Research on the Influencing Factors of Beijing Housing Price

Zihang Wang*

Pocklington school, York, YO42 2NJ, United Kingdom

*Corresponding author: zihangwang1@outlook.com

Abstract. This paper aims to develop a house price prediction system using a multiple linear regression model to improve the accuracy of predicting housing prices by analyzing the linear relationships among various influencing factors. Fluctuations in the real estate market have significant impacts on urban development and residents' quality of life, making precise house price forecasting essential for both real estate research and informed decision-making. Traditional prediction methods are often rooted in economic theories, but advancements in data science have introduced more data-driven approaches that can leverage vast datasets to enhance predictive power. Multiple linear regression, a classic statistical technique, is particularly suitable for examining the linear connections between multiple independent variables and a dependent variable, in this case, housing prices. This study incorporates factors such as location, infrastructure, economic indicators, and property characteristics to construct a predictive model, subsequently validating its effectiveness through real-world data analysis and the effect of Variance Inflation Factor (VIF) on the model is discussed.

Keywords: House price prediction; multiple linear regression; feature selection; model evaluation.

1. Introduction

In today's international situation, the economics situation of various countries in the world is not stable, and house price are a very important indicator of as countries economy. Since 2004, housing prices have continued to rise nationwide, becoming one of the focuses of attention in people's daily lives [1]. Nowadays the real estate market is a standout amongst the most focused regarding pricing and keep fluctuating [2]. As the influence of COVID-19 and the war between Israel-Palestine, Russia-Ukraine the volatility of house prices will be magnified indefinitely so understanding and anticipating house price trends is essential for real estate companies to invest, for the grassroots to spend money in the property market and for the country's economic growth. The real estate market is a huge economic system, in the world housing prices have been of great concern, for the prediction of housing prices is also a major object of discussion in the academic community, many scholars have made predictions on housing prices with different statistical methods House price prediction, or residential property valuation, is a difficult problem, as real estate valuations do not depend on only physical characteristics of the building itself but also its location [3].

The price of commercial property is affected by changes in many factors, firstly, factors affecting the demand for housing, such as per capita income, GDP, and the level of consumption by residents, and secondly, factors affecting the supply of commercial property, such as land prices, construction costs, and bank loans. The second is the influence of the supply of commercial property, such as land prices, construction costs, bank loans [4]. Which is a point Qiao makes in his paper, he uses Back Propagation Neural Network to predict the future housing price and high predictive accuracy and confidence after realistic validation. Yan and Zong used K-nearest neighbor method, decision tree, random forest and support Vector Machine to construct and test the house price and its influencing factors in 35 large and medium-sized cities across China from 1998 to 2019. And they thinkRandom forests should be the optimal model among these 4 types of models but still Insufficient precision [5]. Some research predicted house prices in Liuzhou using Markov chains, but it is only using 3 years data [6]. Some research used Accumulated Long and Short Term Memory Models (AG)-LSTM predict the average housing price of a district, This makes the choice of location more accurate, but the model is based on the average monthly price of the neighborhoods. But is based on the average monthly price of the neighborhoods and does not predict the price of specific houses in the

neighborhoods [7]. In a given city or region, as a result of economic development and population growth, demand increases, resulting in higher equilibrium prices for dwellings, and a resulting in higher equilibrium prices of residential property, Beijing being one of them [8]. After nearly a decade of rapid development, China's property market has grown to become the country's most important pillar industry [9]. Its healthy development is of great significance in boosting GDP growth, adjusting industrial structure and promoting the sustainable and coordinated development of China's economy [10]. So in this paper, the author will use a multiple linear regression model to predict the house prices in Beijing

2. Methods

2.1. Data Source

The dataset used in this essay is fetched from the Kaggle website (Housing Price in Beijing). It was from 2011 to 2017 (and some other years between 1999-2010), fetching from Lianjia.com. This dataset contains 318852 groups of data, and this research selected 100 of them as samples. The original dataset remained in .csv format.

2.2. Variable Selection

The original dataset has close to 300,000, pieces of data and it includes URL, ID, Lng, Lat, CommunityID, Trade Time, DOM (Days on Market), Followers, Total Price, Price, Plaza, Living Room, Number of Living Rooms, Kitchen and Bathroom, Type of Building, Construction Time. Furnishing Condition, Building Structure, Ladder Ratio. It describes how many ladders residents have on average), lifts, five-year ownership, metro, neighborhood, average price of the neighborhood.in this essay this paper will choose 14 Variable to analyzed which is floor, ladder radio, decoration conditions kitchen living room and square illustrate in Table 1.

Table 1. List of variables

Variable	logogram	connotation
square	X1	Total area
Living room	X2	Wheater or not have a subway
Drawing room	X3	The number of drawing room
kitchen	X4	the number of kitchen
Building type	X5	including tower(1) , bungalow(2)， combination of plate and tower(3), plate(4).
Renovation condition	X6	including other(1), rough(2),Simplicity(3), hardcover(4)
Building structure	X7	including unknow(1), mixed(2), brick and wood(3), brick and concrete(4),steel(5) and steel-concrete composite (6).
Elevator	X8	have (1) or not have elevator(0)
Subway	X9	have (1) or not have subway(0)
District	X10	13 area of beijing
Bathroom	X11	Number of bath room
ladder Ratio	X12	the proportion between number of residents on the same floor and number of elevator of ladder. It describes how many ladders a resident have on average.
floor	X13	Height of the house
Construction time	X14	the time of construction

2.3. Method Introduction

Multiple Linear Regression (MLR) is a statistical model widely used to establish the relationship between a dependent variable and multiple independent variables. In this model, it is assumed that there is a linear relationship between the dependent variable and multiple independent variables. The core idea of this process lies in the least squares method and the fitting of linear equations.

3. Results and Discussion

3.1. Multiple Linear Regression (MLR)

Table 2 gives the regression coefficients of the multiple linear regression equation model, in which x3, x4, x13 do not have a significant effect on total price, while the other eleven independent variables have p-values of less than 0.05, so the eleven independent variables x1, x2, x5, x6, x7, x8, x9, x10, x11, x12, x14 have a significant effect on the dependent variable. significant effect, according to the above data, this paper can get the corresponding multiple linear regression equation.

$$E(Y) = 13999.266 + 4.465x1 - 32.354x2 + \cdots - 7.269x14 \quad (1)$$

Table 2. Regression coefficient table

	B	SE	Beta	t	p	VIF
constant	13999.266	2224.073		6.294	0.000	
x1	4.465	0.266	0.702	16.801	0.000	5.143
x2	-32.354	12.432	-0.087	-2.602	0.009	3.308
x3	-21.605	16.040	-0.035	-1.347	0.178	2.018
x4	50.384	40.780	0.024	1.235	0.217	1.118
x5	17.655	7.563	0.057	2.334	0.020	1.767
x6	33.784	7.982	0.080	4.233	0.000	1.043
x7	16.491	7.032	0.076	2.345	0.019	3.100
x8	105.592	28.287	0.129	3.733	0.000	3.523
x9	128.321	16.150	0.161	7.946	0.000	1.203
x10	8.655	2.769	0.059	3.126	0.002	1.034
x11	58.204	18.712	0.098	3.111	0.002	2.897
x12	300.134	45.282	0.164	6.628	0.000	1.794
x13	-0.261	0.167	-0.030	-1.565	0.118	1.102
x14	-7.262	1.114	-0.154	-6.525	0.000	1.646

The square of R shown in the table 2 above represents the strength of the explanation of the dependent variable Y by the analytical phase X. In the multiple linear regression model shown in the table above the R is 0.826, the R-squared is 0.683, and the adjusted R-squared is 0.678 which means that the model has a good fit of 67.8 per cent, which means that the independent variable above explains 68.3 per cent of the variations in the house prices with a good degree of accuracy.

3.2. VIF Problem

The VIF value described in the above table is used to judge the covariance problem between all the dependent variables, the author thinks that the VIF value is less than 5 means that there is no covariance between the dependent variables, from the above table 2 the author can see that the VIF value of square is greater than five, so there is covariance problem between it and other dependent variables (living room, kitchen, drawing room, bath room) and the author thinks the two variables ladder Ratio and elevator have some similarity, so I remove square and ladder Ratio then remove the variables that do not have a significant effect on the model to re-run the multivariate row-limit regression model to reduce the covariance problem on the model interference.

After removing these independent variables, this paper analyzed these remaining variables (x2,x5,x6,x7,x8,x9,x10,x11,x14) again with the multiple linear regression model, as shown in the Table 3, the values of VIF are all lower than five, which excludes the problem of covariance, the author can get the corresponding multiple linear regression equation:

$$E(Y) = 4787.423 + 113.314x2 + 22.631x5 + \cdots - 2.676x14 \quad (2)$$

Table 3. Regression coefficient table after selection

	B	SE	Beta	t	p	VIF
Constant	4787.423	2690.380		1.779	0.075	
X2	113.314	12.223	0.305	9.270	0.000	2.054
X5	22.631	9.029	0.073	2.506	0.012	1.617
X6	26.074	9.932	0.061	2.625	0.009	1.037
X7	14.941	8.758	0.069	1.706	0.088	3.087
X8	154.142	20.061	0.193	7.684	0.000	1.192
X9	5.196	3.444	0.035	1.509	0.132	1.027
X10	253.421	19.661	0.425	12.889	0.000	2.054
X11	163.267	34.909	0.200	4.677	0.000	3.445
X14	-2.676	1.351	-0.057	-1.981	0.048	1.553

The accuracy of the new model changed, this paper found that the R-square of the multiple linear regression model was drastically lowered after this adjustment to only 0.504, and the adjusted R-square was only 0.499, which is a decrease of Nearly 19 percentage points, which shows that the remaining dependent variable is not strong enough to explain the house price, so the author thinks the size of the house has a great influence on the change of the house price.

In order to change the problems generated by the previous model, the author re-ran the multiple linear regression model, this time the author re-added the size of the room and removed the living room and bath room associated with it, the remain variable is(x1,x5,x6,x7,x8,x9,x10,x14) the result illustate in table 4, the author can get the corresponding multiple linear regression equation.

$$E(Y) = 11681.259 + 4.971x1 + 30.208x5 + \cdots - 6.101x14 \quad (3)$$

Table 4. Final Regression coefficient table

	B	SE	Beta	t	p	VIF
Constant	11681.269	2210.226		5.285	0.000	
X1	4.971	0.125	0.781	39.865	0.000	1.064
X5	30.208	7.430	0.098	4.066	0.000	1.602
X6	38.171	8.158	0.090	4.679	0.000	1.024
X7	17.424	7.233	0.080	2.409	0.016	3.081
X8	132.016	16.503	0.165	7.999	0.000	1.180
X9	6.704	2.841	0.045	2.359	0.019	1.023
X10	118.015	28.801	0.144	4.098	0.000	3.431
X14	-6.101	1.111	-0.129	-5.494	0.000	1.536

After this analysis in the VIF decreased at the same time, the R-squared re-increased to 0.660, and the adjusted R-squared of 0.658 regained more than 65% of the Strength of interpretation and the model pass the F-test F (8,940)=228.520, p=0.000 Demonstrated stability of the model (table 4).

4. Conclusion

In this study, 1000 samples were selected from the data set since 2000 and 14 different dependent variables were used to analyses the house prices accurately, efficiently and diversely, this paper uses multiple linear regression model to find out the relationship between different independent variables

and the final total price in order to further reduce the error, this paper takes into account the effect of covariance on the model and by controlling the number of independent variables, as well as screening and analyzing the independent variables that are repetitive and removing similar variables. It is concluded that the type of house, the structure of the house, the number of metros in the neighborhoods, the size of the house, the degree of decoration, whether it has a lift or not, and the area in which the house is situated all have a significant positive impact on the housing price, with size being the main factor affecting the price.

Through the research in this paper, consumers can make reference to the houses they are going to buy and judge their own budget before buying a house through different perspectives, but there are still shortcomings in this study, for example, this paper only uses 1,000 pieces of data to analyses, the sample size is small and the data are not the latest data. In addition to this, there may be a non-linear relationship between some of the independent variables and the dependent variable house price, and the final model still has an inaccuracy rate of close to 35%, in order to improve these problems can be further improved by adding new data and analyzing them using different statistical methods to further improve the accuracy, and as economics is a social science, fluctuations in house prices, may also be affected by social factors such as in the New Crown Epidemic in 2019 has had an impact on the world financial markets, which has had an impact on house prices, which cannot be analyzed by statistics.

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.