Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/38/20240561

Research on housing prices prediction based on multiple linear regression

Qijian Wang

The Second-High School Affiliated to Beijing Normal University, Beijing, 110101, China

100566@yzpc.edu.cn

Abstract. With the steady development of social economy, commercial housing, as an important real estate, occupies a large proportion in family assets. According to the "China Household Wealth Survey Report" (2018) compiled by the Social China Economic Trends Institute, household net worth accounts for 70% of household wealth, including housing prices in Beijing and Shanghai. In higher cities, the proportion is as high as 80%. This paper analyzes the transaction data of about 10,000 second-hand houses in Beijing, constructs a multiple regression model with SPSS software, and obtains the dependent variable (housing price per unit area). The dataset used in this paper is fetched from the Kaggle website (Housing Price in Beijing). The results show that the relationship between the elevator, the floor situation, the decoration method, the administrative division and other independent variables. Also, it is shown that the correlation between the two is significant, so the model can be used. This paper provides reference for the actual transaction of second-hand housing in Beijing.

Keywords: Second-hand house, price, multiple linear regression.

1. Introduction

With the development of China's real estate market, the problem of housing price has become the focus of people's attention, and the commodity housing is related to it. The real estate economy is an important part of China's economy, which supports many industries such as healthcare, railways and public utilities. Real estate has always maintained a high heat in the market, and the real estate industry is one of the important driving forces supporting China's economic growth. Its price is affected by a variety of factors, the function of different locations, public service facilities and housing construction structure have a great impact on the price of second-hand housing. Therefore, it is of great significance to study the influencing factors of second-hand house prices and the uncertainties of each factor in different regions for the regulation of housing prices and the steady and healthy development of the real estate market.

There are many papers that have studied the factors that affect the price of second-hand homes, and here are some main ideas of the other paper. This paper analyzes the monthly average price data of second-hand housing in Shenzhen, and shows that the second-hand housing guidance price policy introduced by Shenzhen government in 2021 will reduce the average price of second-hand housing in the long run [1]. Based on the multiple linear regression analysis of second-hand housing transaction data in Beijing from June 2021 to June 2022, this paper discusses the influencing factors of second-hand housing prices. The results show that the decoration method has a significant impact on the second-hand

DOI: 10.54254/2753-8818/38/20240561

house price. Through proper adjustment, a regression model with good fitting effect was obtained [2]. The author pointed out that in response to the problem of too fast rising housing prices, the central government has put forward the concept of "housing without speculation" and the "three stability" regulation target [3]. This paper introduces the importance of the second-hand housing market and its influence by many factors. The author used GWR model to study the second-hand house price in Wuhan in 2018, and found that architectural features, community environment and public service facilities have significant effects on the second-hand house price, and there is spatial heterogeneity. The effective sample data is obtained through POI data, and the influencing factors are analyzed, which provides decision-making enlightenment for the healthy and steady development of the real estate market [4, 5]. The study pointed out that since the 1990s, the second-hand housing market in Changchun has shown an overall upward trend, but there are significant gaps in different regions [6]. The study uses Internet data to analyze the influencing factors of second-hand house price in Changchun more comprehensively. Based on the random forest theory, the evaluation model of the characteristic price of second-hand houses is used to quantify the influencing factors, in order to provide beneficial decision-making suggestions for the second-hand house trading market in Changchun [7, 8]. Here are some ways to optimize the house market. Improving the evaluation and verification mechanism for guiding the transfer price of second-hand housing as soon as possible. Each local governments should strengthen information communication with the trading market and accurately grasp the housing market the market price, to avoid the transfer of the guide price and the actual transaction price gap is too large.

2. Methodology

2.1. Data 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 amount of data in the original data set is very large, and there are many empty values for the construction time, building type and other variables, and many bad values for the building structure. The data contains 11 variables. The specific description of this dataset is shown in Table 1:

Table 1. Variable introduction.

Variable	Meaning
Direction	Orientation of the house
District	Which street
Elevator	Lift or not
Floor	Floor of a building
Garden	Concrete cell
Layout	The internal structure of the house
Price	Housing prices in Beijing
Region	The area where the house is located
Renovation	Hardcover or not
Size	Area of the house
Year	Years of construction

DOI: 10.54254/2753-8818/38/20240561

2.3. Method introduction

Regression analysis is employed to study the impact of X (quantitative or categorical) on Y (quantitative), including determining whether there is an influence relationship, the direction, and degree of influence [9, 10].

Firstly, the model's fit is analyzed through the examination of the R-squared value, along with the assessment of the Variance Inflation Factor (VIF) to detect any collinearity issues. A VIF value greater than 5 indicates potential collinearity problems, while a tolerance value (tolerance = 1/VIF) below 0.2 also suggests collinearity. This step aims to identify any collinearity problems in the model. Next, the significance of X is analyzed. If the p-value is less than 0.05 or 0.01, it indicates that X significantly influences Y. The direction of this influence is then examined in detail. Thirdly, the influence degree of X on Y is compared and analyzed by considering the regression coefficient (B) values. Finally, the analysis results are summarized, and implications are discussed. Results and discussion: This section provides a comprehensive summary of the analysis findings, discusses their implications, acknowledges any limitations of the model, and suggests potential avenues for future research.

3. Results and discussion

3.1. Model results

As can be seen from the above table, North-South, Dongdan, no elevator, 2 rooms and 1 hall, Dongcheng, hardcover, Year is taken as the independent variable and house price is taken as the dependent variable for linear regression analysis. As can be seen from the table 2.

The values of these coefficients are estimated, and the multiple linear regression model is obtained. the R-square value of the model is 0.703, which means that north-south, Dongshan,6, no elevator, 1.01×10^{11} , 2 rooms and 1 hall, Dongcheng,60, hardcover,1988 can explain 70.3% of the change of 705. During the F test of the model, it is found that the model passes the F test (F=16.572, p=0.000<0.05), which means that at least one item of North-South, Dongshan,6, no elevator, 1.01×10^{11}, 2 rooms and 1 hall, Dongcheng, 60, hardcover, 1988 will have an impact on 705. According to the multicollinearity test of the model, it is found that all the VIF values in the model are less than 5, which means that there is no collinearity problem. Moreover, the D-W value is near the number 2, which indicates that there is no autocorrelation in the model, and there is no correlation between the sample data, and the model is good. The final concrete analysis shows that:

The north-south regression coefficient value is -0.701 (t = -0.310, p = 0.757 > 0.05), indicating that the North-South orientation does not significantly influence the value of house price. The regression coefficient value for Dongdan is -1.281 (t = -2.837, p = 0.006 < 0.01), suggesting that Dongdan has a significant negative impact on the value of house price. The regression coefficient value for "no elevator" is 196.458 (t = 1.931, p = 0.057 > 0.05), suggesting that the absence of an elevator does not significantly influence the value of house price. Similarly, the regression coefficient for "2 rooms and 1 hall" is 10.366 (t = 1.495, p = 0.139 > 0.05), suggesting that it does not significantly affect the value of house price. The regression coefficient value for Dongcheng is 4.527 (t = 0.622, p = 0.536 > 0.05), indicating that Dongcheng does not significantly influence the value of 705. The regression coefficient value for "hardcover" is -31.470 (t = -1.598, p = 0.114 > 0.05), indicating that it does not significantly influence the value of house price. Finally, the regression coefficient value for "Year" is 3.974 (t = 0.953, p = 0.344 > 0.05), suggesting that it does not significantly affect the value of house price. Summary analysis shows that 60 has a significant positive influence on 705. And Dongdan will have a significant negative influence on house price. However, North-South, 6, no elevator, 1.01×10^{11}, 2 rooms and 1 hall, East City, hardcover, year will not have an impact on house price.

DOI: 10.54254/2753-8818/38/20240561

Table 2. Linear regression analysis results.

	Nonnormalized		Standard	t	p	Collinearity diagnosis	
	B	SE	Beta			VIF	tolerance
Constant	-1.218×10^5	6.01×10^6	-	-0.203	0.840	-	-
North and south	-0.701	2.260	-0.023	-0.310	0.757	1.243	0.804
Dongdan County	-1.281	0.452	-0.201	-2.837	0.006**	1.179	0.848
walk-up	196.458	101.724	0.222	1.931	0.057	3.111	0.321
2 rooms and 1 hall	10.366	6.933	0.141	1.495	0.139	2.101	0.476
The eastern part of the city	4.527	7.282	0.045	0.622	0.536	1.244	0.804
hardcover	-31.470	19.689	-0.115	-1.598	0.114	1.230	0.813
Year	3.974	4.168	0.087	0.953	0.344	1.966	0.509
R^2	0.703						
adjust R^2	0.661						
F	F (10,70)=16.572, p=0.000						
D-Wvalue	2.145						

Dependent variable: 705

* p<0.05 ** p<0.01

3.2. Model test

As can be seen from table 3, North-South, Dongdan,6, no elevator, 2 rooms and 1 hall, Dongcheng,60, hardcover, year is taken as the independent variable and house price is taken as the dependent variable for linear regression analysis. As can be seen from the above table, the R-square value of the model is 0.703. It means that north-south, Dongdan,6, no elevator, 2 rooms and 1 hall, Dongcheng, hardcover, year can explain 70.3% of the change of 705.

Table 3. Collinearity test results.

	coefficient	95% CI	Collinearity diag.	
			VIF	tolerance
constant	-1.218×10^5 (-0.203)	1.299×10^6-1.055×10^7	-	-
North and south	-0.701(-0.310)	-5.130 ~ 3.728	1.243	0.804
Dongdan County	-1.281**(-2.837)	-2.166 ~ -0.396	1.179	0.848
walk-up	196.458(1.931)	-2.918 ~ 395.833	3.111	0.321
2 rooms and 1 hall	10.366(1.495)	-3.223 ~ 23.955	2.101	0.476
The eastern part of the city	4.527(0.622)	-9.745 ~ 18.800	1.244	0.804
hardcover	-31.470(-1.598)	-70.060 ~ 7.120	1.230	0.813
year	3.974(0.953)	-4.195 ~ 12.144	1.966	0.509
Sample size		81		
R 2		0.703		
F value		F (10,70)=16.572,p=0.000		

D-W value: 2.145, * p<0.05 ** p<0.01

As can be seen from table 4, when F-test was performed on the model, it was found that the model passed the F-test (F=16.572, p=0.000<0.05), which means that the model construction is meaningful.

DOI: 10.54254/2753-8818/38/20240561

Table 4. Model summary.

	Sum of squares	df	Mean square	F	p value
regression	1.026×10^7	10	1.026×10^6	16.572	0.000
Residual error	4.334×10^6	70	6.191×10^4		
total	1.459×10^7	80			

4. Conclusion

The article sourced data from the home network to analyze the Beijing second-hand housing market from June 2021 to June 2022, constructing a robust multiple regression model to understand the influence of various factors on the unit area housing price. By inputting actual data into this model, accurate second-hand house prices can be obtained. However, a notable limitation of the model is its reliance on a limited number of price factors (variables). In addition to the factors discussed in the article, there exist several other influential variables such as the proximity of the house to subway stations, house type (e.g., standalone villas, bungalows, apartments, school district houses), age of the house, house orientation, surrounding environment, real estate policies, housing supply and demand dynamics, and the overall economic development level. Although these factors are acknowledged for their impact on housing prices, they were not included in the research scope of this paper due to the difficulty in quantifying them accurately. Nevertheless, the author intends to further optimize the model by incorporating these additional influential factors in subsequent research endeavors.

References

[1] Zhou J and Ma S P 2024 SPSSAU research data analysis methods and applications. The 1st edition. Publishing House of Electronics Industry.
[2] Sun D 2000 Selection of the Linear Regression Model According to the Parameter Estimation. Wuhan University Journal of Natural Sciences, 5(4), 400-405.
[3] Zhang H C and Xu J P 2009 Modern psychology and educational statistics. 3rd edition. Beijing Normal University Press.
[4] Gao Y, et al. 2022 The spatial differentiation pattern and influencing factors of housing prices in Shanghai's tourism and accommodation industry. Geography, 42(8), 11.
[5] Pan H 2023 Average of variable coefficient quantile regression model with variable weights. Progress in Applied Mathematics, 12(11), 10.
[6] Wang X J 2019 Research on the impact of second-hand housing prices in Chongqing. Journal of Langfang Normal University (Natural Science Edition), 19(3).
[7] Fan G Z, et al. 2022 Housing property rights, collateral, and entrepreneurship: Evidence from China. Journal of Banking and Finance.
[8] Wang N, et al. 2018 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, 10, 123-130.
[9] Yang C G and Li H B 2019 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, 93-98.
[10] Wang C Y and Zhang Z Y 2023 The Impact of Urban Public Service Allocation on Housing Prices: A Case Study of Chongqing City Journal of Chongqing Jiaotong University (Social Sciences Edition), 23(6), 58-67.