Research Article

Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model

Guangjie Liu

School of Management, Changchun Institute of Technology, Changchun 130021, China

Correspondence should be addressed to Guangjie Liu; liuguangjie@ccit.edu.cn

Received 22 January 2022; Revised 13 April 2022; Accepted 16 April 2022; Published 9 May 2022

Academic Editor: Sheng Bin

Copyright © 2022 Guangjie Liu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Since the housing reform in the 1990s, China’s real estate market has expanded and developed rapidly. It has quickly become the pillar of China’s national economy and made a great contribution to China’s economic growth. However, China’s real estate market started late and its development is not perfect, and house prices show obvious volatility. Accurate prediction of house prices is conducive to the government issuing appropriate regulatory policies, helping investors formulate correct investment strategies, and guiding the healthy and long-term development of the real estate market. Based on the statistical data of the real estate market, this study analyzes the influencing factors of the real estate price, establishes a multiple linear regression model, uses the least square method to solve the unknown parameters of the model, and finally constructs a house price prediction model to predict and analyze the real estate market. The results show that the maximum error of the real estate price prediction model is no more than 8%. The above results show that the house price prediction model has high accuracy and can predict house prices accurately and efficiently.

1. Introduction

In recent years, real estate prices have risen rapidly, which has attracted the attention of researchers in various fields. To some extent, the stability of the real estate market is closely related to the stability of China’s economic development, financial system, and society [1]. Forecasting and analyzing the real estate market can evaluate the stability of the real estate market. On the one hand, it can facilitate the government to carry out macrocontrol on house prices and maintain the healthy and stable development of China’s national economy. On the other hand, the prediction of house prices can also provide a certain basis for real estate investors to formulate investment strategies and avoid losses. Compared with developed countries such as Europe and the United States, China’s real estate market started relatively late and lacked market experience and policy theory, resulting in some deficiencies in policies related to the real estate market. Therefore, many experts conducted theoretical and empirical research on the development of the real estate market [2]. At present, there are many studies on the macrocontrol of the real estate market, but few studies on the prediction of the real estate market. Therefore, based on the statistical data of the real estate market, this study analyzes the influencing factors of the real estate price, establishes the multiple linear regression model, uses the least square method to solve the unknown parameters of the model, and finally constructs a house price prediction model to predict and analyze the real estate market, so as to provide data support for the macrocontrol and investment of the real estate market. It also plays a certain role in the stable development of the real estate market.

The main method of forecasting and analyzing the real estate market is to analyze the real estate data by using the multiple linear regression model and then draw a conclusion. The multiple linear regression model is a common multivariate statistical method. It has the advantages of being a convenient application and a simple model. Therefore, it is widely used in daily production and various scientific research studies. According to the real estate statistical data of the National Bureau of Statistics, combined with the methods of qualitative analysis and quantitative analysis, this study discusses the impact of various factors on the real estate price, such as policy factors, economic factors,
and housing supply factors, and finally constructs the real estate price prediction model to predict and analyze the real estate market.

There are three main innovations in the research: the first is to calculate and analyze the entropy and information gain of various influencing factors of house prices and extract the main factors affecting real estate price, the second is to build a real estate price trend model to facilitate the analysis of the relationship between the main factors, and finally, the multiple linear regression model is established according to the relationship between each factor and the real estate price, and the residual in the model is calculated by the least square method, so as to solve the unknown parameters.

Based on the statistical data of the real estate market, this study analyzes the influencing factors of the real estate price, establishes the multiple linear regression model, uses the least square method to solve the unknown parameters in the model, and finally constructs a house price prediction model to predict and analyze the real estate market. The study is divided into five parts. Section 1 explains that the stability of the real estate market is closely related to China’s economic development, financial system, and social stability. Section 2 mainly summarizes and discusses the research results of scholars on the real estate market in recent years. Section 3 discusses the factors affecting the real estate price and their relationship and establishes a multiple linear regression model to predict and analyze the real estate market. Section 4 is the performance test and analysis of the real estate price prediction model. Section 5 is the summary of the whole research.

2. Related Works

Since the reform and opening-up, China’s national economy has developed rapidly, and the people’s quality of life has also been greatly improved. Real estate is an important industry to stimulate China’s national economy and promote GDP growth, which has been highly concerned by people from all walks of life. Lahav et al. analyzed the large-scale rise in house prices in Israel and believed that changes in the real estate market would impact the stock market [3]. Pirogova et al. studied the driving force of coordinated growth of the real estate market under the condition of digitization [4]. Dumeignil et al. conducted a natural experiment to explore the impact of cross-border labor mobility on real estate price trends [5]. Chernyshova et al. studied the relationship between supply and demand, which is the real estate price formed under the influence of social, economic, and material factors, and predicted the real estate price based on this [6]. Rakhman et al. made theoretical provisions on the specific situation of the real estate market in the Kharkov region and analyzed the changes in its market structure and the dynamic fluctuations of prices [7]. Lee et al. used three machine learning models to build the real estate index prediction model, compared the performance of the three models, and concluded that the real estate index prediction model based on the random forest has high accuracy [8]. Based on the principles of common law, Nwogugu discussed the constitutional, competition law, and economic psychology inherent in the real estate market mechanism [9]. Kang et al. constructed a short-term prediction model of apartment prices based on news article keyword search frequency and machine learning technology and used relevant data for training and testing. The experimental results show that the prediction accuracy of the model can meet the expectation [10].

Baillif et al. studied the case of western Switzerland and proposed a discrete mixed market characteristic pricing model to estimate and predict real estate prices [11]. Saeed studied the impact of green space density on real estate prices in Ramadi city by taking green space area, distance from green space, time to enter the park and public green space, and percentage of urban green space as indicators [12]. Kang et al. constructed the real estate auction price prediction model by using the regression model, artificial neural network, and genetic algorithm, respectively, and tested the accuracy of the model by using real estate auction data in Seoul. The test results show that the real estate auction price prediction model based on a genetic algorithm can be segmented according to the effective area of the auction appraisal price. This further improves the prediction accuracy [13]. Based on the perspective of the world economy, Jaymin discussed the impact of coronavirus disease on India’s real estate industry and the risks and prospects faced by real estate industry participants [14]. Luo studied the relationship between higher education resource allocation and real estate prices [15].

It can be seen from the above research results that there are many studies on the real estate market. Scholars from various countries have predicted and analyzed the real estate market by using machine learning technology or big data analysis technology, but there is little literature to analyze and empirically study the influencing factors of real estate prices. Therefore, the research is based on the statistical data of the real estate market. Based on the statistical data of the real estate market, this study presents an empirical analysis on the influencing factors of the real estate price, establishes a multiple linear regression model, uses the least square method to solve the unknown parameters of the model, and finally constructs a house price prediction model to predict and analyze the real estate market.

3. House Price Forecasting Method Based on the Multiple Linear Regression Model

3.1. Extraction Method of Main Factors Based on the Information Gain Method. The real estate market is closely related to the real estate price. If we predict and analyze the real estate price, we can predict and analyze the real estate market. Commodity prices are closely related to market supply and demand, and there are many index factors affecting commodity supply and demand, and there are complex links between these indexes. Therefore, in order to analyze the real estate price, we first need to select the important index factors that can affect the real estate price and carry out in-depth analysis and research. After investigating the current situation of the real estate market and sorting out the existing literature, the index factors affecting the real estate price are summarized, as shown in Figure 1.
It can be seen from Figure 1 that the factors affecting the real estate price are diverse and complex. The accuracy of the results obtained by using simple quantitative analysis or simple qualitative analysis to predict and analyze the real estate price will be very low and have no reference value. Therefore, it is necessary to extract the main factors affecting the real estate price from these influencing factors and conduct quantitative analysis. Then, it provides reference data for the prediction and analysis of real estate prices. The mainstream principal factor extraction methods include principal component analysis, factor analysis, and information gain method. The information gain method can measure the uncertainty of the event attribute and then measure the correlation contained in the attribute. Therefore, the information gain method is used to extract the main influencing factors of real estate prices. The extraction process of main factors based on the information gain method is shown in Figure 2.

In Figure 2, the operation steps of main attribute extraction based on information gain method are as follows: collect l sample data and set the sample set to Q. If the class label contains n different attribute values, divide these values. Let the number of the ith class be Ci (i = 1, 2, . . ., n). Then, Li represents the number of Ci in the sample set. At this time, if the n samples are to be classified, the required theoretical expectation information can be expressed as

$$I(L_1, L_2, \ldots, L_n) = - \sum_{i=1}^{n} p_i \log p_i,$$  \hspace{1cm} (1)

where $p_i$ represents the probability that any sample value falls into class $C_i$, and its calculation method is shown as

$$p_i = \frac{L_i}{L},$$  \hspace{1cm} (2)

If the attribute $y$ has $k$ different values, denoted as $\{a_1, a_2, \ldots, a_k\}$, the sample dataset can be divided into $k$ subsets by using the attribute $y$, that is,

$$Q = \{Q_1, Q_2, \ldots, Q_k\},$$  \hspace{1cm} (3)

where $Q_j$ contains all samples whose value is $a_j$ on attribute $y$. If $L_{ij}$ is used to represent the number of samples whose
value of 1 in \( Q_j \) falls into class \( C_j \), the entropy divided according to \( y \) can be expressed as

\[
E(y) = -\sum_{j=1}^{k} \frac{L_{ij} + L_{2j} + \cdots + L_{nj}}{L} \log \frac{L_{ij} + L_{2j} + \cdots + L_{nj}}{L},
\]

(4)

where \( L_{ij} + L_{2j} + \cdots + L_{nj}/L \) represents the probability weight of \( Q_j \), which is the ratio of the number of all samples with the value of \( a_j \) in attribute \( y \) in \( Q_j \) to the total number of samples in \( Q \). According to formula (1), the expected information in subset \( Q_j \) can be expressed as

\[
I(Q_1, Q_2, \ldots, Q_n) = -\sum_{j=1}^{n} p_{ij} \log p_{ij},
\]

(5)

where \( p_{ij} \) represents the probability weight that the sample in subset \( Q_j \) belongs to class \( C_j \), and its calculation method is shown as

\[
p_{ij} = \frac{L_{ij}}{L_j}.
\]

(6)

Combining formulas (1) and (5), the information gained on attribute information \( y \) can be obtained:

\[
\text{Gain}(y) = I(Q_1, Q_2, \ldots, Q_n) - E(y).
\]

(7)

According to formula (7), when the entropy value of attribute information is larger, the information gain corresponding to the attribute information is smaller because the entropy value contains the uncertainty of the attribute. In order to ensure the accuracy of subsequent experiments and tests, the attribute with information gain greater than 0.3 is selected as the main factor. The influencing factors affecting the real estate price, i.e., monthly income level of residents, per capita disposable income of residents, per capita housing expenditure of residents and completed area of real estate, are set as \( x_1, x_2, x_3, x_4 \), respectively, and the sales price of real estate is set as \( E \). The information gain corresponding to each influencing factor can be calculated according to the above contents. In order to avoid the lack of centralized distribution of statistical data, the collected data are transformed to show it in the form of 0 and 1, and the formula between transformation rules is

\[
b_{ij} = \begin{cases} 
1 & a_{ij} \geq \frac{a_{i1} + a_{i2} + \cdots + a_{inj}}{n} \\
0 & a_{ij} < \frac{a_{i1} + a_{i2} + \cdots + a_{inj}}{n} 
\end{cases} \quad i = 1, 2, \ldots, L, \ j = 1, 2, \ldots, M,
\]

(8)

where \( a_j \) represents the \( j \) unknown parameter.

3.2. Construction of the House Price Prediction Model Based on Multiple Linear Regression. Among the many factors affecting house prices, some are uncontrollable factors, such as economic growth, and its impact mechanism on real estate prices is shown in Figure 3.

As shown in Figure 3, the impact mechanism of these uncontrollable factors on the growth of real estate prices is very complex. Therefore, in the quantitative analysis of these factors, assumptions need to be made to eliminate the impact of these factors on the research results, such as the price of the target city remains stable. On this basis, we can build a model to predict and analyze real estate prices. The multiple linear regression model can solve the target factors containing a variety of attribute information and make statistics and analysis on the historical data distribution laws of various influencing factors. Therefore, it can well reflect the time-change characteristics of all factors in the model and is suitable for the prediction of real estate prices. The real estate price prediction process based on the multiple regression linear model is as follows: if the extracted main factor is \( x_1 \sim x_4 \), the regression model for quantitative analysis of real estate price is constructed, as shown in the following formula:

\[
E = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + \varphi,
\]

(9)

where \( \varphi \) is an unknown parameter independent of the extraction of the main factor; that is, the random variable introduced by the multiple linear regression model. In the prediction and analysis of real estate prices, because the data collected in the study are from different cities, there is a large gap in various attributes, and the data need to be normalized, as shown in the following formula:

\[
E - \bar{E} = a_1 (x_1 - \bar{x}_1) + a_2 (x_2 - \bar{x}_2) + \cdots + a_4 (x_4 - \bar{x}_4),
\]

(10)

where \( \bar{E} \) is the average house price and \( \bar{x}_i \) is the average value of the \( i \) influencing factor. Formula (10) can also be expressed as the following formula:

\[
\Delta E = a_1 \Delta x_1 + a_2 \Delta x_2 + \cdots + a_4 \Delta x_4,
\]

(11)

where \( \Delta x_i \) represents the data centralization of the \( i \) influencing factor and \( \Delta E \) represents the real estate price centralization. Set a matrix vector:

\[
a = (a_1, a_2, a_3, a_4)^T.
\]

(12)

In order to improve the accuracy of the real estate price prediction model and enable it to correctly reflect the time characteristics of house prices, the matrix vector of formula (12) needs to be estimated by least squares in the multiple linear regression model. In addition, it is necessary to minimize the cumulative square sum \( S(a) \) of the residual value \( \varphi \) in the time domain. The calculation method of \( S(a) \) is shown in the following formula:

\[
S(a) = \sum_{i=1}^{n} (\Delta E_i - a_1 \Delta x_{1i} - a_2 \Delta x_{2i} - a_3 \Delta x_{3i} - a_4 \Delta x_{4i})^2,
\]

(13)

where \( t \) represents time series. Find the partial derivative of \( a_1 \sim a_4 \) in formula (13) and introduce matrices \( X, A, B \), as shown in the following formula:
Based on the above, the solution of the unknown parameter matrix can be obtained, as shown in the following formula:

\[
A = X^{-1}B
\]

where \(Cov(i, j)\) represents the covariance between numerical sequences of centralized data. Based on the above, the solution of the unknown parameter matrix can be obtained, as shown in the following formula:

\[
A = \frac{B}{X}
\]  

According to formula (15), the unknown variables can be solved, and then, the location variables can be substituted into the model. Combined with the fitting curve of the main influencing factors extracted based on the information gain method, the quantitative analysis and prediction of the future real estate price trend can be realized. Among them, the fitting curves of the main influencing factors need to be analyzed according to the actual data, so as to find the curve with the smallest error, so that it can better reflect the time variation law of the influencing factors. Based on the above contents, the real estate price prediction model can be constructed based on the main factor extraction and multiple linear regression model, and then, the real estate market can be predicted and analyzed.

4. Performance Analysis of the Real Estate Price Prediction Model

4.1. Correlation Analysis between Main Influencing Factors and House Price. There are many factors affecting real estate prices, such as market supply and demand, wage income, and economic development level. There is a complex relationship between these influencing factors. If all the influencing factors were substituted into the model, the amount of calculation would be very large and the accuracy of the model would be reduced. Therefore, it is necessary to extract the principal component. The research uses the information gain method to extract the principal components, selects the real estate data of each city from 2010 to 2019 counted by the National Bureau of Statistics for empirical analysis, and selects four factors with an information gain value greater than 0.3 as the main factors, namely, residents’ monthly income level, residents’ per capita disposable income, residents’ per capita housing expenditure, real estate completed area, etc., which are set as \(x_1, x_2, x_3, x_4\), respectively. The correlation between each main factor and real estate price is shown in Figure 4.

It is easy to see from Figure 4 that the monthly income level of residents is significantly positively correlated with the real estate price, \(r = 0.851\); that is, the higher the monthly income level of urban residents, the higher the real estate price of the city and the better the real estate market of the city. There is a significant positive correlation between residents’ per capita disposable income and real estate price, \(r = 0.764\), indicating that the higher the per capita disposable income level of urban residents, the higher the real estate price of the city and the better the real estate market of the city. The per capita housing expenditure of residents is
positively correlated with the real estate price, $r = 0.517$, which is lower than the first two factors. The completed area of real estate is negatively correlated with the real estate price, $r = 0.612$, indicating that the larger the completed area of real estate in the city, the lower the real estate price in the city. This is because the completed area of real estate will affect the relationship between supply and demand in the real estate market, making it slowly change to the trend of oversupply, thus reducing the real estate price. The above results show that the four main factors extracted in the study have a great correlation with the real estate price, indicating that the principal component extraction method based on the information gain method can effectively extract the main influencing factors of the real estate price. Based on the above, when the government regulates house prices, issues relevant policies, or investors want to invest, it needs to focus on the factors such as the monthly income level of residents, per capita disposable income of residents, per capita housing expenditure of residents, and completed area of real estate.

4.2. Fitting Curve Analysis Based on the Real Estate Price Prediction Model. Using the software SPSS 23.0 and the real estate market data of city B in recent years counted by the National Bureau of Statistics, the fitting curves of two influencing factors of real estate completed area and per capita disposable income of residents in city B and the fitting curve of real estate price in city B are drawn, as shown in Figure 5.

It is easy to see from Figure 5(a) that the completed real estate area in city B is decreasing year by year, which shows that the real estate market in city B is gradually saturated, but generally speaking, the completed real estate area is increasing. According to the correlation analysis of the above contents, the increase in the completed real estate area may either lead to the decline of house prices in city B or cause the growth rate of real estate prices in city B to slow down. As can be seen from Figure 5(b), the per capita disposable income of residents in city B is increasing year by year, which may lead to the growth of house prices in city B in the future.

As can be seen from Figure 6, the house price of city B will still show an upward trend in the next few years, which is consistent with the above content. In view of the future growth of house prices, the government of city B should issue corresponding macrocontrol policies to prevent the rapid growth of real estate prices from damaging the stability.
of the real estate market, maintain the stable development of the real estate market, and ensure the steady construction in the new period.

4.3. Accuracy Analysis of the Real Estate Price Prediction Model. According to the real estate data before 2010, the real estate prediction model was used to predict the house prices of four cities (A, B, C, and D) from 2011 to 2017, and the results were verified by the real estate market data of four cities from 2011 to 2017 counted by the National Bureau of Statistics, so as to verify the prediction accuracy of the real estate price prediction model. The prediction error percentage of the prediction model is shown in Figure 7.

As can be seen in Figure 7, the prediction errors of the real estate price prediction models constructed in the study are small. Among them, the maximum prediction error is the prediction of city A and city D in 2014, and the errors between the predicted value and the real value are 7.6% and −5.9%, respectively. The minimum prediction error is the prediction of city A and city B in 2011. The errors between the predicted value and the real value are −0.2% and 1.8%, respectively. Figure 7 shows the real estate market data of four cities from 2011 to 2017 from the National Bureau of Statistics. It can be seen that the maximum error of the real estate price prediction model is no more than 8%, and the prediction error is very small, indicating that the real estate prediction model can more accurately predict the changes of real estate prices so as to provide data support for real estate investment and government real estate price regulation.

5. Conclusion

Real estate is an important industry to stimulate China’s national economy and promote GDP growth. Predicting and analyzing real estate prices can help the government issue appropriate regulatory policies, help investors formulate
correct investment strategies, and guide the healthy and long-term development of the real estate market. Based on the current situation of the real estate market and the existing literature, this study summarizes the influencing factors of the real estate price, extracts four main factors of the real estate price by using the information gain method, namely, the monthly income level of residents, the per capita disposable income of residents, the per capita housing expenditure of residents, and the completed area of real estate, and constructs the real estate price prediction model by using the multiple linear regression model. The results show that there is a significant positive correlation between the residents’ monthly income level and the real estate price, \( r = 0.851 \); there is a significant positive correlation between per capita disposable income and the real estate price, \( r = 0.764 \); the per capita housing expenditure of residents is positively correlated with the real estate price, \( r = 0.517 \); the completed area of real estate is negatively correlated with the real estate price, \( r = 0.612 \). The accuracy of the real estate price prediction model is tested by using the real estate data of four cities before 2020. The test results show that the maximum prediction error of the real estate price prediction model is 7.6%; the minimum prediction error is –0.2%. The above results show that the prediction model can effectively predict the real estate price and then predict the real estate market and has high practicability. The study did not discuss the impact of national policies on house prices, which needs further research in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This work was supported by “Science and Technology Department of Jilin Province, China (Grant no. 20170418046FG).”

References

[1] V. Yodgorov and D. Mirdjalilova, “Priorities of management of real estate objects by service organizations in the republic of Uzbekistan and forecast parameters of performance indicators,” Bulletin of Science and Practice, vol. 6, no. 12, pp. 313–320, 2020.
[2] X. Chen and J. Tongurai, “The relationship between China’s real estate market and industrial metals futures market: evidence from non-price measures of the real estate market,” Asia-Pacific Financial Markets, vol. 2021, no. 2, pp. 1–35, 2021.
[3] X. Li, M. Sun, and K. Boersma, “Policy spillover and regional linkage characteristics of the real estate market in China’s urban agglomerations,” Journal of Management Science and Engineering, vol. 4, no. 3, pp. 189–210, 2019.
[4] O. Pirogova and N. Temnova, “Dynamics of coworking growth in the real estate market under digitalization,” E3S Web of Conferences, vol. 244, no. 3, Article ID 10052, 2021.
[5] C. Dumeignil, “The impact of cross-border labor mobility on real estate price trends: a natural experiment,” International Regional Science Review, vol. 45, no. 2, pp. 108–132, 2021.
[6] M. Chernyshova, A. Malenkaya, and T. Mezhuyeva, “Analysis of pricing factors in real estate market,” Interexpo GEO-Siberia, vol. 6, no. 2, pp. 79–85, 2019.
[7] M. S. Rakhman, V. V. Malko, and V. V. Malko, “Structural changes in the real estate market in Kharkiv region,” Business Inform, vol. 10, no. 501, pp. 146–143, 2019.
[8] J. M. Lee, S. H. Park, S. H. Cho, and J. H. Kim, “Comparison of models to forecast real estates index introducing machine learning,” Journal of the Architectural Institute of Korea Structure and Construction, vol. 37, no. 1, pp. 191–199, 2021.
[9] M. Nwogugu, “Some constitutional law, competition law and economic psychology issues inherent in some real estate market mechanisms,” Geopolitical risk, Sustainability and “Cross-Border Spillovers” in Emerging Markets, no. 1, pp. 137–206, 2021.
[10] B. Markey-Towler, “The competition and evolution of ideas in the public sphere: a new foundation for institutional theory,” Journal of Institutional Economics, vol. 15, no. 1, pp. 27–48, 2019.
[11] M. Baillif, M. de Lapparent, and E. Kazagli, “A hybrid approach to real estate price definition: a case study in western Switzerland,” Revue Économique, vol. 72, no. 6, pp. 1055–1077, 2021.
[12] A. Saeed and L. Mullahwaish, “Effect of green areas density on real estate price in Ramadi city,” International Journal of Design & Nature and Ecodynamics, vol. 15, no. 2, pp. 253–259, 2020.
[13] J. Kang, H. J. Lee, S. H. Jeong, H. S. Lee, and K. J. Oh, “Developing a forecasting model for real estate auction prices using artificial intelligence,” Sustainability, vol. 12, no. 7, pp. 1–19, 2020.
[14] S. R. Jaymin, “The effect of covid-19 on the real estate industry in India,” Baltic Journal of Real Estate Economics and Construction Management, vol. 9, no. 1, pp. 122–129, 2021.
[15] H. Luo, “Research on the interaction between higher education resource allocation and real estate price,” Open Journal of Social Sciences, vol. 08, no. 4, pp. 58–68, 2020.