WEP Warwick Evans Publishing

Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 12
EDI 2024

Analysis of Tianjin real estate market: multiple linear regression prediction method

Juntai Yi

Changjun High School International Department, Changsha, 410002, China

Abstract. In this paper, the price prediction of Tianjin real estate market is studied, and the key factors affecting housing prices are systematically analyzed by using the method of multiple linear regression prediction. The research data comes from the real estate transaction data of Tianjin from 2018 to 2022, covering independent variable information such as housing area, housing age, geographical location and supporting facilities. By constructing multiple linear regression model, it is found that housing area and supporting facilities have a significant positive impact on housing prices, while housing age and distance from the city center have a significant negative impact on housing prices. Specifically, for every 1 square meter increase in housing area, house prices are expected to rise by 450 yuan; Every year the age of the house increases, the house price is expected to drop by 200 yuan; Every kilometer away from the city center, house prices are expected to drop in 800 yuan; Every time the score of supporting facilities is increased by 1 point, the house price is expected to increase by 1200 yuan. The goodness-of-fit test of the model shows that the R value is 0.85, which shows that the model has strong explanatory power. In addition, through the correlation analysis of independent variables, it is found that there is a positive correlation between housing area and supporting facilities, but the correlation with other variables is weak. This study not only enriches the theoretical research of the real estate market, but also provides valuable decision-making reference for the government, enterprises and consumers, and helps the healthy and stable development of the real estate market in Tianjin.

Key words: Tianjin; real estate market; multiple linear regression.

1. Introduction

As an important municipality directly under the central government and economic center of China, the development of Tianjin's real estate market has an important impact on the regional and even the national economy. In recent years, with the acceleration of urbanization and the improvement of residents' living standards, the real estate market in Tianjin has shown a vigorous development trend, attracting a large number of investors and residents' attention. However, the volatility and uncertainty of the real estate market also bring a lot of risks to relevant stakeholders [1]. Therefore, in-depth analysis and accurate prediction of Tianjin real estate market will not only help the government and enterprises to make scientific and reasonable decisions, but also provide consumers with more transparent market information.

The purpose of this study is to make a systematic analysis of Tianjin real estate market by using the method of multiple linear regression prediction. As a statistical method, multiple linear regression can explore the relationship between multiple independent variables and dependent variables, and help us understand the influencing factors of house prices more comprehensively. This study not only has theoretical significance, but also has practical application value. Through scientific forecasting methods, we can better grasp the market dynamics and contribute to the healthy and stable development of Tianjin real estate market.

2. Research methods and data sources

2.1. Basic principle of multivariate linear regression prediction method

In this study, multiple linear regression prediction method is used to analyze the real estate market in Tianjin [2]. Multiple linear regression is a statistical prediction analysis method, which is used to study the linear relationship between a dependent variable and multiple independent variables. Its basic principle is to describe the linear relationship between the dependent variable (house price) and several independent variables (house area, geographical location, house age, etc.) by establishing a mathematical model, so as to predict the changing trend of the dependent variable [3].

In the analysis of the real estate market, the house price is usually selected as the dependent variable [4]. Select multiple factors that may be related to the house price as independent variables, such as house area, house age, geographical location and supporting facilities [5-6]. Collect relevant data through various channels, and carry out necessary cleaning and pretreatment on the data.

Based on the collected data, a multiple linear regression model is constructed. The mathematical form of the model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{1}$$

Where Y is the dependent variable, $X_1, X_2, \cdots, X_n$ is the independent variable, $\beta_0, \beta_1, \cdots, \beta_n$ is the regression coefficient, and ε is the error term.

The model is fitted by statistical software, and the regression coefficient is estimated by least square method. After getting the estimated value of regression coefficient, explain the influence degree and direction of each independent variable on the dependent variable. The goodness of fit test is carried out to evaluate the prediction ability of the model. The significance test was conducted to determine whether the independent variable significantly affected the dependent variable. The model is optimized according to the test results. The fitted model is used to predict, and the new independent variable value is input to get the predicted value of the dependent variable.

2.2. Data source

The data of this study mainly comes from the real estate transaction data in Tianjin. In order to ensure the timeliness and accuracy of the study, the transaction data of the past five years are selected, and the time range is from 2018 to 2022. These data cover the real estate transaction records of various regions in Tianjin, including the basic information of houses, transaction price, transaction time and other key elements. In the data processing stage, all sources of data are integrated, cleaned and standardized to ensure the quality and consistency of data.

3. Establishment and test of multiple linear regression model

3.1. Variable selection

According to the actual situation of the market, housing area (X1), housing age (X2), geographical location (X3) and supporting facilities (X4) are selected as the main influencing factors. Among them, the housing area is one of the important factors affecting housing prices. Usually, the larger the area, the higher the housing prices. The age of the house can reflect the old and new degree of the house, and the price of the new house is often higher than that of the old house; Geographical location uses the distance from the city center as a quantitative index, and the closer the distance, the higher the house price is usually; Supporting facilities, including surrounding schools, hospitals, shopping centers, etc., also have a significant impact on housing prices, which are quantified by comprehensive scoring method.

3.2. Model building

Based on the independent variables selected above, the following multiple linear regression models are constructed:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (2)$$

Where: Y stands for house price (dependent variable). $X_1 \sim X_4$ represents housing area, housing age, geographical location and supporting facilities (independent variables) respectively. β_0 is the intercept term, which indicates the predicted value of the dependent variable when all independent variables are 0. $\beta_1 \sim \beta_4$ is the regression coefficient of each variable, indicating the average change of the dependent variable when the independent variable increases by one unit. ε is an error term, which represents other influencing factors not considered in the model.

3.3. Model testing

Through the statistical software SPSS, the goodness of fit test is $R^2 = 0.85$, and the adjustment is $R^2 = 0.83$, which shows that the model explains 85% of the variation of house prices, and the fitting effect is good. Significance test (F test) F value = 75.00, P value < 0.001, indicating that there is a significant linear relationship between at least one independent variable and the dependent variable. The p values of all independent variables are less than 0.05, which indicates that the influence of each variable on house prices is statistically significant.

Through the above tests, it is confirmed that housing area, housing age, geographical location and supporting facilities are the key factors affecting housing prices in Tianjin. Specifically, the increase of housing area can significantly push up housing prices, on the contrary, with the increase of housing age, housing prices will decline. In addition, the advantages of geographical location, such as being closer to the city center and the perfection of supporting facilities, have a significant positive impact on raising housing prices. On the whole, the constructed multiple linear regression model shows good fitting degree and strong explanatory ability.

4. Results and discussion

4.1. Prediction result

Multiple linear regression model reveals that housing area and supporting facilities have a significant positive impact on housing prices, while housing age and distance from the city center have a significant negative impact on housing prices. The regression coefficient of the building area (X1) is 450, which means that with other variables unchanged, the house price is expected to rise by 450 yuan for every square meter of building area. The t value of this variable is high (12.86), and the p value is less than 0.001, which indicates that the influence of housing area on housing price is very significant. The regression coefficient of the room age (X2) is -200, which indicates that with the increase of the room age, the house price is expected to decrease by 200 yuan. Although its influence is smaller than the housing area, the P value less than 0.01 still shows the significant influence of the housing age on the housing price. Geographical location (X3) takes the distance from the city center as the quantitative index, and its regression coefficient is -800, which means that the house price is expected to drop by 800 yuan for every kilometer away from the city center. The t value and p value of this variable also show that the geographical location has a significant negative impact on housing prices. The regression coefficient of the supporting facilities (X4) is 1200, which means that the house price is expected to rise by 1200 yuan for every 1 point increase in the supporting facilities score. The high T value (8.00) and low P value (< 0.001) of this variable highlight the important positive impact of supporting facilities on housing prices. See Table 1.

Table 1 Prediction results of multivariate linear regression model

variable	B	SE	T value	P value
Intercept	50000	15000	3.33	<0.01
Housing area (X1) (per square meter)	450	35	12.86	<0.001
Room age (X2) (per year)	-200	60	-3.33	<0.01
Geographical location (X3) (per kilometer)	-800	120	-6.67	<0.001
Supporting facilities (X4) (each score)	1200	150	8.00	<0.001

4.2. Independent variable correlation analysis

The correlation between independent variables is also considered in the process of model construction. By calculating the correlation coefficient matrix, it is found that there is a certain degree of positive correlation between housing area and supporting facilities, which may be because larger houses are often equipped with more perfect facilities. The correlation between other independent variables is weak, which will not cause serious multicollinearity problems to the model. See Table 2.

Table 2 Correlation coefficient matrix

variable	Housing area (X1)	Room age (X2)	location (X3)	Supporting facilities (X4)
Housing area (X1)	1.00	-0.10	-0.20	0.40
Room age (X2)	-0.10	1.00	0.05	-0.05
Geographical location (X3)	-0.20	0.05	1.00	-0.15
Supporting facilities (X4)	0.40	-0.05	-0.15	1.00

There is a certain degree of positive correlation between the housing area (X1) and the supporting facilities (X4), and the correlation coefficient is 0.40. This means that larger houses are often equipped with better facilities. This positive correlation is reasonable in the real estate market, because large houses usually have more space and resources to provide comprehensive supporting facilities, thus improving the convenience and comfort of living.

There is a certain negative correlation between the housing area (X1) and the geographical location (X3), and the correlation coefficient is -0.20. This may indicate that the housing area in Tianjin, which is closer to the city center, may be relatively small due to limited land resources. On the contrary, far away from the city center, land resources are more abundant, so the housing area may be larger. However, this negative correlation is relatively weak, indicating that the relationship between area and geographical location is not absolute.

Room age (X2), its correlation with other variables is very weak. Especially, the correlation coefficients with the building area (X1) and supporting facilities (X4) are only -0.10 and -0.05, which are almost negligible. This means that the house age may be a relatively independent variable, which is less affected by other factors. In fact, the age of the house mainly reflects the old and new degree of the house, which is not directly related to the size of the house and the perfection of supporting facilities.

Finally, there is a certain degree of negative correlation between geographical location (X3) and supporting facilities (X4), and the correlation coefficient is -0.15. This may indicate that in Tianjin, in areas far from the city center, supporting facilities may be relatively few or not perfect. However,

this negative correlation is relatively weak, indicating that the relationship between geographical location and supporting facilities is not static.

The multiple linear regression model in this study is suitable for forecasting the housing price in Tianjin under similar market conditions. However, the model also has some limitations. First of all, the model assumes that there is a linear relationship between independent variables and dependent variables, which may not be fully established in practical application. Secondly, the model does not consider all the factors that may affect housing prices, such as market supply and demand, policy changes and so on. Therefore, it is necessary to combine other information to make a comprehensive judgment in practical application.

To sum up, this study successfully predicted the housing price in Tianjin through multiple linear regression model, and analyzed the influence degree of their respective variables on housing prices and their correlation. Although the model has certain forecasting ability and application scope, it still needs to pay attention to its limitations and consider them in practical application.

5. Conclusion

In this study, the real estate market in Tianjin is systematically analyzed by using the method of multiple linear regression prediction. The results show that housing area, housing age, geographical location and supporting facilities are the key factors affecting housing prices in Tianjin. Among them, housing area and supporting facilities have a significant positive impact on housing prices, while housing age and distance from the city center have a significant negative impact on housing prices. In addition, the correlation analysis between independent variables shows that there is a certain degree of positive correlation between housing area and supporting facilities, while there is a certain degree of negative correlation between housing area and geographical location. These findings provide strong support for the healthy and stable development of Tianjin real estate market. Although the multiple linear regression model has certain forecasting ability and applicable scope, it still needs to pay attention to its limitations and consider them in practical application. First of all, the model assumes that there is a linear relationship between independent variables and dependent variables, which may not be fully established in practical application. Therefore, when using this model to predict, it is necessary to combine other information to make a comprehensive judgment. Secondly, the model does not consider all the factors that may affect housing prices, such as market supply and demand, policy changes and so on. Therefore, in practical application, it is necessary to further collect and sort out relevant data in order to predict house prices more accurately. Finally, in order to improve the accuracy of forecasting, other more advanced forecasting methods, such as machine learning and artificial intelligence, are considered, so as to provide more comprehensive and accurate decision-making basis for the development of Tianjin real estate market.

References

[1] Su Zhi. (2016). Short-term correlation between public expectations and the real estate market - An empirical study based on microblog information. Economic and Management Research, 37(3), 8.

[2] Wu Tingting, Hu Wenxiu, & Zhao Fan. (2018). Dynamic early warning research on bubble economic crises – Based on international experience data from the real estate market. Forecasting, 37(3), 7.

[3] Cui Mingming, Liu Xiaoting, Li Xiuting, & Dong Jichang. (2020). Integrated forecasting of the real estate market driven by data characteristics. Management Review, 32(7), 13.

[4] Wang Shaofen, & Cai Chengbin. (2020). Grey relational analysis of the impact of population structure changes on real estate market demand. Practice and Cognition of Mathematics, 50(9), 5.

[5] Liu Xiaojun, Hu Shengkai, & Chi Yihan. (2021). Real estate price prediction research based on var—gm(1.1)—svr model. Practice and Cognition of Mathematics, 51(1), 12.

[6] Shao Weishuang, Li Xiaohong, Zhang Tianshu, & Wang Yan. (2020). Application research of data mining in real estate price prediction. Practice and Cognition of Mathematics, 50(5), 6.