The Research on Factors Influencing Housing Prices-Take Beijing as an Example

Bowen Chen[1], Yifan Min[2] and Siyu Yu[3, *]

[1]School of Mathematical Sciences, Anqing Normal University, Anqing, 246052, China

[2]High School Affiliated to Fudan University, Qingpu Campus, 201799, China

[3]School of Mathematics and Statistics, Fuyang Normal University, Fuyang, 23600, China

*Corresponding author: yusiyu@stu.fynu.edu.cn

Abstract. This article aims to identify those factors that have an impact on housing prices. The method of Multiple Linear Regression is used to analyze the significant factors with 400 samples of Beijing from 2011 to 2017. Based on an assumption, 12 variables that were chosen do correlate with housing prices. This paper also considers the interaction effects between house squares and the number of living rooms, drawing rooms or bathrooms, and uses Forward Stepwise Regression to solve the covariance problem caused by adding interaction terms. In order to test the effectiveness of this operation, the research compares the VIF value and significance of those variables. It turns out that square, building type, elevator, construction time, renovation condition, and subway have a significant linear relationship with prices, while the number of living rooms, drawing rooms and bathrooms, building structure, property rights, and district fail the significance test. Overall, the volatility of housing prices in Beijing can be considered by the extent to which these factors affect them.

Keywords: Housing prices; multiple linear regression; interaction effects.

1. Introduction

As an indispensable part of the general public's life, housing prices have always been a highly valued issue for people. Since 2004, housing prices have continued to rise nationwide, becoming one of the focuses of attention in people's daily lives [1]. House prices can undergo significant changes in a few years. From 2002 to 2010, the average selling price of houses in China increased by nearly 1.5 times [2]. The fluctuation of commodity housing prices in the real estate industry has already affected the daily consumption of urban and rural residents in China [3]. However, the influencing factors are not specifically known to the public. It is essential to understand the factors contributing to housing prices. Therefore, this paper aims to help people evaluate the expected purchase of a house based on the different potential factors that may lead to house prices in Beijing's housing price research.

The real estate market is a complex system with many factors involved in housing price fluctuation. Predictions of housing prices are also a popular argument among academics. For some internal factors, Lv pointed out that housing area data is of great significance for analyzing and predicting the development of China's real estate market [4]. Yan et al. also used statistics on housing prices in Beijing, and their analysis combined Hedonic models, Lasso regression, and Random Forests, and optimized the merging results. The impact of the number of living rooms, drawing rooms, and bathrooms on house prices has been examined. This analysis is fast and predictive, and the combination of multiple models improves the robustness and accuracy of the results [5]. Zhen et al. illustrated that elevators form part of the characteristic variables of second-hand housing prices in Chengdu [6]. They conducted multiple linear regression, decision tree, and Extreme Gradient Boosting (XGboost) models to fit the prediction curves of housing prices for the influencing factors. Compared to the other two models, the XGboost algorithm was the most accurate, generalized, and robust in terms of data prediction, while avoiding overfitting. There may be a slight lack of precision in analyzing the results. Domestic scholars have also found that construction time [7] and building structure [8] all have an impact on housing prices. Zheng claimed that the degree of renovation

conditions of the house itself is an important factor in price differentiation [9]. He applied multiple linear regression and semi-logarithmic models to analyze and used Ordinary Least Squares (OLS) regression to estimate the parameters, which was to test the accuracy of the models. However, the data he used was only 219 pieces, which was relatively small. Fan et al. found that property rights significantly impact housing prices when studying housing mortgage prices [10]. For some externalities, Wang et al. used global regression models to analyze that residential prices exhibit significant spatial distribution heterogeneity under the influence of subway stations [11]. This paper focuses on twelve variables (Square, Living Room, Drawing Room, Bathroom, Building Type, Elevator, Construction Time, Building Structure, Renovation Condition, Property Rights, Subway, and District) that were studied to determine their impact on housing prices, and further select a suitable model to study the correlation between these factors and housing prices [12].

In summary, this article will use the multiple linear regression model to study the impact of these 12 factors on Beijing's housing prices.

2. Methods

2.1. Deta Source

The dataset used in this paper is fetched from the Kaggle website (Housing Price in Beijing). It was from 2011 to 2017, collected on Lianjia.com by Ruiqurm. This dataset contains 318852 groups of data, and this research selected 400 of them as samples. The original dataset remained in .csv format.

2.2. Variable Selection

The original dataset has a very large amount of data, and there are a lot of nulls for variables such as construction time, building type, and many bad values for building structure. At the same time, due to too many Days on Market (DOM) vacancies, and mixed forms of data for floor, this literature chose to remove these variables. Eventually, a random sampling is done to get 400 observations. The data contains 12 variables (Square, Living Room, Drawing Room, Bathroom, Building Type, Elevator, Construction Time, Building Structure, Renovation Condition, Property Rights, Subway, and District) and one dependent variable (Housing Price). The specific description of this dataset is shown in Table 1:

Table 1. List of Variables

Variable	Logogram	Meaning
Square	x_1	Total housing area
Living Room	x_2	The living room's number
Drawing Room	x_3	The drawing room's number
Bathroom	x_4	The bathroom's number
Building Type	x_5	Tower (1), bungalow (2), plate-tower construction (3) and plate (4)
Elevator	x_6	Whether it has elevator
Construction Time	x_7	Years of construction
Building Structure	x_8	Unknown (1), mixed (2), brick and wood (3), brick and concrete (4), steel (5) and steel-concrete composite (6)
Renovation Condition	x_9	Other (1), rough (2), simplicity (3), hardcover (4)
Property Rights	x_{10}	Whether the property is less than 5 years
Subway	x_{11}	Whether near the subway
District	x_{12}	Sixteen regions in Beijing
Housing Price	Y	Total Housing prices in Beijing

2.3. Method Introduction

The paper uses a multiple linear regression model to compare the situation with and without considering the interaction terms. This section will mainly aim to compare the significance of the two models and the accuracy of the results. Eventually, it will enable the optimized processing of models.

The multiple linear regression model is a linear regression model with multiple explanatory variables. It is used to explain the linear relationship between the explained variable and multiple other explanatory variables. Moreover, its basic principle is to estimate a set of parameters by OLS so that the sum of squares of the residuals between the dependent variables and independent variables is minimized.

3. Results and Discussion

3.1. Multiple Linear Regression

The analysis in this paper shows that there are many factors Influencing Housing Prices. As the graph shows:

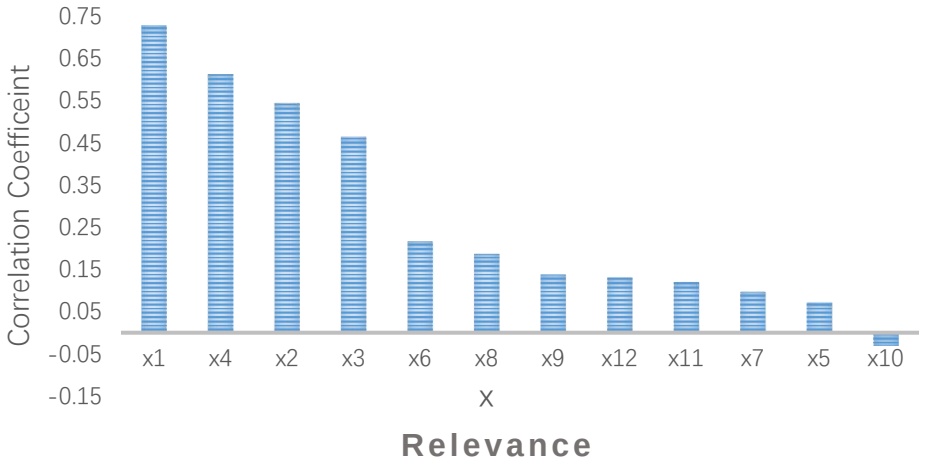

Fig. 1 Relevance Analysis Between Dependent and Independent Variables

From Figure 1, it can be seen the Pearson correlation coefficient between these factors and housing prices. The research data found that the number of living room, drawing room, and bathroom are respectively the factors that correlate most positively with house prices. So maybe nowadays people are very interested in the house type. There is a significant positive correlation between standardized residual and the house prize too. Of course, building type, construction time, renovation condition, building structure, the number of elevators are also positive correlation factors, but they are not as significant as the factors above. Surprisingly, though, the five-year property is negatively correlated with prices in Beijing. From all the above, what affect the housing prices are comprehensive. People nowadays are longing for a perfect house from many different angles. After analyzing the Pearson correlation matrix of various factors, multiple regression analysis was conducted. The general mathematical model for multiple linear regression is:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{13} x_{12} + e \quad (1)$$

In the above formula: β_0 is a constant term, and e is a residual term.

Table 2. Regression coefficient table

	B	S.E.	Beta	T	significance	VIF
Constant	14131.853	2720.631		5.194	0.000	
X_1	3.813	0.392	0.618	9.715	0.000	4.383
X_2	5.077	15.919	0.016	0.319	0.750	2.644
X_3	6.962	20.815	0.014	0.334	0.738	1.909
X_4	65.666	27.411	0.126	2.396	0.017	3.008
X_5	40.302	9.484	0.172	4.250	0.000	1.761
X_6	-7.271	1.367	-0.213	-5.321	0.000	1.726
X_7	29.928	9.798	0.096	3.055	0.002	1.070
X_8	7.130	8.651	0.047	0.824	0.410	3.512
X_9	144.880	35.126	0.250	4.125	0.000	3.959
X_{10}	-19.590	18.295	-0.034	-1.071	0.285	1.107
X_{11}	111.485	19.351	0.194	5.761	0.000	1.220
X_{12}	5.439	3.314	0.052	1.641	0.102	1.097

Table 2 shows the regression coefficients of the multiple linear regression equation model. The p-values of the T-test for the four independent variables x_1, x_4, x_5, x_6, x_7, x_9, x_{10} did not exceed 0.003. Therefore, it can be considered that all seven independent variables have a significant impact on the dependent variable Y. Based on the data above, the relevant multiple linear regression equation can be obtained:

$$E(Y) = 14131.853 + 3.813x_1 + 5.077x_2 + \cdots + 5.439x_{12} \quad (2)$$

The multivariate correlation coefficient R obtained from this model definition is 0.801, the coefficient R-squared for fitting multiple linear regression is 0.642, and the adjusted R-squared is 0.631. The model has a good fit.

The data in the normal distribution P-P in Figure 2 appears to approximately a diagonal straight line, indicating that the cumulative proportion of the data is basically consistent with the cumulative proportion of the normal distribution, and the data exhibits normality.

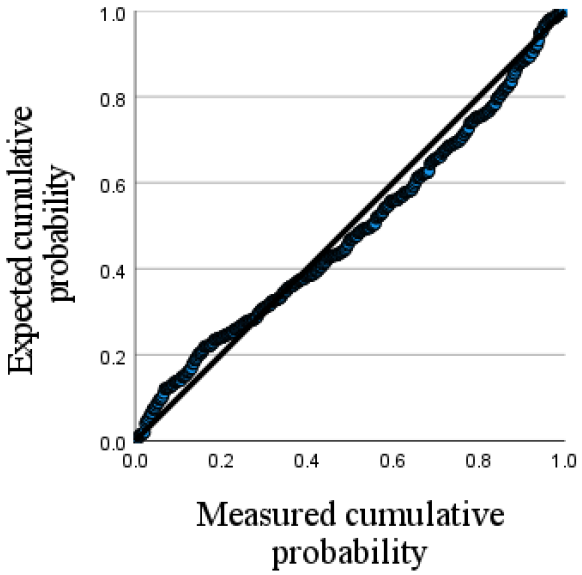

Fig. 2 Normalized P-P plots of regression standardized residuals

3.2. Multiple Linear Regression with Interaction Terms

Interactions between some independent variables may also have some effect on housing prices, and these terms with interactive effects are called interaction terms. Indeed, the number of living

room, drawing room and bathroom probably link to the size of the houses to some extent, i.e., the effect of x_1 on y depends on the values taken by x_2, x_3 and x_4. The solution is to multiply the interaction terms and add the coefficients to the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{14} x_1 x_3 + \beta_{15} x_1 x_4 + \varepsilon \quad (3)$$

Where $\beta_i (i = 1,2,\ldots,15)$ is regression coefficient, $x_1 x_2$, $x_1 x_3$ and $x_1 x_4$ are interaction terms.

If the interaction term regression coefficients are significantly positive, it indicates that larger house squares are expected to result in higher housing prices when the number of these three types of rooms is greater. With the addition of the interaction term, the significance of the original regression coefficients of the independent variables becomes less important than it was originally. The regression coefficient for x_1 is $\beta_1 + \beta_{14} x_2 + \beta_{15} x_3 + \beta_{16} x_4$, so the significance of β_1 alone does not reflect whether the overall effect of x_1 on y is significant.

The results of the analysis using the multiple linear regression model are shown in the table below:

Table 3. Multiple Linear Regression Model analysis results with interaction terms

Variables	β_i	Coefficient	T Value	P Value	VIF	Tolerance
Constant	12378.531	-	4.471	0.000**	-	-
x_1	3.188	0.517	4.084	0.000**	17.725	0.056
x_2	21.554	0.067	0.709	0.479	9.844	0.102
x_3	86.016	0.174	1.895	0.059	9.281	0.108
x_4	-58.933	-0.113	-1.214	0.225	9.639	0.104
x_5	38.76	0.165	4.115	0.000**	1.776	0.563
x_6	150.41	0.259	4.309	0.000**	3.996	0.25
x_7	-6.359	-0.186	-4.563	0.000**	1.834	0.545
x_8	5.86	0.039	0.684	0.495	3.523	0.284
x_9	28.543	0.092	2.936	0.004**	1.076	0.929
x_{10}	-20.413	-0.036	-1.12	0.263	1.122	0.892
x_{11}	105.681	0.183	5.49	0.000**	1.234	0.81
x_{12}	6.058	0.058	1.84	0.067	1.106	0.904
$x_1 x_2$	-0.083	-0.061	-0.34	0.734	35.87	0.028
$x_1 x_3$	-0.67	-0.28	-1.578	0.115	34.716	0.029
$x_1 x_4$	0.946	0.5	3.049	0.002**	29.742	0.034

Note: ** denotes $p < 0.01$, which shows a significant effect.

There are many variables with VIF values greater than 5 and low correlation coefficients, which indicates that the inclusion of interaction terms leads to severe covariance problems in x_1, x_2, x_3, x_4, $x_1 x_2$, $x_1 x_3$, $x_1 x_4$. To solve this problem, Forward Stepwise Regression is used. There is the obtained from the regression analysis of the independent and dependent variables (Table 4):

Table 4. Results of Forward Stepwise Regression

Variables	β_i	Coefficient	T Value	P Value	VIF	Tolerance
Constant	13104.294	-	5.143	0.000**	-	-
x_1	3.028	0.491	6.555	0.000**	6.165	0.162
x_5	38.007	0.162	4.088	0.000**	1.719	0.582
x_6	159.323	0.274	6.143	0.000**	2.192	0.456
x_7	-6.673	-0.195	-5.209	0.000**	1.541	0.649
x_9	28.164	0.09	2.965	0.003**	1.021	0.979
x_{11}	101.979	0.177	5.51	0.000**	1.133	0.882
$x_1 x_4$	0.524	0.277	3.691	0.000**	6.194	0.161

The model identifies x_1, x_5, x_6, x_7, x_9, x_{11} and x_1x_4 as explaining 64.3% of the change in y. The model passes the F-test ($F = 100.881, P = 0.000 < 0.05$), which indicates that the model is valid. Then the model formula is:

$$y = 13104.294 + 3.028x_1 + 38.007x_5 + 159.323x_6 - 6.673x_7 + 28.164x_9 + 101.979x_{11} + 0.524x_1x_4 \quad (4)$$

However, x_1 and x_1x_4 still have some covariance issues, and they both have smaller regression coefficients. The next step is to choose to remove them and perform the regression analysis again:

Table 5. Results of the Improved Forward Stepwise Regression

Variables	β_i	S.E.	Coefficient	T Value	P Value	VIF	Tolerance
Constant	13773.926	2582.187	-	5.334	0.000**	-	-
x_1	4.571	0.199	0.742	22.939	0.000**	1.112	0.899
x_5	43.207	9.337	0.184	4.627	0.000**	1.68	0.595
x_6	169.33	26.204	0.292	6.462	0.000**	2.168	0.461
x_7	-7.057	1.297	-0.206	-5.439	0.000**	1.531	0.653
x_9	30.405	9.63	0.098	3.157	0.002**	1.017	0.983
x_{11}	98.06	18.771	0.17	5.224	0.000**	1.13	0.885

According to the table 5, the significance of all variables is high but not obviously changed from the previous value. x_1, x_5, x_6, x_9 and x_{11} have a significant positive effect on y, and x_7 has a significant negative effect on y. Not only do the corresponding regression coefficients all become larger, but there are no covariance problems. Therefore, it will derive a revised model:

$$y = 13773.926 + 4.571x_1 + 43.207x_5 + 169.33x_6 - 7.057x_7 + 30.405x_9 + 98.06x_{11} \quad (5)$$

Furthermore, the fit and the adjusted R^2 of the model are 63.1% and 0.625 respectively, which is relatively great. The model passes the F-test ($F(6,393) = 111.832$, $P = 0.000$) indicating its validity.

4. Conclusion

The study selected 400 samples from 2011 to 2017 from the data set, which has 12 variables. Our method (Multiple linear regression analysis) is accurate, effective, and comprehensive. Because it performs a multifactor analysis and then gets the Pearson correlation coefficients of each variable.

During the analysis stage, the article uses a multiple linear regression model, which is used to find out the possible relationship between the variables and housing prices. To figure out more, the research takes interaction effects into account and adds interaction terms with coefficients to the equation. Therefore, the factors that positively impact on house prices are the number of living rooms, drawing rooms, and bathrooms, building type, construction time, renovation condition, building structure, and the number of elevators. The five-year property is negatively correlated with prices in Beijing. From all of these, square, building type, construction time, renovation condition, and subway are the main factors.

With the research, people longing for dream houses can have a reference from different angles, and then have an overall determination on the budget of house price. However, there are still some deficiencies such as the drawbacks are that causal relationships between variables can't be found, the sample size is relatively small, and the data is not the latest version. To improve this, searching for new data and using the control variable method to find out possible causalities between house price and factors.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Wu Zhenkui, Tang Wenguang, Wu Bin. Using the Priority Factor Method to Analyze the Impact of House Price Factors on Buyers' Orientation. Journal of Tianjin University of Commerce, 2007, 27(3).

[2] Hu Qiang. Analysis of housing price factors based on the SVAR model. Times Finance, 2017.

[3] Yang Dianxue, Zhang Zhimin. An empirical study on incorporating housing price factors into China's CPI. Statistics＆Information Forum, 2013, 28(3).

[4] Lv Chenyue, Liu Yingxin, Wang Lidong. Analysis and Forecast of Influencing Factors on House Prices Based on Machine Learning. Proceedings of 3rd International Symposium on Information Science and Engineering Technology, 2022, 117-121.

[5] Yan Ziyue and Zong Lu. Spatial Prediction of Housing Prices in Beijing Using Machine Learning Algorithms. In Proceedings of the 2020 4th High-Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence (HPCCT & BDAI '20). Association for Computing Machinery, New York, NY, USA, 2020, 64-71.

[6] Peng Zhen, Huang Qiang, Han Yincheng. Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm. 2019 IEEE 11th International Conference on Advanced Infocom Technology (ICAIT). IEEE, 2019.

[7] Pan Jia, Luan Yaoyao, Hong Xiaoqing, Li Min. Analysis and prediction of second-hand housing prices in Qingdao based on integrated algorithms. Advances in Applied Mathematics, 2023, 12(4): 1671-1682.

[8] Wang Xiaojuan. Research on the impact of second-hand housing prices in Chongqing. Journal of Langfang Normal University (Natural Science Edition), 2019, 19(3).

[9] Zheng Yongfeng. Research on the spatial difference of housing prices in different urban areas of Hangzhou. Economic Forum, 2007, 20: 32-34.

[10] Fan Gangzhi, Li Han, Li Jiangyi, Zhang Jian. Housing property rights, collateral, and entrepreneurship: Evidence from China. Journal of Banking and Finance, 2022.

[11] Wang Nan, Wu Wei, Hu Xiying, et al. The Heterogeneity of the Impact of Major Transportation Facilities on Residential Prices under Urban Crossing Rivers: A Case Study of Binjiang New City in Nanchang City. Urban Studies, 2018, 10: 123-130.

[12] Yang Chenggang, Li Haibin. Population Migration, Changes in Residential Supply and Demand, and Regional Economic Development: An Economic Analysis of the Current "Man Snatching War" in Domestic Cities. Theoretical Investigation, 2019, 3: 93-98.