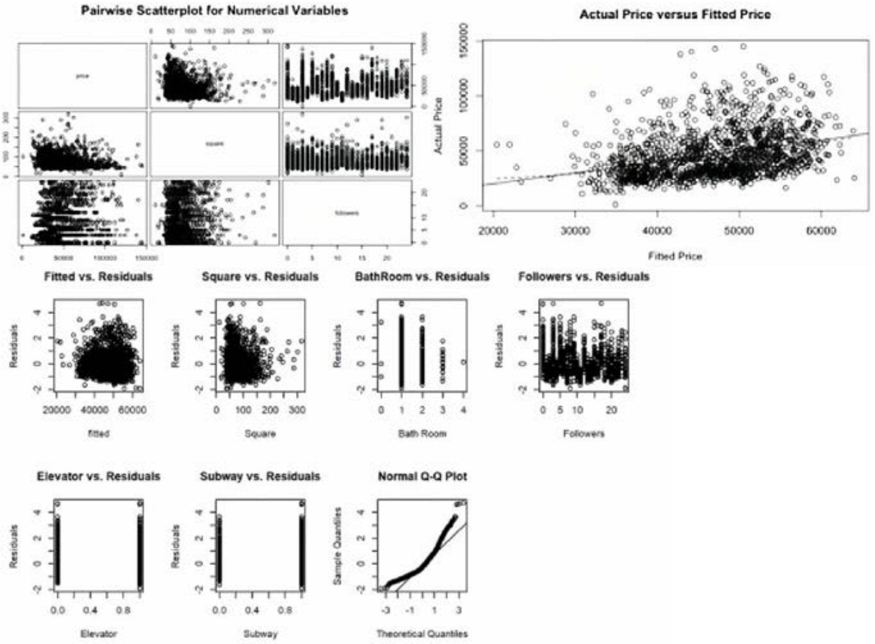

**Figure 2.** Check Additional Conditions and Model Assumptions for Model 2.

We investigate the two additional conditions and four assumptions necessary for linear regression. In Figure 2, the linear pattern between the fitted and actual "price" shows the satisfaction of condition 1; there are nonon-linear patterns in the pairwise scatterplot of predictors, therefore satisfying condition 2. However, the points in some of the residual plots are not equally spread indicating the violation of the constant variance; not all points are on the QQ line in the normal QQ plot indicating the violation of the normality.

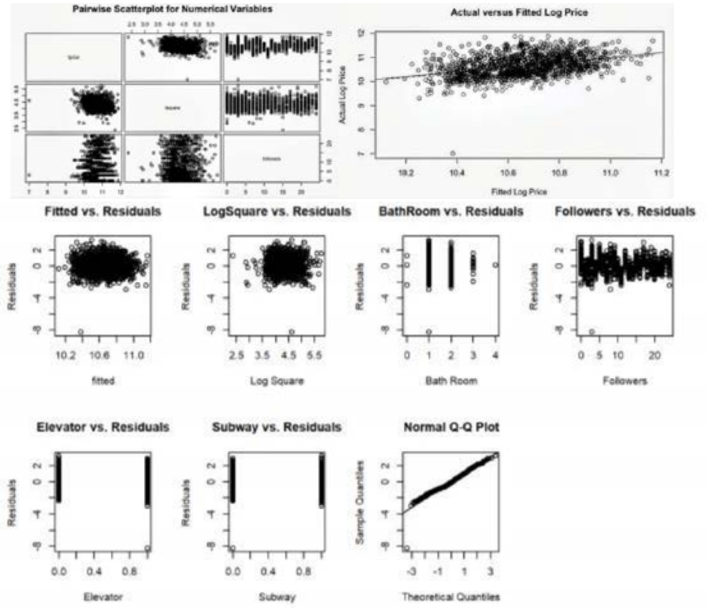

**Figure 3.** Check Additional Conditions and Model Assumptions for Model 3.

To solve assumption violations, a Box-Cox transformation is applied to response and predictors. Log transformations are applied to the response "price" and predictor variable "square"; these transformed variables are then used to fit a third model ("**model3**"). In Figure 3, despite an outlier, the two additional conditions and four assumptions are satisfied which means the transformation has improved the model.

The p-values obtained from t-tests and the F-test of Model 3 are all less than the significance level of 0.05, illustrating the individual predictors and the overall model are statistically significant.

A partial F test was applied to check if the two binary dummy variables ("elevator" and "subway") can be removed from the model or not. The resulting p-value is < 2.2e- 16 which is significant, so we reject the null hypothesis (both coefficient parameters are 0). Therefore, Model 3 is preferred.

Then, check the problematic observations in Model 3. Among the 1500 observations, there are 44 leverage points and 1 outlier. There are no influential observations on all fitted values.

The VIFs for each predictor in Model 3 are calculated to check multicollinearity. The outputs indicate that no numerical predictors have aVIF greater than 5, therefore no multicollinearity problem.

The model is applied to the test dataset. The two additional conditions and four assumptions are satisfied. All predictors and the model are significant, and the estimated coefficients are within one standard error of the results in the training dataset. Therefore, the model is considered valid as the model performance in the test dataset is similar to the training dataset. (More details in Appendix)

## 4. Discussion

### *4.1. Final Model*

Logged Price=11.416201-0.280086[Logged square]+0.122337[bathroom]+0.012222 [followers]+ 0.101952[elevator]+ 0.227567[subway]

Table 2 presents the coefficients of the final model used to predict the logged house price based on five variables. When all five variables are fixed, the predicted logged house price is the value of intercept 11.416201. In addition, when the other 4 variables are fixed, a 1% increase in square footage leads to a 0.280086% decrease in house price; for each additional bathroom, the predicted house price increases by 0.122337%;

for each additional follower, the predicted house price increases by 0.012222%; if the building has an elevator, the predicted house price increases by 0.101952%; the building near subway stations increases the predicted logged house price by 0.227567%.

**Table 2.** Coefficient of Final Model.

| | Intercept | log(square) | bathRoom | followers | Elevator | Subway |
|---|---|---|---|---|---|---|
| **Final Model 3** | 11.416201 | -0.280086 | 0.122337 | 0.012222 | 0.101952 | 0.227567 |
| **Testing Model** | 11.241908 | -0.222648 | 0.078985 | 0.015103 | 0.061504 | 0.180790 |

The final model shows transit accessibility has positive impacts on housing prices, which aligns perfectly with the relevant article (Dai, 2016). However, it seems unexpected that data suggests the larger the square of a house, the lower the price. There are several reasons for this: large properties typically have lower unit costs due to economies of scale. Factors like the allocation of fixed construction costs and diminishing marginal costs contribute to this. Additionally, smaller homes might be more popular in the housing market, especially in densely populated urban areas, because they are more economical, which could result in a relatively higher unit price for smaller houses. Moreover, larger properties may be in less populated areas where land prices are usually lower. Finally, larger houses often come with higher maintenance costs, which may make buyers hesitant to purchase, thus sellers might need to lower prices to attract buyers. These factors impact the housing price and also are more likely to interact and collectively influence the unit price of a property (Leguizamon, 2010).

In conclusion, our final model answers the research questions: factors impacting housing prices in Beijing include square footage, bathroom count, followers, elevator presence, and subway availability.

### *4.2. Limitations ofthe Model*

It is worth noting that there are slight deviations from the 45-degree diagonal line observed in the QQ plot of our final model. However, these minor deviations are not significant enough to conclusively indicate a violation of the normality assumption. Also, 44 leverage points have the potential to shift the estimated line.

Moreover, the dataset can be enriched. The inclusion of additional data points from some other markets or some different time periods of the market can further refine the model and enhance its prediction and its accuracy which will allow a wider range of variables affecting house prices to be captured.

In the future development of models, more and more advanced models can be used to capture and analyze much more complex relationships in data, which may reveal deep dynamics that current models fail to show. For example, Convolutional Neural Networks can learn deeper features making it possible to uncover more subtle non- linear relationships that affect house prices.

## 5. Conclusion

We opted for manual selection methods over automated selection methods in our analysis. Although automated selection methods offer a systematic approach to selecting a model from a plethora of predictors, they may still run even in the presence of model violations or multicollinearity. Additionally, automated selection lacks control over the inclusion or exclusion of variables, potentially leading to the removal of predictors, such as subway and square variables, which we consider essential for our analysis.

Both manual and automated methods pose ethical concerns. The manual selection method may introduce "Human Overriding Algorithms," leading to unconscious biases that impact the fairness and objectivity of the model. Furthermore, humans may be negligent, potentially affecting the model's integrity.

On the other hand, the automated selection method might result in overfitting. When people delegate responsibility to automated programs, these programs can operate regardless of the presence of model violations or multicollinearity, jeopardizing the model's fairness and accuracy. Moreover, this method lacks transparency and may prioritize data significance over hypothesis testing.

In conclusion, both methods present ethical considerations, and we view them similarly in terms of ethical implications. Both methods involve potential biases and lack complete transparency. Our choice of manual selection was driven by the need for greater control over variable inclusion and exclusion, especially concerning predictors we deemed crucial for our analysis.

## References

1. Chen, J., & Wu, D. (2016). Evaluation of Beijing low-income housing security policy: An analysis of the public policy. Open Journal of Social Sciences, 4(2), 129-139. https://doi.org/10.4236/jss.2016.42017
2. SCIRP. (2016). Evaluation of Beijing low-income housing security policy: An analysis of the public policy. https://www.scirp.org/journal/paperinformation?paperid=63831
3. Dai, X., Bai, X., & Xu, M. (2016). The influence of Beijing rail transfer stations on surrounding housing prices. Habitat International, 53, 453-461. https://doi.org/10.1016/j.habitatint.2015.12.005
4. ScienceDirect. (2016). The influence of Beijing rail transfer stations on surrounding housing prices. https://www.sciencedirect.com/science/article/abs/pii/S0197397515302198?via%3Dihub
5. Selim, H. (2008). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. Expert Systems with Applications, 36(2), 2843-2852. https://doi.org/10.1016/j.eswa.2007.10.013
6. ScienceDirect. (2008). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. https://www.sciencedirect.com/science/article/abs/pii/S0957417408000596
7. Leguizamon, S. (2010). The influence of reference group house size on house price. Real Estate Economics, 38(3), 507-527. https://doi.org/10.1111/j.1540-6229.2010.00275.x
8. Wiley Online Library. (2010). The influence of reference group house size on house price. https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6229.2010.00275.x
9. Yang, H., Fu, M., Wang, L., & Tang, F. (2021). Mixed land use evaluation and its impact on housing prices in Beijing based on multi-source big data. Land, 10(10), 1103. https://doi.org/10.3390/land10101103
10. MDPI. (2021). Mixed land use evaluation and its impact on housing prices in Beijing based on multi-source big data. https://doi.org/10.3390/land10101103

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of SOAP and/or the editor(s). SOAP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.